Application of chemometric and experimental tools for monitoring processes of industrial importance

by

Kaushik Sivaramakrishnan

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

Department of Chemical and Materials Engineering
University of Alberta

Abstract

The common theme of this work was to investigate the relationships between multi-dimensional experimental data and the physical and chemical properties of certain reacting systems that have industrial relevance. Three particular systems of increasing complexity in terms of composition of the reaction mixture were considered, namely: catalytic oligomerization of an aliphatic olefin (propylene), oxidation of a naphthene-aromatic (tetralin) and thermal conversion of Canadian oilsands bitumen, which is a mixture of a compound classes. The overall objective of my thesis was to evaluate the statistical and analytical information obtained from a variety of chemometric and experimental tools to understand the chemical behavior of each system from a reaction chemistry viewpoint, with simultaneously validating the potential for online monitoring. The entire research conducted in four phases and though the first two studies were independent from those on bitumen, all works were aimed at addressing gaps in related literature.

Propylene, obtained as a by-product of pyrolysis of organic material is usually converted to heavier hydrocarbons in the gasoline or distillate boiling range. In the first phase of my research, the ability of nonlinear least squares-support vector regression (LS-SVR) adopting kernel transformations to predict concentrations of hydrocarbon products at the outlet of a flow reactor from H-ZSM-5 catalyzed propylene oligomerization was assessed in comparison with linear partial least squares (PLSR) and its variant, interval-partial least squares regression (*i*-PLSR). The classification of product concentrations through dimension reduction and pattern recognition techniques like principal component analysis (PCA) and hierarchical clustering analysis (HCA) reflected different elements of the oligomerization chemistry quite well. The possibility of

eliminating offline concentration measurement and online tracking of product concentration distribution was the key contribution of this work.

The role of mass transfer was considered to be important in the liquid-phase oxidation of hydrocarbons to produce essential petrochemicals. In the second phase, the use of simple and multiple linear regression in identifying the most significant parameter related to oxygen availability in the tetralin oxidation process was investigated. It was conducted in a microfluidic reactor under Taylor flow conditions, where the relative importance of mass transfer over hydrodynamic parameters in controlling product selectivity was established. The challenge was to tackle the multicollinearity that was detected among the explanatory variables through F-tests, t-tests, partial and Pearson correlations, changes in regression coefficient estimates and their standard errors. A pathway to predict product selectivity and tetralin conversion from the only user-controlled variable (inlet flowrate of tetralin) was developed to potentially eliminate the use of gas chromatography for offline measurements of the outputs.

Thermal conversion of Athabasca bitumen was studied in detail with the intention of providing scientific reasoning for the peculiar increases in viscosity observed by certain previous researchers at less severe reaction conditions than employed in industry. A number of product characterizations by straightforward and hyphenated experimental techniques including vibrational and nuclear magnetic resonance spectroscopy, elemental analyses, distillation profiles, compositional analysis of gaseous products density and refractive indices were conducted. A comprehensive review and analysis of the plausible factors that could have affected the viscosity was provided. The third phase of my research was critical in identifying that post-thermal reaction procedures like solvent extraction and rheological conditions of viscosity measurement had a

major impact on viscosity. This study highlighted some detrimental effects of the presence of halogens in bitumen and was important for simulation of partial upgrading at the laboratory-scale.

The final study of my thesis comprised of deriving a credible reaction pathway from the quantitative statistical results of multivariate curve resolution and Bayesian clustering and learning methods on infrared spectra of the products from thermal conversion of Athabasca bitumen over a range of temperatures from 300 – 420 ºC. Minimum external chemical knowledge was used for the chemometric part, while the consistency of the conversion chemistry along with the proposed reaction mechanisms as indicated from the chemometric results was inspected and compared with that of Cold Lake bitumen as well. In summary, explanations for chemical changes in each of the three systems were derived from process data in different forms like spectra and process conditions by means of chemometric and experimental tools with a potential for online control.

Keywords: Support vector regression, interval-partial least squares, acid-catalyzed olefin oligomerization, hydrocarbon oxidation, mass transfer parameters, multicollinearity, multiple linear regression, viscosity trends, solvent extraction, rheological conditions, multivariate curve resolution, Bayesian structure learning, reaction chemistry.

Preface

(Mandatory due to collaborative work)

Most parts of Chapter 2 of this thesis were published as a section of a review article as "Sivaramakrishnan, K.; Puliyanda, A.; Tefera, D. T.; Ganesh, A.; Thirumalaivasan, S.; Prasad, V. Perspective on the Impact of Process Systems Engineering on Reaction Engineering. *Ind. Eng. Chem. Res.* Article ASAP, April 4th **2019**". This chapter titled "Relevance of chemometrics as a process-systems engineering technique with implications in analytical chemistry" was my original work. The submission of the article was invited as a part of the Sirish Shah Festschrift special issue and different sections were contributed by the other co-authors. Vinay Prasad was the corresponding and supervisory author for the article and was also involved in concept formation and composition of the chapter.

Chapter 3 of this thesis was published as "Sivaramakrishnan, K.; Nie, J.; De Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, 57, 13156 – 13176". I was responsible for concept formation, complete statistical analysis of the experimental data, subsequent chemical interpretation and entire manuscript composition as well. I was also the corresponding author. Jinjun Nie had performed all the oligomerization experiments and collected the data previously. He also granted permission to use his data for this work and my thesis. Vinay Prasad acted as the supervisory author and was involved in idea generation, concept formation, data interpretation and manuscript composition. Arno de Klerk was involved in the chemical interpretation and manuscript composition/edition as well.

Chapter 4 of this thesis was published as "Siddiquee, M. N.; Sivaramakrishnan, K.; Wu, Y.; De Klerk, A.; Nazemifard, N. A statistical approach dealing with multicollinearity among predictors in microfluidic reactor operation to control liquid-phase oxidation selectivity. *Reaction Chemistry & Engineering* **2018**, 3, 972 – 990". I was responsible for the complete data analysis, its interpretation and manuscript composition. The experiments and data collection for this paper

were performed by Muhammad Siddiquee and Yucheng Wu. Muhammad Siddiquee was also involved in concept formation and writing the experimental parts in the manuscript. The necessary co-author permissions have been acquired for inclusion of this work in my thesis. Arno de Klerk and Neda Nazemifard acted as the supervisory authors and were involved in interpretation of the regression models and manuscript composition/edition as well.

Chapter 5 of this thesis was published as a book chapter in the American Chemical Society Symposium Series volume as "Sivaramakrishnan, K.; De Klerk, A.; Prasad, V. Viscosity of Canadian Oilsands Bitumen and Its Modification by Thermal Conversion. In *Chemistry Solutions to Challenges in the Petroleum Industry*; Rahimi, P., Ovalles, C., Zhang, Y., Adams, J. J., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2019; Vol. 1320, pp 115–199". I was responsible for concept formation, experimental design and data collection, and manuscript composition. I presented a part of this work at the 255th ACS National Meeting, New Orleans in March, 2018 and was invited to contribute a chapter as a part of the book. Arno de Klerk acted as the supervisory author and was involved in concept formation, data interpretation and manuscript composition/edition. Vinay Prasad was also involved in manuscript composition/edition.

Chapter 6 of this thesis was submitted to *Reaction Chemistry & Engineering* as "Sivaramakrishnan, K.; Puliyanda, A.; De Klerk, A.; Prasad, V. A data-driven approach to generate pseudo-reaction networks for thermal conversion of Athabasca bitumen". I was responsible for the collection of experimental data, performing the chemometric analysis of multivariate curve resolution including local and global models, the full chemical interpretation and manuscript composition as well. Anjana Puliyanda carried out the Bayesian hierarchical clustering and structure learning for the data. Both Vinay Prasad and Arno de Klerk acted as the supervisory authors. Vinay Prasad was involved in the chemometric analysis and manuscript composition/edition. Arno de Klerk was involved in derivation of the reaction chemistry and manuscript composition/edition as well.

To my roommates, Amit and Hemant,
for always being there to motivate me.

Acknowledgements

I want to wholeheartedly thank my supervisors, Dr. Vinay Prasad and Dr. Arno de Klerk for guiding me throughout the course of my PhD, both academically and personally. Both of you have been and will be a constant inspiration in my life moving forward. I will remember the numerous discussions that we have had and your feedback has definitely added value to my scientific approach. I thank you for listening to my cries and supporting me during my hard times. I am so glad to have had the opportunity to work with you over the years.

I wish to thank my colleagues from both research groups that I have been involved with for your scientific suggestions, your encouragement, your positive and negative opinions of me as a person and just your amazing company. I have space to mention only a few names here but I have enjoyed sharing this journey with each and every one of you. Thanks for bearing with me being loud in the office. Lina, you were the first person I interacted with in Dr. Arno de Klerk's group and I want to thank you for training me on many of my experiments and equipment. Joy, I consider you as my sister from another mother and I hope with never stop cracking jokes on each other. My presentation skills improved significantly after discussions with Natalia and I am grateful to you for that. It was a wonderful experience to be in a lab managed by Cibele and Cloribel, whom I look upon as my mentors. It was a pleasure to work with Dereje and Anjana during the final phase of my research and getting to know you better. Thanks Ajay and Garima for those mouth-watering dishes that you cooked for me.

I cannot express in words how much I am indebted to my roommates Amit and Hemant Charaya for being there for me all the time. All those discussions on life are etched in my memory and I will carry them with me wherever I go. To my other two friends, Hemant Bansal, Prashant, and Ananthan you have been through my thick and thin and managed to put a smile on my face from time to time.

I want to thank my parents, my cousin Swathi and my aunt for their constant encouragement and backing my decision to do a PhD.

I would also like to thank Alberta Innovates for providing me financial support through the Alberta Innovates Graduate Student Scholarship for three years of my PhD.

CONTENTS

TABLES

concentrations. LV is the optimum number of latent variables for the PLSR, t is the constant term in the polynomial kernel (equation 3.4 in Chapter 3), d is the degree of the polynomial, γ is the regularization parameter in the LS-SVM cost function (equation 3.2 in Chapter 3) and σ^2 is the square of the width of the Gaussian RBF kernel (equation 3.5 in Chapter 3).

Table A.9. Summary of the number of factors in PLS and optimized intervals in *i*-PLS for combination 3.

Table A.10. Tuned hyper parameters for the LS-SVM calibration models and the RMSECV for combination 3 corresponding to the entire outlet product concentrations (Table 3.8 in Chapter 3). t is the constant term in the polynomial kernel (equation 3.4 in Chapter 3), d is the degree of the polynomial, γ is the regularization parameter in the LS-SVM cost function (equation 3.2 in Chapter 3) and σ^2 is the square of the width of the Gaussian RBF kernel (equation 3.5 in Chapter 3).

Table A.11. Results of combination 3: For each experiment in the validation set.

Table B.1. Parameter estimates, significances and output variances of SLR models with CR as the output and individual explanatory variables as the input. Compare the results with Table 4.7 in Chapter 4 that shows the parameter estimates for the MLR models.

Table B.2. Parameter estimates, significances and output variances of SLR models with S as the output and individual explanatory variables as the input. Compare the results with Table 4.9 in Chapter 4 that shows the parameter estimates for the corresponding MLR models.

Table B.3. Physicochemical properties of tetralin and oxygen at different experimental conditions.

Table B.4. FID response factors of various compounds.

Table B.5. Conversion data for oxidation of tetralin with air at 150 ºC conducted in a semi-batch reactor. [15]

Table D.1. Common strategies for inertia weight employed in the PSO literature.

Table D.2. LOF and R^2 values (% contribution to variance) on reconstruction of the original matrix after performing SVD for the datasets at 300 ºC, 350 ºC, 380 ºC and 420 ºC.

FIGURES

corresponding to each component in the system. These results correspond to data obtained at 400 ºC.

Figure 6.5. Plots of initial estimates of change in concentration of the three pseudo-components with process flow (reaction time in min) at the following temperatures: (a) 420 ºC; (b) 400 ºC; (c) 380 ºC; (d) 350 ºC.

Figure 6.6. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 300 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Figure 6.7. Methyl transfer from an isopropyl group attached to an aromatic (1) followed by hydrogen abstraction from the matrix leading to increased CH_2 content (compound (4)) possible occurring at 300 ºC.

Figure 6.8. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 350 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Figure 6.9. Sequence of reactions speculated to be occurring at 350 ºC based on SMCR results.

Figure 6.10. Pathway for 350 ºC showing the increase in mono-substituted aromatic content from a naphthene, keeping the di-substituted content constant. The bond dissociation energy (BDE) for homolytic cleavage of the indicated bonds is also shown in kJ/mol.

Figure 6.11. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 380 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Figure 6.12. Proposed mechanism corresponding to the changes in derived quantitative parameters observed at 380 ºC. The energies for homolytic bond cleavage of the C-C bonds in (15) and (16) are given in kJ/mol.

Figure 6.13. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 400 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Figure 6.14. Plausible type of reaction happening at 400 ºC where cracking of the weaker benzylic C-tertiary C (in compound (15)) followed by intramolecular hydrogen transfer and hydrogen abstraction to yield the mono-substituted aromatic (compound (24)) and the conjugated free radical (23). This can crack further to give lighter aliphatic products. Possibility of free-radical recombination to form compound (28) is also shown.

Figure 6.15. Plots of: (a) concentration profiles and (b) spectral profiles in the region 900 – 650 cm^{-1} for curve resolution applied on the 400 ºC dataset using 4 pseudo-components.

Figure 6.16. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 420 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Figure 6.17. Rate of convergence in terms of standard deviation of residual vs. number of iterations for ALS and ALS-PSO algorithms used in SMCR in this work.

Figure 6.18. Bayesian network structure produced by hill climbing and Tabu search algorithms describing the causal relationships between different groups clustered by BHC.

Figure 6.18. Bayesian network structure produced by hill climbing and Tabu search algorithms describing the causal relationships between different groups clustered by BHC.

Figure 6.20. Proposed reaction pathway for the thermal conversion of Athabasca bitumen based on the results from Bayesian network structure.

Figure 6.21. Proposed reaction pathway in continuation with Figure 6.20 for the thermal conversion of Athabasca bitumen based on the results from Bayesian network structure.

Figure 6.22. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at temperatures in the range 300 – 420 ºC (global model). The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and

Figure D.3. Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 380 ºC; (c) residual after smoothing.

Figure D.4. Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 300 ºC; (c) residual after smoothing.

Figure D.5. Plot of importance index of the selected 1550 wavenumbers.

Figure D.6. Residuals obtained after performing SVD on the 400 ºC data set considering: (a) 2 components and (b) 4 components.

Figure D.7. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 300 ºC.

Figure D.8. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 350 ºC.

Figure D.9. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 380 ºC.

Figure D.10. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 420 ºC.

Figure D.11. Initial concentration estimates for S_1, S_2 and S_3 at 300 ºC.

Figure D.12. ALS residuals for datasets obtained at: (a) 300 ºC; (b) 350 ºC; (c) 380 ºC; (d) 400 ºC; (e) 420 ºC.

Figure D.13. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 300 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

Figure D.14. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 350 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

Figure D.15. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 380 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

Figure D.16. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 400 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shows as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

Figure D.17. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 420 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shows as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

Figure D.18. Effective intensity for each wavenumber in the fifth cluster (Table 6.14 in Chapter 6). Some of the important peaks are indicated.

Figure D.19. Plots of: (a) ROD vs. number of components and (b) initial estimates of concentration obtained through EFA for the 35 samples at various process conditions used in the SMCR-ALS global model.

1. Introduction

1.1 Background and scope

One of the main challenges in escalating a pilot-scale chemical process to the industrial level is automation. In other words, eliminating the requirement for any analytical equipment that requires significant human intervention would be beneficial in a number of ways for the corresponding industrial process. It cuts human and capital cost, improves overall process time, and reduces the human component of error. At the same time, the process chemistry needs to be fathomed and uncovered. This feature would particularly be useful in an industry that involves a lot of analytical science and focuses on increasing the value of a chemical by conversion to a higher-value product at minimum cost and lesser time. In addition, synthesis of chemicals involving a catalyst are more capital-intensive and quite sensitive to external factors like temperature, pressure and change in other process conditions like flowrate of the reactant.

The backbone of any chemical process in the laboratory setting or in the industry is the data generated from the constituent analytical instruments. Data can be qualitative or quantitative and related to the chemical environment apart from the system itself. [1] Meta-data containing supporting information like file size, date of performing the experiment, etc. is recommended to accompany the raw data for assisting the user. The most common sources of data in a pilot-scale process is from spectroscopy and chromatography, whose elimination can reduce run-times significantly. However, due to the multidimensional nature of the data, there is a lot of overlapping that needs to be dealt with in order to obtain useful information that can be interpreted in a chemical sense. [2]

Chemometric tools that adopt statistical approaches through mathematical methods translate raw data from analytical instruments into reliable information that can further assist in linking the measured property and reaction chemistry. [3] They can substitute for a characterization instrument and speed up the scale-up operation, also. On the other hand, chemoinformatics consists of a further step where the information from chemometrics is converted to global knowledge of the population set of similar systems through building master curves and kinetic models, but is not within the scope of this work. [4]

Furthermore, it is important to consider the nature of the chemical reacting system in which the value of a substance is required to be increased. The world demand for petrochemicals has

been growing tremendously over the years. [5] Most of the derived products from petrochemicals are used on a daily-basis as plastics for food packaging, ammonia for fertilizers, synthetic rubber for tires, detergents/surfactants in households and a number of other commodities. The demand for plastics has almost doubled since 2000 and one of the major starting materials for their manufacture is lighter molecular weight olefins like ethylene and propylene, whose production alone increased by 5 %wt. in the decade from 1990-2000 and their demand is expected to increase by 3.5 %wt. over the next 15 years. [6,7]

Since these gaseous olefins mostly occur as by-products from thermal processing of organic material like heavy oil and bitumen, their conversion to higher molecular weight compounds in the gasoline and distillate boiling range was considered to be worthwhile. Organizations like Mobil Research Development Corporation developed a catalytic method using shape-selective acidic zeolite, ZSM-5, to carry out this conversion. [8] Since the mechanism involves the formation of a carbocation which facilitates cracking, isomerization and polymerization reactions, it can lead to a distribution of carbon numbers in the products and keeping track of them through experimental means alone is a major challenge. Kinetic models comprising lumped mixtures would seem a practical solution but the limitation is that this requires a predefined reaction network and an established link between product properties and oligomerization chemistry. [9,10] Reactor modeling was also shown to predict concentration profiles for ethane cracking but the number of parameters to be estimated was large. [11] There was a need to monitor the product composition through a more versatile method.

Another important reaction involving petroleum-derived feedstocks are oxidation of aromatic species to produce petrochemicals that are crucial in the polymer and pharmaceutical industry. [12,13] Conversion of *p*-xylene to dimethyl terephthalate (DMT), that is later copolymerized with a linear olefin to produce polyethylene terephthalate (PET) was implemented by DuPont and finds application in fibers, resins and films. [14] Though oxidation catalysts are highly selective, not all feeds are suited for catalytic conversion and moreover, the catalysts deactivate over time and generate hazardous wastes that can contribute to an increase in costs. [15] Non-catalytic oxidation is less capital-intensive but the main issue is controlling selectivity since some products are more reactive than the feed itself. Oxidation of naphthene aromatics like tetralin produce 1-tetralone which serves as a starting material for pharmaceutical agents and in agriculture as well. But the reaction needs to be selective to the ketone rather than the corresponding alcohol due as 1-tetralol

is of lesser industrial value than 1-tetralone. It is obvious that oxygen availability should be maximized but the challenge in oxidation of chemicals like tetralin is that it becomes a two-phase flow. Thus, determining the relative importance of a single parameter among others in the complex system that can control the oxygen availability becomes critical in designing an industrial reactor for this process.

One of the major reasons for the increase in world oil demand can be attributed to the petrochemical industry. To put things into perspective, the petrochemical industry accounts for a third of the total oil supply and will continue to do so till 2030, as projected by the International Energy Agency. [16] But with the crude oil demand surging up to 34 billion barrels per year (93 million barrels per day), sources of conventional light oil are fast-depleting and there is a steady shift towards heavy crude oil and bitumen extraction, transportation and processing. [17] Canada owns the world's largest deposit of oilsands of ~2400 billion barrels out of which only ~170 billion barrels are estimated to be recoverable. [18] Once recovered, transportation through pipelines is an arduous task due to its low fluidity. The main challenge is the reduction of viscosity by several orders of magnitude to meet the pipeline specifications, which consist of density and olefin content specifications as well. [19] Due to the higher cost of diluents and issues of compatibility with bitumen, partial upgrading through thermal conversion is considered a viable alternative with the intent to achieve sufficient viscosity reduction at the site of extraction without significantly upgrading bitumen properties not related to transport. Viscosity of a complex mixture is dependent on a number of factors like composition and molecular structure, phase behavior and external factors like interaction with solvents and shear rates of measurement.

Keeping in mind the above-mentioned challenges in different sectors of the petrochemical industry and the usefulness of chemometrics in addressing some of these in analytical science, this work was carried out in four phases considering three systems of increasing complexity in terms of composition of the reaction mixture. These systems were: (i) acid-catalyzed oligomerization of an aliphatic alkene, propylene in the temperature range of 350 – 480 ºC in a flow reactor; (ii) non-catalytic oxidation of a naphthene aromatic, tetralin at 150 ºC in a microfluidic reactor; (iii) non-catalytic thermal conversion of Athabasca bitumen at lower than industrial temperatures of 300 – 420 ºC in batch reactors covering both visbreaking and coking regions. The first two phases were conducted independently and corresponded to propylene oligomerization and tetralin oxidation systems and utilized chemometric tools involving supervised learning like nonlinear and linear

regression to monitor the reaction online so as to eliminate the use of offline concentration-measuring chromatography instruments. The final two phases dealt with thermal conversion of bitumen, where extensive experimental techniques were used in the third phase to identify reasons for viscosity changes followed by application of advanced quantitative chemometric methods like multivariate curve resolution (MCR) and Bayesian approaches in the fourth phase to derive a credible reaction network for the thermal conversion of Athabasca bitumen.

The main objectives and highlights of study for each phase is highlighted. In the first phase, the goal was to establish spectrum-property relationships for tracking the composition of the hydrocarbon products from H-ZSM-5 catalyzed propylene oligomerization at the reactor outlet to alleviate the use of gas chromatography-mass spectrometry (GC-MS) equipment. Least squares-support vector machine (LS-SVM) adopting kernel transformations and partial least squares (PLSR) regression were used to predict product concentrations using a variety of inputs including inlet reaction conditions, raw and dimension reduced Fourier transform infrared (FTIR) spectra of the products through unsupervised clustering. The predictive power was compared with a variant of PLSR, interval-PLSR, when infrared spectra were used as the input. The potential for online tracking of the product composition, learning efficiencies of the regression methods, their ability to capture the oligomerization chemistry and nonlinearity in the data were also examined.

In the second phase, it was essential to quantitatively determine whether oxygen availability played a meaningful role in controlling ketone-to-alcohol selectivity in the petrochemical products obtained from oxidation of tetralin in a micro-fluidic reactor under Taylor-flow conditions. The calibration data was based on a previous study by Siddiquee et al., [20] where only a qualitative analysis was performed. Simple and multiple linear regression (SLR and MLR) models were used to identify the relative importance of each parameter (mass transfer or hydrodynamic) among five identified from the previous study [20] to predict tetralin conversion and product selectivity. The ability of significance tests like t-tests, analysis of variance study like F-tests, partial correlations, standardized coefficients along with multicollinearity diagnostics to identify the most important parameter and whether it was relates to oxygen availability was evaluated. Since the inlet flow-rate of tetralin was the only user-manipulated variable, the feasibility of a regression pathway to predict product selectivity from inlet flow-rate was investigated for potential online monitoring by eliminating image analysis and GC procedures. The result has possible implications for building a larger industrial reactor for hydrocarbon oxidation purposes.

In the third phase, it was important to account for the non-monotonic changes in viscosity of Cold Lake bitumen during its thermal conversion at 400 ºC as observed by certain previous studies, from the partial upgrading perspective. [21-23] This study was conducted on Athabasca bitumen at 400 ºC since a higher liquid yield could be obtained. A number of factors that could potentially affect viscosity were evaluated including changes in composition in terms of aromatic and non-aromatic content, free-radical content, distillation profiles, dependence on post-reaction procedures like the solvent used for product extraction and its effect on infrared spectra and viscosity, and rheological conditions of viscosity measurements like shear rates. The hypothesis was that dominant aggregation forces between the constituent asphaltenes increase viscosity and vice versa. An extensive review of the literature in each of these fields was also provided. The role of viscoelasticity during shear and sol-gel transformation during thermal conversion was also examined and critically analyzed. Only experimental techniques were used in this part of the study.

The final phase of this research involved the application of advanced quantitative chemometric tools like self-modeling multivariate curve resolution (SMCR) and Bayesian hierarchical clustering (BHC) and structure learning methods on the FTIR spectra of liquid products obtained from thermal conversion of Athabasca bitumen in the temperature range of 300 – 420 ºC. The aim was to propose a conceivable reaction pathway for cracking of Athabasca bitumen at less severe conditions than employed in the industry. The resolved concentration and spectral profiles from SMCR serve to provide a hint to the plausible reaction mechanisms occurring at each temperature. Furthermore, BHC classifies similar compound classes based on probability density in the same group and the structure learning methods establish causal relationships between these groups, thus providing a means to building a chemical reaction network. No prior chemical information was used for the chemometric methods but basic knowledge of bitumen conversion chemistry and the proposed reaction mechanisms from SMCR assisted in building a feasible reaction pathway. The derived pathway was compared with that of Cold Lake bitumen [24] to explain changes in viscosity by relating to chemical composition and reactivity as well. The potential of SMCR and Bayesian approaches in monitoring a continuous thermal cracker was also explored.

The point of analyzing reacting systems of increasing chemical complexity was that the application of chemometric and experimental tools could be extended to other colloidal systems, slurries and renewable energy sources like biomass, which are complex mixtures as well.

1.2 Objective

The purpose of this work was to understand the chemical behavior of moderately complex to complex reacting systems under different process conditions by applying a variety of appropriate chemometric and experimental tools with exploration of the possibility of potential online monitoring for each system.

1.3 Chapter-wise highlights

A short description of the work covered in each chapter is given below:

Chapter 2:

A review of important chemometric methods with relevant applications in different fields of analytical chemistry like petroleum, material science, food science, pharmaceutical science and medicine is provided in this chapter. It covers most of the statistical methods including preprocessing, supervised and unsupervised learning techniques like regression and clustering, respectively, used in different phases of the research conducted as a part of this thesis except for Bayesian methods, which are covered in Chapter 6. Though the key aspect of each data analytical technique is described in each study of this thesis, this chapter attempts to compile all methods together and provide context for their use, without providing too much mathematical detail. The need for chemometric approaches in analytical science is also discussed.

Chapter 3:

This chapter describes the work in the first phase on propylene oligomerization as given in the introduction. The usefulness of LSSVM regression adopting kernel transformations in modeling nonlinear spectrum-property relationships combined with dimension reduction of the infrared spectral data is studied. The accuracy of prediction of product composition and implications for oligomerization chemistry with potential for online tracking is also evaluated.

Chapter 4:

The work on investigating the significance of a single parameter in describing oxygen availability in non-catalytic tetralin oxidation in a micro-fluidic reactor through linear regression is covered in this chapter. This was the second phase of my research. The potential for online control by prediction of product selectivity from the only user-controlled variable (inlet flowrate) is also discussed.

Chapter 5:

The work in this chapter corresponded to the third phase of this study. A comprehensive study of the different factors affecting the viscosity of Athabasca bitumen during low-temperature thermal conversion was conducted using only experimental characterization tools. The assumption of a colloidal model for bitumen and the role of aggregation/de-aggregation forces along with hydrodynamic effects of the maltenes with respect to impacting viscosity is also reviewed and critically elaborated.

Chapter 6:

This chapter explains the work performed in the final phase of this research and deals with the application of SMCR and Bayesian methods like BHC and structure learning on FTIR spectra of liquid products from low-temperature thermal conversion of Athabasca bitumen. The objective was to explore the temperature-specific reaction chemistry and to derive a chemical reaction network for the thermal conversion at temperatures lower than in industrial practice.

Chapter 7:

In this chapter, the main conclusions and limitations of each phase of the research are provided. Possible future works that were out of scope for the current works and recommended improvements in each study are also suggested.

Note. The work presented in Chapters 2 – 6 was either published, accepted or submitted for publication in different scientific journals or books. No major modifications were made when including the papers or short communications in the present thesis document. As a result, there might be some similarities in the introduction of each chapter.

1.4 References

(1) Kumar, N.; Bansal, A.; Sarma, G. S.; Rawal, R. K. Chemometrics Tools Used in Analytical Chemistry: An Overview. *Talanta* **2014**, *123*, 186–199.

(2) Liang, Y.Z.; Xie, P.; Chan, K. Quality Control of Herbal Medicines. *J. Chromatogr. B* **2004**, *812*, 53–70.

(3) Willet, P. *Similarity and Clustering in Chemical Information*; Wiley: New York, 1987.

(4) Van de Vijver, R.; Devocht, B. R.; Van Geem, K. M.; Thybaut, J. W.; Marin, G. B. Challenges and Opportunities for Molecule-Based Management of Chemical Processes. *Curr. Opin. Chem. Eng.* **2016**, *13*, 142–149.

(5) Speight, J. G. Refining Heavy Oil and Extra-Heavy Oil. In *Heavy and Extra-heavy Oil Upgrading Technologies*; Gulf Professional Publishing: Boston, 2013.

(6) Belov, G. P.; Matkovsky, P. E. Processes for the Production of Higher Linear α-Olefins. *Pet. Chem.* **2010**, *50*, 283–289.

(7) Cai, T. Studies of a New Alkene Oligomerization Catalyst Derived from Nickel Sulfate. *Catal. Today* **1999**, *51*, 153–160.

(8) Meisel, S. L.; Weisz, P. B. Hydrocarbon Conversion and Synthesis Over ZSM-5 Catalysts. In *Advances in Catalytic Chemistry II Symposium*; Salt Lake City, Utah, 1982.

(9) Mohammed, A. A.; Fateen, S.-E. K.; Ahmed, T. S.; Moustafa, T. M. A Kinetic Model for Ethylene Oligomerization Using Zirconium/Aluminum- and Nickel/Zinc-Based Catalyst Systems in a Batch Reactor. *Appl. Petrochemical Res.* **2014**, *4*, 287–295.

(10) Xu, C.; Al Shoaibi, A. S.; Wang, C.; Carstensen, H.-H.; Dean, A. M. Kinetic Modeling of Ethane Pyrolysis at High Conversion. *J. Phys. Chem. A* **2011**, *115*, 10470–10490.

(11) Garg, R. K.; Krishnan, V. V.; Srivastava, V. K. Prediction of Concentration and Temperature Profiles for Non-Isothermal Ethane Cracking in a Pipe Reactor. *Korean J. Chem. Eng.* **2006**, *23*, 531–539.

(12) Suresh, A. K.; Sharma, M. M.; Sridhar, T. Engineering Aspects of Industrial Liquid-Phase Air Oxidation of Hydrocarbons. *Ind. Eng. Chem. Res.* **2000**, *39*, 3958–3997.

(13) Della Pina, C.; Falletta, E.; Rossi, M. Liquid Phase Oxidation of Organic Compounds by Supported Metal-Based Catalysts with a Focus on Gold. In *Liquid Phase Oxidation via Heterogeneous Catalysis: Organic Synthesis and Industrial Applications*; Wiley: Hoboken, NJ, 2013; pp 221–262.

(14) Emanuel, N. M.; Denisov, E.; Maizus, Z. K. *Liquid-Phase Oxidation of Hydrocarbons*; Plenum Press: New York, 1967.

(15) Hill, C. L.; Kholdeeva, O. A. Selective Liquid Phase Oxidations in the Presence of Supported Polyoxometalates. In *Liquid Phase Oxidation via Heterogeneous Catalysis: Organic Synthesis and Industrial Applications*; Wiley: Hoboken, NJ, 2013; pp 263–319.

(16) *The Future of Petrochemicals*; 2018.

(17) Ancheyta, J.; Trejo, F.; Rana Singh, M. Asphaltenes: Chemical Transformation during Hydroprocessing of Heavy Oils; CRC Press: Boca Raton, FL, 2010.

(18) Gray, M. R. *Upgrading Oilsands Bitumen and Heavy Oil*; The University of Alberta Press: Edmonton, Canada, 2015.

(19) de Klerk, A.; Gray, M. R.; Zerpa, N. Unconventional Oil and Gas. Oilsands. In *Future Energy: Improved, Sustainable and Clean Options for our Planet*; Elsevier: Amsterdam, 2013; pp 95–116.

(20) Siddiquee, M. N.; de Klerk, A.; Nazemifard, N. Application of Microfluidics to Control Product Selectivity during Non-Catalytic Oxidation of Naphthenic-Aromatic Hydrocarbons. *React. Chem. Eng.* **2016**, *1*, 418–435.

(21) Zachariah, A.; De Klerk, A. Thermal Conversion Regimes for Oilsands Bitumen. *Energy Fuels* **2016**, *30*, 239–248.

(22) Yañez Jaramillo, L. M.; De Klerk, A. Partial Upgrading of Bitumen by Thermal Conversion at 150-300 °C. *Energy Fuels* **2018**, *32*, 3299–3311.

(23) Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A. Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C. *Energy Fuels* **2014**, *28*, 5014–5022.

(24) Tefera, D. T.; Agrawal, A.; Yañez Jaramillo, L. M.; De Klerk, A.; Prasad, V. Self-Modeling Multivariate Curve Resolution Model for Online Monitoring of Bitumen Conversion Using Infrared Spectroscopy. *Ind. Eng. Chem. Res.* **2017**, *56*, 10756–10769.

2. Relevance of chemometrics as a process-systems engineering (PSE) technique with implications in analytical chemistry [1]

ABSTRACT

The applications of chemometric methods involving classification and regression in different disciplines of analytical science are manifold. This chapter focuses on key elements of the various chemometric approaches that are mainly used to decouple overlapping spectra from unknown components in a mixture and extract key features from redundant data to facilitate further chemical interpretation. The aim was to present an initial understanding of most of the multivariate analytical techniques employed in further works of this thesis. The borderline between chemometrics and chemoinformatics is also highlighted. Qualitative methods that are a means of unsupervised classification like principal component analysis, hierarchical clustering analysis, multi-dimensional scaling and parallel factor analysis applicable for linear, nonlinear and tri-directional data are reviewed. Quantitative supervised classification techniques like linear discriminant analysis, k-nearest neighbors, and partial least squares-discriminant analysis are also discussed. Calibration and regression techniques capable of modeling linear and nonlinear relationships like partial least squares, support vector machines, artificial neural networks along with variable selection and cross validation techniques are examined. Resolution methods with the corresponding optimization routines for analyzing complex mixtures, especially with their applications in environmental science are also discussed. Other applications in the fields of material science, medicine, petroleum, biological science and biochemistry, and food science are also provided. Possible future developments in the current chemometric approaches like modifications to the existing algorithms and combination of more than one method are also explored.

Keywords: Classification, regression, calibration, resolution, linear and nonlinear relationships, applications in petroleum, environmental and food science.

[1] Most parts of this work were published as a section in 'Sivaramakrishnan, K.; Puliyanda, A.; Tefera, D. T.; Ganesh, A.; Thirumalaivasan, S.; Prasad, V. Perspective on the Impact of Process Systems Engineering on Reaction Engineering. *Ind. Eng. Chem. Res.* Article ASAP, April 4th **2019**'.

2.1 Introduction

With the increasing amounts of data arising from the continuous evolution of advanced experimental techniques in the scientific literature, there is a critical need to convert this data to useful information with the potential of further interpretation. Data are mainly generated in digitized form from the analytical techniques used for characterization. [1] The volume and variety of data is increased significantly by the development of modern hyphenated techniques that combine separation and spectroscopy with detection methods. This offers further scope for qualitative and quantitative analysis to extract embedded and hidden information from the high-throughput data. [2,3]

In the field of heterogeneous catalysis, Medford et al. [4] classified data into four major types namely, relating to the chemical environment, the catalyst material, the active sites involved and the kinetics of the reaction taking place. However, these must be accompanied by meta-data that include information on the conditions and assumptions made during the data acquisition process. Chromatographic and spectroscopic techniques are the most common sources of data in most fields of analytical science. [5] They aid in the compositional analysis of complex mixtures in different fields of application like petroleum, foods, drugs, environmental and biological samples. Overall, data from analytical techniques can be qualitative as well as quantitative. [6] Qualitative data consists of nominal (types of categories), dichotomous (data that can be classified as two major branches) and ordinal (criterion-based classification) variables. On the other hand, data that are measured in continuous, interval or ratio scales like temperature, pressure, concentration are said to be quantitative.

For the results from analytical experiments to be accurate, the instrument- and technique-specific factors like the chemical nature and type of solvent used, detector settings like gain and integration time, the conditions of measurement including sample, column temperature and pH need to be optimized and standardized for enabling comparison between different laboratories. Though modern hyphenated techniques aim to enhance the quality of data obtained by means of reducing instrument noise, increasing separation strength, providing correction for equipment- and sample-related factors like retention time and frequency shifts, anisotropy, and using reference databases for compound identification, the challenge is their interpretation due to the multi-dimensional nature of the data. [7–15]

Chemometrics is a scientific discipline that applies statistical tools involving mathematical methods to enable the conversion of data to valuable information. [16,17] This information, in combination with some prior and intuitive knowledge about the system provides significant interpretation that was not apparent from the stand-alone experimental data. It has a certain degree of overlap with the field of chemoinformatics where knowledge is extracted in the form of advanced models and master equations from the information provided by chemometrics and helps in system automation and further discovery. [18,19]

Analytical chemistry is one of the most common areas that benefits from the use of chemometric tools. This encompasses a variety of specialized disciplines like material science, [20] pharmaceutical science, [7] food science, [21] environmental science, [22] biochemistry, [23] and geochemistry. [17] For example, in environmental science, it can used for tracking of impurities and pollutants in water and air. Perhaps the most important practical application of chemometrics is in food chemistry where it can be used to validate the authenticity of food products, trace origin of food classes, detection of adulteration and quality control, monitoring effects of processing on food components, and detection of food spoilage due to microbial growth. There has been extensive and critically reviewed research relating to the applications of chemometrics to food chemistry. [24–26]

In biochemistry, the anti-bacterial potency of certain drugs can be established quantitatively. [27] Accurate information of source-rocks of certain crude oils and splitting of complex heavy oils into different pseudo-components and tracking their evolution during thermal reactions render chemometrics a paramount application in geochemistry and petroleum disciplines. [28–30] Apart from nullifying the experimental errors, it was also seen that chemometrics was effective in mitigating human effort by eliminating the need for the use of hyphenated techniques like gas chromatography-mass spectrometry (GC-MS) by predicting reactor-output concentrations and product selectivity in processes involving conversion and oxidation of industrially important petrochemical-derived hydrocarbons like propylene [31] and tetralin, [32] respectively.

The origin of the term chemometrics can be traced back to 1971 when the Swedish scientist Svante Wold named the groups of methods as ‘kemometri’. [33] The purpose of chemometric approaches are mainly two-fold: classification and regression. Classification also includes pattern recognition. Furthermore, qualitative chemometric methods can be considered to fall under two broad categories, namely unsupervised and supervised learning techniques. They differ in the way

in which the objects in a dataset are classified, i.e. unsupervised learning techniques perform function classification without prior knowledge of class memberships of the existing data points while supervised learning methods aim to classify new samples into their classes based on a previously trained model. These techniques are generally used as precursors for quantitative methods, wherein relationships between response and input feed variables are established and used for prediction purposes. [34–36] Multivariate resolution methods that involve deconvoluting spectra of complex mixtures into their respective pseudo-components and individual chromatograms also fall under the class of quantitative methods. Specific qualitative and quantitative techniques are briefly reviewed in the next section and interesting applications in various important fields will be examined.

It is not the objective of this chapter to carry out a detailed review of the plethora of chemometric methods used in different fields. The main aim is to realize the vast analytical science-application potency of chemometrics and also look into the future of these battery of techniques and their combinations in better analyzing the huge amount of high dimensional data along with assisting in their conversion to invaluable information. All methods used in various phases of this study are covered in this chapter except Bayesian methods of clustering and structure learning, which have been provided in sufficient detail in Chapter 6.

2.2 Methods

Data handling is the key aspect of any chemometric approach. Every method has a mathematical basis through which the raw data is dealt with in order to make it meaningful and interpretable. This section highlights and briefly describes some of the important techniques commonly used in chemometrics starting from data pre-processing to the various supervised, unsupervised and curve resolution approaches for both qualitative and quantitative analyses of experimental data. The focus of this section will be to understand the key aspect of each method and its effect on the data without diving into too much mathematical details and equations, for which the relevant references will be cited. At the end of this section, it is anticipated that the reader would be equipped to visualize and understand the complexity of each technique. It has to be mentioned that modern computational packages like MATLAB, SPSS, R and languages like

Python, C, C++ have the respective mathematical models embedded in the system and have made the application of chemometric tools user-friendly, time-efficient and tractable.

2.2.1 Data Pre-processing

In a typical dataset, for example data from infrared (IR) spectroscopy that measures the absorbance of each sample at each allocated wavelength channel, the rows generally represent the samples and the columns depict the variables or features (wavelength channel in the case of IR data). A variable can be any attribute that can be measured and is characteristic for a sample, like temperature (T), pressure (P), concentration (C), or molar mass (M), food quality. In multivariate data analysis (MDA) problems, some variables are considered input/explanatory variables or regressors and the desired properties that are to be estimated are labeled output or dependent variables. The inputs can also be correlated with each other so labelling them independent may not be suitable in all cases.

The first step in data analysis is to pre-treat the data to filter out instrument noise and possible human error. It should be noted that some error called imbedded error may be present in the data always and cannot be removed through any pre-processing or factorization method detailed in Chapter 6 (section 6.3.3.4). All variables might not be measured in the same units and real data are mostly heteroscedastic and have significant spread. The different methods of data pre-treatment are aimed to address these issues but care should be taken so as to not lose relevant information as noise. Some pre-treatment methods with their major aims, specific advantages and calculation formulas are summarized in Table 2.1.

Further techniques such as curve smoothing by Savitzky-Golay filter which uses a least-squares polynomial of a lower degree and a particular window-size may also be employed for spectroscopic datasets for noise removal. [37] Multiplicative signal correction (MSC) and multiple imputation that substitute missing data by the average of the variables are other techniques that can be used to quantify extensive chemical and physical variations in the data. [38]

Table 2.1. Common data pre-processing techniques employed for MDA.

Method	Sub-methods	Major aim	Other remarks
Transformation	Can be logarithmic ($\log_{10} x_{ij}$), power ($x_{ij}^{a}, a \in R$), polynomial function ($(t + x_{ij})^{a}$, $a \in I$)	(i) To improve fit of nonlinear variables and reduce effect of heteroscedasticity. [39] (ii) Log has a pseudo scaling effect. (iii) Make skewed distributions symmetric	The problem with log is that it cannot deal with 0 values. Power transformations are better for data with 0 values but cannot convert multiplications to additive effect. Post-scaling needed.
Centering	Mean centering ($x_{ij} - x_{imean}$)	To remove offsets in the data. Focuses on differences in the data while not altering the variance.	Applied on the covariance matrix. Need further processing for heteroscedastic data.
Scaling	Variance-based: auto, [40] pareto, [41] range [42] and vast [43] scaling; Average-based: Level scaling [40]	(i) Auto: to standardize data using a scaling factor to remove effect of measurement units. (ii) Pareto decreases large-fold changes more (iii) Vast scaling emphasizes on samples having small relative standard deviation (iv) Range scaling is based on maximum and minimum values of data points	Since range scaling uses only two points, it is sensitive to outliers. Employing coefficient of variation as scaling factor makes variables dimensionless. Level scaling used when relative changes are more important than absolute ones.

2.2.2 Qualitative methods: Unsupervised classification

These class of methods require pre-processing as a step prior to their application on analytical data. As mentioned in the introduction (section 2.1), unsupervised learning methods search for the underlying data structure without prior knowledge of class memberships of the objects in the data set. A brief description of some of these techniques are given below.

2.2.2.1 Principal component analysis (PCA)

This is one of the most common methods used for dimension reduction of multivariate data in diverse disciplines like engineering, science, meteorology. [44–46] The data matrix is subjected to eigenvector decomposition in the case of a diagonalizable square matrix or singular value decomposition (SVD) in other cases. [47,48] The main objective is to remove data redundancy in the original matrix by feature extraction, to generate a bilinear model of loadings and scores. Loadings indicate the contribution of that variable to the respective principal components and scores are the projected objects in the lower dimensional space. A comprehensive review is given by Jolliffe. [49] In complex systems, PCA is used to determine the number of underlying components that can be a good representation of the overall system and a scree plot of the eigenvalues vs. the number of components is used for this purpose. The results of PCA are paramount for further multivariate calibration methods where transformed uncorrelated variables are used to predict the response variables with minimal interference. [50,51] It is to be noted that Barlett's test can be performed to check the extent of correlation among the variables. [52]

Non-iterative partial least squares (NIPALS) algorithm is sometimes used to carry out PCA where the direction of maximum variance is chosen using a normalized loading vector. [47,53,54] Overlapping between the different elements in scores is checked using sum squared differences to determine the principal components.

2.2.2.2 Hierarchical clustering analysis (HCA)

Clustering is majorly implemented as a pattern recognition technique in different fields of application like medical science and engineering. [54,55] HCA is agglomerative or divisive in nature where a condensed hierarchy algorithm is adopted in most cases. [17] One of the distance metrics among Euclidean, Mahalanobis, Chebychev, cityblock, etc. is used to merge individual objects

into clusters which are then further linked using average, centroid, complete, median or ward metrics that consider all objects within the clusters. HCA is not generally useful as a stand-alone method since the final cluster division is determined by human intervention by placement of the similarity line in the dendrogram though the number of clusters is not predetermined as in *k*-means clustering, which is disadvantageous. [56] One more limitation of HCA is that it does not account for the direction of the data so it is recommended to use PCA before. [5]

2.2.2.3 Multi-dimensional scaling (MDS)

It is a non-linear dimensional reduction method that has been recently developed as an improvement over PCA. [57] It reveals the true data structure in the original space even after converting to the lower dimensional space and is based on the Shepard-Kruskal algorithm. [58,59] In a particular application, MDS was able to identify the origin of certain Chinese oils better than PCA and was able to explain 8% more variance for the first two principal components. [60] Further applications are given in the Applications section (section 2.3) of this chapter.

Apart from MDS, self-organizing maps (SOM) that are similar to artificial neural networks (ANN – section 2.2.4.2) where they map the original data onto set of nodes in the lower dimension is another nonlinear dimension reduction method belonging to the unsupervised category. [17]

2.2.2.4 Parallel factor analysis (PARAFAC)

It is a technique that is borrowed from psychometrics and similar to PCA but applicable to trilinear data like excitation-emission maps from fluorescence. [61–63] The data matrix is decomposed into one score and two loading vectors as opposed to one score and one loading matrix in PCA. Sum of squared errors is used as the basis for determining the number of components and an optimization technique like alternating least squares (ALS) (refer section on resolution methods) can be used to obtain the results of the PARAFAC model. [64] Unlike PCA, all component parameters are calculated together in PARAFAC which reduces computational time.

2.2.3 Quantitative methods: Supervised classification

Two types of supervised classification are reviewed in this work: discriminant analyses (DA) and one-class classifiers (OCC). [65] Discriminant analyses assigns object to predetermined classes while OCC methods confirm the right or wrong allocation of objects to their respective classes. In relation to supervised methods, two terms are of importance, namely class sensitivity and class specificity. Class sensitivity is the inverse of type one error and is the quantity of samples accurately assigned to a class, while on the other hand specificity is the percentage of samples that are recognized as not belonging to the target class and is related to type II error. Both sensitivity and specificity can be increased by raising the sample size. [66] Brief descriptions of important supervised classification methods are given below:

2.2.3.1 Linear discriminant analysis (LDA)

It is quite similar to PCA but hyperplanes of separations are created based on maximum likelihood estimates. It is more applicable in cases where the variance of classes is asymmetric. [67] However, the difference from PCA is that a discriminant function is created for each class and projected on a lower dimensional space, where confidence ellipsoids are created identifying the space for each class. If the number of features exceeds the number of samples, PCA is applied first and overfitting needs to be checked for. [68] Some limitations are that it doesn't apply when data is not normally distributed, or if the covariance matrix is singular. It is sometimes used in combination with a confusion matrix to detect the percentage of correctly classified samples.

2.2.3.2 Partial least squares - discriminant analysis (PLS – DA)

As in most classification problems, the response (y) is a categorical variable and the elements describe membership values of the classes. [69] Partial least squares regression (PLSR) is applied to the data set where the inputs and outputs are decomposed into their respective scores and loading matrices and an inner relation is built relating the scores. Predictions are done based on the PLSR model and compared with the response matrix that have information of the class memberships of the objects. [70] It can deal with singular covariance matrices but the results can depend on the number of latent variables chosen in the regression analysis prior to class allocation.

2.2.3.3 Recursive partitioning

These are classification-regression trees like bagging and boosted trees, random forests where no new canonical variables are created. [71] Based on inequality or equality conditions on the values of the original variables, decisions are taken for allocating the objects to the respective classes. The leaves arising out of the trees point towards the class assignment.

2.2.3.4 Object target rotation (OTR)

A least sum of variance criterion is used to select the central objects in each group in this rare but effective method. [72] A particular example of application of this method is given in Christie et al. [73] Once the central object is selected, the other samples are projected on the selected object and assigned to the respective classes and the result is given as percentage belonging. OTR is considered advantageous over other methods like PCA and PLS-DA and efficient in the case of skewed or bimodal data.

2.2.3.5 Soft independent modeling of class analogies (SIMCA)

It is a soft classification technique where an object can be assigned into more than one classes or even need not be allocated into any class at all. [74] This can be quantified by calculating an inconclusive ratio which compares the number of samples that are not assigned to any of the target classes and those which are assigned to more than one target classes. [75] It is one of the popular OCC methods used in chemometrics, where the target classes are specified for each classification problem, like the chemical compound classes identified in chromatography. This method requires PCA to be performed initially to represent the samples within the score space and the distance to the score space. A new sample is determined to be belonging to a particular class depending on its closeness to a PC space for which defined boundaries are created prior. [76] It is advantageous over the *k*-nearest neighbors method as that does not have a defined boundary for classifying unknown samples. One peculiar disadvantage of SIMCA is that it does not work if the number of samples is too large and leads to overfitting. [77]

2.2.3.6 *k*-nearest neighbors (*k*-NN)

This approach can be used as both classification and regression for categorical and continuous variables, respectively. [78] For a new sample to be classified into predetermined classes in the original data matrix, the Euclidean distance is calculated from the new object to all the objects in the data matrix. Groups of nearest *k* neighbors are created based on minimal distances in the original data set and a new object is allocated to one of the created classes that it most represents with. Dimension reduction is necessary prior to application of *k*-NN if the number of features is larger than the observations. A higher value of *k* is recommended for a lower classification error. [79]

2.2.4 Quantitative methods: Calibration and regression

These methods are employed to establish a mathematical relationship between the input and output variables of the system in order to predict the desired property for new input samples. The output variable is a continuous variable so that the predicted value from the discriminant function in the regression model can be compared with the experimental value to calculate the estimation errors and check the accuracy of the developed model. The samples in the data set are generally divided into three regions: a set for variable selection, a training set for model construction and a validation set for prediction. [1] It is a common practice to split the data set into only 2 parts with the variable selection being carried out on the training set itself and validation performed on a separate data set. [80] The main task in regression analysis is to calculate estimates of regression coefficients of the different variables, and to verify their accuracy through their standard errors and importance through significance (p-values). In the ordinary least squares (OLS) approach, coefficient estimates are unbiased to enable meaningful statistical inferences.

2.2.4.1 Correlations and variable selection

An important aspect to be taken care in regression is correlations and inter-relationships between explanatory variables. A particular example related to one of the works performed in the second phase of this thesis was considered relevant to this context. Siddiquee et al. [32] investigated

a liquid-phase petrochemical oxidation system where multicollinearity among the different explanatory variables was investigated by applying simple (SLR) and multiple linear regression (MLR) models. MLR is not useful when collinearity exists between the regressors but in this case, the goal was to extract information about the relative importance of the input variables in predicting the product selectivity and was performed by evaluating the effect of an added variable to the SLR models. Correlations can be detected with one of Pearson's, Spearman's, Kendall, or γ correlations. [81]

In the case of multivariate data like spectroscopic datasets, variable selection is done to improve the robustness of the model, guard against overfitting and significantly reduce computational time by removing redundant variables. This is achieved by jackknifing where a variable can be omitted if its contribution is insignificant from the calculated p-values in t-tests. Variables can also be assigned weights to realize their contribution in the method of variable importance. Genetic algorithms, that are inspired by natural processes can also perform variable selection. [82] Another way of dealing with collinearity is using ridge regression (RR) where an extra parameter is used while calculating the regression coefficient estimates. [83] However, RR renders the coefficients biased in order to reduce the inflated standard errors due to multicollinearity. A better variable selection method is LASSO (least absolute shrinkage and selection operator) regression which also incorporates an additional parameter in determining the coefficient estimates. [84]

Recently, *i*-PLS developed by Norgaard and Leardi [85] has been employed as an effective technique for variable selection for spectroscopic data sets. It splits the range of input values into a number of segments as given by the user and applied PLSR on each interval to find out the best prediction segment for the desired response. Optimum number of latent variables are calculated for each interval before performing regression analysis. In this way, noisy regions of the input spectrum can be identified and removed.

2.2.4.2 Calibration and regression methods

Linear relationships between inputs and responses in a system can be readily established by MLR, PLSR and principal component regression (PCR) approaches. In PCR, the data matrix is decomposed by SVD into scores and loading vectors and used for developing the regression

function. PLSR can be used when collinearity is detected in the inputs and shows improvement over PCR as both the input and output data are decomposed by PCA (outer relation) after which inner relationships are developed between the respective scores to predict the output. [86] The number of latent variables are chosen by minimization of error of cross validation and is elaborated in the next section. A comparison of these linear methods were made by Frank and Friedman [87] where they worked with different data sets that were ill-conditioned and collinear. In some problems, they found RR to be better than PLSR or PCR and significantly improved over SLR based on OLS.

Machine learning tools like support vector machine-based regression (SVR) and artificial neural networks (ANNs) are quite useful in modeling nonlinear relationships between input and response variables for the system under consideration. [88] SVR is based on the principle of support vector machines as first developed by Vapnik for classification purposes. [89] Discriminant functions representing separation hyperplanes are created by transforming the data to a higher dimensional space using a kernel function. [90] Kernel functions can be linear, sigmoid, polynomial or Gaussian radial basis functions (RBF). RBF and polynomials kernels have been reported to be better for modeling nonlinear relationships than linear and sigmoid kernels in different studies in literature. SVR requires two (RBF) and three (polynomial) parameters to be tuned that include the regularization term to account for the trade-off between margin maximization and estimation error minimization. [80] The kernel width parameter in RBF gives an indication of the degree of nonlinearity of the data, i.e. a lower value indicates a more nonlinear relationship between input and output variables. [91] Tikhonov regularization is a common term given to all methods that utilize a regularization term in the modeling process. [84] SVMs possess certain advantages over ANN as they guarantee a definite solution minima and training repeatability. [88]

2.2.4.3 Cross-validation

The prediction accuracy of a regression model can be significantly improved by cross-validation methods. It is commonly conducted within the training set but exclusive from the prediction set to avoid increase in model bias. [92] Some of the common procedures followed for cross-validation are: Venetian blinds, leave-one-out cross validation (LOOCV), hold-out cross validation, contiguous blocks, etc. [93,94] In each of these types, a certain amount of the samples in

the training set is reserved for validation during the model building process itself. LOOCV leads to a lesser bias but more variance and hold-out CV results in the opposite effect. LOOCV was also shown to be preferable for small sample sized data sets and to guard against model mis-specification. [95,96]

The error of cross validation is calculated for each step during model construction and the tuning parameters in case of SVMs, number of latent variables in case of PLSR and *i*-PLSR, etc. are selected based on the least value of the cross-validation error. Root mean square error (RMSE) concentrates on the larger errors and is most commonly used in regression analysis also to investigate the prediction accuracy of the calibration models by comparing with the experimental output values for new objects. Other types of cross-validation techniques are randomization where the output data points are shuffled, simulation where input data are scrambled, bootstrap where one sample can be selected repeatedly but bias is generated in the process. [97] To deal with the selection bias, repeated double cross validation can also be conducted where the samples are split into two loops with PCA applied on the inner one and prediction done on the outer one. [98]

An interesting method for larger sized data sets and nonlinear systems is Gaussian process regression (GPR) that deals with maximizing posterior probability. [99] A particular variant of SVMs employed in recent works is least squares-support vector machine regression (LSSVM) where the cost function to be minimized becomes linear due to conversion of inequality constraints to a single equality constraint. [31] Specific differences between LS-SVM and GPR have been outlined in the work by Cui and Fearn. [99]

2.2.5 Quantitative methods: Resolution

This class of methods is mainly applied to complex mixtures and aims to extract information on the number of components that significantly contribute to the mixture properties, the concentration of the components and their respective spectra in the case of different hyphenated analytical techniques employed. The advantage is that they do not require any prior knowledge about the system. However, with this knowledge, the reaction network, further chemical interpretation and possible hypotheses can be generated. [100]

The original data matrix consists of the raw experimental data that can be two way or three way, i.e. retention/reaction times vs. wavelength channels vs. intensity of absorbance is an example

of a three-way data. Almost always, the matrix is decomposed to extract the number of components contributing to majority of data variance and then one of the following optimization algorithms employed to calculate the individual concentrations and the spectra of the extracted pseudo-components: (i) rank annihilation factor analysis (RAFA); [101] (ii) Generalized rank annihilation method (GRAM); [5] (iii) Evolving factor analysis (EFA); [102] (iv) Alternating least squares (ALS); [103] (v) Orthogonal projection analysis (OPA). [104] The rank annihilation methods consist of bilinear data sets from which the relative concentrations of components in a mixture are derived. GRAM is a non-iterative method. EFA can provide initial estimates of concentration and spectra of the pure components in a mixture by an evolutionary process of applying PCA in backward and forward time directions to the data set. In this way, active concentration regions are detected. Fixed size moving window EFA (FSMW-EFA) is an improvement over the conventional EFA algorithm where the backward and forward movement is substituted by a continuous moving window of fixed size. [105] FSMW-EFA was suggested to be more robust against noise and thus employed in the final phase of this study to find the initial concentration profiles for the pseudo-components in thermal conversion of Athabasca bitumen (Chapter 6). Another method quite similar to EFA is eigen structure tracking analysis (ETA) where the window size starting with 2 is increased by a unit until the number of components is reached. [105]

In this way, the active and zero concentration regions are revealed for the components over time and subsequently the spectra are calculated by the above-mentioned algorithms. In certain cases, inverse least squares (ILS) is used in resolution where the concentration is considered to be dependent on absorbance. All methods utilize plots of eigen-values of the data matrix covariance that varies with the time points. The advantage of these optimization algorithms is that the underlying noise structure is also estimated. The following section will highlight interesting applications of the various chemometric methods and it has to be noted that most systems require the use of more than one method in combination to yield fruitful information about system properties. The source of the data from different analytical equipment was also highlighted.

2.3 Applications of chemometrics in analytical science

2.3.1 Integration of computational processes with analytical instruments

The association of analytical techniques with computers has increased the scope for control of experiments and further data processing. Computation aids not only in data analysis but also with the performance of the experimental techniques in speeding up the process, eliminate the requirement of constant monitoring and sample feeding. One of the early attempts in this quest was taken by a number of researchers including Christiansen et al., [106] Wu and Malmstadt. [107] The approach was to try to integrate microcomputers with titrators for photometric and potentiometric purposes. A particular application was in nylon manufacturing where the titrant could be added at a user-determined flowrate automatically to find out the equivalence points.

Flow-injection analysis, which enables the study of interaction of a sample with a reagent, the kinetics of the reaction and the dispersion effects can also be optimized by computer-supported techniques. [108] In a particular experiment, temperature was controlled within 0.02 °C, smoothing and peak detection was applied to the product peaks, thus permitting better system performance by the aid of computation. [109] Experimental conditions and parameters can also be optimized using computational software. Early examples of these kinds of studies were done by Stieg and Nieman [110] in 1980 and Sly et al. [111] in 1982 where they worked on optimizing a chemiluminescent reaction and a flow-injection system in a tubular reactor. Even in a complex flow-injection system, automation of the simple processes that are easily enabled by human intervention is challenging. In older UV-Visible spectrometers, wavelengths needed to be scanned mechanically but modern manufacturers have replaced the moving parts by fast electronic scanners that obtain the entire spectrum simultaneously. [112] Modern detectors employed in spectrometers are multi-channeled and constitute a charge-coupled device (CCD) that converts light photons to electric charges for easy detection. Recent developments have made the use of auto-samplers for sample feeding in instruments like inductively coupled plasma-optical emission spectroscopy (ICP-OES) for detection of trace-elements in complex mixtures, gas chromatography-mass spectrometry (GC-MS) for concentration calculations and other techniques that require repetitive sample doses in the same procedure.

2.3.2 Chemometric applications in data processing in various disciplines

As conducive as it seems, digitization of analytical data has further challenges of interpretation and extracting useful information, for which chemometrics employing computational approaches was explored.

2.3.2.1 Material science

An interesting usage was in the analysis of signals emitted from stressed materials to detect their failure. [113] Initially, the spectra were checked for correlations but no particular pre-processing techniques as described in the 'Methods' section (section 2.2.1) were applied. Later, HCA was applied on emissions from materials like polymers to identify spectral groupings, followed by supervised techniques. They were also later applied to composites like glass fiber and polyester to identify damaged material. These chemometric approaches saved time and human effort to identify classes in complex spectra and relate them to damaged material. It has been recorded in the literature that online pattern recognition techniques had been employed in the development of the present-day IR spectrometer by matching the spectral groups created with structural units of compound classes in standard libraries. [112]

2.3.2.2 Medicine

An important application in the medical field is tomography assisted with spectroscopy to detect cancers and tumors. The advantage is that the technique is non-destructive and the absorption data of the biological sample is transformed to its cross-sectional image by some chemometric process. [114] Nuclear magnetic resonance (NMR) is the commonly used characterization technique in tomography. Malign growth can be revealed by changes in the NMR frequency or signal strength and the cross-sectional images can provide information about the size of the tumors. [115]

2.3.2.3 Physical chemistry

Nonlinear regression has also been used for estimation of physical properties like the pKa of a dibasic aromatic compound from its UV absorbance spectra. [116] At a constant wavelength of 304

nm, numerous spectra were obtained at different pH of the compound in a buffer solution after which the following equation proposed by Albert and Serjeant [117] relating the absorbance to the concentration of the compound with the two dissociation constants and molar absorptivity of the ionic and neutral forms being the parameters to be estimated:

$$A = \frac{c_t a_d [H^+]^2 + c_t a_m [H^+] K_1 + c_t a_n K_1 K_2}{[H^+]^2 + [H^+] K_1 + K_1 K_2} \quad \text{Equation 2.1}$$

where c_t, a_n, a_m, a_d, K_1, K_2 are the compound concentration, molar absorptivity of the neutral, mono-cation and di-cation forms of the compound, 1st and 2nd acid dissociation constants of the compound, respectively.

Curve fitting with algebraic manipulation was used to obtain the pKa of the compound with a high coefficient of determination (R^2) of 0.9997.

2.3.2.4 Petroleum

In the field of petroleum, differentiation of crude oils and identification of their sources are significant areas where chemometric tools play a role. Eide et al. [118] used chemometric curve resolution methods on GC-MS characterization to identify individual components from diesel engines exhausts. Eide and Zahlsen [119] utilized electrospray ionization mass spectrometry (ESI-MS) as an improvement over GC as it ionizes the molecules without fragmentation. Mass spectra of crude oil mixtures are complex where each line can indicate different compound isomers of the same molecular weight. Score plots between the first two principal components in PCA was used to classify four types of oils (two from the North sea near Norway, a South American and a central Asian). The first principal component (PC) accounted for 53 % while the second PC accounted for 23 % of the variance in the original data.

Furthermore, PLS regression was used to predict the distribution of the two types of Norwegian oils from ESI-MS spectra and resulted in an excellent prediction accuracy (R^2 = 0.996). For calibration purposes, 5 different volume mixtures of the Norwegian oils were used. Thus, they confirmed that this method could be employed for chemical fingerprinting. However, the single quadrupole ESI-MS is inferior to high-resolution MS in terms of line resolution and thus, requires significant assistance of pattern recognition and calibration techniques. [120]

The impact of auto-scaling and range scaling on the PCA output was studied by Wang et al. [121] on biomarker data for matching the crude oils with their respective source rocks located in the

Northern Wuerxun Depression, Hailar Basin. The plots of the first two principal components for both types of pre-processing methods employed are given in Figure 2.1.

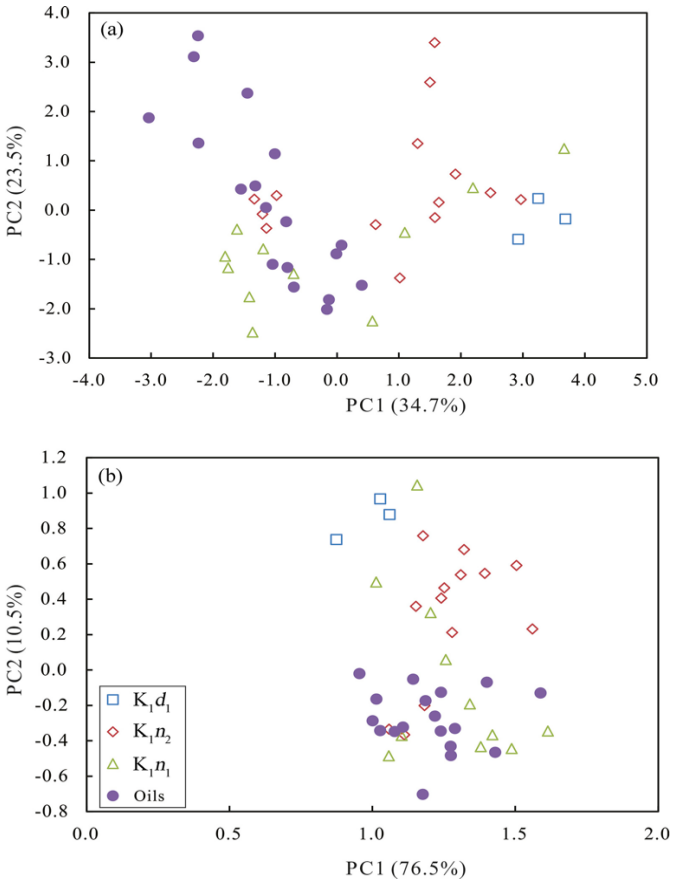


Figure 2.1. PC1 vs. PC2 for auto-scaling (top) and range scaling (bottom) applied on oil and source-rock samples from the Northern Wuerxun Depression, Hailar Basin. [17] Reproduced with permission from Science Direct.

It can be seen that in the case of range scaling, the first two principal components put together could explain 30 % more variance than for the case of auto-scaling. In addition, the first component alone contributes to 76 % of the variance in range scaling while the first component in the case of auto-scaling could only explain 35 % of the variance. Mudge [122] utilized PLSR to identify the possible sources for the hydrocarbons that constitute the sediment deposits in the Gulf of Alaska. A clear mapping of the sources to the hydrocarbons could not be identified in this analysis.

Bylund et al. [123] suggested the use of PARAFAC to deal with slight shifts in liquid chromatography-mass spectrometry to align the multidimensional data in the chromatographic

direction. An interesting application of interval-PLS (*i*-PLS), which is a variant of PLSR was in the work by Sivaramakrishnan et al. [31] where they compared the forecasting ability of LS-SVM adopting RBF and polynomial kernels, PLSR, *i*-PLSR for predicting the concentrations of the products from acid-catalyzed oligomerization of propylene in the temperature range of 346 – 477 °C at molar flowrates of 172 – 440 mL/min based on FTIR spectra. The products were in the gasoline boiling range. Four product streams based on PCA and HCA were identified and regression models were built to establish relationships between inlet operating conditions of temperature and flowrate as well as the FTIR spectra of the products and the product concentrations, respectively. HCA was also used to find groups in the infrared spectra which were later used to predict the output concentrations as well. Among the different chemometric methods used, LS-SVM with RBF kernel was found to have the best prediction accuracy in most cases with *i*-PLSR performing better than PLSR and even LS-SVM based on polynomial kernels in some situations. The best performing interval identified by *i*-PLS was in the 1396 – 1535 cm^{-1} range that corresponded to sp^3 C-H bending frequencies and also some aromatic C=C stretching. Though this could not be directly related to the carbon number composition of the products, a rough idea was obtained through confirmation with HCA classification as well.

2.3.2.5 Archeology

OTR (section 2.2.3.4) was utilized to relate the chemical composition of coins used by Europeans to track their history. [73] It was found that there was no ambiguity in the historical periods but a sudden change in the composition was detected during the intermediate period.

2.3.2.6 Environmental science

The application of chemometric resolution methods in environmental chemistry is paramount. In a study by Comas et al., [124] the aim was to determine the concentration of humic and fulvic acids in river and wastewater using HPLC as the analytical technique. PARAFAC, as a trilinear dimension reduction technique, is of immense importance in the analysis of three-directional data and was used in this study. PARAFAC in combination with GRAM and MCR-ALS was also used to decompose chromatograms that are highly overlapped and drifted due to mixtures of humic and

fulvic acids in water. Due to the large quantity of analytes present, GRAM was utilized which quantifies all components simultaneously. The challenge here was the interference from undesired material like sulfites in HPLC. MCR-ALS was seen to be robust against a time-shift algorithm used to align the shifted peaks to make the concentration predictions more reliable.

In order to resolve the overlapping excitation-emission maps from the fluorescence spectra of a complex poly-nuclear aromatic solution consisting of 6 constituent components, RAFA as a resolution technique was employed by Ho et al. [125] In another work, impurities were able to be detected in samples of tetracycline hydrochloride by decomposition of their HPLC-DAD data into the respective spectra and concentration profiles. [126] FSMW-EFA and OPA were used to obtain the initial concentration estimates and checking the homogeneity of chromatogram peaks while the final resolution was performed by MCR-ALS optimization. MCR-ALS has been shown to solve issues of co-elution in chromatographic techniques in different kinds of compositionally complex samples.

It was noticed by Salau et al. [127] that the pesticides, carbofuran, propoxur and pirimicarb have comparable mass spectra but MCR-ALS could deal with the co-elution problem. The suppression effects on the ion formation of these compounds in the thermospray mass spectra were studied using the resolution methods. MCR-ALS was also particularly useful to resolve the FTIR spectra of Canadian oilsands-derived Cold Lake bitumen into 3 pseudo-component spectra and their change in concentration at temperatures between 150 – 300 °C and reaction times of 1 – 8 h conducted in batch reactors. [30] EFA was used to obtain the initial estimates of concentration and spectral profiles. Plausible reaction mechanisms at different temperatures were proposed according to the results from the MCR-ALS routine and a nature-inspired optimization technique like particle swarm optimization (PSO) was suggested as an improvement to the ALS approach.

Application of calibration methods like PCR based on classical least squares and inverse least squares (ILS) were used on UV spectrometer data for multi-component resolution to determine colorants in soft drinks. [128] The results were validated with data from HPLC with good accuracy.

2.3.2.7 Biological science and biochemistry

Chemometrics has a huge role to play in the classification and quantification of micro-biological samples as shown in the next few examples. Though PLS-DA and SIMCA have

different goals, both were individually applied to identify the sources of the *Ganoderma lucidum* sample. [129] Their performance could not be compared as PLS-DA is a hard classification and SIMCA is a soft classification technique as outlined in section 2.2 of this chapter. FTIR spectra in combination with PLSR as the calibration method was used successfully to track adulteration of fish oil by specific vegetable oils in the work by Rohman and Che Man. [130]

Electron spin resonance (ESR) spectroscopy has been an area of increasing interest for application of chemometric techniques in biological systems. ESR spectra identifies different kinds of free radicals present in a system, namely those with peaks around the same g-value of ~2.003 for organic free radical on carbon centre, free radicals on nitrogen and oxygen centres and those with g-values in the range 1.980 – 2.020 for oxo-vanadyl (VO^{2+}) spins. [131] The output from the ESR instrument is in first derivative form of the absorption intensities against the varying external magnetic field. The multi-dimensional nature of this spectra provides scope for multivariate data analysis to enable easier interpretation. In the work by Koskinen and Kowalski, [132] the objective was to study model membrane systems replicative of those used in biological systems by attaching a long chain spin label nitroxide compound with the radical centre on oxygen to track changes in properties of the system in focus. Particularly, one system used 3-doxyl-5α-cholestane (I) as the spin label to investigate temperature changes in an inclusion thiourea crystal through changes in ESR spectrum. Fourier transform, baseline correction and smoothing were applied on the raw spectra consisting of 980 magnetic field channels to reduce it to 64 variables that contain most of the initial information. ESR spectra were obtained at sixteen different temperatures from -82 °C to 59.2 °C. The chemometric tools employed were correlations between the reduced features and temperature followed by stepwise regression to extract the most important variable that describes the temperature best. Information about restricted anisotropic motion and rotation of the molecule were also obtained from the spread of the regression plots.

Another area of application was the use of the same statistical approaches outlined above in quantifying the concentration of phospholipid spin labels in a protein and extract information about its mobility as a mimic of biological systems. [132]

2.3.2.8 Food science

Applications of multivariate data analysis methods in food science are aplenty and reviewed critically by various researchers. [24–26] Specific areas include studying the effect of processing on food properties, food safety and authenticity in the form of detecting adulterants, spoilage by microbial growth, etc. [52,133] Simple cases like use of HCA for identifying clusters of peaks in FTIR spectra of frozen chicken samples were investigated by Grunert et al. 2016. [134] It was suggested that HCA was more commonly used than *k*-means or *k*-median clustering for pattern recognition. In the work by D'Archivio et al., [135] a total of 144 different Italian samples of saffron were classified according to the concentrations of safranal, picrocrocin, crocin and some of their derivatives by applying LDA on HPLC data. LDA was shown to be beneficial over PCA and resulted in 88 % specificity of the unknown samples in this work. *Cis*-crocetin bis(b-D-glucosyl) ester and *trans*-crocetin bis(b-D-glucosyl) ester were identified to be major contributors in the multivariate data analysis (MDA) process due to their distinct HPLC peaks. Advanced hyphenated techniques like high resolution magic angle spinning proton nuclear magnetic resonance (HR-MAS ^{1}H NMR) were used for tracking the source of 60 samples of Forasteiro cocoa beans to their respective locations. PLS-DA was shown to perform well for this system and the latent variables explained 65 % of the variance in the data.

Gondim et al. [75] used SIMCA in association with IR-spectra to identify impurities like water, sodium hydroxide, chloride, starch, carbonate, etc. in a number of milk samples and it was found that only 17 % of the samples were incorrectly classified. An interesting work to identify the origins of Mexican coffees on the basis of the constituent cations of group 1 and group 2 elements was conducted using ANN by Muniz-Valencia et al. [136] They reported a good prediction accuracy of 93 %. Moreover, ANN was proved to be a better approach than *k*-NN in establishing a calibration relationship between color of cooked food and its quality though computationally more complex. [137] Support vector regression along with regression trees were shown to be more robust than other multivariate calibration methods in the classification of different types of tequilas based on UV-Visible spectroscopic data. [79]

In order to investigate the effects of plasma treatment and subsequent pasteurization on the stability chokeberry samples consisting of hydroxycinnamic acids, anthocyanins and flavonols which are complex polyphenolic compounds, PCA followed by regression techniques was used for parameter estimation. They reported that hydroxycinnamic acid concentrations increased with a simultaneous reduction of anthocyanins in the samples treated with plasma while pasteurization

had the opposite effect. The chemometric methods enabled optimization of the treatment process for determining the phenolic stability of the chokeberry samples.

Studies have also been conducted on the detection of food spoilage using chemometric methods. With the help of dimension reduction through PCA and pattern recognition through HCA in combination with PLSR, compounds responsible for spoilage of foods through promotion of bacterial growth and volatility were identified to be CH_3COOH, $(CH_3)_2S$, $N(CH_3)_3$ by Kuuliala et al. [138]

All these above-mentioned examples show the importance of chemometrics for analysis of data from basic and advanced hyphenated experimental techniques in different fields of analytical science. The next section will highlight the future prospects of chemometric approaches so as to explore avenues to reduce the gap between experimentalists and data scientists.

2.4 Future prospects in chemometrics

Chemometrics offers numerous pathways to progress by building on the existing methods due to its strong mathematical basis. This can be in the form of improvements and tweaks in current algorithms, incorporating additional constraints, tactical combinations of different methods and tests to check their reliability as well. This section will focus on reviewing some of these advancements in chemometric techniques with an eye on their future applications.

A common technique used for function optimization is simplex, which is based on the Nelder and Mead algorithm. [139] It essentially fine-tunes the initial values of the function parameters as provided by coupled simulated annealing (CSA) [140] or by a trial and error approach. [51] In an *m*-dimensional space, $m + 1$ vectors are taken as the vertices for the search space in simplex, i.e. a triangle in a 2-D, a pyramid in a 3-D and so on. Some research works suggested an improvement to this method where in the updating of the search space, the better (n-1) points needed to be considered rather than replacing one vertex with the new point. [141,142] In addition, if it was found that the function optimum lay within the present simplex space, a Lagrange interpolative fit could be employed. Multiple optima and boundary conditions could be treated better in this algorithm. This updated method could be employed in tuning the hyper-parameters like regularization and kernel parameters in a machine learning approach like SVM regression adopting the various kernels mentioned in the previous section.

In the MCR-ALS algorithm, the deconvoluted spectra of the individual pseudo-components contain regions that cannot be interpreted directly by assigning to particular compound classes. Tefera et al. [30] recommended the use of Borgen plots that checks the validity of the optimized solution by restricting the data to be nonnegative and constructing appropriate geometrical regions. Borgen and Kowalski [143] reported that they could be particularly useful for three-component systems. The other issue in curve resolution techniques is rotational ambiguity, where multiple combinations of concentration and spectra can reproduce the original data matrix. To deal with this, additional constraints incorporating chemical knowledge and intuition about the system in focus can be included in the modeling apart from the commonly used closure and nonnegativity constraints.

Genetic algorithms (GA) provide a vast window of opportunities for variable selection and function optimization. They are derived from the motion pattern of living species like the structure of a flock of birds, ant colony formation, swarm of honey bees locating honey, etc. Koljonen et al. [82] provide a thorough review of the GAs suitable for near infrared spectra in combination with other chemometric techniques to process the data. They provide insights into the future application of GAs in the field as well. GAs enable the user to gain good control over the chemometric techniques they operate on. Two interesting algorithms that are gaining popularity over recent times are ant colony optimization (ACO) and particle swarm optimization (PSO). ACO takes the idea from the fact that ants leave traces of chemical compounds called pheromones during the process of food collection so that other ants can follow. [144] It is similar to CSA and an improvement over gradient descent optimization. Particles resembling real ants are involved in finding the solution.

PSO is partly based on the motion of a swarm of bees that act as the particles and their number and the corresponding search-space in which they operate are predetermined. [145] The objective is to find the best value for the cost function based on the current position of the particle. At each step, the position and velocity of each particle is updated and the value of the cost function is calculated and compared with the values determined using the local and global best particle positions. Both PSO and ACO are metaheuristic in nature as they involve minimum assumptions but in PSO, an optimal solution is not guaranteed. Instead of ALS, PSO can be incorporated with curve resolution to analyze complex mixtures like bitumen and heavy oil. PSO is also different from a classical optimization technique as it is not based on a gradient method that requires

differentiation of the data. PSO embedded in ALS has been used as an optimization technique to resolve the concentration and spectral profiles for pseudo-components from the FTIR spectra for thermally cracked products from Athabasca bitumen in the final phase of this study (Chapter 6). The results were compared to that of ALS with the objective of proposing logical reaction pathways for cracking conversion of Athabasca bitumen.

An alternate method to the classical least squares regression (CLS) is least median squares (LMS) regression that minimizes the median of the squared residuals rather than the sum. [146] The disadvantage of this estimator is its slow convergence but can be improved by a least trimmed squares (LTS) estimator proposed by Rousseuw [146] but it requires larger computation time. An interesting study by Shinzawa et al. [147] showed that LMS along with PSO was more robust and accurate than the CLS for optical spectra and was able to deal with larger noise and data contamination. LMS can be considered as a substitute for CLS in further calibration methods for multidimensional data.

Moving ahead, one important focus of chemometric techniques is the integration of chemometrics with cheminformatics ultimately deriving chemical knowledge from the developed statistical models. For example, in the context of catalytic reactions, information deals with predicting binding energy of the adsorbate on the catalyst material. On the other hand, knowledge means applying the information on binding energy to reveal the underlying reaction mechanism and the corresponding structure of the active sites by developing a chemical master equation describing the entire system including the environmental interaction. [4] Chemical intuition can also be incorporated into the chemometric models and one of the methods is facilitated by the software, rule input network generation (RULE) that is inspired from electric circuits and is based on graph theory. [148]

Artificial intelligence (AI) systems utilizing machine learning algorithms have been leading the way in sensor building technology over the last few years. Process automation is one of the most important goals of AI and interestingly, chemometric techniques like pattern recognition, calibration, regression and machine learning form the building blocks of AI. However, the advantages are its ability to deal with big data effectively and that they do not require any human intervention whatsoever. Moreover, they also facilitate reinforcement learning (apart from supervised and unsupervised learning methods), where feedback is given after each classification step though class memberships are not defined in the original data matrix.

Thus, all the current chemometric methods in combination with recent advancements highlighted in this section (though not comprehensive) make the future of chemometric practices in analytical chemistry bright.

2.5 Conclusions

The techniques in chemometrics fall under the fundamental categories of classification and regression and are reviewed in this chapter. The focus was not to explore the mathematical detail of each method but to know the key aspects of each one and obtain a general insight on their applications. Specific highlights based on the reviewed literature in this work is given below:

a. Preprocessing methods like transformation, centering and scaling are applied before classification or regression to reduce heteroscedastic effects, remove offsets and bring the data to the same scale irrespective of the units.
b. PCA and HCA perform unsupervised classification and extract linear models. MDS was shown to be a development over PCA and is a nonlinear dimension reduction technique. PARAFAC, which is also an unsupervised method, factorizes three-way data like fluorescence and chromatograms at different process conditions.
c. Supervised classification techniques like LDA, PLS-DA and *k*-NN are hard classification methods that require the objects to have predetermined class assignments but can model nonlinear data. On the other hand, SIMCA is a soft-classification technique and in general, supervised classification methods utilize PCA initially when variables exceed the number of samples by a large quantity.
d. Correlation analysis and variable selection increase model robustness and decrease computational time and possibility of overfitting when performed before calibration and regression techniques. Cross validation is performed within the training dataset to improve the generalizability of the model and also to tune any hyper-parameters for the regression like those for SVM.
e. The advantage of resolution methods was that no prior chemical knowledge is required for extracting the spectral and concentration optimized profiles though simple mathematical constraints like nonnegativity and closure are usually applied. Inverse least

squares criterion can also be applied for optimization for cases where concentration depended on absorbance rather than the conventional way.

f. Integration of experimentation with computational processes were shown to be advantageous in terms of facilitating optimization of process conditions and enabling the use of autosamplers in hyphenated techniques like GC-MS and ICP-OES.

g. Applications of chemometrics in various fields of analytical chemistry were profound. Pattern recognition methods like HCA in combination with PCA was shown to detect failure in material science and identify the presence of cancer cells and tumors in medicine. Dissociation constants for dibasic compounds could be estimated through nonlinear regression in the field of physical chemistry. Important uses of supervised classification in petroleum was to differentiate crude oils, identify their source rocks and also compound identification from exhaust fumes. OTR was also shown to track the archeological history through analysis of coin colors. MCR techniques were found to be most useful in environmental chemistry while quantifying ESR spectra through spin labeling was important in biological samples. LDA and ANN were found to be better than PCA in various parts of food science including detecting adulterants, tracking rate of food spoilage and geographical classification of food constituents.

h. Improvements in current minimization criterions like use of least median squares instead of classical least squares were suggested. Additions to the existing optimization techniques like Borgen plots to MCR-ALS, SIMPLEX optimization for tuning in SVM and embedding PSO along with ALS for resolution methods were also considered to be useful in improving the respective models. Chemoinformatics and AI were seen to be the future of chemometric methods.

2.6 References

(1) Rácz, A.; Bajusz, D.; Héberger, K. Chemometrics in Analytical Chemistry. In *Applied Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2018; pp 471–499.

(2) Farrusseng, D.; Baumes, L.; Mirodatos, C. Data Management for Combinatorial Heterogeneous Catalysis: Methodology and Development of Advanced Tools. In *High-*

Throughput Analysis; Springer: Boston, MA, 2003; pp 551–579.

(3) Caruthers, J. M.; Lauterbach, J. A.; Thomson, K. T.; Venkatasubramanian, V.; Snively, C. M.; Bhan, A.; Katare, S.; Oskarsdottir, G. Catalyst Design: Knowledge Extraction from High-Throughput Experimentation. *J. Catal.* **2003**, *216*, 98–109.

(4) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403–7429.

(5) Kumar, N.; Bansal, A.; Sarma, G. S.; Rawal, R. K. Chemometrics Tools Used in Analytical Chemistry: An Overview. *Talanta* **2014**, *123*, 186–199.

(6) Szymańska, E.; Gerretzen, J.; Engel, J.; Geurts, B.; Blanchet, L.; Buydens, L. M. Chemometrics and Qualitative Analysis Have a Vibrant Relationship. *TrAC Trends Anal. Chem.* **2015**, *69*, 34–51.

(7) Liang, Y.Z.; Xie, P.; Chan, K. Quality Control of Herbal Medicines. *J. Chromatogr. B* **2004**, *812*, 53–70.

(8) Larson-Hall, J. Levels of Measurement of Variables in a Guide to Doing Statistics. In *Second language research using SPSS*; Taylor & Francis: New York, 2010; pp 33–35.

(9) Osborne, B. G. Near-Infrared Spectroscopy in Food Analysis. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: Chichester, UK, 2000.

(10) Igne, B.; Ciurczak, E. W. *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy*; CRC Press: Boca Raton, 2014.

(11) Ertaş, M.; Hakki Alma, M. Pyrolysis of Laurel (Laurus Nobilis L.) Extraction Residues in a Fixed-Bed Reactor: Characterization of Bio-Oil and Bio-Char. *J. Anal. Appl. Pyrolysis* **2010**, *88*, 22–29.

(12) Guo, X.; Wang, S.; Guo, Z.; Liu, Q.; Luo, Z.; Cen, K. Pyrolysis Characteristics of Bio-Oil Fractions Separated by Molecular Distillation. *Appl. Energy* **2010**, *87*, 2892–2898.

(13) Owen, T. *Fundamentals of UV-Visible Spectroscopy: A Primer*; Hewlett Packard, 1996.

(14) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds.*; John Wiley & Sons: New York, 2014.

(15) Mathiesen, M. D.; Lubeck, A. J. Improving Accuracy in the Determination of Aromatics in Gasoline by Gas Chromatography-Mass Spectrometry. *J. Chromatogr. Sci.* **1998**, *36*, 449–456.

(16) Willet, P. *Similarity and Clustering in Chemical Information*; Wiley: New York, 1987.

(17) Wang, Y.P.; Zou, Y.R.; Shi, J.T.; Shi, J. Review of the Chemometrics Application in Oil-Oil and Oil-Source Rock Correlations. *J. Nat. Gas Geosci.* **2018**, *3*, 217–232.

(18) Van de Vijver, R.; Devocht, B. R.; Van Geem, K. M.; Thybaut, J. W.; Marin, G. B. Challenges and Opportunities for Molecule-Based Management of Chemical Processes. *Curr. Opin. Chem. Eng.* **2016**, *13*, 142–149.

(19) Dellamorte, J. C.; Barteau, M. A.; Lauterbach, J. Opportunities for Catalyst Discovery and Development: Integrating Surface Science and Theory with High Throughput Methods. *Surf. Sci.* **2009**, *603*, 1770–1775.

(20) Carter, S.; Fisher, A.; Garcia, R.; Gibson, B.; Lancaster, S.; Marshall, J.; Whiteside, I. Atomic Spectrometry Update. Review of Advances in the Analysis of Metals, Chemicals and Functional Materials. *J. Anal. At. Spectrom.* **2015**, *30*, 2249–2294.

(21) Xie, L.; Ying, Y.; Ying, T.; Yu, H.; Fu, X. Discrimination of Transgenic Tomatoes Based on Visible/near-Infrared Spectra. *Anal. Chim. Acta* **2007**, *584*, 379–384.

(22) Kowalkowski, T.; Zbytniewski, R.; Szpejna, J.; Buszewski, B. Application of Chemometrics in River Water Classification. *Water Res.* **2006**, *40*, 744–752.

(23) Lankmayr, E.; Mocak, J.; Serdt, K.; Balla, B.; Wenzl, T.; Bandoniene, D.; Gfrerer, M.; Wagner, S. Chemometrical Classification of Pumpkin Seed Oils Using UV–Vis, NIR and FTIR Spectra. *J. Biochem. Biophys. Methods* **2004**, *61*, 95–106.

(24) Danezis, G. P.; Tsagkaris, A. S.; Camin, F.; Brusic, V.; Georgiou, C. A. Food Authentication: Techniques, Trends & Emerging Approaches. *TrAC Trends Anal. Chem.* **2016**, *85*, 123–132.

(25) Munck, L.; Nørgaard, L.; Engelsen, S. B.; Bro, R.; Andersson, C. A. Chemometrics in Food Science—a Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 31–60.

(26) Nunes, C. A.; Alvarenga, V. O.; de Souza Sant'Ana, A.; Santos, J. S.; Granato, D. The Use of Statistical Software in Food Science and Technology: Advantages, Limitations and Misuses. *Food Res. Int.* **2015**, *75*, 270–280.

(27) Martin, Y. C.; Hackbarth, J. J. Theoretical Model-Based Equations for the Linear Free Energy Relations of the Biological Activity of Ionizable Substances. 1. Equilibrium-Controlled Potency. *J. Med. Chem.* **1976**, *19*, 1033–1039.

(28) Telnaes, N.; Dahl, B. Oil-Oil Correlation Using Multivariate Techniques. *Org. Geochem.* **1986**, *10*, 425–432.

(29) Øygard, K.; Grahl-Nielsen, O.; Ulvøen, S. Oil/Oil Correlation by Aid of Chemometrics. *Org. Geochem.* **1984**, *6*, 561–567.

(30) Tefera, D. T.; Agrawal, A.; Yañez Jaramillo, L. M.; De Klerk, A.; Prasad, V. Self-Modeling Multivariate Curve Resolution Model for Online Monitoring of Bitumen Conversion Using Infrared Spectroscopy. *Ind. Eng. Chem. Res.* **2017**, *56*, 10756–10769.

(31) Sivaramakrishnan, K.; Nie, J.; de Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, *57*, 13156–13176.

(32) Siddiquee, M. N.; Sivaramakrishnan, K.; Wu, Y.; de Klerk, A.; Nazemifard, N. A Statistical Approach Dealing with Multicollinearity among Predictors in Microfluidic Reactor Operation to Control Liquid-Phase Oxidation Selectivity. *React. Chem. Eng.* **2018**, *3*, 972–990.

(33) Kiralj, R.; Ferreira, M. M. C. The Past, Present, and Future of Chemometrics Worldwide: Some Etymological, Linguistic, and Bibliometric Investigations. *J. Chemom.* **2006**, *20*, 247–272.

(34) Martens, H.; Næs, T. Multivariate Calibration. In *Chemometrics*; Springer Netherlands: Dordrecht, 1984; pp 147–156.

(35) Sjöström, M.; Wold, S.; Lindberg, W.; Persson, J.-Å.; Martens, H. A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables. *Anal. Chim. Acta* **1983**, *150*, 61–70.

(36) Haykin, S. S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall: Upper Saddle River, N.J., 1998.

(37) Press, W. H.; Teukolsky, S. A. Savitzky-Golay Smoothing Filters. *Comput. Phys.* **1990**, *4*, 669–672.

(38) Afseth, N. K.; Kohler, A. Extended Multiplicative Signal Correction in Vibrational Spectroscopy, a Tutorial. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 92–99.

(39) Kvalheim, O. M.; Brakstad, F.; Liang, Y. Preprocessing of Analytical Profiles in the Presence of Homoscedastic or Heteroscedastic Noise. *Anal. Chem.* **1994**, *66*, 43–51.

(40) Van Den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M.

J. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genomics* **2006**, *7*, 142.

(41) Eriksson, L. *Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*; Umetrics: Sweden, 1999.

(42) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J. C.; Jellema, R. H. Fusion of Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2005**, *77*, 6729–6736.

(43) Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Improved Analysis of Multivariate Data by Variable Stability Scaling: Application to NMR-Based Metabolic Profiling. *Anal. Chim. Acta* **2003**, *490*, 265–276.

(44) Dobos, L.; Abonyi, J. On-Line Detection of Homogeneous Operation Ranges by Dynamic Principal Component Analysis Based Time-Series Segmentation. *Chem. Eng. Sci.* **2012**, *75*, 96–105.

(45) Lavine, B. K.; Davidson, C. E.; Moores, A. J. Genetic Algorithms for Spectral Pattern Recognition. *Vib. Spectrosc.* **2002**, *28*, 83–95.

(46) Smith, G. L.; Mlynczak, P. E.; Potter, G. L. A Technique Using Principal Component Analysis to Compare Seasonal Cycles of Earth Radiation from CERES and Model Computations. *J. Geophys. Res. Atmos.* **2012**, *117*.

(47) Mandel, J. Use of the Singular Value Decomposition in Regression Analysis. *Am. Stat.* **1982**, *36*, 15–24.

(48) Camacho, J.; Picó, J.; Ferrer, A. Data Understanding with PCA: Structural and Variance Information Plots. *Chemom. Intell. Lab. Syst.* **2010**, *100*, 48–56.

(49) Jolliffe, I. T. Mathematical and Statistical Properties of Population Principal Components. In *Principal Component Analysis*; Springer Series in Statistics; Springer-Verlag: New York, 2002.

(50) Shao, Y.; He, Y. Visible/Near Infrared Spectroscopy and Chemometrics for the Prediction of Trace Element (Fe and Zn) Levels in Rice Leaf. *Sensors* **2013**, *13*, 1872–1883.

(51) Yeganeh, B.; Motlagh, M. S. P.; Rashidi, Y.; Kamalan, H. Prediction of CO Concentrations Based on a Hybrid Partial Least Square and Support Vector Machine Model. *Atmos. Environ.* **2012**, *55*, 357–365.

(52) Granato, D.; Putnik, P.; Kovačević, D. B.; Santos, J. S.; Calado, V.; Rocha, R. S.; Cruz, A. G. Da; Jarvis, B.; Rodionova, O. Y.; Pomerantsev, A. Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing. *Compr. Rev. Food Sci. Food Saf.* **2018**, *17*, 663–677.

(53) Wold, S.; Sjöström, M. Chemometrics, Present and Future Success. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 3–14.

(54) Zimmermann, B. Characterization of Pollen by Vibrational Spectroscopy. *Appl. Spectrosc.* **2010**, *64*, 1364–1373.

(55) Mok, P. Y.; Huang, H. Q.; Kwok, Y. L.; Au, J. S. A Robust Adaptive Clustering Analysis Method for Automatic Identification of Clusters. *Pattern Recognit.* **2012**, *45*, 3017–3033.

(56) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An Efficient K-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892.

(57) Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling*; Springer: New York, 2005.

(58) Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. *Psychometrika* **1962**, *27*, 125–140.

(59) Kruskal, J. B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* **1964**, *29*, 1–27.

(60) Wang, Y. P.; Zhang, F.; Zou, Y. R.; Zhan, Z. W.; Peng, P. Chemometrics Reveals Oil Sources in the Fangzheng Fault Depression, NE China. *Org. Geochem.* **2016**, *102*, 1–13.

(61) Aqualog Software User's Guide for Version 3.6. Horiba Scientific 2012.

(62) Bro, R. PARAFAC. Tutorial and Applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171.

(63) Marx, B. D.; Eilers, P. H. C.; Li, B. Multidimensional Single-Index Signal Regression. *Chemom. Intell. Lab. Syst.* **2011**, *109*, 120–130.

(64) ten Berge, J. M. *Least Squares Optimization in Multivariate Analysis*; DSWO Press: Leiden, 1993.

(65) Tax, D.; Duin, R. Outlier Detection Using Classifier Instability. *Lect. notes Comput. Sci.* **1998**, *1451*, 593–601.

(66) Berrueta, L. A.; Alonso-Salces, R. M.; Héberger, K. Supervised Pattern Recognition in Food Analysis. *J. Chromatogr. A* **2007**, *1158*, 196–214.

(67) Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: San Diego, 1990.

(68) Subramanian, J.; Simon, R. Overfitting in Prediction Models – Is It a Problem Only in High Dimensions? *Contemp. Clin. Trials* **2013**, *36*, 636–641.

(69) Ståhle, L.; Wold, S. Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem: A Monte Carlo Study. *J. Chemom.* **1987**, *1*, 185–196.

(70) Cebi, N.; Durak, M. Z.; Toker, O. S.; Sagdic, O.; Arici, M. An Evaluation of Fourier Transforms Infrared Spectroscopy Method for the Classification and Discrimination of Bovine, Porcine and Fish Gelatins. *Food Chem.* **2016**, *190*, 1109–1115.

(71) Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification And Regression Trees*; Routledge: New York, 2017.

(72) Kvalheim, O. M. Latent-Structure Decompositions (Projections) of Multivariate Data. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 283–290.

(73) Christie, O. H. J.; Rácz, A.; Elek, J.; Héberger, K. Classification and Unscrambling a Class-inside-Class Situation by Object Target Rotation: Hungarian Silver Coins of the Árpád Dynasty, AD 997-1301. *J. Chemom.* **2014**, *28*, 287–292.

(74) Wold, M. S. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In *Chemometrics theory and application, American Chemical Society Symposium Series*; American Chemical Society: Washington, D. C., 1977; pp 243–282.

(75) Gondim, C. D. S.; Junqueira, R. G.; Souza, S. V. C. D.; Ruis´anchez, I.; Callao, M. P. Detection of Several Common Adulterants in Raw Milk by MID-Infrared Spectroscopy and One-Class and Multi-Class Multivariate Strategies. *Food Chem.* **2017**, *230*, 68–75.

(76) Gurbanov, R.; Gozen, A. G.; Severcan, F. Rapid Classification of Heavy Metal-Exposed Freshwater Bacteria by Infrared Spectroscopy Coupled with Chemometrics Using Supervised Method. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2018**, *189*, 282–290.

(77) Kanik, E. A.; Orekici Temel, G.; Erdogan, S.; Ersoz Kaya, I. Affected States Soft Independent Modeling by Class Analogy from the Relation Between Independent Variables, Number of Independent Variables and Sample Size. *Balkan Med. J.* **2013**, *30*, 28–32.

(78) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley & Sons: New York, 1998.

(79) P´erez-Caballero, G.; Andrade, J. M.; Olmos, P.; Molina, Y.; Jim´enez, I.; Dur´an, J. J.; Miguel-Cruz, F. Authentication of Tequilas Using Pattern Recognition and Supervised Classification. *TrAC Trends Anal. Chem.* **2017**, *94*, 117–129.

(80) Niazi, A.; Goodarzi, M.; Yazdanipour, A. A Comparative Study between Least-Squares Support Vector Machines and Partial Least Squares in Simultaneous Spectrophotometric Determination of Cypermethrin, Permethrin and Tetramethrin. *J. Braz. Chem. Soc.* **2008**, *19*, 536–542.

(81) Asuero, A. G.; Sayago, A.; González, A. G. The Correlation Coefficient: An Overview. *Crit. Rev. Anal. Chem.* **2006**, *36*, 41–59.

(82) Koljonen, J.; Nordling, T. E. M.; Alander, J. T. A Review of Genetic Algorithms in near Infrared Spectroscopy and Chemometrics: Past and Future. *J. Near Infrared Spectrosc.* **2008**, *16*, 189–197.

(83) Hoerl, A. E.; Kennard, R. W. Ridge Regression:Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.

(84) Kalivas, J. H. Overview of Two-Norm (L2) and One-Norm (L1) Tikhonov Regularization Variants for Full Wavelength or Sparse Spectral Multivariate Calibration Models or Maintenance. *J. Chemom.* **2012**, *26*, 218–230.

(85) Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. Interval Partial Least-Squares Regression (IPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419.

(86) Geladi, P.; Kowalski, B. R. An Example of 2-Block Predictive Partial Least-Squares Regression with Simulated Data. *Anal. Chim. Acta* **1986**, *185*, 19–32.

(87) Frank, I. E.; Friedman, J. H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **1993**, *35*, 109–135.

(88) Balabin, R. M.; Lomakina, E. I. Support Vector Machine Regression (SVR/LS-SVM)—an Alternative to Neural Networks (ANN) for Analytical Chemistry? Comparison of Nonlinear Methods on near Infrared (NIR) Spectroscopy Data. *Analyst* **2011**, *136*, 1703.

(89) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 2000.

(90) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.*; MIT Press: Cambridge, MA, 2002.

(91) Chauchard, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V. Application of LS-

SVM to Non-Linear Phenomena in NIR Spectroscopy: Development of a Robust and Portable Sensor for Acidity Prediction in Grapes. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 141–150.

(92) Miller, A. *Subset Selection in Regression*; Chapman and Hall: London, 1990.

(93) Norgaard, L.; Leardi, R. *IToolbox Manual*; 2005.

(94) Rajalahti, T.; Arneberg, R.; Berven, F. S.; Myhr, K.-M.; Ulvik, R. J.; Kvalheim, O. M. Biomarker Discovery in Mass Spectral Profiles by Means of Selectivity Ratio Plot. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35–48.

(95) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: River Edge, NJ, 2002.

(96) Wahba, G. Numerical Methods. In *Spline Models for Observational Data*; Society for Industrial and Applied Mathematics: Philadelphia, Pennsylvania, 1990; pp 135–143.

(97) Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Chapman and Hall: New York, 1993.

(98) Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated Double Cross Validation. *J. Chemom.* **2009**, *23*, 160–171.

(99) Cui, C.; Fearn, T. Comparison of Partial Least Squares Regression, Least Squares Support Vector Machines, and Gaussian Process Regression for a near Infrared Calibration. *J. Near Infrared Spectrosc.* **2017**, *25*, 5–14.

(100) de Juan, A.; Tauler, R. Chemometrics Applied to Unravel Multicomponent Processes and Mixtures: Revisiting Latest Trends in Multivariate Resolution. *Anal. Chim. Acta* **2003**, *500*, 195–210.

(101) Ho, C. N.; Christian, G. D.; Davidson, E. R. Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer. *Anal. Chem.* **1978**, *50*, 1108–1113.

(102) Maeder, M. Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks. *Anal. Chem.* **1987**, *59*, 527–530.

(103) Tauler, R.; Marques, I.; Casassas, E. Multivariate Curve Resolution Applied to Three-way Trilinear Data: Study of a Spectrofluorimetric Acid–Base Titration of Salicylic Acid at Three Excitation Wavelengths. *J. Chemom.* **1998**, *12*, 55–75.

(104) Sánchez, F. C.; Toft, J.; van den Bogaert, B.; Massart, D. L. Orthogonal Projection

Approach Applied to Peak Purity Assessment. *Anal. Chem.* **1996**, *68*, 79–85.

(105) Keller, H. R.; Massart, D. L. Peak Purity Control in Liquid Chromatography with Photodiode-Array Detection by a Fixed Size Moving Window Evolving Factor Analysis. *Anal. Chim. Acta* **1991**, *246*, 379–390.

(106) Christiansen, T. F.; Busch, J. E.; Krogh, S. C. Successive Determinations of Calcium and Magnesium in Drinking Water by Complexometric, Potentiometric Digital Titration to Two Equivalence Points. *Anal. Chem.* **1976**, *48*, 1051–1056.

(107) Wu, A. H. B.; Malmstadt, H. V. Versatile Microcomputer-Controlled Titrator. *Anal. Chem.* **1978**, *50*, 2090–2096.

(108) Ruzicka, J.; Hansen, E. H. *Flow Injection Analysis*; Wiley: New York, 1981.

(109) Betteridge, D.; Cheng, W. C.; Dagless, E. L.; David, P.; Goad, T. B.; Deans, D. R.; Newton, D. A.; Pierce, T. B. An Automated Viscometer Based on High-Precision Flow Injection Analysis. Part II. Measurement of Viscosity and Diffusion Coefficients. *Analyst* **1983**, *108*, 17.

(110) Stieg, S.; Nieman, T. A. Application of a Microcomputer Controlled Chemiluminescence Research Instrument to the Simultaneous Determination of Cobalt (II) and Silver (I) by Gallic Acid Chemiluminescence. *Anal. Chem.* **1980**, *52*, 800–804.

(111) Sly, T. J.; Betteridge, D.; Wibberley, D.; Porter, D. G. An Improved Flow-through Phototransducer. *J. Automat. Chem.* **1982**, *4*, 186–189.

(112) Belchamber, R. M.; Betteridge, D.; Chow, Y. T.; Sly, T. J.; Wade, A. P. The Application of Computers in Chemometrics and Analytical Chemistry. *Anal. Chim. Acta* **1983**, *150*, 115–128.

(113) Marec, A.; Thomas, J.-H.; El Guerjouma, R. Damage Characterization of Polymer-Based Composite Materials: Multivariable Analysis and Wavelet Transform for Clustering Acoustic Emission Data. *Mech. Syst. Signal Process.* **2008**, *22*, 1441–1464.

(114) Moran, M. G.; Kowalski, B. R. Image Analysis. *Anal. Chem.* **1979**, *51*, 776A-788A.

(115) Gordon, R.; Herman, G. T.; Johnson, S. A. Image Reconstruction from Projections. *Sci. Am.* **1975**, *233*, 56–71.

(116) Martin, Y. C.; Hackbarth, J. J. Examples of the Application of Nonlinear Regression Analysis to Chemical Data. In *Chemometrics: Theory and Application ACS Symposium Series1*; American Chemical Society: Washington, D. C., 1977; pp 153–164.

(117) Albert, A.; Serjeant, E. P. *The Determination of Ionization Constants*; Chapman and Hall: London, 1971.

(118) Eide, I.; Neverdal, G.; Thorvaldsen, B.; Grung, B.; Kvalheim, O. M. Toxicological Evaluation of Complex Mixtures by Pattern Recognition: Correlating Chemical Fingerprints to Mutagenicity. *Environ. Health Perspect.* **2002**, *110*, 985–988.

(119) Eide, I.; Zahlsen, K. A Novel Method for Chemical Fingerprinting of Oil and Petroleum Products Based on Electrospray Mass Spectrometry and Chemometrics. *Energy Fuels* **2005**, *19*, 964–967.

(120) Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. Reading Chemical Fine Print: Resolution and Identification of 3000 Nitrogen-Containing Aromatic Compounds from a Single Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrum of Heavy Petroleum Crude Oil. *Energy Fuels* **2001**, *15*, 492–498.

(121) Wang, Y. P.; Zhang, F.; Zou, Y. R.; Sun, J. N.; Lin, X. H.; Liang, T. Oil Source and Charge in the Wuerxun Depression, Hailar Basin, Northeast China: A Chemometric Study. *Mar. Pet. Geol.* **2018**, *89*, 665–686.

(122) Mudge, S. Aspects of Hydrocarbon Fingerprinting Using PLS—New Data From Prince William Sound. *Environ. Forensics* **2002**, *3*, 323–329.

(123) Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. Chromatographic Alignment by Warping and Dynamic Programming as a Pre-Processing Tool for PARAFAC Modelling of Liquid Chromatography–Mass Spectrometry Data. *J. Chromatogr. A* **2002**, *961*, 237–244.

(124) Comas, E.; Gimeno, R. A.; Ferré, J.; Marcé, R. M.; Borrull, F.; Rius, F. X. Quantification from Highly Drifted and Overlapped Chromatographic Peaks Using Second-Order Calibration Methods. *J. Chromatogr. A* **2004**, *1035*, 195–202.

(125) Ho, C. N.; Christian, G. D.; Davidson, E. R. Application of the Method of Rank Annihilation to Fluorescent Multicomponent Mixtures of Polynuclear Aromatic Hydrocarbons. *Anal. Chem.* **1980**, *52*, 1071–1079.

(126) De Braekeleer, K.; de Juan, A.; Massart, D. L. Purity Assessment and Resolution of Tetracycline Hydrochloride Samples Analysed Using High-Performance Liquid Chromatography with Diode Array Detection. *J. Chromatogr. A* **1999**, *832*, 67–86.

(127) Salau, J. S.; Honing, M.; Tauler, R.; Barceló, D. Resolution and Quantitative Determination of Coeluted Pesticide Mixtures in Liquid Chromatography–Thermospray Mass Spectrometry by Multivariate Curve Resolution. *J. Chromatogr. A* **1998**, *795*, 3–12.

(128) Dinç, E.; Baydan, E.; Kanbur, M.; Onur, F. Spectrophotometric Multicomponent Determination of Sunset Yellow, Tartrazine and Allura Red in Soft Drink Powder by Double Divisor-Ratio Spectra Derivative, Inverse Least-Squares and Principal Component Regression Methods. *Talanta* **2002**, *58*, 579–594.

(129) Chen, Y.; Zhu, S. B.; Xie, M. Y.; Nie, S. P.; Liu, W.; Li, C.; Gong, X. F.; Wang, Y. X. Quality Control and Original Discrimination of Ganoderma Lucidum Based on High-Performance Liquid Chromatographic Fingerprints and Combined Chemometrics Methods. *Anal. Chim. Acta* **2008**, *623*, 146–156.

(130) Rohman, A.; Che Man, Y. B. Application of Fourier Transform Infrared (FT-IR) Spectroscopy Combined with Chemometrics for Authentication of Cod-Liver Oil. *Vib. Spectrosc.* **2011**, *55*, 141–145.

(131) Eaton, G. R.; Eaton, S. S.; Barr, D. P.; Weber, R. T. *Quantitative EPR*; Springer: Vienna, 2010.

(132) Koskinen, J. R.; Kowalski, B. R. Analysis of the Electron Spin Resonance of Spin Labels Using Chemometric Methods. In *Chemometrics: Theory and Application ACS Symposium Series*; American Chemical Society: Washington, D. C., 1977; pp 117–126.

(133) Nascimento, K. S. do; Gasparotto Sattler, J. A.; Lauer Macedo, L. F.; Serna González, C. V.; Pereira de Melo, I. L.; da Silva Araújo, E.; Granato, D.; Sattler, A.; de Almeida-Muradian, L. B. Phenolic Compounds, Antioxidant Capacity and Physicochemical Properties of Brazilian Apis Mellifera Honeys. *LWT* **2018**, *91*, 85–94.

(134) Grunert, T.; Stephan, R.; Ehling-Schulz, M.; Johler, S. Fourier Transform Infrared Spectroscopy Enables Rapid Differentiation of Fresh and Frozen/Thawed Chicken. *Food Control* **2016**, *60*, 361–364.

(135) D'Archivio, A. A.; Giannitto, A.; Maggi, M. A.; Ruggieri, F. Geographical Classification of Italian Saffron (Crocus Sativus L.) Based on Chemical Constituents Determined by High-Performance Liquid-Chromatography and by Using Linear Discriminant Analysis. *Food Chem.* **2016**, *212*, 110–116.

(136) Muñiz-Valencia, R.; Jurado, J. M.; Ceballos-Magaña, S. G.; Alcázar, Á.; Hernández-Díaz,

J. Characterization of Mexican Coffee According to Mineral Contents by Means of Multilayer Perceptrons Artificial Neural Networks. *J. Food Compos. Anal.* **2014**, *34*, 7–11.

(137) O'Farrell, M.; Lewis, E.; Flanagan, C.; Lyons, W.; Jackman, N. Comparison of K-NN and Neural Network Methods in the Classification of Spectral Data from an Optical Fibre-Based Sensor System Used for Quality Control in the Food Industry. *Sensors Actuators B Chem.* **2005**, *111–112*, 354–362.

(138) Kuuliala, L.; Abatih, E.; Ioannidis, A.-G.; Vanderroost, M.; De Meulenaer, B.; Ragaert, P.; Devlieghere, F. Multivariate Statistical Analysis for the Identification of Potential Seafood Spoilage Indicators. *Food Control* **2018**, *84*, 49–60.

(139) Olsson, D. M.; Nelson, L. S. The Nelder-Mead Simplex Procedure for Function Minimization. *Technometrics* **1975**, *17*, 45–51.

(140) Xavier-De-Souza, S.; Suykens, J. A. K.; Vandewalle, J.; Bolle, D. Coupled Simulated Annealing. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **2010**, *40*, 320–335.

(141) Routh, M. W.; Swartz, P. A.; Denton, M. B. Performance of the Super Modified Simplex. *Anal. Chem.* **1977**, *49*, 1422–1428.

(142) Ryan, P. B.; Barr, R. L.; Todd, H. D. Simplex Techniques for Nonlinear Optimization. *Anal. Chem.* **1980**, *52*, 1460–1467.

(143) Borgen, O. S.; Kowalski, B. R. An Extension of the Multivariate Component-Resolution Method to Three Components. *Anal. Chim. Acta* **1985**, *174*, 1–26.

(144) Shmygelska, A.; Hoos, H. H. An Ant Colony Optimisation Algorithm for the 2D and 3D Hydrophobic Polar Protein Folding Problem. *BMC Bioinformatics* **2005**, *6*, 1–22.

(145) Shinzawa, H.; Jiang, J.-H.; Iwahashi, M.; Noda, I.; Ozaki, Y. Self-Modeling Curve Resolution (SMCR) by Particle Swarm Optimization (PSO). *Anal. Chim. Acta* **2007**, *595*, 275–281.

(146) Rousseeuw, P. J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **2012**, *79*, 871–880.

(147) Shinzawa, H.; Jiang, J.-H.; Iwahashi, M.; Ozaki, Y. Robust Curve Fitting Method for Optical Spectra by Least Median Squares (LMedS) Estimator with Particle Swarm Optimization (PSO). *Anal. Sci.* **2007**, *23*, 781–785.

(148) Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Description of RING. *Comput. Chem. Eng.* **2012**, *45*, 114–123.

3. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid Catalyzed Propylene Oligomerization [2]

ABSTRACT

This work is concerned with the development of multivariate calibration models to establish spectrum-composition relationships for the hydrocarbon products in the H-ZSM-5 catalyzed oligomerization of propylene. Regression models based on two multivariate methods were investigated in this work: least squares-support vector machines (LS-SVM) and partial least squares (PLS) regression. The performance of two nonlinear kernels, radial basis function (RBF) and polynomial, is compared with PLSR as well as its variant, interval-PLS regression (*i*-PLSR). For comparing with *i*-PLSR, the Fourier Transform Infrared (FTIR) spectra of the products served as inputs and the respective C_1-C_{10} concentrations, obtained from gas chromatography (GC) were the outputs. The sensitivity of the product distribution to inlet operating conditions was also evaluated through the calibration methods. Spectral clusters having distinct chemical character were identified using principal component analysis (PCA) and hierarchical clustering analysis (HCA) and also used as inputs to the different regression techniques to compare with the full spectrum models. It was found that the best performing spectral regions from *i*-PLSR had chemical relevance and agreed with findings from HCA, improving the predictive capabilities significantly. The decreasing order of performance of the chemometric methods evaluated was: LS-SVM-RBF > LS-SVM-Polynomial > *i*-PLS > PLS. The prediction accuracy of RBF kernel-based LS-SVM regression technique was the highest, indicating its suitability for effective online monitoring of moderately complex processes like acid catalyzed propylene oligomerization.

Keywords: Support vector regression; interval-partial least squares; acid-catalyzed olefin oligomerization; infrared spectra; online monitoring of product composition

[2] This work was published as 'Sivaramakrishnan, K.; Nie, J.; De Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, 57, 13156 – 13176'.

3.1 Introduction

Gaseous light olefins are generally obtained as one of the by-products from the high temperature pyrolysis of organic material. These light olefins can be used to produce heavier hydrocarbons by oligomerization or polymerization processes. Over the years, olefin oligomerization has been studied in great detail and the use of shape-selective catalysts like the synthetic zeolite, ZSM-5 has been developed. [1,2] Though a plethora of reaction mechanisms have been proposed in the literature, the system chemistry is predominantly dependent on temperature, pressure and residence times in a continuous process. [3–6]

At mild operating conditions, oligomerization of the feed olefin to distinct oligomers occurs, followed by isomerization and re-cracking to form a range of light olefins and finally, the olefins re-polymerize to a distribution of heavier olefins. [2–4] Skeletal and double bond isomerization add to the product uncertainty and complexity. At more severe operating conditions, increased cracking rates, disproportionation through carbenium ion intermediates (due to the protonic nature of the catalyst), hydrogen transfer, cyclization and aromatization reactions become more pronounced and makes the reaction pathway complicated. Overall, the product molecular weight distribution depends on which pathway is dominant: cracking or polymerization.

These types of complex reacting mixtures present significant challenges for online monitoring and control. These challenges arise because it is often not possible to have complete characterization of the products because of the existence of a large number of species, which pose difficulties in developing algorithms for tracking their evolution with time. The pragmatic solution is to employ lumped mixture properties that are readily measurable to describe reaction networks and its kinetics. But the main drawback of this approach is that there is not always a clear link between the measured property and the reaction chemistry.

Information on reaction chemistry is provided by spectroscopic methods, but the spectra of complex mixtures should be interpreted with caution. Vibrational spectroscopy techniques like Fourier transform infrared spectroscopy (FTIR) and near-infrared spectroscopy (NIR) have been used in many applications to determine chemical characteristics of reaction products. [7–10] Ultraviolet-visible (UV-Vis) and nuclear magnetic resonance (NMR) spectroscopy provide insight into the electronic and atomic/nuclear environment of the constituent molecules in the analyte, respectively. [11,12] The lower cost and easier operational procedures make vibrational spectroscopy-

based techniques the preferred choice for online application. Separation-based approaches like gas chromatography (GC) with mass spectroscopy (MS) detectors and thermo-gravimetry analyses (TGA) provide quantitative information on product composition and yield. [13,14] Although the spectroscopic methods can be used online, the GC-MS and TG methods work offline and are time consuming.

The purpose of this work, therefore, is to devise a methodology to track the change in composition of the product stream from propylene conversion over H-ZSM-5 conducted at different operating conditions. Chemometric tools provide a viable and faster means for achieving this objective by avoiding the use of offline analytical instruments like GC and deriving product composition directly from online spectroscopic data.

Multivariate calibration methods applied to analytical chemistry have evolved tremendously and been the focus of attention in recent years. [15] The inherent multivariate nature of the spectral data combined with overlapping of the spectral bands of multiple components creates challenges for interpretation. This calls for extracting only the relevant and non-redundant information from the spectral data in order to improve the calibration model relating the outputs of spectroscopy and the desired property of the sample. Partial least squares regression (PLSR) does exactly this and is one of the most widely used techniques in the chemometrics literature. [16,17] PLS is a development over multiple linear regression (MLR) in that it can deal with strongly correlated explanatory variables that can predict the response. In other words, it extracts latent variables in the input space that explain maximum variance in the output. This is done by projecting both the predictors and the outcome variables to a new space. The major limitation of PLSR is its assumption that the spectrum-property relationship is linear, which may not be true in majority of the chemical reactions. Some modifications of the linear PLS exist, namely Poly-PLS and Spline-PLS but they have not proved to describe the nonlinear reacting systems adequately. [18,19] A more recent variant of the PLS method is the interval PLS (*i*-PLS), introduced by Norgaard et al. [20] and applied on NIR data, that aims at selecting only those variables from the spectrum that significantly affect the response (desired property) by building PLS regression models on user-defined sub-intervals in the input space. The predictive power increased compared to PLSR, but the linearity assumption still existed. The *i*-PLS approach will also be employed in our study.

Artificial Neural Networks (ANN) and Support Vector Machines (SVM) are some of the alternatives used for regressing nonlinear data. [21–23] SVMs address some of the disadvantages

posed by ANN methods like the stochastic nature, repeatability of the training process and to an extent, non-uniqueness of the final solution. [24] SVMs are a supervised learning technique, first developed for statistical classification by Vapnik [25] and later extended to solve regression problems. [26,27] The only disadvantage is computational complexity because of the requirement to solve a quadratic optimization problem. To mitigate this issue, the least squares support vector machine (LS-SVM) was proposed by Suykens [28] where a linear optimization is solved. The performance of LS-SVM based techniques was found to be superior as compared to other regression methods and it was successfully implemented with spectroscopic data in determining the required sample characteristic in an array of applications. [29–32] Both SVMs and LS-SVMs use kernel functions to build nonlinear models by mapping to a higher dimensional space. [28] Gaussian process regression (GPR) has also been employed recently in describing nonlinear systems and it has been compared with the performance of LS-SVMs. [33] Cui and Fearn [29] highlighted some of the differences between LS-SVM and GPR in their work. GPR looks at maximizing posterior probability, which might not work well with a smaller sized training set, while LS-SVMs minimize the classical squared error loss function that depends only on the posterior mean and gives importance to accurate forecasts.

It has to be noted that multivariate data analysis has not been implemented extensively in olefin oligomerization processes to monitor product composition. Keeping in mind our long-term interest to develop online monitoring tools for residue and biomass conversion processes, [34,35] we decided to investigate LS-SVM regression as a means of multivariate data analysis on a simple, yet chemically complex propylene oligomerization system that is reasonably well understood. In this work, the performance of the LS-SVM approach utilizing 2 types of kernels, namely, the Gaussian RBF and the polynomial kernel, [36] is compared with PLSR and *i*-PLSR calibration models in terms of prediction accuracy and learning efficiency with varying size of the training set for predicting the concentrations of the products from propylene conversion. Though other types of kernels like linear and sigmoid have been explored in literature, the RBF and polynomial kernels are known to capture nonlinearity in data better (as required in our case). [31] Section 3.4.4 provides a detailed explanation along with the corresponding references, on the reasoning behind the use of these two kernels for this work.

Three kinds of inputs are fed to the regression models: operating conditions (inlet temperature and flowrate), the full spectrum of IR absorbance intensities and lastly, clustered spectral

intensities. The clustered intensities (corresponding to the respective wavelengths) were obtained through the application of HCA on PCA loadings, to the IR spectral data so as to obtain a reduced spectral dimension that can describe the system chemistry better than the whole spectrum. [37] Multivariate data analysis (MDA) techniques have been shown to perform better and faster when combined with spectra in the reduced dimension. [32,38] The clustered intensities, though representing the same data have lesser number of variables with each variable corresponding to specific chemical groups. Hence, they have also been used as inputs to compare with the performances of the full spectrum inputs, in terms of prediction accuracy and training time. They also allow for better chemical interpretation since the clustered intensities represent important groups of specific wavenumbers rather than the full spectrum IR where peaks tend to overlap. Target product concentrations were also clustered and used as regression outputs combined with operating conditions and IR spectra as inputs and compared with the performance of their full concentration counterparts. In summary, we present a comprehensive review of the LS-SVM and PLS-based regression methods to identify the best approach to monitor similar and more complex processes online. [39]

3.2 Experimental

3.2.1 Materials

The oligomerization experiments were conducted using pure propylene (≥ 99.5 %) obtained as a cylinder gas from Praxair Canada Inc. Nitrogen (99.99 %), also supplied by Praxair Canada Inc. was employed as purge gas. The ZSM-5 catalyst (MFI zeolite) was acquired from Zeolyst International Inc. Table A.1 in Appendix A provides the important properties of the catalyst as reported by the supplier.

The catalyst powder was supplied in its ammonium form and the catalyst was activated by calcining at 550 ºC for 5 hours along with dry air purging at 50 mL/min to obtain the H-ZSM-5 used in the experimental work. Silicon carbide (≥ 99.0% SiC and 400-mesh), procured from Karmer Industries Inc., was employed as inert diluent material in the reactor.

3.2.2 Equipment and Procedure

A flow reactor, 34 cm long and 1.9 cm in diameter, was employed for the oligomerization experiments. The H-ZSM-5 catalyst (0.15 g) was blended with 400-mesh silicon carbide (SiC) to dilute the catalyst in a 1:20 ratio by mass. The relatively high dilution ratio for the catalyst to SiC risked bypassing in the reactor, but this served as a trade-off made to improve heat management. Other sizes of SiC were also used to pack the reactor to obtain a smooth, continuous tubular axial flow and restricting radial flow. The length of reactor used was 34 cm and made of ¾" stainless steel tubing (316 grade), purchased from Swagelok. The whole experimental setup is shown in Figure A.1 in Appendix A. The feed to the reactor was supplied directly from the propylene cylinder and the flow rate of propylene was controlled using a Brooks Automation, Inc. gas flow controller. The feed was preheated in the top part of the reactor before reaching the catalyst bed. The temperature inside the catalyst bed was controlled by adjusting the duty of the tubing furnace.

The system pressure was controlled by a back-pressure regulator in the gas line and all experiments were conducted at near-atmospheric pressure. The reaction products were cooled down indirectly using continuous chilled water flow. After cooling and condensation, the products were separated into gas and liquid. The product gas flow rate was measured through an ADM1000 flowmeter from Agilent Technologies and the temperature of the gas flow after separation at the outlet was also measured for calculation of the mass flow of the gaseous products. The liquid products were collected in liquid containers (Fisher Scientific) and the gas products in gas bags (Saint Gobain Chemware FEP). Online flow cells were also installed on the path of liquid product flow and gas product flow for online spectroscopy as shown in Figure 3.1. In each flow cell, one end was connected to a light source and the other end was connected to a sensor for transmission spectra detection, with a potassium bromide (KBr) window of size 13 × 2 mm, being used for all the light paths.

The liquid product, after condensation, first was allowed to flow between two windows through a flow cell with 5 cm optic path length for UV-Vis spectroscopy, followed by a flow cell with 0.2 cm path length for FTIR spectroscopy. The gaseous products flowed through a 50 cm long tubing cell for IR and NIR spectroscopy.

The axial temperature profile was measured during reaction and it was found to vary by 3 ºC or less from top to bottom for all test conditions, i.e. acceptable for near-isothermal operation. The experimental conditions studied in this work covered the temperature range of 346–477 ºC and

normal propylene volumetric flow rate range of 172–440 ml/min (refer section 3.2.4). The pressure used was near-atmospheric, i.e. around 100 kPa absolute.

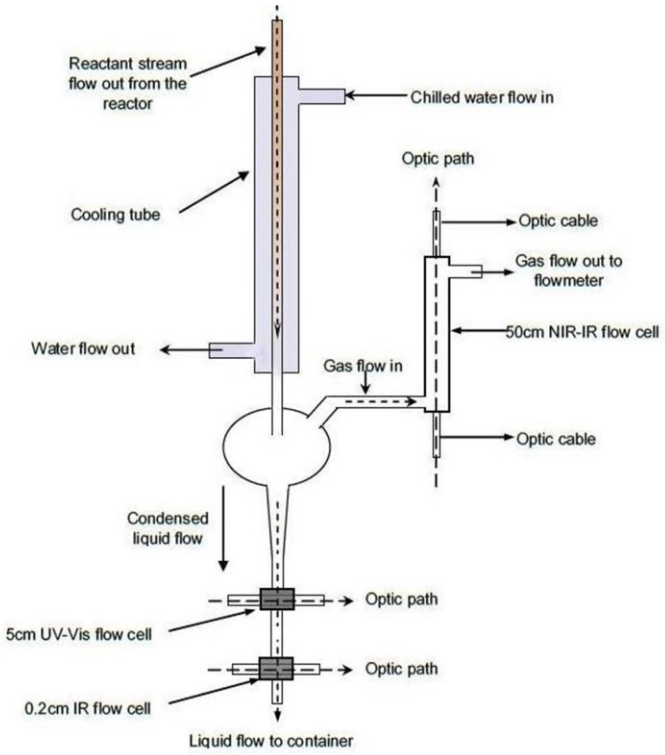


Figure 3.1. Schematic of the condensation of the products and the online flow cells.

3.2.3 Analyses

An ABB MB-3000 infrared spectrometer with liquid flow cell was employed for the IR analysis of the liquid samples collected in the experiments. Deuterated triglycine Sulfate (DTGS) detector, with a gain of 81 dB was employed by the instrument. The spectral parameters employed were: Resolution: 8 cm^{-1}; Wavenumber range: 3400 – 700 cm^{-1}; Number of scans: 120; Mode: Transmittance. Air was used as the reference material for each measured spectrum. Since the UV-Visible and NIR spectroscopic data were not used in this study, the details of the respective spectrometers are not reported.

Off-line product characterization was done by gas chromatography (GC). The gaseous product was analyzed using an Agilent 7890A Gas Chromatograph (GC-FID-TCD) equipped with Flame Ionization Detector (FID) for detecting hydrocarbons, combined with a Thermal Conductivity Detector (TCD) for detecting gases like CO_2, H_2, N_2, etc. The liquid products were

also analyzed using an Agilent chromatograph flame ionization detector (GC-FID). A mass selective detector (GC-MS) was also utilized to assist in assignment of chromatographic peaks to different carbon numbers. The details of the gas chromatographic analyses are given in Table A.2 in Appendix A. In order to calculate the mole ratio of a certain carbon number in the product, the possibility of the carbon number species existing both in the gas and liquid sample is taken into consideration.

3.2.4 Data available

Twenty-five sets of experiments were conducted in the temperature range of 346-477 ºC at five different volumetric flowrates between 172 and 440 mL/min for propylene conversion over H-ZSM-5 catalyst. Since the catalyst mass (0.15 g) used was constant for all reactions, the weight hourly space velocities (WHSV) are directly proportional to the volumetric flowrates and vary from ~120/h (172 mL/min) to ~300/h (440 mL/min). The reaction conditions along with the corresponding reactant conversions and the residence times are reported in Table A.3 in Appendix A. The statistics for the concentrations (in mole percent) of the hydrocarbon products at the reactor outlet calculated from GC analyses are given in Table 3.1. The full set of hydrocarbon product (C_1 to C_{10}) concentrations is available in Table A.4 of Appendix A for the reader's reference.

FTIR spectra of the products were obtained in the wavenumber range of 3400 - 700 cm^{-1} (2942 – 14286 nm) as these were adequate to identify the type of functional groups present in the products. A detailed analysis of the identified groups corresponding to the spectral peaks (after pre-processing) is given in section 3.4.3.2. The feed volumetric flow rates were converted to molar flow rates assuming ideal gas properties and in relation to standard conditions. The residence times in the catalyst bed were calculated from dividing the volume of the catalyst (constant for all reactions) by the volumetric flowrate of the reactant. Since this occurs at the reaction temperature, a correction term involving the ratio of room temperature to reaction temperature (in K) was multiplied to the previously obtained value. Pressure was maintained atmospheric always.

Table 3.1. Concentration data (in % mol) for the products of propylene conversion obtained from GC-FID/MS.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
Minimum	0.002	0.32	8.26	14.39	14.67	20.75	6.02	5.18	0.69	0.00
Maximum	0.025	1.38	28.67	21.74	26.73	28.15	10.18	9.86	3.47	0.51
Average [a]	0.010	0.78	17.16	19.47	22.50	23.19	8.30	6.81	1.56	0.14

[a] Averaged over 25 samples.

3.3 Multivariate data analysis theory

Pertaining to the discussion in section 3.1, the theory behind the different MDA techniques that are utilized in this work is given in this section. The following are discussed: principal component analysis (PCA), hierarchical clustering analysis (HCA), least squares-support vector regression (LS-SVR) and interval-Partial least squares regression (*i*-PLSR) as a specialized case of Partial least squares regression (PLSR).

3.3.1 Principal component analysis (PCA)

PCA has been used widely as an exploratory method for multivariate data analysis in diverse areas of science and engineering for detecting operational changes, as pre-filters for spectral library matching, pattern recognition for wavelength extraction in vibrational spectroscopy and also in meteorology. [40–43] An exhaustive review of the PCA approach is given by Jolliffe [44] and only the important points are mentioned in this section. Conceptually, PCA is a feature extraction method rather than a feature selection method. This is because it creates a set of linearly independent principal components from the existing variables (which may be correlated) by means of singular value decomposition of the data matrix. It does not select subsets of variables from the original data, but projects the original data into the principal component space. PCA is also commonly employed as data pre-processing for subsequent analysis such as clustering analysis, which is also done in this work. PCA reduces the dimensionality of a data set consisting of a large number of interrelated variables through an orthogonal transformation to produce a set of uncorrelated

variables and subsequent truncation of that set. The transformed variables that are retained (the principal components) are linear combinations of the original variables and represent most of the variation present in them. The principal components are ordered such that the first component accounts for the largest fraction of the variance, followed by the second principal component and so on.

For choosing the number of components, a scree plot was considered where the variance explained by each principal component was plotted against the component number. The plot for each PCA might be different in terms of where the curve flattens out. The 'elbow' point after which further components did not contribute in increasing the variance explained significantly compared to previous components was chosen as the number of components in all cases investigated in this work. This was also checked with the plot of cumulative variance explained (sum of the eigen values of the covariance matrix) against the number of components. Details of how the loadings and scores are extracted from the data matrix are provided in section A.5 of Appendix A.

In the current study, the chemical nature and composition of the reaction products depend on many factors such as temperature, propylene flow rate and conversion, residence times, etc. PCA was applied to examine which of these factors affected the products significantly and also to visualize product streams of similar nature, subsequently confirmed by HCA. The FTIR spectral data was also subjected to PCA as a pre-filter to forming clusters through HCA. PCA was performed using the Statistics and Machine Learning Toolbox in MATLAB R2017b (9.3.0) (refer section 3.4.3.1).

3.3.2 Hierarchical clustering analysis (HCA)

Clustering is an unsupervised grouping technique that has found application as a pattern recognition method in vibrational spectroscopy, a machine learning tool in medical science and other areas. [45–48] Clustering, unlike statistical classification where new objects are allotted into categories based on a training set of already categorized instances, assembles objects within a set based on a similarity measure. This results in objects in the same group or cluster being more similar to each other than to those in other clusters. The algorithm employed in this study for clustering was connectivity-based clustering, also called hierarchical clustering. [37] HCA was preferred over *k*-means clustering as it was difficult to determine the number of clusters to be

formed beforehand for the spectral dataset. [49] It uses a distance metric (Euclidean in our case) in the high-dimensional space of variables and provides a hierarchy of clusters represented in a dendrogram. Clusters that are distinct at a lower level in the dendrogram merge at higher levels as the threshold of distance used for distinction between clusters is changed.

Hierarchical clustering analysis (HCA) can be carried out using an agglomerative or a divisive algorithm. Agglomerative (bottom-up) HCA takes each object as a separate cluster in the first step, and merges clusters step by step to form a higher-level hierarchy. Once the distance between existing observations are calculated, clusters are formed by linking pairs of objects closer to each other. This process continues till all clusters are merged together, thus forming the cluster tree, also called the dendrogram. '*Linkage*' function in MATLAB is used for the purpose of linking clusters together based on '*average*' linkage criterion in this work. The other common distance criteria are '*single (minimum)*', '*complete* (*maximum)*', *'ward'* and '*weighted average linkage*'.

Finally, the individual clusters are created by either specifying the number of clusters to the 'cluster' function or deciding where to cut off visually from the dendrogram. Our current study utilizes HCA to find groups of similar objects in the spectral data and product streams, in order to be used as inputs and outputs to the regression models, respectively. HCA was performed employing the agglomerative approach with built-in functions in the Statistics and Machine Learning Toolbox in MATLAB R2017b (9.3.0).

3.3.3 Least squares – support vector machine (LS-SVM) regression

It is easier to derive LS-SVM from the basic knowledge of support vector machines (SVM). The use of SVMs for classification as a supervised learning technique was first introduced by Vapnik [25] (in 1995) as a linear method whereas the LS-SVM formulation was proposed by Suykens [28] in 2002. SVM maps a given sample set to a higher dimensional space and further separates the samples into categories with a clear divide. In simple terms, this process enables new test samples to be recognized and allocated to their respective categories.

Support Vector Regression (SVR) is an extension to SVM and utilizes the same principle as SVM. SVR bears a clear advantage over other regression methods like principal component regression (PCR) and partial least squares regression (PLSR) due to its ability to tackle non-linear problems, which can model real-world systems better. [50] While developed as a linear technique,

the use of the so-called kernel trick allows it to be applied to nonlinear problems. A kernel function is used to map the original data onto a higher-dimensional feature space, where the regression can be treated as a linear problem. This transformation is depicted in Figure A.2 in Appendix A.

For a typical classification problem on a given set of training data with observations X = $\{x_i \in R^n, i = 1, 2, .., n)\}$ (R^n is the n-dimensional vector space) and corresponding outcomes y = $\{y_i \in r, i = 1, 2, .., n)\}$ (r is a one-dimensional vector space), the discriminant function is given in equation 3.1 as follows:

$$y^*(x_i) = w^T \emptyset(x_i) + b \qquad \text{Equation 3.1}$$

where $\emptyset$ is the function used for the nonlinear transformation to the higher dimensional space, w is the parameterized weight vector that can also be viewed as a normal vector to the decision hyperplane that facilitates maximum margin between the separated classes, b the model bias or the offset and y^* is the predicted value. The objective function that needs to be optimized remains the same in both LS-SVM and SVM except that a classical squared error function (similar to MLR and PLSR) is used in LS-SVM as compared to Vapnik's ε-insensitive error in SVM. The ε-insensitive error determines whether a calibration sample is redundant or a support vector depending on the residual error. The final prediction of the SVM model is dependent on the support vectors.

The cost function for the optimization problem in LS-SVM becomes

$$\text{Minimize } G = \frac{1}{2} w^T \mathrm{w} + \frac{\gamma}{2} \sum_{\mathrm{i}=1}^{\mathrm{n}} e_i^2$$
$$\text{where } e_i = w^T \emptyset(x_i) + b - y_i \text{ is the estimation error}$$
$$\text{subject to: } y_i(w^T \emptyset(x_i) + b) = 1 - e_i;\ i = 1, 2, \ldots, n$$
$$\gamma \text{ is the regularization parameter} \qquad \text{Equation 3.2}$$

LS-SVM replaces inequality constraints in SVM with equality constraints that is reflected in the error term in equation 3.2. The first term penalizes the weight vector (w) and intends to maximize the separation margin, while the second term directs the error towards a minimum.

Since it is difficult to classify all training samples in the case of nonlinear data with a linear boundary, the regularization parameter (γ) is introduced to account for the balance between maximizing the hyperplane margin (characterized by w) and minimizing the error of

approximation for the training samples (e_i) and a higher value can have a direct effect on reducing the model offset, b. Since the nonlinear feature space has potentially infinite dimensions, the cost function is combined with the constraints and expressed in Lagrangian dual form as in equation 3.3:

$$L(w, b, e; \alpha) = G - \sum_{i=1}^{n} \alpha_i (w^T \emptyset(x_i) + b + e_i - y_i) \qquad \text{Equation 3.3}$$

where $0 \leq \alpha_i \leq \gamma$ are the Lagrange multipliers learned from the data and describe each x_i. A major advantage of the SVM framework is that there are no issues with local minima. [51] Further details regarding the solution of equation 3.3 is given in section A.6 in Appendix A.

The typical kernel functions reported in literature are linear, polynomial, sigmoid and Gaussian radial basis function (RBF). However, only polynomial and RBF kernels are utilized in this work, as given in equation 3.4 and equation 3.5, respectively. The notion behind applying these two kernels is given in section 3.4.4. A conjugate gradient (CG) algorithm is generally applied for the case of large data sets to iteratively solve equation 3.3 keeping in mind the requirement that K is positive definite. The reader is referred to Schölkopf and Smola [36] for further details regarding the solution to the above optimization.

$$K = \emptyset(x_i)^T \emptyset(x_j) = K(x_i, x_j) = \left(x_i^T x_j + t\right)^d \qquad \text{Equation 3.4}$$

$$K = \emptyset(x_i)^T \emptyset(x_j) = K(x_i, x_j) = \exp\left(-\frac{\left||x_i - x_j|\right|_2^2}{2\sigma^2}\right) \qquad \text{Equation 3.5}$$

Hence, the prediction for a new observation, x, becomes:

$$y^*(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \qquad \text{Equation 3.6}$$

Equation 3.5 suggests that the RBF kernel function is a measure of similarity between 2 training samples in the nonlinear space with similar samples (x_i, x_j) resulting in the RBF kernel moving towards 1 and dissimilar samples making the kernel move towards 0. The hyper-parameters t, d, γ and σ^2 are computed from the training dataset by tuning the model for each output dimension and their optimized values are reported along with the results for each input-output combination (Table 3.2 in section 3.4.2). The disadvantage with SVM is that an added parameter, ε needs to be estimated in addition to the kernel parameters in all cases. d represents

the degree of the polynomial, t denotes the constant term and σ refers to the width of the RBF kernel. σ (values reported as σ^2 in this work) is a measure of how sensitive the output (y) is to any change in the input (X) in the calibration data. A higher kernel width is indicative of a more linear relationship with a less complex solution, whereas a lower σ is a sign of high sensitivity and nonlinearity between X and y. The training time increases linearly with the number of dimensions in the dataset, which served as an additional motive to carry out dimension reduction. However, the computational times observed in this work were well within the scope of realizing a real-time monitoring protocol for the propylene oligomerization process. The LSSVMlab1.8 toolbox was used in our study to implement the LS-SVR in MATLAB R2017b (9.3.0). [52]

3.3.4 Full spectrum partial least squares regression (PLSR) and interval-partial least squares regression (*i*-PLSR)

The theory of PLSR is well known and comprehensively documented in the literature and also shown to be useful for NIR calibration. [15–17] Since this also forms the basis for *i*-PLSR, only the main idea of PLSR and subsequently *i*-PLSR is discussed in this section. In PLSR, the multivariate output variable or the property of interest (y) is predicted from the multivariate input spectra (for example) ($X = (x_{ik}, i = 1, 2, .., n; k = 1, 2, .., K)$) where K is the number of spectral channels and n is the number of observations, by means of a simple linear predictor similar to equation A.1:

$$y = b\mathrm{X} \qquad \text{Equation 3.7}$$

The central idea of PLS regression is that after implementing PCA on both y and X to obtain the latent variables, the individual loading axes are rotated to create a situation where maximum variance in the output is explained by the input. A way to do this is to maximize the covariance between the respective scores. Representing highly correlated spectral ensembles with redundant data in the form of orthogonal, independent variables that are linear combinations of their predecessors reduces model complexity, thereby helping in improving the predictive ability of the developed models. In the case where ≤ 3 latent variables are able to explain majority of the output variance, the data can be plotted in a 2-D or 3-D graph to visualize the input-output relationships.

Interval PLS regression is more useful for concentrating on sections of the original spectral channels that can influence the output or desired characteristic to a greater degree by removing possible interferences from other variables. The whole spectrum is divided into a pre-decided

number of intervals of equal width on which PLS regression is applied and the results of each model on the sub-intervals are compared based on the model performance parameters (described in section 3.4.5). The comparison would make sense only if the number of underlying components for the individual sub-models are optimized. Further optimization of the intervals by shifting and varying the interval width on either side was out of scope for this work but was implemented by Norgaard et al. [20] Out of all the calibration models inspected in this work, *i*-PLSR was applied only with spectral data, both as a means of wavelength selection to reveal carbon number specific reaction chemistry as well as an improvement in the predictive power as compared to the full spectrum PLSR. The built-in function, '*plsregress*' belonging to Statistics and Machine Learning toolbox in MATLAB R2017b (9.3.0) was used for PLSR. The iToolbox (version 1.1 – April 2005) developed by Norgaard and Leardi [53] was used for building the *i*-PLSR models in this work.

3.4 Modeling methods

3.4.1 Training and validation sets

The data shown in Table A.3 in Appendix A is split into 2 sets for model building and performance estimation as follows:

a. Calibration or training dataset: Experiments 6 – 20 were used for constructing all models in this work. This was not an entirely random selection because care was taken to ensure that the product compositions in the calibration set spanned a similar range as in the validation dataset, though the difference was minimal as the concentrations were more sensitive to temperature change than to flowrate variation and the experiments 6 – 20 comprised of almost the entire temperature range used in this work.
b. Validation dataset: Experiments 1 – 5 and 21 – 25 were used as independent test sets to investigate the prediction ability of the calibrated model in terms of accuracy and speed. Though two samples involving the highest and lowest flowrates (0.426 and 1.090 mol/h) and temperatures (346 ºC and 477 ºC) are included in this set, the other samples consist of temperatures within the range of the calibration set. It was seen that temperature had a more prominent effect on product yield and nature (as discussed in the Results and Discussion section), so ensuring that the temperature range in the validation set was within the range

covered by the calibration set was important. Since the number of samples in the training set and validation set is limited, the learning efficiency of the model is tested with a growing number of samples in the training set (increased from 15 – 20) as well as a decreasing number of training set samples (decreased to 10). The results of this procedure are reported in section 3.5.9.

3.4.2 Types of models: Inputs and outputs

To test the prediction capability and efficiency of the regression techniques utilized in this study, six different combinations of inputs and outputs are tested, as communicated in Table 3.2.

Table 3.2. Regressor and output combinations for regression models constructed in this work.

Combination No.	Input/Regressor	Output	Regression methods compared
1	Temperatures & Molar flowrates (Table A.3)	Full $C_1 - C_{10}$ concentrations (Table A.4)	PLS, LS-SVM-Poly [a] & LS-SVM-RBF [b]
2	Temperatures & Molar flowrates	Lumped concentrations (Table 3.3)	PLS, LS-SVM-Poly & LS-SVM-RBF
3	Full spectrum FTIR intensities	Full $C_1 - C_{10}$ concentrations	PLS, *i*-PLS, LS-SVM-Poly & LS-SVM-RBF
4	Full spectrum FTIR intensities	HCA Lumped concentrations	PLS, *i*-PLS, LS-SVM-Poly & LS-SVM-RBF
5	HCA (4 cluster) intensities (Table 3.4)	Full $C_1 - C_{10}$ concentrations	PLS, LS-SVM-Poly & LS-SVM-RBF
6	HCA (6 cluster) intensities (Table 3.5)	Full $C_1 - C_{10}$ concentrations	PLS, LS-SVM-Poly & LS-SVM-RBF

[a] Polynomial kernel-based LS-SVM abbreviated as LS-SVM-Poly; [b] RBF kernel-based LS-SVM abbreviated as LS-SVM-RBF.

3.4.3 Data pre-processing

3.4.3.1 Software tools and other operations for pre-processing

MATLAB R2017b, Mathwork Inc. (9.3.0) software that utilized the University of Alberta academic license and run on a MacOS (Version 10.13.3) was employed for all analytical techniques in this work. Due to the difference in the ways of implementing the preprocessing techniques for different dimension reduction and regression models, it was deemed essential to specify how they were carried out in this study.

The transmission spectra collected were converted to absorption spectra, followed by baseline correction and data smoothing prior to dimension reduction and clustering (PCA, HCA) or regression (LS-SVM & PLS). Baseline correction was performed using the '*msbackadj*' function in the Bioinformatics toolbox in MATLAB. It estimates the baseline points for multiple regions of the spectra every 200 wavenumber units. The regression method used for baseline fitting was '*pchip*', which uses a piecewise cubic polynomial. Smoothing of the baseline corrected intensities was done with '*mssgolay*' function in Bioinformatics toolbox in MATLAB. It is based on the well-known Savitzky – Golay filter that eliminates noise using a least-squares lower degree polynomial (degree 2 in this study). [54] Since the wavenumbers in our case were equally spaced, the window-size sliding frame for smoothing was based on the wavenumber index and 15 samples were used. It was found that smoothing removed certain troughs in the peaks though it did not have a significant effect on the data analysis results. Visual inspection did not reveal any significant noise present in the baseline that warranted the use of a filter. Nevertheless, smoothing was carried out.

Normalization was performed by standardizing the data, which is a combination of mean centering (variable mean is 0) and auto-scaling (making variable standard deviation as 1). This was done by using the '*zscore*' function belonging to Statistics and Machine Learning toolbox in MATLAB. Before conducting PCA, HCA and PLSR, all input features were standardized so as to bring the features on the same scale to account for variability in the data. On the other hand, standardization of the variables in LS-SVM was done during initialization of the object-oriented structure representing the LS-SVM model by supplying the function argument '*preprocess*'. Particular details of how the inputs and outputs were supplied to the calibration models are given in section A.7 of Appendix A.

3.4.3.2 Identification of spectral features

Initially, it was considered worthwhile to assign the spectral features to potential functional groups in the products. The absorption spectra after baseline correction and standardization are presented in Figure 3.2. It should be pointed out that the reaction conditions for propylene conversion were such that at all conditions hydrogen transfer and aromatization reactions were anticipated. [3] This was confirmed by the infrared spectra (Figure 3.2) where all instances contained aliphatic and aromatic absorption peaks. The main absorption bands were identified based on literature. [55] The most prominent features in the infrared spectra were due to the aliphatic sp^3 C-H (methyl and methylene groups) stretching vibrations at ~2960, ~2930 and ~2875 cm^{-1}, as well as the sp^3 C-H bending deformation at ~1465 and ~1380 cm^{-1} respectively. The first overtone of the aliphatic C-H deformation is seen at ~2735 cm^{-1}.

The products produced were mainly olefins with a methylene group ($R_2C{=}CH_2$) and trans-internal olefins. Methyl branching was proved to be much lesser in a ZSM-5 catalyzed oligomer product than from a non-shape selective oligomerization. [2,3] The main absorption bands of the olefins with methylene group can be seen as sp^2 =C-H_2 stretch at ~3008 cm^{-1} and ~2985 cm^{-1} (seen as unresolved shoulder at ~3000 cm^{-1} in Figure 3.2), alkene -C=C stretch at ~1650 cm^{-1} and -CH_2 wag at ~890 cm^{-1}. The presence of the trans di-substituted internal olefins can be seen as sp^2 =CH bend at ~990 cm^{-1} and trans -CH wag at ~965 cm^{-1}. There was also evidence of tri-substituted internal olefins as recognized from the existence of peaks at lower frequencies (~820 cm^{-1}).

The absorption peaks of the aromatic products were not as prominent as the absorption peaks of the aliphatic compounds. This was due to the relative absence of significant peaks in the 800-690 cm^{-1} that would be typical of substituted benzenes. However, many of the spectral features of the substituted benzenes were obscured by the substituted olefin bending deformations in the 900-820 cm^{-1} region and also close proximity of the aromatic -C=C- stretch to the aliphatic -C=C- stretch in the 1650-1500 cm^{-1} domain. This provided additional justification for the use of PCA and a wavelength selection method like *i*-PLS to assist with data analysis.

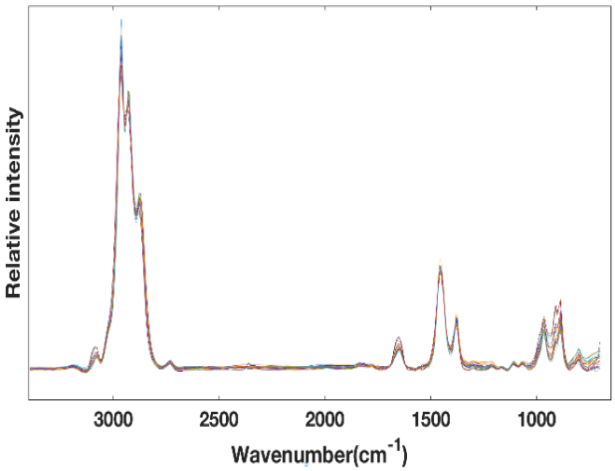


Figure 3.2. The IR spectra after pre-processing (without smoothing). The twenty-five spectra correspond to the products from each of 25 experimental conditions (Table A.3 in Appendix A).

3.4.4 Model tuning and optimization

A comparison between different calibration models for regression makes sense only when consummate results are achieved by optimizing the respective model tuning parameters corresponding to each output variable in the training dataset. In every case, the criteria for arriving at the optimum parameters is assessed through minimization of the root mean square error in cross-validation (RMSECV - refer section 3.4.5). Leave-one-out (LOO-CV) or *n*-fold cross validation (for *n* samples/observations in the calibration set) was used for training the models (except in *i*-PLS) as it has been shown to be robust against model mis-specification and preferable for smaller sized training set, as in our case (15 calibration samples). [28,56] The reason for this advantage is because LOO-CV focuses on point prediction only and no prior knowledge of the observation data is required. But this can lead to overfitting, which is also evaluated in this study by comparing with the root mean square error of prediction (RMSEP). However, *k*-fold CV was also evaluated with varying *k* between 2 and 14 for the training set and was found to yield very similar results as LOO-CV. The parameters to be tuned for each regression method are as follows:

(i) **PLSR** – The number of latent variables (LV) or factors extracted from the training set is the only key parameter in the PLS calibration model. A balance needs to be achieved between the number of components that describe the variance of the regressors and the prediction ability of the PLS model. Using higher than the required number of factors

complicates the model and can reduce its generalizability, also leading to increased computational time. The prediction error is drastically increased when using lower than the optimum number of components in PLSR. *'plsregress'* function gave the mean squared error for each added latent variable for all the 6 combinations investigated in this work (Table 3.2). The variance explained by the number of components that showed least mean squared error was also verified to fall in the 'elbow' region of the scree plot (as explained in section 3.3.1). Some chemometricians also prefer to use the predicted residual errors sum of squares (PRESS) statistic to assess the model performance, but RMSE minimization of the LOO-CV was chosen for this work. [33,57]

(ii) ***i*-PLS** – The 350 wavenumber channels are divided into equidistant subintervals and the number of latent variables in the PLS model for each interval was optimized in *i*-PLSR. This was carried out by '*ipls'* function in the i-Toolbox. [53] The performance of each spectral interval, in terms of the minimized RMSECV, is compared with the global RMSECV to decide the best performing interval that can be used for independent prediction. The optimum number of LV for the individual intervals is different (mostly lower) than that for the global model and they are obtained through a five segmented cross validation (Venetian blinds) method. [53,58] Specific details of this method are given in section A.8 of Appendix A.

(iii) **LS-SVM** – Irrespective of the type of kernel used, the regularization parameter (γ) needs to be optimised during model training. The kernel parameters necessary to be tuned, along with γ, pertaining to this work are: (a) the square of the kernel width (σ^2) for the RBF kernel (equation 3.5); (b) the intercept (t) and degree (d) for the polynomial kernel (equation 3.4). As in previous methods, the tuning process was based on minimization of a target squared error cost function with subsequent RMSE determination for the LOO-CV process. This optimized combination of the hyper-parameters (γ & σ^2 (or) γ, t & d) was obtained by SIMPLEX method, for which the starting values were provided by coupled simulated annealing (CSA) [59] rather than a trial and error approach as adopted in certain studies. [38] The lower and upper bounds over which the search is conducted in CSA are e^{-10} and e^{10}, respectively. [52] It is superior over a gradient descent optimization, since it avoids local minima and reduces sensitivity of the algorithm to initialization parameters, while at the same time carrying the process towards unique and ostensibly optimal values. This

behavior is seen in our work as the starting and the final optimized values of the hyper-parameters are of the same order of magnitude. Refer SI for a brief description of the SIMPLEX method adopted in this work.

Both in this study as well as in previous works in chemometrics involving LS-SVMs, [24,38] RBF kernels were found to give lowest prediction errors for majority of the outputs, among the different kernels used in the literature. They were reported to be suitable in modeling spectral-concentration relationships quite well. Having said that, there has been evidence in some cases that the performance of polynomial kernels was equivalent to that of RBF, which motivated us to inspect them as well. [26,60] In addition, since spectrum-property relationships are mostly nonlinear in nature, [31] RBF and polynomial kernels suited our data better than a linear kernel, for example. An exhaustive comparison with other types of kernels (linear, sigmoid, etc.) was not pursued. The RBF kernel lies within the limits [0,1] while the polynomial kernel varies between [0, ∞), which makes computation involving the polynomial kernel more complex.

3.4.5 Model performance evaluation

Once the parameters for each regression model were estimated based on the above-mentioned methods (section 3.4.4), the performance of the developed calibration model for the prediction of the validation dataset samples was evaluated on the basis of the following statistical parameters as shown in Table A.5 in Appendix A. In addition to the RMSECV, two other types of RMSE are compared in this work: RMSEC (error of calibration for the training set) and RMSEP (error of prediction for the validation dataset). Ideally, for a perfectly optimal model, the three values should be equal. However, in reality, the RMSEP is mostly greater than RMSEC and the criteria for determining extent of overfitting is their difference as well as the gap between RMSECV and RMSEP. If RMSEP is vastly greater than RMSECV or RMSEC, it means that the model is not able to predict accurately for new, real-time and independent datasets and is suggestive of overfitting and too much flexibility.

In this work, the accuracy of multivariate regression techniques is investigated with predictions for the individual carbon number concentrations with the validation set (section 3.5.3.2), in addition to the product composition distribution as a whole for each condition in the

validation set (section 3.5.3.3). RMSE is particularly useful in these cases as it places a significant weight on larger errors due to squaring of the absolute error as reflected in the formula (Table A.5 in Appendix A). A fair comparison of model performance can also be made by bringing the errors on the same scale. RAE, unlike RMSE, not only considers the averaged difference between the predicted and the target values, but also calculates a normalized absolute error. Obviously, since RMSE and RAE are negatively oriented scores, a lower value is suggestive of a more accurate model.

Meanwhile, the coefficient of determination (R^2) is purely a measure of the correlation between the observed and the predicted values, i.e. the higher, the better. From a statistical perspective, it indicates the fraction of output variance explained by the calibrated model inputs. According to this definition (Table A.5 in Appendix A), R^2 can take negative values, which simply means that the fit is worse than a horizontal line. The regression equation is always indicated in the correlation plots as it can be extrapolated to give the likelihood of future outcomes falling within the predicted outcomes as well. R^2 values are also reported for the calibration data and is referred to as R_C^2.

3.5 Results and Discussion

3.5.1 Principal component analysis (PCA)

As a preliminary step, it was important to understand not only the relationship between the different variables in the process, namely, temperature, flow rates of propylene, residence times, conversion and product composition but also variance in the spectral dimension. Hence, PCA was applied to two sets of data: (1) to the 4 columns in Table A.3 in Appendix A along with the 10 product compositions (Table 3.1), all acting as the variables; and (2) to the FTIR spectra, considering the intensities at the 350 wavenumbers as the variables. In both situations, the 15 sets of operating conditions in the training set acted as the samples.

Let us consider case (1). The loading vectors possess two attributes, namely, the direction and the magnitude. The direction is determined by the sign of the loading, i.e., a positive or a negative loading implies that the contribution of that variable to that principal component is in the positive or negative direction, respectively. The magnitude indicates the amount of contribution of that

variable towards the variance of that principal component. The loadings along with the scores for the first two principal components are shown in a biplot in Figure 3.3. The first two principal components account for 93 % of the total variance with the first component alone contributing 84 %. According to the scree plot for this PCA, the contribution of more than two components was deemed insignificant as compared to the first two. For the first component, temperature and feed molar flow rate have high positive loadings (Figure 3.3), respectively along with the product concentrations, C_1, C_2, C_3 and C_6 having loadings in the range of 0.21-0.31. This indicates that increasing the temperature and feed flow rate would increase the concentration of those product streams.

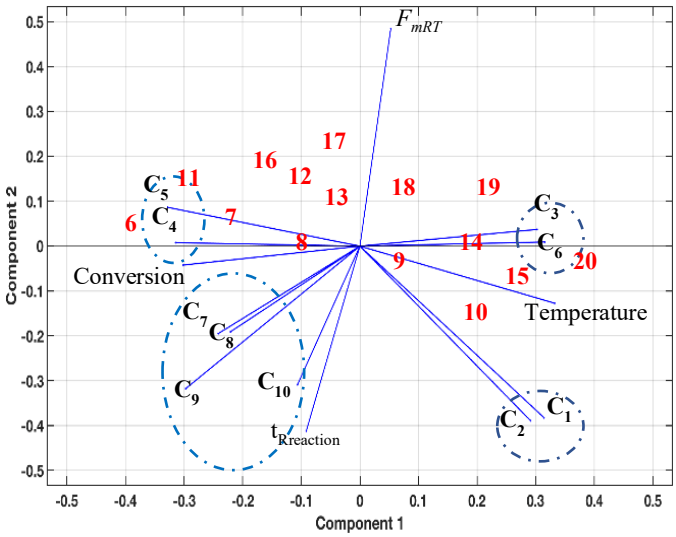


Figure 3.3. Biplot for PCA done on set (1) with 14 variables (follow blue lines) and 15 observations (red values).

An increase in flow rate decreases the residence time (opposite signs of loading for $t_{R_{reaction}}$ and $F_{m_{RT}}$) as well as the conversion of propylene (-0.30 loading). This might suggest that C_1 and C_2 are formed mainly with an increased temperature and a decreased residence time which lowers the conversion of propylene to C_1 and C_2. Figure 3.3 also shows that C_1 and C_2 are quite close to each other as are C_3 and C_6. These groupings give an indication of the reaction chemistry. C_1 and C_2 are mostly by-products formed by cracking of propylene and heavier molecules, which is explained by correlation with temperature. C_6 is mostly formed by dimerization of C_3 but further conversion of C_6 is limited by decreasing the residence time (present in vicinity of experiments 19

and 20). This also explains the higher C_6 concentrations in the product (Table A.4 in Appendix A). These observations are in-line with the mechanism proposed by Quann et al. [3]

The second principal component, which accounted for 9 % of the total variance explained, gives opposite loading directions for $F_{m_{RT}}$ (0.48) and $t_{R_{reaction}}$ (-0.43), which follows the interpretation from the first PC. Temperature is in the same direction as C_1 and C_2, which supports the fact that cracking is a major cause for their formation. An interesting observation is that an increase in the residence time is directly correlated with increase in the C_7 - C_{10} concentrations for both PCs as seen from Figure 3.3. Addition reactions (involving conjunct polymerization, i.e. cyclization and aromatization) are a possibility at higher residence times and can be a potential cause for formation of heavier products. [2] The product streams are, thus, grouped into 4 categories based on PC loadings as given in Table 3.3. However, this grouping needed to be confirmed with HCA as described in section 3.5.2.

Table 3.3. Product stream groups based on PCA.

Carbon numbers associated in a product stream group	**Description**
C_1, C_2	Light stream
C_3, C_6	Dimer product
C_4, C_5	Mid-stream
C_7, C_8, C_9, C_{10}	Heavy stream

The presence of any distinct outliers in the experiments conducted was discarded based on the scores, thus no sample was left out while performing the PCA. Furthermore, they are well separated on the molar flow rate direction, illustrated by samples in the second and third quadrant (Figure 3.3), where the experiments conducted at lower flow rates (6, 7, 8) are in the vicinity of heavier hydrocarbons (C_4 - C_{10}) as compared to the feed (C_3). The experimental points with higher flow rates are not as close to the lighter products (C_1, C_2), thus reiterating the claim that temperature rather than residence time is the major driving force for their formation. It has to be mentioned that if the loadings and scores are accommodated in the same plot, the scores are scaled down by a scalar factor as in Figure 3.3.

The loadings plot for the first two principal components in case (2) with the respective wavenumbers are given in Figure 3.4. A graph of both PCs plotted against each other is given in Figure A.3 in Appendix A. The scores plot for case (2) between the first two principal components is shown in Figure 3.5.

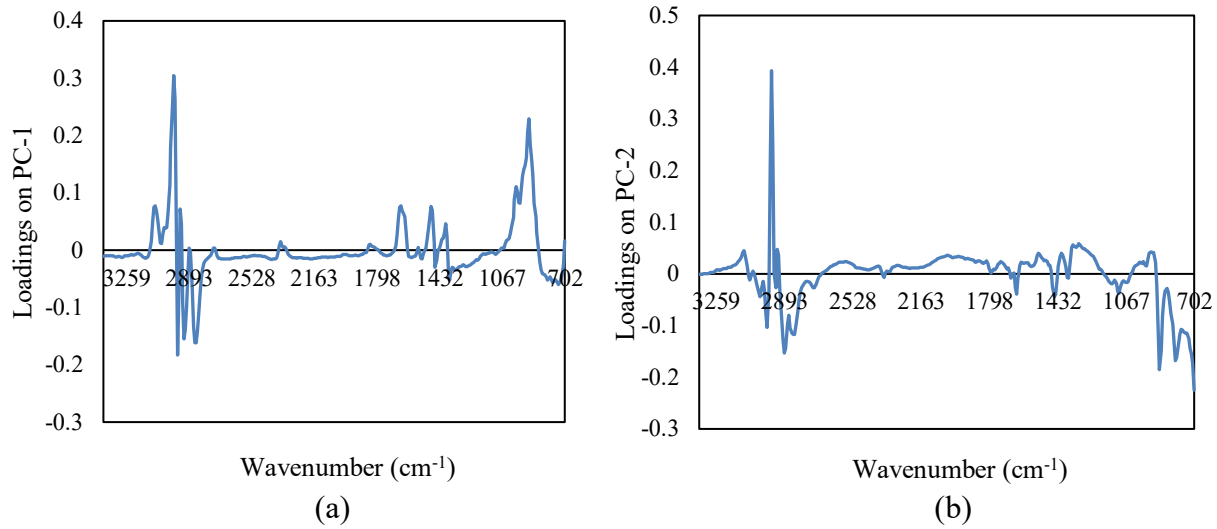


Figure 3.4. Loadings plot with 350 wavenumbers as variables. (a) First principal component loadings vs. wavenumbers; (b) Second principal component loadings vs. wavenumbers.

PCA on spectral data clearly showed the need for dimension reduction in the wavenumber direction. The first principal component accounted for 82 % of the variance, while the second and third component explain a further 7.5 % and 4.8 % of the variance, respectively. The wavenumbers corresponding to baseline intensities have very small loadings, i.e. those which are grouped closer to the origin in Figure A.3 in Appendix A. These can also be seen as near to the *x*-axis in Figure 3.4a and Figure 3.4b. On the other hand, prominent directions in the spectral data are determined by certain wavenumber indices that represent the appropriate functional groups quite well.

The first PC shows higher negative loadings for sp^3 C-H stretches that can be seen as a group from the index 280 in Figure A.3 with 2962 cm^{-1} (point 294) being the highest contributor in the second PC as well. These are visible as lines below the *x*-axis in the 2970-2850 cm^{-1} region in Figure 3.4. To the right side of the plot (Figure A.3 in Appendix A), indices 296-299 (3005-2977 cm^{-1} in Figure 3.4) have a high positive loading on the first PC and likely represent alkene sp^2 C-H stretches. These wavenumbers showed particularly high loadings of 0.13-0.23 on the third PC, thus providing a meaningful direction to the spectral interpretation and indicating the dominance

of olefinic products. The points 23-33 in Figure A.3 in Appendix A (950-870 cm^{-1} in Figure 3.4) most likely correspond to sp^2 C-H bend from substituted alkenes along with -CH_2 wag.

Aromatic C-H stretch can be attributed to the points 309-310 (~3080 cm^{-1}) that have significant positive loadings only for the first PC. There is also a group of wavenumber points in the third quadrant (Figure A.3 in Appendix A) that had equivalent contributions on both first and second PCs (indices 2-15, i.e. 810-710 cm^{-1} – right end of both the plots in Figure 3.4), indicating aromatic C-H bend in mono and di-substituted benzenes, which are a result of ring-closure phenomena. Overall, this plot echoes the discussion in section 3.4.3.2 where the focus was on assignment of the peaks to possible functional groups that could point toward the chemical nature of the products.

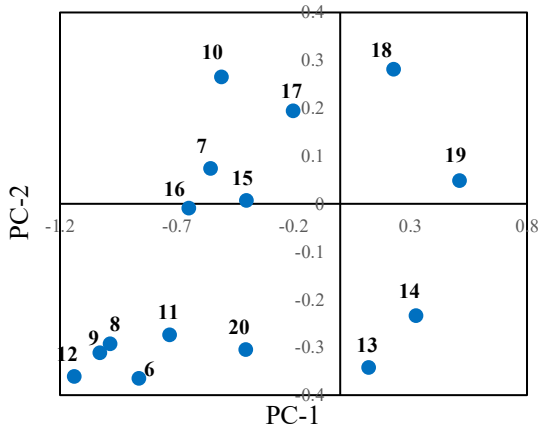


Figure 3.5. Scores plot for first two principal components for the case where PCA was applied to spectral data.

The scores (Figure 3.5) distribution does not warrant the removal of any observation as an outlier, like in case (1). However, HCA was required to be done on the qualitative information extracted by PCA to ascribe cluster numbers to the variables that were used as inputs and outputs for regression analysis in this work.

3.5.2 Hierarchical clustering analysis (HCA)

The next step was to perform HCA for both the cases described in the previous section in order to quantitatively inspect the hypothesized groupings that resulted from PCA. For case (1),

HCA was applied on the loadings of the first two principal components (that captured most of the variance in the variables as detailed in section 3.5.1) with the product composition from C_1 to C_{10} as the objects and clustered into 4 groups. The linkage method used for grouping clusters was '*average*' since it gave the highest cophenetic correlation coefficient of 0.94, i.e. a measure of how well the distances between the clusters containing the product composition objects correlate with the Euclidean distance between those objects. The default '*single*' linkage method gave the least cophenetic correlation coefficient of 0.90.

From the dendrogram of this clustering as shown in Figure 3.6, it is clear that the 4 groups of product streams hypothesized in Table 3.3 are possible when the cut off line is drawn at any distance value between 0.194 and 0.285. Interestingly, a 5-cluster grouping is also possible if the line is drawn between 0.137 and 0.194, the difference being the further splitting of the C_7-C_{10} group into C_{10} as one and C_7-C_9 as the other. However, to decrease the complexity of the problem, only the four-cluster combination is considered for regression in this work. As an aside, clustering purely based on raw data of product composition (after standardization) also yielded a similar cluster distribution.

In case (1), where only product compositions were considered, it was straightforward to visualize the cluster formations from PCA (Figure 3.3) due to the lesser number of variables that were taken into consideration. In contrast, it was difficult to assign the wavenumbers to particular groups just from the PCA loadings plot (Figure A.3 in Appendix A or Figure 3.4) for case (2) due to increased number of features (350 spectral channels) and cluttered nature of the data.

The clustering of the spectral data was important to this work because it was used as one of the inputs in the regression methods for predicting the product composition as will be discussed in section 3.5.7. HCA was applied to the loadings of the first three principal components obtained from PCA of the spectral data (case (2) in section 3.5.1) and divided into 4 and 6 clusters. The individual division of the 4 and 6 clusters and their chemical resemblance is provided in Table 3.4 and Table 3.5, respectively.

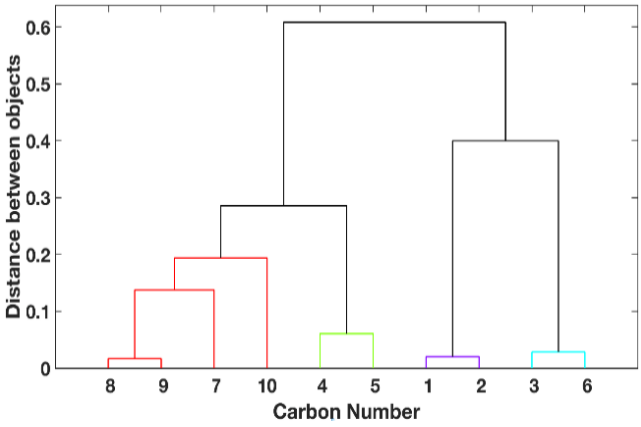


Figure 3.6. Hierarchical cluster tree for product composition based on PCA loadings.

Table 3.4. Summary of the 4-cluster division on FTIR spectra given by HCA.

Cluster number	Index	Wavenumber (cm^{-1})	Chemical description
1	288 – 295	2970 – 2916	Methyl sp^3 C-H stretch
2	277 – 287	2908 – 2831	(i) Methylene sp^3 C-H stretch (2908–2850 cm^{-1}) (ii) 1st overtone of sp^3 C-H bend (2840–2831 cm^{-1})
3	a) 1–23 b) 40–122 c) 127–276 d) 311–350	a) 871–702 b) 1635–1002 c) 2823–1674 d) 3394–3093	(i) Aromatic sp^2 C-H bend (minor) (840–702 cm^{-1} region) (iii) Baseline intensities
4	a) 24–39 b) 123–126 c) 296–310	a) 995–879 b) 1666–1643 c) 3085–2977	(i) Alkene sp^2 C-H bend (mono and di-substituted olefins) (995–879 cm^{-1}) (ii) Alkene C=C stretch (minor) (1666–1643 cm^{-1}) (iii) Alkene sp^2 C-H stretch (3008–2977 cm^{-1}) (iv) Aromatic sp^2 C-H stretch (3085–3008 cm^{-1})

Table 3.5. Summary of the 6-cluster division on FTIR spectra according to HCA.

Cluster number	Index	Wavenumber (cm^{-1})	Chemical description
1	288–295	2970–2916	Methyl sp^3 C-H stretch
2	277–287	2908–2831	(i) Methylene sp^3 C-H stretch (2908–2850 cm^{-1}) (ii) 1st overtone of sp^3 C-H bend (2840–2831 cm^{-1})
3	a) 1–23 b) 40–88 c) 104–121 d) 127–276 e) 311–350	a) 871–702 b) 1373–1002 c) 1627–1486 d) 2823–1674 e) 3394–3093	(i) Aromatic sp^2 C-H bend (minor) (840–702 cm^{-1} region) (ii) Aromatic C=C stretch (1627–1486 cm^{-1} region) (ii) Baseline intensities (b, d, e)
4	a) 24–39 b) 300–310	a) 995–879 b) 3085–3008	(i) Alkene sp^2 C-H bend (mono and di-substituted olefins) (995–879 cm^{-1}) (ii) Aromatic sp^2 C-H stretch (3085–3008 cm^{-1})
5	296–299	3001–2977	Alkene sp^2 C-H stretch (3008–2977 cm^{-1})
6	a) 122–126 b) 89–103	a) 1666–1635 b) 1488–1380	(i) Alkene C=C stretch (minor) (1666–1635 cm^{-1}) (ii) sp^3 C-H bend (1465–1380 cm^{-1})

'*Average*' method, with the highest cophenetic correlation coefficient of 0.96 was used for the linking procedure based on Euclidean distance. Since there were 350 wavenumbers present, the hierarchical cluster tree output consisted of only 30 base leaf nodes by collapsing the lower level nodes for easier visualization. The cut off margin in the dendrogram was quite low for the 6-cluster formation as compared to 4 clusters. It is worthwhile to mention here that there are a small number of wavenumbers for which no meaningful bonds/functional groups have been identified based on the knowledge about the products. These points are classified as baseline intensities (3rd cluster). Furthermore, the assignments were not unambiguous, but were made by identifying the most likely origin of the absorption band, since some wavenumber regions were crowded and there could be more than one plausible assignment.

A similar distribution of wavenumbers was obtained when HCA was performed directly on the spectral data before PCA with a small change that the aromatic C-H stretch wavenumbers

shifted to the third cluster merging with the baseline intensities for both 4 (Table 3.4) and 6 cluster (Table 3.5) distributions. The clusters obtained provide a complete picture of the functional groups present in the products and also seem to connect well with the PCA model. In the case of 4 clusters, the HCA distinguished between aromatic C-H deformations (cluster 3) and aliphatic alkene sp^2 C-H deformations (cluster 4) by grouping them into separate clusters. However, the bend for tri-substituted olefins overlapped with the aromatic C-H bends in cluster 3. Although sp^3 stretches were clearly identified in the first two clusters, the entire sp^2 C-H stretching for both alkenes and aromatics were grouped into the same cluster (4^{th}). The stretching vibrations of the carbon-carbon double bond in the formed alkenes were clearly identified (cluster 4) but the aromatic C=C stretch merged along with the baseline intensities in cluster 3.

The 6-cluster arrangement addressed some of the shortcomings of the 4-cluster distribution with more accurate classification of the functional groups. The main improvements were: (i) sp^3 C-H bend was separated from the baseline intensities and grouped into the 6^{th} cluster; (ii) the sp^2 C-H stretches for the aromatics and alkenes were split and grouped into 4^{th} and 5^{th} cluster, respectively; and (iii) the stretch for the C=C bond in aromatics were clearly singled out in the 3^{rd} cluster. Nevertheless, the first cluster in both distributions and the 6^{th} cluster in the 6-cluster division represented the highest intensity region in the FTIR spectra (Figure 3.2).

Although each cluster cannot be directly associated with the carbon number groups in the products as given in Table 3.3, the functional groups can be distributed in the products according to carbon number. For example, the third and fourth cluster representing the aromatics (Table 3.5) could correspond to C_7-C_{10}, the second, fifth and sixth cluster could represent the midstream hydrocarbons (C_3-C_6) as the methylene groups can be present as those inserted in between two C=C bonds in alkenes and the clusters represent aliphatic and alkene functional groups. The first cluster does not conform to any particular group as it is most likely possible that all the products can potentially contain a methyl group. But the first can mostly be associated with C_1 and C_2 due to the presence of methyl groups. Among C_1 and C_2, ethylene or ethane is the dominant product as β-scission in the carbocation-ZSM intermediate is much easier than α-scission in a mono-molecular cracking mechanism of heavier hydrocarbons. [61] The formation of ethane can occur at the temperatures used in this work through hydrogen transfer to ethane from a cyclic olefin. Another important feature of the six-cluster division is the separation of sp^3 C-H bending vibrations at 1465 – 1380 cm^{-1} (6^{th} cluster) as it can be characteristic of the entire product

distribution (as also shown by *i*-PLSR in section 3.5.5). The main value additions of clustering to this work are: (1) identification of key chemical features in the products that was otherwise difficult to comprehend from direct visualization in Figure 3.2 or from PCA in Figure 3.4; and (2) utilizing the clustered divisions as regressor inputs to the combinations 5 and 6 (Table 3.2) for assessing the performance of the calibration models.

3.5.3 Combination 1: Inlet operating conditions as predictors and entire product composition as outcomes

3.5.3.1 Factor selection for PLSR

The only parameter to be optimized for the prediction of concentration of the carbon numbers in the product is the number of latent variables. Since the rank of the input matrix was 2, a maximum of 2 LVs could be extracted that represented the data well. As expected, the first LV lowered the RMSECV values significantly but the difference in contribution of the second LV decreased with increase in carbon number. However, the amount of variance explained by the LV should also be considered jointly. For C_7 and C_{10}, the variance explained by the 2 LVs together amounted to only 78 % and 40 %, respectively. It is much higher for lower carbon numbers: 97.5 % for C_2, 99.1 % for C_3, 95 % for C_5 but decreases to 87 % for C_6. Increasing system complexity due to interference from a mixture of chemical species like aromatics, mono-olefin and higher substituted isomers is a plausible reason for the drop in the variance explained. [4] One LV was chosen for predicting C_7 (RMSECV for 1st LV: 0.5796; for 2nd LV: 0.5742), C_8 (RMSECV for 1st LV: 0.3124; for 2nd LV: 0.3104) because of close proximity of the errors of cross-validation but C_{10} (RMSECV for 1st LV: 0.1901; 2nd LV: 0.1932) actually showed a minimum at the 1st LV. [29] Two LVs were chosen for the remaining 7 outputs in PLSR. All the minimum RMSECV values are reported in Table 3.6 together with other model performance parameters. For this model, the PLS cross validation did not achieve satisfactory results for C_{10} (RMSECV for C_{10} in Table 3.6). By grouping the concentrations as in combination 2 (section 3.5.4), the performance was improved to an extent.

3.5.3.2 Analysis of predictive power: Concentration-wise

Table 3.6 summarizes the results of all the regression techniques used for the calibration models built with operating conditions as the inputs and entire product composition as the outputs. Only the model performance measures along with cross-validation errors are given in this table. Refer to Table A.6 in Appendix A for the tuned hyper parameters of the calibration models.

From the table, we can see that on the whole, the LS-SVM based methods perform better than PLSR. The average decrease in RMSEP (across all outputs) from PLSR to LS-SVM-Poly was $30.71 \pm 22.73\%$ and a much larger decrease of $51.18 \pm 16.31\%$ from PLSR to LS-SVM-RBF. It is clear that among the two kernels used for LS-SVM, the RBF kernel outperformed the polynomial kernel for all outputs except C_7 (the polynomial kernel exhibited 12.5% lower prediction error than the RBF) and C_{10} (RMSEP for the polynomial kernel was lower by a meagre 2.97% compared to the RBF). On an average, the prediction error while using the RBF kernel was 26.37 ± 16.54 % lower than when using the polynomial kernel.

Examining the tuning parameters revealed that the operating condition-concentration dependence is highly nonlinear. This is conspicuous from the low values of σ^2, with a maximum value of 41 for C_3 (Table A.6 in Appendix A). Chauchard et al. [31] reported that similar values for σ^2 in their LS-SVM regression model used for predicting acidity of grapes from NIR spectra was suggestive of a high degree of nonlinearity. The optimized degree of the polynomial kernel increased, indicating a higher extent of nonlinearity of the model (C_3 had the highest σ^2 of 41 and smallest d of 3; C_{10} had the lowest σ^2 of 4 and a high d of 6 common to 5 other outputs). The superior capacity of the RBF kernel-based LS-SVM technique to model extremely nonlinear systems was apparent from the fact that the decline in RMSEP from PLSR to LS-SVM-RBF was high for the heavier hydrocarbons (C_5+) with the largest error decrease being for C_5 (74.17% with a low σ^2 value of 5). This was true with the polynomial kernel as well.

It was also interesting to see that the calibration error of C_7 and C_{10} was greater than the respective validation set numbers in the PLS model. This can be attributed to the inadequate variance explained by the first LV. LS-SVM-Poly also displayed similar characteristics for C_7 and C_{10}, thus decreasing the stability of the polynomial LS-SVM over a range of datasets that are not linearly correlated. This observation is not new and has been observed by previous researchers [20,38] who concluded that the model with a lower RMSEC was more robust.

Table 3.6. Regression results of combination 1 for outlet concentrations of each carbon number. R_P^2, RMSEP and RAE_p are the variance explained in the outputs, the prediction error and the Relative Absolute Error in the validation set, respectively. R_C^2 and RMSEC are the variance explained in the outputs and root mean square error of the calibration set, respectively. RMSECV is the error of cross validation calculated for the training set.

C No	Model [a]	R_P^2	RMSEP	RMSECV	RAE_p [b]	R_C^2	RMSEC
C_1	PLS	.85	.0026	.0012	.39	.93	.0017
	Poly	.88	.0024	.0002	.27	1.00	.0001
	RBF	.96	.0013	.0010	.17	.99	.0007
C_2	PLS	.87	.1056	.0499	.38	.97	.0775
	Poly	.82	.1254	.0275	.37	.99	0.032
	RBF	.95	.0677	.0261	.23	.99	.0112
C_3	PLS	.95	1.27	.52	.19	.99	.39
	Poly	.98	.91	.40	.1461	.99	.29
	RBF	.98	.87	.39	.1414	.99	.27
C_4	PLS	.45	1.66	.76	.63	.85	.72
	Poly	.78	1.05	.19	.52	.99	.11
	RBF	.86	.83	.21	.41	.99	.10
C_5	PLS	.83	1.51	.72	.36	.95	.69
	Poly	.94	.91	.21	.23	.97	.43
	RBF	.99	.39	.19	.10	.99	.13
C_6	PLS	.59	1.73	.94	.61	.87	.79
	Poly	.81	1.16	.42	.39	.98	.29
	RBF	.89	.79	.40	.32	.99	.23
C_7	PLS	.83	.38	.57	.48	.68	.47
	Poly	.93	.24	.49	.29	.78	.39
	RBF	.91	.27	.49	.32	.93	.25
C_8	PLS	.59	.93	.31	.71	.83	.34
	Poly	.91	.44	.12	.27	.99	.13
	RBF	.97	.26	.12	.18	.99	.11
C_9	PLS	.82	.32	.23	.36	.89	.13

	Poly	.89	.25	.17	.27	.99	.08
	RBF	.95	.17	.11	.22	.99	.05
	PLS	.53	.17	.19	.86	.39	.23
C_{10}	Poly	.81	.0605	.0519	.49	.75	.0728
	RBF	.79	.0623	.0533	.55	.95	.0141

[a] Polynomial-based LS-SVM methods abbreviated as Poly in table and LS-SVM-Poly in text. RBF-based LS-SVM methods abbreviated as RBF in table and LS-SVM-RBF in text; [b] RAE values computed on the validation set.

Finally, the gap between RMSECV and RMSEP as well as RMSEC and RMSEP seemed to decrease while moving from PLSR to LS-SVM-Poly to LS-SVM-RBF in the majority of the outputs, showing a curtailed tendency to overfit and increased consistency in the RBF-based LS-SVM methods. Also, the highest correlations between measured and forecasted values were obtained for the LS-SVM-RBF calibration model, except for C_7 and C_{10} where the polynomial LS-SVR achieved slightly better results for R^2 in the validation dataset.

3.5.3.3 Analysis of predictive power: Validation set experiment-wise

The main objective of this paper was to inspect the prediction ability of the multivariate calibration techniques not only for individual concentrations but also the product composition taken as a whole for each experimental condition. Table A.7 in Appendix A shows the results of all three regression methods organized experiment-wise.

Here again, the improved performance of the LS-SVM-RBF is quite palpable. As compared to PLSR and LS-SVM-Poly, the averaged RMSEP values for LS-SVM-RBF were reduced by $40.59 \pm 26.18\%$ and $25.23 \pm 19.56\%$, respectively averaged over the entire validation set. The numbers are also very similar in terms of RAE, where LS-SVM-RBF lowered the averaged RAE_p by $41.43 \pm 26.12\%$ and $22.11 \pm 18.95\%$ from PLSR and LS-SVM-Poly, respectively. The range of inlet reaction conditions for the feed (T & $F_{m_{RT}}$) in the validation set aids in assessing the performance of the LS-SVM-RBF method. While LS-SVM-RBF outperforms PLS and LS-SVM-Poly in the entire range, the efficiency of the polynomial kernel decreases at higher temperatures and flowrates of propylene. For experiments 23, 24 and 25, the polynomial based LS-SVM performs worse than PLSR by 10, 20 and 16%, respectively. An insight into the reaction chemistry

might provide the explanation for this behavior. [2] C_6 is one of the major products obtained in the oligomerization process and at increased flowrates (experiments 21-25), the formation of C_7+ hydrocarbons is constrained by the lower residence times in the catalytic zone. Moreover, the higher temperature can lead to significant cracking rates, leading to the formation of more light and mid-stream products (C_2 to C_6). Table 3.6 indicated that RBF works better than the polynomial kernel for C_1-C_6 predictions (29.07 % lower RMSEP averaged over C_1-C_6). Prediction errors averaged for the calibration set (experiments 6-20) were very similar for both the polynomial and RBF kernel-based LS-SVM methods (0.14 ± 0.10 for both) while being much larger for PLSR (0.42 ± 0.20).

3.5.4 Combination 2: Inlet reaction conditions as predictors and lumped product composition as outcomes

Variable selection methods combined with SVMs have been employed in the past to improve the working of the regression models. [32] In this section, we have incorporated the product streams, grouped on the basis of PCA (refer section 3.5.1), as outputs to the developed regression techniques to look for any improvements over combination 1.

3.5.4.1 Factor selection for PLSR

As discussed in section 3.4.4, the RMSECV criterion was employed for choosing the number of LVs for the PLSR calibration model on the training set for this combination as well. Though a significant decrease in the RMSECV was observed for the first LV for all 4 product stream groups, the light product stream (C_1, C_2) showed the largest decrease in RMSECV (58.72%) when the second LV was added while the heavy stream showed the smallest difference (6.02%) between the first and second LV. This was similar to what was seen with the full concentration RMSECV (described in section 3.5.3.1). Interestingly, the variance described by the 2 LVs (Table A.8 in Appendix A) was greater for this combination as compared to combination 1. For the heavy stream (C_7-C_{10}), 93.5% of the total variance was explained by 2 LVs in PLSR, while only ~40% was accounted for if the second LV was considered for C_{10} in combination 1. This showed a benefit in classifying the products into groups according to their reaction pathways.

Since the decrease in RMSECV on addition of the 2nd LV for the heavy stream was still significant (-0.036), 2 factors were chosen for constructing the PLSR model as was the case for other outputs as well. The minimum RMSECV values are reported in Table 3.7 along with the other model performance parameters.

3.5.4.2 Analysis of predictive power: Lumped concentration stream-wise

Table 3.7 comprises the performance features of all three regression techniques. Table A.8 in Appendix A reports the corresponding tuned hyper parameters.

Complementary to the discussion in the previous section (section 3.5.4.1), a striking improvement in the performance of PLSR is seen when the output variables are grouped according to PCA & HCA. This can be attributed to the increased capture of the variability in the grouped composition when compared to combination 1. The average squared correlations for PLSR between the predicted and observed samples across all outputs spiked by 14.5% for combination 2 as compared to combination 1. The RMSEP, RMSECV and RMSEC are of the same order of magnitude for the lighter stream for LS-SVM-Poly, indicating a more stable calibration in combination 2 than combination 1, where the RMSECV and RMSEC were much lower than RMSEP (Table 3.6).

Though the polynomial kernel performs better in the case of light and dimer streams, lower than satisfactory performance was achieved with the heavier streams (streams 3 and 4 - Table 3.7). A plausible reason for this lies in the optimized hyper-parameter d. The training process tunes the polynomial to a high degree that the calibration becomes flexible [62] to such an extent that it fits the calibration samples very well but fares badly with the test data (R_C^2 = 0.99 & 0.97; R_P^2 = 0.62 & 0.42 for the mid- and heavy streams, respectively). In combination 1, the polynomial kernel performed well for the validation set possibly due to the higher γ values for C_7 (31) and C_{10} (22) than for the entire heavy stream (0.23).

Table 3.7. Regression results for combination 2 for the associated stream-wise outlet concentrations. R_P^2, RMSEP and RAE_p are the variance explained in the outputs, the prediction error and the Relative Absolute Error in the validation set, respectively. R_C^2 and RMSEC are the variance explained in the outputs and root mean square error of the calibration set, respectively. RMSECV is the error of cross validation calculated for the training set.

S. No [a]	Model [b]	R_P^2	RMSEP	RMSECV	RAE_p [c]	R_C^2	RMSEC
1	PLS	.87	.1074	.0341	.38	.97	0.0388
	Poly	.91	.0879	.0278	.29	.99	.0167
	RBF	.95	.0693	.0265	.23	.99	.0099
2	PLS	.91	2.53	1.41	.28	.97	1.06
	Poly	.99	.73	.36	.09	.99	.21
	RBF	.99	.49	.28	.06	.99	.16
3	PLS	.62	3.14	1.81	.54	.92	1.37
	Poly	.62	3.18	1.74	.65	.99	.81
	RBF	.73	2.67	1.52	.57	.99	.92
4	PLS	.94	.81	.56	.24	.94	.31
	Poly	.42	2.45	.52	.61	.97	.28
	RBF	.96	.73	.51	.22	.97	.31

[a] 1 – Light stream, 2 – Dimer stream, 3 – Mid stream, 4 – Heavy stream (Table 3.3); [b] Polynomial-based LS-SVM methods abbreviated as Poly in table and LS-SVM-Poly in text; RBF-based LS-SVM methods abbreviated as RBF in table and LS-SVM-RBF in text; [c] RAE values computed on the validation set.

It is quite lucid that the LS-SVM-RBF outperforms the other two methods for all 4 output predictions ($35.23 \pm 32.22\%$ decrease in RMSEP from PLSR to RBF and $35.07 \pm 24.46\%$ from polynomial to RBF based LS-SVM). The reason for the large standard deviation for the error decrease from PLSR to RBF was that the smallest decrease was 9.87% for the heavy stream but the largest decrease was for the dimer stream (80.63%), also implying an improved efficiency for the PLSR. RBF was seen to be consistently efficient in not overfitting the calibration samples as the gap between RMSEP and RMSECV is the least for LS-SVM-RBF (0.41) among the three techniques, though PLSR was not far behind (0.69). The nonlinear relationship between inlet reaction conditions and the lumped concentrations at the outlet is confirmed by the low values of

σ^2, with the smallest being 1 for the heavy stream (Table A.8 in Appendix A). The largest decline in prediction error from PLSR to LS-SVM-RBF was 80.63% for the dimer stream with a similarly sparkling decrease of 70.20% from Polynomial to RBF based LS-SVM for the heavy stream.

The analysis of the regression results for the experiment-wise arrangement showed the best performance for the RBF kernel-based LS-SVM (RAE_p lowered by 60.86 $\pm$ 36.92% from PLSR and 52.54 $\pm$ 8.98% from LS-SVM-Poly). There seemed to be an inherent advantage in classification of products based on PCA & HCA as is evident from the higher reduction in prediction errors for LS-SVM-RBF from PLSR and LS-SVM-Poly compared to combination 1 (section 3.5.3.3). Similar to the results of combination 1, LS-SVM-Poly performed worse than PLSR for experiments 23 and 24, exposing its weakness to capture the changes in the system at higher temperatures and lower residence times. The difference between the averaged prediction errors and calibration errors over the 15 training experiments for PLSR, LS-SVM-poly and LS-SVM-RBF were 0.73, 0.90 and 0.44, respectively. This showed that the least overfitting characteristics were obtained for RBF-based LS-SVM.

3.5.5 Combination 3: FTIR full spectrum intensities as predictors and full product composition as outcomes

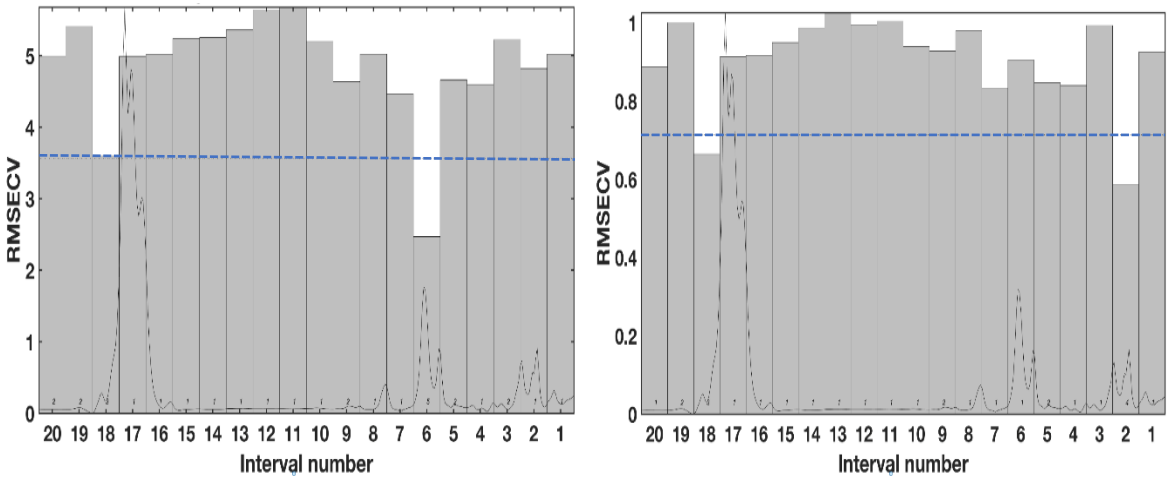
The precision of LS-SVM-RBF in the previous models is perhaps surprising, but it should be noted that the operating conditions vary between 650-750 K, where the gasoline-range equilibrium carbon number distribution does not alter dramatically (compare with Figure 8 in Tabak et al. [2]). Furthermore, the presence of aromatics, though not tracked explicitly, is known to affect the distribution. It should be kept in mind that the hydrocarbon aromatics contribute only to the C_6 and heavier part of the distribution. An issue with the predictions based on inlet operating conditions is that any small changes that affect reactor operation (caused, for example, by channels developing in the reactor, or catalyst deactivation) could negate the validity of the calibration relation built between the operating conditions at the inlet and the product distribution at the outlet. It would be preferable to develop a more reliable model based on the spectroscopic information, since that is gathered online and is an accurate representation of the chemical nature of the products.

3.5.5.1 Optimum factor selection for PLSR and *i*-PLSR

The prime purpose of the interval PLS model was to identify whether any particular region of the IR spectrum captured the variation in the output better than the full spectrum, where the chances of interference are higher due to absorbance from multiple species. The entire spectral range was divided into 20 subintervals, each comprising 17 wavenumbers. For each carbon number concentration output, in addition to optimizing the number of LVs with the RMSECV minimization for the global/full spectrum PLS model, the best performing interval was also chosen based on the comparison of the minimum RMSECV for each divided spectral domain to the global RMSECV minimum. The cross-validation errors obtained for C_3 and C_8 are shown in Figure 3.7a and Figure 3.7b, respectively. The RMSECV plots revealed that only interval 15 for C_3 (from the left in Figure 3.7a) and intervals 3 and 19 for C_8 (from the left in Figure 3.7b) could compete with the respective 10-component full spectrum models. The optimized LV for the sub-intervals are indicated at the bottom of the bars in the figures. In a similar way, the details of the best performing intervals for each carbon number are outlined in Table A.9 in Appendix A.

As can be seen from Table A.9 in Appendix A, there was a decrease in the cross-validation errors of the *i*-PLS calibration for all outputs except C_9 and C_{10}, where no particular range of wavenumbers could give better predictions than the full spectrum PLS model. In other words, for C_1 to C_8, the maximum amount of information in the spectrum that correlates with the respective output concentrations is concentrated within the spectral region indicated by the best performing regions in Table A.9 in Appendix A. Moreover, as the carbon number of the product increased (specifically observed for C_4 and C_6), more intervals showed improved performance compared to the global PLS model, which was a sign of a rise in the number of chemical species absorbing due to isomerization. [5] It can be argued that in the case of C_9 and C_{10}, every interval in the whole spectrum was more important than any particular region.

Furthermore, though *i*-PLS enables the analyst to qualitatively interpret the chemical nature of the products better than PLS, a direct cause-effect assignment for the spectrum-concentration dependence is a challenging task. Here, the developed *i*-PLS model provides a tentative indication of the predominant functional groups that are representative of the products as illustrated in the right column of Table A.9 in Appendix A. Products from C_1 to C_5 mainly correlate with the presence of sp^3 C-H bonds, though the spectral region of 1535 – 1396 cm^{-1} might also contain absorbance due to aromatic C=C stretching, which can obviously only start appearing from C_6.



(a) (b)

Figure 3.7. Bar graphs depicting RMSECV vs. each wavenumber interval in *i*-PLS calibration model for combination 3 with outputs: (a) C_3 and (b) C_8. The dotted line is the RMSECV for the global PLS model with 10 LV (optimized) for both the outputs.

However, an additional interval that corresponded to aromatic C-H stretching vibrations performed equally well for C_6, as with C_8. Expectedly, the presence of mono and di-substituted olefins and aromatics are also seen for C_7 and C_8. On the whole, the benefit of employing interval PLS methods for regression was seen and the best performing intervals were chosen for concentration prediction in the validation set for this model. The results seemed to match the discussion in the section on section 3.5.2 as to why the 6-cluster division explained the product chemistry better. Combinations 5 and 6 will further emphasize the importance of wavelength selection and dimension reduction to maximize input-output correlation.

3.5.5.2 Analysis of predictive power: Concentration-wise

The outcomes of the PLS-based and LS-SVM-based calibrated models are compiled in Table 3.8. Refer to Table A.10 in Appendix A for the corresponding tuned hyper parameters.

On the whole, the comparison of the statistical parameters in Table 3.8 for the regression models re-emphasized that LS-SVM-RBF achieved better results than other MDA techniques. To put things in perspective, the prediction errors for LS-SVM-RBF, as compared to PLS, *i*-PLS and LS-SVM-Poly were smaller by 51.27 ± 7.38%, 30.61 ± 23.98% and 24.41 ± 13.99%,

respectively. Furthermore, one can see that the numbers indicate a slightly superior performance of the techniques in combination 1 (section 3.5.3.2) in terms of the absolute RMSEP and R^2 values. The reason for this is that the higher dimensionality of the spectroscopic data that affected the degree of the fit. [63] However, the standard deviations tell a marginally different story. For combination 1, the average RMSEP decrease from PLS to LS-LSVM-RBF was 51.18 $\pm$ 16.31% and from Poly to RBF was 26.37 $\pm$ 16.54%. The lower variability in the improved performance of RBF indicated that it outperformed PLS and LS-SVM-Poly on a more consistent basis across all outputs than in combination 1.

Moreover, the performance of the interval PLS model matched the general objective of this work, i.e. it not only gave a much better performance than the PLSR (RMSEP lowered by 28.61 $\pm$ 8.14% averaged over C_1 to C_8) but the better performing intervals hinted at the possible reaction chemistry of the system. A more detailed discussion was presented in section 3.5.5.1. It was also able to compete with the LS-SVM methods (RMSEP for Polynomial and RBF were lower than *i*-PLS only by 11.4% & 30.6% on an average respectively) to a higher degree than PLSR (RMSEP for Polynomial and RBF were lower by 51.3% and 34.2%). In fact, in the case of C_1 and C_2, the RMSEP of the RBF kernel was lower than *i*-PLS by a mere 0.1% and 1.66%, respectively, while the difference was larger for higher carbon numbers, signifying better capability of RBF kernels for modeling nonlinearity. One difference between *i*-PLS and the clustering approach is that the spectrum is divided into sub-intervals in a contiguous manner in *i*-PLS whereas the clusters that will be used as inputs to combinations 5 and 6 (section 3.5.7) are a counterpoint to this approach and not contiguous but still chemically meaningful.

Moving ahead to the tuning parameters, the γ values in combination 3 (Table A.10 in Appendix A) are an order of magnitude higher than those in combination 1 (section 3.5.3.2). This has to be accounted for here. Though the complete spectral data may contain correlated variables within the different wavenumber channels, there is no guarantee of any level of covariance with the output concentrations as no latent variables have been extracted prior to application of LS-SVM-RBF. Consequently, the RBF kernel would have a low value, signifying lower sample-sample similarity, which is compensated by increase in γ. This type of phenomenon has also been observed previously by Chauchard and coworkers. [31]

Table 3.8. Regression results of combination 3: Carbon number-wise concentration output. R_P^2, RMSEP and RAE_p are the variance explained in the outputs, the prediction error and the Relative Absolute Error in the validation set, respectively. R_C^2 and RMSEC are the variance explained in the outputs and root mean square error of the calibration set, respectively. Only data for specific carbon numbers in the products are presented (C_1, C_2, C_4, C_6, C_8 and C_{10}).

C No.	Model [a]	R_P^2	RMSEP	RAE_p [b]	R_C^2	RMSEC
C_1	PLS	.66	.0040	.39	.75	.0035
	i-PLS	.84	.0027	.38	.80	.0028
	Poly	.87	.0024	.30	.89	.0017
	RBF	.93	.0018	.18	.96	.0014
C_2	PLS	.75	.1968	.46	.77	.1132
	i-PLS	.87	.1058	.36	.98	.0381
	Poly	.84	.1165	.41	.96	.0461
	RBF	.91	.0892	.21	.98	.0303
C_4	PLS	.34	1.82	.73	.80	.85
	i-PLS	.67	1.29	.64	.90	.59
	Poly	.77	1.08	.49	.92	.55
	RBF	.82	.95	.45	.94	.48
C_6	PLS	.58	1.74	.63	.87	.80
	i-PLS	.74	1.37	.45	.89	.74
	Poly	.80	1.20	.38	.94	.56
	RBF	.92	.76	.26	.97	.39
C_8	PLS	.55	.97	.60	.86	.30
	i-PLS	.76	.70	.43	.94	.20
	Poly	.90	.47	.27	.96	.16
	RBF	.93	.37	.18	.93	.21
C_{10}	PLS	.38	.1098	.72	.99	.004
	i-PLS	-	-	-	-	-
	Poly	.78	.0646	.54	.90	.0178
	RBF	.75	.0692	.48	.91	.0170

[a] Polynomial-based LS-SVM abbreviated as Poly in the table and as LS-SVM-Poly in the text. RBF-based LS-SVM abbreviated as RBF in the table and as LS-SVM-RBF in the text; [b] RAE values computed on the validation set.

However, there is no change in the interpretation for σ^2, whose values are on the lower side, thus amplifying the nonlinearity in the spectrum-concentration relationship. C_7 has the lowest σ^2 of 4 with a corresponding high degree of 6 for the polynomial kernel. The next lowest σ^2 value of 5 was shared by C_6, C_8 and C_{10} (Table 3.8).

Finally, the quality and applicability of a regression model is judged mainly by its capability to fit all kinds of datasets equivalently. Alternatively put, the difference between RMSEP and RMSECV was lowest for LS-SVM-RBF on an average across all outputs (0.16) as opposed to PLS (0.53), *i*-PLS (0.32 – averaged from C_1 to C_8) and LS-SVM-Polynomial (0.21). The gap between errors of calibration and prediction also showed similar trends - LS-SVM-RBF (0.20) < LS-SVM-Poly (0.34) < *i*-PLS (0.35) < PLS (0.45). It should be noted that although the average differences are reported, one can verify that the individual output-wise gaps between the aforementioned error values are also least for LS-SVM-RBF. This indicated that a regression model built with LS-SVM-RBF method demonstrated the least tendency to overfit.

3.5.5.3 Analysis of predictive power: Validation set experiment-wise

Table A.11 in Appendix A provides the results of the experiment-wise performance of the PLS, LS-SVM-Poly and LS-SVM-RBF. *i*-PLS was not included because of its inability to improve over PLSR for C_9 and C_{10} as detailed in the previous section. The superior performance of LS-SVM-RBF is established from the results as seen in Table A.11 in Appendix A, since it outperformed the other 2 models for all the new inlet operating conditions in the test set. The RMSEP and RAE_p for LS-SVM-RBF were lower than its counterparts: Polynomial kernel-based LS-SVM by 36.73 $\pm$ 15.21% and 34.47 $\pm$ 16.06%, respectively; PLS by 55.37 $\pm$ 17.56% and 56.81 $\pm$ 17.19%, respectively. These values were somewhat higher than that for combination 1 although the absolute values of errors and coefficients of determination were better for combination 1. It shows that the nonlinearity caused by the high dimensional spectroscopic data is incorporated more effectively by the LS-SVM methods, particularly with the RBF kernel than by

the PLS based methods. LS-SVM-Poly showed ineffective performance for experiments 4, 24 and 25, exhibiting higher errors than PLSR, emphasizing the effect of higher temperature on the nature of product distribution. [2] The LS-SVM techniques had smaller differences between the RMSEP and RMSEC in combination 3 as compared to combination 1 but this was reversed for PLSR. To add to this, among the three calibration methods for combination 3, LS-SVM-RBF possessed the least average gap between RMSEP and RMSEC (0.19 as compared to 0.45 for LS-SVM-Poly and 0.56 for PLSR), which again showed the diminished propensity to overfit the calibration data in the LS-SVM-RBF approach.

3.5.6 Combination 4: FTIR full spectrum intensities as predictors with lumped product composition as outcomes

The prediction accuracy of the LS-SVM-RBF was the best for the 4 grouped product streams among all the MDA methods compared for this input-output combination namely, PLSR, *i*-PLS and LS-SVM-Polynomial kernel. The best performing intervals extracted by *i*-PLS are as follows: 1535 – 1396 cm^{-1} for the light stream and mid-stream; 3139 – 3008 cm^{-1} & 1535 – 1396 cm^{-1} for the dimer stream and 3139 – 3008 cm^{-1} for the heavy stream. These results complement combination 3 and the reader is referred to section 3.5.5.1 for the chemical interpretation. For brevity, further detailed results of this model are not shown in this chapter.

3.5.7 Comparing the results of combinations 5 and 6: HCA clustered intensities as inputs and full product concentrations (C_1-C_{10}) as outputs

The FTIR spectrum (350 wavenumbers) of the reaction products were grouped into 4 and 6 clusters to investigate the trade-off between input dimensionality and the amount of information contained in each input variable. In this section, the performances of the combinations 5 and 6 are compared among themselves as well as with combination 3, where the full spectrum intensities were considered as inputs. The aim was to see the effect of PCA and HCA on the performance of the regression models, specifically LS-SVM-RBF. The results of combinations 5 and 6 are summarized in Table 3.9.

First of all, one can clearly see that the LS-SVM RBF technique achieved better results than LS-SVM-Poly and PLSR models for all outputs listed in Table 3.9. Looking at the numbers for combination 5, RMSEP values for the RBF kernel were lower than PLS and polynomial kernel-based LS-SVM by 49.61 ± 12.17% and 29.66 ± 15.34% averaged across all carbon number outputs, respectively. For combination 6, LS-SVM-RBF outperformed PLSR and LS-SVM-Poly by 51.74 ± 13.51% and 33.09 ± 14.81% in terms of prediction errors, respectively. In both cases where clustered intensities were used as inputs to build the calibration models, it was noticed that the polynomial kernel performed slightly worse than PLSR for C_2 (5.59% and 0.89% higher error for combination 5 and combination 6, respectively) as was also observed in combination 1. On similar lines, *i*-PLS not only had a lower RMSECV but also predicted better than LS-SVM-Poly in combination 3 for C_2 (Table 3.8). This could be attributed to a higher degree fit of the polynomial for C_2 (6 for combination 1 and 5 for combination 3) that resulted in overfitting akin to the heavy stream in combination 2 (read section 3.5.4.2).

Table 3.9. RMSEP values for independent test set of combinations 5 and 6, concentration wise.

C. No.	PLSR		LS-SVM-Poly		LS-SVM-RBF	
RMSEP	**4 clusters [a]**	**6 clusters [b]**	**4 clusters**	**6 clusters**	**4 clusters**	**6 clusters**
C_2	.1016	.1007	.1073	.1017	.0636	.0614
C_3	1.09	1.03	.73	.67	.34	.32
C_4	1.28	1.19	.80	.77	.70	.62
C_5	1.06	.95	.79	.71	.69	.59
C_6	1.35	1.14	1.01	.89	.66	.57
C_7	.37	.33	.21	.20	.17	.15
C_8	.73	.66	.49	.44	.32	.27
C_{10}	.0855	.0782	.0551	.0455	.0287	.0253

[a] Combination 5; [b] Combination 6.

Next, the drawbacks of the RBF model that were seen in combinations 1 and 3 were well addressed by applying clustering to the spectra. LS-SVM-Poly performed better than RBF for C_7 and C_{10} outputs in combination 1 and for C_{10} in combination 3. Classifying the outlet carbon number concentrations on the basis of PCA and HCA and developing a regression model on the

lumped concentration outputs (combination 2) had alleviated this issue and the RBF fared better than the polynomial kernel for the higher carbon number stream (Table 3.7). Likewise, in combination 5, the prediction errors for RBF were lower than the polynomial kernel by 16.96% & 47.77% and in combination 6, by 22.85% & 44.22% for C_7 and C_{10}, respectively. In addition to this, the average gap across all outputs between RMSEP and RMSEC in combination 6 was lowest for LS-SVM-RBF (0.0685 for RBF < 0.21 for LS-SVM-Poly < 0.41 for PLSR). Conjointly, RMSEP differed from RMSECV, on an average by the smallest amount for LS-SVM-RBF (0.0612 for RBF < 0.10 for Polynomial < 0.23 for PLSR). These numbers clearly imply that the RBF kernel-based LS-SVM approach has a minimized risk of overfitting compared to the other two methods used in this study.

Lastly, the advantage of clustering the wavenumbers into 6 groups was clear from the results reported in Table 3.9. Though it might seem practical to expect better results from combination 5 with lesser number of input variables (4 clusters), it is thought that the utility of six clusters lies in its chemical meaning. Table 3.4 and Table 3.5 illustrate that the aromatic stretching vibrations were better identified in the 6-cluster division than in the 4-cluster division, where it merged with the baseline intensities. Also, the sp^3 –C-H bending deformations were also separated from other overlaps and this frequency band was also identified by *i*-PLS to be present in a majority of the products. The 4-cluster arrangement was not able to do this. This perhaps resulted in the improved prediction ability that was observed for C_6-C_{10} hydrocarbons on the whole, where the averaged decrease in RMSEP values in the 6-cluster combination as compared to 4-cluster division for C_6-C_{10} and C_1-C_5 was 17.89 $\pm$ 11.36% and 7.85 $\pm$ 5.01%, respectively. Remarkably, for the entire C_1-C_{10} range, the decreasing order of prediction ability for LS-SVM-RBF (averaged RMSEP in brackets) was also: Combination 6 (0.27) > Combination 5 (0.31) > Combination 1 (0.37) > Combination 3 (0.43). Similar trends were seen for LS-SVM-Poly as well. There was also a significant enhancement in the predicting ability of PLSR (increasing averaged RMSEP): Combination 6 (0.58) > Combination 5 (0.64) > Combination 1 (0.81) > Combination 3 (0.93). Even on an individual output basis, the regression models with clustered spectra as the inputs had lower prediction errors than in other combinations. Within the two cluster approaches, a pivotal factor to which the superior performance of the MDA techniques can be attributed is the separate formation of a cluster representing sp^3-C-H bending vibrations (6^{th} cluster - Table 3.5). The enhanced possibility of the presence of this bond in the entire product distribution combined with

the reduced dimensionality of the 6-cluster division improved the overall predictive power of combination 6 over other combinations.

On the other hand, when the outputs were organized experiment-wise, the RAE_p and RMSEP values for the RBF kernel-based LS-SVM method were the lowest among PLSR, LS-SVM-Poly and LS-SVM-RBF, which is suggestive of better forecasting performance of the entire product composition for LS-SVM-RBF in combinations 5 and 6 also. Thus, it would appear that the dimensionality of the input can be increased to improve the predictive capability of the regression model, as long as the input contains groups that are developed in a statistically or chemically meaningful sense.

Finally, though the tuning parameters of LS-SVM-RBF for combinations 5 and 6 were not explicitly reported, the σ^2 values were found to be somewhat lower and in the same order of magnitude as in combinations 1 and 3. This was indicative of a nonlinear relationship between clustered intensities and outlet concentrations of the carbon numbers. C_{10} had the highest γ of 794 for combination 6 and the lowest σ^2 value of 3. The γ value was lesser than in combination 3 because of the reduced dimensions and tapered noise levels in the clustered data.

3.5.8 Correlation plots between measured and predicted values for all combinations

The correlation plots for envisioning the relationship between the experimentally observed values and LS-SVM-RBF forecasted values for each of the 6 combinations are depicted in Figure 3.8.

Since there are multiple outputs, only the plots for C_5 hydrocarbons (higher outlet concentration) in combinations 1, 3, 5, 6 and heavy stream in combinations 2 and 4 are shown. All values of coefficients of determinations are reported in the respective sections where each regression model was discussed previously. The only additional comment that can be made from the correlation plots in Figure 3.8 is that caution needs to be applied while extrapolating the prediction correlation for lower concentrations of the heavy stream in combination 4. Overall, it can be stated with confidence that the RBF kernel-based LS-SVM method was able to predict equally well at all the concentration ranges for the outputs dealt with in this study.

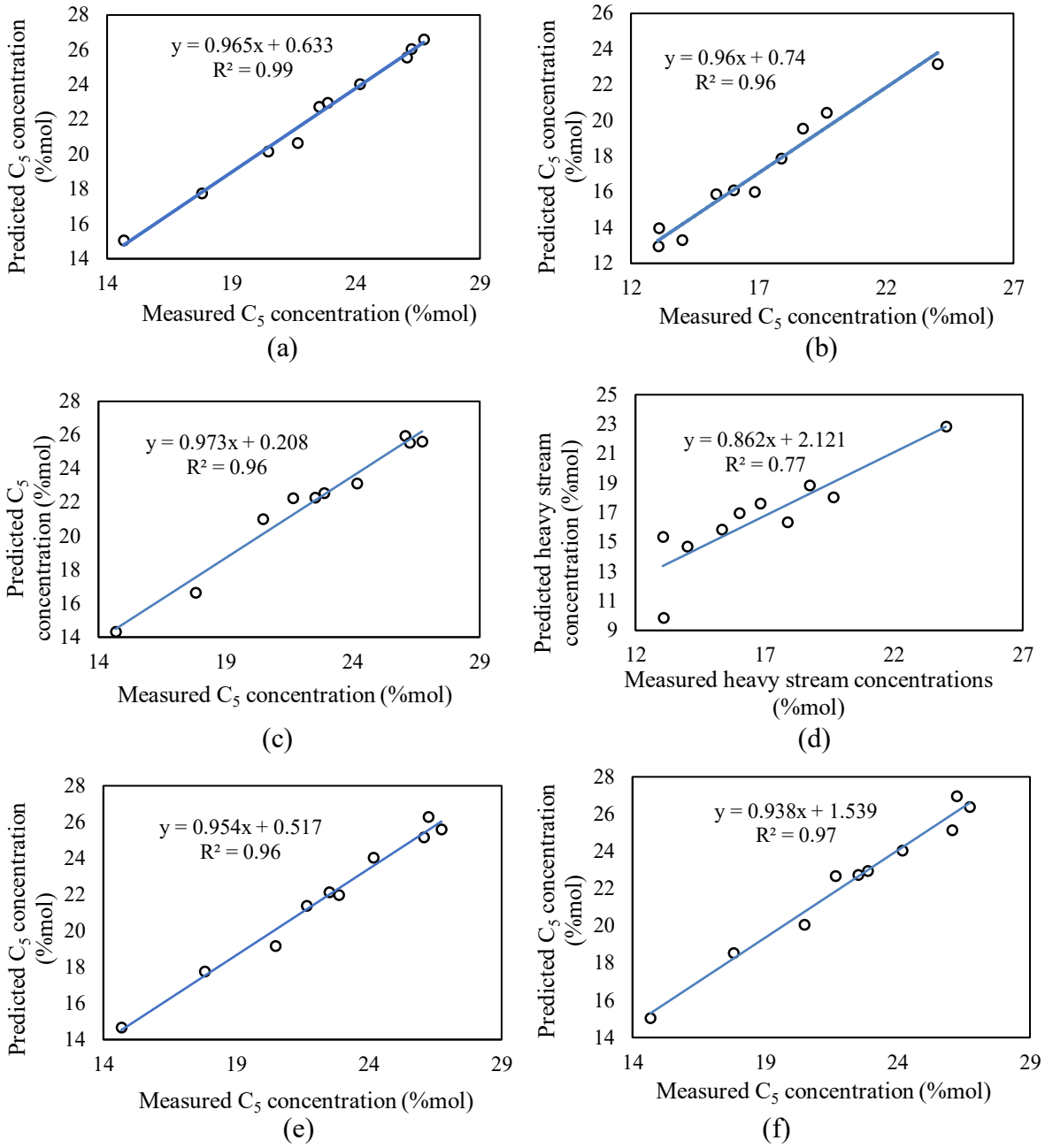


Figure 3.8. Correlations between experimentally measured and LS-SVM-RBF predicted values for concentrations of: (a) C_5 in Combination 1; (b) Heavy stream in Combination 2; (c) C_5 in Combination 3; (d) Heavy stream in Combination 4; (e) C_5 in Combination 5; (f) C_5 in Combination 6.

In order to visualize the product distribution profile, Figure 3.9 shows the product compositions obtained from GC analysis for experiments 1 (346 ºC and 172 mL/min) and 23 (433

ºC and 440 mL/min – Table A.3 in Appendix A) along with the corresponding predicted values from the LS-SVM-RBF using 6-clustered intensities (combination 6). This distribution is similar to Figure 8 in Tabak et al. [2] for the reaction conditions between 600-700 K. It can be seen that as the temperature rises, the concentrations of lighter carbon numbers (C_1-C_3) increases (Figure 3.9b). C_6 is also increased in experiment 23 due to oligomerization with smaller residence time. At higher residence times (experiment 1), due to lower WHSV, the C_7-C_{10} concentration is higher in the product (Figure 3.9a). The main purpose of showing these plots is to demonstrate that the prediction model resulted in slight overestimation of the product concentrations, especially for the heavier hydrocarbons. The explanation for this can be traced to the work by Tabak et al., [2] where they reasoned that it could be difficult for ZSM-5 to accommodate the exponential rise in the number of isomers as carbon number increases, leading to a slight decrease in the measured concentration value.

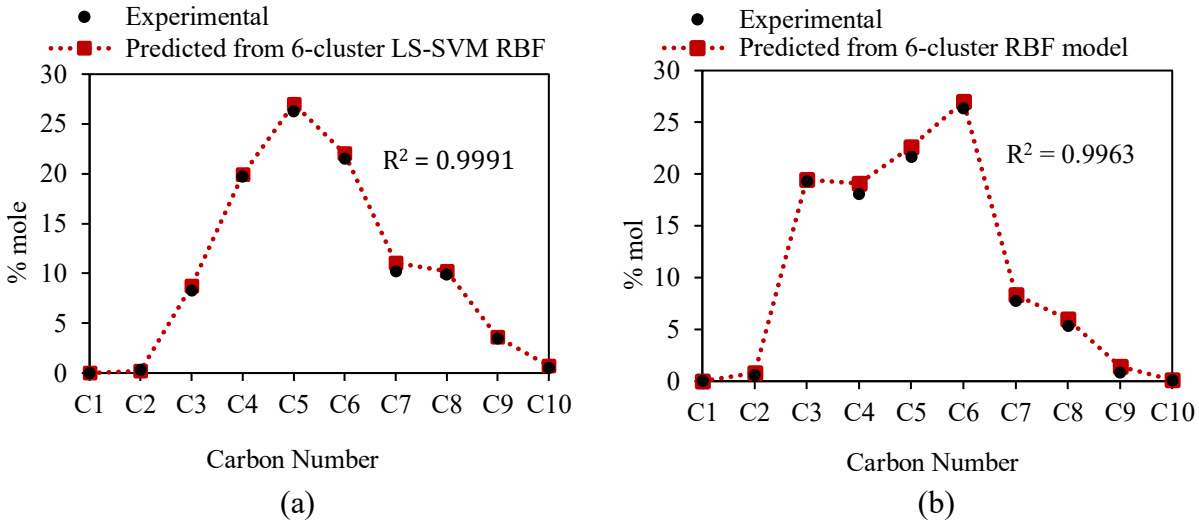


Figure 3.9. Product distribution profile for: (a) Experiment 1 at 346 ºC and 172 mL/min; (b) Experiment 23 at 433 ºC and 440 mL/min flowrate of propylene.

3.5.9 Effect of variation in training set size on model performance

As highlighted in section 3.4.1, all calibration models were built using 15 samples (experiments 6 – 20 in Table A.3 in Appendix A). Both the validation and training sets were composed of the entire concentration range of the products and the validation set comprised of the lowest and highest flowrates of propylene but with similar temperature ranges. This was not a

problem because the chemistry of the system was such that temperature seemed to have a greater control on composition of reaction products. [5] The main limitation of this work is the smaller number of samples available for training. Even with this constraint, it was important to assess the rate of model learning by testing the sensitivity of the model performance to the number of training samples. The benefit of this procedure can be considered as the future work of this study where the number of training samples that gave the least prediction error can be utilized for calibration. Also, reactions can be conducted at more operating conditions at lower temperatures, thereby increasing the number of training samples to more than 25.

A similar study was performed by Cui and Fearn [29] with NIR spectra where they investigated up to 420 training samples and found that the RMSEP attained a near constant value after about 250 samples. However, in our case, we increased the number of samples from 10 (where the remaining 15 samples used for validation) to 20 (where the remaining 5 samples used for validation) and calculated the RMSEP for C_6 concentration as the output in combination 6. For the sets that included 16, 17, 18, 19 and 20 samples for calibration, the corresponding validation sets consisted of samples from experiment numbers 12-20 (9 samples), 13-20 (8 samples), 14-20 (7 samples), 15-20 (6 samples) and 16-20 (5 samples) (refer to Table A.3 in Appendix A for the respective conditions indicated by the experiment numbers), respectively. This would be a case where the validation sets included samples that were interpolated in terms of both temperature and flowrate. The previous analyses included samples where the temperature ranges in the validation set were within that of the calibration set but the two flow rates were out of range of the values used for calibration. But since it was seen that the yield of the products was more sensitive to temperature than to flow rates, the validity of the constructed regression models still hold.

Figure 3.10 shows the effect of the number training samples on the RMSEP. It can be seen that the prediction errors decay with increasing number of samples in the calibration set for all 3 regression techniques. As expected, LS-SVM-RBF shows lowest error among the three methods and notably, both the LS-SVM approaches display slightly faster learning rates than PLSR (slight dip at 12 training samples), before the trends become very similar but did not attain a constant value. This particular test reiterates the need for a higher number of training samples to further optimize the LS-SVM model, at the same time giving more confidence for the analysis performed in the previous sections for the chosen number of samples in the calibration set.

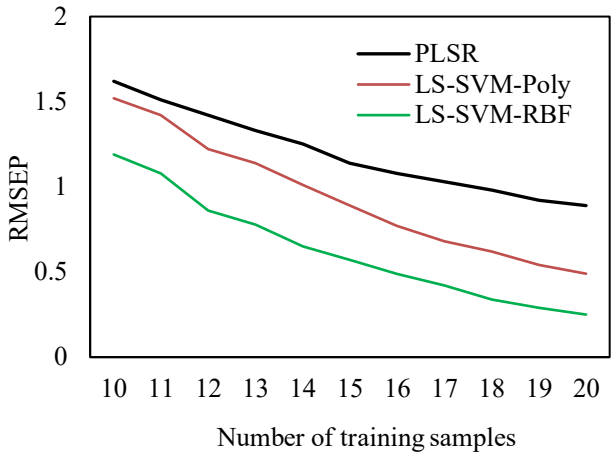


Figure 3.10. RMSEP vs. number of samples in used for building the calibration model tested for PLSR, LS-SVM-Poly and LS-SVM-RBF.

3.6 Conclusions

In this work, we have demonstrated the utility of using online spectroscopic information and inlet reaction conditions for monitoring of the moderately complex propylene oligomerization process over ZSM-5 catalyst, with the potential for real-time application. On the basis of evaluating different multivariate calibration techniques to track the product composition distribution, the following can be concluded:

(a) For the different predictor-output combinations tested, the performance of the RBF kernel-based LS-SVM method was found to be superior over other regression models employed, namely polynomial kernel-based LS-SVM, PLSR and *i*-PLSR (only for the full spectral input)

(b) In addition to the higher prediction accuracy, the training time (calculated inclusive of initialization, tuning and training across all outputs together) was smaller for the RBF kernel as compared to the polynomial kernel in the LS-SVM based methods. For example, when the full IR spectrum was used as input (combination 3), if the time for training for RBF kernel is taken as 1 time unit, polynomial kernels took 1.2 time units. In the case of combination 6 (six-division clustered spectral input), training times were reduced to 0.8 and 0.9 time units for RBF and polynomial kernels, respectively.

(c) The benefits of dimension reduction by feature extraction and using clustered inputs were evident from the fact that all regression models provided the most precise forecasted values for the product concentrations in the order (in terms of inputs to the models as): six-clustered spectrum > four-clustered spectrum > inlet operating conditions (temperature and flowrates of propylene) > full IR spectral intensities. The grouping of the product concentrations resulted in improved performance of PLSR.

(d) *i*-PLS regression was not only more effective than the simple PLSR for accurate concentration prediction, but also was able to provide a link between the measured property and possible reaction chemistry, supplementing the identified spectral groups in PCA and HCA.

(e) The hyper-parameters for LS-SVM-RBF and LS-SVM-Poly needed to be optimized carefully and in certain cases of higher carbon numbers, the efficiency of the polynomial kernel fell below PLSR due to a larger tuned degree that led to overfitting. On the other hand, LS-SVM-RBF exhibited the least tendency to overfit among all the methods investigated in this work.

(f) An interesting observation was that with increasing product molecular weight, the concentrations were slightly overestimated more often, though not by a large amount. This could be attributed to the limited pore-size of ZSM-5 that reduces the ability to accommodate bulkier isomers of higher carbon number products.

(g) Furthermore, the learning efficiency of LS-SVM based methods were higher than PLSR, tested for the case of clustered inputs. It implied that a higher number of calibration samples could improve the prediction accuracy of the models further. It should also be pointed out that the other shortcoming of this work (apart from limited number of samples) is the unavailability of a separate dataset (real time data) to test the applicability of the developed models on a different experimental setup.

To conclude, LS-SVM-RBF calibration was proved to be a versatile and efficient approach that can satisfactorily be used to monitor the acid-catalyzed propylene conversion process online.

3.7 References

(1) Ipatieff, V.; Corson, B. B.; Egloff, G. Polymerization, a New Source of Gasoline. *Ind. Eng.*

Chem. Res. **1935**, *27*, 1077–1081.

(2) Tabak, S. A.; Krambeck, F. J.; Garwood, W. E. Conversion of Propylene and Butylene over ZSM-5 Catalyst. *AIChE J.* **1986**, *32*, 1526–1531.

(3) Quann, R. J.; Green, L. A.; Tabak, S. A.; Krambeck, F. J. Chemistry of Olefin Oligomerization over ZSM-5 Catalyst. *Ind. Eng. Chem. Res.* **1988**, *27*, 565–570.

(4) Garwood, W. E. Conversion of C2-C10 to Higher Olefins over Synthetic Zeolite ZSM-5. In *Intrazeolite Chemistry*; American Chemical Society: Washington, D. C., 1983; pp 383–396.

(5) Pines, H. Acid-Catalyzed Reactions. In *The Chemistry of Catalytic Hydrocarbon Conversions*; Academic Press: New York, 1981; pp 1–122.

(6) Nicholas, C. P. Applications of Light Olefin Oligomerization to the Production of Fuels and Chemicals. *Appl. Catal. A Gen.* **2017**, *543*, 82–97.

(7) Osborne, B. G. Near-Infrared Spectroscopy in Food Analysis. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: Chichester, UK, 2000.

(8) Igne, B.; Ciurczak, E. W. *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy*; CRC Press: Boca Raton, 2014.

(9) Ertaş, M.; Hakki Alma, M. Pyrolysis of Laurel (Laurus Nobilis L.) Extraction Residues in a Fixed-Bed Reactor: Characterization of Bio-Oil and Bio-Char. *J. Anal. Appl. Pyrolysis* **2010**, *88*, 22–29.

(10) Guo, X.; Wang, S.; Guo, Z.; Liu, Q.; Luo, Z.; Cen, K. Pyrolysis Characteristics of Bio-Oil Fractions Separated by Molecular Distillation. *Appl. Energy* **2010**, *87*, 2892–2898.

(11) Owen, T. *Fundamentals of UV-Visible Spectroscopy: A Primer*; Hewlett Packard, 1996.

(12) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds.*; John Wiley & Sons: New York, 2014.

(13) Mathiesen, M. D.; Lubeck, A. J. Improving Accuracy in the Determination of Aromatics in Gasoline by Gas Chromatography-Mass Spectrometry. *J. Chromatogr. Sci.* **1998**, *36*, 449–456.

(14) van den Berg, J. P.; Wolthuizen, J. P.; van Hooff, J. H. C. Reaction of Small Olefins on Zeolite H-ZSM-5. A Thermogravimetric Study at Low and Intermediate Temperatures. *J. Catal.* **1983**, *80*, 139–144.

(15) Martens, H.; Næs, T. Multivariate Calibration. In *Chemometrics*; Springer Netherlands:

Dordrecht, 1984; pp 147–156.

(16) Sjöström, M.; Wold, S.; Lindberg, W.; Persson, J.-Å.; Martens, H. A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables. *Anal. Chim. Acta* **1983**, *150*, 61–70.

(17) Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

(18) Frank, I. E. A Nonlinear PLS Model. *Chemom. Intell. Lab. Syst.* **1990**, *8*, 109–119.

(19) Wold, S. Nonlinear Partial Least Squares Modelling II. Spline Inner Relation. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 71–84.

(20) Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. Interval Partial Least-Squares Regression (IPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419.

(21) Haykin, S. S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall: Upper Saddle River, N.J., 1998.

(22) Cortes, C.; Vapnik, V. Support Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.

(23) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, 2000.

(24) Balabin, R. M.; Lomakina, E. I. Support Vector Machine Regression (SVR/LS-SVM)—an Alternative to Neural Networks (ANN) for Analytical Chemistry? Comparison of Nonlinear Methods on near Infrared (NIR) Spectroscopy Data. *Analyst* **2011**, *136*, 1703.

(25) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 2000.

(26) Thissen, U.; Pepers, M.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. Comparing Support Vector Machines to PLS for Spectral Regression Applications. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 169–179.

(27) Gunn, S. R. *Support Vector Machines for Classification and Regression*; School of Electronics and Computer Science, University of Southampton, 1998.

(28) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: River Edge, NJ, 2002.

(29) Cui, C.; Fearn, T. Comparison of Partial Least Squares Regression, Least Squares Support Vector Machines, and Gaussian Process Regression for a near Infrared Calibration. *J. Near Infrared Spectrosc.* **2017**, *25*, 5–14.

(30) Niazi, A.; Goodarzi, M.; Yazdanipour, A. A Comparative Study between Least-Squares Support Vector Machines and Partial Least Squares in Simultaneous Spectrophotometric Determination of Cypermethrin, Permethrin and Tetramethrin. *J. Braz. Chem. Soc.* **2008**, *19*, 536–542.

(31) Chauchard, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V. Application of LS-SVM to Non-Linear Phenomena in NIR Spectroscopy: Development of a Robust and Portable Sensor for Acidity Prediction in Grapes. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 141–150.

(32) Shao, Y.; He, Y. Visible/Near Infrared Spectroscopy and Chemometrics for the Prediction of Trace Element (Fe and Zn) Levels in Rice Leaf. *Sensors* **2013**, *13*, 1872–1883.

(33) Gibson, N. P.; Aigrain, S.; Roberts, S.; Evans, T. M.; Osborne, M.; Pont, F. A Gaussian Process Framework for Modelling Instrumental Systematics: Application to Transmission Spectroscopy. *Mon. Not. R. Astron. Soc.* **2012**, *419*, 2683–2694.

(34) Tefera, D. T.; Yañez Jaramillo, L. M.; Ranjan, R.; Li, C.; De Klerk, A.; Prasad, V. A Bayesian Learning Approach to Modeling Pseudoreaction Networks for Complex Reacting Systems: Application to the Mild Visbreaking of Bitumen. *Ind. Eng. Chem. Res.* **2017**, *56*, 1961–1970.

(35) Tefera, D. T.; de Klerk, A.; Prasad, V. A Bayesian Learning and Data Mining Approach to Reaction System Identification: Application to Biomass Conversion. In *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*; IEEE, 2017; pp 178–183.

(36) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.*; MIT Press: Cambridge, MA, 2002.

(37) Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, 1990.

(38) Yeganeh, B.; Motlagh, M. S. P.; Rashidi, Y.; Kamalan, H. Prediction of CO Concentrations Based on a Hybrid Partial Least Square and Support Vector Machine Model. *Atmos. Environ.* **2012**, *55*, 357–365.

(39) Ge, Z.; Song, Z.; Ding, S. X.; Huang, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access* **2017**, *5*, 20590–20616.

(40) Dobos, L.; Abonyi, J. On-Line Detection of Homogeneous Operation Ranges by Dynamic Principal Component Analysis Based Time-Series Segmentation. *Chem. Eng. Sci.* **2012**, *75*, 96–105.

(41) Lavine, B. K.; Nuguru, K.; Mirjankar, N.; Workman, J. Development of Carboxylic Acid Search Prefilters for Spectral Library Matching. *Microchem. J.* **2012**, *103*, 21–36.

(42) Lavine, B. K.; Davidson, C. E.; Moores, A. J. Genetic Algorithms for Spectral Pattern Recognition. *Vib. Spectrosc.* **2002**, *28*, 83–95.

(43) Smith, G. L.; Mlynczak, P. E.; Potter, G. L. A Technique Using Principal Component Analysis to Compare Seasonal Cycles of Earth Radiation from CERES and Model Computations. *J. Geophys. Res. Atmos.* **2012**, *117*.

(44) Jolliffe, I. T. Mathematical and Statistical Properties of Population Principal Components. In *Principal Component Analysis*; Springer Series in Statistics; Springer-Verlag: New York, 2002.

(45) Zimmermann, B. Characterization of Pollen by Vibrational Spectroscopy. *Appl. Spectrosc.* **2010**, *64*, 1364–1373.

(46) Mok, P. Y.; Huang, H. Q.; Kwok, Y. L.; Au, J. S. A Robust Adaptive Clustering Analysis Method for Automatic Identification of Clusters. *Pattern Recognit.* **2012**, *45*, 3017–3033.

(47) Fiser, K.; Sieger, T.; Schumich, A.; Irving, J.; Dworzak, M. N.; Vormoor, J. MRD Monitoring of Childhood ALL Using Hierarchical Clustering and Support Vector Machine Learning of Complex Multi-Parameter Flow Cytometry Data. *Blood* **2008**, *112*, 1508.

(48) Fan, C. Y.; Fan, P. S.; Chan, T. Y.; Chang, S. H. Using Hybrid Data Mining and Machine Learning Clustering Analysis to Predict the Turnover Rate for Technology Professionals. *Expert Syst. Appl.* **2012**, *39*, 8844–8851.

(49) Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An Efficient K-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892.

(50) Balabin, R. M.; Safieva, R. Z.; Lomakina, E. I. Gasoline Classification Using near Infrared (NIR) Spectroscopy Data: Comparison of Multivariate Techniques. *Anal. Chim. Acta* **2010**, *671*, 27–35.

(51) Fernández Pierna, J. A.; Baeten, V.; Renier, A. M.; Cogdill, R. P.; Dardenne, P. Combination of Support Vector Machines (SVM) and near-Infrared (NIR) Imaging

Spectroscopy for the Detection of Meat and Bone Meal (MBM) in Compound Feeds. *J. Chemom.* **2004**, *18*, 341–349.

(52) Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. LS-SVMlab Toolbox User's Guide. *Pattern Recognit. Lett.* **2003**, *24*, 659–675.

(53) Norgaard, L.; Leardi, R. *IToolbox Manual*; FOSS, Data & Algorithms in R&D, 2005.

(54) Press, W. H.; Teukolsky, S. A. Savitzky-Golay Smoothing Filters. *Comput. Phys.* **1990**, *4*, 669–672.

(55) Colthup, N.; Daly, L.; Wiberley, S. *Introduction to Infrared and Raman Spectroscopy*, 2nd Edn.; Academic Press: Boston, 1990.

(56) Wahba, G. Numerical Methods. In *Spline Models for Observational Data*; Society for Industrial and Applied Mathematics: Philadelphia, Pennsylvania, 1990; pp 135–143.

(57) Lazraq, A.; Cléroux, R.; Gauchi, J. P. Selecting Both Latent and Explanatory Variables in the PLS1 Regression Model. *Chemom. Intell. Lab. Syst.* **2003**, *66*, 117–126.

(58) Snee, R. D. Validation of Regression Models: Methods and Examples. *Technometrics* **1977**, *19*, 415–428.

(59) Xavier-De-Souza, S.; Suykens, J. A. K.; Vandewalle, J.; Bolle, D. Coupled Simulated Annealing. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **2010**, *40*, 320–335.

(60) Thissen, U.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. Multivariate Calibration with Least-Squares Support Vector Machines. *Anal. Chem.* **2004**, *76*, 3099–3105.

(61) Den Hollander, M. A.; Wissink, M.; Makkee, M.; Moulijn, J. A. Gasoline Conversion: Reactivity towards Cracking with Equilibrated FCC and ZSM-5 Catalysts. *Appl. Catal. A Gen.* **2002**, *223*, 85–102.

(62) Stasinopoulos, M. D.; Rigby, R. A.; Heller, G. Z.; Voudouris, V.; De Bastiani, F. *Flexible Regression and Smoothing*; CRC Press: Boca Raton, FL, 2017.

(63) Krier, C.; Rossi, F.; François, D.; Verleysen, M. A Data-Driven Functional Projection Approach for the Selection of Feature Ranges in Spectra with ICA or Cluster Analysis. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 43–53.

4. A statistical approach dealing with multicollinearity among predictors in microfluidic reactor operation to control liquid-phase oxidation selectivity [3]

ABSTRACT

The current study concerns with applying chemometrics involving regression techniques to identify a single most important parameter that directly affects oxygen availability and has a key influence on tetralin conversion (CR) and product selectivity (S) for tetralin oxidation in a microfluidic reactor at constant temperature and pressure. Five parameters (predictors) identified previously were gas-liquid interfacial area (a), length of oxygen gas bubble (L_G), length of liquid slug (L_S), two-phase superficial velocity (U_{TP}) and liquid flowrate to the reactor (Q), where 'a' was suspected to be directly related to oxygen availability. CR and S were regressed on all the predictors by fitting separate simple linear regression (SLR) models. The decreasing order of explained variance in the outputs based on the calibration model was: In CR: $a^2 > L_G > U_{TP}^3 > L_S > Q$; In S: $a > L_G > U_{TP}^2 > Q$. Multicollinearity issues among predictors were detected through Pearson's correlation coefficients and diagnostics like variance inflation factors (VIF) and eigenvalues of the correlation matrix. This was addressed through multiple linear regression (MLR) by considering a second input in addition to the best predictor from the SLR (a). Drastic changes in regression coefficient estimates and inflated standard errors rendered the coefficients of all other variables (except 'a') insignificant in the MLR models. The incremental contribution of 'a' towards improving output variance was also evaluated through F-tests and partial correlations with the outputs, controlling for other variables as well. The applications of the findings from this study in industrial reactor design (for example – loop reactors) to control product selectivity were also highlighted. In addition, through chemometrics, the possibility of monitoring reaction progress online by predicting reactant conversion and product selectivity, thereby eliminating the need for offline gas chromatographic (GC) measurements was also examined.

Keywords: Oxygen availability; simple and multiple linear regression; multicollinearity; relative importance of parameters; significance tests and analysis of variance; online monitoring

[3] This work was published as 'Siddiquee, M. N.; Sivaramakrishnan, K.; Wu, Y.; De Klerk, A.; Nazemifard, N. A statistical approach dealing with multicollinearity among predictors in microfluidic reactor operation to control liquid-phase oxidation selectivity. *Reaction Chemistry & Engineering* **2018**, 3, 972 – 990'.

4.1 Introduction

Liquid phase oxidation of hydrocarbons is industrially important to produce petrochemicals. [1–3] The main challenge of the non-catalytic free radical oxidation is to achieve good product selectivity. Industrially the non-catalytic liquid phase oxidation is performed at low conversion to control the product selectivity, for example, oxidation of cyclohexane. [4] Oxidation product selectivity depends on conversion, temperature and oxygen availability in the liquid phase. [5–8]

Liquid phase oxidation follows initiation, propagation and termination steps. Once the free radical (R·) is formed during the initiation step, it reacts with local oxygen or other free radicals very fast following zero order kinetics. [5,9] Oxygen transfer to the liquid phase and oxygen availability in the liquid phase are critically important in order to control the product selectivity. Microfluidic reactor, also known as microreactor, is advantageous to ensure higher local oxygen availability. The main advantages of such a miniaturized reactor are: (i) the higher surface-area-to volume ratio that facilitate the improved mass and heat transfer in the liquid phase; (ii) the exact control of the gas to liquid ratio in the reactor that facilitate the manipulation of gas-liquid interfacial area; and (iii) the well-defined flow properties in the microstructure reactor. [5,10,11] These advantages combined with other benefits such as small radial diffusion length leading to good mixing and enhanced safety for oxygen use caused the microfluidic reactor to receive attention in the study of liquid phase oxidation. [10,12–18]

The flow regime in a microchannel depends on the relative gas and liquid properties, flow rates and channel geometry. The five main flow regimes are: bubble, slug, churn, slug annular and annular. [5,19,20] Slug flow, also known as Taylor flow, has its unique hydrodynamic characteristics, where two adjacent liquid slugs are separated by the gas bubbles and are connected only via a thin liquid film. [5,10,11,19,20] This thin liquid film contributes to create a higher gas-liquid interfacial area and hence improves oxygen availability. Taylor flow also has a Marangoni effect within the liquid slug that is beneficial not only to ensure local oxygen availability by proper mixing but also to bring the surface active oxygenates to the liquid phase to prevent over oxidation at the gas-liquid interface. [5,10,11]

Of the parameters that influence oxygen availability, it was not clear which ones affect it the most in a hydro-dynamically complex Taylor flow system. Oxygen availability at a constant pressure depended on several parameters such as gas-liquid interfacial area (a) based on unit cell

volume (volume of gas bubble and liquid slug), the film attached to the wall, length of liquid slug (L_S), length of gas bubble (L_G), two-phase superficial velocity (U_{TP}), and liquid flowrate to the reactor (Q). [5] Small changes in the design of the microfluidic reactor and its operation could dramatically affect the relationship between these different parameters.

Mass transfer in Taylor flow can be explained well by the film theory. [5,10] The gaseous component, oxygen in our case, is transferred to the liquid phase where it is consumed during the reaction. Works on mass transport at Taylor flow conditions have been well documented in literature, but the most cases focused on simulation and/or experiments considering water as the liquid at ideal conditions (no gas consumption). [11,19–24] A few liquid phase oxidation studies at Taylor flow conditions are reported which dealt mainly with conversion enhancement, but mass transfer characteristics were not discussed. [13,14,16–18] In our previous oxidation study, a qualitative description of the mass transport effects on oxidative conversion and product selectivity was provided. [5] However, the quantitative dependency of the parameters affecting the mass transport, oxidative conversion and product selectivity were not deeply analyzed.

Regression is a fast and a cost-effective tool for analyzing the complex dependencies between the different variables existing in the system. One of the main advantages of using regression methods is that it can alleviate the requirement of analytical equipment like gas chromatography (GC) which is used to conduct offline measurements on the reaction products to obtain the conversion and selectivity. Instead, predicting these outputs for new inputs based on developed calibration models on previously acquired empirical data can save time for processes on an industrial scale.

Simple linear regression (SLR) is a very common chemometric method used to examine relationships between a continuous quantitative outcome and a quantitative explanatory (input) variable. [25] It enables an analyst to fathom the effect of each explanatory variable one at a time on the output by looking at the magnitude of the regression coefficient estimates and the overall variance explained by the model. In most real-life situations, the outcome is influenced by more than one variable, in which case multiple linear regression (MLR) suits best as it considers all the input variables simultaneously. [26] The solutions of both SLR and MLR are based on ordinary least squares (OLS) principle where sum of squares of the residuals from the regression equation are minimized to obtain the intercept and coefficient estimate for each explanatory variable. [27]

However, one critical assumption in linear regression is that the errors are required to be normally distributed so that the OLS is equivalent to the maximum likelihood estimate. [28]

An unwanted obstacle to linear regression is the presence of a high degree of linear dependencies between the input variables, termed as multicollinearity. This is relevant to this study because some variables have been calculated from others and considered in the regression model. Particularly, it makes the parameter estimates and the standard errors of the correlated variables unstable by degrading their significance, thus rendering the regression model unreliable. [29] If two variables are correlated among themselves as well as with the desired outcome variable, there always exists a portion of shared variance in the output variable that is common to both input variables. Disentangling the variance purely explained by each variable becomes difficult and this concept can be understood well by means of Venn diagrams. [30]

The objective of this work was to determine the most important hydrodynamic/mass-transfer parameters and the interactions among them that control conversion rate and product selectivity in the primary oxidation products through a chemometric approach. For this purpose, the autoxidation of tetralin was selected as a test reaction. SLR models were implemented to detect the potential predictors with their best fits (linear regression implies linearity only in the coefficient estimates and not necessarily between the variables involved) based on certain model performance statistic measures. The additional contribution of a second variable was investigated through MLR models and inspecting for drastic changes in the coefficient estimates, their standard errors and significance arising due to collinearity by comparing with the corresponding SLR models. Statistical tests, namely t-test and F-test were employed to determine the significances of the regression coefficients and overall models, respectively.

A positive aspect of multicollinearity is that the coefficient estimates remain unbiased (as with the OLS method) as this was crucial to establish relative importance of the explanatory variables in impacting the outputs. Though multicollinearity among input variables inflates the variance of the coefficient estimates and causes the estimates to move away from the true population value (unknown most of the times), the expected value of the estimates across random samples is not altered. [31–33] But the changes in the relatively important variable would be minimal and thus, recognized. Since the coefficient estimates are unbiased, the statistical inferences would still be meaningful. However, the variance in the estimates can also depend on the sign of the correlations between the explanatory variables as shown by Mela and Kopalle, [34] where they indicated that

multicollinearity was not always detrimental (detailed further in section 4.4 of this chapter). Thus, by the use of conventional statistical techniques, the reaction parameters could be analyzed for inter-dependencies with the possibility of simultaneous monitoring of progress of the reaction.

4.2 Experimental

4.2.1 Materials

The hydrocarbon that was oxidized to perform experimental calibration and validation of the regression model was tetralin (99 % purity, Sigma-Aldrich). Hexachlorobenzene (99 % purity, Supleco) was used as internal standard and chloroform (98 % purity, Fischer Scientific) was used as a solvent for gas chromatography. Extra dry oxygen (O_2 99.6 % molar purity) was used as an oxidizing agent and nitrogen (N_2 99.999 % molar purity) was used as inert to control back-pressure in the reactor. Praxair supplied both cylinder gases. In addition to the materials used for the experimental work, some authentic compounds were used to confirm key oxidation products, namely, 1,2,3,4-tetrahydro-1-naphthol (alcohol of tetralin) and α-tetralone (ketone of tetralin). These were used as described before. [35]

4.2.2 Equipment and Procedure

A glass rectangular microfluidic reactor (Dolomite Microfluidics, Charlestown, MA, USA) was used in this study to conduct the oxidation studies. The reactor volume was 1000 µl having a mixing channel of depth = 1240 µm, width = 161 µm, length = 536 mm and a reaction channel of depth = 1240 µm, width = 391 µm, length = 1844 mm. It had three inlet ports and one outlet port. Oxygen and tetralin were injected into the reactor using fluid input port 1 and fluid input port 2, respectively, and port 3 was blocked. The hydraulic diameter of the reaction channel was d_H = 6.0×10^{-4} m and aspect ratio of the reaction channel of width/depth was 0.32. Details on the individual components are given in section B.1 of Appendix B.

Oxidations were performed under Taylor flow conditions at 150 ºC and an average pressure of 90 kPa gauge to control product selectivity. In Taylor flow conditions, tetralin liquid slugs were

separated by elongated oxygen gas bubbles. The detailed experimental procedure of oxidation in microfluidic reactor was provided in our previous study. [5]

Briefly, in a typical experiment, tetralin was loaded into a 5 mL syringe and the system was pressurized to 90 kPa gauge by introducing oxygen into the setup. Tetralin was then allowed to flow through the system at a specific volumetric flow rate (1.5, 2, 4, 7 or 12 µL/min for new experiments in this work) by using a syringe pump. The co-feed of tetralin and oxygen and application of back-pressure using a back-pressure regulator and nitrogen gas facilitated the gas-liquid slug formation. A digital camera mounted above the microfluidic reactor was used to monitor the flow patterns of the gas and liquid during the experiment. The experiments were conducted for twenty minutes. The system was then depressurized and the oxidized tetralin was collected from the pressure vessel at the outlet of the reactor for analysis by gas chromatography.

4.2.3 Analyses

Chromatographic analyses were performed using a gas chromatograph equipped with a flame ionization detector (GC-FID). The separation, response factor corrections used for analyses and method of product identification were reported before. [5] A summary of the analytical details is presented in section B.2 of Appendix B.

4.2.4 Calculations

The parameters that potentially influenced oxidation conversion and product selectivity and are shown as explanatory/input variables in Table 4.1. Some of them could be obtained directly from the images captured during experiments. For deriving the other variables, the same image analysis protocol and calculations that was used in previous work [5] was employed for the validation set of experiments in this study (Table 4.4) and details can be found in section B.3 of Appendix B.

Table 4.1. Different independent and dependent parameters used in regression analysis.

Variable	Type	Symbol	Unit
Tetralin flow rate to the reactor	Explanatory (input)	Q	µL/min.
Length of liquid slug	Explanatory (input)	L_S	m
Length of gas bubble	Explanatory (input)	L_G	m
Two-phase superficial velocity	Explanatory (input)	U_{TP}	m/s
Gas-liquid interfacial area	Explanatory (input)	a	m^2/m^3
Conversion rate	Dependent (output)	CR	mol/s
Ketone-to-alcohol Selectivity	Dependent (output)	S	-

4.3 Regression modelling methods

This section presents the statistical details of the modelling process conducted in this study. The different assumptions involved in constructing the regression models, the software employed to carry out the mathematical analyses, the methodical approach followed along with pertaining equations and certain model performance evaluation measures used in SLR models are given in this section. The relevant experimental data from our previous study [5] that served as the calibration set for building the regression models is also included in this segment.

4.3.1 Assumptions in regression model building

Simple linear regression (SLR) and multiple linear regression (MLR) models were constructed over the course of this work for quantifying the different relationships between the input/output parameters and identifying the best predictor for conversion and selectivity. Before proceeding with the regression analysis, the following set of assumptions were checked for validity of the model:

(i) **Linearity.** Figure B.1 in Appendix B shows the scatter plots for variation of CR and S with each of the explanatory variables. The effect of each input variable can be seen roughly through these plots and though there are linear regions in between, the data dependencies are mostly non-linear as expected. A nonlinear relationship was shown to be a better fit in

some cases (refer to section on SLR results and discussion) by always considering statistical measures (refer to section 4.3.4). Deviations from linearity is tolerable in real data since linear regression allows for transformation of the variables and is linear in the coefficient estimates only.

(ii) **Normality.** This is regarded as a classical regression assumption but it should be noted that a regression model is quite robust to its violation. [36] The errors or residuals from the regression analysis need to be normally distributed and one way to confirm this assumption is by looking at the residual plots for the regression models. This is elaborated in the section 4.4.4.4. The only shortcoming in this process was that there were limited points available to build the model (refer to section 4.3.5 for the calibration dataset (Table 4.2) but the model was also verified by prediction with a new, independent validation set (Table 4.4) for choosing the best fits and moving forward. At the end of the paper, a statistical pathway to predict the outputs directly from the user-controlled parameter (Q) is also derived where the predicted values are compared with those in the validation set.

(iii) **Equal spread.** This refers to the residuals having equal variance in the regression model and can be verified from the residual vs. predicted value plots as shown in Figure B.2 of Appendix B. Explanation of the plots is provided in the 'Residual plots' section under SLR analyses results (section 4.4.4.4). The other implication of this assumption is that the mean of the residuals must be close to zero. Any point that violates this condition can be termed as an outlier and its omission can significantly improve the fit of the regression. Omission of the intercept term forces the line to pass through 0, in which case there can be a violation of this assumption. In our case, no point was omitted in the regression calculation and intercept term was included for further prediction.

(iv) **Minimum variation in the measurement of input variables (Fixed *X*).** The scatter plots do not reflect this property but it depends on the experimental way that the input variables were measured. The inputs in this analysis are taken from the previous study [5] and it can be asserted with confidence that they were measured and calculated with minimal error, thus confirming the validity of this assumption. Each experimental sample was done in triplicate and the reported values were the mean of the observed measurements.

(v) **Independent error.** Each of the five experimental data points in the calibration and validation sets were not dependent on each other and were conducted as separate

experiments by manipulating the flow rate each time and conducting GC analysis to obtain the outputs. For example, the conversion rate and selectivity at 3 μL/min was not predicted from the outcome at a previous flowrate (1 μL/min) and was calculated separately (see 'Calculations' section B.3 in Appendix B). This confirms that the individual errors are independent for each experiment.

4.3.2 Software used

IBM SPSS Statistics (Version 25) was utilized to perform all statistical analyses in this work. The software was run on MacOS (Version 10.13.3).

4.3.3 Regression calculations: Relevant equations and formulae

It is important to note that all correlations and regression models in this work are based on sample data, which can be approximated for the entire population since the population data is not known. Here, the population refers to the set of all possible tetralin oxidation studies conducted at constant temperature (150 ºC) and pressure (90 kPa gauge), but at a larger number of flowrates, simultaneously maintaining Taylor flow conditions in the microfluidic reactor. The parameter estimates obtained based on the experimental data in the previous [5] and from the current study can be considered an approximation for the population data since the true population parameters are unknown. [37] The following statistical parameters are used in this work:

(i) **Pearson's product-moment correlation coefficient.** As a preliminary step, inter-relationships between the various variables were inspected by means of sample Pearson's correlations (r). [38] The mathematical expression for this measure is given in equation B.15 in Appendix B. Always pairwise bivariate correlations are calculated and interpreted in this study due to the availability of limited number of observations for model construction and prediction. The correlation coefficients were investigated for two reasons: (1) To gain knowledge regarding the presence of multicollinearity among the variables for guarding against redundancy in output estimation; (2) To know which input parameters have a higher and significant linear correlation with CR and S to provide a basis for building further regression models. Though it provided only a quantitative estimate of the strength of the

linear relationship between the variables, other non-linear fits with 2 regression coefficients (intercept and the non-linear term) were also evaluated to obtain a comprehensive regression model.

(ii) **Significance of a correlation.** Since the sample size in our work is small (n = 5), it was also imperative to check the representativeness of this sample data to model the relationships in the whole population, i.e. testing the significance of the correlation by means of a t-test. It should be noted that all tests of significance are two-tailed in this study to cope with positive as well as negative values of the test statistic. By means of P-value plots (P-P plots), the variables were checked for the level of adherence to a bivariate normal distribution. [39] It is a means of descriptive statistics that consists of a scatter plot between the cumulative distribution functions of the observed function and the expected function (normal distribution). Since the number of data points were limited (5 samples in calibration set), some amount of deviation from a straight line passing through the origin in the P-P plots was acceptable. P-P plots suited better than Quantile-Quantile (Q-Q) plots for our study because our focus was more on seeing the agreement between the distributions and not on estimating location and scale parameters from the plot (for which Q-Q plot was useful). A quantile-normal plot (Q-N) of the regression residuals can also be used similarly to detect normality of the errors. [39]

The details of hypothesis testing and the formula for calculation of the t-statistic (following a t-distribution with 3 degrees of freedom (DF)) is given in section B.5 of Appendix B. The decision rule used for the hypothesis testing based on the p-value is that if it falls below 0.05 (5%), the null hypothesis can be rejected because there is less than 5% chance that the estimate of population correlation coefficient is near to 0. In other words, the correlation can be stated to be significant at the 95% confidence level if it is > 0.878 or < -0.878 (critical value from Pearson's Correlation Table, calculated in turn from the t-table). [40]

(iii) **SLR and MLR.** The main equations for simple and multiple linear regression analyses, the standard errors and the corresponding t-statistic for evaluating the significance of the regression coefficient estimates are provided in section B.5 of Appendix B. The explained variance in the output by all the predictors simultaneously is calculated through coefficient of determination (R^2 for MLR and r^2 for SLR – refer section 4.3.4 in Chapter 4).

It can be conceptualized that the standard errors of the regression coefficient estimates (equation B.19 and equation B.20 in Appendix B) increase if the errors of predicting **y** increase or if the correlation between the predictors increase or even if the variance of the predictors decrease. Palpably, $r_{X_1-X_2}$ becomes 0 in SLR. The t-statistic for the slope estimates (equation B.22 in Appendix B) is compared with the t-distribution with $n-k-1$ degrees of freedom (3 for SLR and 2 for MLR in this study) to find the range of critical values between which this statistic value lies. The exact p-value for this statistic is given by SPSS. Similarly, all regression coefficient estimates (standardized and unstandardized), their standard errors, t-statistics and the standard error of prediction reported in this study are given by SPSS. They were also checked by manual calculations using the respective equations for consistency.

(iv) **F-statistic.** This is a statistic for testing the significance of analysis of variance (ANOVA) or the overall model R^2. The exact p-value is obtained from SPSS. The value of the statistic can be computed from equation B.23 in Appendix B and verified against that obtained in SPSS. It is interesting to note that in SLR, the p-value for the F-statistic is equal to the p-value of the regression coefficient of the explanatory variable.

A second type of F-statistic is used to test the incremental contribution of an added predictor in improving the overall model r^2. In this study, one variable was added to the already developed SLR to test for the incremental importance of that predictor in explaining the output variance. This statistic is calculated manually (using equation B.26 in Appendix B) and compared with the critical value of F (F_{crit}) from the F-table [41] corresponding to (1, 2) DF in all cases. If the obtained F value is greater than F_{crit} we can say that the contribution of the second variable in improving the overall model r^2 of the SLR is significant.

4.3.4 Model performance evaluation

Once the regression coefficient estimates for the explanatory variables were calculated based on the calibration set data, they were used to predict the outcomes in the validation set. The performance of the models was evaluated using two statistical parameters, root mean squared error

(RMSE) and coefficient of determination (R^2), given by equation B.27 and equation B.28 in Appendix B.

Two types of RMSE expressions are used in this work: RMSEP (error of prediction in the validation set) and RMSEC (error of calibration in the calibration set). Likewise, R_P^2 (r_P^2) and R_C^2 (r_C^2) refer to explained variance in the outputs for the validation set and the calibration set of the MLR (SLR) models, respectively. The primary criteria for choosing the best fits in the linear regression models (section 4.4.4) are lower RMSEP and higher R_P^2 values. Ideally, for a perfectly optimal model, RMSEP should equal RMSEC. However, in reality, RMSEP tends to be greater than RMSEC in most cases and in this work, a smaller absolute difference between RMSEP and RMSEC values is also considered in choosing the best fits. A larger gap between RMSEP and RMSEC implies that the model doesn't predict accurately for new and independent datasets and is suggestive of overfitting and too much flexibility.

4.3.5 Data for model building/calibration

The purpose of this investigation was to evaluate the impact of hydrodynamic parameters on oxidation conversion and selectivity and identify their plausible best predictors. Therefore, only those data from Siddiquee et al. [5] were used where temperature and pressure were controlled and not varied. As indicated in the 'Experimental' section, tetralin was oxidized isothermally at 150 ºC and 90 kPa gauge at different flow conditions in a microfluidic reactor. Hydrodynamic and mass transfer data observed and collected from five different injection flowrates of tetralin were used for constructing the regression models in this study. These data are provided in Table 4.2 and served as the calibration set for this work.

From a statistical point of view, it was important to primarily investigate whether oxygen availability in particular had a superior effect on the outcomes as hinted in the previous work. [5] t_R was not considered in model building because it was indirectly a representation of U_{TP} (as given in section B.3 in Appendix B). Once the regression models were built based on this calibration set, an independent validation data set (Table 4.4) that was collected at flowrates different to the calibration ones) was used to predict the outputs from the SLR models to determine best fits and to establish a prediction pathway from Q to CR and S. In all experiments total conversion at the

outlet of the reactor was low (< 1 %), so that it was possible to consider conversion rate and selectivity as two independent outputs that were not affected by the total conversion.

Table 4.2. Experimental data used for construction of model building. [a]

Exp ID	T (ºC)	Q [b] (µL/ min)	L_S (m)	L_G (m)	U_{TP} (m/s)	t_R (min)	a (m^2/m^3)	k_La (s^{-1})	CR (mol/s)	S [c]
A	150	1	.0016	.2100	.0260	1.5	300000	1900	5.95×10^{-07}	14.0
B	150	3	.0027	.0980	.0210	1.8	150000	960	1.40×10^{-07}	7.00
C	150	5	.0049	.0510	.0200	2	62000	390	1.03×10^{-07}	1.60
D	150	10	.0043	.0110	.0160	2.4	16000	100	6.00×10^{-08}	1.30
E	150	15	.0049	.0040	.0110	3.5	5400	34	5.14×10^{-08}	1.30

[a] Data obtained from our previous study [5]; [b] Inlet tetralin flowrate into the reactor; [c] Ketone-to-alcohol selectivity in primary oxidation products.

4.4 Results and Discussion

4.4.1 Bivariate correlations among input and output variables

Ideally, it would be expected that some independent variables (IV) have higher correlation with the output (**y**) so that the proportion of the variability accounted for in y_i is distinct for each $x_i \in$ **X**. With real data, this condition doesn't occur in most of the cases and there is some amount of overlap of the explained variance in the output by the IVs, arising due to the presence of collinearity among them. Table 4.3 reflects this phenomenon. The P-P plots for the explanatory variables, a and U_{TP} and outcome variables, CR and S are given in Figure 4.1.

Table 4.3. Bivariate Pearson correlation coefficients between different parameters considered in this study.

Variable	CR	S	Q	a	L_G	U_{TP}	L_S
CR	1	0.947 (0.014) [a]	-0.683 (0.203)	0.945 (0.015)	0.954 (0.012)	0.803 (0.102)	-0.853 (0.066)
S		1	-0.759 (0.137)	0.988 (0.002)	0.981 (0.003)	0.832 (0.080)	-0.966 (0.007)
Q			1	-0.839 (0.076)	-0.843 (0.073)	**-0.980 (0.003)** [b]	0.766 (0.131)
a				1	**0.998 (0.0001)**	**0.900 (0.037)**	**-0.945 (0.016)**
L_G					1	**0.909 (0.033)**	**-0.925 (0.024)**
U_{TP}						1	-0.813 (0.094)
L_S							1

[a] The number in brackets indicates the respective p-values; [b] Bold values indicate correlations > 0.9 between explanatory variables.

Although the P-P plots for each of the variables shown in Figure 4.1 do not indicate a straight line ((b) was the closest to the straight line $y = x$) passing through the points of the respective cumulative distribution functions, the closeness of the distributions was determined acceptable envisaging a five-point system. This implied that the application of t-tests are still valid. [28]

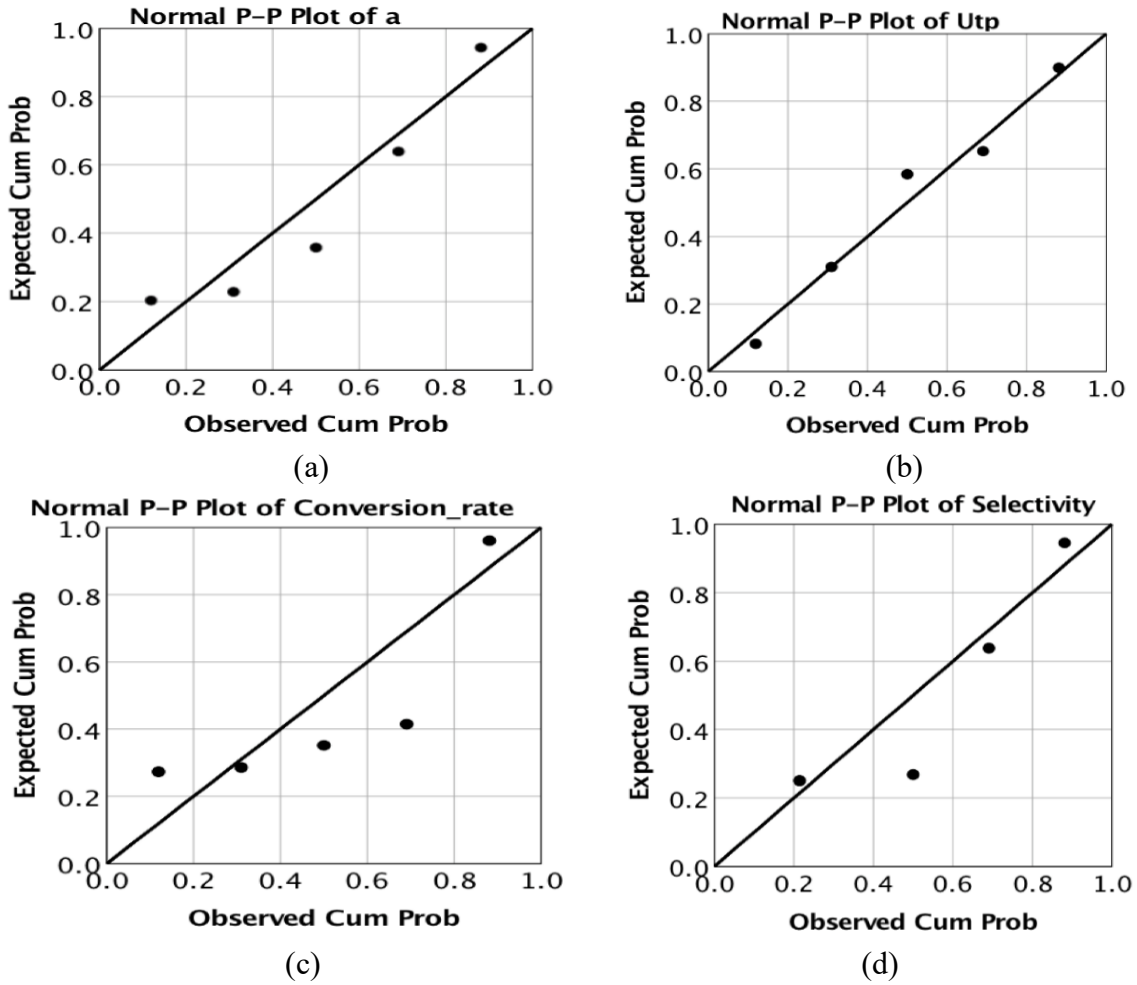


Figure 4.1. P-value plots shown for 2 explanatory variables: (a) gas-liquid interfacial area; (b) two-phase liquid velocity; and the 2 outcome variables: (c) tetralin conversion rate; (d) ketone-alcohol selectivity.

Let us consider the outcome-predictor variable relationships first. Conversion rate (CR) of tetralin was found to have a strong linear relationship with L_G (r_{CR-L_G} = 0.954) and a (r_{CR-a} = 0.945) with p-values of 0.012 and 0.015 respectively (Table 4.3), entailing that the relation is significant at the 5% level. A consensus has not yet been arrived in the literature regarding a cut-off value that describes the strength of a linear relationship between 2 variables. Values of 0.9, [42] 0.8, [43] 0.7 [44] and even as low as 0.35 [45] have been suggested as threshold levels for a bivariate correlation to be considered strong enough to cause collinearity effects in regression results.

However, as mentioned previously, the critical value of 0.878 can be used as an indicator for a significant relationship. The direction of the correlation between CR with U_{TP} was positive as

was with a and L_G but a p-value of 0.102 labels it insignificant. Negative correlations were observed with other input variables like L_S and Q, with r_{CR-L_S} being more significant than r_{CR-Q} but still unsubstantial at the 5% level. Similarly, a and L_G exhibited very strong correlations with the other output variable (Table 4.3), selectivity with $p < 0.01$ for both the explanatory variables. Interestingly, the relationship between S and L_S came out to be the third strongest with $r_{S-L_S}(3) = -0.966$, $p = 0.007$, while U_{TP} and Q were not found to be significantly related to the product selectivity, though they had slightly higher correlation coefficients than with CR.

Having inspected the correlations between the output and input variables alone, it could lead us to conclude that a and L_G are potentially good predictors for both CR and S, with L_S as another possible predictor for S. But the correlation matrix given in Table 4.3 also indicated that the input variables are significantly correlated among each other (correlations ≥ 0.9 are specified in bold). Absolute values of r_{a-L_G}, $r_{a-U_{TP}}$ and r_{a-L_S} are all greater than or equal to 0.900 and p-value corresponding to r_{a-L_G} stooped as low as 0.0001.

The high correlations between the explanatory variables was a first sign of the existence of elevated levels of multicollinearity in the data [46] and had to be dealt with appropriately with further diagnosis and rectification. Reasons for some of these increased inter-relationships among the input variables can be traced back to the 'Calculations' section (section 4.2.4).

For example, a is indirectly calculated as the ratio of linear combinations of L_G and L_S (equations B.1 – B.4 in Appendix B) and is the origin of the high correlation between them. U_{TP} depends on the volume fraction of the gas bubble (equation B.6 in Appendix B), which in turn is related to the ratio of volumes of the oxygen bubble and liquid slug, i.e. ratio of linear combination of L_G in the numerator and L_S in the denominator for ε_G (equation B.7 in Appendix B), to which a is also related. L_G and L_S have high correlation plausibly due to the Taylor flow conditions in the microreactor. Also, $r_{Q-U_{TP}}$ (DF = 3) = -0.980, p = 0.003 can be understood in terms of the tetralin flowrate negatively impacting the superficial velocity since Q increases the volume of liquid present in the reactor at a given time that can decrease the gas bubble volume fraction, thus lowering U_{TP}.

The other correlations not indicated in bold, though insignificant, tell us that increasing the only manipulated variable, Q decreases the parameter responsible for mass transfer, a as well as

the hydrodynamic parameters, L_G and U_{TP} but has a positive relation with L_S. Finally, Q is moderately correlated with the outcomes (CR and S) in the negative direction.

4.4.2 Impact of multicollinearity on regression analysis and its diagnostics: Path forward

Linear dependencies within the predictor variables are a definite cause for concern. Given the fact that our main aim was to identify the relative importance of each explanatory variable in predicting the conversion rate and selectivity, evaluating the effect of collinearity between the input variables on the reliability of the regression models was deemed essential. Multicollinearity can be defined as a statistical phenomenon where more than one predictor variable in a linear regression model is associated with the other variables and can lead to an ill-conditioned system. [47] Real data, as in this study, majorly falls in between a situation of perfect collinearity ($r_{x_1-x_2} = 1$ or -1 where x_1 and x_2 are two example explanatory variables) and that of no correlation ($r_{x_1-x_2} = 0$).

The different sources responsible for the origin of this problematic phenomenon can be majorly classified into two types: (i) sampling methodology of the data collected purely due to the experimental observations and (ii) extraction of new variables that are transformations of existing variables in the system. In the second type, using the transformed variables along with the existing variables for construction of regression models can raise issues in interpreting the model estimates.

In this study, Q is manipulated by the user and L_G and L_S are observed during the reaction, so can be labelled as existing variables. On the other hand, a and U_{TP} were calculated indirectly from the reactor geometry and the existing variables, which was the cause for multicollinearity in our system. It was imperative to ensure that this did not affect the conclusions and interpretations of the regression models employed in this study and thus treat the problem in an effective manner by considering the potential hazards of regressing with collinear variables and ways to tackle it. Multicollinearity can affect the regression model in the following ways:

a. Increase sensitivity of the magnitude of regression coefficient or slope estimates (corresponding to the correlated variables) to addition or removal of explanatory variables and slight changes in data point values
b. Cause reversal in signs of the regression coefficient estimates of the correlated variables involved in the regression model

c. Inflate the standard errors of the regression coefficient estimates resulting in unstable t-statistics and thus altering the significance of the predictors
d. Expansion in the width of confidence intervals of the intercept and predictor slope estimates

The disadvantage of multicollinearity relevant to our data is that it becomes quite strenuous in assessing as to which explanatory variable contributes most in explaining the share of variance in the response variable (CR and S) due to inaccurate predictor slope estimates and escalated standard errors as compared to the case of uncorrelated variables. In other words, it does not cause the coefficients to be biased, but renders them unstable. This was recognized as the inherent challenge in this work. As highlighted in the previous section, the first step in the detection of multicollinearity is through analysis of the pairwise product-moment Pearson correlations, which is necessary but not a sufficient condition. Interpreting the strength of a correlation can be subjective too. Hence, certain other diagnostics suggested in literature are utilized in this work to confirm the presence of interdependencies among the predictor variables and can be useful when proper conclusions cannot be made from the correlation matrix alone. These are given in section B.7 in Appendix B.

At this juncture, it is worthwhile to mention the study by Mela and Kopalle, [34] where the authors note that the standard error associated with regression coefficient estimates of a negatively correlated variable can *contract* if the relationship with the outcome is positive and higher. Also, omitting a negatively correlated variable from a multiple regression model can result in a higher *omission bias* for the parameter estimates (a direct function of $r_{x_1-x_2}$), suggesting that the model was better off having correlations among the explanatory variables. However, it was also observed that there was no substantial change to the collinearity diagnostics like VIF, CI and the determinant while omitting a positively and a negatively correlated variable in separate situations.

Hence, an important learning from their study was that it was important to assess the presence of multicollinearity related problems only if a large change in the diagnostics mentioned above as well as significant changes in slope estimates and their standard errors was observed. Once detected, the incremental contribution of adding a correlated variable to the output variance needed to be established in order to quantify the relative importance of a correlated variable. Increasing the sample size would increase the reliability of the coefficient estimates but was not feasible in

our case. Small sample size increases the standard error of the coefficient estimates, but since the focus in this work was to determine the relative importance of each input variable in predicting output variance, a change in the standard errors and significance was more significant than the absolute value itself. Despite the limited sample size, the findings from this study were found to be consistent with literature that a single parameter (a) could influence oxygen availability significantly to control product selectivity even in reactors of other sizes as well. [10,12,14,16,48,49] Both these aspects will be clear under the MLR model analyses in later sections.

Simple regression models were constructed considering each input variable (Table 4.1) separately. Linear, quadratic, cubic and exponential fits were evaluated along with prediction on the validation set to identify the best fit for each individual predictor variable. The best predicting variables were identified based on the performance parameters like RMSEP, R^2 (equation B.27 and equation B.28 in Appendix B, respectively) and the tendency to overfit was also assessed by means of the difference between RMSEP and RMSEC. Recognizing that sample size was an apparent limitation, an elaboration of this drawback is given in the next section on SLR models.

Going forward, multiple regression models with a maximum of 2 explanatory variables were built, where the incremental contribution of the added variable to the output variance was evaluated and regression coefficient estimates as well as standard errors were compared between the multiple and simple regression models. In addition, standardized coefficients and partial correlations were also evaluated. In this way, harmful effects of multicollinearity were mitigated to an extent so as to draw meaningful chemometric inferences and tenable interpretations about the relative importance of the regressors in predicting the output variables.

4.4.3 Independent dataset for validation/prediction

To validate the formulated calibration models on the basis of the training set of experiments (refer to the experimental data from previous study used as calibration set – Table 4.2), a separate dataset was collected for tetralin oxidation conducted at five other flowrates but at the same temperature and pressure as the calibration set.

The data are summarized in Table 4.4. The predictive power of the constructed models is examined on this validation set with the help of the statistical performance indicators, given previously. This data set is entirely independent of the calibration set but falls within the range of

the flowrates used in the calibration set to guard against extrapolation. The other parameters like L_S, L_G and the outputs accordingly change with the user-controlled flow rates to generate this dataset (compare with Table 4.2).

Table 4.4. Experimental data collected at 150 ºC and 90 kPa (gauge) used for prediction.

Exp ID	Q (μL/min)	L_S (m)	L_G (m)	U_{TP} (m/s)	a (m^2/m^3)	CR (mol/s)	S
V_1	1.5	0.0013	0.1590	0.0282	280000	7.27×10^{-07}	13.0
V_2	2	0.0014	0.1270	0.0254	250000	5.53×10^{-07}	10.0
V_3	4	0.0014	0.0850	0.0226	170000	1.33×10^{-07}	7.00
V_4	7	0.0022	0.0210	0.0205	56700	8.63×10^{-08}	1.30
V_5	12	0.0037	0.0043	0.0153	7100	5.85×10^{-08}	1.40

4.4.4 Simple Linear Regression analysis (SLR): Model performance parameters

The relationships between the output and explanatory variables were evaluated through linear, quadratic and cubic fits with the calibration set experiments by considering only the intercept term and the power of the variable term in the regression calculation. Exponential fits are not shown due to largely inconsequential results.

The coefficient estimates obtained through the calibration model were used to predict the validation set outputs obtained at new flowrates of 1.5, 2, 4, 7 and 12 μL/min and the results of the model performance on both the calibration and validation sets of experiments are shown in Table 4.5 and Table 4.6 for CR and S, respectively.

The first section in 'Results and Discussion' not only highlighted the potentially important predictors for the outcomes purely based on Pearson's pairwise correlation coefficients, but also identified significant collinearity between the explanatory variables that could impact the relative contribution to the output variance. Since the scatter plots (Figure B.1 in Appendix B) indicated that a nonlinear relationship could be present between output-input variables, it was important to find out the best fit for each predictor with the outcome before evaluating effects of multicollinearity.

4.4.4.1 Model prediction for conversion rate of tetralin (CR)

Table 4.5. Model performance parameters for simple linear regression of conversion rate of tetralin on the input variables considering each of them separately.

Input Variable	Nature of fit	RMSEP (mol/s)	r_P^2 [a]	p-value [b]	RMSEC (mol/s)	r_C^2 [c]	RMSEP-RMSEC (mol/s) [d]
a	Linear	1.78E-7	0.75	0.015	8.67E-8	0.89	9.17E-8
	Quadratic	1.53E-7	0.82	0.001	3.14E-8	0.99	1.21E-7
	Cubic	1.69E-7	0.77	0.0002	2.05E-8	0.99	1.49E-7
L_G	Linear	2.34E-7	0.56	0.012	7.95E-8	0.91	1.55E-7
	Quadratic	2.75E-7	0.40	0.0002	2.05E-8	0.99	2.55E-7
	Cubic	3.24E-7	0.16	0.0002	2.11E-8	0.99	3.03E-7
U_{TP}	Linear	2.10E-7	0.65	0.102	1.58E-7	0.65	5.27E-8
	Quadratic	1.68E-7	0.77	0.053	1.31E-7	0.76	3.76E-8
	Cubic	1.35E-7	0.86	0.024	1.00E-7	0.86	3.46E-8
L_S	Linear	3.10E-7	0.24	0.066	1.38E-7	0.73	1.72E-7
	Quadratic	3.17E-7	0.20	0.111	1.62E-7	0.63	1.55E-7
	Cubic	3.38E-7	0.11	0.158	1.82E-7	0.73	1.56E-7
Q	Linear	2.79E-7	0.38	0.203	1.93E-7	0.47	8.57E-8
	Quadratic	3.21E-7	0.18	0.350	2.23E-7	0.29	9.81E-8
	Cubic	3.42E-7	0.07	0.430	2.35E-7	0.21	1.07E-7

[a] Coefficient of determination for the validation set in the SLR; [b] Represents the significance of the overall simple regression model involving the constant and the variable term; [c] Coefficient of determination for the calibration set in the SLR; [d] Always reported as absolute value (positive).

It can be seen from Table 4.5 that RMSEP for a is lowest for quadratic fit with a corresponding highest r_P^2 (0.82), which tells us that the highest variance in the validation set is explained by a quadratic relationship between CR and a. Even though the significance of the cubic term is higher, the p-value for quadratic term is still significant at the 0.1% level which is acceptable. The quadratic model fits the calibration set quite well and shows a lower tendency to overfit than a

cubic model since the gap between RMSEP and RMSEC is 18.5% lower for quadratic fit. Linear fit showed the least propensity to overfit with the difference between RMSEP and RMSEC being 24.4% and 38.4% lower than quadratic and cubic fits, respectively. But since the RMSEP of a^2 term is 14.5% and 9.8% lower than linear and cubic fits respectively, it was chosen as the best fit model for a in predicting CR.

A similar interpretation for the model with L_G as the predictor showed that the linear fit has the least RMSEP (2.35×10^{-7} mol/s), highest r_P^2 (0.56) and minimum tendency to overfit based on the smallest difference between RMSEP and RMSEC (1.55×10^{-7} mol/s). The closeness of both the prediction and calibration error values invariably portray consistency of model performance over multiple ranges of the data points. The p-value for the linear term was significant (1.2%) and gave a good fit for the calibration set as well with r_C^2 of 0.91 and an RMSEC of 7.95×10^{-8} mol/s. On the other hand, a cubic fit worked best for the regression model with U_{TP} as the explanatory variable with the lowest RMSEP and highest r_P^2 values among other fits. Strikingly, it also had the lowest p-value (2.4%), smallest RMSEC and largest r_C^2 and thus simultaneously possessing the least propensity to overfit (least gap between RMSEP and RMSEC). Regression models with L_S and Q revealed that a linear fit could be chosen as the best model for their relationship with CR (Table 4.5).

A comparison of the best fits and not just the linear models for each explanatory variable, in terms of the statistical performance parameters, provided initial information on their individual ability to predict the output, which will further be confirmed moving ahead with multicollinearity analyses. The errors of prediction for the best fits of the 5 variables identified in this section can be arranged in increasing order as: $U_{TP}^3 < a^2 < L_G < Q < L_S$. The terms in the decreasing order of significance for the best fit models were found to be: $a^2 > L_G > U_{TP}^3 > L_S > Q$. This was also a measure of the overall significance of explained variance in CR with the calibration set, in terms of decreasing order of the F-statistic values. Overall, quadratic fit in a contributed to maximum variation in CR (r_C^2 = 0.99), followed by L_G (r_C^2 = 0.91) and U_{TP}^3 (r_C^2 = 0.86), whereas L_S and Q accounted for the least output variance with r_C^2 values of 0.73 and 0.47, respectively.

Similarly, a^2 and U_{TP}^3 contributed to almost equal explained variance in CR for the validation set (r_P^2 = 0.82 and 0.86 respectively) followed by L_G with a r_P^2 value of 0.56. As with the calibration set, Q and L_S accounted for the least variance in CR with values of 0.38 and 0.24 for the coefficient of determination. On an average across all fits, the prediction error was the least when a was used

as the explanatory variable in the simple regression model. Compared to a, the averaged RMSEP values were larger for U_{TP}, L_G, L_S and Q by 1.2%, 67%, 81% and 86% respectively. Likewise, in comparison to a, the averaged calibration errors over all the fits for U_{TP}, L_S and Q were higher by 209%, 257% and 401% respectively while L_G showed a lower RMSEC value than a by a paltry 18%. In summary, keeping the aforementioned discussion in mind, it can be partially concluded that the overall variance in CR is well explained by mass transfer parameter, a and the hydrodynamic parameters, U_{TP} and L_G.

4.4.4.2 Model prediction for ketone-to-alcohol selectivity (S)

The results for this section are compiled in Table 4.6. Unlike the model with CR as the outcome, linear fit had lowest RMSEP of 0.94 with the highest coefficient of determination of 0.98 for a in the validation set. The significance of the linear term was the highest with a p-value of 0.002, but it was to be noted that the quadratic and cubic terms were also significant at the 5% level. The numbers for RMSEC and r_C^2 further testified the accuracy of the linear model. The RMSEC of the linear fit was lower than the quadratic and cubic model by 27% and 53%, respectively. Likewise, the calibration coefficient of determination for the linear fit was higher than the quadratic and cubic fit by 2% and 9%, respectively. The linear fit also displayed the least tendency to overfit due to minimum gap between the prediction and calibration errors, followed closely by the quadratic fit. Although there was not much to differentiate between linear and quadratic fit based on these prediction errors and explained variance of the output variable in the validation set, the linear model was chosen as the best fit for the simple regression between a and S.

For the model with L_G as the regressor, a linear fit was found to be the best with lowest RMSEP and RMSEC values. The linear fit also sported highest coefficient of determination for the calibration and validation sets with the lowest p-value. Expectedly, it depicted least chance of overfitting. In the case of U_{TP} as the explanatory variable, a quadratic fit showed least prediction error and highest r_P^2 though the corresponding numbers for the linear fit were very close. The difference between RMSEP and RMSEC values were the smallest for quadratic fit and thus was chosen as the best fit of U_{TP}.

Table 4.6. Model performance parameters for simple linear regression of selectivity of primary oxidation products on the input variables considering each of them separately.

Input Variable	Nature of fit	RMSEP	r_P^2 [a]	p-value [b]	RMSEC	r_C^2 [c]	RMSEP-RMSEC [d]
a	Linear	0.94	0.98	0.002	0.98	0.98	0.04
	Quadratic	1.41	0.95	0.004	1.35	0.96	0.06
	Cubic	1.78	0.91	0.015	2.08	0.90	0.30
L_G	Linear	1.95	0.89	0.003	1.25	0.96	0.70
	Quadratic	3.58	0.64	0.007	1.59	0.94	1.99
	Cubic	4.87	0.34	0.018	2.23	0.88	2.64
U_{TP}	Linear	3.13	0.73	0.080	3.56	0.69	0.47
	Quadratic	3.09	0.73	0.043	2.93	0.79	0.20
	Cubic	3.61	0.64	0.023	2.39	0.86	1.22
L_S	Linear	7.05	-0.38	0.007	1.65	0.93	5.40
	Quadratic	6.73	-0.26	0.021	2.34	0.87	4.39
	Cubic	6.60	-0.21	0.040	2.88	0.80	3.72
Q	Linear	3.22	0.71	0.137	4.19	0.58	0.97
	Quadratic	4.40	0.46	0.276	5.10	0.37	0.70
	Cubic	5.02	0.30	0.366	5.48	0.27	0.46

[a] Coefficient of determination for the validation set in the SLR; [b] Represents the significance of the overall simple regression model involving the constant and the variable term; [c] Coefficient of determination for the calibration set in the SLR; [d] Always reported as absolute value.

Next, an intriguing observation was that all polynomial fits displayed negative r_P^2 values when L_S was used as the input variable, which meant that L_S was not a suitable predictor for selectivity. The interesting fact was that the fits explained the variance of selectivity in the calibration set quite well with r_C^2 values in the range 0.80-0.93. This portrayed a clear case of model overfitting. The linear term was shown to have highest significance (0.7%) but could not perform well with the validation set.

Lastly, a linear fit was shown perform best when Q was the explanatory variable in the model with a lowest prediction error of 3.22 and highest squared correlation coefficient of 0.71 with S in

the validation set. It was interesting to note that the calibration R^2 for quadratic and cubic fits were very low (0.37 and 0.27 respectively) for Q and the r_C^2 value of 0.58 for the linear fit was the lowest among all the input variables. This was supportive of the discussion in the previous section that Q bore the lowest correlation with the outputs.

The comparison between fits of input variables for the prediction of selectivity is provided in section B.8 in Appendix B. Based on the above analysis, it can be stated that the gas-liquid interfacial area (a) and the length of the gas bubble (L_G) can be considered as better potential predictors than other explanatory variables for selectivity as the output. Nevertheless, their exclusive contributions to the variance in the output variables remains to be evaluated and will be examined in the next few sections.

4.4.4.3 Sample size limitation

As already pointed out, it is quite apparent that the limited number of samples in the calibration and validation sets can be seen as a drawback in this study. Before moving ahead with further analyses, this point needs to be addressed and clarified. From an experimental viewpoint, it was difficult to maintain Taylor flow conditions at flowrates lower than 1 µL/min and higher than 15 µL/min due to practical constraints. The liquid build-up was too high to form liquid slugs and gas bubbles at higher flowrates. On the other hand, at lower flowrates, the lower amount of liquid in the microreactor made it difficult to monitor the reaction in the liquid phase. Also, the maximum conversion attained was 0.74 %wt. at 1 µL/min and there was minimal change anticipated at further lower flowrates. The selectivity was almost constant at flowrates higher than 7 µL/min (Table 4.2 and Table 4.4). Thus, the minimal variability in the conversion and selectivities did not warrant more experiments to be conducted.

From a statistical point of view, with the available 10 samples (5 calibration and 5 validation), a separate study was conducted where the number of samples in the calibration set was increased to 6 and 7 (4 and 3 samples in validation set, respectively) to check the variation of the prediction and calibration errors with conversion rate and selectivity as the outputs and a as the input. It was seen that the best fits highlighted in the previous section remained the same across sample sizes and considering RMSEP, RMSEC and tendency to overfit all together, the choice of 5 calibration and 5 validation sets can be justified. Cross validation was not conducted as specific regression

model parameters were not required to be tuned or optimized as in the work by Sivaramakrishnan et al. [50] The validation dataset given in Table 4.4 can also be used for prediction from the constructed regression models.

4.4.4.4 Residual Plots

It is always a good practice to inspect the residuals to detect any violation from the equal variance assumption, as highlighted in the assumptions section previously. The standardized residuals for the SLR models of CR and S were plotted against the standardized fitted/predicted values of the explanatory variables and are provided in Figure B.2 of Appendix B. It should be mentioned that since there are only 5 points, a rigorous interpretation cannot be carried out. The following can be visualized from the plots:

a. The mean of the residuals is roughly zero.

b. In every case, the residuals represented as scatter points do not appear to follow any known pattern, which means that the linearity assumption between the output and the input term indicated in the figure caption (transformed variable for two-phase velocity) is valid.

c. The residual points can be considered distributed equivalently above and below an imaginary line passing through 0 parallel to the x-axis. This satisfies the assumption of equal variance of the errors.

Thus, based on these observations, the regression models developed can be considered valid and we proceed to multiple linear regression with two predictor variables in the following section.

4.4.5 Multiple linear regression (MLR) vs. SLR models: Comparing parameter estimates, their standard errors and significance

As discussed in the previous section, gas-liquid interfacial area was able to explain maximum variance in both the outputs for the calibration set with the highest significance among other explanatory variables. Excellent predictive ability of the model with a^2 and a as the input for CR and S as the output variable, respectively have been established. But it was also found that L_G and U_{TP} were close competitors with respect to explaining the variance in the output and moreover, were highly correlated with a. Thus, some amount of disentangling the combined effects of the

correlated variables on the output was needed to illustrate the relative importance of each variable in contributing to the output variance significantly.

4.4.5.1 Analysis of the effect of explanatory variables on CR

Table 4.7 and Table B.1 in Appendix B show the results of MLR and SLR for regression of CR on the explanatory variables, respectively. Multicollinearity diagnostic measures are also shown.

Table 4.7. Parameter estimates and multicollinearity diagnostics from the multiple linear regression models with 2 explanatory variables and CR as the output. The eigenvalue column has been omitted due to space constraints. Kindly refer the published article if required.

Model [a]	Var. involved [b]	b_n [c]	SE [d]	Pr > $\|t_{crit}\|$ [e]	Std. coeff. *	R_C^2	F [f]	VIF	CI	Variance Proportions		
										b_0	T_1	T_2
M1	b_0	6.3E-8	2.9E-8	.165		.988	80 (.01)		1.0	.03	.01	.0
	a^2 (T_1)	6.9E-18	1.9E-18	.071	1.16			17	2.3	.43	.02	.0
	L_G (T_2)	-4.7E-7	8.9E-7	.649	-.173			17	12	.53	.97	1
M2	b_0	4.8E-8	4.4E-8	.382		.986	70 (.01)		1.0	.02	.01	.0
	a^2 (T_1)	5.8E-18	1.4E-18	.050	.975			7	2.4	.20	.09	.0
	U_{TP}^3 (T_2)	7.2E-4	.0084	.939	.019			7	9.7	.78	.90	1
M3	b_0	4.0E-8	4.8E-8	.487		.986	72 (.01)		1.0	.03	.03	.0
	a^2 (T_1)	6.1E-18	6.9E-19	.013	1.01			2	1.5	.00	.26	.1
	Q (T_2)	1.2E-9	4.7E-9	.818	.030			2	5.5	.97	.71	.9
M4	b_0	-1.2E-7	6.0E-8	.174		.997	384 (.00)		1.0	.00	.01	.0
	a^2 (T_1)	7.3E-18	5.1E-19	.005	1.22			6	1.7	.00	.12	.0
	L_S (T_2)	3.9E-5	1.3E-5	.097	.251			6	16	1.0	.87	1

[a] Prefix 'M' refers to multiple linear regression models; [b] b_0 represents the constant or intercept term in the regression equation. T_1 and T_2 indicate the respective input variables used to specify the variance proportions associated with that variable; [c] Regression coefficient or slope estimates; [d] Standard errors associated with the respective coefficient estimate; [e] p-value indicating the probability that the t-statistic for the term is greater than the critical value which depends on the degrees of freedom (DF) of the model. For MLR, DF = 2 & t_{crit} = 4.303; [f] F-statistic, distributed as F with 2 & 2 DF for MLR ($F_{critical}$ = 19.00). Statistic for the test of overall R^2 of the model. Value in bracket is the p-value for the significance of the overall model; * Standardized coefficient for the input variable in the MLR model.

(i) Effect of gas-liquid interfacial area and length of gas bubble

First, it is to be noted that a maximum of 2 variables were considered for the regression analysis because of the limited number of samples available in the calibration set. Even with only two variables in the system, three parameters needed to be estimated with 5 data points. Care was taken to restrain the system from being over-defined, i.e. having more variables than observations. Consider model M1, S1 and S3 in Table 4.7 and Table B.1 in Appendix B. The sign of the regression coefficient estimate for L_G was changed from positive (S3) to negative (M1), while it remained same for a^2. This reversal for L_G is an indication that any mutation in this variable would negatively affect the changes in CR when a^2 is kept constant and thus, its contribution to the output variance is diminished compared to a^2 in the combined model. The standard errors for a^2 and L_G when modelled in combination increased by 377% and 88%, respectively which is a direct consequence of the high correlation between them.

Moreover, despite the overall model (M1) being significant with a p-value of 0.012 corresponding to the F-statistic, the individual coefficient estimates were not significant at the 5% level. However, the coefficient estimate for L_G displayed a much higher p-value (0.649) than that of a^2 (0.071), providing a hint that a^2 term was relatively more important than L_G. Critical values of F-test [41] and t-tests [51] are obtained from literature.

Since both the predictors are measured in different units, it was useful to look at the dimensionless standardized coefficients, where a higher value would imply more relative significance of that variable. In other words, when the input variable changes by one standard deviation keeping the other variable constant, the outcome changes by 'standardized weight' number of standard deviations corresponding to that variable. It can be seen from Table 4.7 that the standardized weight of a^2 (1.16) is an order of magnitude higher than L_G (-0.173), again reiterating that the a^2 has a higher weightage in affecting CR. This was not as apparent in the individual models S1 and S3 (Table B.1) since the standardized weights were not too different, though a^2 had a higher value. In addition, the variables involved had the highest value for VIF (17) as compared to models M2, M3 and M4, complementing the fact that a^2 and L_G possessed the highest collinearity among all other combinations with a^2. Finally, both the predictors displayed high variance proportions on the third eigenvalue dimension (Table 4.7).

Almost all the variance in their regression coefficients is associated with this dimension (97% and 100%), thus confirming the large dependency between the variables. Attention is to be brought to the fact that though the CI was below 15 [33] or 20, [34] the question of multicollinearity still prevailed. This further exemplifies the notion in the literature that no consensus has yet been reached regarding the threshold values of any diagnostic measure explaining multicollinearity. It is obvious that the R^2 never decreases when a regressor variable is added to the existing model as can be comprehended from S1, S3 and M1. R^2 of M1 is higher than that of S1 and S3 by 0.002 and 0.078 units, respectively. The incremental effect of each predictor in increasing the total variance in CR is investigated through a modified F-statistic (equation B.26 in Appendix B). The values of the statistic for the different combination of variables analyzed in models M1 - M4 are given in Table 4.8.

The F-curve in this case is distributed with 1 and 2 degrees of freedom when the null hypothesis (that there is no incremental effect of the added variable) is true. When L_G was added to a^2 (S1), the F value was 0.33, while if a^2 is added to L_G (S3), F increased to 13.00. Neither occasion was significant since the critical value for F (1, 2) is 18.51 but it definitely indicated that the incremental effect of adding a^2 to L_G was much greater than the other way around. Lastly, the partial correlation between CR and a^2, while controlling for L_G was 0.929 (a decrease of 6.4% from the zero-order correlation of 0.993). Contrarily, the partial correlation between CR and L_G, while controlling for a^2 stooped down to -0.351 (a sharp decrease of 137%).

This clearly certified that the relationship between conversion rate of tetralin and gas-liquid interfacial area was not affected much by the length of the gas bubble, even though they were highly correlated (second column from the right in Table 4.8). Put differently, in the presence of a^2 term in the model M1, the contribution of L_G to the variation in CR was negligible.

Table 4.8. F-statistic values for addition of one explanatory variable to SLR models with CR and S as outcome.

Output variable	Variables involved	R^2 MLR	r^2 SLR [a]	F-statistic	F-critical (1,2)	r [b]	VIF
CR	L_G over a^2 a^2 over L_G	.988 (M1) .988	.986 (S1) .910 (S3)	.33 13.00	18.51	.971	17.29
	U_{TP}^3 over a^2 a^2 over U_{TP}^3	.986 (M2) .986	.986 .856 (S2)	0.00 18.57	18.51	.929	7.32
	Q over a^2 a^2 over Q	.986 (M3) .986	.986 .467 (S4)	0.00 74.14	18.51	-.704	1.98
	L_S over a^2 a^2 over L_S	.997 (M4) .997	.986 .727 (S5)	7.33 180.00	18.51	-.905	5.51
S	L_G over a a over L_G	.986 (SM1) .986	.977 (SS1) .962 (SS2)	1.29 3.43	18.51	.998	250.25
	U_{TP}^2 over a a over U_{TP}^2	.995 (SM2) .995	.977 .792 (SS3)	7.20 81.20	18.51	.946	9.52
	Q over a a over Q	.993 (SM3) .993	.977 .576 (SS4)	4.57 119.14	18.51	-.839	3.37
	L_S over a a over L_S	.987 (SM4) .987	.977 .934 (SS5)	1.54 8.15	18.51	-.945	9.35

[a] Calibration coefficient of determination of SLR for CR and S on the existing variable. Reference to the corresponding model in Table B.1 and Table B.2 in Appendix B is given in brackets; [b] Pairwise bivariate correlation between the two variables involved.

(ii) Effect of gas-liquid interfacial area and two-phase velocity

No sign reversal was observed in the regression coefficient estimate of U_{TP}^3 in model M2 (Table 4.7) when compared with model S2 (Table B.1 in Appendix B), but it decreased by 2 orders of magnitude (by 98%) from S2. On the contrary, the change in the slope estimate for a^2 was imperceptible. The interpretation of the estimates would be more meaningful by looking at the

corresponding standard errors. The standard error for U_{TP}^3 in M2 is an order of magnitude higher than the coefficient itself, while in S2, it is an order of magnitude smaller.

This indicates that the coefficient estimate of U_{TP}^3 is unstable and not very reliable because the lower bound for the 95% confidence interval range of values for this estimate was negative (-0.0354), which means that 0 is also a possible value for U_{TP}^3. Though the standard error for a^2 increased by 231% as compared to S1, the slope estimate obtained in M2 was still significant (p-value = 0.050). The standardized coefficients for the two predictors signified that changes in a^2 affected the output much more than changes in U_{TP}^3, thus emphasizing the higher importance of a^2 in predicting the output.

Since the pairwise correlations between L_G & CR (0.954 – Table 4.3) and U_{TP}^3 & CR (0.925 – Table B.1) are both positive, it is expected that the regression coefficient estimates (standardized included) should also be positive in the MLR. As already reviewed, this was not the case with L_G (previous section), plausibly due to the high correlation with a^2 (0.971 – Table 4.8). In the case of U_{TP}^3, the coefficients are still much lower compared to a^2 in the MLR, but at least still maintain the positive effect on the output, unlike L_G. The reason for this can be attributed to the lower correlation between U_{TP}^3 and a^2 (0.929 - Table 4.8) in comparison with that between L_G and a^2.

The overall model (M2) is significant at the 5% level with a F (2, 2) = 70.85 and a p-value of 0.014. With reference to the discussion on multicollinearity, a change in the VIF is a better indicator of multicollinearity than the absolute value itself. VIF was 7 between U_{TP}^3 and a^2, which meant that according to Kutner et al., [52] multicollinearity was not a problem. But as was observed in Vatcheva et al., [37] VIF increased to 7 as compared to the SLR models S1 and S2, thus implying that collinearity can be a hindrance to model interpretation. The VIF was lower than that of M1 and it was interesting to note that it followed the order of the pairwise correlations between the transformed explanatory variables as given in Table 4.8.

On a relative scale, the problem of collinearity between U_{TP}^3 and a^2 was found to be lesser than that existing between L_G and a^2 and was reflected in the results just elaborated. CI value for M2 was also lower than that for M1, but still higher than those for S1 and S2, indicating the presence of collinearity. This was also confirmed by the higher concentration of the variance proportions of the regression coefficients of both the predictors on the third eigenvalue dimension in M2 (90% and 100%).

The results from Table 4.8 for CR were further affirmative of the relative significance of a^2 as compared to U_{TP}^3 in explaining output variance. The increment to r^2 of S1 provided by U_{TP}^3 was insignificant (F-statistic = 0.00), while the incremental effect of adding a^2 to model S2 in determining CR values proved to be significant with F (1,2) = 18.57 and cleared the critical value of 18.51. Additionally, the partial correlations between a^2 & CR, controlling for U_{TP}^3 and that between U_{TP}^3 and CR, controlling for a^2 were 0.950 and 0.060, respectively. This reflected that in the presence of a^2 term, the incremental contribution of U_{TP}^3 to the output variance is slim. Thus, it can be stated that the gas-liquid interfacial area was the more important parameter compared with two-phase velocity as well.

(iii) Effect of gas-liquid interfacial area and injection flow-rate of tetralin

For brevity, details of this section are provided in section B.9.1 in Appendix B. To summarize, it can be asserted that the gas-liquid interfacial area (transformed quadratic form) was the more important variable in predicting conversion rate of tetralin, relative to the injection flow rate.

(iv) Effect of gas-liquid interfacial area and length of liquid slug

The reader is referred to section B.9.1 in Appendix B for the detailed analysis of this section. These results gave a different picture of collinearity, where this phenomenon actually increased the individual contribution of a predictor in describing the overall variance of the output, that was previously observed in literature. [34]

But, the underlying observation in all four MLR models was that the relative importance of the gas-liquid interfacial area was substantial in explaining the variance in conversion rate.

4.4.5.2 Analysis of the effect of explanatory variables on S

Table 4.9 and Table B.2 in Appendix B show the results of SLR and MLR for regression of S on the explanatory variables, respectively. Multicollinearity diagnostic measures are also shown.

Table 4.9. Parameter estimates and multicollinearity diagnostics from the simple and multiple linear regression models with 2 explanatory variables and S as the output.

Model [a]	Var. involved [b]	b_n [c]	SE [d]	Pr > $\|t_{crit}\|$ [e]	Std. coeff. *	R_C^2	F [f]	VIF	EV [g]	CI	Variance Proportions		
											b_0	T_1	T_2
SM1	b_0	.376	1.05	.590					2.6	1	.05	.00	.0
	a (T_1)	1E-4	6E-5	.203	2.63	.986	71 (.014)	250	.4	3	.92	.00	.0
	L_G (T_2)	-108	92	.364	-1.64			250	.0	54	.03	1.0	1
SM2	b_0	2.55	.89	.103					2.7	1	.01	.01	.0
	a (T_1)	6E-5	7E-6	.011	1.39	.995	211 (.005)	9.5	.3	3	.13	.08	.0
	U_{TP}^2 (T_2)	-1E4	3982	.107	-.42			9.5	.0	14	.86	.91	1
SM3	b_0	-2.30	1.22	.200					2.2	1	.01	.02	.0
	a (T_1)	5E-5	5E-6	.008	1.19	.993	146 (.007)	3.4	.8	2	.00	.11	.1
	Q (T_2)	.2320	.105	.158	.24			3.4	.0	8	.99	.87	.9
SM4	b_0	5.89	4.64	.332					2.4	1	.00	.01	.0
	a (T_1)	3E-5	1E-5	.106	.70	.987	74 (.013)	9.3	.6	2	.00	.06	.0
	L_S (T_2)	-1154	942	.345	-.31			9.3	.0	23	1.00	.93	1

[a] 'SM' refers to multiple linear regression models for selectivity as the output; [b] b_0 represents the constant or intercept term in the regression equation. T_1 and T_2 indicate the respective input variables used to specify the variance proportions associated with that variable; [c] Regression coefficient or slope estimates; [d] Standard errors associated with the respective coefficient estimate; [e] p-value indicating the probability that the t-statistic for the term is greater than the critical value of t-stat, that depends on the degrees of freedom (DF) of the model. For MLR, DF = 2 & t_{crit} = 4.303; [f] F-statistic, distributed as F with 2 and 2 DF for MLR ($F_{critical}$ = 19.00). Statistic for the test of overall R^2 of the model. Value in bracket is the p-value for the significance of the overall model; [g] Eigenvalue for the respective dimension; * Standardized coefficient.

(i) Effect of gas-liquid interfacial area and length of gas bubble

Best fits for a and L_G based on predictive performance on selectivity were already identified as linear for both the predictors in the results of SLR as discussed previously. One observation that was very apparent in this case was the very high simple correlation between a and L_G (0.998 – Table 4.8). It was supported by the highest VIF (250.25) for SM1 among all MLR models with a correspondingly large CI value of 54.12 (Table 4.9). Almost 100% of the proportion of variances of the regression coefficient estimates of both the predictors belonged to the third eigenvalue dimension, indicating exorbitant collinearity.

The steep change in the magnitude as well as the sign reversal (positive to negative) of the regression coefficient estimate of L_G from model SS2 to SM1 can be attributed to its high correlation with a. This was accompanied by an acute upsurge of 1250% in the standard error of L_G. Concurrently, both the regression coefficient estimate and the standard error of a increased by an order of magnitude from model SS1 to SM1, rendering the coefficient insignificant in the process (p-value = 0.203). But, the L_G term was relatively less significant than a, having a p-value of 0.364. The standardized coefficients clearly point towards the higher relative importance of a (2.63) as compared with L_G (-1.64). Nevertheless, the overall model was significant with a p-value (two-tailed) of 0.014 for the probability of a new F-statistic being more extreme than the observed value of 70.92 for SM1.

The F-statistic for the test of incremental effect of adding a to L_G (3.43) for improving the overall r^2 from the SLR to MLR, was greater than the adding L_G to a (1.29) by a small margin (Table 4.8). The collinearity between the explanatory variables involved in model SM1 resulted in a downturn of 19.3% from the zero-order correlation to the partial correlation between a and S, while controlling for L_G. a had a more dramatic effect on the relationship between L_G and S as a zero-order correlation of 0.981 dropped down to a partial correlation of -0.636, thus amplifying the relative importance of a over L_G in influencing selectivity of the products in tetralin oxidation.

(ii) Effect of gas-liquid interfacial area and two-phase velocity

A quadratic model was determined as best fit for the prediction of S with U_{TP} (refer SLR results). When regressed along with a in predicting S as in model SM2, the overall significance of the model was very high (p-value = 0.005 for the F-statistic of 210.90), with a corresponding R^2 value of 0.995 (Table 4.9). The VIF and CI numbers stood at 9.52 and 14.34, respectively, which were both near the respective threshold values highlighted by Kutner et al. [52] and Midi et al. [32] The CI value for a did not change by much but compared to SS3, it increased 3-fold for U_{TP}^2. Major proportion of the variances (91% and 100%) of the regression coefficients of the predictors were found to be concentrated on the 3rd eigenvalue dimension. A high correlation coefficient of 0.946 (Table 4.8 for S) between the predictors, along with the collinearity diagnostic statistics indicated that the collinearity could be a problem in regression.

Conceivably, the regression coefficient estimate of U_{TP}^2 moved from positive (23740.8) to negative (-11194.0) in model SM2 from SS3, revealing the negative impact of collinearity on U_{TP}^2. This drastic difference made the term insignificant (p-value = .107), even though the standard error reduced approximately by half. On the other hand, the slope estimate and the corresponding standard error of a in model SM2 increased by 1.4 times and 1.75 times from SS1 respectively, but the coefficient was still significant (p-value = 0.011). Additionally, the increased standardized coefficient of a (1.39) in comparison with that of L_G (-0.42) elucidated the higher influence of the effect of a change in a causing a change in S.

The incremental contribution in elevating the overall variance of the model SM2 by appending a to U_{TP}^2 (in model SS3) was significant, as indicated by a F-statistic value of 81.20. Adding U_{TP}^2 to a did not prove significant since the F-value was lower than the critical value of 18.51. The partial correlation between a and S (0.989) signified that their relationship was not affected by U_{TP}^2 at all since it was 0.001 units higher than the zero-order correlation between them. On the other hand, the presence of a had a negative impact on the relationship between U_{TP}^2 and S as the partial correlation was -0.893 while the zero-order correlation was 0.890. These results tell us that gas-liquid interfacial area was the more important predictor for selectivity.

(iii) Effect of gas-liquid interfacial area and injection flow-rate of tetralin

Detailed discussion for this section is provided in section B.9.2 in Appendix B. The analysis suggested that a was the more relevant variable than Q in predicting the selectivity.

(iv) Effect of gas-liquid interfacial area and length of liquid slug

Section B.9.2 in Appendix B provides the analysis for this section as well. On the whole, it can be stated with confidence that gas-liquid interfacial area was the most significant contributor to overall variance in the outcomes, conversion rate and selectivity based on the above analyses.

4.4.6 Predicting the outputs from Q: Identification of the best regression pathway

Since Q was the only manipulated variable in the tetralin oxidation system, it was considered lucrative to trace out the best possible statistical path to predict the output variables. The outcomes of this new path were compared with the SLR models, S4 and SS4 which directly predicted CR and S respectively from Q to look for improvements in model performance parameters (RMSE and R^2). This is a faster way to monitor the reaction's progress without the use of offline GC measurements, which was another aim of this study.

The results of this section are compiled in Table 4.10. Based on Table 4.3, it can be seen that Q had maximum correlation with U_{TP} ($r_{Q-U_{TP}}$ = -0.980). A simple linear regression model was built with U_{TP} as the outcome and Q as the explanatory variable based on the calibration set and a linear model was determined the best fit based on least RMSEP and highest r_P^2 for the validation set.

Next, a regression relationship was established between U_{TP} and a, since the gas-liquid interfacial area was proved to be the best predictor for both the outputs as discussed before. With the help of data points in the calibration set, a linear model was chosen as the best fit between these two variables. Afterwards, the predicted U_{TP} values in model P1 (Table 4.10) with the validation set were used as input to the regression relation built in model P2 in order to predict a. Lastly, these new a values were utilized to predict CR and S using the already determined best fit quadratic and linear regression relationships (SLR models for CR and S). The variance explained in CR through this process was higher (r_P^2 = 0.46) than model S4 (r_P^2 = 0.38) for the validation set (Table 4.10). Also, the prediction error in model P3 was lower than model S4 by 7%, indicating a clear advantage of this pathway. Likewise, the prediction error and the coefficient of determination for

the validation set in model P4 was lower by 8% and higher by 7% than the corresponding values in model SS4, respectively.

The previous discussions on SLR and MLR models proved that the outputs can be predicted with acceptable accuracy from the gas-liquid interfacial area, which needed to be derived from analysis of the images captured during the tetralin oxidation experiments. [5] It discarded the need for GC analysis to calculate conversion rate and selectivity. The new pathway discussed in this section proved beneficial and showed potential to predict conversion rate and selectivity from inlet flowrate of tetralin by eliminating the need for calculating the hydrodynamic parameters through image analysis, in addition to avoiding GC measurements. This also saves time for the researcher.

Table 4.10. Results for prediction of CR and S from Q through the best regression pathway.

Input variable	Output variable	Model ID	RMSEP *	r_P^2	RMSEC *	r_C^2	p-value [a]	RMSEP – RMSEC [b]
Q	U_{TP}	P1	.0031	0.70	.0013	0.96	0.003	0.0018
U_{TP}	a	P2	58412.4	0.82	61427.0	0.81	0.037	3014.6
a	CR	P3	**2.60E-7**	**0.46**	3.14E-8	0.99	0.001	2.29E-7
a	S	P4	**2.95**	**0.76**	0.98	0.98	0.002	1.97
Q	CR	S4	2.79E-7	0.38	1.93E07	0.47	0.203	8.57E-8
Q	S	SS4	3.22	0.71	4.19	0.58	0.137	0.97

* The values are in the respective units of the output variable; [a] p-value relates to the significance of both the coefficient of the predictor as well as the overall model in SLR; [b] Always reported as absolute values.

4.4.7 Implications of the current work

This study proved that the oxidation process of an important naphthene-aromatic petrochemical, tetralin conducted at 150 ºC and near-atmospheric pressure could be monitored (tracking of reactant conversion and product selectivity) using conventional chemometric tools. The statistical analyses performed in this work signify that the most important parameter that controlled product selectivity is the gas-liquid interfacial area (a). The two-phase velocity (U_{TP})

that is closely related to the Marangoni effects, comparatively constitutes a lesser weightage in influencing the system outputs.

A key implication is that a reactor could be designed to create a higher interfacial area, for example, a loop reactor, [48,49,53] for industrial applications on a larger scale. However, small-scale productions of high value products, valuable pharmaceuticals or chemicals [14,54] can be produced by using the microfluidic reactor due to its capability to maintain higher interfacial area. There was a definite value addition to our previous work [5] in employing mathematical tools to study the dependencies between the different parameters that affected oxygen availability and deducing the most significant parameter.

4.5 Conclusions

Application of statistical regression models to monitor tetralin oxidation reaction in a microfluidic reactor by identifying the parameters of paramount importance that affect the outputs (conversion rate and selectivity) was the central focus of this work. IBM SPSS (Statistical Package for the Social Sciences) software was used for all statistical analysis conducted in this work. Some key findings were as follows:

a) The magnitude and direction of the relationships between the explanatory variables and the outcomes (CR and S) were investigated by means of Pearson's correlation coefficients. It was found that L_G and a showed maximum positive correlation with CR and S respectively. Q and L_S displayed a negative correlation with both the outputs, whereas U_{TP} had a moderate positive correlation. There were also indications on the presence of strong multicollinearity between the predictors as evident from values of correlations > 0.9. This was confirmed further by diagnostic measures like elevated VIF and CI values.

b) Best fits for each predictor were determined by examining separate SLR models and evaluating the results through model performance parameters, i.e. chosen based on lower RMSEP, higher r_P^2 and lesser propensity to overfit. Specifically, the order of best-fitting predictors based on SLR models and considering the statistical performance parameters were: $a^2 > L_G > U_{TP}^3 > L_S > Q$ for CR as the output; $a > L_G > U_{TP}^2 > Q$ for S as the output. The model with L_S as the explanatory variable gave poor results with the validation set, indicating overfitting. A major limitation in this study was the small number of calibration

samples due to experimental and practical constraints but as elaborated in the discussion, it did not hinder the consensus on a meaningful interpretation.

c) In order to identify the relative importance of the predictors in explaining output variance, it was essential to resolve the issue of multicollinearity among the explanatory variables. For each output, four MLR models were explored by adding one input variable to the best predictor from SLR (gas-liquid interfacial area – a). Sign reversals and inflated standard errors were observed for the added variable, thus rendering it insignificant due to collinearity effects. The standardized regression coefficients from MLR models further gave evidence that a influenced the outputs the most.

d) Partial correlations between a and CR and S, controlling for other predictors further indicated that even in the presence of other variables, a shared maximum variance with the outputs. Incremental contribution of a in improving the overall model r_C^2 by adding to the SLR models with other predictors was proved to be significant through F-tests distributed with (1, 2) degrees of freedom.

e) A faster way to monitor the reaction by eliminating the need for image analysis, in addition to the GC measurements, was attempted. The best pathway to predict system deliverables from the lone manipulated variable in the system (Q) was identified to be: Q -> U_{TP} -> a -> CR and S. This pathway gave lower RMSEP and higher r_P^2 as compared to model S5 and SS5, where the outputs were directly predicted from tetralin injection flowrate.

Thus, by means of the chemometric approach undertaken in this study, it can be concluded that oxygen availability in terms of gas-liquid interfacial area is one of the major factors affecting conversion and selectivity in tetralin oxidation. It also indicated potential application of the same principles to larger reactors (loop reactors for example) for industrial applications.

4.6 References

(1) Centi, G.; Cavani, F.; Trifiro, F. *Selective Oxidation by Heterogenous Catalysis*; Kluwer/Plenum: New York, 2001.

(2) Suresh, A. K.; Sharma, M. M.; Sridhar, T. Engineering Aspects of Industrial Liquid-Phase Air Oxidation of Hydrocarbons. *Ind. Eng. Chem. Res.* **2000**, *39*, 3958–3997.

(3) Clerici, M. G.; Kholdeeva, O. A. *Liquid Phase Oxidation via Heterogeneous Catalysis*; John Wiley & Sons, Inc.: Hoboken, NJ, 2013.

(4) Hermans, I.; Peeters, J.; Jacobs, P. A. Autoxidation of Hydrocarbons: From Chemistry to Catalysis. *Top. Catal.* **2008**, *50*, 124–132.

(5) Siddiquee, M. N.; de Klerk, A.; Nazemifard, N. Application of Microfluidics to Control Product Selectivity during Non-Catalytic Oxidation of Naphthenic-Aromatic Hydrocarbons. *React. Chem. Eng.* **2016**, *1*, 418–435.

(6) De Klerk, A. Continuous-Mode Thermal Oxidation of Fischer-Tropsch Waxes. *Ind. Eng. Chem. Res.* **2003**, *42*, 6545–6548.

(7) Govindan, V.; Suresh, A. K. Modeling Liquid-Phase Cyclohexane Oxidation. *Ind. Eng. Chem. Res.* **2007**, *46*, 6891–6898.

(8) Siddiquee, M. N.; de Klerk, A. In Situ Measurement of Liquid Phase Oxygen during Oxidation. *Ind. Eng. Chem. Res.* **2016**, *55*, 6607–6618.

(9) Emanuel, N. M.; Denisov, E.; Maizus, Z. K. *Liquid-Phase Oxidation of Hydrocarbons*; Plenum Press: New York, 1967.

(10) Gemoets, H. P. L.; Su, Y.; Shang, M.; Hessel, V.; Luque, R.; Noël, T. Liquid Phase Oxidation Chemistry in Continuous-Flow Microreactors. *Chem. Soc. Rev.* **2016**, *45*, 83–117.

(11) Sobieszuk, P.; Aubin, J.; Pohorecki, R. Hydrodynamics and Mass Transfer in Gas-Liquid Flows in Microreactors. *Chem. Eng. Technol.* **2012**, *35*, 1346–1358.

(12) Neuenschwander, U.; Jensen, K. F. Olefin Autoxidation in Flow. *Ind. Eng. Chem. Res.* **2014**, *53*, 601–608.

(13) Vanoye, L.; Aloui, A.; Pablos, M.; Philippe, R.; Percheron, A.; Favre-Reguillon, A.; Bellefon, C. De. A Safe and Efficient Flow Oxidation of Aldehydes with O 2. *Org. Lett.* **2013**, *15*, 5978–5981.

(14) Hamano, M.; Nagy, K. D.; Jensen, K. F. Continuous Flow Metal-Free Oxidation of Picolines Using Air. *Chem. Commun.* **2012**, *48*, 2086–2088.

(15) Marre, S.; Adamo, A.; Basak, S.; Aymonier, C.; Jensen, K. F. Design and Packaging of Microreactors for High Pressure and High Temperature Applications. *Ind. Eng. Chem. Res.* **2010**, *49*, 11310–11320.

(16) Jevtic, R.; Ramachandran, P. A.; Dudukovic, M. P. Capillary Reactor for Cyclohexane

Oxidation with Oxygen. *Chem. Eng. Res. Des.* **2010**, *88*, 255–262.

(17) Leclerc, A.; Alam, M.; Schweich, D.; Pouteau, P.; Bellefon, C. De. Gas – Liquid Selective Oxidations with Oxygen under Explosive Conditions in a Micro-Structured Reactor. *Lab Chip* **2008**, *8*, 814–817.

(18) Fischer, J.; Lange, T.; Boehling, R.; Rehfinger, A.; Klemm, E. Uncatalyzed Selective Oxidation of Liquid Cyclohexane with Air in a Microcapillary Reactor. *Chem. Eng. Sci.* **2010**, *65*, 4866–4872.

(19) Günther, A.; Khan, S. A.; Thalmann, M.; Trachsel, F.; Jensen, K. F. Transport and Reaction in Microscale Segmented Gas–Liquid Flow. *Lab Chip* **2004**, *4*, 278–286.

(20) Triplett, K. A.; Ghiaasiaan, S. M.; Abdel-Khalik, S. I.; Sadowski, D. L. Gas–Liquid Two-Phase Flow in Microchannels Part I: Two-Phase Flow Patterns. *Int. J. Multiph. Flow* **1999**, *25*, 377–394.

(21) Yao, C.; Dong, Z.; Zhao, Y.; Chen, G. An Online Method to Measure Mass Transfer of Slug Flow in a Microchannel. *Chem. Eng. Sci.* **2014**, *112*, 15–24.

(22) Yue, J.; Luo, L.; Gonthier, Y.; Chen, G.; Yuan, Q. An Experimental Study of Air – Water Taylor Flow and Mass Transfer inside Square Microchannels. *Chem. Eng. Sci.* **2009**, *64*, 3697–3708.

(23) Fries, D. M.; Trachsel, F.; Rohr, P. R. Von. Segmented Gas – Liquid Flow Characterization in Rectangular Microchannels. *Int. J. Multiph. Flow* **2008**, *34*, 1108–1118.

(24) Yun, J.; Lei, Q.; Zhang, S.; Shen, S.; Yao, K. Slug Flow Characteristics of Gas-Miscible Liquids in a Rectangular Microchannel with Cross and T-Shaped Junctions. *Chem. Eng. Sci.* **2010**, *65*, 5256–5263.

(25) Yan, X.; Xiaogang, S. Simple Linear Regression. In *Linear Regression Analysis: Theory and Computing*; World Scientific, 2009; pp 9–39.

(26) Freedman, D. *Statistical Models: Theory and Practice*; Cambridge University Press: Cambridge, 2009.

(27) Lai, T. L.; Robbins, H.; Wei, C. Z. Strong Consistency of Least Squares Estimates in Multiple Regression II. *J. Multivar. Anal.* **1979**, *9*, 343–361.

(28) Lange, K. L.; Little, R. J. A.; Taylor, J. M. G. Robust Statistical Modeling Using the t Distribution. *J. Am. Stat. Assoc.* **1989**, *84*, 881–896.

(29) Adeboye, N. O.; Fagoyinbo, I. S.; Olatayo, T. O. Estimation of the Effect of

Multicollinearity on the Standard Error for Regression Coefficients. *J. Math.* **2014**, *10*, 16–20.

(30) Brannick, M. *Regression with Two Independent Variables*; University of South Florida.

(31) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons, Inc.: New York, 1980.

(32) Midi, H.; Sarkar, S. K.; Rana, S. Collinearity Diagnostics of Binary Logistic Regression Model. *J. Interdiscip. Math.* **2010**, *13*, 253–267.

(33) Johnston, J. *Econometric Methods*, 3rd ed.; McGraw-Hill: New York, 1984.

(34) Mela, C. F.; Kopalle, P. K. The Impact of Collinearity on Regression Analysis: The Asymmetric Effect of Negative and Positive Correlations. *Appl. Econ.* **2002**, *34*, 667–677.

(35) Siddiquee, M. N.; De Klerk, A. Hydrocarbon Addition Reactions during Low-Temperature Autoxidation of Oilsands Bitumen. *Energy Fuels* **2014**, *28*, 6848–6859.

(36) Montgomery, D. C.; Peck, E. A.; Vining, G. G. *Introduction to Linear Regression Analysis*, 5th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2012.

(37) Vatcheva, K. P.; MinJae, L.; McCormick, J. B.; Rahbar, M. H. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiol. Open Access* **2016**, *6*, 227.

(38) Kowalski, B. R. *Chemometrics*; Springer Netherlands: Dordrecht, 1984.

(39) Thode, H. C. *Testing for Normality*; Marcel Dekker: New York, 2002.

(40) Zaiontz, C. Pearson's Correlation Table http://www.real-statistics.com/statistics-tables/pearsons-correlation-table/ (accessed May 7, 2018).

(41) Dinov, I. D. Statistics Online Computational Resource (SOCR). *UCLA Statistics, Neurology, LONI* 2016.

(42) Green, P. E.; Tull, D. S.; Albaum, G. S. *Research for Marketing Decisions*, 5th ed.; Prentice Hall: Englewood Cliffs,NJ, 1988.

(43) Berry, W.; Feldman, S. *Multiple Regression in Practice*; SAGE Publications, Inc.: Thousand Oaks, CA, 1985. https://doi.org/10.4135/9781412985208.

(44) Tull, D. S.; Hawkins, D. I. *Marketing Research: Measurement and Method*; Maxwell Macmillan International: New York, 1993.

(45) Lehmann, D. R.; Gupta, S.; Steckel, J. H. *Marketing Research*; Addison-Weasley: Reading, MA, 1998.

(46) Field, A. P. *Discovering Statistics Using SPSS*; SAGE Publications, Inc.: London, 2005.

(47) Stewart, G. W. Collinearity and Least Squares Regression. *Stat. Sci.* **1987**, *2*, 68–84.

(48) Stitt, E. H. Alternative Multiphase Reactors for Fine Chemicals: A World beyond Stirred Tanks? *Chem. Eng. J.* **2002**, *90*, 47–60.

(49) Havelka, P.; Linek, V.; Sinkule, J.; Zahradník, J.; Fialová, M. Hydrodynamic and Mass Transfer Characteristics of Ejector Loop Reactors. *Chem. Eng. Sci.* **2000**, *55*, 535–549.

(50) Sivaramakrishnan, K.; Nie, J.; de Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, *57*, 13156–13176.

(51) Piegorsch, W. W. *Tables of P-Values for t- and Chi-Square Reference Distributions*; University of South Carolina, Columbia, 2002.

(52) Kutner, M.; Nachtsheim, C.; Neter, J. *Applied Linear Statistical Models*; McGraw-Hill: Irwin, 2004.

(53) Cramers, P. H. M. R.; Beenackers, A. A. C. M. Influence of the Ejector Configuration, Scale and the Gas Density on the Mass Transfer Characteristics of Gas-Liquid Ejectors. *Chem. Eng. J.* **2001**, *82*, 131–141.

(54) Snead, D. R.; Jamison, T. F. A Three-Minute Synthesis and Purification of Ibuprofen: Pushing the Limits of Continuous-Flow Processing. *Angew. Chemie - Int. Ed.* **2015**, *54*, 983–987.

5. Viscosity of Canadian oilsands bitumen and its modification by thermal conversion [4]

ABSTRACT

The dependence of Canadian oilsands-derived bitumen viscosity on different factors like the microstructure, chemical composition, compatibility and interactions between the different constituent components has been documented in literature but a comprehensive review is missing. The goal of this chapter is to explore the sensitivity of bitumen viscosity on some of these parameters in depth and to provide scientific explanation for the observed changes. Non-catalytic-thermal conversion is used as tool for realizing this objective. Previous studies on thermal treatment of bitumen reported contrasting trends in viscosity with reaction time, but a substantial explanation was not provided and served as motivation for the experimental investigations in this chapter. Specifically, post-reaction procedures like nature of solvent used to extract the products from the reactor and rheological conditions of viscosity measurement like shear rate had a significant impact on viscosity. Methylene chloride introduced hydrogen bonding interactions due to some solvent remaining after evaporation, which plausibly increased viscosity of thermally converted bitumen significantly. This was not observed with toluene as the extraction solvent. Tracking changes in other properties like free radical content, boiling point distribution, aromatic and aliphatic content during thermal conversion as well as before and after solvent addition and removal assisted in accounting for the viscosity changes. Chemical composition of the feed and its geological origin also seemed to have an effect on its viscosity. Assuming a colloidal representation for bitumen, a theory was proposed on the effect of the nature of asphaltene aggregation on viscosity in terms of changes in the effective volume of the aggregate clusters during thermal conversion. The onset of elastic nature on application of shear is also explored.

Keywords: Thermal conversion; Athabasca bitumen viscosity; solvent extraction; effect of shear rates; product characterization; asphaltene aggregation.

[4] This work was published as a book chapter in the ACS Symposium Series as "Sivaramakrishnan, K.; De Klerk, A.; Prasad, V. Viscosity of Canadian Oilsands Bitumen and Its Modification by Thermal Conversion. In *Chemistry Solutions to Challenges in the Petroleum Industry*; Rahimi, P., Ovalles, C., Zhang, Y., Adams, J. J., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2019; Vol. 1320, pp 115–199".

5.1 Introduction

The extraction and transportation of Canadian oilsands bitumen has posed a long-standing challenge due to the low fluidity of bitumen. [1,2] In addition to the high viscosity of bitumen, it is also fouling in nature, like some heavy crude oils. What are the problems associated with high viscosity fluids, especially in pipeline transport? The quandary lies in the fact that high viscosity leads to elevated pressure drop compared to less viscous fluids in pipelines during transportation that escalates the pumping loads which in turn increases the cost of transportation and reduces capacity. [3]

Part of the problem is due to the presence of 'heavy' compounds called asphaltenes that constitute between 15 – 20 % (expressed as *n*-pentane insoluble material) of bitumen extracted from Canadian oilsands. [4] Asphaltenes are a solubility class of material that are precipitated from solution on addition and mixing with lighter paraffinic compounds like *n*-pentane or *n*-heptane. Though there has been no particular acceptance regarding a generalized representation of their structure in the literature, [5] they are prone to aggregation and there is a risk of subsequent precipitation under conditions that exist in pipelines due to de-pressurization and incompatibility with the some oils that make up the surrounding medium. Precipitation of asphaltenes leads to clogging due to the formation of solid deposits and is detrimental to transportation.

The aggregation susceptibility of asphaltenes can also be a significant cause for viscosity increase of bitumen. [3] All asphaltenes may not be in the aggregated state and the converse need not be true either. Any cluster of molecules existing in the nanometer to micrometer size range can be considered as an aggregate. [6] An obvious strategy to decrease the risk of precipitation is to reduce or remove the heavy asphaltene-type material through deasphalting, which has also shown to be beneficial in reducing viscosity. [7]

Pipeline specifications in North America require viscosity of the liquid to be a maximum of 0.33 Pa.s or 350 cSt at a minimum pipeline temperature of 7.5 ºC (winter temperature) and density less than 940 kg/m^3 (or $\geq$ 19 ºAPI) at 15.6 ºC. [8] It also poses restriction on the total olefin content to be $\leq$ 1 %wt. as 1-decene equivalent. This meant that a viscosity reduction of several orders of magnitude is required from a feed bitumen viscosity in the range 10 – 100 Pa.s measured at 25 ºC along with a significant decrease in density from > 1000 kg/m^3 for the bitumen feed.

Current methods to decrease viscosity after extraction include upgrading and the use of diluents. Natural gas condensate, which is a mixture of C_5+ liquids is a common diluent, as is light naphtha, but employing diluent has certain disadvantages. Due to their paraffinic nature, they increase the risk of agglomeration and precipitation of asphaltenes. Their usage is also limited by their insufficient availability of diluents at oilsands production sites and higher cost due to the installation requirement of additional pipelines for their transport and recovery. [9]

A full upgrader at the extraction site converts the bitumen to synthetic crude oil (SCO) which is quite low in sulfur and other heavy metals and is suitable for feeding into downstream refineries. But installation and operation of a full-scale upgrader is much more expensive than using diluents to make dilbit or a partial upgrader. To put things into perspective, the cost of full upgrading (including capital) would be much higher than a partial upgrader and also slightly more than dilbit production. [10] Field or partial upgrading aims to find the middle ground between bitumen dilution and full-scale upgrading. Here, the intent is to achieve sufficient viscosity reduction at the site of extraction with minimum focus on changing associated chemical properties.

Visbreaking is the lowest cost per capacity residue conversion technology [11] and therefore of interest as a potential field upgrading technology. Visbreaking, as one of the forms of thermal conversion processes apart from coking and gasification was found to be more successful than catalytic conversion due to the risk of catalyst deactivation by coke deposition and metal contamination as the thermal cracking progresses. [12] Industrially, visbreaking is conducted at temperatures of 430 – 490 ºC and residence times depending on the type of visbreaker employed. Coil visbreakers operate for shorter residence times while soaker visbreakers employ longer times. [13] The pressures are generally less than 2 MPa, with a minimum of 0.3 MPa. [14] However, the conversion based on the decrease in vacuum residue material in the feed is limited by coke formation in the industrial visbreaker. [15] Viscosity reduction rather than conversion increase is of primary importance in visbreaking for partial upgrading and it is to be noted that viscosity of complex mixtures like bitumen is not dependent on a single dominant parameter like that of molecular weight for polymers.

The severity of the reaction conditions in coking is much higher than that of visbreaking. A higher temperature of 480 – 510 ºC combined with residence times up to 24 hours and a pressure of ~0.6 MPa are typical conditions for delayed coking. Delayed coking is the most common type where the advantage is the rejection of metal content along with the carbon is highly efficient. [16]

However, the main disadvantage compared to a hydroconversion process is that the liquid yield decreases. An alternative to delayed coking is flexicoking (temperature of 870 ºC) where a fluidized bed is used to convert the heavy feed to lighter products with higher yields and the formed coke is converted to flexigas, which can be further used to product useful gases like H_2, N_2 and CO_2. [17] Fluid coking is quit similar to flexicoking but employs lower temperatures in the range 490 – 560 ºC but shorter residence times and much lower pressures (~0.1 MPa) than delayed coking.

However, the reactions as a part of this chapter were conducted over longer range of reaction times, i.e. between 15 min and 24 h and at much higher pressures of 4 MPa but at a significantly lower temperature of 400 ºC than those used in the industry. This range of reaction times comprised of both the visbreaking region where minimum coke was formed and the coking region from where the liquid yield decreased with coke formation occurring in significant amounts. Further implications of the effect of the higher pressure used is elaborated in the discussion section.

A combination of factors like its microstructure and chemical composition, molecular weight (also a consequence of chemical composition), physical and chemical interactions and phase compatibility between the different components all are known to affect viscosity. When subjected to thermal treatment, the reaction temperature and time of exposure are key parameters that control physical properties like viscosity and density. As a part of this chapter's investigation, it was also seen that procedures carried out after reaction like the type of solvent used to dissolve and extract the products from the reactor impacted viscosity significantly. Rheological conditions of viscosity measurements like shear rate and temperature used in the rheometer/viscometer also influenced viscosity of the sample in different ways.

These observations were based on certain results reported by previous researchers while conducting thermal conversion on bitumen obtained from two major geological deposits in Alberta – Cold Lake and Athabasca. Wang et al., [18] Zachariah & De Klerk [19] and Yañez & De Klerk [20] conducted thermal cracking reactions on Cold Lake bitumen in the range 150 – 400 ºC for 8 h maximum and observed a non-monotonic trend in viscosity with an initial decline and an increase at later times. The solids, that included the originally present mineral matter and the coke formed during reaction, were separated out by dissolving the products in a di-halogenated solvent, methylene chloride (CH_2Cl_2) and the liquid and solid products were separately characterized. On the contrary, Shu and Venkatesan [21] reported that viscosity decreased monotonically when Cold

Lake bitumen was thermally cracked at 300 ºC even times at excess of 24 h, but no solvent was used to extract the products. A monotonic decline in the viscosity of Athabasca bitumen was seen by Henderson and Weber [22] when subjected to thermal conversion at 371 ºC, where they did not employ a solvent for product recovery. Work by Castillo & De Klerk [23] also reported continuous decrease in viscosity wherein no solvent was used in product recovery. A comparison of some of these data are compiled in Table 5.1.

Table 5.1. Comparison of reported viscosity data in select previous thermal cracking studies on bitumen.

Henderson & Weber (Athabasca) [22]			Shu & Venkatesan (Cold Lake) [21]			Yañez & De Klerk (Cold Lake) [20]		
T [a] **(°C)**	***t*** [b] **(h)**	***μ*** [c] **(Pa.s) at 65.5 ºC**	***T*** **(°C)**	***t*** **(h)**	***μ*** **(Pa.s) at 35 ºC**	***T*** **(°C)**	***t*** **(h)**	***μ*** **(Pa.s) at 30 ºC**
371	0.00	5.242		Feed	13.8		Feed	353
371	2.85	0.33	300	6	10.9	300	1	19.64
371	9.75	0.089	300	24	7.12	300	2	1.344
371	4.00	0.536*	300	72	4.82	300	3	2.37
371	10.0	0.067*	300	120	2.22	300	4	0.37
371	24.0	0.04*	300	480	0.646	300	6	9.28

[a] temperature; [b] reaction time; [c] viscosity; * reaction conducted on bitumen not separated from sand and water.

All of these reactions were conducted in batch reactors, which explains the need for the use of an external solvent for product recovery in some of the studies. Toluene was also used as a common solvent for product recovery after thermal conversion by some researchers where other reactors such as autoclaves and column flow reactors were employed in a pilot-scale as well. [24–26] Even when advanced catalysts such as natural zeolites were employed for upgrading and viscosity reduction of Athabasca bitumen by Junaid et al., [27] toluene was employed for product extraction. A higher polyhalogenated compound, chloroform, was used to extract the liquid products from thermal cracking of a Russian bitumen.

Industrially, steam stripping is commonly used to extract the different product fractions out of the fractionator after the soaker drum. [28] A combination of high-conversion soaker cracking (HSC) with the traditional residuum oil supercritical extraction (ROSE) to enhance liquid product recovery together with deasphalted oil was suggested by Washimi and Limmer. [29]

Another notable difference (apart from reaction conditions) in the post-reaction procedure was the viscosity measurement parameters. The works that separated the solids from the liquid products used shear rates of 10 s^{-1} and reported viscosity at different temperatures in the range of 20 – 60 ^{o}C but no information on shear rates were available in the other works where a continuous viscosity decrease was observed. [21,22] Though the reasons for initial viscosity decrease and increase at later reaction times were speculated to be related to de-aggregation of the aggregated material and addition reactions in bitumen, the fundamental cause of viscosity change was not established.

The lingering question from these observations was whether the viscosity change depended just on the reaction conditions during thermal conversion or did the procedures after the reaction have an effect? Two points of concern that caught the eye were (a) the chemical nature of the solvent used to extract the products after reaction; (b) the shear rate and temperature used for viscosity measurement. Though bitumen dissolves in a number of solvents, the choice of solvent, be it polar (methylene chloride) or aromatic non-polar (toluene) becomes important. It was also shown previously through a test experiment on raw Cold Lake bitumen that the process of methylene chloride addition, mixing and subsequent removal resulted in a viscosity increase and the reason was attributed to loss of lighter boiling material during evaporation of the solvent. [18] Whether this holds true for the products in the case of Athabasca bitumen as well remained to be seen.

Differences in chemical nature and molecular weight of the solvents employed for product removal could also cause physical and chemical changes in the liquid products due to < 100 % efficiency of the evaporation process. Some solvent could still remain in the samples that can be responsible for these changes. Physical attractions like hydrogen bonding do not exist at high temperature [30] but they may be introduced or strengthened at room temperature by the presence of halogenated solvents like methylene chloride. Evidence of such type of interactions was shown by Smith [31] and similar changes were noted by Prado & De Klerk. [32] With the hypothesis that attractive forces increase viscosity, these associations can potentially affect viscosity and will be investigated in this chapter. Nevertheless, it should be recognized that when multiple forces are

active in a system, it is difficult for all of them to be disengaged in a concerted fashion even at high temperature.

Non-Newtonian behavior combined with viscoelastic effects have been shown to exist in bitumen and depend on the shear rate and temperature range of operation. [20,33,34] This can have implications during viscosity measurements and impact the resulting viscosity. The shear rate and time-dependence of viscosity of the thermally converted products will also be discussed in this chapter.

Moreover, bitumen and its thermally treated products have been shown to possess persistent free radicals that are quite stable. [35,36] Since thermal conversion of bitumen is generally accepted to follow a free radical mechanism, [37] Electron Paramagnetic Resonance (EPR) studies provides valuable insights into the reaction chemistry and was used in our study. Apart from detecting free radicals on carbon centres, EPR also gives information about paramagnetic oxo-vanadyl species, whether it is associated with porphyrin ring ligands or exists in relatively free states. [38] Quantification of organic radical spin concentration and a qualitative analysis of vanadyl EPR spectra can be used to determine the chemical environment and also related to the local viscosity changes occurring during thermal treatment. [39]

Other spectroscopic techniques like Fourier Transform infrared (FTIR) and proton nuclear magnetic resonance (^{1}H-NMR) spectra are not only regarded as useful tools for detection of hydrogen bonding [40] but also proved inevitable to be able to track the chemical changes occurring during thermal conversion. Furthermore, the chemical composition of the feed bitumen, the variation of viscosity from sample-to-sample and among different feed types will also be inspected in this work.

The primary objective of this chapter is to identify and investigate the possible factors that bitumen viscosity depends on and to develop the understanding necessary to account for the observed trends in viscosity with conversion time during thermal conversion of oilsands bitumen. Though it is recognized that both attractive and repulsive-type interactions are present at all times, their relative amounts with respect to each other are hypothesized to affect viscosity. If attractive forces overweight the repulsive interactions, it can be said that it could lead to a viscosity increase and vice versa. To realize this, most of the results provided in this chapter are related to characterization of the feed and thermally cracked products obtained from thermal conversion of Athabasca bitumen at 400 ºC and upto 24 h reaction time in a batch reactor. More importantly, the

effect of post-reaction procedures like nature of extraction solvent used and the rheological conditions of viscosity measurement like shear rate and temperature on viscosity are inspected and discussed. Comparison between changes in physical and chemical properties for the solvent-free mechanically extracted products and solvent-extracted products reveal interesting results and provide avenues for discussion.

Other aspects like solute-solvent interaction forces under the assumption that bitumen can be represented in the colloidal form, changes in the effective volume of the asphaltene clusters with progress of thermal conversion and how this could potentially affect the aggregation process of the dispersed phase were considered to have a significant impact on viscosity of a compositionally complex bitumen and are elaborated on in the discussion section, though no additional experimental data are not provided. It is the global aim of this chapter to explore and highlight the different facets of viscosity and its complicated relationship with the various parameters that can influence it.

5.2 Experimental

5.2.1 Materials

All the experiments in this study were performed on Athabasca bitumen supplied by Suncor Energy. Specific properties of the feed are provided in Table 5.2. Nitrogen (99.99%), provided by Praxair was used to pressurize and maintain an inert atmosphere during thermal reaction in a batch micro-reactor. The solvents employed to extract the product from the reactor were methylene chloride (Dichloromethane: 99.5%) and toluene (99.9%), both supplied by Fisher Scientific. The *n*-pentane (99.5%) used for precipitating asphaltenes from the bitumen and the corresponding thermally converted products was also obtained from Fisher Scientific. Acetone (99.6%), obtained from Fisher Scientific was used to clean the crystal Attenuated Total Reflectance (ATR) system attached to the FTIR spectrometer, before and after measurements.

Table 5.2. Characterization of Athabasca bitumen.

Property	Value
Viscosity at 40 ºC (Pa.s) [a]	36.2 ± 2.8
Density (kg/m^3) [a]	
At 25 ºC	1007.89 ± 0.26
At 40 ºC	998.56 ± 0.45
Refractive Index	
At 25 ºC	1.5737 ± 0.0005
At 40 ºC	1.5678 ± 0.0003
nC_5-Asphaltene content (% wt.) [b,c]	19.9 ± 1.6
Mineral matter % [d]	0.82 ± 0.18
Elemental Analysis (% wt.) [e]	
C	83.44 ± 0.01
H	10.36 ± 0.01
N	0.58 ± 0.005
S	4.80 ± 0.03
O	0.83 ± 0.01 [f]
Nature of hydrogen (from ^{1}H NMR)	
Nonaromatic (%) [g]	90.34 ± 0.01
Aromatic (%)	9.66 ± 0.01
Free radical content (spins per g) (× 10^{-17}) [h]	6.45 ± 2.19

[a] done in duplicate; [b] n-pentane insoluble; [c] done in triplicate; [d] ashing in oven at 500 ºC for 2 h; [e] mineral matter free; [f] determined by difference; [g] includes terminal methylic, mid-chain methylenic, benzylic and olefinic hydrogen; [h] determined in duplicate by EPR.

5.2.2 Equipment and Procedure

The thermal reactions were conducted in batch micro-reactors attached to a self-built tubing system. The batch reactor was manufactured using Stainless Steel Grade 316 tubing and fittings which were supplied by Swagelok Inc. A typical batch reactor is 0.086 m long from the attachment

to the tubing system to the bottom lid. The inner diameter and the tube wall-thickness together determine the pressure capacity of the reactor and in our case 0.0254 m and 0.0021 m were used, respectively that yielded a maximum allowable pressure of 18.6 MPa. For the schematic representation of a typical batch reactor, the reader is referred to Figure S1 in Yañez & De Klerk. [20] The procedure followed from reaction to product characterization is shown in Figure 5.1.

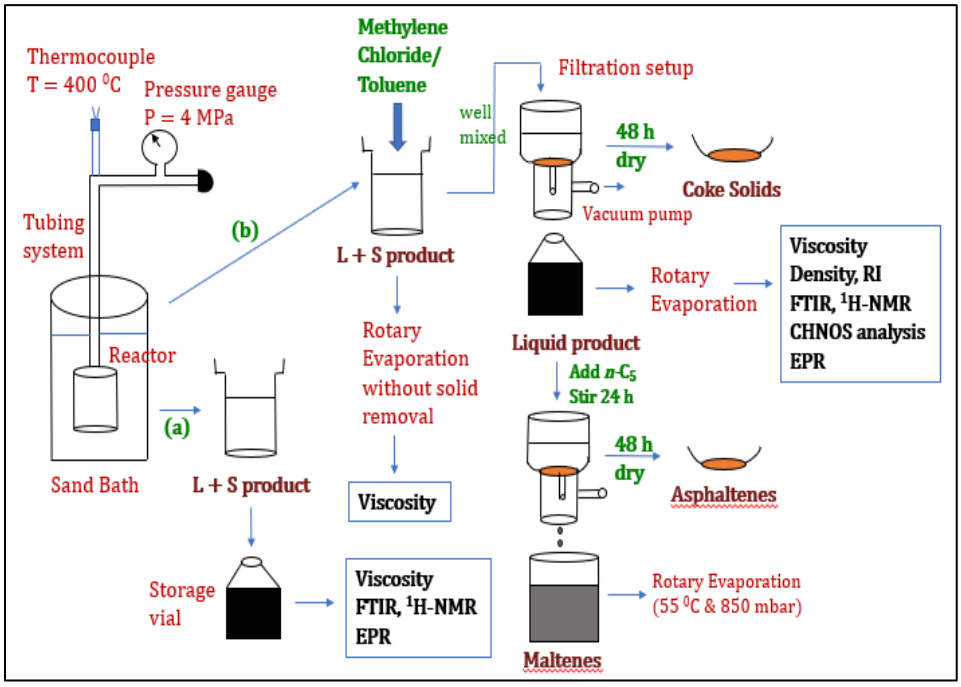


Figure 5.1. Experimental procedure and related characterization followed in this work. L – Liquid; S – Solid.

For heating purposes, a fluidized sand bath (Model: Omega Engineering FSB-3) was employed, the temperature of which was set through a temperature controller. The airflow was monitored in order to maintain a stable temperature, once the set value was attained. An 8 mL shell glass vial, supplied by Fisher Scientific, was utilized to hold the sample inside the reactor so as to avoid side reactions and product contamination due to direct contact with the reactor inner walls. To elaborate, if bitumen sample was directly filled inside the reactor, parts of the inner walls peeled off after reaction, which was not desired as their separation from the formed coke solids was very tedious. This could have occurred due to contact of sulfur present in the bitumen with the stainless-steel walls of the reactor. The reactor was pressurized to 4 MPa using nitrogen and checked for leaks using Snoop soap solution, supplied by Fisher Scientific. Industrial visbreakers usually

employed < 1 MPa but it was reported previously that coke yield was found to decrease when higher pressure was used. [41] Also, since a back-pressure regulator was not used to control the increase in pressure during the reaction and considering the maximum capacity of the batch reactor as well (18.6 MPa for the given reactor dimensions), a pressure of 4 MPa was chosen to start the reaction. Further explanation of the effect of a higher pressure in a closed system on coke yield us given in the discussion section (section 5.4).

For repeatability purposes, all reactions were conducted in duplicates and 4 reactors reactors were immersed in the fluidized sand bath heater at a time of which two were used for one reaction time and the remaining two were reacted for a different reaction time. After the system of 4 reactors was immersed in the sand bath, it took ~18 min to reach 400 ºC inside each reactor (effectively the sample temperature). During this time, the pressure increased to 6-8 MPa on an average depending on the reaction time, which was well within the pressure limit. Once the desired reaction time was reached, the reactor was removed from the sand bath and cooled down to room temperature in two steps: (i) first, blowing compressed air in order to expedite the cooling process to reach ~100 ºC (this step took ~6 min); (ii) next, dipping in water at room temperature to bring down from ~100 ºC to 25 ºC (this step took ~10 min).

Subsequently, the pressure reached 4.5 - 5.5 MPa, where the increase in pressure beyond the original pressure was due to the gases formed during the reaction which increased with reaction time, as will be shown in the results section. Stuck sand was removed from the reactor, following which the system was depressurized by collecting the gases in a gas bag. The reactor system was weighed at each stage using a Mettler Toledo ML 3002 balance (3200 g capacity with a readability of 0.01 g), for mass balance purposes. The reactor and inner parts of the tubing were cleaned using solvents (methylene chloride and toluene – according to which solvent was used as given in procedure (b) in the following paragraph) and pipe brush (acquired from Fisher Scientific), respectively.

After completion of the reaction, the product from the reactor essentially followed two paths for comparison (conducted as separate experiments for the whole product) as shown in Figure 5.1: (a) directly collected without adding any solvent; (b) extracted by adding a solvent with a day of mixing. This also includes the cleaned material mentioned in the previous paragraph. In case (a), the solid coke formed during reaction was retained in the product and the viscosity of the total product was measured. While characterizing only the liquid samples taken from the storage vial in

which solids were present, the interference of solid particles was minimized by allowing them to sediment at the bottom (except for viscosity measurements where solids were allowed to be present along with the liquids). As a separate measurement, the solids were allowed to settle for a week at room temperature after collection in the sample container and only the liquid product was taken to measure viscosity and for other characterizations as well. Case (b) constituted the following two studies: (1) the solvent-insoluble solids were separated through vacuum filtration on a 2.7 μm filter paper, following which the solvent was removed from the filtrate liquid in a rotary evaporator; (2) the solids were present and only the solvent was removed from the product by rotary evaporation. Methylene chloride was evaporated at 55 ºC and 65000 Pa while the conditions for toluene were 55 ºC and 8500 Pa, the time of evaporation being 5 h for each.

When bitumen or its thermally converted products are mixed with a lighter solvent like methylene chloride or toluene and subsequently the solvent is evaporated, some amount of mass of the system is likely to be lost either with the solvent or during transfer of material within the laboratory glassware. This was accounted for while calculating the yield as the gravimetrically measured sample weights were used for the yield calculations. However, care was taken as to not to expose the samples to air for a long time. In case (b), after the solvent was added to the sample in the reactor after each reaction, it was closed with a flexible parafilm (supplied by Bemis) and flushing nitrogen in the space between top meniscus of the solvent and the closing film to minimize oxygen exposure. This set up was left for a day after which, in the case of no solid removal, the product with sample was transferred to a conical flask from which solvent was evaporated for 5 h. During this transfer, the time of exposure to air was around 10 seconds and since the sample was at room temperature, this time of oxygen exposure would not have caused any detrimental effects like oxidation or polymerization. During evaporation of toluene, the rotary evaporator is vacuum sealed due to the very low pressure used (8.5 kPa absolute). However, during the evaporation with methylene chloride, though its boiling point at atmospheric pressure is 39.6 ºC, a pressure lower than atmospheric pressure of 65 kPa (absolute) is used to partially seal the chamber. As a test experiment, evaporation using methylene chloride was conducted under nitrogen medium as well. The characterization of products evaporated under nitrogen and partially sealed chamber yielded very similar results for all analyses conducted. In the evaluation of boiling point distribution of the samples conducted as detailed in a later section in Results (section 5.3.10), it was seen that the lighter ends were not evaporated along with methylene chloride or toluene as the distillation

profiles for the solvent-extracted products were very similar to the solvent-free ones. This meant that evaporation as such did not have an impact on sample quality and other properties. Oxidation of bitumen at these lower temperatures (~55 ºC) has not been reported to be significant in vacuum or nitrogen atmosphere. [42,43] After evaporation, the samples were allowed to cool down in under vacuum itself and immediately transferred to a cylindrical amber glass 120 mL container with a tapered mouth (obtained from Fischer Scientific). This transfer was done in the fumehood and took approximately 2 minutes for viscous samples and one minute or lesser for samples of lower viscosity obtained at higher reaction times. For the directly extracted samples without solvent, the products were directly transferred to the 120 mL amber glass containers in the fumehood and this took an average time of 4 minutes depending on the viscosity and amount of coke formed. All samples, whether solvent-extracted or solvent-free, were stored under nitrogen in the amber vials.

The mass loss on evaporation was calculated as follows: (i) the weight of the flask containing the solvent and extracted sample was measured before evaporation; (ii) the amount of solvent evaporated was collected and weighed; (iii) the amount of sample that should be remaining in the flask was calculated by subtracting the weight of the solvent evaporated from the total weight in (i); (iv) the actual amount of sample remaining was measured by weighing the solvent-evaporated flask. It must be noted that this sample also contained a small amount of unevaporated solvent (quantified to be ~1 %wt. in the FTIR section), so the amount of actual sample would have been slightly lesser.

5.2.3 Analyses

5.2.3.1 Viscosity

Viscosity measurements were carried out using an Anton Paar RheolabQC viscometer for constant shear rates and an Anton Paar MCR 102 rheometer for variable shear rates. Viscometers and rheometers generally operate on the principle of application of a shear force on the fluid and calculating the viscosity from the resulting torque generated. The instruments come in different geometries depending on the application required. In our work, a concentric cylinder arrangement consisting of a cup and bob with the fluid filled in the cup upto a certain level is used for viscosity measurements. A review of the different types of rheometers and viscometers used for viscosity

measurements of bitumen and heavy oils and some of their key findings are given in section 5.4.3.3 in the Discussion of this chapter.

In the viscometer, a concentric cylinder measuring cup (CC17/QC-LTC) was maintained at 40 ± 0.1 ºC for all the measurements. The system consists of a cup (outer cylinder) and bob (inner cylinder) arrangement and the fluid fills the annular space between them (Figure A3 in Appendix). The bob was 0.02497 m long with a diameter of 0.01664 m. The inner diameter of the cup with upto where the fluid was in contact was 0.01700 m, thus the annular space being 0.00036 m wide. It follows the Searle system where the bob is rotated by a motor and the cup is kept stationery. Further details regarding torque calculations are given in section C.3 of Appendix C. A Julabo F25-EH circulating heater/chiller was employed for the temperature control and took ~20 min to reach the set temperature before each run. About 4 g of sample was required for accurate viscosity measurement. The shear applied was varied between 0.1 s^{-1} and 1544 s^{-1} depending on the sample and care was taken to maintain the torque always above the lower measuring limit of the instrument. The viscometer was calibrated using a Newtonian viscosity standard (mixture of hydrocarbons), which was traceable to the national standard of viscosity.

The rheometer (Modular compact rheometer - MCR 102) worked on the same principle as the viscometer but it required four times more sample for each run. The main component was a pressure cell (CC25/Pr150/ln/A1/SS with a maximum sustainable pressure of 15 MPa and a maximum operating temperature of 300 ºC), consisting of a measuring cup and a measuring cylinder (similar to the bob in viscometer) that was attached to a pressure head over which the magnetic coupling rested. Here also, the bob was rotated by the magnetic coupling attached to a motor with the outer cup being stationery. The bob was 0.0420 m long and 0.0245 m wide while the inner diameter of the measuring cup was 0.0265 m. The annular space was wider (0.0020 m) and thus could hold more sample as compared to the viscometer. The temperature in the cup was controlled by a heater in a flange ring affixed around the measuring cup, which was cooled down by compressed air after each measurement. Minimum torque required for reliable measurements in the viscometer was 0.26 mNm but a much lower limit existed for the rheometer.

5.2.3.2 Density and Refractive Index

Densities of the samples were measured in an Anton Paar DMA 4500M densitimeter. It consists of a U-tube that oscillates at a characteristic resonant frequency and since the volume is fixed, the density is calculated from the resonant frequency, which is inversely proportional to the square root of its mass. A U-View camera displays live images of the U-tube and the sample filling as well. The instrument is also equipped with bubble detection technology during measurement and has the advantage of minimum temperature fluctuations. Viscosity-related errors are corrected automatically.

Measurements were conducted at 25 ºC and 40 ºC with temperature accuracy of 0.01 ºC. The density values were accurate to 2 decimals (0.01 kg/m^3) for the results in this work. Ultra-pure distilled water (provided by Anton Paar) was used for calibrating the instrument at 20 ºC.

An Anton Paar Abbemat 200 refractometer was used to obtain the refractive indices of the bitumen samples. The reference was air and measurements were made using sodium D-line at 589 nm. The accuracy of the refractive index values and temperature were 0.0001 and 0.01 ºC, respectively. All measurements were conducted at 25 ºC as well as at 40 ºC. The equipment was factory calibrated with official standards from the National Metrology Institute of Germany.

5.2.3.3 Free radical concentration

Electron paramagnetic resonance (EPR) technique was utilized to detect free radicals in the samples for this study. The spectra were obtained using an Active Spectrum benchtop EPR spectrometer, operating at 9.7 GHz (X-band microwave frequency range). At this constant frequency, in the presence of an external magnetic field, the energy levels of parallel and anti-parallel electron magnetic moments are split by a value that is proportional to the strength of the field, with the constant of proportionality being the *g*-factor, which is characteristic for each paramagnetic centre. It is important to keep in mind that a much lower energy is required for EPR as compared to techniques like UV-Vis spectroscopy and fluorescence spectroscopy since only electron spins are excited in EPR, whereas the whole electron is excited to the next electronic energy level in UV-Vis and other spectral techniques where the electromagnetic radiation interacts with the whole electrons in the shells of the atoms. However, EPR requires a higher energy than NMR due to lower mass of electron as compared to nuclei.

For qualitative and quantitative measurements, 30 %wt. solution of the sample in toluene were made and 0.7 mL of this solution was taken in 5 mm outer diameter (OD) EPR tubes (701PQ7 – 7" long), supplied by Wilmad-LabGlass. The spectra of pure sample (without dissolution in toluene) were also compared among different products for qualitative analysis.

Calibration for quantitative measurements was performed by preparing 0.1-16 mM DPPH (2, 2-Diphenyl-1-picrylhydrazyl) solutions in toluene. Double integrated intensity (DII) values for the organic radical peaks in each spectrum were converted to number of spins/gram (N_g) of the sample using the concentration of the DPPH solution, avagadro number and density of toluene. From the calibration equation that related N_g with DII, the number of spins in the unknown sample was determined from its measured DII value.

In order to calculate the number of spins per gram of the unknown analyte, the following procedure was followed: (i) The EPR spectra of the unknown sample was obtained; (ii) The area under the absorption curve was calculated through double integration (DII); (iii) This value of DII was input into the calibration equation given in Figure C.1 in Appendix C to get the number of spins per gram of the solution; (iv) This value was divided by the concentration of the unknown analyte (in % wt.) to obtain the solvent-free number of spins per gram of the unknown analyte. Different concentrations of the unknown samples were prepared (5 %wt. to 50 %wt.) and it was determined that 30 %wt. solutions of thermally converted products in toluene gave noise-free spectra and at the same time was not too high a concentration where the density approximation (given in section C.1 of Appendix C) would not be valid. This procedure makes comparison between different thermally converted samples possible by bringing all quantified spin concentrations to the same scale.

The following spectral parameters were used in the EPR: sweep magnetic field range - 3200-3600 G; microwave power - 15 mW; modulation coil amplitude - 1.2 G; number of scans - 10; number of points for resolution - 4096; digital gain - 12 dB.

5.2.3.4 FTIR spectra

This is a type of vibrational spectroscopy where the sample molecule absorbs a photon of particular energy from light of wavelength in the near-infrared to mid-infrared range such that the molecule is excited to a higher vibrational energy level. It is required that the dipole moment of

the molecule changes during vibration for it to be IR active. This is a complementary technique to Raman spectroscopy where polarizability of the molecule plays a role.

FTIR spectra were collected using an ABB MB3000 spectrometer. It consisted of a single reflection PIKE MIRacle^{TM} ATR accessory with a diamond crystal plate supplied by PIKE Technologies. Acetone was used to clean the crystal after each measurement and complete drying was ensured before obtaining each spectrum. The spectral parameters used were: reference - air; resolution - 4 cm^{-1}; number of scans - 120; wavenumber range - 4000-600 cm^{-1}; detector gain - 81 dB; mode - transmittance.

5.2.3.5 ^{1}H NMR spectra

This is a technique that is used to study the different types of protons present in a molecule. It can detect aliphatic methyl, methylene, and benzylic hydrogen, olefinic as well as aromatic hydrogen. In the presence of an external magnetic field, nuclei that have non-zero nuclear spin are differentiated based on their chemical environment. Each type of hydrogen has a respective chemical shift value that is calculated such that it is made independent of the instrument frequency. Aromatic hydrogen is said to be more de-shielded than aliphatic hydrogen, and this de-shielding effect commonly occurs due to electron withdrawing effect of electronegative atoms that move the electron density away from the protons thus minimizing the strength of the local opposing magnetic field created due to the presence of electrons.

^{1}H-NMR spectroscopy was performed in a Nanalysis 60 MHz NMReady-60 spectrometer, pre-calibrated with chloroform-D (Deuterated Chloroform). The nature of hydrogen in the bituminous species were determined through this method. Solutions made by dissolving ~130 mg of the sample in 0.7 mL of Chloroform-D were poured into a 5 mm OD Norell^{R} Standard Series NMR tube (supplied by Sigma Aldrich) for each run. The spectral parameters adopted were: spectral width (δ) = 12 ppm; Number of scans/sample = 64; time per scan = 23.4 seconds; sufficient time delay between scans was allowed for magnetization recovery of protons.

5.2.3.6 Elemental analysis

This technique is used for detection of major elements present in the sample, namely carbon, hydrogen, and the heteroatoms including sulfur and nitrogen. The sample is flash combusted in excess of oxygen after which the evolved gases (CO_2, H_2O, NO and SO_x) are collected and their masses noted in order to back-calculate the chemical composition of the sample. In most cases, the results are presented in terms of ratio of the elements in the sample.

In our work, the liquid products (solid allowed to settle in case of 'a' and 'b_1' shown in Figure 5.1) and the feed bitumen were analysed for C, H, S, N and O content. The elemental analysis was conducted using a Thermo Scientific Flash 2000 CHNS-O organic elemental analyzer which required only a minute amount of the sample. Oxygen content was determined by difference.

5.2.3.7 Asphaltenes content

Asphaltenes are a solubility class that can be measured by mixing the bitumen or crude oil sample in an anti-solvent such as *n*-pentane or *n*-heptane. Ancheyta et al. [44] compared the properties of asphaltenes extracted from different heavy crude oils using both solvents and suggested that extraction with *n*-pentane was more complete and enables recovery of more asphaltenic material compared to *n*-heptane. The difference in the asphaltene content is attributed to the presence of resin-type material and low molecular weight asphaltenes. [45] Patino et al. [46] separated C_5 and C_7 asphaltenes from a South American heavy oil and found abundance of island-type structures in C_5 asphaltenes and archipelago-type structures in C_7 asphaltenes. This seemed in contrast to the result by Ancheyta et al. [44] where they found that C_7 asphaltenes exhibited a higher aromaticity with a lower H/C ratio and higher N/C, S/C and O/C contents as compared to C_5 asphaltenes derived from Maya crude oil. This means that the properties of these asphaltenes were sample dependent.

The ASTM D6560 standard test method [47,48] recommends the use of *n*-heptane as the anti-solvent at a solvent to oil ratio of 30:1 for precipitating asphaltenes. In this work, a slightly modified method based on the ASTM D2007 standard test method [49] was employed by using *n*-pentane instead of *n*-heptane at a higher ratio of 40:1. There are other factors that are considered when choosing the solvent for deasphalting in industries as given by Speight. [47] Liquid propane is commonly used for lighter oils but for heavier feedstocks, *n*-butane and iso-butane are more suitable due to their higher critical temperature. It is also observed that yield of asphaltenes

decreased and molecular weight increased when solvents with higher carbon number are used to for deasphalting. [4,50] Also, a ratio of *n*-pentane to bitumen higher than 30:1 yielded a constant asphaltenes content. [48] Keeping all these factors in mind and since the industrial sponsor of our research, CNOOC Inc., uses *n*-pentane in its BituMax Partial Upgrading Technology process, *n*-pentane with a solvent to oil ratio of 40:1 were chosen for separation of asphaltenes in our work. [47]

The asphaltene content was determined for the liquid products after separation of the toluene-insoluble and methylene chloride-insoluble coke solids (case 'b_2' shown in Figure 5.1). It was important that no solids interfered with the determination process. This procedure is also depicted in Figure 5.1. Typically, a volume ratio of 40:1 between *n*-pentane and the sample was mixed in an Erlenmeyer (conical) flask obtained from Fisher and magnetically stirred for 24 h. Next, the non-homogeneous mixture was then vacuum filtered through a 2.5 µm Whatman filter (Grade 42 obtained from Sigma Aldrich) and the retained asphaltenes were dried in a vacuum oven at 60 ºC. *n*-pentane from the filtrate was evaporated at 55 ºC and atmospheric pressure in a rotary evaporator and the maltenes were collected.

5.2.3.8 Microcarbon Residue (MCR) content

This test is used for determining the quantity of carbonaceous residue after the sample is subjected to pyrolysis and evaporation of the lighter and heavy end components over a wide temperature range. In other words, it is a measure of the coke formation tendency of the sample. It is observed that this method produces very similar results as compared to the Conradson Carbon Residue (CCR) test as mentioned in ASTM D189 method. [51] As detailed by Noel, [52] CCR determination has some disadvantages: (i) large amount of sample required (5 – 10 g) as opposed to milligrams of sample in the MCR; (ii) it is conducted under open conditions with evolution of smoke.

In this work, a Mettler Toledo TGA/DSC1 equipped with a LF1100 furnace, sample robot for carrying the crucible with sample to the combustion chamber, and a MX5 internal microbalance for measuring weights of the sample and reference. STARe System software was used for data retrieval and processing. Around 15 mg of the sample was taken in a 70 µL alumina crucible and temperature was increased form 25 ºC to 600 ºC at a rate of 10 ºC/min. A N_2 flow of 50 mL/min

controlled with Mettler GC10 gas controller was employed as the inert atmosphere. The MCR content was calculated as the amount of material remaining with respect to the initial sample after the temperature ramp was applied.

5.2.3.9 Simulated Distillation

This is a quick method to determine the boiling ranges of different distillation fractions of a sample and uses a GC separation method. A sample is separated into fractions of different boiling ranges by passing through a column and are subsequently correlated with their retention time by developing a calibration curve using a mixture of *n*-alkane hydrocarbons covering the entire boiling range. The output can be viewed in the form of boiling point vs. volume of sample eluted/distilled from the column.

Specifically, an Agilent 7890B high temperature gas chromatograph was used to determining the distillation profile of the feed and thermally converted products to a temperature equivalent of eluting ~720 ºC boiling point material. The ASTM D7169 standard test method was followed. [53] The GC was equipped with a flame ionization detector (FID). For sample preparation, 100 mg of the liquid was dissolved in 10 mL CS_2 with carefully recorded gravimetric measurements. Polywax655 (supplied by Agilent) and 5010 standard reference material (obtained from Supelco) were used for calibrating the boiling point and the FID response, respectively. The peaks for each carbon number in Polywax655 are distinct to help with the retention time calculation. The 5010 reference standard is a mixture of heavy hydrocarbons that give overlapped peaks and the area under the peak is used to calculate the response factor. Blanks were run at the start of a sequence and after each sample or standard injection and subtracted to ensure minimum carryover from previous runs.

5.2.3.10 Analysis of gaseous products with Gas Chromatography

This technique is used for identifying components present in gas samples. The instrument consists of a small sized-column through which the sample is carried by means of a gas-phase inert carrier and the column consists of a specific stationary phase with which the gas sample interacts. Different components get adsorbed at different stages, thus resulting in different retention times

inside the column. The difference between other common techniques like liquid chromatography (LC) is that in GC, the mobile phase is a gas and in LC, the mobile phase is a liquid.

In our work, an Agilent 7890A gas chromatograph combined with a flame ionization detector (FID) as well as a thermal conductivity detector (TCD) was utilized to analyze the composition of the gas phase from the thermal conversion process in this study. The FID detected hydrocarbons (mostly C_1-C_6) while the TCD was responsible for detecting other gases like CO_2, H_2S, CO, N_2 and Ar. The gases were injected into a HayeSep R column (2.44 m × 0.003 m) to perform the separation. Helium at a constant flowrate of 25 mL/min was employed as the carrier gas and the injector temperature was set at 200 ºC. The following are the details for the temperature method exercised to evaluate the gaseous products. The initial temperature was 70 ºC, held isothermally for 7 min after which it was increased to 250 ºC at the rate of 10 ºC/min and held at that temperature for 2 min once reached. Finally, the temperature was decreased at the rate of 30 ºC/min till 70 ºC was attained and was maintained constant for 8 min.

5.2.3.11 Optical Microscopy

This is a very simple technique where solid or liquid samples are analyzed under a microscope to observe the visible characteristics of a sample like its shape, size and color. Magnification can be achieved till the micro-scale level.

Optically zoomed images of the coke solids formed during various stages of thermal conversion were taken using a Carl Zeiss SteREO Discovery V20 microscope.

5.3 Results

5.3.1 Product yield from thermal conversion

In this section, the total product yields are shown for two sets of reactions performed at the same conditions (400 ºC, 4 MPa), but with different product work-up procedures.

5.3.1.1 Product work-up without solvent

Table 5.3 shows the yields for gaseous products and liquids + solids products together as in these reactions, the solids were not removed from the products, which corresponds to case 'a' in Figure 5.1. The yields are calculated as a percentage of the total feed in each case. All reactions were performed in duplicates.

Table 5.3. Yields of gaseous and liquid products retaining the solids mechanically removed without the aid of an extraction solvent.

Reaction time (min)	Gas %wt. [a]	SD Gas %wt. [b]	Liquid + Solids %wt. [a]	SD Liquid + Solids %wt. [b]	Total %wt.
15	1.29	0.07	95.64	1.62	96.93
30	2.44	0.21	94.90	0.46	97.34
45	3.59	0.93	93.68	0.84	97.26
60	6.47	0.21	90.27	0.04	96.74
75	6.77	0.30	89.55	2.29	96.33
90	9.87	0.12	87.04	0.46	96.92
120	9.05	0.95	86.04	3.96	95.09
135	9.16	0.68	88.52	0.04	97.69
150	9.94	0.06	88.83	1.17	98.77
180	10.56	0.13	83.28	0.15	93.84
210	11.28	0.30	85.73	2.08	97.01
240	12.65	0.17	81.63	2.98	94.29
360	14.00	0.63	79.11	4.72	93.11
480	15.00	0.19	79.36	0.88	94.36
1170	14.72	1.13	78.49	5.32	93.21
1440	16.03	0.44	80.02	2.38	96.05

[a] averaged value over two experiments for each reaction time;

[b] standard deviation of the duplicates at each reaction time.

5.3.1.2 Product work-up with methylene chloride

The material balance for the case where methylene chloride was used to extract the products from the reactor and evaporated without separation of solids (pathway 'b_1' in Figure 5.1) is shown

in Table 5.4. Also, Table 5.5 shows the yields for liquid-solid separation of products for all instances where solids were separated from the solvent recovered samples.

Table 5.4. Product yield when methylene chloride was used to extract liquid and solid products together without solids removal.

Reaction time (min)	Gas %wt. [a]	SD Gas %wt. [b]	Liquid + Solids %wt. [a]	SD Liquid + Solids %wt. [b]	Total %wt.
15	1.23	0.01	99.57	0.56	100.79
30	1.85	0.65	98.32	0.32	100.18
45	3.55	0.37	96.04	0.91	99.59
60	6.54	0.03	91.74	0.06	98.29
75	7.08	0.35	93.05	1.90	100.13
90	8.45	0.57	89.33	0.92	97.79
120	8.96	1.17	91.06	2.45	100.02
135	9.47	0.62	89.13	0.20	98.60
150	9.94	0.19	90.13	0.87	100.07
180	10.97	0.19	85.18	0.76	96.79
210	11.23	0.87	88.64	0.87	99.87
240	12.65	0.17	85.45	0.19	98.11
360	14.06	0.33	84.36	1.33	98.42
1170	14.72	1.13	84.25	1.41	98.98
1440	16.03	0.55	81.58	2.05	97.61

[a] averaged value over two experiments for each reaction time;

[b] standard deviation of the duplicates at each reaction time.

Table 5.5. Yields of solid-separated products with methylene chloride as the extraction solvent.

Reaction time(min)	Liquid %wt. [a]	SD Liquid %wt. [b]	Solid %wt. [a]	SD Solid %wt. [b]	Total %wt. of Liquids + Solids
15	96.93	0.63	0.92	0.18	97.85
30	96.34	1.16	1.16	0.04	97.50
45	93.53	0.55	2.21	0.12	95.74
60	88.27	0.93	2.69	0.43	90.96
75	87.28	1.08	3.80	0.56	91.08
90	84.27	0.86	4.81	0.31	89.08
120	84.01	3.00	6.20	0.88	90.21
135	81.98	1.41	6.41	0.11	88.39
150	82.31	0.45	6.65	0.35	88.96
180	77.99	0.92	7.44	0.32	85.43
210	78.28	0.15	8.60	0.80	86.88
240	77.12	1.21	7.73	1.11	84.85
360	74.98	0.47	8.77	0.38	83.75
1170	68.20	0.82	11.71	0.73	79.91
1440	68.78	1.60	12.27	0.91	81.05

[a] averaged value over two experiments for each reaction time;

[b] standard deviation of the duplicates at each reaction time.

5.3.1.3 Product work-up with toluene

The material balance tables for the products separated with toluene as the solvent without solids removal is given in Table 5.6. Table 5.7 provides the data for the case where solids were separated from the liquid products. They were quite similar to the results obtained with methylene chloride as the extraction solvent except that in certain cases, the solids content was marginally higher for the toluene-extracted products (15 min, 90 min in Table 5.5 and Table 5.7).

Table 5.6. Product yield when toluene was used to extract liquid and solid products together without solids removal.

Reaction time (min)	Gas %wt. [a]	SD Gas %wt. [b]	Liquid + Solids %wt. [a]	SD Liquid + Solids %wt. [b]	Total %wt.
15	1.21	0.02	99.64	1.02	100.85
30	1.86	0.55	97.77	0.41	99.63
45	3.50	0.41	96.14	0.88	99.64
60	6.56	0.09	92.25	0.54	98.81
75	7.09	0.29	93.09	0.87	100.18
90	8.41	0.63	90.12	0.82	98.53
120	8.96	1.25	91.25	2.02	100.21
135	9.51	0.69	89.24	0.24	98.75
150	9.90	0.22	90.04	1.24	99.94
180	11.02	0.29	85.24	0.83	96.26
210	11.22	0.88	87.62	1.52	98.83
240	12.71	0.16	85.55	0.55	98.26
360	14.14	0.43	84.31	1.21	98.45
480	15.06	0.29	82.72	0.76	97.78
1170	14.65	1.20	85.78	1.12	100.43
1440	16.44	0.65	83.68	1.56	100.12

[a] averaged value over two experiments for each reaction time;

[b] standard deviation of the duplicates at each reaction time.

Table 5.7. Yields of solid-separated products with toluene as the extraction solvent.

Reaction time(min)	Liquid %wt. [a]	SD Liquid %wt. [b]	Solid %wt. [a]	SD Solid %wt. [b]	Total %wt. of Liquids + Solids
15	96.94	0.98	0.98	0.29	97.92
30	96.44	0.55	1.21	0.21	97.65
45	93.58	0.98	2.32	0.25	95.90
60	88.29	0.93	2.71	0.54	91.00
75	87.22	1.45	3.87	0.32	91.09
90	84.32	1.10	4.98	0.11	89.30
120	84.14	2.25	6.24	1.02	90.38
135	82.09	1.42	6.45	1.11	88.54
150	82.22	0.76	6.72	0.34	88.94
180	78.54	0.87	7.52	0.98	86.06
210	78.34	0.35	8.66	1.04	87.00
240	77.19	1.22	7.84	0.54	85.03
360	75.04	0.54	9.01	0.76	84.05
480	71.87	0.32	9.22	0.35	81.09
1170	68.22	0.67	11.75	0.98	79.97
1440	68.89	1.21	12.34	0.66	81.23

[a] averaged value over two experiments for each reaction time;

[b] standard deviation of the duplicates at each reaction time.

5.3.1.4 Observations about product yield

As can be seen from Table 5.3, the overall mass balance was between 95-98 % since a complete recovery of the product samples were not possible due to absence of a solvent. At lower reaction times, the product was not fluid enough to flow out on its own from the reactor, whereas at higher reaction times, some formed solids were noticed to be stuck to the reactor walls but the liquid was quite fluid. The SD values for this procedure (each reaction performed in duplicate) were relatively higher for certain reaction times like 15 min, 75 min, 120 min, 240 min, 360 min

and 1170 min, which were all after coke started forming in significantly amounts (> 45 min) except for the product at 15 min. This meant that the formation of coke hindered the product removal ability to an extent but in all cases, more than 88 %wt. of the products was recovered (Table 5.3).

Though peculiar results were obtained when methylene chloride was used as the solvent (as is the focus of this study), the fact that nature of the products extracted with toluene were very similar to the mechanically extracted products indicated that the loss of material in the reactor (2 – 7 %wt. – Table 5.3) did not cause a change in composition that would have an effect on the results. The mass balance was offset the maximum of ~7% for the 360 min and 1170 min products (Table 5.3) when the product was mechanically extracted. But, the mass balance for these products was more than 98 %wt. when extracted with solvent with or without solids removal. Some other sources of error during product recovery were: (i) further separation of solids resulted in a minor loss of liquid product of within ~0.5 %wt. (comparing Table 5.4 and Table 5.5 for methylene chloride, Table 5.6 and Table 5.7 for toluene) due to attachment to filter paper. However, these values were accounted for due to gravimetric measurement and was not found to impact the results in a major way; (ii) 100 % evaporation of solvent was not possible as detailed below; (iii) small amount of material being stuck to the bottom of the conical flask after solvent evaporation and transfer to storage container. This occurred only for some of the more viscous products at lower reaction times (< 90 min) and was accounted for in the material balance as gravimetric measurements were conducted at every stage.

The total mass balance in the approach where a solvent (either methylene chloride or toluene - Table 5.4 & Table 5.6) was utilized for product recovery was ~1 – 2 % higher than when the products were collected without the solvent and solids retained. In some of the instances, the total mass balance was above 100 %wt. (15 min, 30 min, 2 h, 2.5 h in Table 5.4, 15 min, 75 min, 1170 min in Table 5.6) which gave indications that some solvent remained in the sample and could not be fully evaporated. In fact, even though other mass balances were below 100 %wt., some solvent remained in all samples and formed the basis for the arguments developed in this study and was quantified with FTIR spectra, detailed later. As evident from Table 5.3, the gas yields followed a steady increasing trend for the reaction times up to 24 h (16 % gaseous product) with acceptable standard deviations. The liquid + solid yield decreased from 95 – 80 %wt. over 24 h of reaction time with the highest variance in the values at 19.5 h. Comparing these results with Table 5.4 and Table 5.6, the total liquid + solids content is observed to decrease from 98 % to 81 %, the values

being 1 – 3 %wt. higher than for the previous case discussed. The overall decrease in the non-gaseous products with increase in reaction time is due to the faster decline in the liquid content in comparison to the increase in solids formed. The liquids content decreased from 97 %wt. at 15 min to ~68 %wt. at 24 h, whereas the difference between the maximum and minimum solids content was only ~11 %wt. and the reminder of the difference was due to the formation of gas.

Feed Athabasca bitumen comprised of 0.82 %wt. solids, which are predominantly mineral matter (Table 5.2). There was minimal increment in the first 30 min of reaction, typically due to solids formation on the mineral matter, [19] but a prominent increase in the solids content was observed at 45 min reaction time, where the amount of coke formed was almost doubled as compared to that at 30 min (Table 5.5). It should be noted that the solids were measured gravimetrically so can be considered robust in its measurement. The higher variation in solids content at longer reaction times (for example SD of 1.11 %wt. at 240 min in Table 5.5) can be attributed to phase behavioral changes in bitumen, although some amount of experimental error cannot be excluded. There was also a noticeable dip in the liquid content at this time (2.81 % decrease from 96.34 %wt. at 30 min to 93.54 %wt. at 45 min as compared with 0.59 % decrease at 30 min from 15 min).

Standard deviations for the composition of solids, liquids and gaseous products were low, which indicated that a meaningful statistical comparison could be made with different reaction times.

5.3.2 Viscosity

5.3.2.1 Viscosity of thermally converted products with reaction time: No solvent used

Feed Athabasca bitumen was found to have a viscosity of 36.2 Pa.s measured at 40 ºC (Table 5.2). The variability in the feed viscosities depending on their source types was relevant to this study and a separate section is dedicated for its discussion.

As already deliberated in the Introduction, both monotonic and non-monotonic trends in product viscosity during thermal conversion at 300 ºC were observed with Cold Lake bitumen as the feed. [20,21] On the other hand, a constant decrease in viscosity was observed by for Athabasca bitumen [22] but it should be noted that only 3 reaction times at large intervals were reported (0, 2.85

and 9.75 h). However, in this study the trends in the viscosities of the thermally converted products at 400 ºC were investigated at shorter time intervals from 15 min to 24 h and depicted in Figure 5.2. The viscosity values correspond to the liquid products with the formed coke present (case 'a') in Figure 5.1. The products collected directly without the use of a solvent will be referred to as 'straight' thermally converted products in the rest of the chapter.

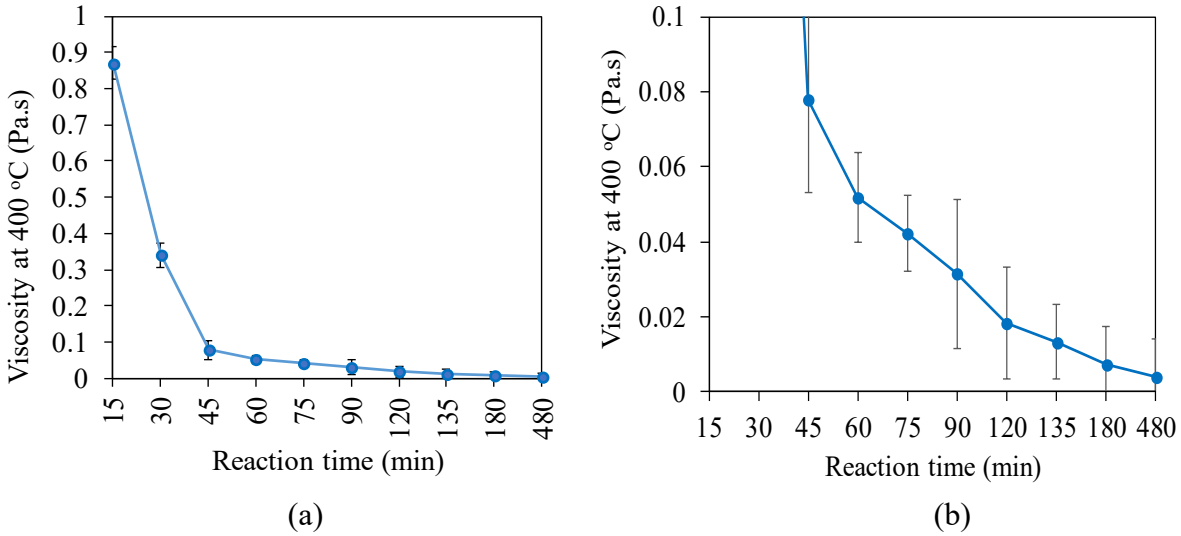


Figure 5.2. (a) Product viscosity after thermal conversion at 400 ºC for the indicated reaction times. All viscosities were measured at 40 ºC; (b) The same data is shown with a zoomed y-axis to visualize the graphic details. The shear rates at which all the viscosities were measured are given in Table 5.8.

It can be seen that the viscosity decreased monotonically for the entire range of reaction times. The reduction in viscosity was rapid, with a decrease of 97.6% from the feed when thermally converted for 15 min. The rate of viscosity decrease subsequently dropped at higher reaction times but did not seem to level out. This is evidenced by the diminishing slope of the straight lines connecting two data points moving from left to right in Figure 5.2 and also shown in Figure 5.3. The liquid products were quite fluid but recovery without solvent was made difficult once coke started forming in significant amounts from 45 min (Table 5.5).

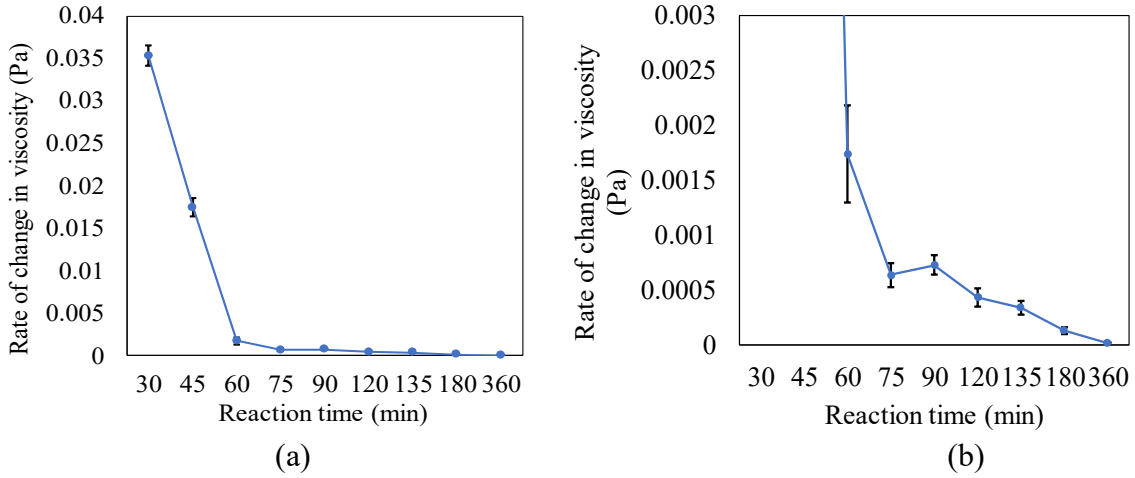


Figure 5.3. (a) The rate of change of viscosity with respect to reaction time for the thermally converted samples; (b) The same data is shown with a zoomed y-axis to visualize the graphic details.

With a low standard deviation, the decreasing monotonic trend in thermally converted product viscosity is evident. The specific viscosity values and the shear rates at which they were measured are given in Table 5.8. These values were obtained in adherence to the lower device limit of torque for the viscometer. The effect of shear rates on viscosity is elaborated in section 5.4.3.3 in the discussion.

Table 5.8. Minimum shear rates required and actual shear rates employed for viscosity measurement of the thermally converted products extracted without the aid of a solvent.

Reaction time (min)	Viscosity (Pa.s)	Minimum shear rate required (s^{-1})	Actual shear rate employed (s^{-1})
15	0.870 ± 0.045	30	35
30	0.340 ± 0.032	150	150
45	0.078 ± 0.009	310	320
60	0.052 ± 0.006	500	500
75	0.042 ± 0.004	550	550
90	0.0314 ± 0.005	600	600
120	0.0184 ± 0.003	1000	1000
135	0.0133 ± 0.002	1500	1544 (Max. possible)
180	0.0073 ± 0.002	1500	1544 (Max. possible)
360	0.004 ± 0.001	1500	1544 (Max. possible)
480	0.003 ± 0.001	1500	1544 (Max. possible)

5.3.2.2 Viscosity of thermally converted products with reaction time when methylene chloride was used as extraction solvent

The viscosity values reported in this section belong to the thermally converted products obtained from step 'b_1' in Figure 5.1. As described in the experimental section 5.2.2, methylene chloride was added as an extraction solvent to retrieve the products inside the reactor and evaporated afterwards. The solids were not separated from the liquid products. The viscosities of the liquid products with the retained solids are shown in Figure 5.4.

The specific viscosity values and the shear rates at which they were measured for the methylene-chloride separated products are given in Table 5.9. It can be seen that the shear rates vary according to the viscosity of the sample as was for the solvent-free products.

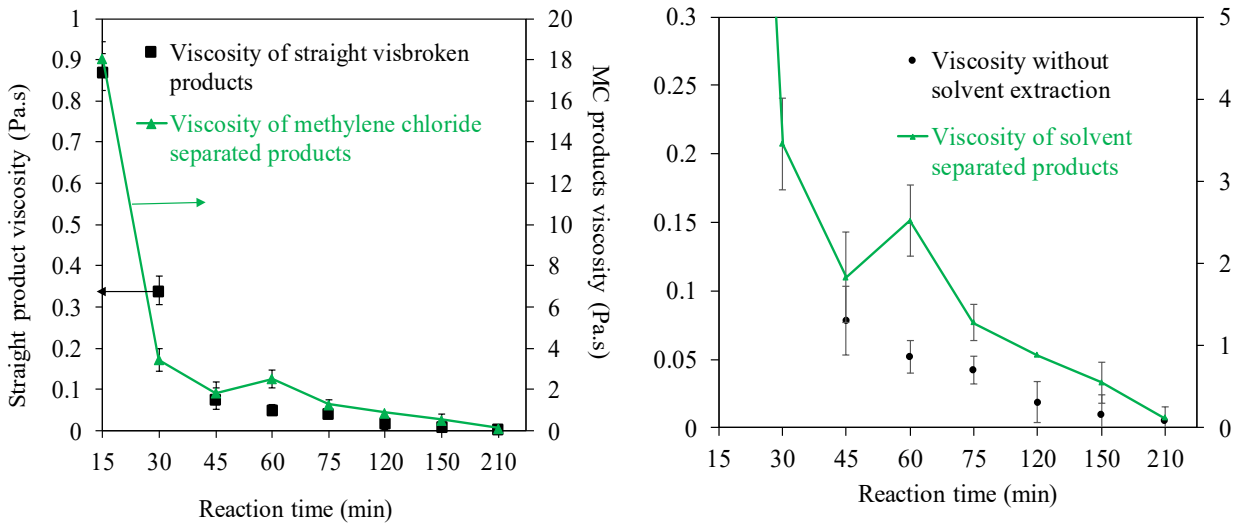


Figure 5.4. (a) Comparison of the viscosity of straight thermally converted products with the viscosity of those extracted with methylene chloride (MC products viscosity); (b) The same data is shown with a zoomed y-axis to visualize the graphic details.

Table 5.9. Actual shear rates employed for viscosity measurement of the thermally converted products extracted with methylene chloride.

Reaction time (min)	**Viscosity (Pa.s)**	**Actual shear rate employed (s^{-1})**
15	18.05 ± 0.83	5
30	3.45 ± 0.56	10
45	1.83 ± 0.55	10
60	2.52 ± 0.44	10
75	1.28 ± 0.22	20
120	0.88 ± 0.11	30
150	0.55 ± 0.09	43
240	0.08 ± 0.01	300

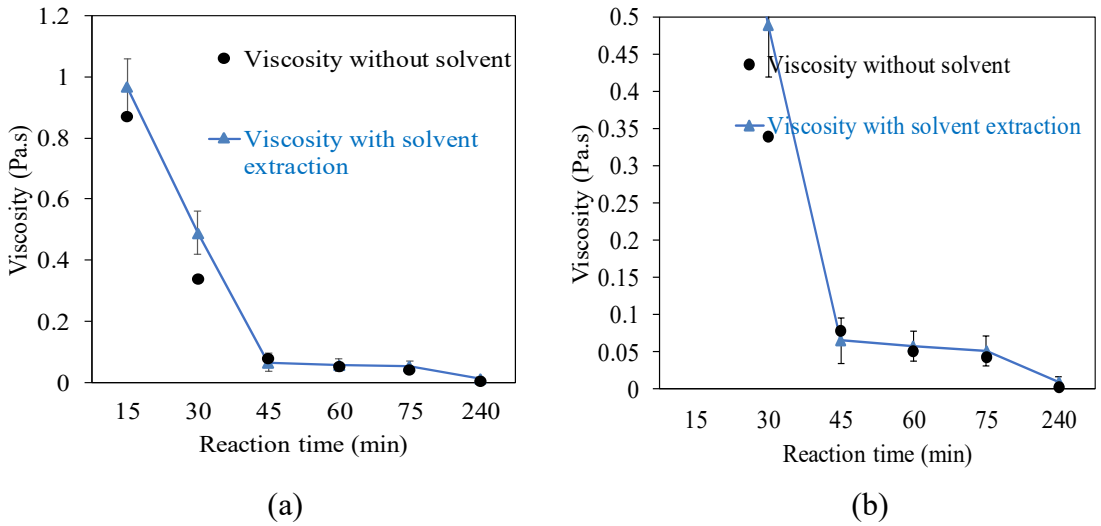
At each reaction time, the change of procedure in extracting the products caused viscosity to swell by an order of magnitude at the minimum. For example, at 15 min., the original viscosity was 0.87 Pa.s which increased to 18.05 Pa.s when the products were dissolved in methylene

chloride and the solvent evaporated before viscosity measurement. At 3.5 h (210 min), the viscosity rose to 0.12 Pa.s from 0.005 Pa.s, resulting in an increase by 24 fold. The escalation in viscosity for the solvent-separated products was maximum at 2.5 h (~60 times higher) and a minimum at 30 min (~10 times higher). With the exception of a local maxima at 60 min for the methylene chloride-separated products (Figure 5.4), the overall trend was a monotonic decrease with reaction time.

5.3.2.3 Viscosity of thermally converted products with reaction time with toluene as the extraction solvent

Surprisingly, the use of toluene to separate the thermally converted products did not alter the viscosity by much and was very much comparable to that of the straight thermally converted products. Figure 5.5 shows this comparison. It can be seen that except for the reaction times, 15 min and 30 min, where there was a mild increase in viscosity of the toluene-separated products by 0.10 and 0.15 units, respectively, the viscosities at higher reaction times were very similar to that of the straight thermally converted products. This was intriguing because a change in the nature of the solvent caused the viscosity to vary drastically. Surely, this was an effect that occurred after thermal cracking had taken place and adhesive forces between the solvent (methylene chloride) and the product matrix were suspected to be present, as discussed in later sections.

Table 5.10 shows the specific viscosity values and the shear rates employed for the toluene-extracted products.



(a) (b)

Figure 5.5. (a) Comparison between viscosities of toluene separated thermally converted products and of those collected without the use of a solvent; (b) The same data is shown with a zoomed y-axis to visualize the graphic details.

Table 5.10. Actual shear rates employed for viscosity measurement of the thermally converted products extracted with toluene.

Reaction time (min)	Viscosity (Pa.s)	Actual shear rate employed (s^{-1})
15	0.97 ± 0.09	23
30	0.49 ± 0.07	47
45	0.065 ± 0.030	340
60	0.057 ± 0.020	490
75	0.051 ± 0.020	500
240	0.01 ± 0.007	1520

5.3.3 Density and refractive index

The density and refractive index (RI) of the directly collected thermally converted products, measured at 40 ºC are shown in Figure 5.6. Only the liquid was taken after allowing the solids to settle in the sample container so as to not interfere in the measurement. However, with or without the use of the solvent, the densities of the samples decreased slightly when solids were present.

This is represented in Table 5.11. This might be due to the lower effective density of the solids due to possibly large differences in densities of the material that fill the pores and micropores of the solids as compared to the solid material itself. Though specific solid characterization has not been done in this work, the reader is referred to the work by Zachariah and De Klerk, [19] where the particle size distribution for the solids obtained from thermal conversion of Cold Lake bitumen was found to be bimodal.

The refractive index was not majorly affected by the potential presence of solids because the sample amount used to measure the property was really small that only liquid was in contact with the crystal surface of the refractometer. But overall, for the straight thermally converted products, the physical properties measured can be taken as belonging to the liquid itself.

Table 5.11. Density values of thermally converted products with and without solids (illustrated in Figure 5.6).

Reaction time (min)	Density without solids (Figure 5.6 - kg/m^3)	Density with solids (kg/m^3)
Feed	998.57	998.23
15	989.94	989.02
30	977.09	976.55
45	953.75	952.98
60	942.93	942.44
75	925.58	925.05
90	932.43	931.87
120	934.70	934.23
150	888.18	887.24
180	871.29	871.12
240	822.49	821.40
360	816.18	815.65
1170	793.45	793.33
1440	789.08	787.99

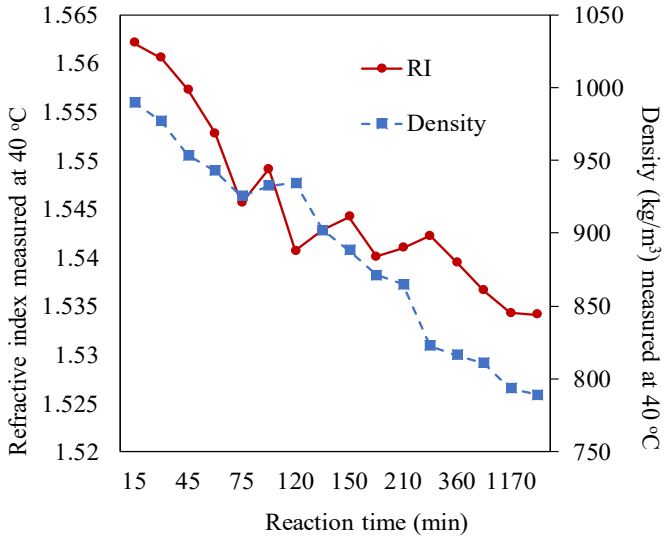


Figure 5.6. Plots of density and refractive index of the straight thermally converted liquid products (without solids) at different reaction times. All values are measured at 40 ºC.

It can be seen that both the density and refractive index follow a decreasing pattern and all values are less than those of the feed bitumen. This certainly indicates some compositional changes occurring during thermal conversion and will be elaborated in the discussions section. It is quite interesting to observe that in the first 30 min, the density decreased from 999 kg/m^3 in the feed to 977 kg/m^3 but showed a faster decrease once coke started forming at 45 min to 953 kg/m^3. The density for the 1440 min product was ~790 kg/m^3 which is much less than that of toluene. The faster decline in the density values in the coking regime might be due to the increase in lighter boiling material formed, as shown later in the boiling point distributions of the products. This went hand-in-hand with the monotonic decrease in viscosity as shown in the previous section. There was a slight increase in both density and RI for the product at 90 min (Figure 5.6), but this pattern was observed for RI at 135-150 min and 210-240 min also.

These properties were also measured for select samples in the case 'b_1' and 'b_2' (Figure 5.1), when the liquid and solid products were extracted with a solvent (methylene chloride and toluene separately). These values are reported in Table 5.12 and Table 5.13, respectively. The values were not much different to account for a different interpretation for the case 'b_2' so are not provided. The densities of other samples were not measured because of a large amount of sample was required in the densitimeter and the effect of the solvent on the density was able to be captured in

these measurements itself. However, only a small amount of sample was required for RI measurement but since it did not change by much, further samples were not measured.

Table 5.12. Densities of solvent-extracted products measured at 40 ºC.

Sample at reaction time (min)	Case 'a' – without solvent addition	Case 'b_1' – after MC evaporation	Case 'b_1' – after toluene evaporation
60	942.93	950.12	943.02
120	934.70	944.21	935.21

Table 5.13. Refractive indices of solvent-extracted products measured at 40 ºC.

Sample at reaction time (min)	Case 'a' – without solvent addition	Case 'b_1' – after MC evaporation	Case 'b_1' – after toluene evaporation
60	1.5528	1.5535	1.5527
120	1.5407	1.5415	1.5410

It can be seen that the density increased by 7.2 and 9.52 kg/m^3 for 60 min and 120 min samples when methylene chloride was used as the extraction solvent as compared to the straight thermally converted products. On the other hand, there was only a minor increase in the densities of the toluene-extracted products as seen in Table 5.12. There was not much change in the RI values for the solvent-extracted products as much as compared to the clear increase in density.

5.3.4 Free radical content

Apart from analyzing the straight thermally converted products, EPR was also used as a tool to inspect the changes caused in the products by addition and evaporation of the solvents, methylene chloride and toluene. The spectra are obtained in the first derivative form and the signals appear as intense lines, each consisting of a combination of two peaks (positive and negative). Two types of signals are commonly exhibited by bituminous samples: (i) one that corresponds to an organic free radical at ~g = 2.0032 (a symmetrical singlet); (ii) the set of 8 hyperfine lines

related to oxo-vanadyl species (VO^{2+}) that consists of 8 parallel and 8 perpendicular components, depending on the orientation of the V-O symmetric axis to the external magnetic field.

Some of these components (similar to Figure 1 in Niizuma et al. [35]) are indicated in Figure 5.7, where the EPR spectrum of thermally converted oil at reacted at 400 ºC and 15 min reaction time is shown. While operating the instrument, it was taken care that the microwave power was within saturation (all measurements done at 15 mW and room temperature). The saturation point for an undiluted bitumen was found to be 30 mW at room temperature.

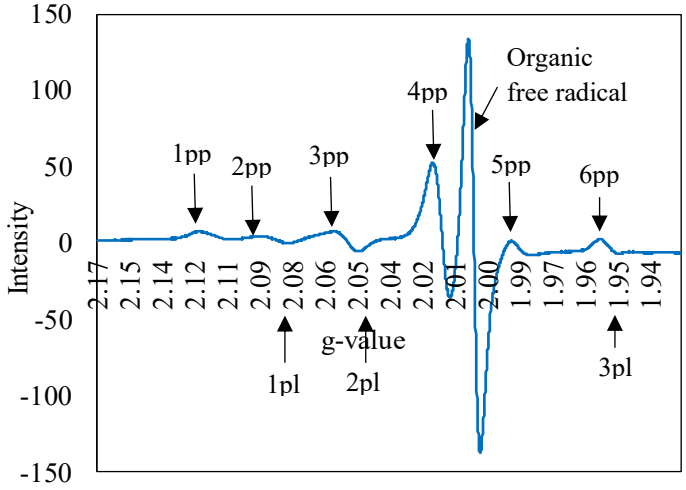


Figure 5.7. EPR spectra of thermally converted bitumen at 400 ºC and 15 min acquired without dissolution in solvent. The parallel (pl) and perpendicular features (pp) of the vanadyl peaks are indicated along with the organic free radical line.

The hyperfine spectrum (HFS) of vanadyl species is anisotropic in nature, where the perpendicular and parallel features can be attributed to the different orientations of the spin of ^{51}V nucleus with respect to the external magnetic field. For the HFS of the vanadyl species, the line of highest intensity occurs at g = 2.0160. The *g*-value of the 5th perpendicular peak as interpreted from the EPR spectra of thermally converted samples (Figure 5.7) came out to be 1.9850 $\pm$ 0.0017, which fell within the range of reported values in literature. Some of these were 1.984 $\pm$ 0.002 by Trukhan et al., [38] 1.974 $\pm$ 0.002 by Bielski and Gebicki, [54] 1.9943 by Cui et al. [55] and 1.9840 in the work by Malhotra & Buckmaster. [56] These spin-Hamiltonian parameters were derived by 2nd order perturbation theory the discussion of which is out of scope of this work. Out of all the 8 perpendicular peaks for VO^{2+}, the 4th peak was the most prominent.

Our focus in this study is the single, intense line corresponding to the combination of overlapping lines of the organic free radicals. It is important to note that all of the EPR signals arise from the concentration of paramagnetic species in the asphaltenic material in bitumen. [36,55] Maltenes show less intense peaks. Quantification of the organic radical peak in EPR spectra was done using the 2, 2 – diphenyl – 1 – picrylhydrazyl (DPPH) as the calibration standard. The details of the calibration and the corresponding equation are given in section C.1 Appendix C. The *g*-value for the DPPH as measured from the spectra is 2.0033 ± 0.0002 as given in the caption of Figure C.1 in Appendix C. It fell within the acceptable limits of the literature value (2.0037 ± 0.0002). [57]

5.3.4.1 Quantitative EPR analysis for the straight thermally converted products

The spin concentrations of the organic free radical in the straight thermally converted liquid products are shown in Figure 5.8.

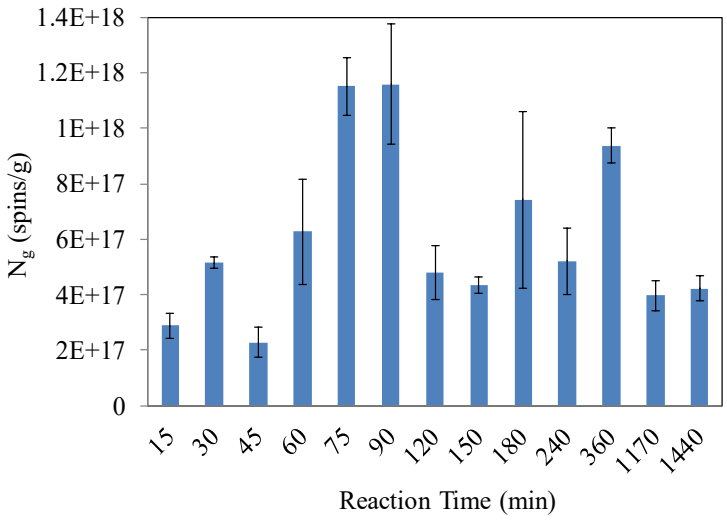


Figure 5.8. Quantitative EPR data for the thermally converted products (organic free radical) collected without using a solvent at the indicated reaction times. All reactions were conducted at 400 ºC and EPR spectra taken at room temperature. The average *g*-values of the organic free radical peak was 2.0025 ± 0.0005.

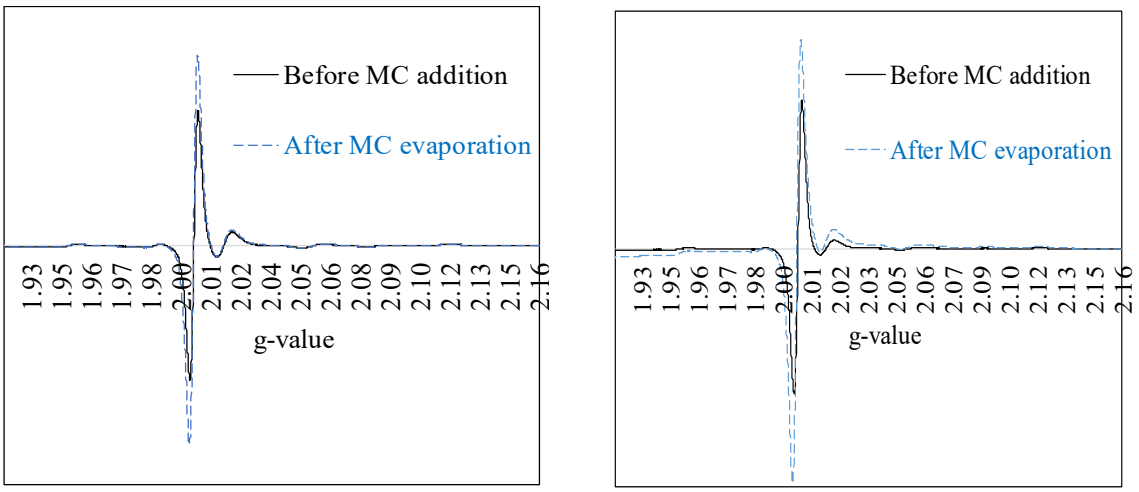
A literature value of the *g*-factor for the organic free radical peak was reported to be 2.0030 by Trukhan et al. [38] However, Malhotra & Buckmaster [56] observed shifts in calculated *g*-values for

vanadyl complexes and attributed it to the structure of the coordinating ligand. This might not be possible for organic free radicals as it is observed as an overlap of peaks with very less hyperfine splitting constant rather than a separate individual peaks. Khulbe et al. [58] reported values of 2.003 for the *g*-value for organic free radicals in bitumen obtained from Athabasca oilsands.

Interestingly, a non-monotonic trend was observed as against the expected increase in the free radical concentration in the liquid products with thermal cracking. After the initial increase in N_g, a local minima was observed 45 min. This is the time when coke started to form in determinable amounts. The concentration of the organic free radicals in the thermally converted products increased till 90 min reaction time which was the highest point in the graph reaching an order of magnitude higher than the feed free radical content of $6.4 \times 10^{17} \pm 3.1 \times 10^{16}$ spins (Figure 5.8). Furthermore, the free radical content dropped by 58% at 120 min and remained around this value even at higher reaction times of 19.5 h and 24 h, with two local maxima at 3 h and 6 h. The fact that spin concentration decreased below the feed level at large reaction times could be due to free radical termination with the available hydrogen and methyl radicals, which forms a part of the reported thermal cracking mechanism. [37] It could also be due to the shift to the singlet state from the doublet mono-radical and triplet biradical states. This will be further elaborated in the discussion (section 5.4.3.2).

5.3.4.2 EPR spectra of methylene chloride-separated thermally converted products

Some of the products from the procedure 'b_1' in Figure 5.1 were characterized through EPR spectroscopy to determining the free radical content. As given in the previous section, only the solid-free product was taken which meant that it was also equivalent to procedure 'b_2' in Figure 5.1. The data for two samples is shown in Figure 5.9 to illustrate the comparison between the solvent-extracted and the solvent-free products. Concentrations of 30 %wt. in toluene of the thermally converted products at 75 min and 210 min were prepared to obtain the spectra. The spin concentrations were calculated with the calibration equation, as described before. There was a marked increase in the signal intensity of the organic radicals in both cases after methylene chloride was added and removed as noted from Figure 5.9.



(a) (b)

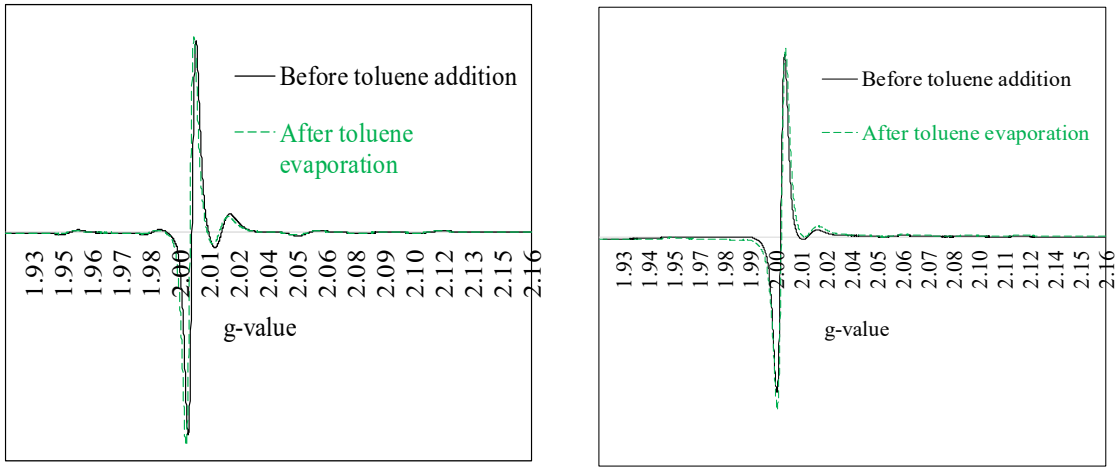
Figure 5.9. Comparison of the EPR spectra of methylene chloride-extracted products with the straight thermally converted products reacted at 400 ºC and (a) 75 min; (b) 210 min. The *g*-value of the organic free radical line is 2.0027 and 2.0025 at 75 min and 210 min, respectively. The y-axis is the first derivative of signal intensity and is on the same scale.

In the EPR spectra of the methylene chloride-separated product at 210 min, the 4th perpendicular hyperfine line of the vanadyl species also showed an increase in intensity as compared to the straight-product. On a quantitative scale, the spin concentration for the methylene-chloride extracted products increased to 1.38×10^{18} from 1.15×10^{18} spins/g for the 75 min product and from 6.42×10^{17} to 7.33×10^{17} spins/g for the 210 min product. The standard deviations for the straight products were 1.04×10^{17} and 9.83×10^{16} spins/g for the 75 min and 240 min products, respectively. So, the increase in spin concentration for the methylene chloride-separated product appeared to be true and outside the upper bounds of the corresponding straight thermally converted products.

5.3.4.3 EPR spectra of toluene-separated thermally converted products

The results of EPR analyses on the thermally converted products at (75 min and 240 min) that were separated from the reactor with toluene are depicted in Figure 5.10. There was minimal change in intensity for organic and vanadyl free radicals when toluene is used as extraction solvent.

Quantitatively, the spin concentration underwent a minor change that is within the measurement uncertainty of the technique from 1.15×10^{18} spins/g to 1.18×10^{18} spins/g (Std. Dev. = 7.46×10^{16}) for the 75 min product and from 5.21×10^{17} to 5.26×10^{17} spins/g (Std. Dev. = 8.81×10^{15}) for the 240 min product. The negligible change when toluene was used to extract the products and subsequently evaporated in comparison with the increase observed for methylene chloride-separated products warranted for additional discussion.



(a) (b)

Figure 5.10. Comparison of the EPR spectra of toluene-extracted products with the straight thermally converted products reacted at 400 °C and 75 min (left), 240 min (right). The *g*-value of the organic free radical line is 2.0028 and 2.0024 at 75 min and 240 min, respectively. The y-axis is the first derivative of signal intensity and is on the same scale.

5.3.4.4 Mass loss during evaporation of solvent from thermally converted products

Table 5.14 shows the loss of mass that occurred just after evaporation for three reaction times dealt with in Figure 5.9 and Figure 5.10.

The mass losses were minimal and only slightly lower when toluene was evaporated from the products compared to methylene chloride. This made sense because methylene chloride is lighter than toluene with a boiling point difference of about 80 ºC. Since the mass losses were not vastly different when both type of solvents were used, this test served as a confirmation that the comparison of the chemical characterization of the samples extracted using the two types of solvents was not due to the differences in mass loss due to solvent evaporation.

Table 5.14. Mass losses incurred during solvent evaporation. The data is given for the thermally converted samples obtained at 400 ºC and at the reaction times where EPR spectra was reported in the previous section.

Solvent	75 min (g)	210 min (g)	240 min (g)
Methylene Chloride	0.78 ± 0.1	0.83 ± 0.06	0.74 ± 0.06
Toluene	0.70 ± 0.1	0.72 ± 0.04	0.65 ± 0.08

5.3.5 FTIR spectra of the thermally converted products

The chemical nature of the products was further investigated through vibrational spectroscopy of the feed as well as the thermally converted products. Although the structure of the absorption peaks for the different samples remained the same there were notable differences in the FTIR spectra. The spectra of the feed and the straight thermally converted product obtained at 90 min is shown in Figure 5.11.

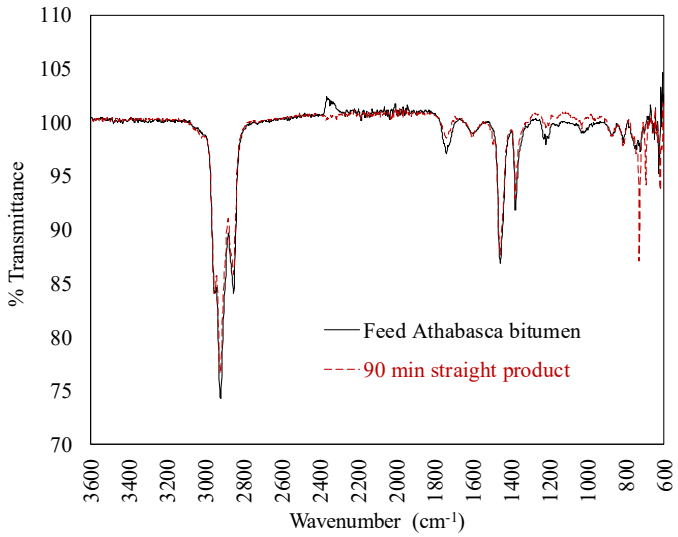


Figure 5.11. Comparison of the FTIR spectra of the feed bitumen and the thermally converted product obtained at 90 min reaction time.

5.3.5.1 Aromatic content of the straight thermally converted products

Since an ATR attachment was utilized along with the spectrometer, only ratios of absorption/transmittance intensities at the different frequencies of interest were compared so as to nullify the effect of the path lengths. Here, the peaks corresponding to aromatic C-H bends at 727, 810 and 860 cm^{-1} along with the peak for the ring bend at 694 cm^{-1} were taken together relative to the largest intense peak of the spectra at 2920 cm^{-1} that corresponded to aliphatic -CH_2 stretch. Among the different peaks that represent the aromatics, the absorption at 727 cm^{-1} represents majority of the monosubstituted aromatics (more than 4 adjacent H atoms), the absorptions at 810 and 860 cm^{-1} represent mostly *m*- and *p*- disubstituted aromatics along with some overlap with tri-substituted aromatics as well. [40] These data are given in Table 5.15.

Table 5.15. Variation of the aromatic absorption intensities relative to the aliphatic peak in the FTIR spectra for thermally converted samples obtained at different reaction times. All intensities were taken in terms of transmittance %.

Reaction time (min)	Sum of intensities of aromatic peaks/Intensity of methylene group stretch [a]	Extent of mono-substitution [b]
Feed	0.31	1.71
15	0.45	1.67
30	0.47	1.81
45	0.43	1.13
60	0.61	2.06
75	0.99	2.37
90	1.24	4.11
120	0.66	1.86
135	0.80	3.22
150	1.22	4.84
180	1.01	3.41
210	1.12	4.93
240	0.57	2.11
360	1.22	3.30
480	0.62	2.41
1170	0.53	1.90

1440	0.80	2.04

[a] aromatic ring bend - 694 cm^{-1}, Ar-C-H bends at 727, 744, 810 & 860 cm^{-1}, aliphatic -CH_2 stretch – 2920 cm^{-1}; calculated as (I694+I727+I744+I810+I860)/I2920; [b] intensity of peak at 727 cm^{-1} relative to the sum of intensities at 810 and 860 cm^{-1} calculated as I727/(I810+I860).

Beer's law is assumed to be valid for interpretation of the ratio of FTIR absorbance measurements. The first observation is that the aromatic content of all the products appears to be more than that of the feed. Interestingly, the ratio of the aromatic to aliphatic absorption peaks of the thermally converted products seemed to follow the trend in the free radical content given in Figure 5.8. This agrees with the general speculation that most of the free radicals are concentrated in the aromatic region of the material. [36]

Another quantity shown in Table 5.15 is the extent of substitution which signifies the mono-substituted aromatic content relative to the higher substituted aromatics (mostly di- with some overlap with tri-). [59] It is seen that with thermal conversion, the concentration of the mono-substituted aromatics also increased and was more than the feed except for the 45 min product. The dip in the aromatic content at 45 min correlated with the dip at the same time in the free radical content (Figure 5.8). The aromatic content seemed to peak at around 1.5-2.5 h reaction time and then decrease at higher reaction times of > 8 h (at 19.5 h and 24 h). The similar trends of the mono-substituted to di-substituted aromatics ratio with the total aromatic content provided information about the reaction chemistry of thermal cracking.

5.3.5.2 Estimating the remnant solvent content of the thermally converted products

It was speculated that the entire solvent could not be evaporated from the thermally converted products after reaction (procedure 'b' in Figure 5.1). It was essential to estimate the amount of solvent remaining in the samples after solvent evaporation. The residual solvent content was quantified using the broadened peak at 734 cm^{-1} (shifted from the normal C-Cl stretches at 748 and 717 cm^{-1}) representing the C-Cl stretch for the methylene chloride-extracted products and the 727 cm^{-1} aromatic C-H stretch (no shift) for the toluene-extracted products. Further details of this procedure are provided in section C.2 of Appendix C.

The amount of methylene chloride remaining in the products fluctuated between 0.90 %wt. at 1170 min and 1.34 %wt. at 30 min reaction time, with an average value of 1.11 ± 0.14 %wt. As the product became less viscous and less dense, the amount of methylene chloride remaining also slightly decreased. On a similar note, the average amount of toluene remaining in the products was slightly more at 1.27 ± 0.16 %wt.

5.3.5.3 Frequency shift for C-Cl stretching and C-H wagging bands in methylene chloride-extracted products

As indicated in the details for obtaining the calibration curve, the FTIR spectra for the thermally converted products after evaporation of methylene chloride showed a broader band with a shift to lower wavenumber for the C-Cl stretch (initially at 744 cm^{-1} for pure CH_2Cl_2 and shifted to 734 cm^{-1}). Meanwhile, in some products the band for C-H deformation (initially a sharp band at 1265 cm^{-1} for pure CH_2Cl_2) shifted to a higher frequency (1274 cm^{-1}) with a broadening of the C-O (from alcoholic species) at 1224 and 1216 cm^{-1} without any shift. These shifts in frequency bands were not observed in the straight thermally converted products that were extracted without methylene chloride. These effects are illustrated in Figure 5.12 and Figure 5.13 for the 900 – 600 cm^{-1} region and 1300 – 1200 cm^{-1} region, respectively.

On the other hand, no such frequency shifts or broadening of the bands were observed in the case of toluene-extracted products. These observations provide hints of the presence of hydrogen bonding due to the remnant polyhalogenated hydrocarbon in the thermally converted products. It is elaborated further in section 5.4.3.1 in the discussion.

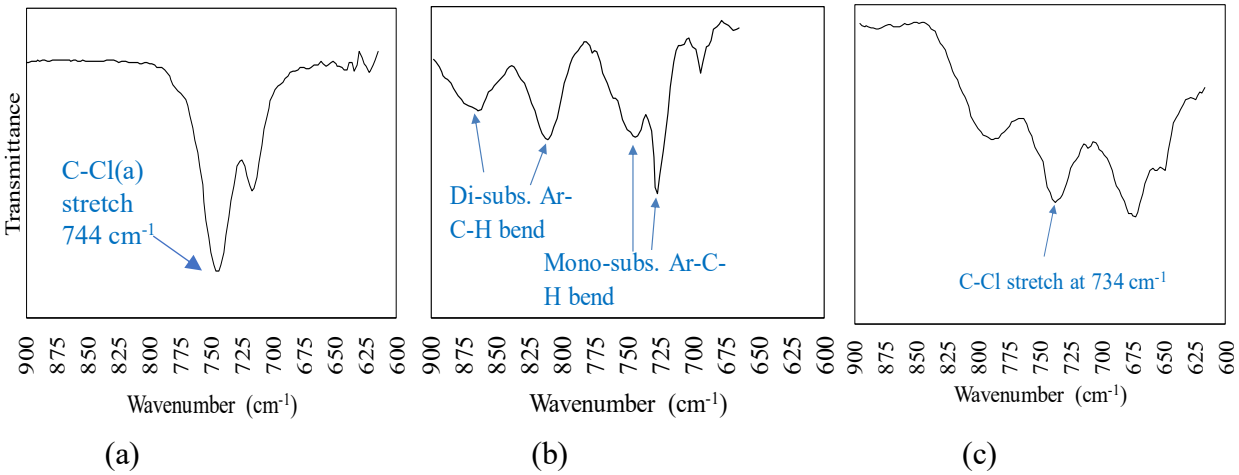


Figure 5.12. FTIR spectra of (a) pure CH_2Cl_2; (b) the solvent-free 15 min thermally converted product; (c) the methylene chloride-extracted thermally converted product at 15 min showing absorption bands in the 900 - 600 cm^{-1} region. The bonds responsible for the respective bands are indicated.

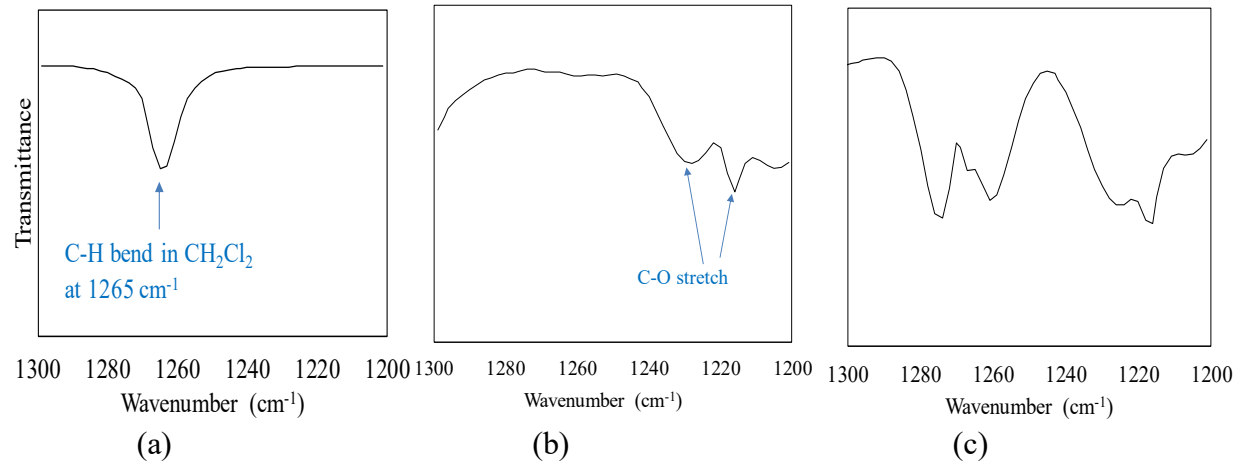


Figure 5.13. FTIR spectra of (a) pure CH_2Cl_2; (b) the solvent-free 30 min thermally converted product; (c) the methylene chloride-extracted thermally converted product at 30 min showing absorption bands in the 1300 - 1200 cm^{-1} region.

5.3.6 Proton NMR (^{1}H-NMR) spectra of the thermally converted products

The nature of hydrogen that was distributed among the aromatics and non-aromatics (including saturated cyclic and acyclic material) in the straight thermally converted products as compared to the feed bitumen was evaluated through proton NMR spectroscopy. The changes in

the products as indicated by EPR and FTIR spectra evoked the necessity to explore the changes caused in the products by the extraction solvent through NMR spectra.

5.3.6.1 Hydrogen distribution in the thermally converted products with no solvent

As given in Table 5.2, the non-aromatic and aromatic hydrogen content of the feed bitumen was 90.34 % and 9.66 %. For both the feed and the thermally converted products, there were three distinct peaks for the aliphatic range (0.5 – 4.5 ppm): (i) at δ = ~0.9 ppm indicating the aliphatic -CH_3 content; (ii) δ = ~1.3 ppm indicating methylene (-CH_2) that is part of an aliphatic chain or belonging to a naphthenic ring; (iii) δ = ~2.3 ppm that mostly is combined with a broader peak at ~2.6 ppm. The peak at the shift of 2.3 ppm indicates the benzylic protons while the broader peak which is more de-shielded could be assigned to a H attached to the α-Carbon of a carbonyl carbon or an alkynyl proton. [40] But these kinds of functional groups are quite rare in bitumen and due to the possibility of overlap, we intend to omit this peak (at 2.6 ppm) during the interpretation and focus more on the benzylic and methyl protons in this work. Figure 5.14 shows the comparison between the ^{1}H-NMR spectra of feed bitumen and the straight thermally converted product after reaction for 90 min.

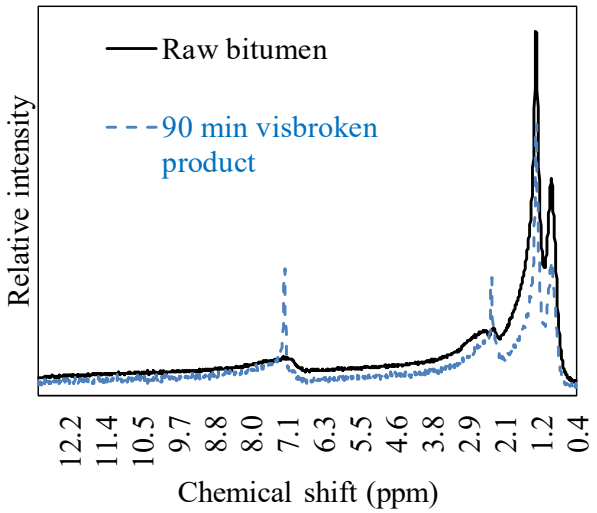


Figure 5.14. ^{1}H-NMR spectra of feed Athabasca bitumen and 90 min solvent-free thermally converted product.

It can be seen that for the unreacted bitumen, the aliphatic peaks are reasonably distinct, whereas the aromatic region (6.5-8.5 ppm) is much broader than in the 90 min thermally converted product. Figure 5.15 depicts the aromatic hydrogen content along with the -CH_3 proton and benzylic hydrogen contents of the aliphatics for the straight thermally converted products. Only the liquid part of the product was taken for sample preparation.

It can be seen that there was an increase in aromatic hydrogen with thermal cracking reaction time. Though the increase was not monotonic, the overall curve had an upward slope. Similar observations were seen from the FTIR data as reported in Table 5.15. In addition, the variation of aromatic hydrogen appeared to be similar to the trend of free radical content as seen in Figure 5.8. A dip was seen at 45 min and local maxima appeared at 90 min and 360 min. The only difference was that the highest aromatic hydrogen content was obtained at 360 min but the highest free radical content was reported at 90 min. As with the concentration of the organic paramagnetic species, the aromatic hydrogen showed a decrease at the highest reaction times investigated in this work (1170 min and 1440 min).

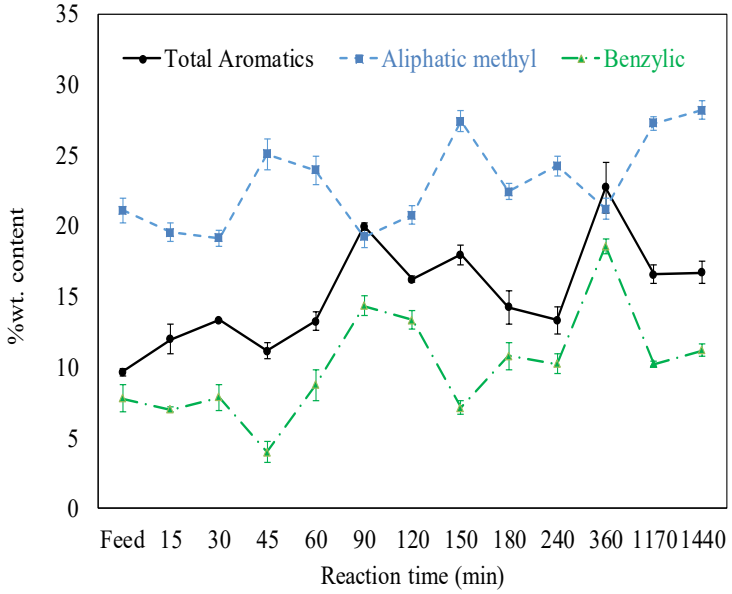


Figure 5.15. Hydrogen distribution in the aromatics and those attached to benzylic and methyl carbons in aliphatic groups.

Also, the aromatic hydrogen content showed a very good correlation with the benzylic hydrogen. The FTIR data showed that the ratio of aromatic intensities with respect to the aliphatic methylene band had a very similar trend as the mono-substituted aromatic peaks. From the ^{1}H-

NMR spectra, it is difficult to determine whether the peak of the benzylic proton is indicative of the number of substituents in the aromatic ring. For example, a decreasing peak intensity can indicate a decrease in the number of hydrogens on one methyl substituent while the other substituent remains unchanged. So, the trend of the benzylic hydrogen cannot be strictly assigned to mono-substituted aromatics like in the FTIR but, in consideration with the other types of hydrogen (terminal methyl and chain-methylene), it provides information on the reaction chemistry for thermal cracking and will be discussed in further detail in section 5.4.2 in the discussion.

There was an initial decrease in the aliphatic -CH_3 protons (Figure 5.15) till 30 min and then a local maxima at 45 min (that corresponded to a local minima for the aromatic and benzylic hydrogen). At all later times except 90 and 360 min, the methyl hydrogen content of the thermally converted products was greater than or equal to that of the feed bitumen. This suggested that there was a decrease in aliphatic chain length attached to the aromatic groups or the naphthenes as thermal conversion proceeded. Within the first 60 min, there was a noticeable minimum in the aromatic hydrogen in the liquid product when coke started forming at 45 min reaction time. The observation that the trends of methyl protons and benzylic protons across the reaction times appeared to be mirror images of each other was intriguing.

5.3.6.2 Change in the ^{1}H-NMR spectra for the solvent-extracted thermally converted products

When compared with the spectra for the straight thermally converted products, there were changes in line shapes, shift in frequencies of some proton signals and changes in aliphatic methyl and aromatic proton contents for the methylene chloride-extracted thermally converted products but no such changes were found when toluene was used as extraction solvent. This is shown in Figure 5.16.

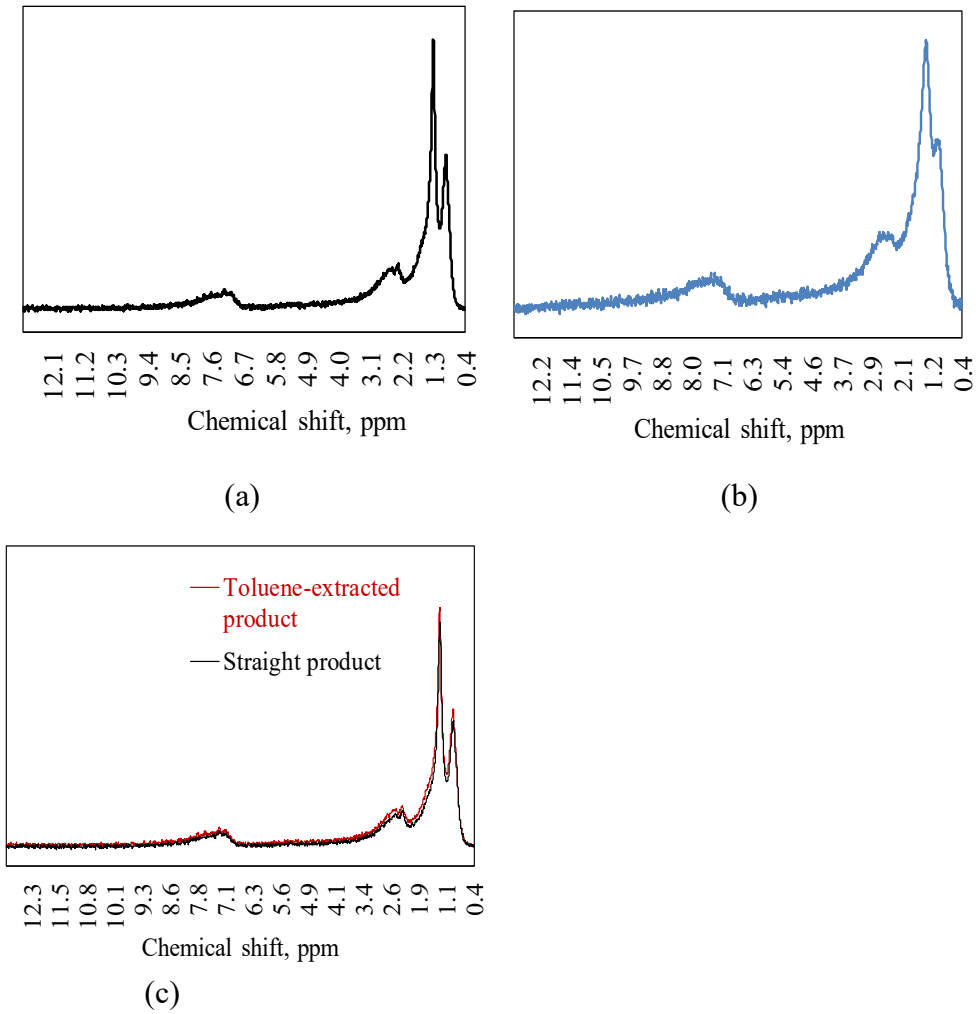


Figure 5.16. ^{1}H-NMR spectra of thermally converted product: (a) At 60 min reaction time mechanically extracted without the use of solvent compared with that of the same thermally converted product extracted with (b) methylene chloride and (c) toluene.

It was seen in the previous section that as compared to the feed bitumen, the NMR spectra of the straight thermally converted product obtained at 90 min exhibited sharper peaks for the benzylic and aromatic protons (Figure 5.14). Though these signals were not as sharp for the 60 min product (Figure 5.16a) as the 90 min spectrum, the peaks corresponding to the aliphatic methyl and methylene protons showed a clear broadening when the same product was extracted with methylene chloride (Figure 5.15b). There was also increased area observed for the spectral region between 4.5 – 6.5 ppm for the solvent-extracted product (from 0.012 to 0.073 units) due to residual

methylene chloride. Though this increase in area was not quantified, it can be related to the quantified value of the remnant solvent from FTIR as done in the previous section.

The ^{1}H-NMR spectra of the 120 min, 150 min and 360 min thermally converted products mechanically extracted are compared with that of the solvent-extracted products in Figure 5.17. In the 2 h solvent-free product, the aromatic and benzylic hydrogen peaks are sharp, whereas the methylene chloride-extracted product peaks for these protons appear to be broadened. There was no change in line shape for the aliphatic methyl and methylene proton signals. Also, there was a slight shift in the resonant frequency for the methylene protons from 1.34 (in the straight thermally converted product) to 1.45 ppm (in the solvent-extracted product) for the 2 h product. There was no shift in the trimethylsilane (TMS) peak for the solvent-extracted products as compared to the straight products. The integral area in the spectral region 4.5 – 6.5 ppm showed a spike for the methylene chloride-extracted product.

Similarly, the thermally converted products at 150 min exhibited broadening in the aliphatic and aromatic hydrogen peaks when extracted with methylene chloride. The aliphatic -CH_2- proton peak for the solvent-extracted product shifted by 0.11 ppm to a higher frequency, with the rest of the peaks showing no changes in frequency as compared with the solvent-free product. The interesting fact about the spectra for the 360 min thermally converted product was that the peak corresponding to -CH_2- proton that most plausibly belongs to CH_2Cl_2 that was remaining in the products was seen (marked with an arrow in Figure 5.17c). In pure methylene chloride, this peak appears at 5.34 ppm, but in this thermally converted product it is seen at 5.41 ppm, with no change in the frequencies at which other proton peaks were obtained. This peak belonging to methylene chloride was not clearly visible in other solvent-extracted products, but only increase in integral areas in the 4.5 – 6.5 ppm region was seen. A slight broadening of the aliphatic methyl and methylene peaks were also seen for solvent-extracted 360 min thermally converted product.

In the study by Yañez & De Klerk, [20] the ^{1}H-NMR peaks for aliphatic methyl and methylene hydrogen were sharper in the 8 h thermally converted product as compared to the feed Cold Lake bitumen, but there was no improvement in resolution of the aromatic hydrogen. In their case, all thermally converted products were extracted by using methylene chloride and the feed itself showed broadened peaks in the NMR spectra. In the findings of NMR specific to this chapter, there seemed to be a shift in the peaks for aliphatic methyl and methylene hydrogen region for the methylene chloride-extracted products at lower reaction times with line broadening for the

aromatic and benzylic peaks. At higher reaction times, line broadening was more prominent in the aliphatic methyl and methylene proton region with a shift for the CH_2Cl_2 proton peak location to a higher value. On the other hand, no such peculiar behavior was seen when the reaction products were extracted with toluene. These observations pointed towards a definite chemical change in the nature of the methylene chloride-extracted products and demanded for a deeper discussion.

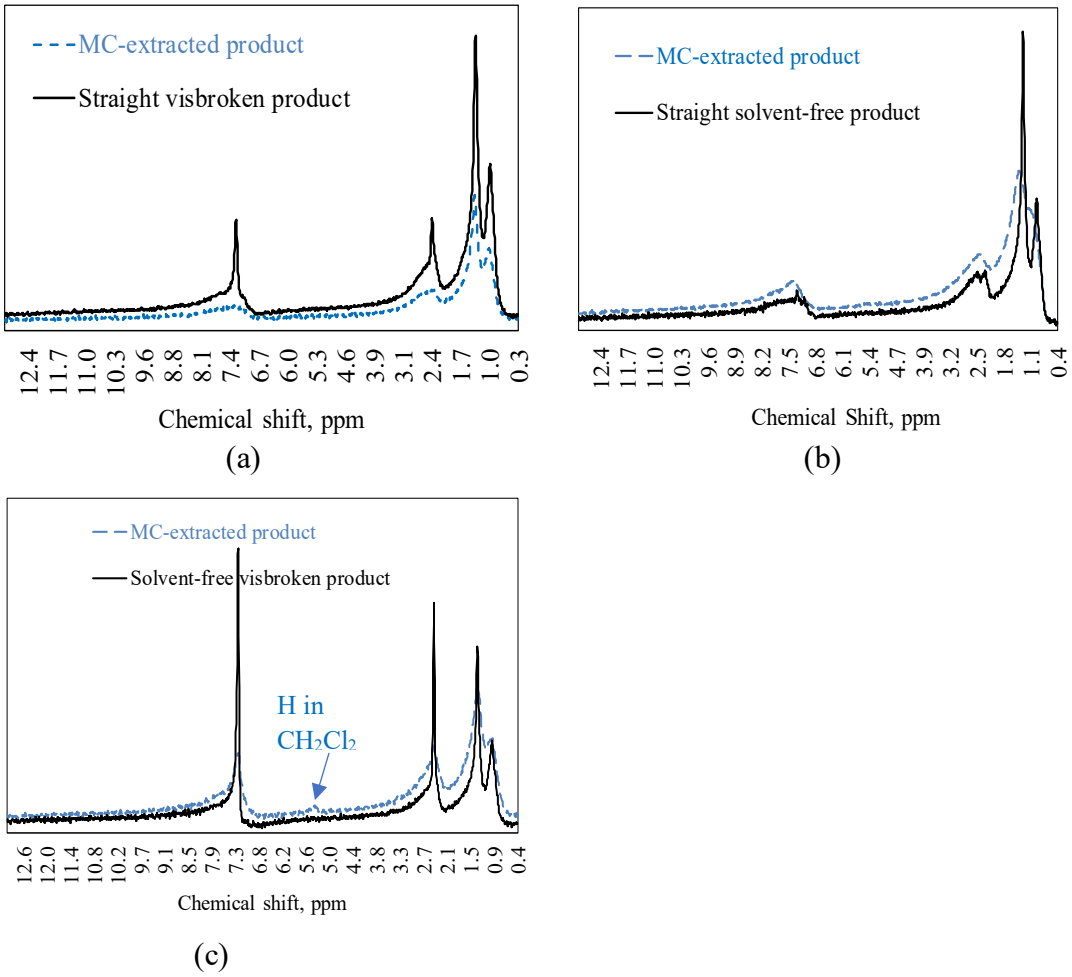


Figure 5.17. Comparison of the ^{1}H-NMR spectra of straight and methylene chloride-extracted thermally converted product at: (a) 120 min; (b) 150 min & (c) 360 min reaction times.

5.3.7 *n*-Pentane insoluble content of the thermally converted products

Since the experimental procedure in this work involved the use of two different types of solvents, one polar aliphatic and a relatively non-polar aromatic one, it was worthwhile to investigate whether this procedure caused any change in the *n*-pentane insoluble content obtained

from the liquid product. This comparison will further relate to the question whether chemical changes are occurring in the methylene chloride-extracted product as opposed to the toluene-extracted product. An interesting observation was that the content of toluene-insoluble solids was similar to that of methylene chloride-insoluble solids. Figure 5.18 depicts the change in *n*-pentane insoluble content with reaction time for the solvent-separated products.

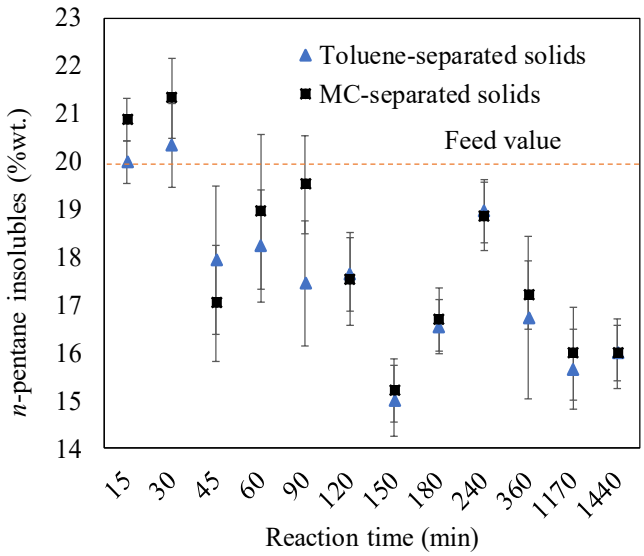


Figure 5.18. Comparison of *n*-pentane insoluble material in the liquid products when the solids were separated using methylene chloride (black squares) and toluene (blue triangles) as the solvents.

The asphaltenes fraction is a solubility class and this work employed *n*-pentane as the anti-solvent for precipitation. For this reason, asphaltenes will be referred to as *n*-pentane insoluble in this chapter. Also, since two different solvents were used to separate the solids, causing plausible a difference in the chemical nature of the liquid products as evident from previous characterizations like EPR, FTIR and NMR, it was preferred to refer to this class of material as *n*-pentane insoluble instead of referring to it as asphaltenes.

A monotonous variation was not seen for the *n*-pentane insoluble content with increase in reaction time as with the viscosity (Figure 5.2). The overall trend for both sets of points appeared to be decreasing. Initially, the *n*-pentane insoluble material increased till 30 min and then decreased in concentration at 45 min reaction time, similar to that observed in the free radical concentration and aromatic hydrogen content. The *n*-pentane insoluble content when the solids were separated

from the liquid product with both the solvents covaried with each other except at 90 min, where opposite local optima were observed. Although the trend in the *n*-pentane insoluble material can be viewed as being similar to the free radical content in the products (Figure 5.8), there were differences in local minima and maxima attained. A local minima was attained at 240 min in the free radical content while that was the point of local maximum in the *n*-pentane insoluble content.

The least value of *n*-pentane insoluble material was 15.0 %wt. at 150 min when the solids were toluene-insoluble, while the maximum value reached reached to 20.3 %wt. (at 30 min). This suggested that the decrease was not as drastic as reported in the work by Henderson & Weber [22] where the asphaltene content in thermally cracked Athabasca bitumen (conducted at 397 ºC) fluctuated between values as high as 70.5 %wt. and as low as 0.02 %wt. The intriguing finding was that the *n*-pentane insoluble content when the solids were separated with methylene chloride was on an average ~0.4 %wt. higher than when the solids were separated with dissolution of the liquid products in toluene, except at 45 min, 120 min and 240 min. At lower values, it can be ascertained that the difference between the *n*-pentane insoluble material of the liquid products extracted by both solvents was minimal. This could also be one of the reasons for the increased viscosity for the methylene chloride-extracted product, in combination with other observations.

5.3.8 MCR content

Figure 5.19 shows the MCR content of the Athabasca bitumen feed and thermally converted products. As already mentioned in the experimental procedure, this value indicates the propensity for coking in the feed and products.

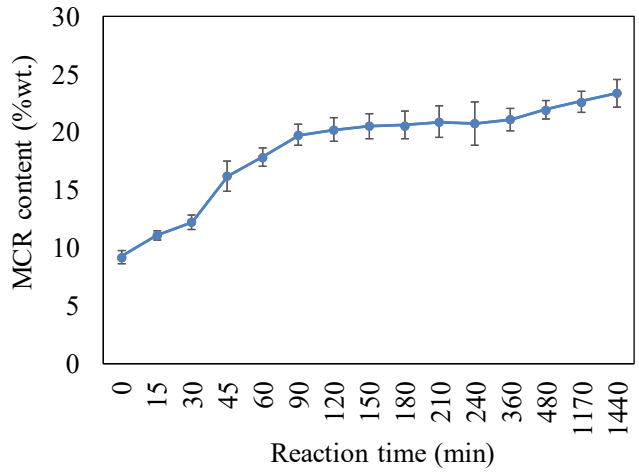


Figure 5.19. Microcarbon residue content of the bitumen feed and thermally converted products at different reaction times. The value for the raw bitumen feed corresponds to 0 min.

It can be seen that the MCR content gradually increases with reaction time from a feed value of 9.2 %wt. to ~23.5 %wt. at the final reaction of 24 h. There was a relatively steep increase at 45 min from 12 to 16 %wt. which corresponded with the start of significant coke formation in the 45 min product (Table 5.5, Table 5.7). The standard deviation for the product at 240 min was the maximum with the 45 min product exhibiting the next highest deviation at 1.3 %wt. There was also a 0.2 % drop in the MCR value from 210 min to 240 min but the overall trend can be said to be increasing.

5.3.9 Elemental composition

The elemental composition of the feed Athabasca bitumen was given in Table 2. Elemental analysis of the straight thermally converted products (only liquid) was carried out and the results are compiled in Table 5.16, in terms of the H/C ratio, nitrogen (N) and oxygen (O) content. A plot of the variation of H/C ratio across reaction times is given in Figure 5.20.

The elemental analysis was important in two aspects: the hydrogen to carbon ratio and the heteroatomic content. It can be seen from Figure 5.20 that the H/C ratio is lesser than the feed at all reaction times for the thermally converted products that do not contain any solvent. The trend was non-monotonic. At the point where carbonaceous solids started to form, the H/C ratio went through a global maximum (at 1.48) among only the products (excluding the feed).

Table 5.16. Results of elemental analysis of the straight thermally converted products.

Time (min)	H/C	O (%wt.) [a]	N (%wt.)	S (%wt.)
Feed	1.489 ± 0.002	0.83	0.58 ± 0.05	4.80 ± 0.12
15	1.459 ± 0.003	1.89	0.55 ± 0.04	3.95 ± 0.23
30	1.466 ± 0.002	2.17	0.51 ± 0.06	3.36 ± 0.12
45	1.476 ± 0.004	1.17	0.48 ± 0.03	3.50 ± 0.15
60	1.460 ± 0.001	1.50	0.46 ± 0.04	3.41 ± 0.05
90	1.435 ± 0.004	1.61	0.42 ± 0.03	3.47 ± 0.06
120	1.443 ± 0.003	1.81	0.36 ± 0.02	3.28 ± 0.09
150	1.466 ± 0.003	2.35	0.31 ± 0.01	2.98 ± 0.10
180	1.462 ± 0.002	2.68	0.35 ± 0.05	2.73 ± 0.08
240	1.431 ± 0.002	3.09	0.35 ± 0.04	2.81 ± 0.04
360	1.433 ± 0.002	1.42	0.35 ± 0.06	2.54 ± 0.09
1170	1.444 ± 0.005	3.61	0.31 ± 0.02	2.93 ± 0.10
1440	1.458 ± 0.004	3.79	0.24 ± 0.07	2.63 ± 0.07

[a] determined by difference.

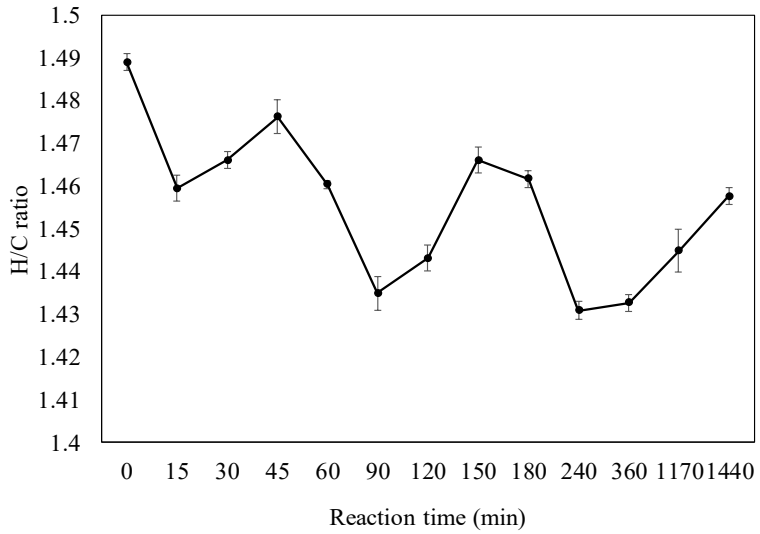


Figure 5.20. H/C ratio of the straight thermally converted products with reaction time. '0' min corresponds to the feed bitumen at room temperature (25 ºC).

This corresponded with a local minimum in the *n*-pentane insoluble material (Figure 5.18). It was interesting to note that there was an immediate drop in the hydrogen-to-carbon ratio to 1.46 at the minimum reaction time of 15 min, at which the aromatic hydrogen (Figure 5.15) and the *n*-pentane insoluble content (Figure 5.18) showed an increase. At 90 min, where the free radical spin concentration (Figure 5.8) and the aromatic hydrogen showed a maximum, the hydrogen-to-carbon ratio of the product passed through a minimum at 1.43. After dropping low at 240 min and 360 min, it seemed to increase at further reaction times of 1170 and 1440 min.

With the heteroatoms, there was a decrease in the sulfur and nitrogen contents in the thermally converted products (Table 5.16). The feed bitumen was quite sour with a sulfur content of 4.80 %wt. On thermal cracking, it decreased to 2.63 %wt. at the highest reaction time of 1440 min explored in the experiments conducted in this work. Similar observations were seen with the variation in nitrogen in the products with the lowest being at 1440 min (0.24 %wt.). All values in the products were less than the feed bitumen. Interestingly, there was an increase in the oxygen content observed in the thermally converted products as compared to the feed. The increase was not monotonous but all products had a higher oxygen content than the feed. It should be added that oxygen was determined by difference and the change could reflect a systematic increase in CHNS material not quantified by analysis.

The extraction of the thermally converted products with both solvents effected a mild increase in the hydrogen-to-carbon ratio, but no significant change in the heteroatomic content. It was tested on three select samples and shown in Table 5.17.

Table 5.17. Change in H/C ratio and heteroatomic contents on solvent-extraction.

Time (min)	H/C straight [a]	H/C (MC) [b]	H/C (Toluene) [c]	O (%wt.) [d]	N (%wt.)	S (%wt.)
90	1.434	1.450	1.452	0.83	0.58 ± 0.05	4.80 ± 0.12
180	1.461	1.467	1.471	1.89	0.55 ± 0.04	3.95 ± 0.23
360	1.432	1.440	1.445	2.17	0.51 ± 0.06	3.36 ± 0.12

[a] straight thermally converted product; [b] methylene chloride-extracted product; [c] toluene-extracted product; [d] determined by difference.

5.3.10 Boiling point distributions using simulated distillation

Figure 5.21 shows the distillation profiles of the feed compared with two of the straight thermally converted products at 60 min and 1440 min. There are a number of key observations that can be seen from the boiling point distribution curves. The vacuum residue cut point is taken as 525 ºC as marked in Figure 5.21 and a clear increase in the lighter fractions was seen as the amount of material that could be distilled for the 60 min and 1440 min products were greater than the feed by 19 %wt. and 28 %wt., respectively. When the material balance is considered (Table 5.3), the actual yield of < 525 ºC liquid products after 60 min and 1440 min reaction time was 55 and 56 %wt. of the feed. At 720 ºC, about 27 %wt. of the feed bitumen did not elute. Similarly, about 20 %wt. of the 60 min and 1440 min products did not elute at lower boiling temperatures of 709 ºC and 688 ºC, respectively.

The increase in the amount of lower boiling material on thermal cracking was concomitant with the decrease in density and refractive index of the products (Figure 5.6). Indications that 400 ºC was a sufficient temperature to perform cracking (and coking) were seen from the increment in the gas yield too. It also reflected the overall decrease in the *n*-pentane insoluble material, though it was non-monotonic. The amount of material boiling in the naphtha range (88 – 190 ºC) had values of 0, 2 and 16 %wt. in the feed, 60 min and 1440 min thermally converted products, respectively. Similarly, formation of the distillate fraction (190 – 360 ºC) also rose from 15 %wt. in the feed to 38 %wt. in the 1440 min thermally converted product. These increments in the lighter boiling material could have played a major role in causing a massive reduction in viscosity of the thermally converted products, as will be discussed later.

It was also important to check whether evaporation of the extraction solvent altered the distillation profile of the thermally converted product. Since there was minimum mass loss during rotary evaporation (Table 5.14), it was expected that the distillation profile would not be altered by much due to this procedure. If a significant fraction of the lighter boiling material was gone with the solvent during evaporation, that would have automatically caused the viscosity to increase due to the dominant heavier fraction, but this was not the case here. Even with toluene, no significant change in the boiling point distribution curve was observed. These are shown in Figure 5.22. This meant that there was some other aspect relating to a chemical or a physical change that caused the viscosity to increase post methylene chloride evaporation from the thermally converted products as shown in Figure 5.4.

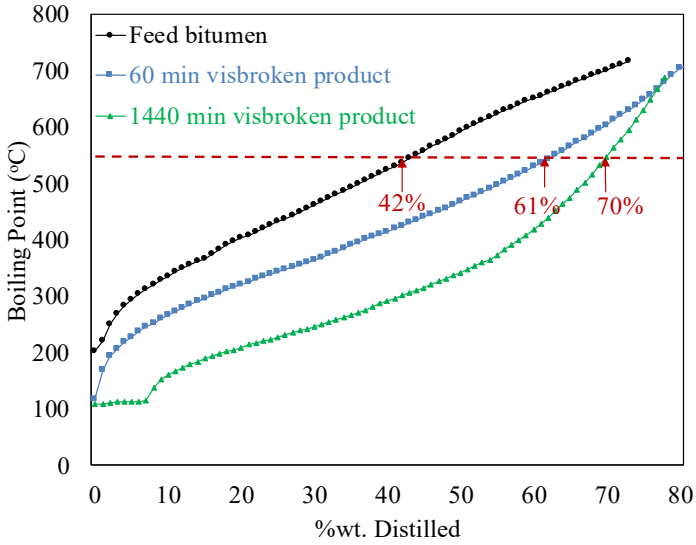


Figure 5.21. Simulated distillation curves for the feed bitumen and the thermally converted products obtained at 60 min and 1440 min with no solvent.

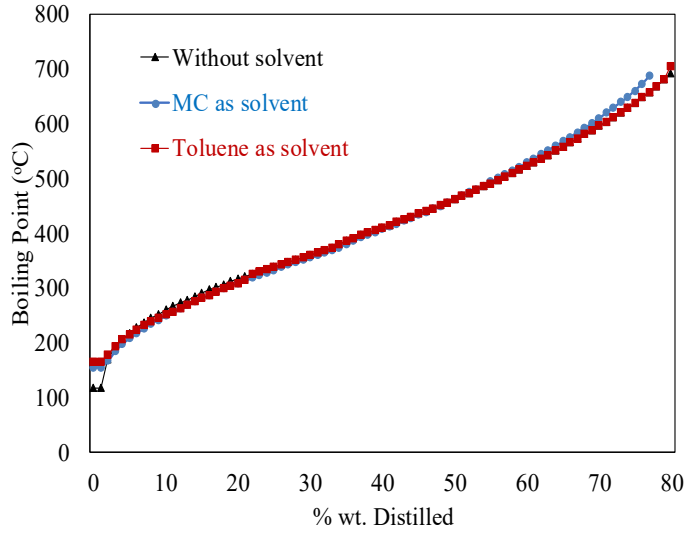


Figure 5.22. Comparison of the distillation profiles for the thermally converted product at 60 min obtained with methylene chloride extraction (blue curve), toluene extraction (red curve) and without any solvent (black curve).

5.3.11 Images of the coke solids

Microscopic images of the toluene-insoluble solids formed on thermal conversion were taken at select reaction times and are shown in Figure 5.23. The solids obtained on separation with methylene chloride did not appear much different and thus not shown in this section of the chapter.

It can be seen that the solids changed in shape and color with increase in reaction time. At lower reaction times (30 min), the shape resembled that of platelets and could ascribed to mostly mineral matter present in the feed that carried over to the products. Some nonmineral solids and carbonaceous content could also be associated with them. A zoomed in figure of a single coke particle is shown at 45 min, when the "coke-like" solids started forming significantly on the surface of the mineral matter. Once the solids content started rising (Table 5.5 and Table 5.7), more organic matter started depositing on the mineral matter, thus changing the shape of the solids as visible through microscopy (Figure 5.23). With a further increase in reaction time, the solids swelled in size and the organic matter deposited became clearly visible as a shiny layer at 240 min and 360 min.

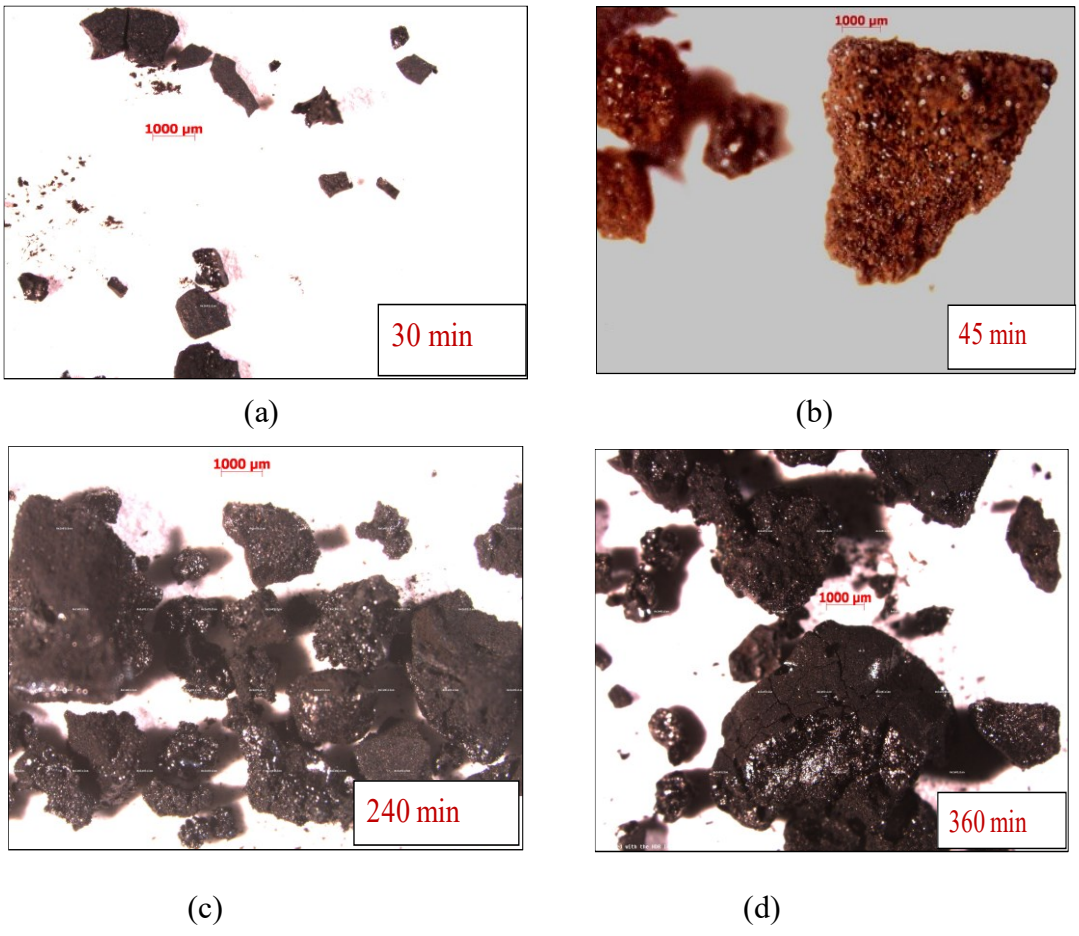


Figure 5.23. Optical images of the carbonaceous solids formed after thermal cracking and separated from the liquid products by using toluene.

5.3.12 Gas analyses

Only a qualitative investigation of the gaseous products was carried out for the purpose of this study in this chapter. For the experiments conducted at 400 ºC, it was found that methane (CH_4) was the dominant hydrocarbon product at all reaction times. The composition of methane in the gaseous products increased gradually with an increase in the reaction time. C_2, C_4, C_5 and C_6 were the other hydrocarbons in the gaseous phase. Both alkanes and alkenes of these carbon numbers were observed. The intensity of the C_4 peaks increased with time with C_5 and C_6 being lesser in concentration as compared to C_4. Interestingly, for the C_2 hydrocarbon, minor amounts of acetylene at higher reaction times was also seen along with ethane and ethene. No CO_2 was formed initially and with no noticeable increase at later times.

5.3.13 Effect of shear rates on product viscosity

Viscosities for the different thermally converted products were measured at the indicated shear rates as given in Table 5.8 (mechanically extracted without any solvent), Table 5.9 (methylene chloride-extracted products) and Table 5.10 (toluene-extracted products). These were in accordance with viscometer measurements. All the shear rates were more than that used in recent works on Cold Lake bitumen. [18–20] The rheometer enabled us to better visualize the change in viscosity with shear rate for the different straight thermally converted products than the viscometer. Results for selected reaction times are given in Figure 5.24a. The data for 30, 45, 60 and 150 min is also reported in the logarithmic scale in Figure 5.24b.

Two types of rheological characteristics are commonly exhibited by complex fluids and were checked for the bitumen and its thermally converted products in this study. These are the shear-dependent viscosity and the time-dependent viscosity at a constant shear. Other rheological properties like normal stress differences, dynamic measurements such as storage and loss modulus were not inspected for these samples but the reader is referred to specific works [34,60,61] for discussion on these topics for different kinds of bitumen.

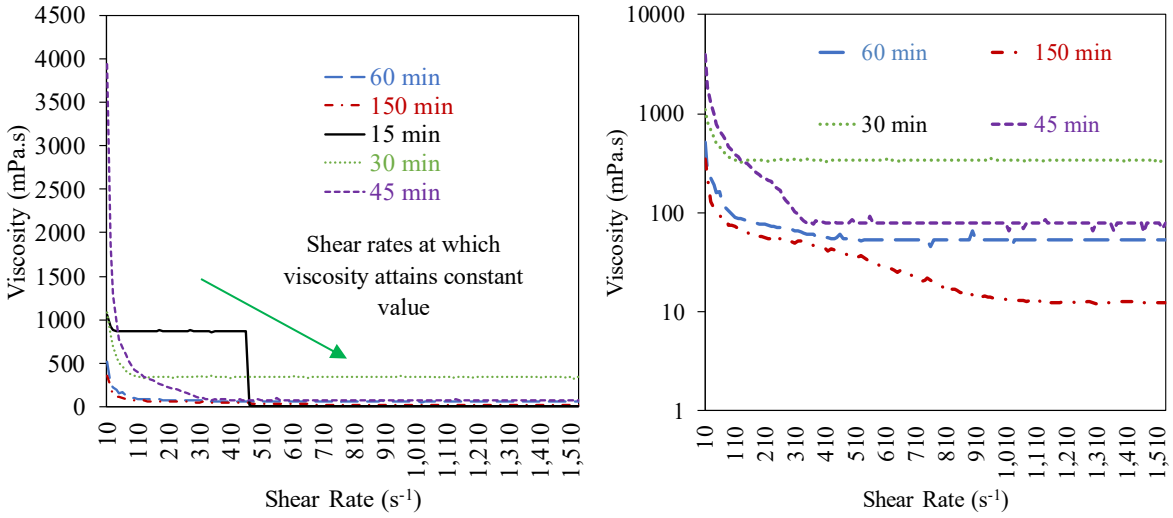


Figure 5.24. Variation of viscosity with shear rate for the thermally converted products obtained at 15, 30, 45, 60 and 150 min (left). The figure on the right represents the data for 30, 45, 60 and 150 min with viscosity on the logarithmic scale. All measurements were taken at 40 ºC.

There are a number of observations that can be made from the plot of shear rate-dependence of viscosity (Figure 5.24). In the shear rate range of 10 – 1544 s^{-1} (1544 s^{-1} was the upper limit of the rheometer), the viscosity decreased with increasing shear rate at all the indicated reaction times. This is reminiscent of shear thinning behavior, and very high shear rates did not have an effect on the viscosity. For the 15 min product, the low shear viscosity was 1.05 Pa.s which reached a constant value of 0.87 Pa.s in the shear range of 30 – 470 s^{-1}, after which there a sudden drop to < 10 mPa.s. This drop was not seen in products obtained at higher reaction times, where viscosity attained a constant value at higher shear rates.

The green arrow indicates that the shear rates at which viscosities reached a constant value, increased as the product becomes more fluid. The torque was also found to increase as the shear rate increased. This meant that as the sample became less viscous, a higher shear rate would yield stable values of viscosity in the Newtonian-like region. This was also reflected in the data reported in Table 8. Before the region of viscosity decrease, an initial Newtonian-like region was observed for various heavy oils by Abivin & Taylor [62] and also for bitumen by Dealy [34] at shear rates lower than 1 s^{-1}. This regime was not investigated in this work.

Another purpose of this investigation was to see whether the shear thinning behavior continued in the lower shear rate region also (1 – 10 s^{-1}). It could be seen from Figure 5.25 that the there was a steeper reduction in viscosity in the 1 – 10 s^{-1} shear domain for the thermally cracked products obtained at different reaction times. The low shear viscosity (compared at 1 and 10 s^{-1}) for these products are given in Table 5.18. Table 5.18 shows that there is an order of magnitude decrease in viscosity (in mPa.s) while moving from 1 – 10 s^{-1}. This transcribed to ~1 – 3 Pa.s decline in the viscosity as compared to a similar decrease for Athabasca bitumen and ~4 – 11 Pa.s that was observed in the same region for Cold Lake bitumen by Dealy. [34]

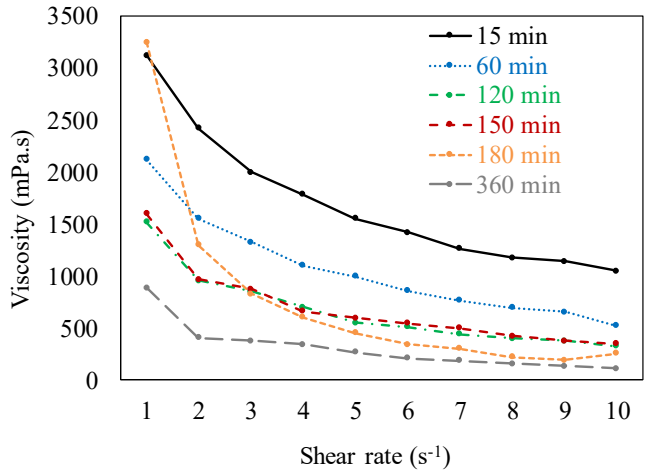


Figure 5.25. Variation of viscosity with shear rate for thermally converted products obtained at 15, 60, 120, 150, 180 and 360 min in the 1-10 s^{-1} region.

Table 5.18. Comparison of low shear viscosities for select thermally converted products.

Reaction time (min)	Viscosity at 10 s^{-1} (mPa.s)	Viscosity at 1 s^{-1} (mPa.s)
15	1049	3123
60	521	2120
120	323	1521
150	351	1600
180	256	3252
360	110	886

The reduction in viscosity at higher shear rates (> 10 s^{-1}) was much lesser (~0.3 – 0.6 Pa.s) as compared to that observed at lower shear rates (< 10 s^{-1}). The rate of decrease in viscosity with shear rate was higher for the less viscous products as compared to those obtained at lower reaction times. The other question is regarding the molecular mechanism responsible for this shear thinning behavior. Is it due to pure de-aggregation or any conformational changes involving spatial orientation of the atoms or stretching of the bonds under shear? This would be elaborated in section 5.4.3.3 in the discussion of this chapter.

The possible time-dependence of viscosity measurements was determined in a separate analysis where viscosity was measured at constant shear rate over time, with a period of no shear in between. Figure 5.26 demonstrates this. No significant variation in viscosity with time was seen with the gap between the least and the largest value being 0.06 Pa.s. These observations were consistent with those of Behzadfar & Hatzikiriakos. [63]

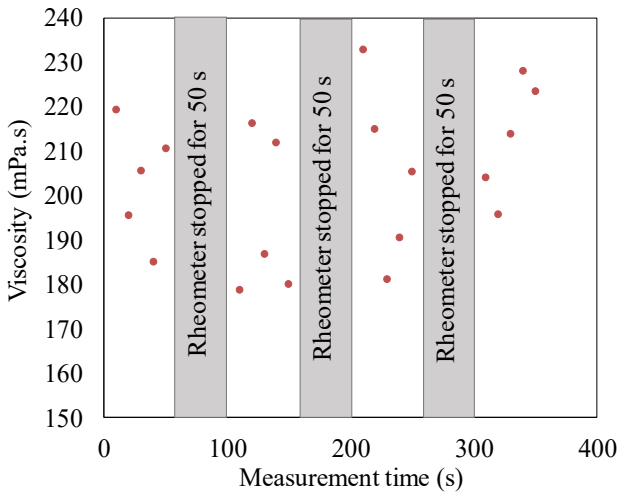


Figure 5.26. Variation of viscosity with measurement time at a constant shear rate of 10 s^{-1} for the thermally converted product at 240 min. The instrument was stopped at intervals of 50 s for the same time and then the process repeated again.

For the first 50 s, the viscosity varied between 219 mPa.s and 210 mPa.s. After the rheometer was stopped for 50 s and started again, the viscosity value dropped to 178 mPa.s and should a paltry increase to 179 mPa.s at the 150th s but reached up to a value of 212 mPa.s within this time domain. During the next two intervals of stopping and starting, the viscosity varied between 232 and 206 mPa.s and 204 and 223 mPa.s, respectively. For practical purposes, these variations can be considered as nearly stable with respect to time at a particular shear rate. This was true for all

other thermally converted products as well. A similar study was conducted by Yañez & De Klerk [20] for the thermally converted product of Cold Lake bitumen at 4 h reaction time for which they found a decrease of more than 200 Pa.s, whereas for other products the variation was considerably lesser (within 1 – 3 Pa.s) and was treated as a part of the standard deviation for that sample.

5.4 Discussion

For a simple liquid, viscosity can be considered as a kind of frictional resistance offered on application of an external force. In most cases, the general notion that heavier the substance, the higher the viscosity and lower the fluidity holds true. However, for the case of complex mixtures like oilsands-derived bitumen, viscosity interpretation is more complicated. There are a number of factors that can influence bitumen and heavy oil viscosity directly or indirectly. Furthermore, the contribution of these factors in modifying the viscosity on thermal reaction could change with time.

The mechanical and thermal history of the feed sample, its source, chemical composition and molecular mass could affect both the feed viscosity as well as the products and play a role in determining the changes happening during pyrolysis. [64] Operating conditions of temperature, pressure and the residence time are the only parameters under the control of the experimentalist during thermal conversion. The reaction causes changes in the inter and intra-molecular associations between the diverse components in bitumen and the reaction chemistry can also modify the phase behavior of the sample. It would be expected that the viscosity would be greatly affected by these factors and thermal treatment was a viable tool to cause these changes.

One of the contributions of this chapter is investigation of the work-up processes that are performed post-thermal conversion to the cracked sample. As highlighted in the Introduction, the speculation from previous observations was that the use of an external solvent to recover the products from the batch reactor can result in changes in some physical and chemical characteristics of the thermally converted sample, separate from the pyrolysis step. The results section (section 5.3) of this chapter outlined some of the findings with respect to the properties of the thermally converted samples when differently natured solvents were used to extract products from the reactor. It was seen that the viscosity was altered in the case of one of the solvents (methylene

chloride) with evidence pointing towards increase a particular type of intermolecular force arising due to the post-processing procedure only.

Viscoelastic and non-Newtonian characteristics of feed and thermally converted products were also reported previously and observed in this study as well. The variation of viscosity with the shear rate of measurement also falls under post-reaction processes that were investigated. Thus, the aim of the discussion part of this chapter was to explain changes in viscosity that were experimentally observed. It was hypothesized that the viscosity is also majorly affected by aggregation propensity of the asphaltenes and a possible theory involving the change in the effective volume of the asphaltene clusters during thermal conversion was discussed. For this purpose, the assumption of a colloidal model for bitumen was adopted from literature.

5.4.1 Source-dependent viscosity of feed bitumen

5.4.1.1 Viscosity range of Athabasca and Cold lake bitumen

Before considering the different processes responsible for viscosity change during thermal conversion, it was worthwhile to concentrate on the feed itself. Bitumen derived from the oilsands in two major geological deposits in the Alberta region of the Western Canadian Sedimentary Basin, i.e. Cold Lake and Athabasca will be the main focus of discussion in this chapter. An intriguing but well-accepted fact in both the heavy oil industry and academia is the sample-to-sample variability in the nature of the oilsands. This is due to a combination of their location, depth and age. [4] In other words, the time spent by the deposits in the formation till recovery by mining or in-situ processes, the temperature and pressure that they are exposed to due to the depth and surrounding materials are critical in determining the characteristics of the extracted bitumen from the oilsands, in addition to the extraction process itself. This would typically affect subsequent processing like upgrading also.

It is not the objective of this chapter to carry out a deep review of the different formations of Canadian oilsands and their depositional processes. Different viscosity values for bitumen feeds obtained from the two deposits measured by various works in thermal conversion is summarized in Table 5.19.

Table 5.19. Feed viscosity values reported in the thermal conversion literature.

Author(s)	Source deposit	T (°C) [a]	Viscosity of feed bitumen (Pa.s)	Shear rate (s^{-1}) [c]
Shu & Venkatesan (1982) [21]	Cold Lake	35.0	13.8	NI [3]
Henderson & Weber (1965) [22]	Athabasca	65.5	5.242	NI
Yañez & De Klerk (2018) [20]	Cold Lake	40.0 60.0	88 9	10 10
Zachariah & De Klerk (2016) [19]	Cold Lake	60.0	9.6	10
Wang et al. (2014) [18]	Cold Lake	40.0 60.0	92 9.2	10 10
Dealy (1979) [34]	Athabasca	42.5	17.5	0.15
	Cold Lake	42.5	22.3	0.20
Present study	Athabasca	40.0	36.2	10

[a] temperature at which viscosity was measured; [b] shear rate at which viscosity was measured; [c] no information available.

Let us look at Cold Lake bitumen first. Viscosity values for feed viscosity reported in recent times were found to be reasonably consistent at ~9 Pa.s and ~90 Pa.s measured at 60 ºC and 40 ºC, respectively. [18–20] But, Shu & Venkatesan [21] recorded a lower value of 13.8 Pa.s at 35 ºC, which would be even lower at 40 ºC and 60 ºC. The value reported by Dealy [34] was also lower than the recent works [18–20] on Cold Lake bitumen measured at 40 ºC. Since feed bitumen was also shown to be mildly non-Newtonian in the 0.1-10 s^{-1} shear rate range [34] and also considering the results from section on effect of shear rates in this study, it could be expected that the viscosity at 10 s^{-1} would be still lower than 22.3 Pa.s (Table 5.19). Though the shear rate data was not supplied in some works, it could be anticipated that nothing higher than 10 s^{-1} would have been used due to the high viscosity of the feed and the instrument constraints like torque limitations that was

highlighted before. The values for Cold Lake bitumen from different sources spanned an order of magnitude when compared at the same reference temperature.

For the case of Athabasca bitumen, a similar variability in viscosity data was noted. A value of 5.24 Pa.s was given by Henderson & Weber [22] measured at 65 ºC, meaning that it would translate to ~80-90 Pa.s if measured at 40 ºC. The temperature dependence of viscosity of Athabasca bitumen was reported to be exponential in nature and described fairly well by the Andrade equation in the study by Schramm & Kwak. [65] It was derived from an extensive database of viscosity data for Athabasca bitumen at temperatures as high as 100 ºC. This was similar to that of Cold Lake values reported in recent times (Table 5.19) but again the shear rate was unknown. On the other hand, the present study and that by Dealy [34] found quite similar range of values for the feed Athabasca bitumen measured at 40 ºC. The value observed in this study (Table 5.2 and Table 5.19) was less than half the value as reported for Athabasca bitumen by Henderson & Weber. [22]

Another reported value for viscosity of Athabasca bitumen at 40 ºC was 14 Pa.s by Scramm & Kwak, [65] which was of the same order of magnitude as reported in this study and that by Dealy. [34] Suncor bitumen is also another type of Athabasca bitumen and since this study used the bitumen supplied by Suncor, it was also worthwhile to compare its viscosity with that of Suncor bitumen reported by Wallace & Henry. [66] They reported a value of 42.3 Pa.s at 25 ºC which when translated to 40 ºC, would be definitely an order of magnitude lower than that obtained in our current study (36 Pa.s at 40 ºC). Thus, even though the samples were obtained at different times (1965 – 2018) and different locations in the two regions, the change in the viscosity of bitumen obtained from the same geological deposit is quite interesting.

The bitumen extraction process from the oilsands might also have had an impact on the viscosity of the final sample used for experimentation. [67] The variance can also be simply due to the different sampling, i.e. it is difficult to obtain a sample representative of the whole type of bitumen. Next, comparing the two kinds of bitumen from Athabasca and Cold Lake, if 88-92 Pa.s was considered as the reference value for Cold Lake bitumen viscosity at 40 ˚C, then according to this work and in literature, [34] the viscosity of Athabasca bitumen is lower, which does not remain in line with the the general notion that Athabasca bitumen has an order of magnitude higher viscosity than Cold Lake. [3] On the whole, a particular reason could not be attributed to these findings of viscosity variations in the various deposits. But they can definitely play a role in affecting their processing and further reactions and some of those are discussed in this chapter.

Instrument variability may have played a role but cannot be claimed to be a definite reason. For example, different instruments have been utilized across the works cited in Table 5.19 like Rheolab QC viscometer obtained from Anton Paar, [18–20] a Mechanical Spectrometer [34] and in some works there was no information about the type of instrument adopted for viscosity measurements. [21,22]

5.4.1.2 Variation of viscosity with composition

Physical factors aside, it was useful to examine the chemical composition of Cold Lake and Athabasca bitumen and relate it to the aforementioned viscosity changes. Bitumen contains hydrocarbons and heteroatoms in the order S > O > N depending on the type of bitumen. An exhaustive review of both kinds of bitumen was carried out by Selucky et al. [68,69] and their key findings are compiled in Table 5.20. On a broad scale, they reported that both bitumen were quite similar in their chemical constitution but on a deeper scale, some differences were apparent.

Table 5.20. Comparison of the Cold Lake and Athabasca bitumen with respect to their chemical composition as reviewed by Selucky et al. [68,69]

Property	Cold Lake	Athabasca
Molecular weight of whole bitumen (g/mol)	Lower (490)	Higher (540)
Molecular weight of nC_5-asphaltenes (g/mol)	Higher (8000)	Lower (4800)
Asphaltene content	Lower	Higher
H/C whole bitumen	Similar (1.495)	Similar (1.486)
H/C asphaltenes	Lower (1.10)	Higher (1.16)
Polar material whole bitumen	Lower	Higher
Saturates – straight chain paraffins	Higher	Lower
Naphthenes	Lower	Higher
Acidic compounds	Higher	Lower
Material boiling above 360 °C	Less	more

It is worthwhile to draw the attention of the reader to the molecular weight values given in the above table, which were determined in accordance with protocols available in the 1970s. The reported molecular weights for the asphaltenes were high. Mass spectrometry and time-resolved fluorescence depolarization (TRPD) suggest that the molecular weights of asphaltenes from different types of bitumen are much lower and in the range 250 – 3000 g/mol as indicated by McKenna et al. [70] but some studies like Patino et al. [46] suggest the range to be having even a lower maximum value of 1200 g/mol. [46,70–72] However, molecular weight measurements were not carried out in this study and in the past works highlighted before. [18–20] This said, the *n*-pentane insoluble material was lower in Cold Lake as compared to Athabasca, which was also the case as observed with this study (19.9 ± 1.6 %wt. - Table 2, Figure 18) in comparison to Yañez & De Klerk's (16.5 ± 0.9 %wt.). [20] In view of the lower molecular weight and asphaltene contents of Cold Lake bitumen, it should be expected that the viscosity also be lower than Athabasca. [3] This is further supported by the fact that the heavier boiling material is more for Athabasca than in Cold Lake (Table 5.20). Still, there were contrasting observations with respect to viscosity values of raw bitumen as given in Table 5.19. This can still be attributed to the much heavier molecular weight of the asphaltenes in Cold Lake, assuming this characteristic of the sample did not change much over time.

The heteroatomic contents of nitrogen and oxygen were more in Athabasca as compared to Cold Lake bitumen. [69] Comparing the results for this chapter with that of Zachariah & De Klerk's, [19] the nitrogen contents were similar but the oxygen content of the Cold Lake was double that of Athabasca (comparing Table 2 and Table 1 in Zachariah and De Klerk's work [19]). It should be noted that oxygen was obtained by difference and not by analysis. Higher the oxygen content, the higher the electronegative groups available for intermolecular hydrogen bonding, which can contribute to an increase in viscosity. [73]

It is known that under the assumption of the colloidal model, resin-type material stabilizes the asphaltenes which results in a reduced aggregation tendency. By means of SARA and TAN measurements that give an idea of the resin and acidic content in bitumen, if a quantitative relationship can be established between concentration of acidic compounds and resin content, and since majority of the acids are reported to be in the form of naphthenic acids and can play the role of amphoteric material in the resins, it can be said that a higher acidic content may stabilize the asphaltenes in Cold Lake and thereby, reducing the aggregating tendency and thus, the viscosity.

Smith et al. [74] also reported that Athabasca bitumen has a high TAN value of 3.5 – 5.5 mg of KOH/g of sample but accurate TAN measurements were difficult in bitumen due to its higher viscosity. They also observed that ESI FT-ICR MS could selectively ionize acidic species in Athabasca bitumen and with the use of ion exchange resins, carboxylic acid content can be determined. But these measurements were not carried out in our work.

Still, no particular reason can be attributed to the increased feed viscosity of Cold Lake over Athabasca bitumen based on their respective chemical compositions. These observations indicate that there is more to viscosity than just molecular weight and asphaltene content.

Another striking inference from the work by Selucky et al. [69] was the higher straight chain paraffin content in Cold Lake bitumen (21.3 % against 18.3 % in Athabasca) and the higher naphthenic content in Athabasca bitumen. They were also of the opinion that Cold Lake bitumen was less biodegraded by natural forces than Athabasca. However, thermal conversion of Athabasca bitumen gave rise to products with increasing fluidity with reaction time at 400 ºC (Figure 5.2), but showed lesser fluidity for Cold Lake at higher reaction times. [18,19]

Bitumen, when subjected to thermal cracking, was shown to display the characteristics of a 'young' crude oil with respect to its reactivity and ease of conversion, despite its long geological history. The high reactivity of bitumen at lower than industrial temperatures of visbreaking ($\leq$ 400 ˚C) has been well documented in literature. [64,75,76] But, the trends in viscosity reductions were quite peculiar and prompted us to investigate the reasons behind the observed changes in product properties depicted in the Results section (section 5.3) of this chapter.

5.4.2 Effect of thermal conversion at 400 ºC on product properties

This section will discuss the results of thermal conversion conducted at 400 ºC, when the products were removed mechanically without the use of any solvent (pathway 'a' in Figure 5.1).

5.4.2.1 Regions of visbreaking and coking

It was important to realize that the reactions explored in this study comprised essentially of two regions: (i) visbreaking for the first 30 min, until which there was minimum coke formation; (ii) coking from 45 min which led to substantial yields of coke (Table 5.5 and Table 5.7). To

capture the whole process, the term thermal conversion rather than visbreaking or coking is used in this chapter.

5.4.2.2 Formation of gaseous products

The gas yield was seen to increase gradually and continuously at all reaction times, which meant that thermal conversion via cracking and bond breakage was certainly taking place at 400 ºC (Table 5.3). The major gaseous product was found to be methane, while $C_2 – C_6$ hydrocarbons were also present (section 5.3.12). One way of forming methane gas is the abstraction of hydrogen from the parent bitumen matrix by methyl radicals. This is quite possible at this temperature as the energy requirement for this process is less (30 – 45 kJ/mol). [77] Like hydrogen transfer, methyl transfer was also a possible phenomenon that was postulated to occur at lower temperatures (300 ºC). [20] At higher temperatures (400 ºC), it was speculated that methyl transfer from light gases evolved in a closed reacting system to the free radical intermediates resulted in reduction of coke yield. [41] This said, methyl transfer does not necessarily facilitate the formation of free methyl radicals in the reaction mixture. Methyl radicals can be formed by cracking of aliphatic side chains attached to aromatics or naphthene rings and this seemed to be quite prevalent with increase in reaction time. Similar observations that methane was a significant gaseous product at temperatures as low as 300 ºC were reported by Jha et al. [78] for Athabasca bitumen. A detailed discussion of the prevalence of methyl transfer and formation of methane is also given in Chapter 6.

It was seen that the sulfur content of the liquid products decreased with reaction time (Table 5.16). Sulfur in bitumen is present mostly in two forms – as aliphatic sulfides with S attached to sp^3 carbon and thiophenes where the S atom is a part of the aromatic ring. Athabasca bitumen consists of ~1 %wt. of aliphatic sulfides, where the C-S bond is most susceptible to cleavage at temperatures of 400 ºC. [3] It would result in the formation of H_2S gas. Since the majority of sulfur is present in aromatic form, the thiophenic sulfur is plausibly removed as a part of the solid coke precipitate instead of H_2S. This corroborates with the increase in solid content and a decrease in sulfur content in the reaction products with reaction time (Table 5.5).

CO_2 was not produced much during the entire time range of thermal conversion conducted in this study. There was only a slight evolution observed at later times. In contrast, thermal cracking of Cold Lake bitumen at milder conditions resulted in CO_2 being released. FTIR data displayed a

peak occurring at 1220 cm^{-1} even at larger times (Figure 5.13b), indicating the presence of some C-O type functional group. But there was complete disappearance of the peak for C=O at 1740 cm^{-1} which indicated that ester-type or anhydride-type functional groups were fully converted at 400 ºC. It is speculated that conversion takes place by hydrolysis, to form carboxylic acid and alcohol/phenol functional groups. The chemical environment of carboxylic acids are known to affect their thermal decomposition process. [79] Thermal decarboxylation is generally viewed as a complex reaction pathway that results in the release of CO_2. [80]

The formation of gases in a closed reaction system like that employed in this work can affect the coke yield. Thermal conversion in this study needs to be distinguished from common visbreaking process and its effect on gas and coke yields due to the higher pressure and lower temperature used. In this context, it is worthwhile to refer to the work by Zachariah and De Klerk [41] where they found that at pressures higher than 1 MPa, the coke yield decreases from 5 %wt. to ~2 %wt. at 8 MPa for thermal conversion of Cold Lake bitumen at 400 ºC and 1.5 h in a semi-batch system. But the results were not majorly altered from pressures of 4 – 8 MPa where the coke yield decreased from ~2.5 %wt. to ~2.2 %wt. However, in a batch system, it was seen that the coke yield was less at ~1.5 %wt. compared to a semi-batch system. The lower coke yield was a benefit of using higher pressure. In a closed system, the vapor-liquid equilibrium (VLE) governs the concentration of the products in the gas and liquid phase and pressure affects the amount of material in each phase. [81] In bitumen conversion, it was suggested that the lighter products contributed to hydrogen transfer to the liquid phase, thus compensating for the hydrogen deficiency of the medium and contributing to a decrease in the coke yield. This could be more prevalent in a batch reactor, where the lighter products were not removed. Another observation was that liquid yield was higher when thermal conversion was performed in a batch as compared to a semi-batch reactor.

Govindhakannan and Khulbe [82] found that the yields of coke from the delayed coking of Athabasca bitumen increased when with an increase in pressure but the lighter products were removed constantly even with an increase in pressure. Viscosity was shown to be not much dependent on coke yield. [41] A continuous increase in gas yield could also be the result of overcracking, which could be checked by varying the liquid and gas volumetric ratio and conducting further reactions. In this work, gas yields also showed significant increase from 45 min

reaction time which coincided with the coking region. An exact quantification of the effects of gas on coke yield and reaction chemistry was not provided in this work.

5.4.2.3 Change in boiling point distribution

Thermal conversion of Athabasca bitumen at 400 ºC resulted in a continuous viscosity reduction (Figure 5.2). A common measure of conversion is the change in the amount of material belonging to a particular boiling fraction in the oil. [28] Henderson & Weber defined a term called scale of upgrading that was very much related to this definition of conversion. [22]

Viscosity of complex mixtures like bitumen is generally evaluated along with the respective distillation profiles to get an idea of the distribution of lighter and heavier fractions. The amount of material boiling above the vacuum residue cut point (> 525 ºC) was found to decrease with time as compared to the feed (Figure 5.21). In fact, around 25 %wt. of the feed did not elute from the column. On-column cracking is something that has been reported to occur if the sample has a significant heavier boiling material. [83] However, the distillation curves could still be compared with each other. There seemed to be a significant increase in the lighter boiling material as the reaction progressed and it might be tempting to ascribe the viscosity decrease to this observation. Wang et al. [18] did not find a noticeable change in the boiling ranges of the thermally converted products after thermal cracking conducted at 360 ºC and short reaction time of 30 min on Cold Lake bitumen. The product was found to be only slightly lighter than the feed. In contrast, Ball [76] had noted a drastic change in the distillation profile when Athabasca bitumen was subjected to thermal cracking at 360 ºC for 30 min with the production of a significantly lighter product. Similarly, after thermal conversion at 400 ºC in our current study the material in the naphtha and distillate boiling range in the products were significantly more than in the Athabasca feed (Figure 5.21).

5.4.2.4 Change in refractive index and density

The decrease in density and refractive indices of the products also support this observation (Figure 5.6). A change in the refractive index implied a modification of the composition with thermal cracking. Interestingly, aromatic hydrogen in the liquid products showed an increase from

around 10 % wt. to around 17 %wt. over 24 h of reaction time (Figure 5.15). This increase in aromatic hydrogen was not continuous but exhibited a global maximum of 22.7 %wt. at 360 min and a local maximum of ~20 %wt. at 90 min. This was in contrast to the observed decline in refractive index of the products, which is affected by composition, molar mass and density. Paraffinic oils have the lowest refractive indices and aromatic ones have the largest for the same boiling range of compounds.

In a related previous study on Cold Lake bitumen, the aromatic hydrogen was found to increase to 20 %wt. within the first 300 min of reaction at 400 ºC along with a corresponding gain in refractive index of the products from 1.58 to 1.66. [19] They also observed a concomitant viscosity increase, which they attributed to the formation of free radical addition products because the hydrogen-to-carbon ratio did not decrease at the times when refractive index gained in value.

5.4.2.5 Change in composition

In the present study (Figure 5.15), the aromatic hydrogen increased only to 13.3 % at 300 min reaction time as compared to 20 % observed for Cold Lake bitumen. [19] Data from FTIR showed an increase in the mono-substituted aromatic content in the liquid products (Table 5.15). There was an initial decrease in the hydrogen attached to benzylic carbon but overall showed a net increase of ~ 4 % as compared to the feed. Though the benzylic hydrogen does not give an idea of the number of substituents on an aromatic ring, the FTIR data indicated that the mono-substituted aromatic concentration was high in the liquid products.

In a previous study where FTIR data of the thermally converted products of Cold Lake bitumen was analyzed through a technique called self-modeling curve resolution (SMCR), [84] a mechanism was proposed that could potentially explain the observed increase in the extent of substitution in aromatics. The speculated reaction chemistry was the replacement of the aryl C-H bond with an alkyl group that could result in a ring closure and it can be considered to be present in the form of an intermediate where the nearby benzylic carbon on the adjacent aromatic ring replaced the aromatic hydrogen to product a disubstituted aromatic. This process was also responsible for the observed increase in CH_2/CH_3 ratio which pointed towards free radical addition as a cause of viscosity gain. The key aspect to look for was the plausible link between the reaction chemistry and the viscosity as a result of thermal conversion of different types of bitumen. The

possible factors contributing to viscosity change were bond formation/scission that affected the chemical composition and the nature of the feed bitumen itself (discussed in the previous section). A further elaboration of these aspects is provided in Chapter 6 along with detailed explanation of the proposed reaction mechanisms for thermal conversion of Athabasca bitumen.

However, in our current study, Table 5.15 showed that the extent of mono-aromatic substitution of the thermally converted products were higher than the feed at all reaction times. This can be anticipated to occur only when an aryl C-C bond gives way to an aromatic C-H bond (effectively a decrease in di- and tri- substituted aromatics). Also, considering the variation in the density, refractive index, distillation profiles and decrease in the extent of aromatic substitution of all the products compared to the feed (indicated by the spectroscopic data), the formation of lighter products indeed ensued during thermal conversion of Athabasca bitumen at 400 ºC in a batch reactor. Viscosity reduction was a result of the diluent-like effect of the produced lighter fractions present in the liquid products. This should not be interpreted as a non-formation of heavier products. Both processes occur at all stages of thermal conversion but cracking of heavier molecules to lighter ones probably dominated in this case, unlike observed for Cold Lake bitumen previously. This might also be the reason for the observed viscosity decrease in the work by Henderson & Weber [22] on Athabasca bitumen where the decline in the heavier boiling products was reported as an increase in the 'scale of upgrading'.

It was interesting to see the trend in hydrogen-to-carbon ratio of the liquid products in consideration with the *n*-pentane insoluble content. Just looking at the overall change, there was a decrease from 1.49 to 1.46 over 24 h of thermal conversion in the H/C ratio for Athabasca bitumen (Figure 5.20). This was a small change in comparison with the extent of hydrogen depletion experienced by the thermally converted products of Cold Lake bitumen, [19] where the H/C ratio fell from ~1.45 to 1.09 after 360 min. Another difference was that in this study (Figure 5.20), the trend was non-monotonic with the lowest point being 1.43 attained at 90 min and 240 min but all values were lesser than that of the feed. This can be explained as follows. There was a significant formation of gaseous products which contained the hydrogen-rich molecules (Table 5.5). Without the precipitation of solid coke, the liquid products would expect to be very hydrogen depleted. However, it could be anticipated that the solid coke was hydrogen-deficient due to the maintenance of H/C ratio of the liquid products. Typical H/C molar ratio in petroleum coke obtained from

delayed coking process is close to lesser than 1. [85] No proof for the coke composition in our current study was provided but comments can be made from the observed findings in this study.

5.4.2.6 Change in asphaltenes content

The *n*-pentane insoluble content seemed to be reflective of the hydrogen-to-carbon ratio of the liquid products. Except initially before the onset of coking at 45 min, the liquid products always contained lesser *n*-pentane insoluble content than the feed (Figure 5.18). It has to be kept in mind that the calculation was based on the liquid products, whose yield kept decreasing (Table 5.5). This material is referred to as the asphaltenes, which is on a broad scale, a solubility class.

In a study on bitumen upgrading by cavitation processing conducted by Mohapatra & Kirpalani, [86] it was reported that the continuous viscosity reduction was a repercussion of the asphaltene content in the liquid products obtained after sonication. Shu & Venkatesan [21] reported a monotonous decrease in asphaltenes content along with a continuous decrease in viscosity of Cold Lake bitumen though any specific correlation between these properties were not explicitly mentioned. Other studies on bitumen pyrolysis at 400 ºC all reported a monotonous decrease in asphaltenes content though viscosity data were not given. [87,88]

However, the non-monotonic trend in *n*-pentane insoluble content in our case (Figure 5.18) implied that the viscosity was not a direct consequence of the change in asphaltene content but rather a combination of more factors as mentioned in the above paragraphs. There are quite a few studies in literature that are in support of this observation. Brauch et al. [89] observed that the product stability was not exactly correlated with the asphaltenes content during thermal conversion of bitumen. They also were not supportive of the notion that higher asphaltenes content were a cause for higher viscosity. It was also noted in their study that a more asphaltene-like material was formed from low sulfur feeds. A combined interpretation of other works [90,91] showed that viscosity, rather than asphaltenes content varied with the boiling range of bitumen or any heavy crude. In contrast, Kapadia et al. [92] compiled and reported the yields of asphaltenes, coke and gas obtained from the pyrolysis of Athabasca bitumen conducted by various researchers in the past. At all reaction times, the asphaltenes content decreased monotonically with reaction time at 360 ºC and 400 ºC with some formation of H_2 gas also noted. In their modified kinetic model, they believed that asphaltenes reacted to not only yield coke but also lead to formation of heavy molecular weight

gases that included compounds with greater than 2 carbon atoms. The belief that MCR content was also a direct function of asphaltenes content [52] of the sample need not be necessarily true as observed in this work (refer Figure 5.18 and Figure 5.19) since the asphaltenes react to give other products as well. But, both the coke content and the MCR contents increased concomitantly, meaning that MCR content was a good measure of the coking tendency of Athabasca bitumen samples. However, the overall increase in MCR correlated with the overall decreasing trend in the H/C ratio (Figure 5.20). The H_2 formation was difficult to imagine because in hydrogen transfer by abstraction and disproportionation occurring as a part of the pyrolysis mechanism, free hydrogen radicals (their combination yielding H_2) are rarely produced. [93]

5.4.2.7 Summary of different analysis test methods used for bitumen and heavy oil characterization

This section provides some useful and advanced characterization methods for bitumen and heavy oils employed by researchers in recent times. Some of the techniques were not used in the context of analyzing the thermal cracking products but only in the characterization of the feed bitumen/heavy oil. Also, though some of these methods were not employed in our work, they definitely provide areas of improvement and scope for deeper insights on the reaction chemistry. Three major types of analytical methods are discussed here: microscopy, mass spectrometry and thermal analysis methods.

(i) Microscopic techniques:

These techniques have the ability to detect surface properties like morphology, topography, stiffness, elasticity, and interaction forces of adhesion and repulsion between surfaces. These are important characteristics for asphaltenes that enables one to analyze changes at the micro-scale level. Four major types of microscopy are discussed here: atomic force microscopy (AFM), scanning electron microscopy (SEM), scanning tunneling microscopy (STM) and surface force apparatus (SFA). Major differences between SEM, STM and AFM are that the former two methods require the measurements to be conducted in vacuum only but samples in AFM can be present in a gas, liquid or in vacuum. The elemental make-up of a material can be analyzed by SEM but AFM

can only assess the surface topography and other mechanical properties like elasticity, stiffness and has a better resolution. The power of these techniques has been put to use by researchers to study complex molecules like asphaltenes in bitumen and heavy oils.

Hou et al. [94] used AFM to characterize bitumen morphology and found the existence of two phases, namely, a bee-like structure that consisted of bumps dispersed on a second continuous phase. These observations were in agreement with other studies on bitumen and asphalt as well. [95,96] To study further mechanical phenomena at the micro-scale level, they conducted molecular dynamic simulations as did Guo et al. [97] for studying the properties of asphalt. Bitumen is considered as a mixture of 3 major domains, the asphaltenes, the naphthene aromatics and the saturates and a phase-field variable is associated with each of these phases, using which a free energy functional is constructed. Through small angle neutron scattering (SANS) experiments, Schmets et al. [98] verified that a similar kind of phase separation as in the surface existed in the bulk as well.

The relationship of different phases with their chemical composition was established by Fischer and Cernescu, [99] where they made use of a derived technique from AFM called friction force microscopy (FFM), scanning thermal microscopy (SThM) and a scattering near-field IR microscope (NeaSNOM). FFM was able to achieve a higher contrast than AFM between the two types of phases detected, i.e. peri-phase that is at a lower elevation than the perpetua phase. The peri phase, partly consisting of paraffins and cyclic saturates can crystallize on quenching and was shown to have a lower friction coefficient than the more viscous perpetua phase that comprises of aromatic compounds. SThM was performed at 0 ºC to prevent melting of the crystalline phase so as to reduce noise. Thermal contrast between the phases was obtained due to difference in heat conductivity of the crystalline peri phase and the amorphous perpetua phase. From SNOM, it was seen that heteroatomic S=O groups (IR vibrations at ~1037 cm^{-1}) were associated with the aromatic moieties and carbonyl groups (IR vibrations at ~1698 cm^{-1}) were found to be present along with the peri-phase and sulfoxide concentration appeared to increase faster than carbonyl groups during ageing of the bitumen. This approach was found to be better than using chemical force microscopy (CFM) adopted by Allen et al. [100] to study the chemical composition of the different phases in bitumen, though CFM seemed to reveal that there was a direct relationship between the polarity of the overall asphaltenic phase (bee structure) and the saturate content in bitumen.

To address the debate over asphaltenic structure in crude oils and bitumen, a breakthrough study was conducted by Schuler et al., [101] where individual molecules of asphaltenes were isolated by flash-heating from a silicon wafer on the substrate and analyzed by AFM and complemented by orbital imaging through STM that could reveal the polycyclic aromatic hydrocarbon structures which are important sites for intermolecular interactions. They were able to differentiate asphaltenes derived from coal and crude oils and found that the primary structure consisted of aromatic cores attached to side chains as proposed by the Yen-Mullins model [102] was valid but archipelago-type structures were also existent in petroleum-derived asphaltenes, which had longer alkyl side chains as well. This detail of structure could be invaluable to thermal cracking process to analyze the products and develop insights into reaction chemistry.

SFA measurements to investigate the surface interactions between asphaltene molecules were conducted by Zhang et al. [103] They observed that DLVO-type electrostatic forces could not describe asphaltenes extracted from Athabasca bitumen but more steric interactions were responsible for repulsion. However, DLVO-type interactions were existent between the AFM tip and the asphaltene aggregate molecules and this combined with the SFA results demonstrated that the interactions at the micro- and nano-scale between asphaltenes and their surrounding media are of colloidal nature.

(ii) Mass spectrometry (MS) techniques

The advantage of MS approach is that fragmentation in terms of molecular weight is conducted under high vacuum, where the chance of unwanted secondary reactions is significantly lower. It rests on the principle of ionizing the parent sample, fragmenting each ion into recognizable smaller molecules and then detecting them. Different types of ionization are used, for example hard ionization techniques like electron ionization (EI) and soft ionization methods like electron-spray ionization (ESI), chemical ionization (CI), matrix-assisted laser desorption/ionization (MALDI). A particularly common and useful method in identifying the different compounds in complex mixtures is gas chromatography coupled with mass spectrometry (GC-MS). To investigate whether free radical reactions occurred during further treatment of cracked naphtha, obtained from thermal cracking of oil sands-derived bitumen, the reaction products were analyzed using GC-MS in the work by Uzcategui et al. [104] The challenge for in

directly analyzing bitumen and its thermally cracked products via a gas chromatograph is the probability of the heavier components not eluting and clogged in the columns is very high, which would require immense effort to wash out.

Recently, MS approaches with different methods for ionization and fragmentation have found applications in the quest for unravelling the molecular structure of asphaltenes. Island-type structure which consists of a single aromatic core of 7-10 fused rings and attached alkyl side chains have been reported to be dominant in asphaltenes from different techniques like MS with laser desorption and laser ionization, [70,105] and in combination with molecular dynamics, [106] AFM, [107] time-resolved fluorescence depolarization (TRPD), [108] and X-ray diffraction (XRD) [109] methods. Another school of thought was the existence of archipelago-type structures which was hypothesized based on a number of observations including heterogeneous nature of asphaltene aggregation, [110] ability to entrap solvent in voids, [111] and most importantly from the chemical compound classes of the upgraded products especially from thermal cracking which leads to a variety of products including gases, some coke, middle and heavy distillates consisting of 1-4 ring alkyl aromatics. [112,113] On the other hand, if island was the dominant structure, the only cracked products would be gas, coke and light naphtha-range alkanes, which was not the case. [113]

Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS), has gained a lot of application in this area due to its higher sensitivity and level of molecular detail achieved. Patino et al. [46,114,115] used FT-ICR MS combined with positive-ion atmospheric pressure photoionization (APPI) and an infrared multiphoton dissociation (IRMPD) for determining the asphaltene structure derived from a South American heavy oil, Wyoming deposit and Athabasca bitumen in their series of works devoted to Asphaltene Petroleomics. In their analysis of the South American heavy oil, two fragmentation pathways were found: one indicated dealkylation as the major route that was evident from the plot of double bond equivalence (DBE – number of aromatic rings + number of double bonds) vs. carbon number plot with a decrease in the carbon number in the fragments but minimal decrease in DBE and this indicated an island structure; the other pathway indicated loss of carbon number as well as a decrease in DBE which pointed towards an archipelago structure. They were also able to distinguish between *n*-pentane and *n*-heptane asphaltenes, where C_5-asphaltenes had fragments with DBE in the range of 18-27 resembling island-type motifs but C_7-asphaltenes had fragments of much lower DBE (4-15) indicating archipelago motifs. In conlusion, they were of the opinion that island motifs were easily detectable in MS techniques due to their

higher ionization efficiency and lesser tendency to form nano-aggregates but extensive purification or extraction is required to identify archipelago-type structures in the asphaltenes due to their higher potential for aggregation, which was also in agreement with another similar study by Podgorski et al. [116] on deasphalted heavy distillate.

In their following work, Patino et al. [115] investigated the differences between asphaltenes extracted from Wyoming deposit and Athabasca bitumen using the same characterization techniques for MS but conducting the analysis on both whole asphaltenes and their stepwise extrography fractions with a number of solvents including acetone, heptane, toluene, toluene/tetrahydrofuran (THF), THF alone and THF/MeOH. In the analysis of whole asphaltenes, they reported a low abundance of lower molecular weight (m/z) fragments than the precursor ions and other fragments having DBE close to the polycyclic aromatic hydrocarbon (PAH) limit of ~20, indicating island-type structures. Athabasca bitumen had a dramatically different MS spectra where a lower molecular weight distribution was obtained in the m/z = 100 – 300 region along with dealkylated fragments, indicating archipelago-type parent structures. This gained support from recent AFM studies which suggested that the aromatic cores consisted of 4 fused aromatic rings rather than previously postulated 7 rings, though the nature of bridges were difficult to determine. [101,107] They also recommended a new definition for asphaltenes due to this discrepancy on the asphaltenes structure confirmed from analysis on the extrography fractions where they found that polar fractions had a lower DBE and higher carbon number in the homologous series as against the expected higher aromaticity and smaller side alkyl chains. In conclusion, advanced MS methods with enhanced ionization and fragmentation methods have the potential to provide valuable insights into the chemical composition of thermally cracked samples as well and could be used to build complex reaction pathways.

(iii) Thermal analysis methods

Differential scanning calorimetry (DSC) and thermogravimetric analysis (TGA) are two common methods used for thermal characterization of complex mixtures like bitumen. DSC gives information on important thermal events like glass transition, melting, crystallization associated with changes in the respective heat flows for the sample, while TGA provides information on mass loss over the temperature range that the sample is subjected to. Thermal conversion experiments

can be simulated in the TGA to obtain the product yields in different boiling ranges as performed in the work by Prado and De Klerk to investigate the effects of bromination on bitumen, asphaltenes and maltenes. [32]

TGA can also be coupled with FTIR and MS to obtain the chemical composition of the gaseous products. [117] This type of analysis was also performed with biomass and coal by other researchers. [118,119] In the work by Zhao et al., [117] pyrolysis of bitumen supplied by Cenovus Energy Inc. resulted in three stages of mass loss at < 110º C, 110 – 380 ºC and 380 – 600 ºC. Gases were composed of alkane and alkene hydrocarbons and inorganic compounds like H_2O and CO_2. They also employed a pyrolysis-GC/MS system to study the thermal reaction pathway at each stage of cracking. Lower molecular weight compounds evolved in the first stage at lower temperature with cleavage of weak heteroatomic bonds. The second stage resulted in higher molecular weight compounds due to cracking of longer side chains, whereas in the third stage, compounds ranging from lower to higher molecular weights were obtained in equal proportions.

5.4.3 Effect of post-thermal reaction procedure on viscosity of thermally converted products

5.4.3.1 Effect of extraction solvent on product viscosity through molecular association

The dramatic increase in viscosity when the reaction products were removed from the batch reactor by dissolving in methylene chloride (CH_2Cl_2) (Figure 5.4) was a case in point to address. More so, the minimal change in viscosity observed in the instances where toluene was used as the extraction solvent (Figure 5.5) called for a deeper investigation. Interestingly, Wang et al. [18] also noted a 15 % gain in viscosity of raw Cold Lake bitumen when washed and dried with CH_2Cl_2. Though the distillation profiles of the feed before and after methylene chloride addition and removal were not obtained, it was suspected that some of the lighter material was carried away in the drying process and this caused a rise in viscosity. Since there is no naphtha range material in the feed bitumen, this appears to be an implausible explanation.

Prado and De Klerk [32] initially suspected that halogenation of bitumen would disrupt the π- π aromatic stacking in the asphaltenes but eventually found an increase in the hardness of bitumen and asphaltenic material. It was also postulated based on the changes in physical properties that

after bromination, that occurred both on aromatic and aliphatic carbons, further C=C bonds can be formed through dehydrohalogenation (loss of Br and H bonds from neighboring carbon atoms), which can undergo subsequent addition reactions that could contribute to an increase in the molecular mass. [120] Increase in molar mass can also be attributed to the addition of bromine atoms to the molecule. They also noticed a decrease in the boiling point of brominated asphaltenes which could be due to intramolecular hydrogen bonding. In essence, the introduction of halogens (apart from the already present halides in bitumen) had deleterious effects on bitumen properties.

The origin of the concern regarding the effect of halogenated solvents on thermally converted product properties in our study was the presence of remnant solvent even after sufficient rotary evaporation (refer Appendix for quantification procedure). The products could not be recovered completely solvent-free with around 1 – 2 % wt. remaining in the case of both toluene and methylene chloride. There was also a corresponding mass loss observed in the products after evaporation as shown for some select products (Table 5.14). It was not clear whether the reason for this loss was due to actual loss of lighter material or it was a result of the mechanical transfer process between the different glassware. It appeared that there was not much loss of lighter material during evaporation as the distillation profiles of the straight thermally converted product and the solvent-extracted products were similar in nature (Figure 5.22). There was only a minor increase in the heavier fraction as the methylene chloride-extracted product had 3 % less material that boiled at 687 ºC than the straight product. The lighter fractions were not affected.

Increase in density (Table 5.12) with minor change in the refractive indices (Table 5.13) of the methylene chloride-extracted products as compared with the case of solvent-free products pointed to the direction of a physical change rather than a chemical or compositional change. A chemical change could be anticipated if the introduction of the halogen atoms into the sample followed a rigorous method as in the case of Prado & De Klerk. [32] Toluene, being an aromatic and a less polar solvent than the aliphatic methylene chloride did not alter the physical or chemical properties of the thermally converted samples much. Since the evaporation process was done at much lower temperatures (55 ºC) than the thermal conversion temperatures, possibility of inter- and intramolecular hydrogen bonding and other types of complex formation seemed plausible.

Halogen atoms in C-X (X represents a halogen) bonds are electronegative enough to interact with molecules containing hydrogen atoms attached to more electronegative atoms such as O and N to form hydrogen bonds which are quite stable at temperatures in the range of 25 – 55 ºC. Carbon

– halogen bonds can also form complexes in the form of adducts with carbonyl compounds due to their polarizability. Though esters and anhydride-type compounds were shown to be converted during thermal conversion, the presence of carboxylic acids and some amount of ketones cannot be eliminated in the thermally converted products. Evidence for the aforementioned kinds of interactions has been documented by Smith, [31] especially for the case of polyhalogenated solvents.

In our study, the FTIR and ^{1}H-NMR data provided some evidence for the presence of stable hydrogen bonds in the methylene chloride-extracted thermally converted products. Detailed discussion is provided with reference to changes observed in FTIR spectra. FTIR spectra is affected in the following way. The stretching frequency of an IR active bond is determined by the force constant of the bond and the atomic mass of the atoms involved in bond formation. [40] When a hydrogen bond forms (Y-H --- X-C where Y is the electronegative atom), it weakens the force constant of Y-H as well as of the C-X bond, thus shifting the absorption wavenumber of the participating bonds to a lower value. A broadening of the absorption bands can also occur. [121] These changes were briefly discussed in the Results section and were illustrated in Figure 5.12 and Figure 5.13 for select thermally converted samples obtained after reaction for 15 min and 30 min, respectively.

For the comparison to be meaningful, the spectra of the solvent-extracted sample was compared with that of pure solvent as well as the solvent-free sample in terms of wavenumber shifts and line shape changes of the C-Cl stretch and the C-H bend of the CH_2Cl_2 molecules. In the 900 – 600 cm^{-1} wavenumber region, the doublet for methylene chloride was due to the asymmetric stretch (744 cm^{-1}) and the symmetric stretch (717 cm^{-1}) (Figure 5.12a). The Ar-C-H deformations in the mono-substituted and di-substituted aromatics present in the thermally converted sample absorb at 727 & 811, 865 cm^{-1}, respectively (Figure 5.12b). Surprisingly, in the solvent-extracted sample that contained 1.3 %wt. CH_2Cl_2, the peaks are shifted. The peak at 734 cm^{-1} falls right in between the mono-substituted aromatic C-H bends in the solvent-free sample and is 10 cm^{-1} less than the main asymmetric C-Cl stretch occurring at 744 cm^{-1} in pure CH_2Cl_2. Though it may be a result of overlap of these absorptions in this region, it can be considered that the wavenumber of the C-Cl asymmetric stretch shifted to a lower value. Furthermore, the bands appear to be broadened, similar to hydrogen bonded peaks in asphalt shown by Petersen. [121] From this evidence, it can be inferred that the C-Cl bonds from the remaining methylene chloride in the thermally converted sample appear to be intermolecularly hydrogen bonded with molecules that

act as proton donors. The structure of these molecules was not clear from the spectra as absorptions in the 3700 – 3000 cm^{-1} region corresponding to the most probable donors of O-H and N-H were quite noisy and could not be clearly resolved. The shift of the Ar-C-H bends could not be attributed to the hydrogen bonds as Silverstein [40] noted that bending wavenumbers shift to a higher values when in the complexed state, but in our case they seem to have shifted to a lower wavenumber (Figure 5.12c). Nevertheless, peak broadening occurred.

There was no shift observed in the aliphatic C-H stretches (3020 – 2700 cm^{-1}) for the methylene chloride-extracted molecules. In fact, NMR spectra provided evidence of possible involvement of aliphatic C-H bonds being involved in complexation due to line broadening and changes in chemical shifts observed for products at different reaction times (Figure 5.17). The findings with respect to the C-H deformation bands in CH_2Cl_2 were quite interesting (Figure 5.13). The peak for C-H bend occurred at 1265 cm^{-1} in pure methylene chloride but was not seen in the straight thermally converted product, shown for 30 min in Figure 5.13b. In addition, peaks indicating alcoholic and acylic C-O stretches occur at 1216 and 1228 cm^{-1} in the solvent-free product. In the solvent-extracted sample, the C-O stretching bands are broadened with lesser resolution, whereas the absorption for the C-H bend arising from CH_2Cl_2 appeared to be split into equally intense peaks at 1259 cm^{-1} and 1274 cm^{-1}. It is exactly unclear as to what phenomenon caused this shift but the higher wavenumber peak appeared to be a result of the shift in bending wavenumber due to hydrogen bonding.

Evidence from Smith [31] suggests that the hydrogen atoms in CH_2Cl_2 are rendered acidic by the presence of two Cl atoms due to inductive effect. In the work by Allen et al., [122] it was shown that C-H bonds from polyhalogenated solvent molecules like chloroform and dichloromethane form hydrogen bonds with electronegative atoms like N, O, S and even with halogens with a very high propensity of 82 % and 77 % for possible structures, respectively, even if the stronger acidic groups like N-H and O-H were present. They suggested that this bonding could contribute to molecular aggregation in crystal structures, which are quite different from bitumen matrix but still the possibility cannot be ruled out. They used the Cambridge Structural Database (CSB) for this study and a previous study on acetylene acidity as well. The line broadening in ^{1}H-NMR for the methylene-chloride separated products as compared to the solvent-free samples (Figure 5.16) could be attributed to higher viscosity or the existence of significant amount of paramagnetic

species, as was seen in the increase in spin concentration for the methylene chloride-separated products (Figure 5.9).

All this being said, there are certain requirements for hydrogen bonding to occur. It can be envisaged as an association between a proton donor molecule and an electron donor species. The hydrogen atom from the proton donor needs to be sufficiently acidic and is beneficial if it is covalently bonded to an electronegative atom. On the other hand, the electron donor should be sufficiently basic to participate in the hydrogen bond. Spatial orientation of the atoms should be favorable for the *s*-orbital of the hydrogen to effectively overlap with the *p*- or π- orbital of the electron donor atom. If the bond results in a ring formation, five membered rings are entropically favored over six membered ones even though the Y-H---X-C might be strained. Ideally, the 4 atoms (Y, H, X, C) are required to be collinear for a strong hydrogen bond. The energy change associated with the bond formation is proportional to the extent of orbital overlap (which in turn depends on collinearity) and the entropy change depends on the probability of the electron donor approaching the proton. It might be quite difficult to imagine all these factors coming together in a complex species like bitumen but it has to be kept in mind that thermally converted bitumen can be lighter and chemically less complex than feed bitumen. Figure 5.27 illustrates some of these aforementioned possible scenarios.

Figure 5.27a describes a possible scenario of hydrogen bonding between the two hydrogen atoms in methylene chloride with N and O and interestingly, it results in the formation of a favorable 5-membered ring. In addition, since the lone pair on nitrogen and one lone pair of oxygen is not involved in aromatic delocalization of the furan-derivative ring, their electron donating capacity increases towards the acidic hydrogens from methylene chloride. These type of ring formation in hydrogen bonding was also shown by Ding et al. [123] in their study of halogen bonding. The ability of sulfur to donate electrons to form hydrogen bonds with the acidic hydrogen atoms from methylene chloride was also suggested. [124] The orbital overlap shown in Figure 5.27b should not be confused with that of a sigma bond since the overlap in a hydrogen bonding is weaker but a necessity for the formation of the bond. The importance of a collinear arrangement of the N-H---X-C atoms is shown in Figure 5.27c where the hydrogen coming from an aromatic amine is favorably aligned to form a hydrogen bond with one of the Cl atoms attached to CH_2Cl_2 even though this is less probable than scenario (a). The aromatic nature of the amine renders the H atom more acidic. The interaction of Cl and H attached to a carbon atom (as in the benzylic carbon in

Figure 5.27c) cannot be ruled out as these kind of interactions, though weak, were shown to exist in the work by Brammer et al. [125]

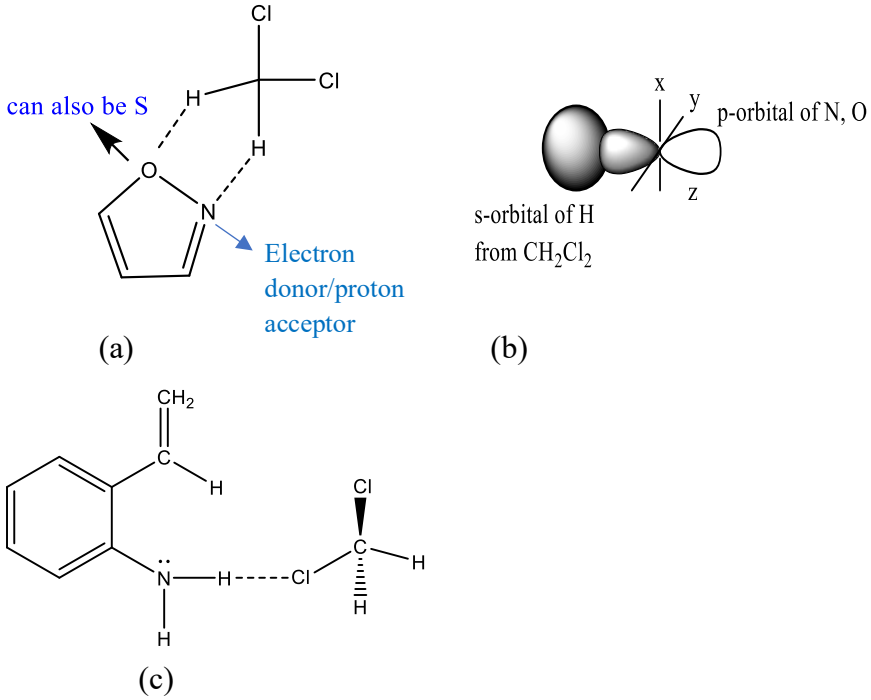


Figure 5.27. Types of hydrogen bonding possible in methylene chloride-extracted products: (a) Most probable scenario of the hydrogens from CH_2Cl_2 being involved in bonding with electronegative atoms in a naphthene aromatic compound; (b) orbital overlap in the formed hydrogen bond; (c) depicts the collinear arrangement of atoms participating in a possible N-H---Cl bond as explained in the preceding paragraph.

Dichloromethane is a di-halogenated compound and consists of 2 electronegative C-Cl bonds along with 2 C-H bonds with much lesser difference in electronegativity, that produce a dipole moment of 1.6 D. Toluene is a non-polar organic solvent with a dipole moment of 0.36 D. On similar lines, the potential to form stable benzyl radical from toluene makes the methyl H also acidic. But, the lack of a stronger electronegative atom and the bulkiness of toluene as compared to dichloromethane makes CH_2Cl_2 better suited to form hydrogen bonding with the bitumen matrix. The absence of hydrogen bonding with toluene was supported by negligible change in the FTIR or ^{1}H-NMR spectra after toluene was evaporated from the products.

Even though the heteroatoms may be concentrated more towards the aromatic core as the reaction proceeds, the possibility of methylene chloride gaining access to N, O sites is higher due

to its small molecular size. The whole implication of these kinds of interactions was that they were shown to have a drastic effect on increasing the viscosity. [73,126] So, it was theorized based on the observations of increased viscosity combined with minor changes in the product characterization results that molecular associations of this kind may cause a rise in viscosity for methylene chloride-extracted products. It should be noted that these interactions would be disrupted at thermal cracking conditions, [30] albeit the absence of solvent during reaction.

Apart from hydrogen bonding, another kind of interaction possible was due to polarization of the C-Cl bond. The C-Cl bond can be associated with a proton donor due to a partial negative charge developed on the Cl atom based on electronegativity difference with carbon. It cannot be considered as an entirely electrostatic or a purely hydrogen bonding type of interaction, thus somewhere in between. Evidence of this type of interaction was proposed by Bellamy et al. [127] and can possibly contribute to increasing the viscosity. Koyo [128] noticed changes in the Raman spectra of the C-Cl bond stretching in CCl_4 when the polar compound NH_3 was added, indicating possible association due to polarizability of the C-Cl bond though the compound as a whole is non-polar. Polyhalogenated solvents have also been found to form polar adducts with carbonyl (C=O) groups through the activated C-H bond. [129]

Halogens can also act as electron acceptors by forming halogen bonds with electron donors such as nitrogen, oxygen, sulfur and the key to such bond formation was polarizability as given in the work by Ding et al. [123] A similar type of interaction where complexes were formed between crown ethers and halogen containing groups like halogenides was documented by Vogtle and Weber. [130] More importantly, Brammer et al. reported that as much as halogens attached to metals were good hydrogen acceptors, this ability decreases when the metal is replaced with carbon. This particular case is relevant for our study as we also discuss about hydrogen bonding involving C-Cl. But other works on similar areas like McBride et al. [131] and Davey et al. [132] highlight the importance of interactions involving C-Cl bonds of the type O-H---Cl-C in the packing nature of solids and nucleation during growth of crystals. Though the application was different, one can imagine a similar kind of interaction, though weak, playing a role in influencing solvent-extracted bitumen viscosity. On the other hand, some interactions of the type C-H---Cl-C were reported as being of the Van der Waals type rather than hydrogen bonding in the work by Thallapally and Nangia. [133] Also, Allen et al. [122] reported that hydrogen bonding involving Cl atoms in solvents

like chloroform and methylene chloride had only a 6 % propensity to hydrogen bond with N-H and O-H groups during molecular aggregation in crystal structures.

Thus, considering all the different postulated and reported observations involving the C-Cl bond, it can be stated that the actual existence of hydrogen bonding with Cl atoms from polyhalogenated solvents is still unresolved. At the same time, there is sufficient evidence in the literature to state with confidence that hydrogen bonding involving the hydrogen atoms in solvents like methylene chloride in complex mixtures like bitumen, exists and can possibly impact viscosity. This is also shown from FTIR spectra in our work as discussed before. Therefore, post-reaction processes can indeed affect product viscosity.

5.4.3.2 Effect of extraction solvent on free radical spin concentration and its possible impact on viscosity

The use of methylene chloride to extract the reaction products from the batch reactor affected the spin concentration of the paramagnetic species in the thermally converted sample. This is illustrated in Figure 5.9. Toluene did not alter the spin concentration by much (Figure 5.10) and since the characterization procedure was the same for both kinds of samples, external factors like exposure to air during transfer of samples could not explain the observed difference. This section attempts to provide a reasoning for the observed increase in the free radical content in the case when methylene chloride was used to extract the products.

A molecule may be paramagnetic depending on the type of spin state that the atom capable of possessing the free radical exists in. These are singlet, doublet and triplet states that also determine whether the species can be detected by EPR technique or not. The singlet biradical represents two paired electrons of opposite spin and thus does not provide an EPR signal. Only the doublet monoradicals (1 unpaired electron) and triplet diradicals (2 unpaired electrons) are EPR active. Biradicals are special cases of diradicals where the unpaired electrons are present on different carbon centres usually concentrated in an aromatic ring system and delocalized with the conjugated π-bond electrons. [4] The key parameters that characterize a hyperfine line are the peak-to-peak width also called line-width (ΔH_{p-p} in Gauss), the Lande g-value (g) and the hyperfine splitting constant (A_{HFI}). [57] In bitumen, about 65 % of the free radicals on organic sites exist as doublets at room temperature and give rise to a HFS with 2 spectral lines but a low splitting constant (A_{HFI})

in the EPR. But the high spin density gives rise to complex phenomena like anisotropy of the g-values and dipole-dipole interactions between them. This results in considerable overlap giving a single line as indicated in Figure 5.7.

V^{4+} has an unpaired electron in d_{xy}-orbital and interaction with a nuclear spin of I = 7/2 yields the set of 8 hyperfine lines. The parallel and perpendicular components are affected by anisotropic interactions and local viscosity also. If the V-O bond is able to orient itself always perpendicular to the external magnetic field, it means that the VO^{2+} species is in the free state (as in pure vanadyl porphyrins) and not complexed or associated in the medium. Though vanadyl free radical concentration was not quantified in this work, the intensity of the parallel components of the HFS peaks can be an indication of the extent of molecular associations in the sample. This could be related to the viscosity. For example, as reaction time increased, the degree of anisotropy seemed to decrease (compare Figure 5.7 for 15 min with Figure 5.9 for 75 min and 210 min) that could have manifested as a decrease in viscosity observed in the thermally converted products (Figure 5.2). This needs further investigation.

In general, EPR detects the organic and vanadyl free radicals though the *g*-values of the unpaired electrons on paramagnetic O and N species occur quite close to that of C. A clear discrimination is not carried out in this study. It can be assumed that the EPR signal intensity is a direct measure of the concentration or spin density of the free radicals present in the sample. [57] During thermal conversion, free radical generation by bond scission and free radical concentration reduction by termination or by recombination (addition) to produce a heavier product would occur simultaneously at all reaction times. If bond breakage dominates at a particular reaction time, then it can lead to increase in free radical intensity and is reflected in the quantitative values of spin concentration (Figure 5.8 – 75 min, 90 min). Initially bond scission dominates leading to increase in the intensity upto the onset of coking (dip at ~ 45 min), then it further seemed to increase till 90 min after which termination reactions appeared to dominate resulting in the overall decrease in the spin concentration at very high reaction times (1170, 1440 min). Viscosity decrease was monotonous (Figure 5.2) and cannot be directly related to the free radical content. Figure 5.28 illustrates the low correlation between spin concentration and viscosity of the thermally converted products obtained at reaction times between 15 and 360 min.

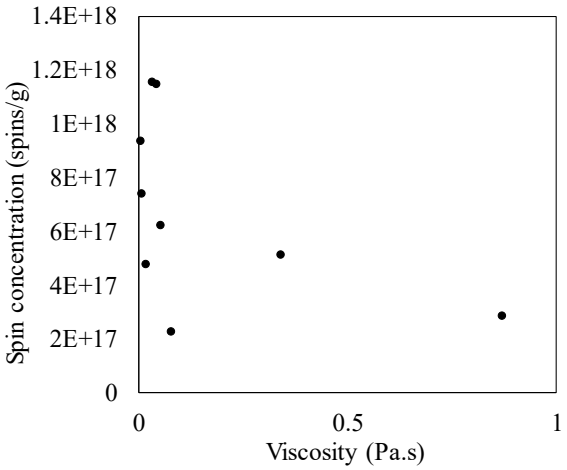


Figure 5.28. Plot of spin concentrations vs. viscosity of the thermally converted samples obtained without the use of any solvent in the reaction time range of 15 – 360 min.

On the other hand, the increase in spin concentration observed on dissolution of thermally converted products in methylene chloride and its evaporation was quite intriguing and it cannot be explained by free radical recombination by addition or dimerization only. It is to be noted that these chemical changes occurred in the temperature range 25 – 55 ºC and not at thermal cracking conditions. According to Niizuma et al., [35] EPR signal intensity depended on the number of paramagnetic species (N), spin quantum number (S) and gyromagnetic ratio (γ) of the electron that arise from the angular momentum and magnetic moment due to its spin. They observed an increase of spin density on passing oxygen gas through the sample without a change in temperature. The speculation was that oxygen interacted with the radical pairs in singlet and the higher energy triplet state radical pairs in the asphaltene forming a complex that decoupled the spins without the formation of any new radical species. Whether a similar explanation can be provided for our case needs to be evaluated.

It was shown that the energy difference between the singlet and the triplet state was small (around 300 cm^{-1}) and an increase in spin concentration can happen if the singlet to triplet transition in the electron spin state can occur. [134] Singlet to doublet conversion can also lead to enhancement in the EPR signal intensity and it was shown to be possible in a complex asphaltene molecule. [135] There is no proof provided in this study that the presence of a foreign molecule such as methylene chloride can decouple the spins the same way as oxygen was postulated to do, [35] but the possibility cannot be ruled out.

Furthermore, on increasing temperature, doublet monoradicals can even convert to triplet state biradicals, that could have led to the increase in EPR intensity during solvent evaporation (Figure 5.9). During evaporation, the temperature was increased from 25 – 55 ºC that could have converted some singlet and doublet radicals to the triplet state but past works have shown that a larger temperature range was required to see a significant change in the intensities and line shapes of EPR spectra of asphaltenes. [38,55]

It can also be argued that though the samples were stored in N_2 atmosphere after thermal reaction, the transfer process during the characterization could have exposed the sample to O_2 from the air. The time of exposure was minimal to cause such a change.

Evidence of increase in spin densities on treatment of bituminous substances with organic and polar solvents have been observed. [58,136] On treatment of char-like material with organic solvents, elevation in the spin concentrations was reported. [4] The explanation was similar to the mechanism of decoupling of the spins by oxygen as discussed previously, where diamagnetic solvent molecules percolated in-between the aromatic sheets and resulted in narrowing of the signal that manifests as intensity increase due to separation of the free radical interaction. Separation of aromatic sheets can also result in charge transfer through the formation of radical ions that could also act as driving forces for asphaltene aggregation through dipole-induced dipole interactions. [4] This could result in a viscosity elevation. Also, when CH_2Cl_2 was used to separate bitumen and its mineral matter, the extracted mineral matter displayed a higher density than the bitumen itself, which was attributed partially to mineral content which gave a signal at the same *g*-value. [58] This was supported by the claim that conservation of spins during solvent extraction was not required to be valid.

However, there was not much effect of solvent extraction on the line shape of either the organic free radical or the paramagnetic vanadyl peaks (Figure 9). Thermal conversion can decrease the extent of anisotropy of the EPR signals from the thermally converted products, that can be viewed as a consequence of the viscosity reduction. As mentioned before, when the VO^{2+} species are liberated from their molecular associations or if they experience a depletion in spins in their vicinity, the parallel components of the HFS peaks reduce in intensity, ultimately transitioning into an isotropic spectrum where the VO^{2+} species are in a free state. This can also be reflected in the peak-to-peak line width of the vanadyl peaks as it is inversely related to the spin-spin relaxation time (T_2). The broader the line-width, the lower the relaxation time, implying

that the spin concentration in the vicinity of the VO^{2+} species is higher, indicating more local viscosity. This type of change may be observed during thermal conversion with increasing reaction time but was not seen in the solvent extracted products, though there was an increase in the intensity with no change in line shape. This suggests that the spin concentration cannot be directly related to the viscosity change observed in the solvent-extracted products.

5.4.3.3 Influence of shear rates and the impact of viscoelasticity on viscosity

The other important parameters that affected viscosity of bitumen and its cracked products were the viscosity measurement conditions of shear rate and temperature. This can also be considered independent to changes occurring during thermal conversion itself. Some interpretations could be derived from the observed results in this study, which are the following:

a. Shear rate-dependent viscosity was exhibited by the thermally converted samples, which eventually led to a Newtonian plateau at high shear rates (Figure 5.24). The transition from shear-thinning to constant viscosity Newtonian-like region occurred at higher shear rates for less viscous samples. In order to maintain consistency in viscosity measurements, the shear rates at which the viscosity is reported should conform to the Newtonian region for a sample. The existence of Newtonian region at low and high shear rates was reported. [34]
b. For less fluid samples (15 min in Figure 5.24), a peculiar phenomena was observed. The viscosity dropped steeply and could be clearly distinguished from shear thinning-type curves, which were more gradual and steady and seemed to follow a power-law-type decrease. It was not necessary that every sample exhibited this behaviour (30 min and 45 min samples did not show this in the high shear Newtonian-like region).
c. The phenomenon of thixotropy (time-dependency of viscosity) was not observed for the thermally converted samples of Athabasca bitumen (Figure 5.26), unlike the case of some thermally converted Cold Lake samples. [20] This was also supported in the work by Behzadfar & Hatzikiriakos [63] where they showed that some bitumen systems showed no thixotropic behavior when the rheological measurements were conducted at 30 ºC.
d. The order of magnitude decrease in low-shear viscosity as shear rate was increased from 1 to 10 s^{-1} (shown for the thermally converted samples – Figure 5.25) could be ascribed mostly to de-aggregation effect of physical nature which require minimal external force

to be disrupted. This also meant that the shear thinning exhibited by the sample was really existent, since the risk of viscous heating was reduced at such low shear deformations.

e. The presence of viscoelastic effects that contribute to an increase in viscosity due to the presence of gel-type structure (explained in the next section) cannot be ruled out due to the temperature of measurement. Behzadfar & Hatzikiriakos [137] also found that viscoelastic effects were prominent in bitumen in the temperature range of 20 – 60 ºC. Since 40 ºC was used as the measurement temperature in this work, the role of shear rates is exemplified in achieving a stable viscosity. But it should be kept in mind that these are still examples of mild shear thinning behavior in comparison with polymers. [138]

It is important to understand the differences in the rheological behavior of bitumen compared to polymers. In one of the model descriptions of bitumen, bitumen can be considered to be represented by means of a colloidal model where the asphaltenes are dispersed in the bulk solvent (maltenes) by resin-type material. [33] However, there was some opposition [139] that claimed that bitumen was a homogeneous fluid with no asphaltene micelles as in the colloidal model. Subsequently, more evidence for the existence of colloidal dispersed particles rather than a homogeneous fluid were shown through a number of characterizations. [33] In the colloidal model for bitumen, asphaltene micelles were shown to possess a fractal dimension of 2, which are very similar to those of polymer coils. [33] A fractal dimension of 2 can be thought of as equivalent to a surface behavior. Though this similarity exists, analogies between bitumen and polymer should be done with caution.

The definition of relaxation time is complex for bitumen and its cracked samples as compared to polymers. Shear force applied to a polymer melt or a diluted polymer causes stretching or elongation of the chains with their simultaneous disentanglement. This can also be viewed as a conformational change where the entropy decreases as they move to a more ordered state. Dealy [34] pointed out that the molecular mechanism involved in viscosity decrease on application of shear was not related to conformational changes and elongation. The first normal stress differences, that arise due to Brownian motion when the liquid is under shear and are an indication of elasticity were seen to be significant at shear rates higher than where shear thinning behavior was observed. The presence of normal stress difference can imply changes in molecular orientation under shear rates similar to polymers. But since Dealy [34] observed that this occurred out of the shear-thinning

range for different types of bitumen, the viscosity change could not be attributed to polymer-like response, i.e. stretching and molecular orientation changes.

The shear rate-dependent viscosity response of colloidal bitumen also depended on the concentration of the dispersed in the continuous phase. Since the concentration of asphaltenes or *n*-pentane insoluble content in feed and thermally cracked bitumen hovered around ~15-20 %wt. (Figure 18), it falls in between a dilute and a concentrated regime of the dispersed phase. Shear flow of a concentrated colloid system of model particles (volume fraction $\phi_c = 0.3 - 0.6$) was studied by Silbert & Melrose [140] where they stated that viscosity of such a system was influenced by the combination of three kinds of forces: hydrodynamic, repulsive or attractive colloidal and Brownian forces. Under shear, the colloidal mixture displayed shear thinning behavior at lower shear rates, eventually transitioning into Newtonian region, which they called as shear ordering. They postulated that colloidal repulsive forces were predominant in the shear thinning region, causing the viscosity to decrease drastically after which hydrodynamic forces come into play at higher shear rates. Aggregated colloidal particles showed a faster viscosity reduction than a non-interacting hard sphere system, which can be seen as one of the reasons for the Newtonian plateau being reached faster for thermally converted products at lower reaction times in which the *n*-pentane insolubles can exist in a more aggregated state than less viscous products at higher reaction times (Figure 5.24).

The viscosity variation with shear rate can be considered in three regimes: (i) Newtonian-like region at shear rates lower than 1 s^{-1}. [34,62] This was not investigated in this study so will not be discussed; (i) Shear thinning region where the interaction potential decreases and the function resembles a power law type relation; [33,140] (ii) Shear ordering region where hydrodynamic interaction of the asphaltenic material with the solvent prevent further aggregation and maintain a constant viscosity. This region can be compared to that of non-interacting hard spheres in a solvent. The transition depends on the volume fraction of the dispersed particles and the onset of elastic nature that arises due to asphaltene particle interactions within themselves as well as with the solvent.

In order to understand viscoelasticity in bitumen, the concept of 'sol' and 'gel' structures is quite relevant. It was proposed in the early 1930s with the asphaltene particles viewed as micelles dispersed in the bulk by the amphoteric resins. The two hypothetical forms are shown in Figure 5.29 as adapted from Lesueur. [33]

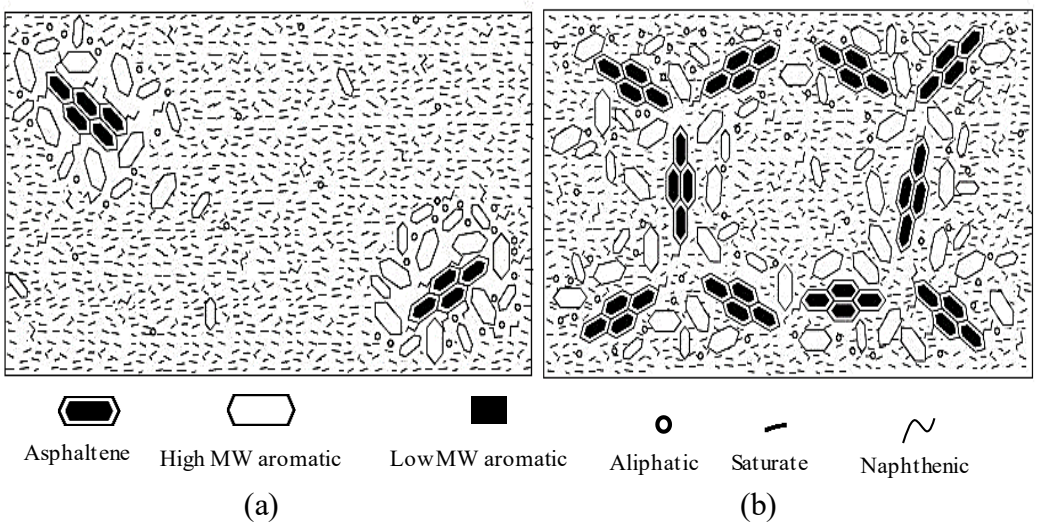


Figure 5.29. (a) Sol and (b) gel representations of bitumen. Reproduced with permission from reference 33. Copyright 2009 Elsevier. [33]

Viscous nature is believed to be because of the 'sol' form while elasticity and Non-Newtonian behavior is exhibited by the 'gel' structure in bitumen. [141] Species that were described as resins act as the transition between the micellar asphaltenes and the non-polar maltenes. The existence of asphaltenes in micelle forms has been confirmed by Small Angle X-Ray Scattering (SAXS) experiments and thermal analysis experiments. [142]

Attractive forces between asphaltenes can be of various types, namely polar forces that can include associations between vanadyl species and nitrogen containing aromatic structures like porphyrins, $\pi - \pi$ aromatic stacking, dispersion forces and other kinds of physical forces like hydrogen bonding. [3] These can co-exist in the gel structure at the same time. The polar regions of the resins can interact with the heteroatoms in asphaltenes while the aliphatic, paraffinic non-polar parts can interact with the bulk solvent, thus forming a link between the dispersed phase and the continuous phase of colloidal bitumen. On application of shear, disruption of the gel structure can result in a structural breakdown and de-aggregation that causes viscosity to decrease. However, the rate of viscosity reduction decreases and attains a stable value at high shear rates. This can be explained by means of two kinds of forces acting at the same time where the de-aggregated asphaltenes are free to associate by means of chemical or physical interactions but the hydrodynamic interaction with the bulk solvent prevents them from agglomerating.

Apart from the temperature dependence of viscoelasticity, some discrepancy regarding the contribution of elastic forces to viscosity change during shearing was found in the literature. [34] At lower shear rates, it was found that steady shear viscosity was equivalent to the real part complex viscosity as determined from the storage modulus (G') vs. frequency (ω) curve indicating minimal elasticity for raw samples of Athabasca and Cold Lake bitumen at these shear rates. This is in contrast to the theory that gel structure breaks down under shear, which can imply that elastic nature is prevalent at lower shear rates. On the other hand, the stress growth curves indicated the presence of elasticity at lower shear rates for different types of bitumen, though not investigated in this study. Elasticity can potentially lead to a viscosity increase. At higher shear rates, the constant viscosity can also be viewed as a result of decrease in the rate of viscosity reduction due to onset of elasticity. Once the aggregates are broken down, their movement can be restricted by the high shear and the ability to store the energy from the external force increases effectively leading to an elastic nature. This can be true for thermally converted samples as well.

In summary, shear thinning can be considered to occur due to breakdown of the gel structure at lower shear rates but due to the emergence of elastic nature of bitumen at higher shear rates, the viscosity change is impacted significantly. However, the description of viscosity reduction in terms of gel breakdown does not appear to be consistent with low shear measurement of the storage modulus.

It was also worthwhile to review the different types of viscometers and rheometers used for rheological measurements of bitumen and heavy crude oils by various researchers in recent times. A summary of some of these works is given below:

- Junaid et al. (2010) [27] used a chamber and spindle viscometer (Brookefield DV-E). Their topic of research was cracking of Athabasca bitumen using Natural Zeolites where they found that higher liquid product yields and viscosity reduction were obtained in catalytic cracking as compared to thermal cracking at 300 ºC, 350 ºC. Viscosity was not shear rate and time-dependent.
- Hasan et al. (2010) [143] made use of a cone and plate viscometer (Haake Rheostress 100) operated in controlled stress (CS), controlled rate (CR) and oscillation (OSC) modes. They worked on viscosity reduction of a Canadian heavy crude oil from Husky Energy for pipeline transport. Shear thinning was observed between shear rates of 50

– 800 s^{-1}. Mixing lighter crude with heavy oil decreased viscosity by 96 %. Yield stress was 0.7 Pa at 25 ºC.

- Mouazen et al. (2011) [144] employed a parallel plate rheometer (25 mm dia., 1 mm gap TA Instruments ARG2) operated in the stress and strain modes separately. Plates of larger diameter were also used to confirm results. They investigated bitumen behavior through rheological and thermal studies. From modulus vs. frequency curves, viscoelasticity was shown to exist with elastic nature at lower frequencies. This was in concurrence with the theory given previously that gel structure exists at lower shear. Consequently, shear thinning was prevalent at shear rates between 1 – 100 s^{-1} measured at 50 ºC. Yield stress was between 0.1 – 1 Pa at 22 ºC.
- Abivin and Taylor (2012) [62] used both parallel plate rheometer (gap 0.8 mm) for measurements at 10 – 50 ºC and a capillary viscometer for measurements at temperatures > 50 ºC as a closed system was needed to maintain uniform composition by preventing evaporation. They studied viscoelasticity and thermal characteristics of heavy oils. Variation of viscosity of heavy oils with temperatures upto 200 ºC was comparable to that of bitumen. Shear thinning was observed between 10 – 100 s^{-1} at 25 – 38 ºC. Importantly, asphaltene content seem to correlate well with viscoelastic behavior and not with viscosity.
- Behzadfar and Hatzikiroakos (2014) [63] utilized three geometries in their Anton Paar MCR501 rheometer, namely, concentric cylinder, parallel plate and vane-cup to study the effect of temperature, pressure and CO_2 saturation on Athabasca bitumen rheology. They suggested that vane cup geometry has better mixing efficiency and minimizes wall slip effects. Onset of shear thinning in bitumen was observed beyond 1 s^{-1}. Not a large dependency of viscosity on pressure was seen in the range 0 – 15 MPa.
- Yanez and De Klerk (2018) [20] used a cup and bob viscometer (Anton Paar RheolabQC), similar to the one used in this work to study the thermal conversion of Cold Lake bitumen at lower temperatures of 150 – 300 ˚C. Mild thixotropic behavior at 10 s^{-1} and at 40 ºC was observed in a limited number of samples. Non-monotonic trend in viscosity with reaction time and poor correlation with asphaltene content were also seen.

It can be seen that a variety of rheometers and viscometers were used for viscosity measurements by previous researchers and some of the conclusions were related to the observations discussed in this chapter in the context of bitumen.

5.4.4 Changes in aggregate content structure during thermal reaction that might affect viscoelastic properties

Wang et al. [18] expressed the opinion that mild thermal conversion did not cause much change in molecular weight of Cold Lake bitumen and attempted to explain the change in viscosity in terms of the effective volume fraction of aggregates. Viscosity depended on the effective volume fraction of the aggregates which included the trapped material and molecules in the boundary layers inside the pockets formed by hydrophobic groups in the aggregates and microstructural pores. It should be noted that not all aggregates are necessarily asphaltenes but can be considered as a representation of *n*-pentane insoluble material. Figure 5.30 shows an illustration of the effective volume of a cluster of asphaltene molecules. The resin-like material surrounds the asphaltene particle leading into the entrapped solvent.

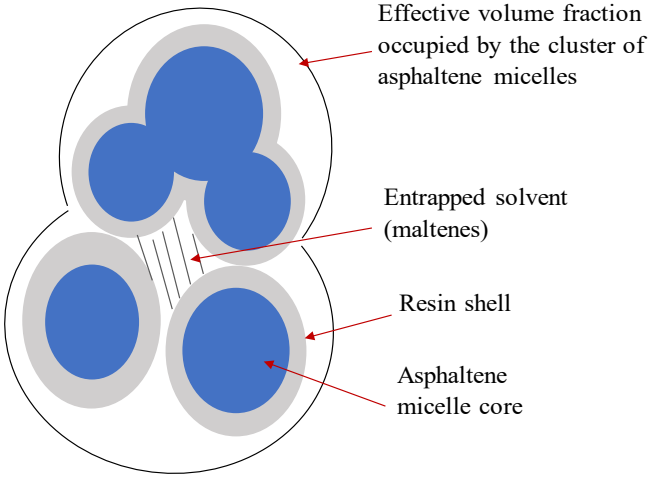


Figure 5.30. Schematic of the effective volume of a cluster of asphaltenes with the entrapped solvent.

It can be seen that the effective volume of the cluster is more than the sum of the volumes of the individual micelles due to entrapped material. It can also be quantified by a parameter called

the solvation constant (K) in relation with the effective volume fraction of the cluster (ϕ_{eff}) in the whole system as in equation 5.1: [33]

$$\phi_{eff} = Kx_a \qquad \text{Equation 5.1}$$

where x_a is the mass fraction of the aggregate content in the mixture. This can be approximated as the asphaltene content with caution. The solvation constant (K) is different from solubility parameter (δ) and one way to quantify it is by measuring the mass fraction of *n*-pentane asphaltenes that dissolve in cyclohexane (y) and applying the following equation: [145]

$$K = \frac{1}{1-y} \qquad \text{Equation 5.2}$$

Lesueur [33] suggested that the value K/ϕ_m (ϕ_m is the minimum volume fraction of the asphaltenes at maximum packing density) can be taken a constant and equal to 5.5 for bitumen with asphaltene contents between 7 – 25 %wt. It was found that K value decreases for an aggregated molecule as compared to the separate monomers. However, in the case of colloidal model assumption, the solvation constant alone would not be sufficient to explain stability and solubility of asphaltenes in the bulk. Other factors like their packing density and spatial orientation with the resin-type material, molecular weight and size of their aggregates would also affect their stability. The solvation constant also influences the relaxation time of the colloidal species, which can be correlated with viscosity.

On thermal cracking, the decrease in effective volume of the cluster (Figure 5.30) potentially contributed to drastic viscosity reduction during initial reaction times in Cold Lake bitumen. [18–20] Thermal conversion may also cause destruction of the solvation layer of resin molecules around the asphaltene core. Separation of the thermally converted products into SARA fractions can give the estimated resin content and though not performed in this work, the variation of resin content with reaction time can be a valuable addition to the future work and can also serve the purpose of testing the assumption of colloidal model for bitumen. With progress of thermal conversion, the observed viscosity increase may not only be due to the suggested addition products formation, [18] but it can also be argued that the asphaltene core that becomes exposed on disruption of the solvation layer can form aggregates again. This could also lead to a lower solvation parameter. Again, the role of solvent cannot be understated in resisting the approach of the dispersed asphaltene molecules. Using a polymer-inspired model to quantify the hydrodynamic resistance of the solvent to aggregation would further assist in understanding the viscosity changes occurring

not only during thermal conversion but also in feed bitumen as well. [146] For example, the Zimm model is used to describe polymer motion in dilute solutions where the whole polymer is considered as a combination of beads connected by springs in a solvent. [147] When this model is extended to bitumen, the beads represent the asphaltene micelles and springs account for the elastic effects.

Shear force may not be strong enough to disrupt chemical interactions between the asphaltene molecules but a temperature of 400 ºC can cause more disruption than just physical forces. This tempts one to think that with increase in thermal cracking reaction time, dispersion and hydrogen bonding forces would have a minimal contribution to aggregation but polar forces and aromatic sheet stacking might play a role. Brownian motion of the asphaltene micelles is required for aggregation but at the same time a lot of encounters might be necessary. Although descriptions in terms of forces and reactions acting upon a cluster as shown in Figure 5.30 is speculative, it provides a useful model framework for exploring the observations.

Certain observations by past researchers emphasized the importance of understanding the asphaltene aggregation process in more detail qualitatively. A change in the nature of asphaltenes after the onset of coking in the thermally converted products formed during 90 – 270 min of thermal reaction time was noted by Zachariah & De Klerk. [19] During this time, the H/C ratio of the liquid phase remained constant, the asphaltenes remained soluble in the liquid and the viscosity passed through a local minimum. A change in the nature of the asphaltenes can imply a change in the nature of interaction between them. Therefore, estimation of the type of interaction forces present in the asphaltenes during thermal conversion can give valuable information on the reasons for the observed changes. Investigating the nature of dominant forces responsible for aggregation on a qualitative basis can provide valuable insights to the impact of aggregation on viscosity change. This can be done by fractionating the asphaltenes of the thermally converted products, performing elemental analysis and obtaining the correlation between heteroatom-carbon content ratios and the solubility parameter of the solvent mixture used to obtain each fraction. Gawrys et al. [148] applied a similar approach to asphaltenes taken from different crude oils but this can be adopted on asphaltenes obtained from thermally cracked samples.

The notion that all asphaltenes are not aggregates finds support from the existential pattern of the ‘gel’ structure. Functions of storage modulus vs. temperature or frequency did not reach a plateau at any point indicating the absence of a completely elastic gel structure. [149] Viscous

character was always present. The transition from sol to gel can be characterized by an instability index (I_c) where values higher than 1.2 indicate a gel structure. [150] The implication to thermal cracking is that since I_c was defined to be dependent on the asphaltene content and amount of paraffin solvent required to precipitate the asphaltenes, it can be calculated for a thermally converted sample to determine the extent of sol or gel character. This can also be an indication of how viscoelastic a particular product is. However, at temperatures below -20 ºC, bitumen becomes a glassy material and can be considered to be elastic. Above 60 ºC, viscous nature predominates and in between this range of temperatures, bitumen tends to be viscoelastic. Viscoelasticity can be linear or non-linear functions between viscosity and shear rate. Power law-type relationships as observed in Figure 5.24 is quite common.

Relaxation times of molecules can also serve as indicators of the elasticity of a sample. For a simpler viscoelastic liquid like a polymer, the intersection point between the storage modulus (G') and the loss modulus function ($G"$) can be easily used to calculate the relaxation time. Relaxation functions were stated to be bimodal in nature for bitumen, [33] where the colloidal asphaltene phase and the maltene phase were shown to possess separate modes of relaxation with some coupling. These were found to be quite sensitive to temperature where α-relaxation of the asphaltene phase was related to the transition from Newtonian region to viscoelastic flow and β-relaxation of the maltene phase corresponded to the shift from viscoelastic to glassy/elastic regime at temperatures < -20 ºC as mentioned before. [151] Relaxation time was shown to be dependent on the maltenes viscosity and the cube of the asphaltene aggregate size combined with the solvation parameter that describes the effective volume of the cluster (equation 5.1).

As a final comment on viscoelasticity, the dispersed asphaltenes in the sol structure may be stabilized by colloidal repulsive forces (as mentioned in the previous section also) that are not necessarily electrostatic but more steric in nature. Viscosity of polymers majorly depend on molecular weight raised to any power between 1 – 3.4 depending on the chemical nature and entanglement, but in the case of bitumen, it is much more complex.

5.5 Conclusions

The modification of viscosity of Canadian oilsands bitumen derived from the Athabasca region was explored by means of thermal conversion at 400 ºC and reaction times in the range 15

– 1440 min. This was compared with the results from thermal conversion of Cold Lake bitumen performed by certain previous studies [18–20] and differences in viscosity trends with reaction time were noted. The aim of this chapter was to account for the changes in viscosity by considering different factors that could potentially contribute in viscosity modification in a complex mixture like bitumen. The products from thermal conversion were characterized by different spectroscopic techniques like infrared and proton nuclear magnetic resonance spectroscopy, electron paramagnetic resonance spectroscopy and other properties like density, refractive index and boiling point distributions were also measured. However, it is required to mention that specific characterization of the solids was not done in this study and could provide further useful insights into the reaction chemistry and product properties. The following were the key conclusions obtained from the results in this chapter:

a. *Difference in viscosity trends for Athabasca and Cold Lake bitumen:*

A monotonous decrease in viscosity was observed during thermal conversion of Athabasca bitumen in this study while a non-monotonic trend was reported for Cold Lake bitumen in certain previous studies. [18–20] This prompted for further investigation and variation in the viscosity of the feed from Athabasca and Cold Lake sources was investigated. It was found that there was variability in feed viscosity across the two deposits and in addition, sample-to-sample variation existed within the same type of bitumen as well. Though small differences in the chemical composition of both type of bitumen, where Cold Lake contained a higher percentage of straight chain paraffins were noted from the work by Selucky et al., [69] no particular reason could be attributed to the difference viscosity of the feeds and its change during thermal conversion.

b. *Effect of post-reaction procedure on viscosity – Nature of extraction solvent:*

When methylene chloride was used to extract the thermally converted products from the reactor, significant differences were observed in product properties before solvent addition and after its evaporation. The viscosity of the solvent-extracted products increased by two orders of magnitude with a slight increase in density but negligible change in refractive index as compared to the solvent-free products. FTIR spectra indicated the possibility of the presence of

intermolecular hydrogen bonding between the remaining methylene chloride in the products (around 1 %wt.) and the bitumen matrix. Evidence of the presence of such interactions involving halogenated solvents was shown by Smith [31] where the acidic hydrogens or lone pairs of chlorine could be associated in hydrogen bonding. Also, the organic free radical intensity of the methylene chloride-extracted products was higher than that of the solvent-free products. Though this could not directly explain the increase in viscosity, complex formation by decoupling of the singlet and triplet spin states of the radicals on carbon centres by CH_2Cl_2 molecules could not be ruled out. No such changes were observed when toluene was used as the extraction solvent.

c. Effect of post-reaction procedure on viscosity – Shear rates and temperature of viscosity measurement:

It was clear from the rheological behavior of the thermally converted samples that bitumen viscosity was dependent on shear rate. In the shear rate range of 1 – 1544 s^{-1}, mild shear thinning at lower shear rates and a Newtonian-like region at higher shear rates were observed for all products and the onset of Newtonian-like regime depended on the fluidity of the sample. A Newtonian-like region was also reported [34] at shear rates lower than 1 s^{-1} but was not investigated in this study. A higher shear rate was required to reach the second Newtonian regime for lower viscous samples and this was plausibly due to their lower asphaltenes content. [140] The temperature of viscosity measurement was 40 ºC, which fell in the viscoelastic range for bitumen. [63] Under the assumption that bitumen is colloidal, it was speculated that the viscosity reduction in the shear thinning region was due to breakdown of a gel-type structure, where the asphaltenes existed in the aggregated state and resulting in a decrease in elasticity. [33] The region of stable viscosity was thought of as a re-occurrence of elastic nature of the sample. Lastly, it was also seen that the thermally converted samples did not exhibit thixotropic behavior.

d. Effect of aggregate structure on viscosity during thermal conversion:

The effective volume of the asphaltene clusters incorporates trapped solvent material and solvation layer of the asphaltenes. Based on the hypothesis that aggregation tendency of the dispersed phase leads to an increase in viscosity, an explanation was proposed to account for the

changes in viscosity during thermal conversion in different studies. It was proposed that de-aggregation could have been responsible for the viscosity decrease during the initial stages of thermal reaction but this also leads to decrease in effective volume of the asphaltene cluster and disruption of the solvation layer of asphaltenes consisting of resin-like material. This could expose the asphaltenes to other similarly exposed asphaltenes and thereby increase their probability of aggregating to form larger aggregates that would lead to an increase in the viscosity. This proposed explanation must still be experimentally evaluated.

Finally, it can be concluded that viscosity of a complex fluid like bitumen depends on a combination of various factors, which need to be dealt with together in order to account for changes in viscosity.

5.6 References

(1) Macchietto, S.; Hewitt, G. F.; Coletti, F.; Crittenden, B. D.; Dugwell, D. R.; Galindo, A.; Jackson, G.; Kandiyoti, R.; Kazarian, S. G.; Luckham, P. F.; Matar, O. K.; Millan-Agorio, M.; Müller, E. A.; Paterson, W.; Pugh, S. J.; Richardson, S. M.; Wilson, D. I. Fouling in Crude Oil Preheat Trains: A Systematic Solution to an Old Problem. *Heat Transf. Eng.* **2011**, *32*, 197–215.

(2) Derakhshesh, M. Asphaltene Aggregation and Fouling Behavior. Ph.D. Thesis, University of Alberta, September 2012.

(3) Gray, M. R. *Upgrading Oilsands Bitumen and Heavy Oil*; University of Alberta Press: Edmonton, AB, 2015.

(4) Strausz, O. P.; Lown, E. M. *The Chemistry of Alberta Oil Sands, Bitumens and Heavy Oils*; Alberta Energy Research Institute: Calgary, 2003.

(5) Mullins, O. C.; Sabbah, H.; Eyssautier, J.; Pomerantz, A. E.; Barré, L.; Andrews, A. B.; Ruiz-Morales, Y.; Mostowfi, F.; McFarlane, R.; Goual, L.; Lepkowicz, R.; Cooper, T.; Orbulescu, J.; Leblanc, R. M.; Edwards, J.; Zare, R. N. Advances in Asphaltene Science and the Yen-Mullins Model. *Energy Fuels* **2012**, *26*, 3986–4003.

(6) Mullins, O. C.; Sheu, E. Y.; Hammami, A.; Marshall, A. G. *Asphaltenes, Heavy Oils, and Petroleomics*; Springer: New York, 2006.

(7) Brons, G.; Yu, J. M. Solvent Deasphalting Effects on Whole Cold Lake Bitumen. *Energy*

Fuels **1995**, *9*, 641–647.

(8) De Klerk, A.; Gray, M. R.; Zerpa, N. Unconventional Oil and Gas. Oilsands. In *Future Energy: Improved, Sustainable and Clean Options for our Planet*; Elsevier: Amsterdam, 2013; pp 95–116.

(9) Hart, A. A Review of Technologies for Transporting Heavy Crude Oil and Bitumen via Pipelines. *J. Pet. Explor. Prod. Technol.* **2014**, *4*, 327–336.

(10) From Diluted Bitumen to Synthetic Crude: Upgrading Explained https://www.oilsandsmagazine.com/technical/bitumen-upgrading (accessed Jan 5, 2019).

(11) Maples, R. E. *Petroleum Refinery Process Economics*; PennWell: Tulsa, OK, 1993.

(12) Rana, M. S.; Sámano, V.; Ancheyta, J.; Diaz, J. A. I. A Review of Recent Advances on Process Technologies for Upgrading of Heavy Oils and Residua. *Fuel* **2007**, *86*, 1216–1231.

(13) Speight, J. G. *The Chemistry and Technology of Petroleum*; Marcel Dekker: New York, 1991.

(14) Biasca, F. E.; Dickenson, R. L.; Chang, E.; Johnson, H. E.; Bailey, R. T.; Simbeck, D. R. Future Technology In Heavy Oil Processing. In *Upgrading heavy crude oils and residue to transportation fuel: Technology, Economics, and Outlook*; SFA Pacific Inc.: Palo Alto, CA, 2009.

(15) Raseev, S. *Thermal and Catalytic Processes in Petroleum Refining*; Marcel Dekker: New York, 2003.

(16) Tokarska, A. Investigations on the Processing Vacuum Residue and Its Mixtures and Coal Tars of Oil with Coal of Crude Materials. *Fuel* **1996**, *75*, 1094–1100.

(17) Mcketta Jr, J. J. *Petroleum Processing Handbook*; Marcel Dekker: New York, 1992.

(18) Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; De Klerk, A. Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C. *Energy Fuels* **2014**, *28*, 5014–5022.

(19) Zachariah, A.; De Klerk, A. Thermal Conversion Regimes for Oilsands Bitumen. *Energy Fuels* **2016**, *30*, 239–248.

(20) Yañez Jaramillo, L. M.; De Klerk, A. Partial Upgrading of Bitumen by Thermal Conversion at 150-300 °C. *Energy Fuels* **2018**, *32*, 3299–3311.

(21) Shu, W. R.; Venkatesan, V. N. Kinetics of Thermal Visbreaking of a Cold Lake Bitumen. *J. Can. Pet. Technol.* **1984**, *23*, 60–64.

(22) Henderson, J. H.; Weber, L. Physical Upgrading of Heavy Crude Oils by the Application

of Heat. *J. Can. Pet. Technol.* **1965**, *4*, 206–212.

(23) Castillo, J.; De Klerk, A. Visbreaking of Deasphalted Oil from Bitumen at 280–400 °C. *Energy Fuels* **2018**, *33*, 159–175.

(24) Dutta, R. P.; McCaffrey, W. C.; Gray, M. R.; Muehlenbachs, K. Thermal Cracking of Athabasca Bitumen: Influence of Steam on Reaction Chemistry. *Energy Fuels* **2000**, *14*, 671–676.

(25) Prado, G. H. C.; Riya; Hyrve, M.; De Klerk, A. Role of Metal Halides in Coke Formation during Bitumen Upgrading. *Fuel* **2018**, *211*, 775–782.

(26) Vilcáez, J.; Watanabe, M.; Watanabe, N.; Kishita, A.; Adschiri, T. Hydrothermal Extractive Upgrading of Bitumen without Coke Formation. *Fuel* **2012**, *102*, 379–385.

(27) Junaid, A. S. M.; Wang, W.; Street, C.; Rahman, M.; McCaffrey, W.; Kuznicki, S. Upgrading of Athabasca Oilsands Bitumen by Low Temperature Natural Zeolite Cracking. *10AIChE - 2010 AIChE Annu. Meet. Conf. Proc.* **2010**, *4*.

(28) Joshi, J. B.; Pandit, A. B.; Kataria, K. L.; Kulkarni, R. P.; Sawarkar, A. N.; Tandon, D.; Ram, Y.; Kumar, M. M. Petroleum Residue Upgradation via Visbreaking: A Review. *Ind. Eng. Chem. Res.* **2008**, *47*, 8960–8988.

(29) Washimi, K.; Limmer, H. New Unit to Thermal Crack Resid. *Hydrocarb. Process.* **1989**, *68*, 69.

(30) Pauling, L. *The Chemical Bond: A Brief Introduction to Modern Structural Chemistry*; Cornell University Press: Ithaca, NY, 1967.

(31) Smith, J. W. Hydrogen-Bonding and Complex-Forming Properties. In *The Carbon-Halogen Bond (1973)*; John Wiley & Sons, Ltd.: Chichester, UK, 1973; pp 265–300.

(32) Prado, G. H. C.; De Klerk, A. Halogenation of Oilsands Bitumen, Maltenes, and Asphaltenes. *Energy Fuels* **2014**, *28*, 4458–4468.

(33) Lesueur, D. The Colloidal Structure of Bitumen: Consequences on the Rheology and on the Mechanisms of Bitumen Modification. *Adv. Colloid Interface Sci.* **2009**, *145*, 42–82.

(34) Dealy, J. M. Rheological Properties of Oil Sand Bitumens. *Can. J. Chem. Eng.* **1979**, *57*, 677–683.

(35) Niizuma, S.; Steele, C. T.; Gunning, H. E.; Strausz, O. P. Electron Spin Resonance Study of Free Radicals in Athabasca Asphaltene. *Fuel* **1977**, *56*, 249–256.

(36) Yen, T. F.; Erdman, J. G.; Saraceno, A. J. Investigation of the Nature of Free Radicals in

Petroleum Asphaltenes and Related Substances by Electron Spin Resonance. *Anal. Chem.* **1962**, *34*, 694–700.

(37) Blanchard, C. M.; Gray, M. R. Free Radical Chain Reactions of Bitumen Residue. *ACS Div. Fuel Chem. Prepr.* **1997**, *42*, 137–141.

(38) Trukhan, S. N.; Yudanov, V. F.; Gabrienko, A. A.; Subramani, V.; Kazarian, S. G.; Martyanov, O. N. In Situ Electron Spin Resonance Study of Molecular Dynamics of Asphaltenes at Elevated Temperature and Pressure. *Energy Fuels* **2014**, *28*, 6315–6321.

(39) Gabrienko, A. A.; Lai, C. H.; Kazarian, S. G. In Situ Chemical Imaging of Asphaltene Precipitation from Crude Oil Induced by *n*-Heptane. *Energy Fuels* **2014**, *28*, 964–971.

(40) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds*, 8th ed.; John Wiley & Sons, 2014.

(41) Zachariah, A.; Wang, L.; Yang, S.; Prasad, V.; De Klerk, A. Suppression of Coke Formation during Bitumen Pyrolysis. *Energy Fuels* **2013**, *27*, 3061–3070.

(42) Babu, D. R.; Cormack, D. E. Effect of Low-Temperature Oxidation on the Composition of Athabasca Bitumen. *Fuel* **1984**, *63*, 858–861.

(43) Petersen, J. C.; Harnsberger, P. M.; Robertson, R. E. Factors Affecting the Kinetics and Mechanisms of Asphalt Oxidation and the Relative Effects of Oxidation Products on Age Hardening. *ACS Div. Fuel Chem. Prepr.* **1996**, *41*, 1232–1241.

(44) Ancheyta, J.; Centeno, G.; Trejo, F.; Marroquín, G.; García, J. A.; Tenorio, E.; Torres, A. Extraction and Characterization of Asphaltenes from Different Crude Oils and Solvents. *Energy Fuels* **2002**, *16*, 1121–1127.

(45) Strausz, O. P.; Peng, P.; Murgich, J. About the Colloidal Nature of Asphaltenes and the MW of Covalent Monomeric Units. *Energy Fuels* **2002**, *16*, 809–822.

(46) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs. *Energy Fuels* **2017**, *31*, 13509–13518.

(47) Speight, J. G. *Fouling in Refineries*; Elsevier: Amsterdam, 2015.

(48) Turuga, A. S. S. Effect of Solvent Deasphalting Process on the Properties of Deasphalted Oil and Asphaltenes from Bitumen. MSc. Thesis, University of Alberta, 2017.

(49) Wang, J.; Buckley, J. *Standard Procedure for Separating Asphaltenes from Crude Oils*; New Mexico, 2002.

(50) Luo, P.; Wang, X.; Gu, Y. Characterization of Asphaltenes Precipitated with Three Light Alkanes under Different Experimental Conditions. *Fluid Phase Equilib.* **2010**, *291*, 103–110.

(51) ASTM D189-06 (Reapproved 2014). *Standard Test Method for Conradson Carbon Residue of Petroleum Products*; ASTM: West Conshohocken, PA, 2014.

(52) Noel, F. Alternative to the Conradson Carbon Residue Test. *Fuel* **1984**, *63*, 931–934.

(53) ASTM D7169-16. Standard Test Method for Boiling Point Distribution of Samples with Residues Such as Crude Oils and Atmospheric and Vacuum Residues by High Temperature Gas Chromatography. ASTM: West Conshohocken, PA 2016.

(54) Bielski, B. H. J.; Gebicki, J. M. *Atlas of Electron Spin Resonance Spectra*; Academic Press: New York, 1967.

(55) Cui, Q.; Nakabayashi, K.; Ma, X.; Miyawaki, J.; Ideta, K.; Tennichi, Y.; Ueda, M.; Al-Mutairi, A.; Marafi, A. M. J.; Park, J. II; Yoon, S. H.; Mochida, I. Studying Rotational Mobility of V=O Complexes in Atmospheric Residues and Their Resins and Asphaltenes by Electron Spin Resonance. *Energy Fuels* **2017**, *31*, 4748–4757.

(56) Malhotra, V. M.; Buckmaster, H. A. 9 and 34 GHz EPR Study of the Free Radicals in Various Asphaltenes: Statistical Correlation of the g-Values with Heteroatom Content. *Org. Geochem.* **1985**, *8*, 235–239.

(57) Eaton, G. R.; Eaton, S. S.; Barr, D. P. *Quantitative EPR*; Springer, 2010.

(58) Khulbe, K. C.; Mann, R. S.; Lamarche, G.; Lamarche, A. M. Electron Spin Resonance Study of the Thermal Decomposition of Solvent Extracted Athabasca Tar Sand Bitumen. *Fuel Process. Technol.* **1992**, *31*, 91–103.

(59) Beauchamp, J. Spectroscopy Tables https://www.cpp.edu/~psbeauchamp/pdf/spec_ir_nmr_spectra_tables.pdf (accessed Oct 12, 2018).

(60) Saoula, S.; Soudani, K.; Haddadi, S.; Munoz, M. E.; Santamaria, A. Analysis of the Rheological Behavior of Aging Bitumen and Predicting the Risk of Permanent Deformation of Asphalt. *Mater. Sci. Appl.* **2013**, *4*, 312–318.

(61) Rek, V.; Barjaktarovi, Z. M. Dynamic Mechanical Behavior of Polymer Modified Bitumen. *Mater. Res. Innov.* **2002**, *6*, 39–43.

(62) Abivin, P.; Taylor, S. D.; Freed, D. Thermal Behavior and Viscoelasticity of Heavy Oils.

Energy Fuels **2012**, *26*, 3448–3461.

(63) Behzadfar, E.; Hatzikiriakos, S. G. Rheology of Bitumen: Effects of Temperature, Pressure, CO_2 Concentration and Shear Rate. *Fuel* **2014**, *116*, 578–587.

(64) McNab, J. G.; Smith, P. V. J.; Betts, R. L. Evolution of Petroleum. *Ind. Eng. Chem.* **1952**, *44*, 2556–2563.

(65) Schramm, L. L.; Kwak, J. C. T. Rheological Properties of an Athabasca Bitumen and Some Bituminous Mixtures and Dispersions. *J. Can. Pet. Technol.* **1988**, *27*, 26–35.

(66) Wallace, D.; Henry, D. A Correlation for Correcting the Viscosity of Solvent-Extracted Bitumen to a Solvent-Free Basis. *AOSTRA J. Res.* **1987**, *3*, 239–248.

(67) Masliyah, J. H.; Xu, Z.; Czarnecki, J. A. *Handbook on Theory and Practice of Bitumen Recovery from Athabasca Oil Sands*; Kingsley Knowledge Publishing, 2011.

(68) Selucky, M. L.; Chu, Y.; Ruo, T.; Strausz, O. P. Chemical Composition of Athabasca Bitumen. *Fuel* **1977**, *56*, 369–381.

(69) Selucky, M. L.; Chu, Y.; Ruo, T. C. S.; Strausz, O. P. Chemical Composition of Cold Lake Bitumen. *Fuel* **1978**, *57*, 9–16.

(70) McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. Heavy Petroleum Composition. 4. Asphaltene Compositional Space. *Energy Fuels* **2013**, *27*, 1257–1267.

(71) Poveda, J. C.; Molina, D.; Martínez, H.; Florez, O.; Campillo, B. Molecular Changes in Asphaltenes within H_2 Plasma. *Energy Fuels* **2014**, *28*, 735–744.

(72) Pomerantz, A. E.; Hammond, M. R.; Morrow, A. L.; Mullins, O. C.; Zare, R. N. Two-Step Laser Mass Spectrometry of Asphaltenes. *J. Am. Chem. Soc.* **2008**, *130*, 7216–7217.

(73) Friend, J. N.; Hargreaves, W. D. Viscosity and the Hydrogen Bond. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **1945**, *36*, 731–756.

(74) Smith, D. F.; Schaub, T. M.; Kim, S.; Rodgers, R. P.; Rahimi, P.; Teclemariam, A.; Marshall, A. G. Characterization of Acidic Species in Athabasca Bitumen and Bitumen Heavy Vacuum Gas Oil by Negative-Ion ESI FT-ICR MS with and without Acid-Ion Exchange Resin Prefractionation. *Energy Fuels* **2008**, *22*, 2372–2378.

(75) Erdman, J. G.; Dickie, J. P. Mild Thermal Alteration of Asphaltic 702. *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **1964**, *9*, 69–79.

(76) Ball, M. W. Development of the Athabaska Oil Sands. *Can. Inst. Min. Metall.* **1941**, *44*, 58–91.

(77) Steacie, E. W. R. *Atomic and Free Radical Reactions*, 2nd ed.; Reinhold: New York, 1954.

(78) Jha, K. N.; Montgomery, D. S.; Strausz, O. P. Chemical Composition of Gases in Alberta Bitumens and in Low-Temperature Thermolysis of Oil Sand Asphaltenes and Maltenes. In *Oil Sand and Oil Shale Chemistry*; Strausz, O. P., Lown, E. M., Eds.; Verlag Chemie: New York, 1978; pp 33–54.

(79) Dunn, G. E.; Thimm, H. F.; Mohanty, R. K. Kinetics and Mechanism of Decarboxylation of Some Pyridinecarboxylic Acids in Aqueous Solution. III. 3-Hydroxy- and 3-Aminopyridine-2-Carboxylic Acids. *Can. J. Chem.* **1979**, *57*, 1098–1104.

(80) Montoya Sánchez, N.; De Klerk, A. Oxidative Ring-Opening of Aromatics: Decomposition of Biphenyl Carboxylic Acids and Zinc Biphenyl Carboxylates. *Energy Fuels* **2015**, *29*, 7910–7922.

(81) Sachanan, A. N. *Conversion of Petroleum. Production of Motor Fuels by Thermal and Catalytic Processes*, 2nd ed.; Reinhold: New York, 1948.

(82) Govindhakannan, J.; Khulbe, C. Effect of Pressure on Delayed Coking of Athabasca Bitumen. *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **2010**, *55*, 44–45.

(83) Carbognani, L.; Lubkowitz, J.; Gonzalez, M. F.; Pereira-Almao, P. High Temperature Simulated Distillation of Athabasca Vacuum Residue Fractions. Bimodal Distributions and Evidence for Secondary "on-Column" Cracking of Heavy Hydrocarbons. *Energy Fuels* **2007**, *21*, 2831–2839.

(84) Tefera, D. T.; Agrawal, A.; Yañez Jaramillo, L. M.; De Klerk, A.; Prasad, V. Self-Modeling Multivariate Curve Resolution Model for Online Monitoring of Bitumen Conversion Using Infrared Spectroscopy. *Ind. Eng. Chem. Res.* **2017**, *56*, 10756–10769.

(85) Lee, J. M.; Baker, J. J.; Rolle, J. G.; Llerena, R. Comparison of Fuel Properties of Petroleum Cokes and Coals Used in Power Generation. *Abstr. Pap. Am. Chem. Soc.* **1999**, 80–89.

(86) Mohapatra, D. P.; Kirpalani, D. M. Bitumen Heavy Oil Upgrading by Cavitation Processing: Effect on Asphaltene Separation, Rheology, and Metal Content. *Appl. Petrochemical Res.* **2016**, *6*, 107–115.

(87) Phillips, C. R.; Haidar, N. I.; Poon, Y. C. Kinetic Models for the Thermal Cracking of Athabasca Bitumen. *Fuel* **1985**, *64*, 678–691.

(88) Millour, J. P.; Moore, R. G.; Bennion, D. W.; Ursenbach, M. G.; Gie, D. N. A Simple Implicit Model for Thermal Cracking of Crude Oils. In *SPE Annual Technical Conference*

and Exhibition; Las Vegas, 1985; pp 1–15.

(89) Brauch, R.; Fainberg, V.; Kalchouck, H.; Hetsroni, G. Correlations between Properties of Various Feedstocks and Products of Visbreaking. *Fuel Sci. Technol. Int.* **1996**, *14*, 753–765.

(90) Hasan, M. A.; Shaw, J. M. Rheology of Reconstituted Crude Oils: Artifacts and Asphaltenes. *Energy Fuels* **2010**, *24*, 6417–6427.

(91) Mehrotra, A. K.; Eastick, R. R.; Svrcek, W. Y. Viscosity of Cold Lake Bitumen and Its Fractions. *Can. J. Chem. Eng.* **1989**, *67*, 1004–1009.

(92) Kapadia, P. R.; Kallos, M. S.; Gates, I. D. A New Kinetic Model for Pyrolysis of Athabasca Bitumen. *Can. J. Chem. Eng.* **2013**, *91*, 889–901.

(93) Poutsma, M. L. Free-Radical Thermolysis and Hydrogenolysis of Model Hydrocarbons Relevant to Processing of Coal. *Energy Fuels* **1990**, *4*, 113–131.

(94) Hou, Y.; Wang, L.; Wang, D.; Guo, M.; Liu, P.; Yu, J. Characterization of Bitumen Micro-Mechanical Behaviors Using AFM, Phase Dynamics Theory and MD Simulation. *Materials (Basel).* **2017**, *10*, 1–16.

(95) Fischer, H.; Stadler, H.; Erine, N. Quantitative Temperature-Depending Mapping of Mechanical Properties of Bitumen at the Nanoscale Using the AFM Operated with PeakForce Tapping TM Mode. *J. Microsc.* **2013**, *250*, 210–217.

(96) Yu, X.; Zaumanis, M.; Dos Santos, S.; Poulikakos, L. D. Rheological, Microscopic, and Chemical Characterization of the Rejuvenating Effect on Asphalt Binders. *Fuel* **2014**, *135*, 162–171.

(97) Guo, M.; Huang, Y.; Wang, L.; Yu, J.; Hou, Y. Using Atomic Force Microscopy and Molecular Dynamics Simulation to Investigate the Asphalt Micro Properties. *Int. J. Pavement Res. Technol.* **2018**, *11*, 321–326.

(98) Schmets, A.; Kringos, N.; Pauli, T.; Redelius, P.; Scarpas, T. On the Existence of Wax-Induced Phase Separation in Bitumen. *Int. J. Pavement Eng.* **2010**, *11*, 555–563.

(99) Fischer, H. R.; Cernescu, A. Relation of Chemical Composition to Asphalt Microstructure - Details and Properties of Micro-Structures in Bitumen as Seen by Thermal and Friction Force Microscopy and by Scanning near-Filed Optical Microscopy. *Fuel* **2015**, *153*, 628–633.

(100) Grover Allen, R.; Little, D. N.; Bhasin, A.; Glover, C. J. The Effects of Chemical Composition on Asphalt Microstructure and Their Association to Pavement Performance.

International Journal of Pavement Engineering. Taylor & Francis 2014, pp 9–22.

(101) Schuler, B.; Meyer, G.; Peña, D.; Mullins, O. C.; Gross, L. Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy. *J. Am. Chem. Soc.* **2015**, *137*, 9870–9876.

(102) Mullins, O. C. The Modified Yen Model. *Energy Fuels* **2010**, *24*, 2179–2207.

(103) Zhang, L.; Xie, L.; Shi, C.; Huang, J.; Liu, Q.; Zeng, H. Mechanistic Understanding of Asphaltene Surface Interactions in Aqueous Media. *Energy Fuels* **2017**, *31*, 3348–3357.

(104) Uzcátegui, G.; Fong, S. Y.; De Klerk, A. Cracked Naphtha Reactivity: Effect of Free Radical Reactions. *Energy Fuels* **2018**, *32*, 5812–5823.

(105) Sabbah, H.; Morrow, A. L.; Pomerantz, A. E.; Zare, R. N. Evidence for Island Structures as the Dominant Architecture of Asphaltenes. *Energy Fuels* **2011**, *25*, 1597–1604.

(106) Wang, W.; Taylor, C.; Hu, H.; Humphries, K. L.; Jaini, A.; Kitimet, M.; Scott, T.; Stewart, Z.; Ulep, K. J.; Houck, S.; Luxon, A.; Zhang, B.; Miller, B.; Parish, C. A.; Pomerantz, A. E.; Mullins, O. C.; Zare, R. N. Nanoaggregates of Diverse Asphaltenes by Mass Spectrometry and Molecular Dynamics. *Energy Fuels* **2017**, *31*, 9140–9151.

(107) Schuler, B.; Fatayer, S.; Meyer, G.; Rogel, E.; Moir, M.; Zhang, Y.; Harper, M. R.; Pomerantz, A. E.; Bake, K. D.; Witt, M.; Peña, D.; Kushnerick, J. D.; Mullins, O. C.; Ovalles, C.; Van Den Berg, F. G. A.; Gross, L. Heavy Oil Based Mixtures of Different Origins and Treatments Studied by Atomic Force Microscopy. *Energy Fuels* **2017**, *31*, 6856–6861.

(108) Zhang, H. T.; Li, R.; Yang, Z.; Yin, C. X.; Gray, M. R.; Bohne, C. Evaluating Steady-State and Time-Resolved Fluorescence as a Tool to Study the Behavior of Asphaltene in Toluene. *Photochem. Photobiol. Sci.* **2014**, *13*, 917–928.

(109) Tanaka, R.; Sato, E.; Hunt, J. E.; Winans, R. E.; Sato, S.; Takanohashi, T. Characterization of Asphaltene Aggregates Using X-Ray Diffraction and Small-Angle X-Ray Scattering. *Energy Fuels* **2004**, *18*, 1118–1125.

(110) Hosseini-Dastgerdi, Z.; Tabatabaei-Nejad, S. A. R.; Khodapanah, E.; Sahraei, E. A Comprehensive Study on Mechanism of Formation and Techniques to Diagnose Asphaltene Structure; Molecular and Aggregates: A Review. *Asia-Pacific J. Chem. Eng.* **2015**, *10*, 1–14.

(111) Gawrys, K. L.; Blankenship, G. A.; Kilpatrick, P. K. Solvent Entrainment in and

Flocculation of Asphaltenic Aggregates Probed by Small-Angle Neutron Scattering. *Langmuir* **2006**, *22*, 4487–4497.

(112) Rueda-Velásquez, R. I.; Freund, H.; Qian, K.; Olmstead, W. N.; Gray, M. R. Characterization of Asphaltene Building Blocks by Cracking under Favorable Hydrogenation Conditions. *Energy Fuels* **2013**, *27*, 1817–1829.

(113) Karimi, A.; Qian, K.; Olmstead, W. N.; Freund, H.; Yung, C.; Gray, M. R. Quantitative Evidence for Bridged Structures in Asphaltenes by Thin Film Pyrolysis. *Energy Fuels* **2011**, *25*, 3581–3589.

(114) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 2: Selective Separation Method That Reveals Fractions Enriched in Island and Archipelago Structural Motifs by Mass Spectrometry. *Energy Fuels* **2018**, *32*, 314–328.

(115) Chacón-Patino, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 3. Dominance of Island or Archipelago Structural Motif Is Sample Dependent. *Energy Fuels* **2018**, *32*, 9106–9120.

(116) Podgorski, D. C.; Corilo, Y. E.; Nyadong, L.; Lobodin, V. V.; Bythell, B. J.; Robbins, W. K.; McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. Heavy Petroleum Composition. 5. Compositional and Structural Continuum of Petroleum Revealed. *Energy Fuels* **2013**, *27*, 1268–1276.

(117) Zhao, H. Y.; Cao, Y.; Sit, S. P.; Lineberry, Q.; Pan, W. P. Thermal Characteristics of Bitumen Pyrolysis. *J. Therm. Anal. Calorim.* **2012**, *107*, 541–547.

(118) Taş, S.; Yürüm, Y. Co-Firing of Biomass with Coals: Part 2. Thermogravimetric Kinetic Analysis of Co-Combustion of Fir (Abies Bornmulleriana) Wood with Beypazari Lignite. *J. Therm. Anal. Calorim.* **2011**, *107*, 293–298.

(119) Shao, J.; Yan, R.; Chen, H.; Wang, B.; Lee, D. H.; Liang, D. T. Pyrolysis Characteristics and Kinetics of Sewage Sludge by Thermogravimetry Fourier Transform Infrared Analysis. *Energy Fuels* **2008**, *22*, 38–45.

(120) Boucher, J. L.; Wang, I.-H.; Martinez, D. F. Elimination Chemistry in Asphalt. *Prepr. Pap. − Am. Chem. Soc., Div. Pet. Chem.* **1990**, *35*, 550–555.

(121) Petersen, J. C. An Infrared Study of Hydrogen Bonding in Asphalt. *Fuel* **1967**, *46*, 295–305.

(122) Allen, F. H.; Wood, P. A.; Galek, P. T. A. Role of Chloroform and Dichloromethane Solvent Molecules in Crystal Packing: An Interaction Propensity Study. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2013**, *69*, 379–388.

(123) Ding, X.; Tuikka, M.; Haukk, M. Halogen Bonding in Crystal Engineering. In *Recent Advances in Crystallography*; InTech, 2012.

(124) Calkins, W. H. The Chemistry of Sulfur in Coal - A Historical Perspective. *Prepr. Fuel Chem. Div., Am. Chem. Soc.* **1993**, *32* (2), 358–368.

(125) Brammer, L.; Bruton, E. A.; Sherwood, P. Understanding the Behavior of Halogens as Hydrogen Bond Acceptors. *Cryst. Growth Des.* **2001**, *1*, 277–290.

(126) Blyler, L. L.; Haas, T. W. The Influence of Intermolecular Hydrogen Bonding on the Flow Behavior of Polymer Melts. *J. Appl. Polym. Sci.* **1969**, *13*, 2721–2733.

(127) Bellamy, L. J.; Hallam, H. E.; William, R. J. Infra-Red Spectra and Solvent Effects. Part 1.—X—H Stretching Frequencies. *Trans. Faraday Soc.* **1958**, *54*, 1120–1127.

(128) Koyo, A. I. D. A. Raman Spectra of Liquid Ammonia Solutions. VI : Liquid Ammonia and a Few Alkyl Chlorides. *Sci. reports Res. Institutes, Tohoku Univ. Ser. A, Physics, Chem. Metall.* **1957**, *9*, 458–466.

(129) Earp, D. P.; Glasstone, S. Dielectric Polarisation and Molecular-Compound Formation in Solution. Parts I and II. *J. Chem. Soc.* **1935**, 1709–1723.

(130) Vogtle, F.; Weber, E. Crown Ethers-Complexes and Selectivity. In *Crown Ethers and Analogs*; Weber, E., Toner, J. L., Goldberg, I., Vgtle, F., Laidler, D. A., Stoddart, J. F., Bartsch, R. A., Liotta, C. L., Eds.; John Wiley & Sons, Inc.: Chichester, UK, 1989; pp 207–304.

(131) McBride, M. T.; Luo, T. J. M.; Tayhas R Palmore, G. Hydrogen-Bonding Interactions in Crystalline Solids of Cyclic Thioureas. *Cryst. Growth Des.* **2001**, *1*, 39–46.

(132) Davey, R. J.; Blagden, N.; Righini, S.; Alison, H.; Quayle, M. J.; Fuller, S. Crystal Polymorphism as a Probe for Molecular Self-Assembly during Nucleation from Solutions: The Case of 2,6-Dihydroxybenzoic Acid. *Cryst. Growth Des.* **2001**, *1*, 59–65.

(133) Thallapally, P. K.; Nangia, A. A Cambridge Structural Database Analysis of the C–H···Cl Interaction: C–H···Cl− and C–H···Cl–M Often Behave as Hydrogen Bonds but C–H···Cl–C Is Generally a van Der Waals Interaction. *CrystEngComm* **2001**, *3*, 114–119.

(134) Murai, H.; Torres, M.; Strausz, O. P. Quintet State Triplet-Triplet Radical Pairs. 2. *J. Am.*

Chem. Soc. **1980**, *102*, 7390–7391.

(135) Soos, Z. G.; Hughes, R. C. Paramagnetic Susceptibilities and Temperature-Dependent Excitation Energies in Linear Organic Crystals. *J. Chem. Phys.* **1967**, *46*, 253–259.

(136) Rudnick, L. R.; Audeh, C. A. Thermal Chemistry of a Utah Tar Sand: An Electron Spin Resonance (ESR) Characterization. *Ind. Eng. Chem. Res.* **1988**, *27*, 1366–1369.

(137) Behzadfar, E.; Hatzikiriakos, S. G. Viscoelastic Properties and Constitutive Modelling of Bitumen. *Fuel* **2013**, *108*, 391–399.

(138) Sunthar, P. Polymer Rheology. In *Rheology of Complex Fluids*; Springer: New York, 2010; pp 171–191.

(139) Little, D. N.; Prapnnachari, S.; Letton, A.; Kim, Y. R. *Investigation of the Microstructural Mechanism of Relaxation and Fracture Healing in Asphalt*; College Station, Texas, 1993.

(140) Silbert, L. E.; Melrose, J. R. The Rheology and Microstructure of Concentrated, Aggregated Colloids. *J. Rheol. (N. Y. N. Y).* **1999**, *43*, 673–700.

(141) Saal, R. N. J.; Labout, J. W. A. Rheological Properties of Asphaltic Bitumen. *J. Phys. Chem.* **1940**, *44*, 149–165.

(142) Dwiggins, C. W. A Small Angle X-Ray Scattering Study of the Colloidal Nature of Petroleum. *J. Phys. Chem.* **1965**, *69*, 3500–3506.

(143) Hasan, S. W.; Ghannam, M. T.; Esmail, N. Heavy Crude Oil Viscosity Reduction and Rheology for Pipeline Transportation. *Fuel* **2010**, *89*, 1095–1100.

(144) Mouazen, M.; Poulesquen, A.; Vergnes, B. Correlation between Thermal and Rheological Studies to Characterize the Behavior of Bitumen. *Rheol. Acta* **2011**, *50*, 169–178.

(145) Schabron, J. F.; Rovani, J. F. J. Method for Determining Asphaltene Stability of a Hydrocarbon-Containing Material. US 8,367,425 B1, 2013.

(146) Yamakawa, H. *Modern Theory of Polymer Solutions*; Harper & Row: New York, 1971.

(147) Iwao, T. *Polymer Solutions: An Introduction to Physical Properties*; Wiley: New York, 2002.

(148) Gawrys, K. L.; Blankenship, G. A.; Kilpatrick, P. K. On the Distribution of Chemical Properties and Aggregation of Solubility Fractions in Asphaltenes. *Energy Fuels* **2006**, *20*, 705–714.

(149) Anderson, D. A.; Christensen, D. W.; Bahia, H. U.; Dongre, R.; Sharma, M. G.; Antle, C. E. *Binder Characterization and Evaluation Volume 3: Physical Characterization*;

Washington, D. C., 1994.

(150) Spiecker, P. M.; Gawrys, K. L.; Kilpatrick, P. K. Aggregation and Solubility Behavior of Asphaltenes and Their Subfractions. *J. Colloid Interface Sci.* **2003**, *267*, 178–193.

(151) Lesueur, D. Letter to the Editor: On the Thermorheological Complexity and Relaxation Modes of Asphalt Cements. *J. Rheol. (N. Y. N. Y)*. **1999**, *43*, 1701–1704.

(152) Koschmieder, E. L. *Bénard Cells and Taylor Vortices*; Cambridge University Press: Cambridge, UK, 1993.

(153) Bird, R. B.; Stewart, W. E.; Lightfoot, E. N. *Transport Phenomena*; John Wiley & Sons, 2001.

6. A data-driven approach to generate pseudo-reaction networks for thermal conversion of Athabasca bitumen [5]

ABSTRACT

This work focuses on the application of self-modeling multivariate curve resolution (SMCR) methods and Bayesian clustering and learning methods on Fourier transform infrared (FTIR) spectra of the liquid products obtained from thermal cracking of Athabasca bitumen in the temperature range of 300 – 420 ºC and reaction times ranging from 15 min to 27 h. The objective was to develop a reaction pathway for the thermal cracking process from the Bayesian methods and identifying key elements of the reaction chemistry that also affected physical properties like viscosity at each temperature that was important from the perspective of partial upgrading of the bitumen. An important aspect of this work was that minimum external chemical knowledge was used for the chemometric techniques. Similar studies were conducted previously for Cold Lake bitumen in the temperature range of 150 – 400 ºC, but the changes in physical properties of the cracked product could not be fully accounted for from the reaction pathway. Therefore, consistency between the results of the two chemometric methods and the accompanying chemical changes was also inspected in this work. The SMCR method employed in our study was applied on both temperature-specific and augmented datasets considering all temperatures to extract resolved concentration and spectral profiles using alternative least-squares (ALS) optimization. The improvements of particle swarm optimization (PSO) over ALS were investigated with regards to resolution quality, convergence speed, residuals and explained variance. The Bayesian methods were used to obtain causal relationships between the identified clusters representing different chemical compound classes and multiple reaction pathways were suggested. The practicality of the Bayesian network algorithms in replicating a real-time bitumen conversion system and potential of SMCR for online monitoring of a continuous thermal cracker is also studied.

Keywords: Curve resolution; Bayesian structure learning; alternative least squares optimization; particle swarm optimization; conversion chemistry; reaction network.

[5] This work was submitted to *Reaction Chemistry & Engineering* as 'Sivaramakrishnan, K.; Puliyanda, A.; De Klerk, A.; Prasad, V. A data-driven approach to generate pseudo-reaction networks for thermal conversion of Athabasca bitumen'.

6.1 Introduction

The chemistry behind the thermal conversion of oilsands bitumen is quite complicated. Most of the proposed reaction networks in the literature for thermal cracking of bitumen over a wide range of temperatures and residence times involve compound classes segregated based on phase classification rather than individual chemical components due to the obvious difficulty in identifying the constituent species in bitumen. [1,2] Though successful attempts to identify the molecular structure and composition of the heavier components of bitumen like asphaltenes have been made, tracking their changes during thermal treatment is a difficult task. [3] Achieving composition control by setting up distributed monitoring networks to measure process variables is expensive and inefficient. [4]

The development of hyphenated analytical techniques based on spectroscopy and chromatography has facilitated the enhanced characterization of analytes in various fields of petroleum, catalysis and analytical chemistry. [5–10] The data from these techniques serve as the building blocks for developing the reaction network for a chemical system since empirical models are more practical to develop than a first-principles model for complex mixtures. Specific to bitumen, Fourier transform infrared spectroscopy (FTIR), [11–13] proton nuclear magnetic resonance (^{1}H-NMR), [14] and electron spin resonance (ESR) [15,16] have been applied frequently to obtain information on physical and chemical properties like the presence of hydrogen bonding, aromatic, nonaromatic and heteroatomic content, and free radical concentration. The data from these measurements can also be used for qualitative and quantitative analysis. The major advantages of applying spectroscopic techniques for complex mixtures are that they require small amounts of samples, have shorter processing times and do not contaminate the sample due to their non-invasive nature. [17–19] The inclusion of accessories like flow cells, quartz windows and immersion probes also facilitate faster characterization. [20] They also provide avenues for online monitoring of the system, which is important if the goal is automation and control. [21] However, the challenge is that the data obtained is multi-dimensional and often represent overlapped spectra from a vast number of components.

The traditional assumption of a reaction pathway for bitumen during thermal conversion is that the maltenes, which are soluble in a light paraffinic hydrocarbon, lead to asphaltenes formation, which eventually convert to coke and gas. [22] Kinetic models provide a good estimate

of the probability of the occurrence of each constituent reaction through calculations of kinetic parameters like rate constant and activation energy. In one such study, Kapadia et al. [22] suggested an improvement to the conventional kinetic model for bitumen by splitting the gas into individual gaseous products of H_2, CH_4, CO, CO_2, H_2S and adding another component called heavy molecular weight gases (HMWG) consisting of carbon numbers upto C_5. However, the drawback of kinetic models is that a reaction network is always required to be assumed prior to performing calculations and lumping of components also creates issues in interpretation. If the model is based on macroscopic properties like viscosity as done by Shu and Venkatesan, [2] sample-to-sample variability in such properties can also contribute to possible sources of error. [23,24]

Chemometric techniques involving statistical approaches have been shown to be quite useful in tackling the challenges of higher dimensional data and the limitations of a kinetic model. [25–28] Their principal benefit is the requirement of minimal prior knowledge of the system, both mathematically and chemically and the ability to operate with fewer assumptions. They are used to convert data to valuable information that assist in further processes that require human intervention like incorporation of chemical knowledge of the system to develop reaction pathways and obtaining an insight into the reaction chemistry. For these reasons, chemometric resolution, clustering and learning methods were employed in this work to analyze spectroscopic data of the liquid products obtained from thermal conversion of bitumen derived from an Athabasca oil sands deposit.

At this point, it is worthwhile to point out certain recent studies by Tefera et al. [29,30] where they utilized chemometric resolution and machine learning tools to investigate thermal conversion process of Cold Lake bitumen in a large temperature range of 150 – 400 ºC. The temperature-wise resolution and subsequent constrained optimization yielded results that suggested the formation of free-radical addition products and the conversion of lower-substituted aromatics to higher-substituted aromatics. This shift towards heavier products was also correlated with a peculiar trend in product viscosity where it was observed to increase at higher reaction times. [31–33] On the other hand, the reaction network developed on the basis of the Bayesian learning approach applied to the same data indicated the formation of lighter aliphatic products. [30] These results were quite intriguing and prompted us to conduct a similar study on Athabasca bitumen that resembles Cold Lake bitumen in overall chemical composition except for a lower saturate content, higher asphaltene content and viscosity. [23,34,35]

Depending on the reaction times for which the bitumen sample was held at each temperature in a batch reactor, two regimes were inspected in this work: visbreaking and coking. Visbreaking corresponds to the times before solid organic particles start forming in significant amounts and the coking regime that follows has measurable coke content. [32] Industrial visbreakers employ temperatures of 430 – 490 ºC, pressures in the range 0.3 – 2 MPa and the residence times are in the order of seconds or minutes in a coil visbreaker (but much larger in a soaker-type visbreaker). [36,37] Typical conditions for industrial delayed coking are 480 – 510 ºC and higher residence times of upto 24 hours under low pressures of ~0.6 MPa. [38] In this work, Athabasca bitumen was subjected to temperatures of 300 – 420 ºC under 4 MPa inert atmosphere (N_2) at residence times ranging from 15 min to 27 hours. FTIR spectra of the reaction products were obtained for a total of 35 samples including the feed bitumen.

Two types of chemometric analyses were conducted on the FTIR data, namely self-modeling multivariate curve resolution (referred to as SMCR or MCR) and Bayesian hierarchical clustering (BHC) combined with a causality detection method using Bayesian learning that employed heuristic search algorithms to recover the underlying network structure in the data. The MCR approach is self-sufficient in that it does not require any additional mathematical or chemical information apart from spectral data to perform the deconvolution. For this reason, the term 'self-modeling' is prefixed and implied when referred to as MCR in this work. The SMCR methods were applied on the datasets comprising of each temperature separately, called local models and also applied on the combined data from all temperatures, referred to as the global model. Certain quantitative parameters from the FTIR intensities are also calculated for each resolved model-spectra to aid with the tracking of chemical changes with time and get an insight into the specific kinds of reactions occurring during thermal conversion. The curve deconvolution process involved three major steps: (i) data matrix decomposition as a means of exploratory analysis based on to find out the number of components that are active and change in concentration during the reaction; (ii) obtaining initial estimates of concentration or spectral profiles for the determined number of active species; (iii) final resolution through a constrained optimization to retrieve the change in concentration with time and the individual spectra for each component.

The number of active components (more appropriately described as pseudo-components since they are model-derived) are extracted through two methods: (i) the conventional principal component analysis (PCA) [39] by means of singular value decomposition (SVD); (ii) Elbergali's [40]

recommendation based on the maximum ratio of derivatives of second and third order of Malinowski's [41] indicator function (which is based on real experimental error). As a separate case, one or two more components than the optimal number were chosen for further optimization and the results were compared. Fixed-size moving window evolving factor analysis (FSMW-EFA), [42] which is an iterative method, was used to obtain the initial estimates of concentration profiles in the SMCR method since it was suggested as a better approach to distinguish concentration regions of a component from the noise as compared to forward/backward EFA employed by Tefera et al. [29] Other contemporary techniques having similar goals to EFA are rank estimation-based methods like generalized rank annihilation method (GRAM), [10] window factor analysis (WFA), [43] and sub-window factor analysis (SFA) [44] that includes elution limits for interfering compounds as well.

Iterative EFA was chosen over other methods like non-iterative EFA since the data used in our study possessed an evolutionary structure and the algorithm did not require much user-mediation as well as it could be automated easily. Alternating least squares (ALS) is a common optimization technique for obtaining the final concentration and spectral profiles and is utilized in this work, as was also adopted by Tefera et al. [29] for the analysis of the spectra of Cold Lake bitumen.

However, there are certain limitations to the algorithm in that it does not always reach the global minimum. Data-related problems like collinearities present among the variables, non-ideally distributed noise patterns, background signals and algorithm-related issues like rank deficiency, intensity and rotational ambiguities can also hinder the accuracy of the final solution. Using multiple initial estimates or Monte-Carlo methods are useful methods to tackle some of these issues. [45]

A key difference in the approach used in this work as compared to that of Tefera et al. [29] is the inclusion of particle swarm optimization (PSO) to improve the ALS-obtained concentration profiles. PSO is a population-based meta-heuristic technique that is inspired by the natural phenomena of bird flocking. [46] Its primary advantage over other optimization methods apart from convergence to the global minimum is that the search space does not constitute any restrictive assumptions. In addition, though it is similar to genetic algorithms (GA), it is computationally simpler and faster. [47] The results from SMCR-ALS-PSO were compared to that of SMCR-ALS in the local models for Athabasca bitumen while only ALS was investigated for the global model in the first part of this work.

In the second part of the analysis, BHC was applied as a data-driven approach to the FTIR data to create 5 groups of wavenumbers that was later identified to represent specific categories of compound classes. BHC is a model-based agglomerative clustering method that operates on the basis of maximization of the marginal posterior probability to identify similarities in the data. [48,49] It is preferred over distance-based clustering methods for high-dimensional data where the number of variables is greater than the number of samples, as is our case (35 reaction conditions as samples and 1738 spectral channels as variables). Once the clusters were identified, it was necessary to find the magnitude and direction of the relationships between each of them in order to develop a plausible reaction network. This was done using a probabilistic graphical modeling approach (Bayesian network) of inferring causal reactions from spectral data to hypothesize the reaction pathways based on the previously identified reaction chemistry. Bayesian networks were used rather than Granger causality or transfer entropy due to the smaller number of observations in the data structure and the unavailability of time-series data for the system. [50] Hill climbing, [51] Tabu search [52] and maximum-minimum hill climbing (MMHC) [53] algorithms based on a score criterion were used to find the optimal solution for the Bayesian learning procedure.

The key contribution of this study was to combine basic chemical knowledge of the system and the results of both classes of chemometric techniques to build a conceivable reaction pathway for the thermal conversion of Athabasca bitumen. The results were compared with that obtained for Cold Lake bitumen in terms of differences in reaction chemistry with variation in chemical composition and possible relation to physical properties like viscosity. The consistency of the results from the global model was verified with that of the Bayesian methods since both operated on the entire dataset comprising of all temperatures. The suitability of the local and global SMCR models for online monitoring of a continuous thermal cracking system compared with Bayesian methods is also briefly discussed. A short discussion on the implication of the reaction pathway on partial upgrading of bitumen at lower temperatures is also provided.

6.2 Experimental

6.2.1 Materials

The thermal conversion experiments in this study were conducted on Athabasca bitumen feed supplied by Suncor Energy. Elemental composition and important properties like density, refractive index and aromatic content of the feed were already provided in Table 5.2 in Chapter 5. The same feed was used for thermal conversion experiments at other temperatures in this study.

During the thermal reaction, nitrogen (99.99%) provided by Praxair, was used to pressurize the batch micro-reactors and maintain an inert atmosphere. All reactions and experimental analyses were conducted in duplicate. After each reaction, methylene chloride (99.5 % dichloromethane) and toluene (99.9 %), both obtained from Fisher Scientific, were used to clean the reactor and tubing system. Solid particles stuck to the inner walls of the tubing were removed using a pipe brush, also supplied by Fisher Scientific. The crystal in the attenuated total reflectance (ATR) system that was attached to the FTIR instrument was cleaned with acetone (99.6 %), also supplied by Fisher Scientific.

6.2.2 Equipment and procedure

The experimental procedure employed for the collection of thermally converted samples subjected to FTIR characterization was quite similar to that employed by Tefera et al. [29,30] in their work on Cold Lake bitumen. The difference was that no solvent was used to extract the products from the batch reactors in our work, while the products in Tefera et al. [29,30] were extracted using methylene chloride. The effects of methylene chloride extraction on the properties of thermally converted samples of Athabasca bitumen were investigated by Sivaramakrishnan et al., [24] in Chapter 5 which was why its use was avoided in this work.

Four stainless steel (Grade 316) batch micro-reactors attached to a self-built tubing system were utilized for the bitumen conversion experiments in this study. The dimensions and the corresponding pressure rating of the reactors were as follows: inner diameter of 0.0254 m, tube wall-thickness of 0.0021 m, length of 0.086 m with a maximum allowable pressure of 18.6 MPa. In this study, reactions were conducted at 5 different temperatures (300 ºC, 350 ºC, 380 ºC, 400 ºC, 420 ºC), for which they were immersed in a fluidized sand bath (Omega Engineering FSB-3), whose temperature could be set by a temperature controller. Care was taken to maintain a near-constant airflow to the sand bath to avoid temperature fluctuations. Since all reactions were conducted in duplicates, 4 reactors were immersed at once, out of which 2 corresponded to one

reaction time and the remaining two to another reaction time. Considering that no back-pressure regulator was used to control the pressure during the reaction, and also a previous observation that an increased pressure could limit the coke formation, [54] all reactions were conducted at an initial pressure of 4 MPa. This increased to 6-8 MPa depending on the reaction conditions but was well below the pressure limit of the batch reactor. A soap solution (Snoop, supplied by Swagelok Inc.) was used to check the presence of leaks in nitrogen gas from the reactor before the reaction. To avoid direct contact of bitumen from the inner side walls of the reactor, an 8 mL glass vial was used to hold the sample inside the reactor.

It took approximately 15 – 19 min for the reactors to reach the set temperature for set points between 300 ºC and 420 ºC. After the reaction was complete, the two reactors were removed from the sand bath and cooled to 25 ºC by using compressed air first and then dipping the reactor in water at room temperature. This entire process took around 16 – 18 min, depending on the reaction temperature. Due to the formation of gases during the reaction, the final pressure after cooling down increased to 5.5 MPa at 420 ºC, depending on the reaction time, but the pressure remained at 4 MPa at 300 ºC. Any sand particles that were stuck to the reactors were removed and then the reactors were depressurized to collect the gases in a gas bag. The liquid and solid products were directly collected in a 120 mL amber glass container and stored under nitrogen. Since the objective of this study was to extract information from the FTIR spectra of the liquid samples through multivariate techniques, the solids in the sample were allowed to settle at the bottom of the storage vial to minimize interference during spectral measurements. Only the liquid part was taken for FTIR measurements. During transfer of samples from the reactor to the storage vials and from the storage containers to the FTIR equipment, the time of exposure to air was kept less than 2 minutes for each sample to minimize any chance of oxidation.

6.2.3 Analyses

6.2.3.1 Feed characterization

A brief description of the characterization procedure and the equipment used for the evaluation of the feed properties (Table 5.2 in Chapter 5) is given in this section. The viscosity was measured using an Anton Paar RheolabQC viscometer that consisted of a stationary cup and

a rotatable bob arranged as concentric cylinders with the bob on the inside. The shear rate employed was 10 s^{-1} and all measurements were conducted at 40 ºC. Viscosity of the liquid products have also been measured in the same viscometer but at different shear rates depending on the fluidity of the product. This was because it was necessary to measure in the Newtonian region of the samples since they were observed to be shear thinning; a detailed study regarding this has been conducted in Chapter 5 on the liquid products from Athabasca bitumen at 400 ºC. Wherever the viscosity values are reported in the Results and Discussion section (section 6.2.4) of this work, the corresponding shear rate is also given.

A U-tube densitimeter (Anton Paar DMA 4500M) with an accuracy of 0.01 kg/m^3 was used to measure the density of the feed at 25 ºC (accurate to 0.01 ºC). The refractive index was measured using an Anton Paar Abbemat 200 refractometer with an accuracy of 0.0001. Sodium D-line at a wavelength of 589 nm with air as the reference was used to take the measurements. Elemental analysis of the feed was performed in a Thermo Scientific Flash 2000 CHNS-O organic elemental analyzer but the oxygen content was determined by difference. Proton NMR (^{1}H-NMR) spectra was obtained using a Nanalysis 60 MHz NMReady-60 spectrometer to determine the nature of the different types of hydrogen present in the feed. Since the instrument was pre-calibrated with chloroform-D, the bitumen sample needed to be dissolved in chloroform-D before making obtaining the spectra. Typically, about 130 mg of the feed sample was dissolved in 0.7 mL of the solvent and poured into a 5 mm OD Norell^R Standard Series NMR tube supplied by Sigma Aldrich. The spectral width, number of scans per sample, and the time per scan used were 12 ppm, 64, and 23.4 seconds, respectively.

6.2.3.2 FTIR spectra

An ABB MB3000 spectrometer fitted with a single reflection ATR accessory (model: PIKE MIRacle^{TM} manufactured by PIKE Technologies) consisting of a diamond shaped crystal plate was used for the measurements. For each run, about 10 mg of sample was required so as to cover the crystal plate, which needed to be cleaned with acetone and completely dried before the next run. Air was used as the reference in the transmittance mode with other spectral parameters as follows: resolution - 4 cm^{-1}; number of scans - 120; wavenumber range - 4000-600 cm^{-1}; detector gain - 81 dB.

6.2.3.3 Gas chromatography for analysis of gaseous products

In our work, an Agilent 7890A gas chromatograph combined with a flame ionization detector (FID) as well as a thermal conductivity detector (TCD) was utilized to analyze the composition of the gas phase from the thermal conversion reactions at various conditions in this study. This was mainly used to identify the dominant species present in the gas phase for assisting with chemical interpretation as detailed in the 'Results and Discussion' section in this chapter. The FID detected hydrocarbons (mostly C_1-C_6) while the TCD was responsible for detecting other gases like CO_2, H_2S, CO, N_2 and Ar. The gases were injected into a HayeSep R column (2.44 m × 0.003 m) to perform the separation. Helium at a constant flowrate of 25 mL/min was employed as the carrier gas and the injector temperature was set at 200 ºC. The following are the details for the temperature method exercised to evaluate the gaseous products. The initial temperature was 70 ºC, held isothermally for 7 min after which it was increased to 250 ºC at the rate of 10 ºC/min and then held at that temperature for 2 min. Finally, the temperature was decreased at the rate of 30 ºC/min to 70 ºC and was then maintained constant for 8 min.

6.3 Methods and parameters used

The theory behind the chemometric techniques used in this work, multivariate curve resolution and Bayesian clustering and network learning methods is provided in this section. In addition, the software tools and the related functions along with the respective important parameters adopted in each step are highlighted. The reasoning behind the steps implemented during the mathematical analysis is also mentioned wherever appropriate.

6.3.1 FTIR data

As alluded in the experimental section, thirty-five samples of liquid products were collected from thermal conversion at 5 different temperatures of 300 ºC, 350 ºC, 380 ºC, 400 ºC and 420 ºC at various reaction times ranging from 15 min to 27 hours. The number of samples at each

temperature and the respective reaction times that were used in this study are summarized in Table 6.1.

The FTIR spectra of these samples were obtained at 1764 spectral channels in the wavenumber range between 4000 – 600 cm^{-1} (16666 – 2500 nm). Out of these, only 1738 points were used for modeling purposes as the wavenumbers in the region 650 – 600 cm^{-1} corresponded to instrument noise due to the ATR and appeared as random peaks with arbitrarily high values of transmittance. The transmittance data was converted to absorbance units by their logarithmic relation, which is also related to Beer-Lambert's law. [55] This raw absorbance data is shown in Figure 6.1.

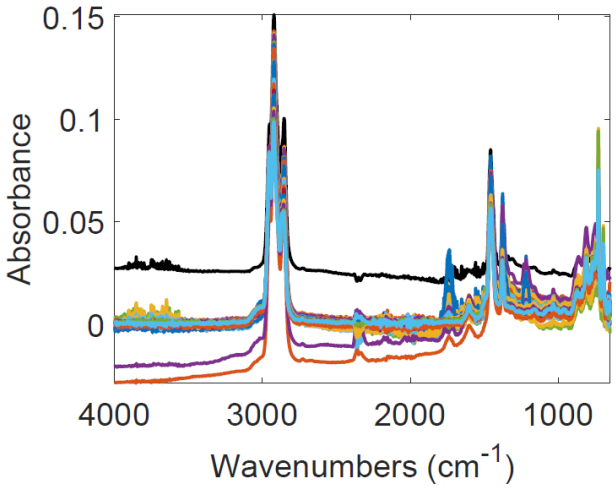


Figure 6.1. Raw FTIR absorbance spectra of 35 liquid products from thermal conversion of Athabasca bitumen at five different temperatures and reaction times before pre-processing.

As can be seen from Figure 6.1, the region from 4000 – 3200 cm^{-1} exhibited mostly baseline intensities with no peaks worthy of chemical interpretation though O-H and N-H groups present in bitumen (both free and hydrogen bonded) absorb in that region. These regions are classified into one cluster in the BHC and elaborated in section 6.4.5.

Table 6.1. Experimental conditions of thermally processed samples used for data analysis.

Temperature	Number of samples	Reaction times at each temperature (min)
Feed	1	-
300 ºC	2	360, 1080
350 ºC	6	30, 60, 180, 240, 360, 480
380 ºC	6	120, 240, 360, 480, 1320, 1620
400 ºC	16	15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 180, 210, 240, 360, 1170, 1440
420 ºC	4	360, 420, 480, 660

6.3.2 Pre-processing of FTIR data

Raw spectroscopic data is multi-dimensional and may have issues such as significant spread, different units of measurement among the variables, heteroscedasticity, possible experimental error and inherent instrument noise that is unavoidable. Existence of these features in the data may hinder further processes of rank determination, curve resolution, clustering and causal structure learning. It was thus necessary to subject the raw data for pre-treatment that consisted of three steps: (i) baseline correction; (ii) smoothing; (iii) normalization. With the objective of identifying major types of reactions occurring at each temperature over time, the SMCR method was applied to datasets that were split temperature-wise as given in Table 2. This helped to identify whether there was a difference in the reaction chemistry at low and high temperatures. However, for the Bayesian learning approach, all temperatures were considered together since the goal was to derive a reaction network for Athabasca bitumen as a whole, as was done with Cold Lake bitumen. [30]

MATLAB R2017b (9.3.0) was used to carry out all the chemometric analysis in this study. *'msbackadj'* function belonging to the Bioinformatics toolbox was used to perform the baseline correction. First, the wavenumbers were split into windows, each of width 200 separation units, which is the default window size. Adjacent windows are located at a distance of 200 wavenumbers units from each other and are given by the step size. A baseline value was found for every window through an expectation-maximization algorithm and these estimated points were regressed further

to smoothen the curve through a piecewise cubic interpolation that is given by the function *'pchip'*. After baseline correction, smoothing was performed using *'mssgolay'* function, also located in the Bioinformatics toolbox, which used the well-known Savitzky-Golay (SG) filter, in which a least-squares two-degree polynomial is used for removing the noise from the spectra every 5 samples (window size). [56] It was important to note that the normal SG filter required the wavenumber units to be equally spaced but the *'mssgolay'* function allows for unequally spaced wavenumbers as well.

Normalization by mean-centering and auto-scaling was executed using the *'zscore'* function from the Statistics and Machine Learning toolbox. Mean-centering removes offsets in the data while auto-scaling is a variance-based scaling method that brings all intensity data between 0 and 1. [57] These two processes are necessary to deal with variability in the variables in the data that will affect the results of further exploratory and regression analysis like rank determination, clustering and curve deconvolution. Other types of scaling include pareto, range, and vast scaling which also act on the variance of the data, out of which range scaling is sensitive to outliers. [58–60] Level scaling is an average-based method that can be used when changes on a relative scale are more significant than absolute values in the data. [57] These have already been highlighted in Chapter 2.

All of these pre-processing steps including baseline correction, smoothing and standardization were applied to the FTIR data before proceeding with curve resolution and Bayesian clustering and learning methods. The pre-processed data (only baseline corrected and smoothed) along with the residual after pre-processing as compared with the raw data for the spectra of liquid samples obtained at 350 ºC is shown in Figure 6.2. Similar results were obtained for other temperatures as well and is supplied in section D.1 in Appendix D.

The advantage of pre-processing the data is clearly visible in Figure 6.2a where the spectral features are more distinctly seen than in the raw data (Figure 6.2c). The stretching vibrations of sp^3 hybridized C-H methylene groups that can belong to either alkyl side chains or naphthenic rings occur at ~2850 cm^{-1}, 2920 cm^{-1} while those of methyl C-H stretches can be seen at slightly higher wavenumbers of 2950 cm^{-1}. These are the highest intensity peaks in the spectrum. The second most intense set of peaks correspond to the bending vibrations of sp^3 C-H groups at 1380 cm^{-1} and 1460 cm^{-1}. The set of aromatic C-H bending vibrations fall in the 690 – 900 cm^{-1} range that comprise of mono-substituted aromatics (with more than 4 adjacent hydrogens) peaks at ~727 cm^{-1}, *o*-disubstituted aromatic peaks at 744 cm^{-1} and 763 cm^{-1}, and the *m*- and *p*-disubstituted

aromatic peaks at 810 cm^{-1} and 860 cm^{-1} that also overlap with substituted alkenes. The C=O stretching that appeared at ~1740 cm^{-1} corresponds to ester-type and anhydride-type (more probable) functional groups that were thought to be converted to carboxylic acids by hydrolysis and eventually decarboxylated. In addition, the peaks at ~1220 cm^{-1} indicate the presence of alcoholic and acyclic C-O groups, whose chemistry during thermal conversion has not gained much clarity yet.

The spectra of the samples at each temperature at different reaction times vary only slightly in intensity and represent a mixture of components whose structures are unknown. Significant overlap of functional groups also occurs in the 1550 – 1650 cm^{-1} region that corresponds to both the aromatic and alkene C=C stretches. No significant chemical interpretation can be derived by viewing the spectra in isolation. Hence, it made sense to deconvolute the spectra to extract the concentration and spectral profiles for a lesser number of pseudo-components that were representative of the change in properties of bitumen with time and further assigning chemical identities to wavenumber groups derived through Bayesian clustering to obtain a reaction network.

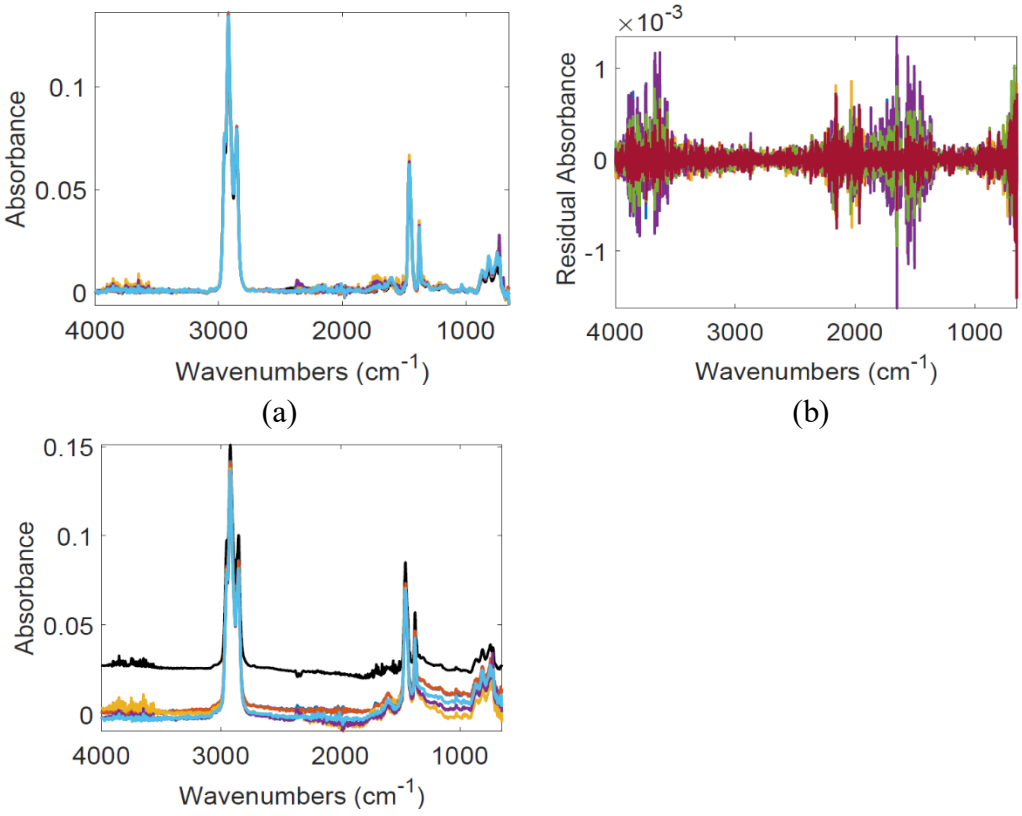


Figure 6.2. Plots of: (a) Baseline corrected and smoothed data; (b) residual after smoothing and (c) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 350 ºC.

6.3.3 SMCR-ALS and SMCR-ALS-PSO

Spectral measurements consisting of multivariate responses are obtained in a number of industrial processes. These measurements are generally cast into data matrices, which can be decomposed in a bilinear or tri-linear fashion using SMCR or PARAFAC, respectively. [61–64] SMCR, is a soft-modelling technique that utilizes factor analytical decomposition and invokes physically meaningful laws like Beer Lambert's law. [64] Beer's law is also a bilinear model that relates the absorbance of a light-irradiated species to its concentration and the path length.

6.3.3.1 Data decomposition

Let D be the data matrix that is composed of the FTIR spectra of the m samples obtained at n spectral channels. In our study, m is 35, n is 1738 for the whole set of spectra while m varies when modelled for temperature-wise data. In the SMCR bilinear model, the data matrix is decomposed as follows:

$$D = CS^T + E \qquad \text{Equation 6.1}$$

where C is an $m \times l$ matrix that contains the concentration of l components, S is a $l \times n$ matrix that consists of their resolved spectra, and E $(m \times n)$ is the residual matrix that contains the error of decomposing D into the constituent C and S profiles through SMCR. l is the rank of the data matrix D that is representative of the number of active components in the bitumen mixture and needs to be found first before proceeding with solving equation 6.1. The procedure for finding the number of components is given in section 6.3.3.4.

6.3.3.2 Multiset structures

If the experiments were analyzed with more than one characterization technique, equation 6.1 can be changed to a row-wise augmented matrix where D and S are arranged as the combination of different measured and resolved intensities while the concentration is unaltered as the process analyzed is the same. On the other hand, D becomes a column-wise augmented matrix when multiple groups of experiments are analyzed with the same characterization technique and C is split up into individual concentrations for each set of experiments. Here, the spectral shape of the component does not change over different conditions, as for example in a high-pressure liquid chromatography with diode-array detection (HPLC-DAD) system. The formulation of augmented matrices is given in equation D.1 and equation D.2 in Appendix D. In this work, a combined dataset formed by combining the data from all 5 temperatures is also analyzed to check for improvements in the final spectral resolution and consistency with the local models.

6.3.3.3 Algorithms to solve SMCR

Either non-iterative or iterative algorithms are used to solve equation 1. [65] Non-iterative methods like heuristic evolving latent projections (HELP), WFA, orthogonal projection analysis (OPA), [66,67] and parallel vector analysis (PVA) [68] are more useful when a single experiment is

analyzed by a single characterization technique. It is essential that the concentration profiles follow a sequential structure for non-iterative methods to work. On the other hand, iterative methods like iterative target transformation factor analysis (ITTFA), [69] ALS and resolving factor analysis (RFA) overcome these limitations of non-iterative methods and just require initial estimates of C or S to arrive at the solution.

ITTFA optimizes the concentration profile with certain constraints and then calculates the spectral profile from D and C while ALS calculates an optimized C and S simultaneously at each iteration. They are adaptive in nature and do not necessarily require the concentration region to follow a sequential structure. They allow for meaningful chemical and mathematical information to be included as constraints, which can tackle the problem of ambiguity that is often prevalent in the MCR solution. [70] The SMCR method can find the solution even without these additional inputs (which is why it is called self-modeling) but these are added by the user based on knowledge of the system. Iterative algorithms also have the added advantage of having the ability to deal with augmented matrices and extracting the solutions for each experimental run and characterization technique. Hence, we chose to optimize the SMCR with an iterative method instead of a non-iterative approach.

The ALS-based iterative optimization approach can be viewed as a block-relaxation algorithm applied to a least squares loss function which is non-convex. Here, the minimization subproblems over blocks of decision variables that are updated alternately in an iteration have convex objective functions. [71] The algorithm is relatively fast and computationally simple compared to other contemporary methods. The ease of incorporating chemical knowledge about the system (some known spectra of a component class that can be present in bitumen), mathematical and natural constraints, and the ability to merge the rank and initial estimate determination process with ALS optimization prompted us to proceed with this method. Details of extracting the number of significant contributing components, obtaining initial estimates for C and S, the associated constraints and some limitations such as ambiguous solutions are provided in the next few parts of this section.

The objective function to be minimized in the SMCR-ALS routine is the 2-norm (entry-wise matrix norm) of the residual obtained from the SMCR solution (given in equation D.5 and equation D.6 in Appendix D). The first task consists of minimizing the residual using a least squares approach to calculate S, given D and an initial estimate of C, which is calculated by FSMW-EFA

(section 6.3.3.5). [72] In our work, this is achieved by using a user-defined function that is similar to the *'lsqnonneg'* function in the Optimization toolbox in MATLAB. Matrix right hand division is used where both the known matrices are required to have the same number of columns (m in this case for D and C). This function includes the application of non-negativity and unimodality constraints. A new estimate of C is obtained from the calculated S in the previous step and the data matrix. The residual is calculated in each iteration and the loop is stopped by assessing the difference between residual values in the current step and the previous step, i.e. when a near-constant value for the residual is obtained.

The final solutions to the concentration profiles were verified for adherence to the mass balance constraint. Since the concentrations were normalized, all values were between 0 and 1 and in no particular unit. The summations of the concentrations of the number of active components involved was found to be closer to 1 when optimized by the ALS-PSO method across all times for the models at each temperature. The exact concentration profiles and further discussion are shown in section 6.4.3 of this chapter.

6.3.3.4 Determination of chemical rank of the system

This is a process of identifying the number of active components that are participating in the thermal reaction and undergoing a chemical change during the time of reaction. SMCR requires this information prior to estimating the initial concentration and applying the constrained optimization algorithm. [73] However, it is always not necessary that the extracted number of principal factors, i.e. the chemical rank (l), is equal to the number of active species in the system and when the former is lesser than the number of actual species, the matrix is said to be rank deficient. This is a common situation in real data that consists of redundant, overlapping spectra along with noisy backgrounds. [74] In the case of bitumen, the composition is quite complex and the number of components constituting bitumen and further taking part in thermal reaction is very difficult to determine. Thus, the challenge is to identify the number of species responsible for the observed spectral and concentration change and separate them from the inert ones.

Malinowski [75] has comprehensively reviewed a number of empirical and statistical methods for the prediction of rank in cases where no assumptions regarding noise as well as situations where full information regarding experimental error was available. The empirical indicator functions

could also be combined with a non-iterative partial least squares (NIPALS) that is sometimes used for PCA where the loading vector is normalized, [76] but the true dimension of the factor space needs to be determined irrespective of the knowledge or presence of error of any type, which can be achieved by looking into the theory of error developed by Malinowski. [41]

In terms of eigenvalues that are used to reproduce the data matrix, primary eigenvalues correspond to the principal factors while secondary eigenvalues consist of the extracted error in the data and should not be included in the MCR model. Most empirical functions that assist in determining the dimension of the factor space are based on real error (RE), which is representative of the experimental error in the system. The RE is also the residual standard deviation (RSD) that is calculated as the difference between the original data and the PCA-decomposed or singular value decomposition (SVD)-decomposed matrix using l components (equation 6.3). [40] One such empirical function, called the indicator function (IND), was shown to exhibit a minimum when the appropriate number of factors were able to best reproduce the original matrix (equation 6.4). However, in the case of excessive error, a second minimum could be produced on the addition of further data points even though the assumption that the error was random and homoscedastic was valid.

Another error function called the imbedded error (IE) quantifies the amount of error remaining after factor decomposition of the spectral data which cannot be eradicated (equation 6.5). However, it was found from real datasets that the IND function was more sensitive to the inclusion of secondary eigenvalues that increased the error than the IE. [40] On the other hand, both IE and IND both increase continuously when the original data matrix is not factorizable and consists of random numbers. Elbergali [40] reported that for simulated data sets, both IE and IND reached a minimum value since the error was uniform for the entire dataset. The problem was with real-time experimental data such as fluorescence and HPLC runs, where finding the point of change in slope for the functions became difficult due to multiple data points showing similar values. The second derivative (SD) of the IND function was shown to display a maximum at the optimum point and have better sensitivity than IND and IE functions themselves (equation 6.6). Apart from these, the ratio of derivatives (ROD) criterion was discovered to be the best indicator for the rank determination (equation 6.7). At the point where the last primary eigenvalue is added, the ROD would show a maximum and combined with the minimum in the IND function, was suggested in the literature to be the best indicator among all the above ones mentioned. [77]

In our work, after the splitting of the data sets temperature-wise, SVD was applied to decompose each data matrix (D) to obtain the diagonal matrix of singular values (S) and the two unitary matrices (U and V). The residual was calculated by subtracting the product of U, S and V^T from D (equation 6.2), followed by its standard deviation considering the residual matrix as a single column vector of elements (equation 6.3). This procedure was also conducted on the augmented multiset structure consisting of all temperatures as well. The IND and ROD were calculated for each iteration up to the number of samples (rows) in D as per the following equations:

$$R(l) = D - USV^T(l) \quad \text{Equation 6.2}$$

$$RE(l) = RSD(l) = \sqrt{\frac{\sum_{j=l+1}^{n} R_j^2}{mn-1}} \quad \text{Equation 6.3}$$

$$IND(l) = \frac{RE}{(n-l)^2} \quad \text{Equation 6.4}$$

$$IE(l) = \sqrt{\frac{l}{n}} RE(l) \quad \text{Equation 6.5}$$

$$SD(l) = \log[IND(l)] - 2\log[IND(l-1)] + \log[IND(l-2)] \quad \text{Equation 6.6}$$

$$ROD(l) = \frac{IND(l-2) - IND(l-1)}{IND(l-1) - IND(l)} \quad \text{Equation 6.7}$$

The ROD and IND were checked for maximum and minimum respectively to determine the number of active components for the local and global models. Once the number of factors was chosen, the error remaining after the reproducing the original matrix from the decomposed data was calculated and reported. As a separate test case, the number of components was chosen as one higher than the SVD-determined value and solution profiles are compared with the optimum case to check whether there was any improvement in in terms of explained variance (R^2) and lack of fit (LOF). All these results are provided in section 6.4.3.

6.3.3.5 Obtaining the initial estimates

Once the number of components to be included in the SMCR model is determined, a key step is to provide the optimization process with an initial guess of the concentration or spectral profile. It is preferred that the initial estimates also satisfy the constraints applied in the optimization algorithm rather than arbitrary profiles. [64] Since the system investigated in this work has a sequence in an ordered variable (reaction time), i.e. experimental data is available at continuously increasing times at each temperature (Table 6.1), EFA was thought to be quite suitable for initial profile estimation. However, methods like simple-to-use self-modeling analysis (SIMPLISMA), [78] key set factor analysis (KSFA) [79] and OPA can handle unstructured data as well.

Forward/backward EFA works by applying PCA to find the eigenvalues (EVs) of sub-matrices of increasing size row-wise from the main data matrix. The difficulty here is to differentiate the EVs belonging to the noise from those indicating the presence of an actual component. FSMW-EFA was suggested to be an improvement over forward/backward EFA by Keller and Massart [42] as the noise EVs are also constant due to the fixed size of each window on which SVD is applied. Tefera et al. [29] used forward/backward EFA for the analysis of Cold Lake bitumen. In our work, a moving window of fixed size 3 was selected for carrying out FSMW-EFA for all temperatures except at 300 ºC where a window-size of 2 was used. This was because only 3 spectra were obtained at that temperature due to minimal gas production and miniscule changes in macroscopic properties like viscosity over long reaction times of up to 18 hours. It should also be noted that the higher the size of the moving window, the more robust the method becomes. The other advantage of FSMW-EFA is that it consumes lesser time as it needs to be carried out in only one direction. In addition, the concentration direction had a lower overlap in the component profiles than the spectral direction for thermally converted samples in our study. Therefore, it was logical to estimate the initial concentration profiles rather than the spectral profile.

Shinzawa et al. [80] compared the use of PSO against the use of EFA to obtain the initial estimate of concentration profiles in their work on application of SMCR on near-infrared (NIR) spectra of a mixture of oleic acid and ethanol. They showed that PSO performed better than EFA and yielded smaller residuals for the final solution. In our work, a hybrid PSO technique was introduced in an attempt to improve the profiles obtained by ALS, still using EFA for obtaining the initial estimates. The theory behind PSO and the procedure followed in our work is detailed in section 6.3.3.8.

6.3.3.6 Limitations: Ambiguities

The uniqueness of solutions from SMCR-ALS method is crucial for its reliability. However, the MCR solution potentially suffers from 3 types of ambiguity: intensity, permutation and rotational. [81] All ambiguities revolve around the fact that different combinations of C and S can produce the same parent data matrix D. Profiles having the same line shape and structure but different relative intensities/scales indicate the presence of intensity ambiguity. This is illustrated in equation D.3 in Appendix D. If concentration is optimized first in the ALS scheme, it is normalized before every iteration to help reduce intensity ambiguity. Permutation ambiguity implies that the order of the components can vary within the concentration and spectral profiles, while still yielding the same parent matrix. This does not affect the solution as much as intensity or rotational ambiguity in terms of further chemical interpretation.

The most commonly present ambiguity in SMCR results is rotational ambiguity, which is specified in equation D.4 in Appendix D. A combination of different line shapes of the component spectra and concentration can reproduce the original data matrix and affect the uniqueness of the solution. One way to deal with this is by appending subsets to D in a row-wise or column-wise manner, which would decrease the possibility of the number of solution profiles obeying the same constraints and reproducing the parent matrix as well. [82]

Since ambiguity is component-specific, it is only required that the profiles related to the active species are unambiguous, even if ambiguity exists in the rest of the profiles. Juan et al. [64] recommended to test the presence of ambiguity by calculating the extent and the location of the ambiguity. They suggested the calculation of two parameters that give the range of an objective function for each contributing component given by:

$$f_{i,min} = min \frac{||c_i s_i^T||}{||CS^T||} \qquad \text{Equation 6.8}$$

$$f_{i,max} = max \frac{||c_i s_i^T||}{||CS^T||} \qquad \text{Equation 6.9}$$

The lower the value of $f_{i,max} - f_{i,min}$, the lower the extent of ambiguity, implying the closeness to the uniqueness of the solution. To identify the location of ambiguity, dyads of profiles

can be plotted corresponding to $f_{i,min}$, $f_{i,max}$ for each component and check for deviations. However, this kind of testing was out of scope for our work.

6.3.3.7 Implementation of constraints

The strength of multivariate curve resolution is the ability to incorporate chemically meaningful constraints that compensate for the limited prior knowledge of the system. Their other function is to suppress ambiguities that arise in the solutions. [83] Constraints can be naturally part of the system like non-negativity (applicable to both concentration and spectra), closure (mass balance), unimodality (concentration specific) and equality (where some columns of the solution are forced to follow known spectra). Mathematical constraints include local rank that specifies where some components are absent and selectivity that indicates where one particular component is present in the time space. In the case of column-wise augmented matrices, correspondence of species and model constraints that facilitate trilinear or multilinear data for components whose spectral shapes do not change much across components can be used as constraints. [84]

Lastly, just like a known spectrum constrains a column of the solution, a physicochemical model with user-input parameters can be integrated with concentration regions. To predict concentrations of unknown analytes, a model relationship can be obtained between the SMCR-derived concentrations and those in calibration samples. This is implemented using the correlation constraint. [85] In our work, a local rank constraint was used during the initial estimate determination, while non-negativity and closure were employed in the ALS and PSO optimization routine. The case of known spectra was not employed in our work because the objective was to propose a reaction mechanism without much input of chemical knowledge of the system into the curve resolution algorithm.

6.3.3.8 PSO and its use as a hybrid technique in this study

ALS is not always robust in avoiding local minima that leads to insufficiency in the performance of the curve resolution process. As highlighted in the introduction (section 6.1), multiple local estimates or the use of Monte Carlo methods can serve as ways to tackle this problem. It was noticed by Tefera et al. [29] that some resolved spectra were uninterpretable and

showed only noisy peaks when ALS was used to resolve the spectra of Cold Lake samples. PSO, which is a nature-inspired technique, was recommended as an alternative and is employed in this study as a hybrid technique. Since they are as effective locally in the quest for neighborhood solutions, they are used in combination with a constrained optimization solver for searching for solutions in the local neighborhood, namely *'fmincon'* in the Optimization toolbox. [86,87] In this study, the merger of PSO and the local optimizer is embedded inside the ALS loop to further improve the ALS-produced concentration profiles with the same initial estimates as obtained by EFA (section 6.3.3.5).

In PSO, particles (which are the candidate solutions) are deployed in a bound search space to look for the best value of a user-defined objective function. [88] The particles are identified with two important parameters, namely their position and their velocity. For a D-dimensional space, the position of the i^{th} particle is given by $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$ and its velocity in different dimensions is represented by $v_i = (v_{i1}, v_{i2}, \dots, v_{iD})$. While performing the search, the value of the objective function is calculated wherever the particles move and compared with the previous values. If the current value is the least or maximum (according to the nature of the objective function) among all previous values in the path of that particle, then it is the personal best position of that particle ($p_i = (p_{i1}, p_{i2}, \dots, p_{iD})$). Similarly, a global best position ($pg = (p_{g1}, p_{g2}, \dots, p_{gD})$) is also calculated based on the comparison with the fitness values of all particles. These parameters are used to update the velocity and position of the particles through the following equations:

$$v_{id}^{new} = w.v_{id}^{old} + c_1 r_1 \left(p_{id} - x_{id}^{old}\right) + c_2 r_2 (p_{gd} - x_{id}^{old}) \qquad \text{Equation 6.10}$$

$$x_{id}^{new} = x_{id}^{old} + \mu.v_{id}^{new} \qquad \text{Equation 6.11}$$

where v_{id}^{new} and v_{id}^{old} are the updated and current velocity of the particle, x_{id}^{new} and x_{id}^{old} are the updated and current position of the particle in the search space, w is the inertia weight parameter, c_1 and c_2 are correction or learning factors also referred to as cognitive and scaling factors respectively, r_1 and r_2 are random numbers obtained from a uniform distribution and vary between 0 and 1, and μ is a time parameter that is used to update the position using the amended velocity.

The parameters in equation 6.10 and equation 6.11 are usually chosen from experience but Bansal et al. [89] reviewed the different strategies for choosing the inertia weight parameter and tested the performance of PSO in various scenarios. The first PSO model developed by Kennedy & Eberhart [90] did not include w to update the particle velocity and instead, the velocity was capped at a maximum value. As an improvement, inertia weight parameter was introduced later to have a trade-off between model exploration and the error, similar to the function of a regularization parameter in SVM. [91,92] In this work, we employ a constant w set to 1 since this strategy was shown to produce minimum error, albeit with a large convergence time. Most of the other strategies for the choice of w depend on the global and local positions of the particles and some common types are summarized in Table D.1 in Appendix D.

A value close to 1 for w allows for a global search but as w decreases towards 0, local search predominates. In this work, the MATLAB function *'particleswarm'* belonging to the Global Optimization toolbox was used to perform PSO. The 2-norm of the residual calculated using the original data matrix, ALS-obtained spectra and PSO-obtained concentration is used as the fitness function. A three-dimensional search space with a swarm size of 150 was utilized within lower and upper bounds of 0 and 1, respectively, since the concentration was already normalized. The parameters of swarm size were given by the function *'optimoptions'* and combined further with a constrained nonlinear optimizer, *'fmincon'*, both of which belong to the Optimization toolbox. The role of *'fmincon'* is to find a better local solution after the PSO terminates. *'fmincon'* employs an interior point algorithm by default, and is a large-scale algorithm that solves a quadratic optimization problem without generating or storing any matrices. Further details of other algorithms used to solve nonlinear constrained optimization problems are given in section D.2 in Appendix D. The default number of iterations for the PSO used is 600.

6.3.3.9 Process flow followed in SMCR-ALS-PSO method

Figure 6.3 provides the work flow followed in this work regarding the SMCR methods applied on the FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen. The reasoning for the use of the different steps employed is already highlighted in previous sections during the description of each method.

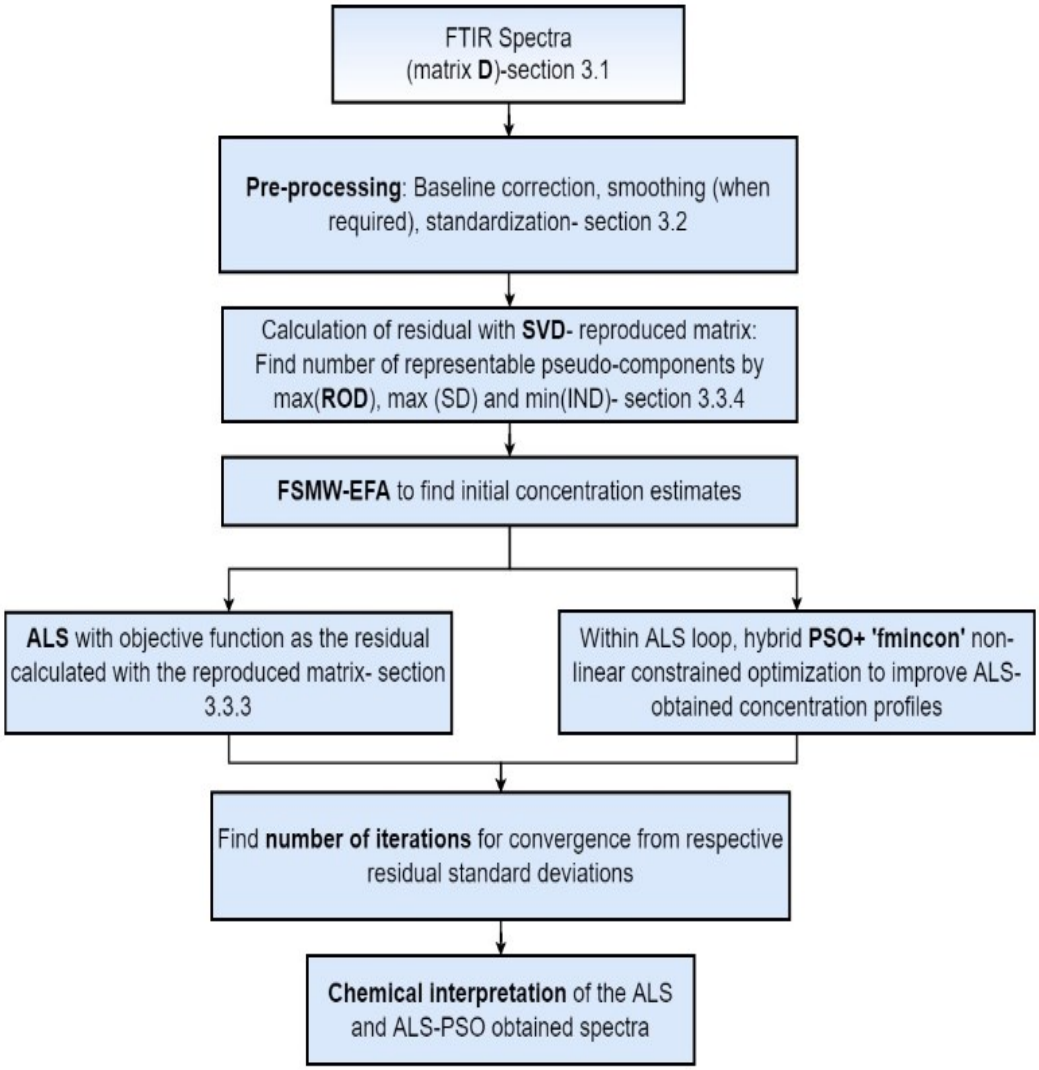


Figure 6.3. Sequence of steps followed in this work for chemometric analysis of the FTIR spectra through curve resolution.

6.3.4 Bayesian networks

Generating causal structures from experimental data in the framework of pairwise conditional independence tests between the process variables, [93] or causal Bayesian networks constructed on a Markov assumption, have gained application in diverse fields. [94] Obtaining quantitative structure-property relationships involves translating the information from chemometric models into knowledge of real chemistry using databases or expert systems. [95] These systems map

properties deduced from the algorithms to predict building blocks such as candidate molecular structures.

The current state-of-the-art methods for representation of reaction networks largely rely on encoding prior knowledge and the use of databases. Methods like the rule input network generator (RING) algorithm make it possible to manually encode reaction rules that generate reaction networks among reactants and products. [96] On the other hand, high throughput reaction prediction (HTRP) uses link predictions for binary reactions in multi-modal graphs whose feasibility is evaluated using filters implemented using the Tanimoto similarity score of fingerprints at each reaction node. [96] Hence there is an imperative need to develop data-driven causal methods to generate hypotheses on reaction networks of complex chemical processes, in the absence of prior knowledge of the process composition or physicochemical/kinetic models so as to develop an understanding of molecular level mechanisms.

Bayesian networks are a mathematically coherent framework for encoding causal relations as probabilistic graphical models in complex systems. They are known to be fairly robust in handling missing data, allowing for combining data with domain knowledge, protecting against overfitting and suitable for a high dimensional space with fewer samples. [97,98] Bayesian networks consist of nodes of random variables with directed acyclic links among them in accordance with the conditional Markov assumption. [99] Values of the random variables are described as probability distributions.

6.3.4.1 Bayesian hierarchical clustering

Keeping in mind the above discussion, most analysis tools used for building causal maps from experimental data are based on clustering algorithms, where groups of entities that have similar expression patterns over a set of experiments are grouped together. [30,98] In this work, a model-based clustering algorithm (BHC) was used to group wavenumbers that have similar absorbances over a set of collected spectral measurements. The Bayesian approach to clustering is based on marginal posterior probability rather than a distance measure to identify similarity in the data. [48,100] The random variables which comprise groups of similar wavenumbers are assumed to follow a multinomial distribution with a Dirichlet conjugate prior, as is typically the case with experimental data that could be noisy or incomplete. [97] All wavenumbers are denoted as data points and grouped

into K cluster nodes, each of sizes $T_1, T_2, \ldots, T_K$ such that D_i points (where $D = \{X_1, X_2, \ldots, X_N\}$ comprising multinominal data of all the random variables X_i that follow a distribution) belong to node T_i wherein the total number of data points is given by $\sum_{i=1}^{K} T_i$. [101] N represents the number of data points before clustering.

Initially, the clustering begins with as many nodes as the number of data points. The number of data points corresponds to the number of spectral channels, which was reduced to remove noisy regions from the raw spectra. In SMCR, the region from 650 – 600 cm^{-1} was not considered (as described in section 6.3.1); that gives a total of 1738 spectral channels, which was further reduced to 1550 wavenumbers for the Bayesian methods. This was done by first implementing PCA on the entire FTIR spectral dataset to obtain the scores (35 × 1764), loadings (1764 × 1764) and the eigenvalues corresponding to each principal component (1764 × 1). A score called the importance index that is the sum of the product of the loadings and the normalized eigenvalue of a principal component (PC) is calculated for each wavenumber. Next, the wavenumbers were arranged in descending order of the importance index and Bayesian networks were obtained from 5 clusters considering the first 1700, 1650, 1600, 1550 and 1500 wavenumbers (a trial and error approach) to check whether at least two of the algorithms for Bayesian learning (hill climbing, Tabu search and max-min hill climbing – detailed in section 6.3.4.2) produced the same network. It was interesting to note that when the first 1550 wavenumbers (according to descending order of importance index) were used, hill climbing and Tabu search produced the same network while maximum-minimum hill climbing gave a network that differed in the connection of one particular group with the rest but also made chemical sense. No two algorithms produced the same network when the first 1700, 1650, 1600 wavenumbers were used. Thus, BHC was carried out considering these 1550 wavenumbers for all the 35 samples in this work. A plot of the importance index of the wavenumbers is shown in Figure D.5 of Appendix D.

The pairwise merge of nodes $T_i + T_j \rightarrow T_k$ is based on whether data points in the nodes maximize the posterior probability (Υ_k) as given in equation 6.14 below:

Hypothesis 1 (H_1): Data in nodes is generated from the same mixture component

Hypothesis 2 (H_2): Data in nodes is generated from distinct mixture components

The maximum likelihood estimates (MLEs) of the above hypotheses are evaluated as:

$$MLE\ (H_1): P(D_k|H_1) = \sum_\theta P(D_k|\theta)P(\theta|\beta) \qquad \text{Equation 6.12}$$

$$MLE\ (H_2): P(D_k|H_2) = P(D_i|T_i)\ P\left(D_j\middle|T_j\right) \qquad \text{Equation 6.13}$$

$$\gamma_k = \frac{P(H_1)\ P(D_k|H_1)}{P(H_1)\ P(D_k|H_1)+(1-P(H_1))P(H_1)\ P(D_k|H_2)} \qquad \text{Equation 6.14}$$

The random variables are assumed to have the following distribution: $P(X_i = x_i \mid \theta) = \theta_i$, where $i = 1,2,\dots,N$ with N being the number of pre-defined clusters. Equation 14 is used to construct the likelihood function $P(D_k|\theta)$ as a product of the probabilities of the mutually independent random variables that represent the data. Here the parameters are $\theta = \{\theta_1, \theta_2, \dots., \theta_N\}$ where $\theta_1 = 1 - \Sigma\theta_i$ and these in turn correspond to the physical probabilities of the random variables that have a Dirichlet distribution as given below: [102]

$$P(\theta|\beta) = Dir(\theta|\beta_1, \beta_2, \dots, \beta_N) = \frac{\Gamma(\sum_{i=1}^{N}\beta_I)}{\prod_{i=1}^{N}\Gamma(\beta_i)} \prod_{i=1}^{N} {\theta_i}^{\beta_i - 1} \qquad \text{Equation 6.15}$$

Here, the vector of β_i, is the non-negative vector of scaling coefficients representing the hyperparameters of the distribution.

It is important to note that Bayes' theorem that states that the posterior probability depends on the product of the prior and the likelihood. The Bayes theorem is defined as:

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} \qquad \text{Equation 6.16}$$

In order to evaluate whether a cluster merge is successful or not, the Bayes' theorem is used in conjunction with the prior distribution to evaluate the maximum a-posteriori probability. Tefera et al. [30] used 5 clusters to represent the different groups in Cold Lake bitumen to investigate the causal relationships between them. They based this value considering that the major products of thermal cracking could be classified into five groups of compound classes namely, alkanes, alkenes, naphthenes, aromatics and polyaromatic hydrocarbons (PAH) that may include heteroatoms as well. Sivaramakrishnan et al. [103] applied LS-SVM regression to predict the

concentration of products from propylene oligomerization conducted at 150 ºC in a flow reactor. They used distance-based hierarchical clustering analysis (HCA) on the FTIR spectra of the products to classify carbon numbers of C_1-C_{10} into groups of 4 and 6 that represented a mixture of compound classes. Though each cluster could not be related to a single type of compound, they were able to capture the reaction chemistry quite well. In HCA, the dendrogram showing the agglomerative path for all data points plots the distance value against the data point number. In BHC, the maximum a-posteriori for each merge is used instead of a distance value. In our work, the dendrogram was cut off at a value of approximately 100,000 for making 5 clusters while the maximum height considering the whole data in one cluster was ~250,000. It should be mentioned that the dendrogram was quite complex and difficult to analyze due to the large number of variables.

6.3.4.2 Bayesian structure learning

Once the wavenumbers were clustered into respective groups, the next step involved employing algorithms to extract a structure or cause-effect relationship among the nodes (random variables). This encodes a probabilistic causal map between clusters of similar wavenumbers that represent a class of pseudo-components and their chemical interactions during the partial upgrading process. It can be seen as a method to facilitate the generation of reaction hypotheses. Working with Bayesian networks is a two-step process as follows:

a. Learning the structure of the network that encodes the conditional independence of the random variables represented by the nodes using either score-based, constraint-based or Bayesian model averaging-based methods. The joint factorization of the directed acyclic graph (DAG) that represents the derived causal relationships between the different groups can be expressed as:

$$P(X_1, X_2, \ldots X_K) = \prod_{i=1}^{K} P(X_i | \, Pa_{X_i}) \qquad \text{Equation 6.17}$$

where X_i indicate random variables represented by the nodes, Pa_{X_i} are the parents of the random variable at a node and K is the number of pre-defined clusters (5 in our case).

b. Estimation of the parameters of the joint factorization of the conditional probability distribution model of the Bayesian networks from the data using an expectation-maximization (EM) algorithm [104] by iteratively calculating the maximum likelihood estimate (MLE) for each of the parameters.

As mentioned above, there are 3 approaches to learn the structure among the nodes: (i) constraint-based, which relies on pairwise conditional independence tests between the random variables and is typically unreliable in high dimensional variable space; [105,106] (ii) score-based, in which a scoring function is used that evaluates how well a structure represents the data. Bayesian networks with N nodes have $2^{O(n^2)}$ possible structures, which makes it intractable to score all of them in order to choose the best one. Hence, we rely on greedy search methods to find the best possible structure from the data; [105] (iii) Bayesian model averaging-based, which uses an ensemble of possible structures from both the constraint-based and score-based techniques and then averages out the prediction of the ensemble, instead of just relying on one best structure. [105]

Constraint-based methods are sensitive to violation by even one data point and fit poorly in the presence of noise. [107] Moreover, a score-based method can handle data with lesser number of observations than the variables, thus making it the preferred choice for structure learning in this work. The Bayesian information criterion (BIC), which is the posterior probability of a structure given the data, is a frequently used metric employed in this work. It is penalized by the dimension of the structure to favor sparser networks. [98,108] The BIC is written in terms of the log likelihood (LL) of the data given the structure (G), which in turn can be expressed in terms of the mutual information (I) and entropy (H) pursuant to the connections among the nodes:

$$BIC(G,X) = LL(G,X) - \frac{log\ m}{2}\ Dim(G) \qquad \text{Equation 6.18}$$

$$LL(G,X) = m\sum_i I(X_i;\ Pa_{X_i}) - m\sum_i H(X_i) \qquad \text{Equation 6.19}$$

$$I(X_i;\ Pa_{X_i}) = \sum_{X_i}\sum_{Pa_{X_i}} P(X_i, Pa_{X_i})\ log\ \frac{P(X_i,Pa_{X_i})}{P(X_i)P(Pa_{X_i})} \qquad \text{Equation 6.20}$$

$$H(X_i) = -\sum_{X_i} P(X_i)\ log\ (P(X_i)) \qquad \text{Equation 6.21}$$

In equation 6.18 and equation 6.19, G indicates a certain graph structure of connections between a node (random variable) and its parents (Pa). 'm' is the number of samples in the data (in our work, it is the number of reduced wavenumbers which is 1550) and $Dim(G)$ implies the dimension of the graph structure in terms of the number of edges among the nodes.

As highlighted in the introduction, the algorithms which are typically used to find suitable structural connections to maximize the BIC include hill climbing, Tabu search, and MMHC. Hill climbing involves making locally optimum search iterations with random restarts while Tabu search is similar except that penalties are enforced when certain searches are repeated. MMHC is a hybrid algorithm that is an amalgam of both constraint-based and score-based methods. Since there is no guarantee that any one of these algorithms by themselves could reveal a structure that maximizes the BIC over an intractable search space of a large number of structures, Bayesian networks are constructed using all the algorithms separately. The belief in a structure to be the best is reinforced when more than one of these algorithms return an identical result. However, there is a chemical component attached to the extracted structure from a Bayesian network. In other words, even if all 3 networks do not match, or two match and the third one differs slightly, the one that makes most chemical sense (based on expert knowledge) would be preferred to infer the reaction network. In this work, the hill climbing with random restarts and Tabu search yielded the same network while MMHC produced a different structure. The chemical interpretation is given in section 6.4.3 and 6.4.6.

The statistical software R (version 3.4.3) was used to execute both the BHC and the structure learning techniques in this study. PCA is used before the BHC algorithm to order the wavenumbers in decreasing order of their contribution to total variance (R^2) to remove the ones that contribute the least. For the Bayesian learning approach, the collective intensity of a particular group was calculated as the square root of the sum of squared intensities corresponding to the wavenumbers in that group.

6.3.5 Performance indices for SMCR models

Two measures of performance based on residuals are used to evaluate the performance of SMCR in two places: (i) choosing the number of components (section 6.4.1) and (ii) final

resolution of C and S profiles (section 6.4.3) in this work. These are the lack of fit (LOF) and total explained variance (R^2), both expressed as a percentage. LOF is calculated by dividing the sum of squared error (SSE) by the total sum of squares (SSM) and taking the square root of the resulting value as given in equation 6.24. In other words, LOF is a measure of the unexplained variance in the respective model. The R^2 is calculated by subtracting the result of division of SSE and SSM from 1 as given in equation 6.25. Similar statistical indicators were also used in Chapter 3 and Chapter 4 to evaluate performance of various regression models employed.

$$SSE = \sum r_i^2 \qquad \text{Equation 6.22}$$

$$SSM = \sum D_i^2 \qquad \text{Equation 6.23}$$

$$LOF = 100 * \sqrt{\frac{SSE}{SSM}} \qquad \text{Equation 6.24}$$

$$R^2 = 100\left(1 - \left(\frac{SSE}{SSM}\right)\right) \qquad \text{Equation 6.25}$$

where r_i is each element of the residual matrix calculated by subtracting the factor-reproduced matrix from the original data matrix, D.

6.4 Results and Discussion

For the purpose of complete comprehension of the results in this work, it was considered useful to accompany the discussion alongside the results. The results of SMCR adopting both the ALS and the ALS-PSO-'*fmincon*' algorithms along with certain quantitative spectra-derived parameters to help with the chemical interpretation for the local models are given in the first four parts of this section. The allocated wavenumbers to the respective BHC clusters along with their possible chemical association are discussed in the fifth part of this section. Following this, the three Bayesian network structures learned using the clustered data and the postulation of plausible reaction pathways for the thermal conversion process of Athabasca bitumen over the temperature range of 300 – 420 ºC is presented. Finally, the results from a global SMCR model on the FTIR data at all temperatures considered together using ALS-optimization is given and the possible implications to partial upgrading of bitumen are discussed.

6.4.1 Rank of each sub-matrix

Out of the unknown number of chemical constituents present in bitumen, it was essential to identify a fewer number of components that were representative of the changes undergone by bitumen during thermal conversion. This assisted in building the SMCR model as well as subsequent chemical analysis. As elaborated in section 6.3.3.4, ROD and SD were the indicators used for this purpose. Figure 6.4 depicts the ROD, SD and the residual after performing SVD with the optimum number of components on the data obtained at 400 ºC. The plot of cumulative contribution of each component to the overall variance is also shown. These plots at other temperatures are provided in section D.4 in Appendix D.

The *x*-axes in Figure 6.4a, Figure 6.4b and Figure 6.4c have 17 components/eigenvalues emerging from 16 experimentally obtained datapoints plus one for the feed. It can be seen that ROD has a maximum value of 19.99 when 3 components are used. This is consistent with the maximum for the second derivative of the IND function as shown in Figure 4b. Further support for a 3-component SMCR model comes from the residual calculated by subtracting the original data matrix from the SVD-reproduced data with 3 components in consideration. The maximum positive value of the residual is ~0.01 and the maximum negative value is -0.013. Figure D.6a and Figure D.6b in Appendix D give the residual plots after conducting SVD with 2 and 4 components, respectively, for the 400 ºC dataset. It can be seen that the residuals calculated using 2 components were higher than those using 3 components, but the residual spectra when a higher than optimal number of components were considered had lower values (Figure D.6b). This could be because the total variance explained (R^2) by 4 components (99.77 %) was higher than that explained by 3 components (99.60 %) as depicted in the eigenvalue plot in Figure 6.4d. The LOF (defined in section 6.3.5), also decreased from 9.57 to 4.78 when 2 to 4 components were employed for the 400 ºC dataset. However, the LOF was 6.37 when the optimal number of components were employed. Values for the LOF and R^2 for other datasets are given in Table D.2 in Appendix D.

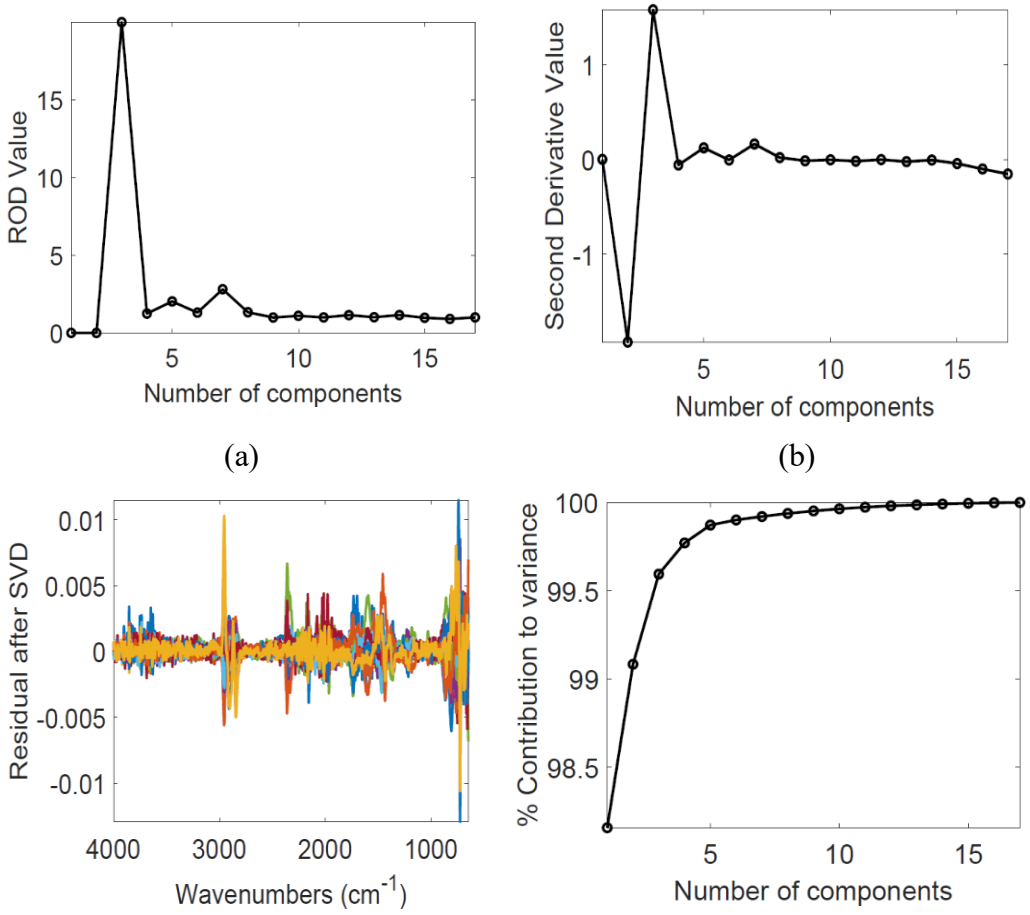


Figure 6.4. Plots for: (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 400 ºC.

For our work, bitumen could be represented by 3 components, and from a chemistry perspective, they are better referred to as pseudo-components since their spectra may not necessarily correspond to a real component in the bitumen mixture. Figure D.7, Figure D.8, Figure D.9 and Figure D.10 in Appendix D show that ROD is maximum when 3 components are chosen for the SMCR models at other temperatures of 300 ºC, 350 ºC, 380 ºC and 420 ºC as well. In all cases, more than 99.5 % of the total variance is explained by 3 components, and thus 3 components fairly represent the original data. However, for the 300 ºC dataset, only 2 data points were available

from the experiment but ROD and SD displayed the highest value as compared to 1 or 2 components. The residuals after performing SVD (Figure D.7c) dataset were 10 orders of magnitude lower than that for other temperatures with 3 components contributing to 99.5 % of the explained variance.

It is to be noted that it would have been difficult to choose between 3, 4 or 5 components from the scree plot (Figure 6.4d) alone since the inclusion of more than 2 components could explain > 99 % of the total variance. It is general practice that the number of optimal factors is chosen at the point where the scree plot flattens or levels out, which would be 2, 6, 5, 5 and 2 for 300 ºC, 350 ºC, 380 ºC, 400 ºC and 420 ºC respectively. But none of these correspond to the optimal number as determined according to the ROD and SD indicators and choosing more than the required number of factors will mean the inclusion of secondary eigenvalues that represent only noise and will unnecessarily increase the computational time of the rest of the algorithm. On the other hand, utilizing a smaller than optimal dimension of the factor space leads to larger residual values, thereby indicating that there is scope for a more accurate reproduction of the original data matrix. Hence, it was decided to proceed with 3 pseudo-components for the rest of the curve resolution process.

It was found that the thermal conversion of Cold Lake bitumen over the temperature range of 150 – 300 ºC could be represented by 3 pseudo-components as well. [29] The splitting of the dataset was done such that the first pseudo-component is likely to represent the feed in all cases. The concentration of the feed was expected to decrease with time on thermal conversion but either the second or third component could represent the final converted products in the reaction mixture depending on the extracted concentration profile (section 6.4.3). Though each pseudo-component cannot be assigned a specific molecular structure, their spectra can provide valuable information about major chemical changes occurring during thermal conversion and the flow of the reaction is determined by the changes in concentration.

6.4.2 Initial concentration estimates

As detailed in section 6.3.3.5, FSMW-EFA was used to obtain the initial estimates for the concentration profiles at each temperature. Figure 6.5 shows the initial concentration estimates obtained through this method at each temperature except 300 ºC, which is shown in Figure D.11

in Appendix D. The profiles depict the conversion of one pseudo-component to another quite clearly. The concentration of the first pseudo-component (S_1 – black line in Figure 6.5 and Figure D.11) appeared to decrease gradually at all temperatures and vanished at reaction times of 480 min, 1170 min, 1320 min and 240 min at 420 ºC, 400 ºC, 380 ºC and 350 ºC respectively. This corresponds most likely to the feed since its concentration is expected to decrease with the progress in thermal conversion.

The concentration of the second pseudo-component (S_2) seemed to follow the path of a downward parabola, i.e. it rose up in the regions where the feed concentration decreased, reached a maximum and then declined at higher reaction times. S_2 started appearing at the first instance where the feed concentration started to decrease, while the third pseudo-component (S_3) came into existence at a later reaction time at all temperatures (Figure 6.5a, Figure 6.5b, Figure 6.5c, Figure 6.5d). Since there were only 3 points at 300 ºC, S_1 disappeared at the intermediate reaction time where only S_2 existed and S_3 remained at the final reaction time. The small number of points for 300 ºC is a limitation and further interpretation of the resolved profiles for this temperature is provided in the next section.

These observations suggest that S_2 is an intermediate product and S_3 would be the final product at all temperatures as indicated by the initial estimates. Any further interpretation based on these initial concentration profiles should be done with caution and we move on to the optimized spectral profiles for the results to make chemical sense. The final ALS- and PSO-'*fmincon*'-optimized concentration profiles can be verified against the initial estimates so as to get an idea of the accuracy of the EFA method.

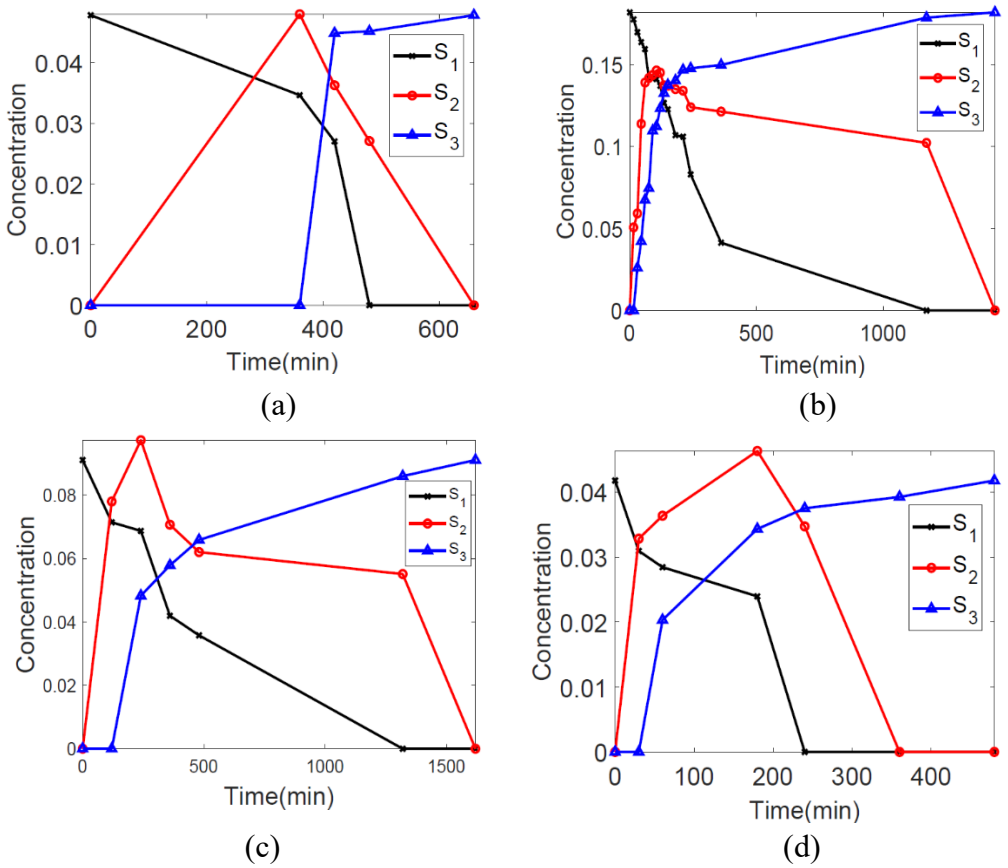


Figure 6.5. Plots of initial estimates of change in concentration of the three pseudo-components with process flow (reaction time in min) at the following temperatures: (a) 420 ºC; (b) 400 ºC; (c) 380 ºC; (d) 350 ºC.

6.4.3 ALS-optimized *C*, *S* profiles and spectra-derived quantitative parameters

In this section, the concentration and spectral profiles of the three pseudo-components obtained from curve resolution using ALS optimization for each temperature are provided and the implications on reaction chemistry are discussed. Some of the trends are also compared with those obtained for thermal conversion of Cold Lake bitumen and the similarity or difference in the possible classes of reactions occurring are explored.

6.4.3.1 Results and analysis at 300 ºC

Figure 6.6 shows the ALS-optimized concentration and spectral profiles for the pseudo-components when the reaction was conducted at 300 ºC. The spectra between 3200 – 650 cm^{-1} are shown as split into 4 regions for easier visualization. The residual plot when the original matrix was subtracted from the product of the profiles obtained from ALS routine is shown in Figure D.12a of Appendix D.

The pattern of the concentration profiles (Figure 6.6a) can be understood from the fact that the 300 ºC dataset consisted of only 3 data points including the feed. The initial estimates (Figure D.11 in Appendix D) that were not normalized, had a similar pattern and the conversion among the pseudo-components appeared to follow the path $S_1 \rightarrow S_2 \rightarrow S_3$. The importance of bringing the concentrations to the same scale before implementing the ALS can be seen from Figure D.11. The concentration of S_2 is much higher than the other two components and when fed as is to the ALS routine, can affect the results. Due to the limited amount of data, there is no information available on the exact concentration values between 0 – 360 min and between 360 – 1080 min but the trend can be seen. S_1, which most likely represents the feed, does not exist at higher reaction times, where only S_3 remains, while the intermediate product (S_2) is present at intermediate times.

Visual inspection of the resolved spectra can help in identifying certain striking features in line shape like band-splitting at a particular wavenumber or broadening and can also identify the presence or absence of absorptions characteristic to some chemical species. We considered four quantitative parameters defined in terms of intensity ratios from the resolved spectra to improve the ability for understanding crucial steps in thermal cracking at each temperature. Ratios of intensity rather than absolute intensities are examined so as to negate the effect of path length that is a source of uncertainty in the ATR attachment to the FTIR spectrometer. These parameters are delineated as follows:

$$n\text{-}CH_2/n\text{-}CH_3 = \frac{I_{2920}}{I_{2950}} \qquad \text{Equation 6.26}$$

$$\text{Overall extent of aromatic substitution (EOS)} = \frac{(I_{744}+I_{763}+I_{810}+I_{860})}{I_{727}} \qquad \text{Equation 6.27}$$

$$\text{Degree of condensation (DOC)} = \frac{A_{1630-1550}}{A_{3130-3050}} \qquad \text{Equation 6.28}$$

$$\text{Complex oxygenate content (COC)} = \frac{I_{1740}}{I_{1740}+I_{1605}} \qquad \text{Equation 6.29}$$

where I is the intensity at the respective wavenumber shown in subscript and A is the area under the regions in wavenumber range specified as the subscript.

The n-CH_2/n-CH_3 parameter, defined in terms of the intensity ratios of the dominant asymmetric stretch of methylene C-H to the stretch of terminal methyl C-H groups, serves as an indicator of the average aliphatic chain length in bitumen. Naphthenic rings attached to aromatics exist in large proportions in Athabasca bitumen and a larger value of this parameter (equation 6.26) can also indicate the presence of non-aromatic cyclic rings. [109] DOC was incorporated in a modified form from the work by Tefera et al. [29] where they had defined it as the area under the C=C stretch of aromatics divided by the area under the entire C-H deformation intensities from 900 – 700 cm^{-1}. In our work, the denominator was changed to include the area under the aromatic C-H stretching wavenumbers only (equation 28). This was because the chance of overlap of C-H bending between the alkenes and di- and tri-substituted aromatics in 900 – 700 cm^{-1} region was much more than in the 3130 – 3050 cm^{-1} region. [110] However, the limitation of DOC is that the decrease in hydrogens in the aromatic ring is not taken into account when non-aromatic substituents are present. Other disadvantages of DOC are discussed in section 6.4.3.4.

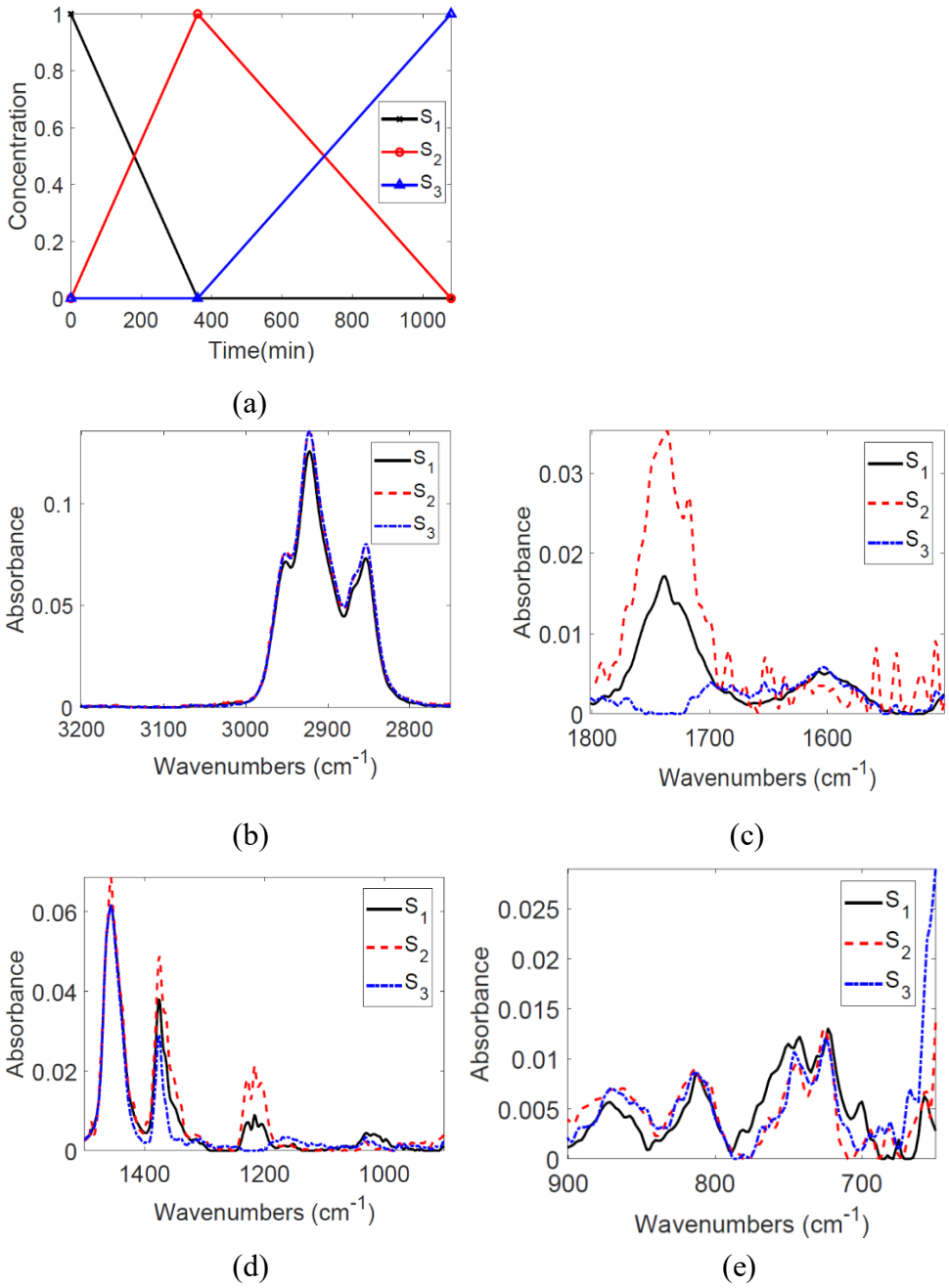


Figure 6.6. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 300 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

To tackle these shortcomings, EOS was introduced in this work (equation 6.27) and was calculated as the ratio of intensities rather than areas (which was the case for DOC) to limit the possibility of overlap with alkenes. The sum of intensities corresponding to ortho, meta and para-disubstituted aromatics (744, 763 and 810 cm^{-1}, respectively) and tri-substituted (~860 cm^{-1}) aromatics were divided by the intensity at 727 cm^{-1} that originated from a mono-substituted aromatic with 5 adjacent hydrogens. In effect, EOS considers the number of adjacent hydrogens in an aromatic compound and could account for both condensed aromatics as well as the presence of acyclic and non-aromatic substituents and was considered in combination with DOC for interpretation of the resolved spectra in this work.

COC focused on the C=O stretching frequencies for the ester and anhydride-type carbonyl compounds (1740 cm^{-1}) and their change during thermal conversion was examined by calculating this parameter for each pseudo-component (equation 6.29). However, this must be interpreted with caution as the exact process of conversion of complex oxygenates to acids, alcohols and further decarboxylation from the acids to yield CO_2 was complicated and still unclear. [111] Also, since phenolic compounds are more prevalent than aliphatic alcohols in bitumen, it is possible that the anhydride content is more than the ester content. [112] This could be because the interaction of carboxylic acids with themselves leading to anhydrides would be more probable than phenols interacting with carboxylic acids yielding esters.

The change in the above-mentioned quantitative parameters across S_1, S_2 and S_3 is summarized in Table 6.2.

Table 6.2. Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 300 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.78	1.79	1.82
Overall EOS	3.15	3.26	2.90
DOC (C=C stretch/C-H stretch wavenumber)	2.77	2.43	2.81
COC value	0.77	0.91	0.00

It can be seen from Table 6.2 that thermal conversion did not alter the ratio of methylene groups to terminal methyl group intensity by much. S_1 and S_2 had a near-constant value of 1.78 and it mildly increased to 1.82 for S_3, which was quite intriguing. This suggests that there was neither a major change in the length of alkyl side chains attached to cyclic moieties nor there was much formation of additional cycloalkanes due to thermal cracking at 300 ºC. This also reflects the minimal change in the spectral intensities and line shapes at 2850 cm^{-1}, 2920 cm^{-1}, 2950 cm^{-1} in Figure 6.6b.

Cronauer et al. [113] suggested that hydrogen disproportionation from potential donors like naphthene-aromatics or benzylic carbon centres to multinuclear aromatics (MNA) would be difficult at temperatures lower than 300 ºC. Hydrogen disproportionation is one form of hydrogen transfer that does not necessarily form a free radical on the parent carbon centre. [114] At the same time, there have been studies in the literature that reported the occurrence of hydrogen transfer to MNAs at even lower temperatures of 150 ºC. [115,116] This is also supported by lower values for energy requirements for hydrogen transfer (~35 kJ/mol), which suggest that the occurrence of this event does not demand a high temperature. [114]

Tefera et al. [29] observed that the results and chemical interpretation for the SMCR-ALS resolved spectra for pyrolysis of Cold Lake bitumen at 300 ºC were similar to that at 150 ºC. nCH_2/nCH_3 was reported to increase from 1.82 to 1.89 from the first to the third pseudo-component, which was a significant increase compared to that observed for Athabasca (Table 6.2). They suggested the possibility of hydrogen transfer from a benzylic carbon that later combined with another free radical to yield a longer side chain. Methyl transfer was also thought to occur, which increases the $-CH_2$ content. The assumption here was that hydrogen transfer reactions occur in bitumen at temperatures as low as 150 ºC, despite contrasting views in the literature as highlighted in the previous paragraph. At 150 ºC, this could explain an increase in viscosity for the liquid products (from 88 Pa.s to 240 Pa.s) due to a gain in molecular weight over time. [31] For complex mixtures, the change in viscosity and density cannot be directly related to the molecular weight when the composition varies. The increase in viscosity and decrease in density was explained through an 'excluded volume effect' that restricted the flow of the formed heavier molecules. It was interesting to note that the viscosity trend followed an opposite pattern (decreased overall from 88 Pa.s to 1 Pa.s over 8 h with non-constant values at 4 h) at 300 ºC despite spectral features in the ALS-resolved profiles for Cold Lake bitumen remaining similar to those at

150 ºC. Though boiling point distributions for the products are not known, the density of most products obtained at 300 ºC was higher than that of the feed. These observations indicate that reasons for viscosity change cannot be attributed to a single reason in complex mixtures as highlighted by Sivaramakrishnan et al. [24] as well.

Shifting our attention back to Athabasca bitumen in our work, the viscosity of the pyrolyzed products at 300 ºC showed a sparse change over 1080 min and gas was hardly produced during the reaction. While the peak at 1460 cm^{-1} corresponds to C-H bending in methyl groups only, the peak at 1380 cm^{-1} can indicate both methyl and methylene group bends. [117] Since the splitting of the bands was more visible at 1380 cm^{-1}, it is considered for examining migration of methyl groups at various conditions in this work. The C-H bending bands for methyl groups occurring at 1380 cm^{-1} (Figure 6.6d) showed clear signs of splitting for S_1, S_2 but appeared to be flat for S_3, indicating the possibility of some methyl transfer. From the perspective of energy demand, methyl transfer from ethane requires the same energy as hydrogen abstraction from the benzylic carbon in toluene, which is quite probable. [118] This suggests that some methyl transfer may have occurred to an extent that led to an increase in the $-CH_2$ content and was reflected in the mild increase of nCH_2/nCH_3 (Table 6.2). This process is illustrated in Figure 6.7.

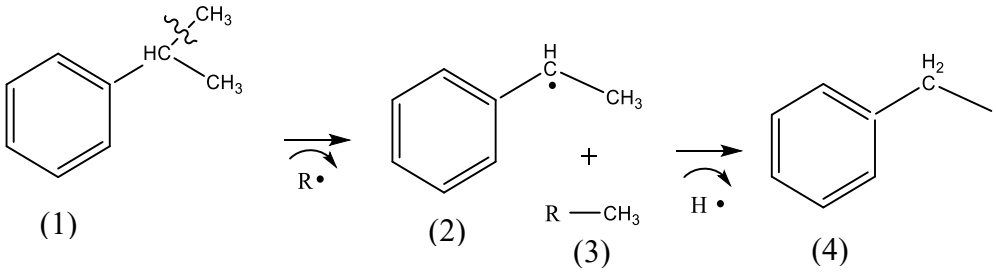


Figure 6.7. Methyl transfer from an isopropyl group attached to an aromatic (1) followed by hydrogen abstraction from the matrix leading to increased CH_2 content (compound (4)) possibly occurring at 300 ºC.

DOC had similar trends for both Athabasca and Cold Lake where it showed a very slight effective increase from 2.77 to 2.81 (S_1 to S_3) for Athabasca (Table 6.3) but a higher net increase from 0.18 to 0.26 (S_1 to S_3) with a minimum at S_2 for both types of bitumen. However, the EOS showed an overall mild decrease from 3.15 to 2.90 from S_1 to S_3 reaching a maximum at S_2 for Athabasca (Table 6.2) but it increased from 2.26 to 2.56 for the pseudo-components in Cold Lake.

The EOS was calculated from the resolved spectra for Cold Lake bitumen for this work, and was not calculated by Tefera et al. [29] The increase in EOS for Cold Lake spectra could have been a result of increase in DOC due to intramolecular ring closure reactions as suggested by Tefera et al. [29] On the other hand, the slight decrease in EOS despite the near-stable DOC for Athabasca was thought-provoking since both were related to aromatic part of the sample. It should be kept in mind that DOC was calculated differently in the two works though the intended meaning was the same. In addition, Figure 6.6e shows that there was no significant change in the intensity at 727 cm^{-1} that indicated significant formation of mono-substituted aromatics in S_3.

On the whole, considering the minimal changes in DOC and nCH_2/nCH_3 parameters and the little to no gas production, it can be concluded that the extent of cracking at 300 ºC was indeed low. This could also be a reason for the reduced effect on viscosity at 300 ºC (which stayed nearly constant at the feed viscosity of 36 Pa.s even after 1080 min reaction time), though the results for viscosity were not shown more elaborately in this work. Selucky et al. [23] reported that the amount of saturates in Cold Lake bitumen is more than in Athabasca bitumen, making it more susceptible to cracking at lower temperature. This is probably another reason for the chemical changes being more significant in Cold Lake bitumen than in Athabasca bitumen when subjected to pyrolysis at 300 ºC.

6.4.3.2 Results and analysis at 350 ºC

The ALS-resolved concentration and spectral profiles for the FTIR dataset at 350 ºC thermal conversion is given in Figure 6.8. The residual plot when the ALS-reproduced profiles are subtracted from the original matrix (D) is given in Figure D.12b in Appendix D.

The concentration profiles of the pseudo-components (Figure 6.8a) suggest that the reaction pathway is not as straightforward as was the case with 300 ºC. Both S_2 and S_3 appeared to exist at large reaction times while the concentration of S_1 decreased but did not vanish until 360 min. As a contrast to the initial concentration estimates (Figure 6.5d) where S_2 rose before S_3, the third pseudo-component appeared before the second in the final resolved concentration profile. The concentration of S_3 briefly dropped below S_2 at 180 min but was equal to or more than S_2 at all other reaction times. Due to the existence of both S_2 and S_3 at larger times and the appearance of S_3 from the moment S_1 started decreasing, the reaction network can be considered as $S_1 \rightarrow S_3$

directly with S_2 as an intermediate product existing in lower concentration. For the purpose of analysis of chemical changes through the resolved spectra, the values of the derived parameters for S_1 as the feed and S_3 as the final product was considered as a fair assumption.

The spectra-derived quantitative parameters for the pseudo-components at 350 ºC dataset are compiled in Table 6.3.

Table 6.3. Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 350 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.78	1.77	1.74
Overall EOS	3.14	3.45	1.30
DOC (C=C stretch/C-H stretch wavenumber)	2.79	3.59	2.22
COC value	0.77	0.00	0.51

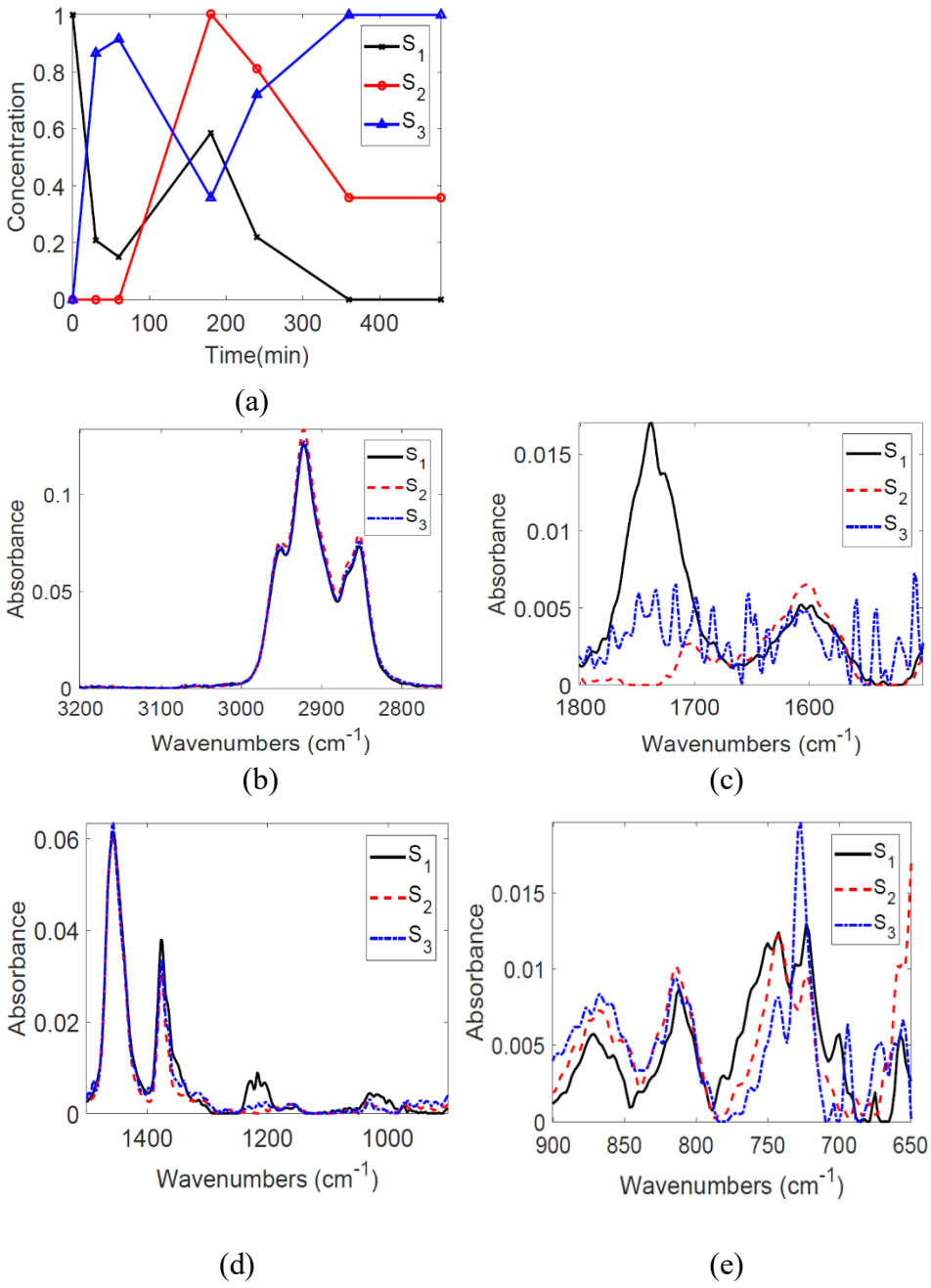


Figure 6.8. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 350 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

It was interesting to note that the ALS-resolved concentration profiles for that of Athabasca bitumen (Figure 6.8a) were similar to that of Cold Lake at 340 ºC [29] where the third pseudo-component started rising from the start when the first component began to decrease and both the second and third pseudo-component existed at larger reaction times with S_3 being the dominant one. The reaction network, though difficult to ascertain for thermal cracking at 340 ºC for Cold Lake bitumen, was considered to follow the path $S_1 \rightarrow S_3$ with S_2 being existent in lower concentration.

The amount of gas produced during reaction at 350 ºC for Athabasca bitumen was ~3-4 %wt. after 480 min and this was more than that produced at 300 ºC (section 6.4.3.1). A mild but continuous decrease in $n\mathrm{CH_2}/n\mathrm{CH_3}$ (Table 6.3) possibly meant that there was a reduction of the length of aliphatic chains attached to cyclic structures during cracking. At the same time, it could also mean a decrease in $-\mathrm{CH_2}$ content by the conversion of naphthenes to aromatics through hydrogen-disproportionation. However, the extents of these changes was still low and were reflected in the minimal change in intensity of methyl and methylene absorption bands at 2950 cm^{-1} and 2920 cm^{-1} respectively in Figure 8b. The release of some gas indicated that cracking was happening to a certain extent.

At the same time, the comparison with the results obtained after ALS-resolution of the spectra for Cold Lake bitumen at 340 ºC (as reactions were not performed at 350 ºC in the work by Tefera et al. [29]) was interesting. They observed an increase in $n\mathrm{CH_2}/n\mathrm{CH_3}$ similar to the trends at lower temperatures and attributed it to the presence of hydrogen transfer to multinuclear aromatics as the major reason. This was opposite to the trend observed with Athabasca bitumen as just stated above. In addition, the C-H deformation band at 1380 cm^{-1} was split for S_1 and S_2 but not for S_3 (Figure 8d). This potentially indicated that carbons having more than one $-\mathrm{CH_3}$ attached to them initially have only one methyl group at longer reaction times. This could occur through methyl transfer, as was indicated at 300 ºC, or by the formation of methyl radicals eventually leading to the production of methane gas by free-radical combination with hydrogen radicals. In this work, methane was found to be the dominant gaseous product at all reaction times at 350 ºC as determined from GC-FID. This was also supported by the work of Jha et al. [119] where they found that methane was the major product in the gaseous phase when Athabasca bitumen was thermally reacted at 300 ºC. On the other hand, if methyl transfer alone took place, it would have led to an increase in the $\mathrm{CH_2}$ content as noticed in the case of Cold Lake bitumen at 340 ºC and 360 ºC, while the opposite was

observed with Athabasca bitumen with a decrease in the nCH$_2$/nCH$_3$ parameter at 350 ºC (Table 6.3). [29] On the whole, it can be said that methane formation happened to a larger degree than methyl transfer at 350 ºC during thermal conversion of Athabasca bitumen.

On the other hand, significant changes were observed for DOC and EOS as is seen from their values for the three pseudo-components. DOC decreased from 2.79 to 2.22 from S_1 to S_3 with a maximum at S_2 (Table 6.3) while the opposite was seen to happen with Cold Lake bitumen at 340 ºC, where it increased gradually from 0.18 to 0.39. [29] Despite this increase, they suggested that ring closure reactions through intra-aromatic coupling were suppressed. [119] In order to make a meaningful interpretation of the DOC, it was important to view it together with EOS values. The EOS for thermal cracking of Athabasca bitumen at 350 ºC decreased sharply from 3.14 to 1.30 (S_1 to S_3) while that of Cold Lake bitumen at 340 ºC rose from 2.27 to 2.63 over the three pseudo components. This is reflected by the higher intensity for mono-substituted aromatic absorption at 727 cm^{-1} for S_3 as compared to the feed and S_2 (Figure 8e). The increase in EOS for Cold Lake bitumen shows that there was an increase in the number of substituents in the aromatic rings that could be due to condensation of a naphthene-aromatic to a complete aromatic with a non-cyclic substituent already on the naphthene ring. The ring closure reactions cannot be ignored as well due to the lower temperature.

Considering the variation in all the parameters at 350 ºC for Athabasca bitumen, a plausible reaction mechanism is illustrated in Figure 6.9.

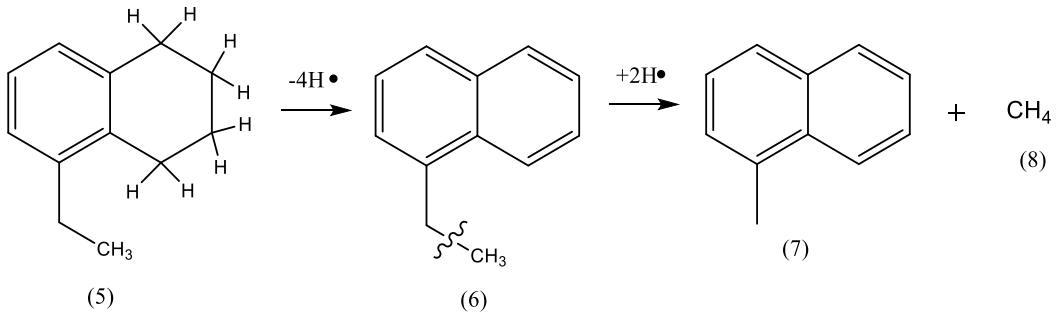


Figure 6.9. Sequence of reactions speculated to be occurring at 350 ºC based on SMCR results.

When compound (5) is converted to (6) through the loss of hydrogen, the -CH$_2$ content is clearly lowered. The ethyl substituent can crack further, leading to benzylic and methyl free radicals that can abstract hydrogens from the matrix and form compound (7) and release methane gas. This accounts for the decrease in the nCH$_2$/nCH$_3$ parameter (Table 6.3) and a possible pathway

to produce methane that was seen to be dominant in the gaseous products. Another interesting observation is that compound (5) is a tri-substituted aromatic while compound (6) is *o*-di-substituted with respect to the newly formed aromatic ring. This does not directly imply a decrease in the EOS with the way it was calculated (equation 6.27). To verify this, another parameter that represents the ratio of tri-substituted, meta-, para-disubstituted aromatic C-H bends at 810, 860 cm^{-1} and that of *o*-disubstituted aromatic C-H bends at 744, 763 cm^{-1} was calculated from the resolved spectra. It was interesting to see that this parameter reduced from 1.53 to 0.93 from S_1 to S_3, which could imply an increased formation of o-disubstituted aromatics as compared to *m*-, *p*- and tri-substituted aromatics. This is explained by Figure 6.9 where the tri-substituted aromatic which is also *m*- substituted is not altered but leads to the formation of an *o*-disubstituted aromatic (compound (6)), while not much is known about mono-substituted aromatics. Thus, this mechanism fairly represents the changes in the quantitative parameters at 350 ºC.

The question remains whether the decrease in EOS for Athabasca bitumen at 350 ºC can point towards the breakage of the relatively strong Ar-C-alkyl C bond, since this is a way to produce mono-substituted aromatics. The dissociation energy of this bond in toluene is 433 kJ/mol at 25 ºC, which is equivalent to abstracting a hydrogen from methane to produce methyl radical, which is quite difficult under these conditions. [118] However, there is evidence for scission of the stronger Ar-C-H bond in benzene when it was used to study the cracking of nC_{16} paraffin in the temperature range of 398 – 450 ºC at pressures of ~13 MPa. [121] Though cracking of the side chain was postulated to occur (Figure 6.9), the scission of the aryl C-alkyl C did not seem feasible at 350 ºC. Figure 6.10 depicts a reaction chemistry that can lead to a decrease in EOS without the cleavage of the aryl C-alkyl C bond and keeping the *o*-disubstituted aromatic (compound (9)) content constant, with the formation of mono-substituted aromatics. This is also reflected in the resolved profile in the 900 – 650 cm^{-1} region where there was a significant increase in the intensity at 727 cm^{-1} for S_3 as compared to 810 cm^{-1} and 860 cm^{-1}. The processes involved in Figure 6.10 are homolytic bond scission to give compound (10), hydrogen transfer from the matrix to produce compound (11) from which the hydrogen is abstracted to yield a tertiary free radical and subsequent hydrogen disproportionation to yield the mono-substituted aromatic (14).

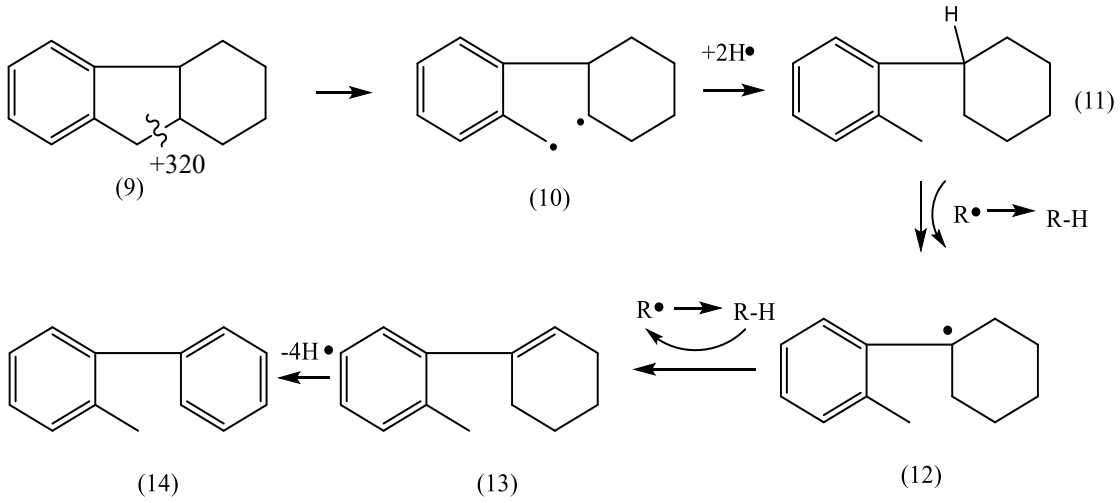


Figure 6.10. Pathway for 350 ºC showing the increase in mono-substituted aromatic content from a naphthene, keeping the di-substituted content constant. The bond dissociation energy (BDE) for homolytic cleavage of the indicated bonds is also shown in kJ/mol.

Little weightage should be given to the COC calculation as the resolved spectrum of S_3 was noisy (Figure 6.8c) and not representative of the spectra arising from real compound classes. The intensity at 1740 cm^{-1} appeared to decrease from S_1 to S_2 signifying conversion of ester-type and anhydride-type compounds, but no interpretation can be assigned to S_3.

A review of the trends in viscosities of Athabasca and Cold Lake bitumen during thermal conversion at 350 ºC and 340 ºC respectively, gave interesting insights. The viscosity of Athabasca bitumen experienced a slight decrease from 36 Pa.s to 15 Pa.s over 480 min reaction time as determined from a separate analysis in our work. This decline in viscosity was monotonic and gradual. A shear rate of 10 s^{-1} was employed for all samples at 350 ºC as this was above the lower device limit for torque in the viscometer and a higher shear rate was not required due to the lower fluidity of the samples. The dependence of the viscosity of cracked samples from Athabasca bitumen on shear rates was explored by Sivaramakrishnan et al. [24] Wang et al. [33] reported the viscosity of pyrolyzed liquid products at 340 ºC to decrease from 88 Pa.s (feed) to ~2.3 Pa.s at 240 min, but it had a local maximum at 480 min (5.1 Pa.s). It was interesting to see the overall decrease in viscosity in both cases even when there were differences in reaction mechanisms as discussed previously. There was not much information available regarding the composition of the products at 340 ºC and 350 ºC for both types of bitumen, which made attribution of a particular reason to

the observed viscosity change difficult. However, at 300 ºC, the lower refractive index of the products from Cold Lake bitumen corroborated with their increased non-aromatic content in the products (as determined by ^{1}H-NMR) compared to the feed. This increase in lighter material content may have caused the viscosity to decrease as was observed by Yanez and De Klerk. [31] Also, while the density of the thermally converted products of Cold Lake bitumen at 300 ºC were ~6 – 12 kg/m^3 higher than that of the feed, [31] the density was barely affected during cracking of the same feed at 340 ºC. [33]

Since properties other than viscosity were not measured for the products at 350 ºC, only speculative comments could be made from the chemometric results and proposed reaction mechanism about the reasons for the observed decrease in viscosity. If the alkyl substituent in compound (5) in Figure 6.9 was longer, cracking could lead to lighter aliphatic products that stay in the liquid phase and contribute to a decrease in viscosity. Based on the above discussion, it can be considered that the extent of cracking was higher at 350 ºC than at 300 ºC for Athabasca bitumen and liquid phase hydrogen transfer plays a major role.

6.4.3.3 Results and analysis at 380 ºC

The final resolved concentration and spectral profiles for the FTIR spectra of the liquid products at 380 ºC are given in Figure 6.11. The residual plot when the matrix reproduced from the ALS-optimized profiles was subtracted from the original data matrix is given in Figure D.12c in Appendix D.

The initial estimates were good starting points for the dataset at this temperature (Figure 6.5c) as the final concentration profiles seemed to follow a similar trend (Figure 6.11a). S_1 decreased continuously and disappeared at 1320 min while S_2 emerged at 0 min and exhibited two local maxima at 240 min and 1320 min. S_3 became non-zero at 240 min and remained at a higher concentration than S_2 at all later times except at 1320 min. Based on this, $S_1 \rightarrow S_2 \rightarrow S_3$ can be regarded as the reaction pathway for the conversion of Athabasca bitumen at 380 ºC from the nature of the concentration profiles of the pseudo-components. Table 6.4 provides the values of the derived quantitative parameters from the resolved spectra at 380 ºC.

Table 6.4. Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 380 ºC.

Pseudo-component	1	2	3
n-CH_2/*n*-CH_3	1.80	1.28	1.50
Overall EOS	3.15	2.33	3.67
DOC (C=C stretch/C-H stretch wavenumber)	2.80	1.29	1.86
COC value	0.81	0.21	0.00

Significant changes can be observed in all the parameters as compared to the previous temperatures (Table 6.2 and Table 6.3). Overall, *n*CH_2/*n*CH_3 decreased from 1.80 (S_1) to 1.50 (S_3) with a minimum of 1.28 for S_2. The resolved spectra in the range 2750 – 3000 cm^{-1} reflected this trend and were different from the ones at 300 ºC and 350 ºC where there was a minimal change in intensity in this wavenumber range. Gas was released in measurable amounts at 380 ºC with ~8 %wt. produced at 1620 min. These observations indicated the formation of lighter products and that cracking was occurring to a higher extent compared to lower temperatures. However, the increase in *n*CH_2/*n*CH_3 from S_2 to S_3 was intriguing and could be due to free radical recombination at higher reaction times, which was also suggested as a reason for the formation of heavier products in Cold Lake bitumen pyrolysis by Wang et al. [33] This was slightly different from that for Cold Lake bitumen conversion at 360 ºC where the *n*CH_2/*n*CH_3 increased from 1.82 to 1.89 for S_1 to S_2 but then decreased to 1.87 for the third pseudo-component. [29] This could indicate the onset of cracking at larger reaction times at 360 ºC with the major types of possible reactions being C-C bond scission in alkyl substituents to yield benzyl radicals that abstract a hydrogen from the matrix to increase the -CH_3 content.

Surprisingly, DOC and EOS showed opposite trends for Athabasca bitumen conversion at 380 ºC in this work (Table 6.4). DOC decreased from 2.79 to 1.86 with a minimum of 1.29 for S_2 while EOS initially decreased to 2.33 from 3.15 and then increased to 3.67 for S_3 that was more than S_1. Hydrogen disproportionation from naphthene aromatics having transferable hydrogen and alkyl side chains could be a reason for the observed changes in DOC, EOS and *n*CH_2/*n*CH_3. This mechanism is illustrated in Figure 6.12.

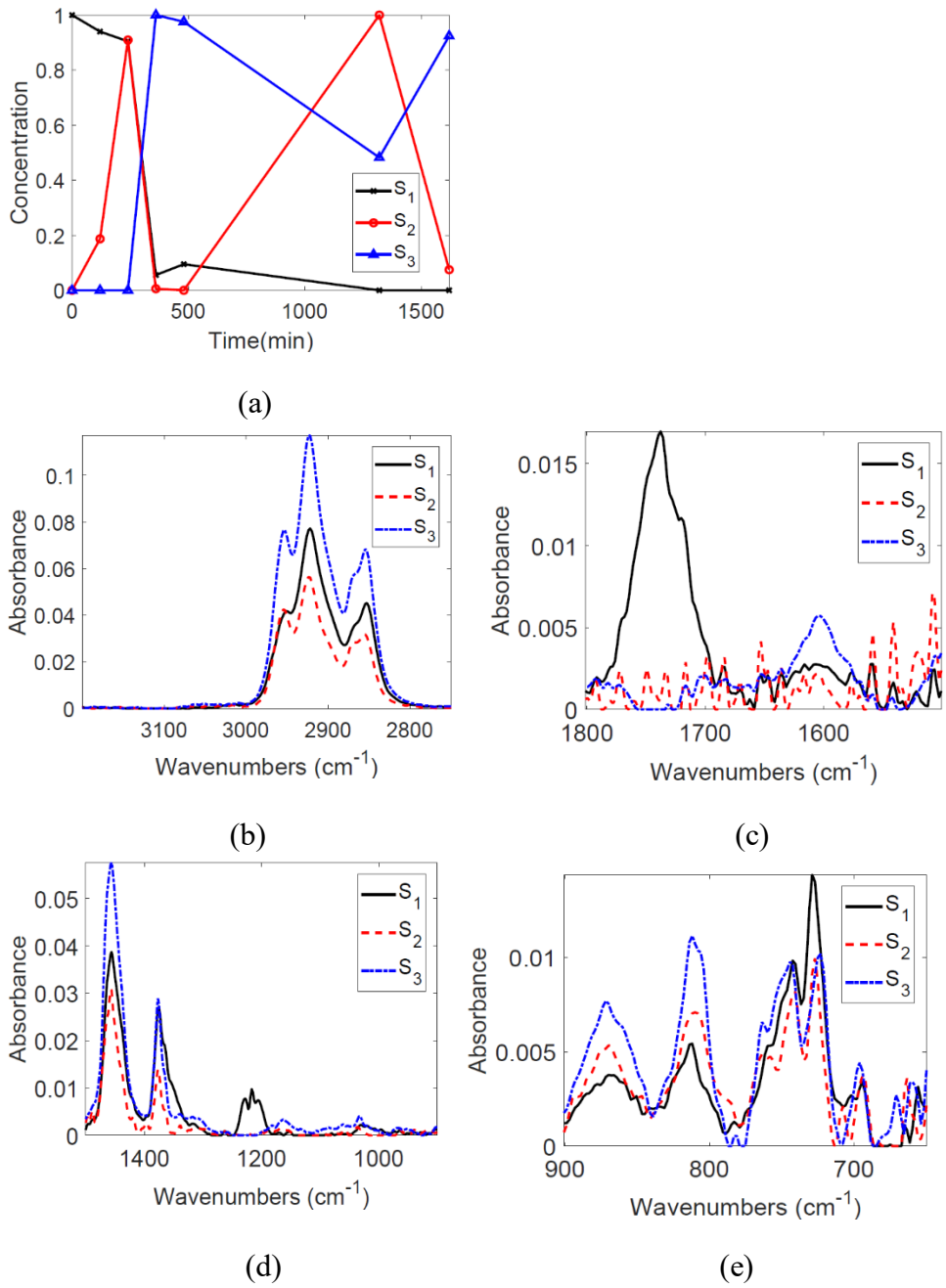


Figure 6.11. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 380 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

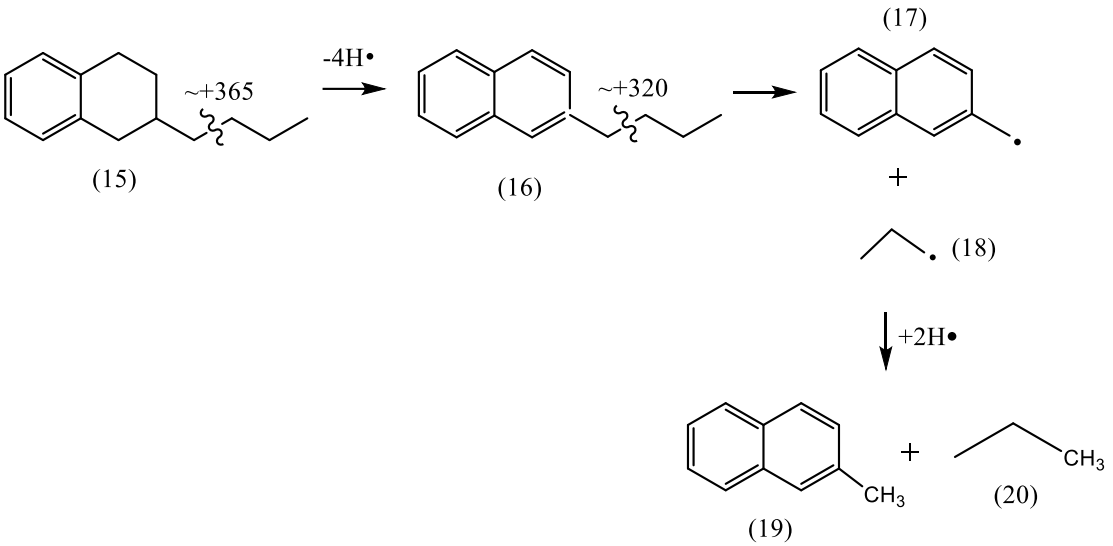


Figure 6.12. Proposed mechanism corresponding to the changes in derived quantitative parameters observed at 380 ºC. The energies for homolytic bond cleavage of the C-C bonds in (15) and (16) are given in kJ/mol.

The naphthenic ring ((15) in Figure 6.12) becomes more stable when it loses 4 hydrogens due to being attached to another aromatic, and involves its π electrons in the delocalization that increases stability. This makes the scission of benzyl C-aliphatic C ((16) in Figure 6.12) less demanding as the BDE for its homolytic cleavage is much smaller than it was for (15) as indicated in Figure 6.12. Once the corresponding free radicals are formed ((17) and (18)), they can stabilize themselves by easily abstracting hydrogens to form the alkane (20) and alkyl aromatic (19). In this entire process, -CH_3 content has increased and chain length decreased that manifests as a decrease in nCH_2/nCH_3 (Table 6.4). Due to the formation of an adjacent aromatic ring to the benzene in ((15) to (19)), the number of aromatic hydrogens in (19) is 3 more than in (15) while the aromatic C=C stretch has only increased by 2. This can potentially reduce the DOC, based on the way it is calculated (equation 6.28).

Also, when compared to (15), the intensities at 810 cm^{-1} and 860 cm^{-1} for (19) would definitely be higher (as observed in Figure 6.11e) due to *m*- and *p*- substitution with respect to the methyl substituent for the second aromatic ring on the right. If the mono-substituted aromatic content is considered constant, this would cause the EOS to increase from the way it is calculated in this work (equation 6.27). Also, it can be seen from Figure 11e that the intensity of ortho substitution did not experience much change at 744 cm^{-1} but the intensity corresponding to the mono-

substituted aromatic C-H bend at 727 cm^{-1} decreased from S_1 and but remained stable for S_2 and S_3. This further could contribute to the increase in EOS and corroborates with the chemistry illustrated in Figure 6.12. However, it should be noted that this mechanism is proposed with model representative compounds based on the trends in the quantitative parameter values and no experimental proof is provided for the exact occurrence of this reaction sequence.

The absorption bands at 1380 cm^{-1} for all three components appeared to be split, indicating that not much methyl transfer had happened. Still, methane gas was found to be dominant in the gaseous products and might have been produced from the cracking of side chains as shown in Figure 6.9 for 350 ºC. On the other hand, DOC did not vary over the three pseudo-components for Cold Lake bitumen conversion at 360 ºC while EOS showed a mild decrease from 2.28 to 2.21 from the first to the third pseudo-component. The dominant reactions that were logically thought to be responsible for these changes were breakage of relatively weaker C-C bonds along with hydrogen transfer to the benzylic free radicals to stabilize the products. At 380 ºC, the viscosity of Athabasca bitumen continuously decreased from 36 Pa.s for the feed to ~2 Pa.s for the product at 1620 min and measured at a shear rate of 10 s^{-1}. Cold Lake bitumen was shown to exhibit a steeper viscosity decrease by Wang et al., [33] where it reached 0.31 Pa.s at 60 min but later increased to 1 Pa.s at 240 min. They attributed this increase to free-radical addition reactions but if the same kind of reactions were responsible for the increase in nCH_2/nCH_3 at 380 ºC for Athabasca bitumen (Table 6.4), viscosity should also have increased at some point; this did not happen. This shows that a combination of different factors like composition, phase behavior, and mainly boiling point distribution of the liquid products would be responsible for a change in viscosity rather than molecular weight alone. [24,122,123]

Though Wang et al. [33] reported that the distillation profiles of the cracked products from Cold Lake bitumen were not altered by much at 360 ºC, other factors might have contributed to the viscosity change. Tefera et al. [29] observed that the second pseudo-component existed at higher concentration at a larger reaction time for thermal cracking of Cold Lake bitumen at 360 ºC. Also, the chain length parameter was maximum for this pseudo-component. If alkyl chain length can be considered an indicator of molecular weight and in turn viscosity at all, this increase for the second pseudo-component could have contributed to the local gain in viscosity at 240 min. Not much information was available regarding the aromatic and non-aromatic contents for the liquid products

at 380 ºC for Athabasca or at 360 ºC for Cold Lake; thus, further interpretation regarding the effect of composition on viscosity was not possible.

6.4.3.4 Results and analysis at 400 ºC

Among all temperatures, the maximum amount of data was recorded for this dataset. Data for different types of characterization like ^{1}H-NMR, ESR, boiling point distribution, viscosity, density, refractive index and asphaltene content for the liquid products obtained from thermal conversion of Athabasca bitumen at 400 ºC was given in Chapter 5 and documented in the study by Sivaramakrishnan et al. [24] Some points from their chapter are useful in supporting the results obtained in this work. The final resolved concentration and spectral profiles for the 400 ºC dataset is provided in Figure 6.13 and the residual plots when the ALS-reproduced matrix is subtracted from the original data matrix is given in Figure D.12d in Appendix D.

The resolved concentration profiles for the three pseudo-components (Figure 6.13a) resembled those from the initial estimates (Figure 6.5b) for most parts though S_1 experienced a steeper decrease and vanished at 360 min. It can also be seen that the concentrations profiles deviated from the closure condition in the intermediate reaction times (135 min, 150 min, 180 min) when the summation reached close to 1.70 but was within 1.20 at the rest of the times (section 6.3.3.7). This could be possibly because the intermediate times mentioned above are where all three pseudo-components exist though S_1 is decreasing and S_3 is increasing while S_2 reaches its peak. S_2 followed the path of an inverted parabola and was higher in concentration than S_3 between the times 90 min and 240 min. Coke started forming in significant amounts from 45 min and the time between 90 – 240 min can be considered as the active coking region for 400 ºC. [24] As expected, S_1 had a much lower concentration during this period but at larger reaction times (> 360 min), S_3 became the dominant product. Overall, $S_1 \rightarrow S_2 \rightarrow S_3$ can be thought of as the reaction network for the cracking of Athabasca bitumen at 400 ºC. Table 6.5 gives the values of the derived quantitative parameters from the ALS-resolved spectra for the 400 ºC dataset.

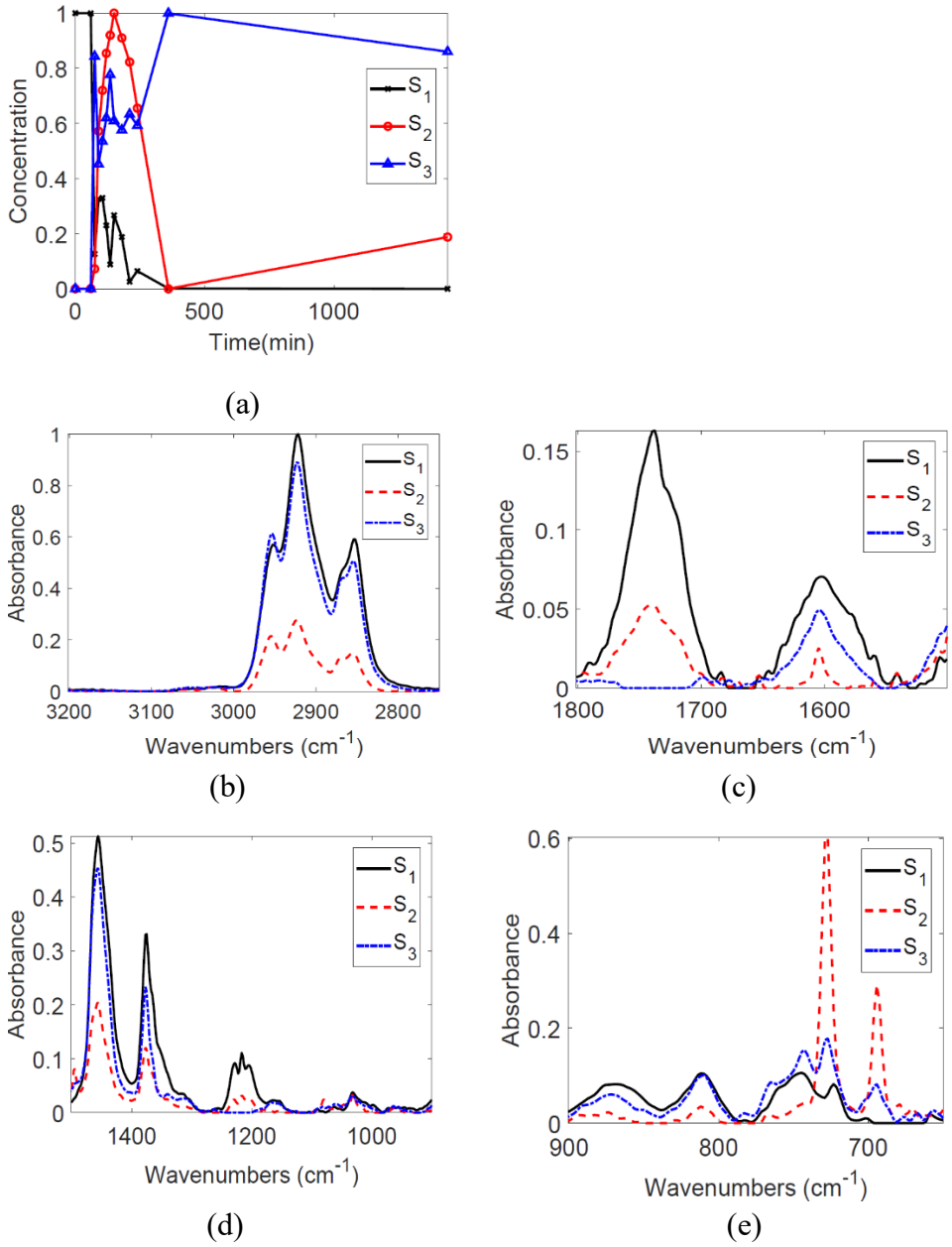


Figure 6.13. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 400 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Table 6.5. Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 400 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.76	1.22	1.41
Overall EOS	3.20	0.16	2.12
DOC (C=C stretch/C-H stretch wavenumber)	2.81	0.58	1.77
COC value	0.70	0.70	0.00

There was a larger decrease in nCH_2/nCH_3 as compared to lower temperatures where the value for the parameter dipped to 1.41 for S_3 from a feed value of 1.76 (Table 6.5). However, S_2 showed a minimum for the chain length parameter at 1.22. The concentration profiles signify that both S_2 and S_3 exist at majority of the reaction times (Figure 6.13a) so the conversion of $S_1 \rightarrow S_2$ was considered as important as $S_1 \rightarrow S_3$ at 400 ºC. Gas production increased to 16 %wt. over 1440 min [24] and this combined with the overall decrease in the -CH_2 content or increase in the -CH_3 content indicated that cracking was taking place significantly. Interestingly, curve resolution on the FTIR spectra of cracked products of Cold Lake bitumen at 400 ºC determined that nCH_2/nCH_3 reduced from the first to the third pseudo-component but reached a maximum for the second component. [29] Hydrogen transfer followed by free-radical recombination was stated as the reason for the initial increase but the rate of bond scission is more important and could have been higher at later reaction times to lead to a decrease in nCH_2/nCH_3. In the case of Athabasca bitumen, the trend in nCH_2/nCH_3 from S_1 to S_2 to S_3 could indicate that the rate of bond breaking was greater than the rate of bond formation, but the increase from S_2 to S_3 was intriguing.

Inspection of the resolved spectral profiles at 1380 cm^{-1} (Figure 6.13d) provided evidence of methyl transfer happening at 400 ºC. The absorption band was split for S_2 but not for S_3 which indicates that methyl transfer from dimethyl carbon centres followed by free-radical stabilization by hydrogen abstraction possibly occurred at higher reaction times, which could be a reason for the increase in -CH_2 content with methyl content being a constant (as was illustrated in Figure 6.7). This could have led to the observed increase in nCH_2/nCH_3 for S_3 compared to S_2 but cracking kept this value lower than in the feed. The aromatic ring in compound (1) can be replaced with an

aliphatic group as well since it depicts changes occurring after significant thermal cracking has progressed. Free-radical recombination of the cracked lighter products with aromatic compounds (as shown in Figure 6.14) can also lead to an increase in -CH_2 content, which is explained later in this section.

DOC decreased dramatically from 2.81 to 1.77 from S_1 to S_3 with a minimum of 0.58 for S_2 (Table 6.5). This was another observation that differed from Cold Lake bitumen conversion at 400 ºC, where the DOC increased from 0.17 to 0.26 across the three pseudo-components. Since DOC was calculated as the ratio of the areas under the aromatic C=C stretch and the aromatic C-H bending in Tefera et al., [29] it was seen in concurrence with the inverse of H/C ratio of the products which decreased from 1.43 to 1.09 over 360 min for Cold Lake bitumen. In contrast, there were similar trends seen for both H/C and DOC for Athabasca bitumen. H/C decreased slightly from a feed a value of 1.48 to 1.46 at 15 min which was also the value for the 1440 min product. [24] It reached a minimum of 1.43 at 90 min and 240 min that were the start and end regions of the existence of S_2. Since hydrogen rich gaseous products were formed in large amounts, it was deemed that the coke was hydrogen deficient in order to maintain the H/C ratio of the liquid products.

Arguing on the same lines as that for Cold Lake bitumen, DOC for Athabasca bitumen should have mildly increased owing to the slight decrease in H/C. The fundamental difference between H/C and DOC was that H/C related to the whole liquid product while DOC conformed to the aromatic region only. The minimal change in H/C also found support in the work by Wiehe, [124] where in the model for coke formation from pyrolysis of Cold Lake bitumen at 400 ºC, he suggested that asphaltenes reached a constant ratio H/C once a second phase called mesophase (that eventually leads to coke) starts forming. On the whole, DOC was thought to be a less reliable parameter for two reasons: (i) since it considered areas and not intensities at single wavenumbers; and (ii) it only takes into account the aromatic part of the spectrum but the H/C ratio corresponds to the entire liquid product.

There was also a noted difference in the EOS for the pseudo-components in converted products of Athabasca and Cold Lake bitumen at 400 ºC. EOS decreased from 3.20 to a surprisingly low value of 0.16 from S_1 to S_2 for Athabasca bitumen and then increased to 2.12 for S_3 while there was a gradual and mild decrease in EOS from 2.29 to 2.26 for Cold Lake bitumen. It can be seen from Figure 6.13e that the peak for monosubstituted aromatics at 727 cm^{-1} was much

higher for S_2 and S_3 than for S_1. It was in fact highest for S_2 among all 3 pseudo-components. The formation of mono-substituted aromatics is possible by two mechanisms: (i) as shown in Figure 6.10, with cracking of a three-ringed naphthene-aromatic followed by hydrogen transfer by hydrogen disproportionation to give the mono-substituted aromatic compound (14); (ii) a radical hydrogen transfer (RHT) mechanism as suggested by Blanchard and Gray [125] through intramolecular hydrogen transfer could facilitate the replacement of an aromatic-C – alkyl C bond from a di-substituted aromatic to yield a mono-substituted aromatic as depicted in Figure 14. A previous study on cracking of hexadecane in the presence of benzene indicated the formation of biphenyl which would have been possible only if a hydrogen was abstracted from benzene. Removal of a carbon attached to a benzene ring would be relatively easier but even though other studies in literature indicate that this temperature could render them susceptible to be broken, it is considered a rare occurrence at 400 ºC. [54,121] The possibility of further cracking of compounds similar to (14) through hydrogen abstraction from benzylic carbon followed by RHT mechanism to produce a mono-substituted aromatic, though less probable, cannot be ignored.

It is fascinating to see how the same compound (15) can follow different paths during cracking depending on the temperature and its structure. The benzylic C – tertiary carbon bond in (15) (BDE = 320 kJ/mol) is weaker than the bond between the carbon attached to the tertiary carbon of the naphthenic ring and the aliphatic side chain (BDE = 365 kJ/mol as shown in Figure 6.12). Homolytic scission of this bond gives compound (21) which can undergo intramolecular hydrogen transfer and the hydrogen radical can arise from the second carbon in the side chain. The delocalization of this radical inside the aromatic ring facilitates cleavage of the aromatic C-aliphatic C bond (BDE = 420 kJ/mol) and produces the stable benzyl radical and the alkene (23) whose primary radical is stabilized by resonance with the double bond. Due to the higher energy available at 400 ºC than at 380 ºC, conversion of the naphthene to an aromatic is not required to aid the cleavage of the C-C bond in the C_4 substituent as was proposed to occur at 380 ºC (Figure 6.12). It also reflects the decrease in EOS as the aromatic in (15) with 2 substituents and 4 adjacent hydrogens undergoes cracking to produce (24) which has 1 substituent and 5 adjacent hydrogens. The alkene (25) can undergo further cracking to give lighter compounds (26) and (27) which can recombine with aromatic free radicals to produce compounds with longer side chains. This occurs at later reaction times and could also be a reason for the observed increase in nCH_2/nCH_3 from S_2

to S_3 (Table 6.5). Though no specific proof of this mechanism is given in this work, it is proposed based on the experimental data and subsequent curve resolution results.

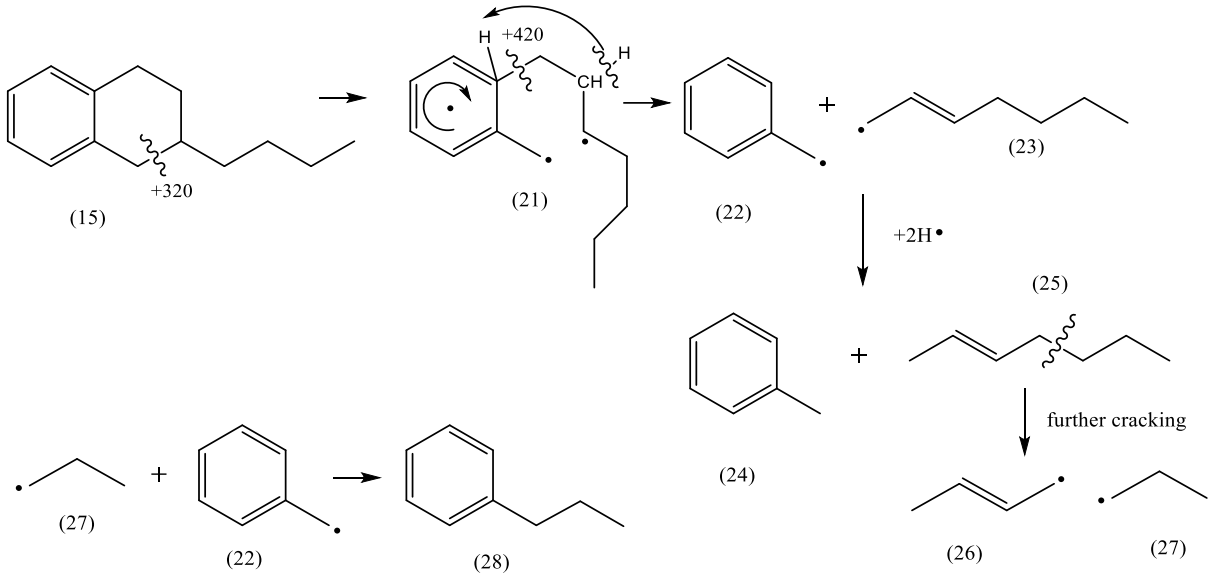
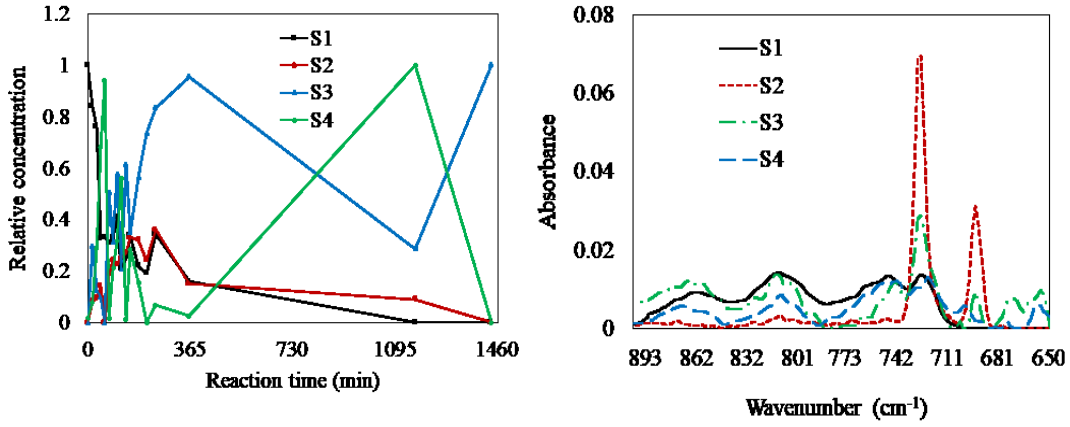


Figure 6.14. Plausible type of reaction happening at 400 ºC where cracking of the weaker benzylic C-tertiary C (in compound (15)) followed by intramolecular hydrogen transfer and hydrogen abstraction to yield the mono-substituted aromatic (compound (24)) and the conjugated free radical (23). This can crack further to give lighter aliphatic products. Possibility of free-radical recombination to form compound (28) is also shown.

Since the value of EOS for S_2 was quite low, an internal test was done to check the validity of this observation. This was done by relaxing the number of components from 3 to 4 and inspecting the resolved concentration and spectral profiles in the 900 – 650 cm^{-1} region. It should be kept in mind that though 4 components gave a lower LOF and higher R^2 compared to the optimum value of 3 (section 6.4.1 and Table D.2), it incorporates a secondary eigenvalue that would make the profiles prone to noise. This is visible in the concentration profiles as they seem to be noisier during the lower reaction times as shown in Figure 6.15a. The spectral profiles (Figure 6.15b) indicate higher intensities at 727 cm^{-1} for both the second and third pseudo-components while the peaks for S_4 were not as clear. Concurrently, the EOS for S_2 and S_3 were 0.25 and 0.91 and this indicated that the lower value for EOS for S_2 did not happen by chance in the three-component spectra (Table 6.5). This lower value for EOS could possibly suggest that the formation of mono-substituted

aromatics occurred by both hydrogen disproportionation (Figure 6.10) and by the RHT mechanism (Figure 6.14), both of which would cause the content of hydrogen donor molecules to depreciate.

For the three-component resolution, the intensities at 744 cm^{-1}, 810 cm^{-1} (Figure 13e) also provided hints towards determining which of the two mechanisms proposed were dominant at different phases of the reaction. For S_2, the intensity for *o*-disubstituted aromatic showed a large decrease when the concentration of mono-substituted aromatic was maximum but again increased for S_3. On the other hand, the intensities for the *m*-disubstituted aromatic at 810 cm^{-1} remained near-constant. This might suggest that formation of mono-substituted aromatics through the RHT mechanism (Figure 6.14) could be prevalent initially, as it decreases the *o*-disubstituted aromatic content, and the mechanism of the type proposed in Figure 6.10 could occur at later times, since the *o*-disubstituted aromatic content is not affected. This can also allow for free-radical recombination from the products of the RHT mechanism to occur at later reaction times (compound (28)). An alternative route for the formation of mono-substituted aromatics apart from RHT mechanism or the pathway shown in Figure 6.10 is suggested in the discussion presented later on Bayesian networks (Figure 6.21 – section 6.4.6.2). No cleavage of the carbon attached to the aromatic ring was required in this route, which seemed quite feasible at temperatures < 400 ºC.



(a) (b)

Figure 6.15. Plots of: (a) concentration profiles and (b) spectral profiles in the region 900 – 650 cm^{-1} for curve resolution applied on the 400 ºC dataset using 4 pseudo-components.

Also, proof for the formation of lighter products was seen from the boiling point distribution profiles obtained from simulated distillation in Chapter 5 from the work by Sivaramakrishnan et al. [24] The amount of material boiling below the vacuum residue cut point (at 525 ºC) increased

from 42 %wt. in the feed to 70 %wt. in the 1440 min product. The formation of lighter products could also explain the monotonic decrease in viscosity from a feed value of 36 Pa.s to ~0.003 Pa.s at 1440 min. The shear rates employed for remaining above the lower device limit of the viscometer and also measuring in the Newtonian region for each sample increased from 10 s^{-1} for the feed to 1500 s^{-1} for the 1440 min product. [24] The decrease in refractive index showed a change in composition of the product and corroborated with the fact that aromatic content did not increase as much as was observed for Cold Lake bitumen in the work by Zachariah and De Klerk. [32] They also observed a net increase in viscosity after the first 30 min of reaction and the viscosity reached a value that was almost equal to the feed value after 360 min of reaction. Apart from a possible difference in reaction chemistry for Athabasca bitumen as compared to Cold Lake bitumen, the use of solvents like methylene chloride to extract the liquid products after thermal reaction from the micro-reactor in the works on Cold Lake bitumen [29,31–33] was investigated as a possible cause for the observed increase in viscosity. For this reason, all liquid products in this work were mechanically extracted with no solvent being used in the process. In addition, a number of previous works on thermal cracking of Athabasca bitumen reported a monotonous decrease of viscosity and other properties like asphaltenes content, which was similar to the data reported in a recent study involving the liquid products obtained in this work. [24,126–128]

6.4.3.5 Results and analysis at 420 ºC

First, it should be noted that this temperature is higher than the lower visbreaking region as defined by Wang et al., [33] which was considered to be 400 ºC and lower. Since the lower temperatures were investigated previously, it was decided to operate close to industrial conditions that function between 430 – 490 ºC. [36] Longer reaction times were explored and are similar to those employed in soaker visbreakers. As expected, a large amount of coke was formed (~17 %wt. at 420 ºC and 660 min compared to ~12 %wt. at 400 ºC and 1440 min) with the liquid yield being much lesser than at lower temperatures. Besides, 4 datapoints were collected at this temperature along with the feed, which gave a better resolution than at 300 ºC. The resolved concentration and spectral profiles obtained through ALS-optimization at 420 ºC are provided in Figure 6.16 and the residual plot is given in Figure D.12e in Appendix D.

There were similarities in the initial concentration estimates (Figure 6.5a) and the final resolved profiles (Figure 6.16a) which indicated that the initial estimates provided by EFA were a good guess for the optimization. S_2 rose at 0 min as soon as S_1 started decreasing in concentration in both the initial estimates and final profiles. At the point of appearance of S_3 (360 min), S_2 decreased and S_3 remained to be highest in concentration for the rest of the reaction times. These observations were indicative of a reaction pathway $S_1 \rightarrow S_2 \rightarrow S_3$ though the reaction mixture consisted of both S_2 and S_3 at higher reaction times. Table 6.6 provides the values of the spectra-derived quantitative parameters for the three pseudo-components at 420 ºC.

Inspection of the resolved spectral profile in the 3000 – 2750 cm^{-1} range suggests a significant change in both methylene and methyl group intensities as one moves from S_1 to S_3 but the relative changes were better indicated by their ratios. nCH_2/nCH_3 decreased to the lowest value for S_2 (1.13) out of all previously investigated temperatures in this work and interestingly, remained almost constant for S_3 as well. This could probably indicate cracking as the dominant reaction with the rate of bond scission much higher than bond formation for most times. The free radicals formed through bond cleavage are being stabilized by hydrogen radicals (such as compounds (19) and (20) in Figure 6.12), thus preventing them from further recombination to yield longer molecules such as (28) in Figure 6.14.

DOC decreased continuously from 2.82 for S_1 to 1.33 for S_3 (Table 6.6) as opposed to previous temperatures where it exhibited a minimum or a maximum. EOS also followed a similar pattern to the chain length parameter where it decreased from S_1 to S_2 (3.13 to 1.97) and then remained nearly constant at 1.98 for S_3 (Table 6.6). This value was lesser than that observed at 400 ºC, suggesting the formation of more monosubstituted aromatics through either mechanism as proposed for 400 ºC in the previous section. The absorption bands for C-H deformation at 1380 cm^{-1} (Figure 6.16d) for all three pseudo-components did not appear to be split and thus the occurrence of methyl transfer could not be confirmed even though the temperature was high enough to facilitate this phenomena. [54] A clear increase in mono-substituted aromatic content (Figure 6.16e) meant that side reactions that decline the value of the product such as intra-aromatic ring closing were suppressed. This also rendered the liquid products sufficiently hydrogen rich and with the formation of hydrogen-enriched gases, coke would likely be depleted in hydrogen in order to maintain the material balance. This argument found support in the Wiehe model [129] for coke formation from bitumen pyrolysis where it was stated that the H/C ratio of the unconverted

asphaltenes in the liquid product did not change much once the second liquid phase (mesophase) started forming and eventually converted to coke. Overall, it can be said that cracking was quite prominent at 420 ºC.

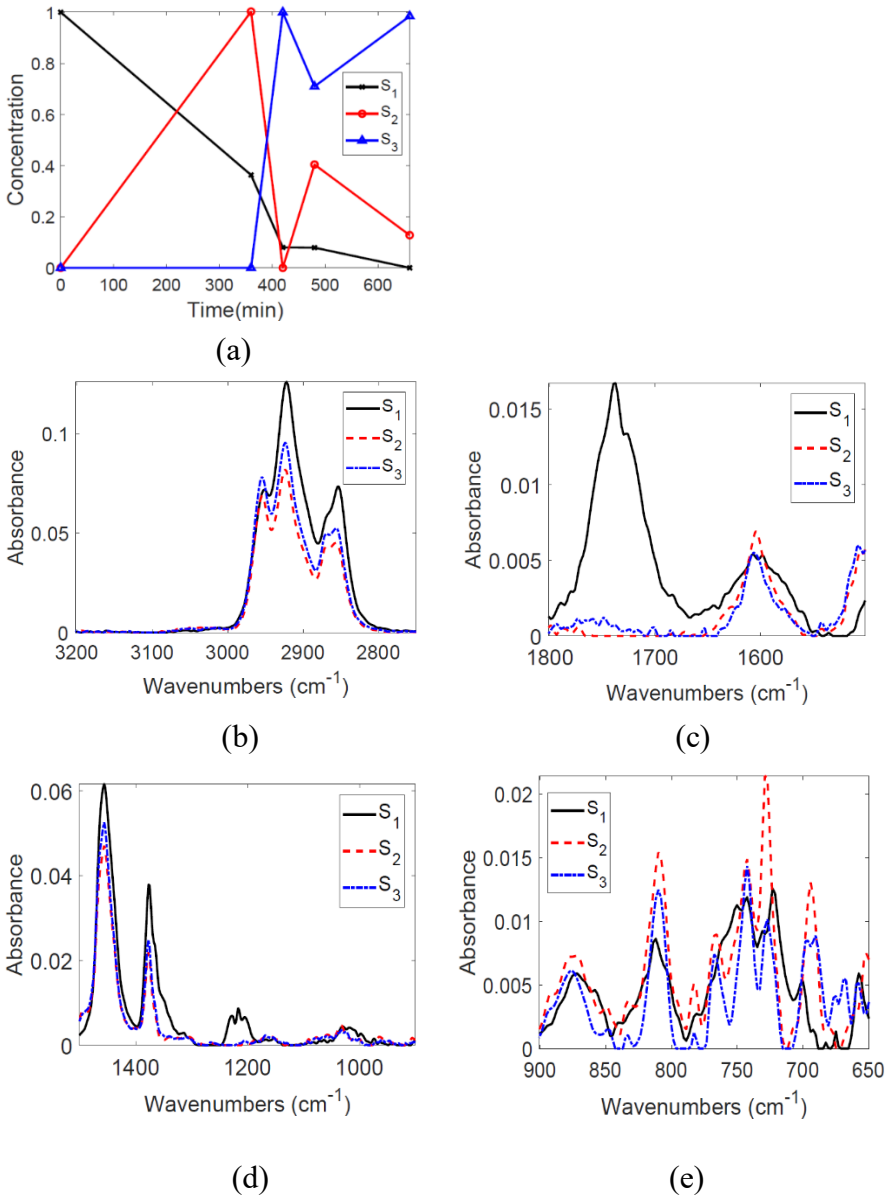


Figure 6.16. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 420 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{-1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

Table 6.6. Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 420 ºC.

Pseudo-component	1	2	3
n-CH_2/*n*-CH_3	1.78	1.13	1.16
Overall EOS	3.23	1.97	1.98
DOC (C=C stretch/C-H stretch wavenumber)	2.82	1.48	1.33
COC value	0.76	0.00	0.12

6.4.4 PSO-optimized C, S profiles and spectra-derived quantitative parameters

In this section, the results of the concentration and spectral profiles using PSO combined with a constrained minimization function called *'fmincon'* as the optimization method are provided for the temperature-wise datasets. The PSO algorithm was embedded inside the ALS loop so that it served as an improvement upon the concentration profiles supplied by ALS and *'fmincon'* identified a further local optimum if any, after the PSO converged. It was compelling to see that there was an enhancement in the resolution of the final profiles for ALS-PSO as compared to ALS profiles as expected and the subsequent chemical interpretation was also very similar. For this reason, no detailed interpretation is provided in all cases for the PSO-optimized profiles but the differences between the ALS and ALS-PSO routines for SMCR in terms of ameliorating noisy regions of ALS-resolved spectra, values in LOF and R^2 and the speed of convergence are discussed in this section.

6.4.4.1 Results at 300 ºC

Due to the similarity in the nature of the results of ALS-PSO with ALS as mentioned in the previous section, all concentration and spectral profiles for the PSO-optimized method are provided in section D.7 in Appendix D. Figure D.13 depicts the concentration and spectral profiles along with the ALS-PSO residual when SMCR was conducted on the 300 ºC dataset.

In comparing the concentration profiles of ALS-PSO (Figure D.13a) and ALS (Figure 6.6a), it can be seen that there was a difference in concentration at 360 min. The relative concentration of S_2 was 0.75 and for that of S_1 and S_3 were 0.13 and 0.13 respectively, while in the ALS-optimized profiles, it was 1, 0 and 0 for S_2, S_1 and S_3, respectively. Though there is no direct measure to specify which concentration profile is better resolved, a value other than 0 and 1 for the pseudo-component concentration at the second data point suggests that the addition of PSO to the ALS method brought the profile closer to reality for the system. Another improvement was in removal of the noisy patterns in the spectral region between 1800 – 1500 cm^{-1} that was evident in the ALS-resolved profiles (Figure 6.6c) but was well resolved in ALS-PSO (Figure D.13d). The residual plot for ALS-optimized profiles (Figure D.12a) extended only in the negative direction, which seemed unusual. But in the case of ALS-PSO-derived profiles (Figure D.13b), the residual had both positive and negative values and was more symmetric than the ALS profiles alone. Table 6.7 shows the ALS-PSO spectra-derived quantitative parameters for S_1, S_2 and S_3 for the resolution performed on the 300 ºC dataset.

Table 6.7. Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 300 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.80	1.79	1.82
Overall EOS	3.48	3.55	3.23
DOC (C=C stretch/C-H stretch wavenumber)	2.91	2.70	2.79
COC value	0.77	0.00	0.00

Although the absolute values of the parameters in Table 6.7 were slightly different compared to the ones derived from the ALS-resolved spectra (Table 6.2), the overall trends were the same. This meant that the interpretation of the chemical changes and types of reactions occurring at 300 ºC based on these parameters is similar to the discussion in section 6.4.3.1 and the reader is referred to that section. Further differences between the ALS and ALS-PSO techniques in terms of LOF, R^2 and convergence speed are given in section 6.4.4.6.

6.4.4.2 Results at 350 ºC

Figure D.14 in Appendix D provides the concentration and spectral profiles and also the residual plot when ALS-PSO approach was used to resolve the FTIR spectra. The concentration profiles for ALS-PSO optimized profiles (Figure D.14a) followed the same pattern as the ALS-optimized ones with the only difference being that the magnitude of the concentration at 180 min for S_1 and S_3 was higher and lower, respectively, for ALS-PSO as compared to the ALS method. This might be well due to the PSO method trying to satisfy the closure constraint at this reaction time (the sum of concentration values was ~1.5 for the ALS-derived profiles).

Also, the noisy spectrum in the 1800 – 1500 cm^{-1} region for S_3 in the ALS-optimized profiles was resolved to a higher extent by the ALS-PSO method (Figure D.14c). While the intensity for S_2 at 1740 cm^{-1} was 0, S_3 showed a mild absorption at this wavenumber which was reflected in the COC value as shown in Table 6.8. The only difference in the trends of the spectra-derived quantitative parameters of the ALS-PSO profiles as compared to the ALS-profiles was the continuous decrease in EOS from 3.50 to 2.15 (Table 6.8) in the PSO case whereas there was a maximum at S_2 for the ALS case (Table 6.3). However, the overall effect was a decrease in EOS for both methods which did not change the chemical interpretation by much. The peaks in the other regions were well resolved and exhibited similar trends to those derived from the ALS-optimized profiles. These values are compiled in Table 6.8.

Table 6.8. Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 350 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.81	1.77	1.76
Overall EOS	3.50	2.98	2.15
DOC (C=C stretch/C-H stretch wavenumber)	2.89	3.15	2.63
COC value	0.77	0.00	0.12

6.4.4.3 Results at 380 ºC

The results of ALS-PSO analysis on the FTIR spectra at 380 ºC along with the residual is provided in Figure D.15 in Appendix D. The concentration profiles indicate a mild difference from the ALS-optimized profile but one can arrive at the same reaction network of $S_1 \rightarrow S_2 \rightarrow S_3$ by inspection (Figure D.15a). The concentration of S_3 showed a monotonic rise from the start which was different from that in the ALS profiles (Figure 6.11a) while S_2 had a similar trend except that it did not rise as sharply at the higher reaction times of 1320 min and 1620 min.

Similar to other temperatures, the noisy region for S_2 in the 1800 – 1500 cm^{-1} region of ALS profiles (Figure 6.11c) was mitigated by the addition of the PSO method as can be seen from Figure D.15d without altering the COC value too much (Table 6.10). The values of the spectra-derived parameters for the 380 ºC dataset are given in Table 6.9.

Table 6.9. Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 380 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.83	1.31	1.39
Overall EOS	3.35	1.83	3.58
DOC (C=C stretch/C-H stretch wavenumber)	2.91	1.86	1.92
COC value	0.87	0.2	0.00

It was interesting to see that the trend in the *n*CH_2/*n*CH_3 was similar to that of profiles obtained through ALS optimization even though the spectra for S_2 in the 3200 – 2750 cm^{-1} region had higher absolute intensities for methylene and methyl C-H stretches. However, the increase in -CH_3 stretch was more than -CH_2 for this pseudo-component, which resulted in a decrease in the chain length parameter (Table 6.9). The trends in other parameters like EOS, DOC and COC were the same as in the ALS profiles, which makes the chemical interpretation the same as discussed in section 6.4.3.3. Another engaging observation was the persistence of band splitting for all three pseudo-

components at 1380 cm^{-1} (Figure D.15e). This assigns more credibility to the proposition of minimal methyl transfer occurring at 380 ºC from the ALS profiles as discussed in section 6.4.3.3.

6.4.4.4 Results at 400 ºC

The concentration and spectral profiles for the 400 ºC data set resolved by ALS-PSO optimization along with the residual plot are given in Figure D.16 in Appendix A. The relative concentrations of S_1, S_2 and S_3 followed similar paths for both ALS and ALS-PSO-optimized profiles as seen in Figure 6.13a and Figure D.16a respectively. In the ALS-PSO-optimized profiles, the concentration of S_1 decreased continuously while S_2 showed a global maximum at 150 min and S_3 peaked at higher reaction times. An important observation was that the sum of relative concentrations of the three pseudo-components at 150 min was much closer to 1 for the PSO approach rather than the ALS method (section 6.4.3.4). This was also true at some other reaction times between 120 min and 210 min as well. This signified that PSO embedded with ALS caused the concentration profiles to adhere to the closure constraint better.

The spectral profiles from the PSO method narrated a similar story as the ALS ones with the trends in the derived parameters confirming this observation as shown in Table 6.10. EOS and DOS had minimum values for S_2 but showed an overall decrease from S_1 to S_3 in these parameters. The peaks at 727 cm^{-1} for S_2 and S_3 (Figure D.16f) were similar to that of the ALS-optimized profile (Figure 6.13e). The spectral bands at 1380 cm^{-1} appeared split for S_1 but straight for S_2 and S_3 indicating no major changes in the reaction chemistry from what was discussed for the ALS-optimized profiles in section 6.4.3.4.

Table 6.10. Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 400 ºC.

Pseudo-component	**1**	**2**	**3**
n-CH_2/*n*-CH_3	1.79	1.12	1.41
Overall EOS	3.45	0.60	2.10
DOC (C=C stretch/C-H stretch wavenumber)	2.89	1.13	1.56
COC value	0.86	0.79	0.00

6.4.4.5 Results at 420 ºC

The results of ALS-PSO optimization at 420 ºC are given in Figure D.17 in Appendix D. The concentration of S_1 was similar for the profiles obtained from both ALS (Figure 6.16a) and ALS-PSO methods (Figure D.17a), but the PSO profiles were seen to adhere to the closure constraint better, especially at 360 min, where the concentration of S_2 was much lower than for the ALS-optimized profile, thus making the total concentration ~1.1 as opposed to 1.4 for the ALS method. At all other times, the sum of concentrations was around ~1.1 for results obtained from both methods, thus signifying the usefulness of PSO.

The residual appeared to fluctuate in a lower range for PSO ($\pm$0.003 in Figure D.16b) while it reached a maximum of ~0.008 for the ALS-optimized profile (Figure D.12e). This led to a lower LOF and R^2 by the slightest of margins, which is shown in section 6.4.4.6. The spectra-derived quantitative parameters for the ALS-PSO profiles at 420 ºC are summarized in Table 6.11.

Table 6.11. Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 420 ºC.

Pseudo-component	1	2	3
n-CH_2/n-CH_3	1.82	1.14	1.17
Overall EOS	3.52	1.97	1.88
DOC (C=C stretch/C-H stretch wavenumber)	2.90	2.11	1.71
COC value	0.79	0.00	0.00

At both 400 ºC and 420 ºC, ALS profiles were well resolved in all regions including the noise-prone 1800 – 1500 cm^{-1} region (Figure D.17d) and reflects the decrease in COC value for the 420 ºC dataset which corroborates with the ALS profiles as well (Table 6.6). The absorption bands at 1380 cm^{-1} appeared to have a split characteristic for all three pseudo-components. This implied that the occurrence of methyl transfer could not be confirmed as was explained for the ALS profiles in section 6.4.3.5. While trends in $n$$CH_2$/$n$$CH_3$ and DOC were similar to the ALS-optimized profiles (Table 6.6), the value of EOS for S_3 slightly decreased from S_2 for ALS-PSO-optimized profiles (Table 6.11) whereas it remained constant for the profile obtained through ALS optimization. If the ALS-PSO method is considered to capture the system changes better, it is just indicative of the formation of mono-substituted aromatics in higher amounts and the temperature might be sufficient to break the Ar-C-alkyl-C bond that can lead to a decrease in DOC as well. The reader is referred to section 6.4.3.5 for an extended explanation on the chemical changes occurring at 420 ºC.

6.4.4.6 Comparison of ALS and ALS-PSO methods

The two methods of optimization used for MCR in this work namely, ALS and PSO-*'fmincon'* embedded in ALS are compared in terms of LOF, R^2 as the performance indicators and the rate of convergence, and the results are discussed in this section. The values of LOF and R^2 as calculated from the residual plots for each method are given in Table 6.12.

Table 6.12. LOF and R^2 values for the dataset at each temperature when ALS and ALS-PSO were employed as the final optimization approach.

*	350 ºC		380 ºC		400 ºC		420 ºC	
	ALS	**ALS-PSO**	**ALS**	**ALS-PSO**	**ALS**	**ALS-PSO**	**ALS**	**ALS-PSO**
LOF	2.779	2.771	3.596	3.587	6.713	6.701	1.741	1.731
R^2	99.923	99.926	99.871	99.876	99.543	99.551	99.975	99.989

*the performance indicators at 300 ºC are not shown due to the lesser number of datapoints.

Tefera et al. [29] reported that the ALS algorithm applied on Cold Lake bitumen converged in about 10 iterations at all temperatures and Shinzawa et al. [80] concluded PSO to be better than EFA for estimating the initial concentrations in SMCR by comparing the squared residual for both the methods. Also, they employed a residual based on global phase angle proposed by Noda [130] and this considered the time sequence of events and the effect of any external perturbation. In our work, we can see from Table 6.12 that though the difference in LOF and R^2 values is quite small for both the methods, ALS-PSO had a lower residual than ALS. The largest difference was for 400 ºC, where the residual for ALS-PSO was an order of magnitude lower than that for ALS. Subtle differences in residual plots are well captured by these indicators and the difference was in the third decimal for both indicators. It should be noted that LOF increased with the number of elements in the dataset.

The rate of convergence for both the methods is depicted in Figure 6.17 and shown for 420 ºC and it was similar for the other temperatures considered. ALS converged in about 30 iterations and the standard deviation of the residual plotted on the *y*-axis is calculated from the sum of squared residual and reached a stable value of 0.00079. On the other hand, PSO combined with ALS converged in half the time (~15 iterations) and the standard deviation of the residual reached a constant value of 0.00056, which was also lower than that of ALS. This confirmed that ALS-PSO converged faster than ALS and could be better suited for online monitoring of the system in focus as highlighted in the introduction section.

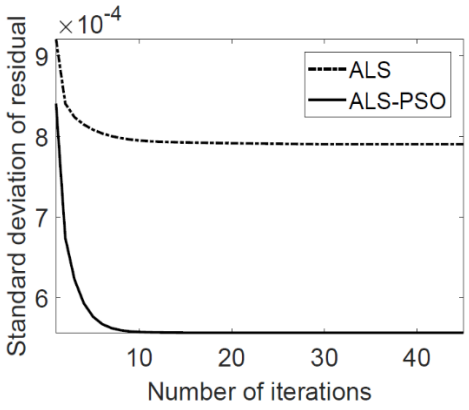


Figure 6.17. Rate of convergence in terms of standard deviation of residual vs. number of iterations for ALS and ALS-PSO algorithms used in SMCR in this work.

6.4.5 Assigning chemical relations to clusters from BHC

For the purpose of Bayesian clustering and subsequent network learning, 214 wavenumbers that were mostly in the 4000 – 3200 cm^{-1} region were omitted from the clustering analysis in order to limit the noise in the system. The procedure for this was provided in section 6.3.4.1. Table 6.13 provides the distribution of the obtained 5 clusters along with their plausible chemical identities and the associated wavenumbers determined by referring to Silverstein and coworkers, [117] Colthup et al. [131] and Beauchamp. [132]

The difference between clustering analysis and the decomposition of the data matrix into the pseudo-components in SMCR is that the user is able to decide the number of groups to work with in clustering based on some initial chemical knowledge of the system whereas in SMCR, the number of pseudo-components are chosen through maximum or minimum of specific indicators like ROD and IND functions. The common trait in both methods is that they lump variables into pseudo-components, and is often done in developing kinetic models. [133] The reason for choosing 5 clusters is based on the classification of compounds existing in bitumen into basic groups of alkanes, alkenes, aromatics, condensed aromatics, and naphthenes as was already mentioned in section 6.3.4.1. It can be seen from Table 6.13 that the chemical entities are not as simple and some groups can consist of a mixture of compound classes and functional groups as well. However, the basis for the choice of 5 clusters was a good starting point as it led to a good separation of species that were hypothesized to participate in the thermal reactions through the SMCR methods.

It was important to note that the wavenumbers were ascribed to chemical species only if they had an absorption in the original spectra for bitumen or thermally processed bitumen at any temperature or reaction time. For example, the band between 1000 – 1100 cm^{-1} was captured into the 5th cluster but the absorption corresponding to alkoxy C-O was quite weak in the original spectrum.

The first cluster consists mainly of ortho (*o*-), meta (*m*-) and para (*p*-) di-substituted or tri-substituted aromatics whose absorptions were at 740 cm^{-1}, 760 cm^{-1}, 810 cm^{-1} and 860 cm^{-1}. The aliphatic alkanes with -CH_3 bending at 1380 cm^{-1}, 1460 cm^{-1} also fall under this group while the wavenumbers 2890 – 2840 cm^{-1} could also represent CH_2 present in cyclic alkanes. A naphthene-aromatic with an attached alkyl side chain on either of the rings would be a good representative model compound of this cluster. If the alkyl group is located on the aromatic ring, it will be tri-substituted with possible absorption at ~860 cm^{-1}. In the molecule (34) shown in Figure 6.20, the R group is in the para and meta position with respect to the adjacent aromatic ring and thus can also absorb at 760 cm^{-1} and 810 cm^{-1} as well. However, if the substituent is on the naphthene ring as in (29), - the left aromatic ring is *o*-di-substituted, thus absorbing at 740 cm^{-1}. The naphthene ring contains the aliphatic CH_2 stretch and the terminal CH_3 would be present in the alkyl side chain, which covers the other absorptions in this cluster. Thus, both molecules (29) and (34) are good representatives of the first group.

Since a bunch of wavenumbers representing specific groups were part of each cluster, it was worthwhile to check the most important wavenumbers by calculating the square root of the sum-squared intensities for all 35 samples at each wavenumber in the cluster. This number can be termed as an effective intensity for the wavenumber. It was seen that all values for the effective intensity in the first cluster were in the range 0.03-0.05 with no particular dominant value. This implied that all wavenumbers are equally important in this cluster, which makes chemical sense as well. If a naphthene-aromatic did not have any side chain or substituent, there would be no -CH_3 group, which is quite rare in bitumen. [112]

On the other hand, if the wavenumbers corresponding to aliphatic -CH_2 (1380, 2890-2840 cm^{-1}) had the maximum effective intensity, it would be similar to the second cluster, and it would not have been logical to have two clusters with the same chemical identity. Cluster 2 represents only the stretching frequencies corresponding to the methylene group, which requires it to be a cyclic aliphatic compound. The third cluster includes mainly mono-substituted aromatics with alkyl side

chains due to the respective absorption bands included in this cluster and the effective intensity does not vary by much (Table 6.13). The fourth cluster possibly consists of just non-cyclic aliphatic compounds due to the methyl and methylene stretching bands at 2850 cm^{-1}, 2920 cm^{-1} and the sole bending band for -CH_3 at 1460 cm^{-1} (Table 6.13).

Table 6.13. Classification of chemical group functionalities identified for the 5 clusters obtained through applying BHC on the FTIR spectra at all 5 temperatures between 300 ºC and 420 ºC.

Group No.	Wavenumber range (cm^{-1})	Associated chemical attributes/classes of compounds
1	740, 760, 810; 860 2840 – 2890; 1370, 1440; 2960 - 2970	*o-, m-, p-**disubstituted or tri-substituted aromatics***, aliphatic/naphthenic-CH_2, aliphatic-CH_3
2	2918 - 2933	***Naphthenic***-CH_2
3	694 – 699; 727 – 730; 2840, 1370, 1440; 2974 - 2979	Aromatic ring bend, ***Mono-substituted*** aromatics Aliphatic, naphthenic-CH_2, CH_3 Aliphatic-CH_3
4	2850 – 2920; 2940 – 2960; 1460	***Aliphatic***, naphthenic-CH_2, ***aliphatic***-CH_3
5	(i) 599 – 649; (ii) 650 – 680; 1150 – 1200; 1230 – 1360; 1750 – 2820; 1620 – 1720; 3200 - 4000 (iii) 1730 – 1740; (iv) 900 – 1000; (v) 1216 – 1234; 1000 – 1100; (vi) 3050 – 3100; 1605 (vii) 1380, 1450; (viii) 2980 – 3050	(i) Noisy; (ii) Baseline intensities; (iii) Ester, anhydride-type; (iv) Mono-substituted, geminal ***alkenes***; (v) phenolic, acyl C-O; alkoxy C-O (weak) (vi) ***Aromatic*** C-H stretch, C=C stretch (weak) (vii) ***aliphatic***-CH_2, -CH_3 (viii) ***alkene***-C-H stretch

Lastly, the allotment of a specific type of chemical species to cluster 5 was found to be difficult due to its diverse range of wavenumbers. The effective intensity plot of this cluster is shown in Figure D.18 in Appendix D. This was the only cluster where particular wavenumbers had distinct peaks in effective intensities as indicated in Figure D.18. It can be said that there were both aliphatic and aromatic compounds present in this cluster with the existence of weak C=C (1605 cm^{-1} in Figure D.18) and C-H stretches for aromatics and strong absorptions relating to aliphatics consisting of an alkene C-H stretch (2980 cm^{-1} in Figure D.18), sp^3 methylene and methyl groups (1370 cm^{-1}, 1460 cm^{-1} in Figure D.18). The type of substitution, whether mono, di or tri, was not evident from the wavenumbers belonging to this cluster. The presence of alkenes was an indication that this could be one of the final clusters in the reaction network (since thermal cracking leads to some alkene formation as well). The presence of heteroatoms were indicated for this group with absorptions at 1740 cm^{-1} being paired with those in the 1230 – 1216 cm^{-1} region, which could indicate ester-type and anhydride-type oxygenates and phenolic compounds as well. The alkoxy C-O group, was also identified as part of this cluster, though the peak was weak. But the conversion or formation of heteroatoms during thermal conversion is not particularly discussed in this work. It is also emphasized that there is no proof provided for the existence or transformations of these model compounds during thermal processing in this work and only possible pathways are explored based on the quantitative changes observed in the chemometric analysis.

Moreover, the distribution of chemical species among the different clusters for Athabasca bitumen was relatively simpler as compared to that of Cold Lake bitumen (Table 1 in Tefera et al. [29]). Aromatics were present in the first three groups apart from cycloalkanes and straight chain alkanes as well. The 5th group was the only single-component group that represented alkanes only. This was similar to the 4th cluster in Table 6.13 that pointed to aliphatic methylene and methyl groups characteristic of alkanes alone. The conversion of one group to another during the thermal reaction is discussed in the next section by considering a particular compound that is representative of the entire cluster.

6.4.6 Causal relationships between the BHC clusters extracted by Bayesian networks and a plausible representative reaction pathway

6.4.6.1 Bayesian network structure

Once the five clusters were identified, they were considered as the nodes to develop directional relationships between them using Bayesian learning methods. As described in section 6.3.4.2, three algorithms were used to build a network separately by maximizing a BIC score (equation 6.18). The terms in equation 6.19 describe the dependence and independence of the current node on the node which gave rise to it (also called the parent node) and constitute the first term in the BIC equation (equation 6.18). The logarithmic term in BIC indicates a penalty term that regularizes between model fit and the ability to generalize the model. In other words, if this term was omitted, the structure would be more complex with more than 3 nodes having the same parent and an increase in the number of connections between nodes that can lead to difficulty in chemical interpretation as well.

For the clustered FTIR spectra from the thermal conversion of Athabasca bitumen, hill climbing and Tabu search gave the same underlying network structure that is depicted in Figure 6.18, while MMHC gave a similar network structure but with some disconnected groups as shown in Figure 6.19. The network structures given in Figure 6.18 and Figure 6.19 are called directed acyclic graphs (DAGs) and the arrows in the structures indicate the direction of information flow from the cause to effect, i.e. from the parent to the descendant node. The thickness of the arrows expresses the strength of the relationship and the magnitude of their strength is given by the negative number on top of the arrows as shown in both the figures.

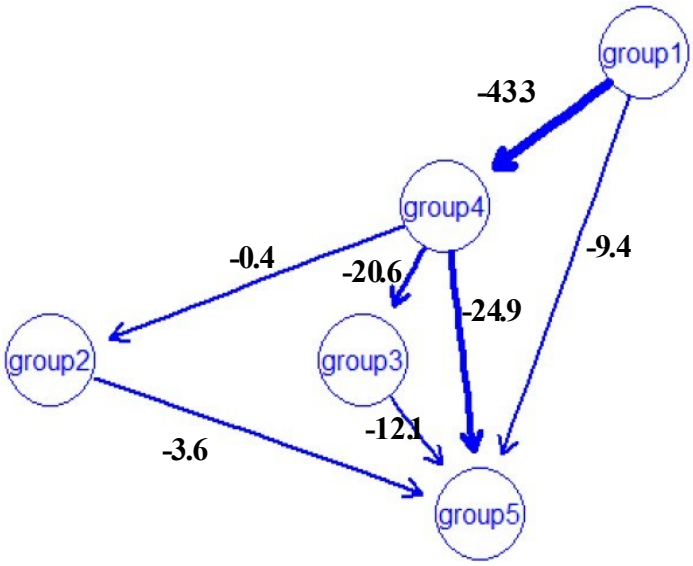


Figure 6.18. Bayesian network structure produced by hill climbing and Tabu search algorithms describing the causal relationships between different groups clustered by BHC.

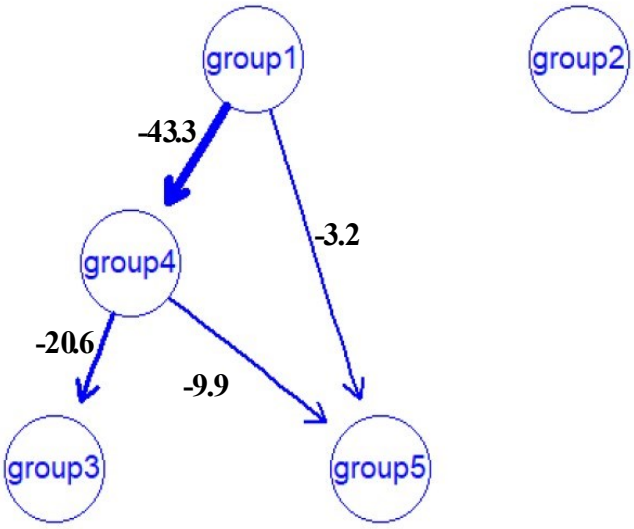


Figure 6.19. Bayesian network structure given by maximum-minimum hill climbing (MMHC) algorithm describing the causal relationships between different groups clustered by BHC.

A higher negative value of the arc strength meant that the existence of the cause-effect pair in the Bayesian structure is more probable. Put differently, if a particular connection is removed from the network, the score of the network will be affected in a negative manner by the value of the arc strength of the connection. The connection from Group 1 (G1) to Group 4 (G4) was the strongest in the networks produced by all three algorithms (arc strength of -43.3) but the next most probable

path was from G4 to Group 5 (G5) in the Tabu search and hill climbing-produced networks (Figure 18 – arc strength of -24.9) while G4 to Group 3 (G3) was the second strongest with G4 to G5 not far behind in the MMHC-produced network (Figure 6.19 – arc strengths of -20.6 and -9.9, respectively). The least probable connections involved Group 2 (G2) in both networks where G4 to G2 had an arc strength of -0.4 and that of G2 to G5 was -3.6. This was reflected in the MMHC-produced network (Figure 6.19) where G2 was classified as a separate node with no parent or descendent. Another interesting aspect to note was that G5 had 4 families (parents) in the Tabu search-produced network but only 2 families in the network structure given by MMHC. Considering the Bayesian network structures produced by the three algorithms as well as the wavenumbers associated with each node, a reaction network was devised, the details of which are provided in the next section.

6.4.6.2 Proposed reaction pathway

Though two algorithms (hill climbing and Tabu search) resulted in the same network (referred to as network 1 – Figure 6.18), it was also worthwhile to compare the proposed chemical reaction pathway against the network produced by MMHC (network 2 – Figure 6.19). The proposed reaction pathways as illustrated in Figure 6.20 and Figure 6.21 could also be validated with the types of reactions that deduced from the SMCR results in section 6.4.3. The main difference is that this reaction network is not specific to a particular experimental condition but is a generic portrayal of the thermal conversion of Athabasca bitumen in the temperature range of 300 – 420 ºC and based on the causal relationships given by the Bayesian network structures.

The most probable route was G1 to G4 which represented the conversion of di- or tri-substituted aromatics to aliphatic products (Table 6.13). The starting compound was considered to be a tricyclic naphthene-aromatic compound (compound (29) in Figure 6.20) with the alkyl substituent on the naphthene ring. This was the same as compound (1) in Tefera et al. [30] since it satisfied the characteristic absorptions of the first group. This molecule was a good representation of G1 in the network structure as indicated in Figure 6.20 as well. This can crack at the indicated position to give compound (30) that has a free radical at the benzylic and the secondary carbon. Hydrogen can stabilize these free radicals to give compound (31), whose side chain can crack either at the benzylic carbon to give an aliphatic alkene or at the carbon next to the R group.

Intramolecular hydrogen transfer from the benzylic carbon occurs to give the aliphatic molecule (33) and the di-substituted naphthalene with a conjugated side chain (32). Compound (33) is representative of G4 since the -R group can have methylene groups and terminal methyl could be present as well. Thus, G1 → G4 is possible and highly probable but need not necessarily happen in a single step.

For the conversion of Athabasaca bitumen at 400 ºC conducted by Sivaramakrishnan et al. [24] in Chapter 5, ^{1}H-NMR data indicated a net increase of 7 %wt. in the aromatic content from 9 – 16 %wt. over 1440 min along with the formation of lighter molecules that contributed to a decrease in viscosity. The probable pathway from G1 to G4 as described, indicates that an increase in the lighter aliphatic molecules (33) cannot occur without the formation of aromatic compounds (32). Also, in this case, the substituted naphthalene is still di-substituted with respect to the left ring as it was in the starting compound (29). This was similar to the mechanism proposed at 380 ºC (Figure 6.12), where the di-substituted ring was not affected but side-chain cracking occurred. In contrast, for the conversion of Cold Lake bitumen at 400 ºC, Zachariah and De Klerk [32] reported a larger increase of aromatic content from 9 to 20 %wt. over 360 min, which translated to a higher decrease in the aliphatic character of the liquid products as compared to Athabasca bitumen. The lower content of lighter cracked products might not have helped with decreasing the viscosity as it was reported to increase after the first 30 min to be almost equivalent to that of the feed after 360 min at 400 ºC for Cold Lake bitumen conversion. [32]

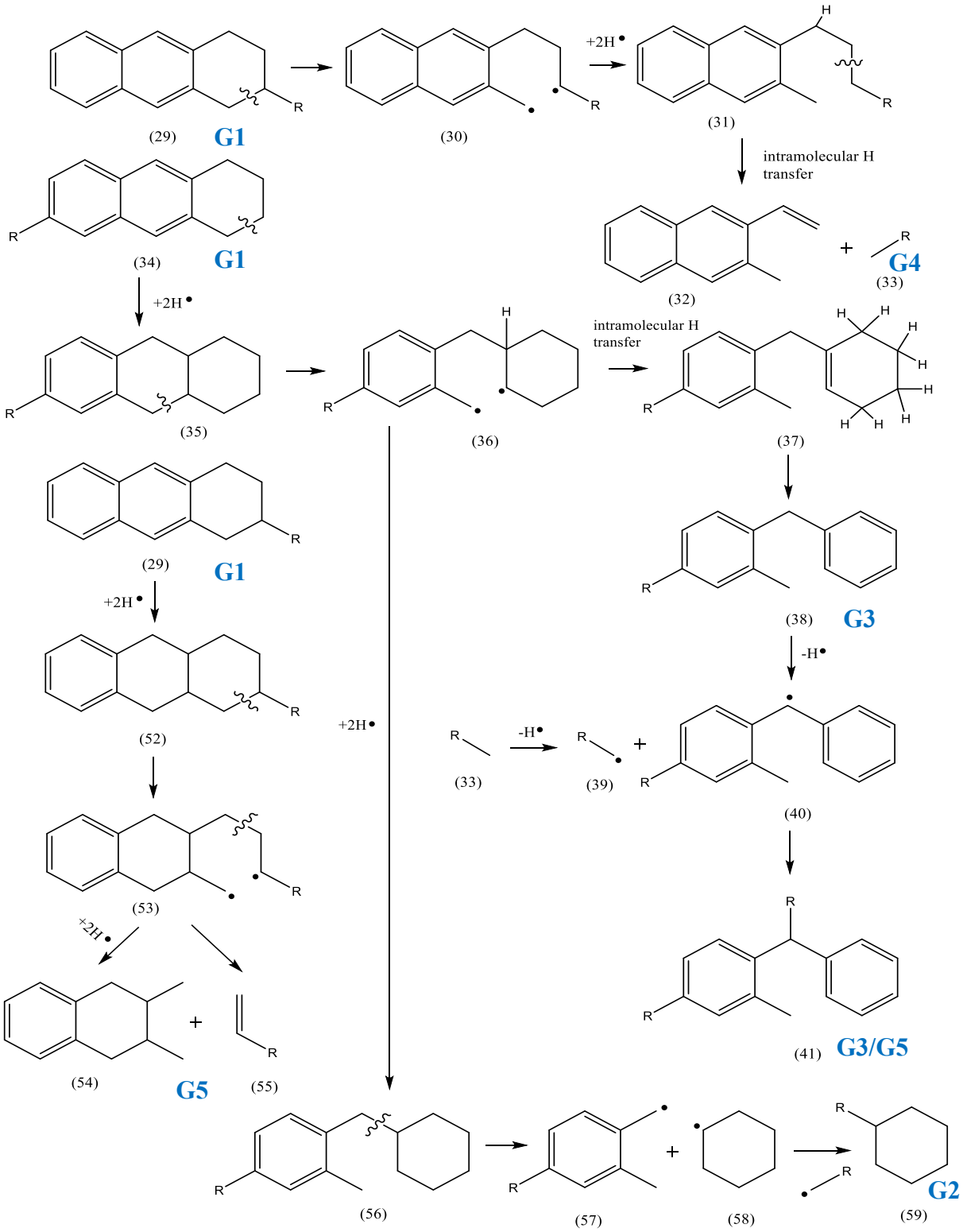


Figure 6.20. Proposed reaction pathway for the thermal conversion of Athabasca bitumen based on the results from Bayesian network structure.

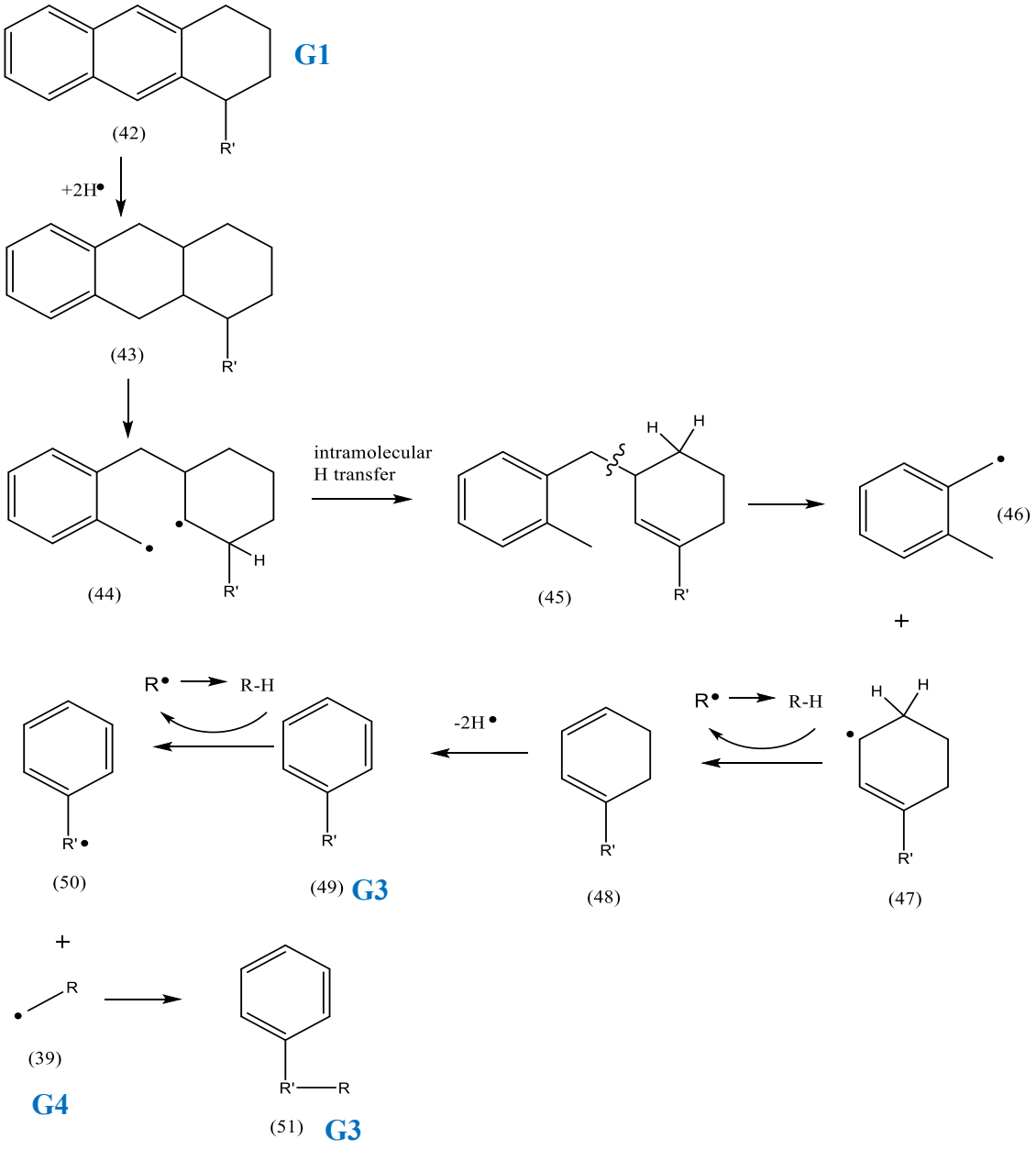


Figure 6.21. Proposed reaction pathway in continuation with Figure 6.20 for the thermal conversion of Athabasca bitumen based on the results from Bayesian network structure.

Though reaction network through Bayesian structure indicated a clear shift towards aliphatic product formation (Group 5 in Tefera et al. [30]), the simultaneous aromatic compound formation cannot be ignored. It can only be stated that the relative formation of aliphatic compounds in Athabasca were more than that for Cold Lake and also some aromatic products were less heavier

than those produced from Cold Lake due to the decreased substitution as shown in Figure 6.21 and discussed below.

Moving ahead, we consider the tri-substituted aromatic (34) as the starting compound representing G1 as described in section 6.4.5. Hydrogen radicals can add to the central aromatic ring to give (35) that can crack at the benzylic carbon to produce compound (36) which has free radicals at the favorable benzylic carbon and the secondary carbon centres as indicated in Figure 6.20. There are two ways in which this compound can lead to mono-substituted aromatics. It is very similar in structure to (10) shown in Figure 6.10 for 350 ºC and can undergo elimination by removal of the hydrogen attached to the tertiary carbon, which adds to the benzylic carbon to yield the alkene-aromatic (37). Four hydrogens can be removed through hydrogen disproportionation from (37) to form (38) that has mono-substituted aromatic absorptions. Essentially, the mono-substituted aromatic content is increased keeping the tri-substituted aromatic content constant without breaking of the aryl C-alkyl C bond. Furthermore, free-radical recombination was also an important reaction during thermal cracking as highlighted in the SMCR discussion (section 6.4.3). The benzylic hydrogen with respect to both aromatic rings in (38) can be abstracted and this free radical can re-combine with the free radical form of (33) to yield (41), which has characteristic absorptions of G3 (Table 6.13). Hydrogen can be abstracted not only from the primary carbon in (33) but also from secondary or tertiary if the R group is branched. The pathway from (33) to (41) can be considered as an explanation for G4 → G3 (3^{rd} and 2^{nd} strongest in network 1 and network 2, respectively), which was equally probable in the Bayesian networks revealed by all three algorithms (section 6.4.6.1).

The only concern with molecule (41) is that it has the wavenumbers included in G1 arising from the already existing tri-substituted aromatic. Hence, if the reaction sequence described in the above paragraph started with compound (29) instead of (34) and the intermediate radical (36) underwent the RHT mechanism through intramolecular hydrogen transfer from the hydrogen attached to the tertiary carbon, toluene would have been formed similar to that shown in Figure 6.14 at 400 ºC. But since this is seen as a rare occurrence at 400 ºC and may be possible at 420 ºC, it is not shown in the proposed reaction networks. However, it should be kept in mind that molecules similar to (37) are easier to crack at the Ar-C-alkyl C bond than for (29) or (35). As alternative route to RHT as suggested at 400 ºC (Figure 6.14), the sequence shown in Figure 6.21 seemed to be a more feasible pathway to produce mono-substituted aromatics where the starting

compound was (42) with the substituent in the benzylic position on the naphthene ring. It was interesting to note that though compounds (45) and (37) were formed through the same process, the possibility of (45) cracking at the indicated position was far more than (37) due to the position of the double bond facilitated by the presence of the R' group. This would lead to the stable free-radical (46), while (47) can undergo hydrogen abstraction and disproportionation to yield the mono-substituted aromatic (49). The benzylic hydrogen in (49), if any, can be abstracted and combined with (39) to give (51), which is a good example of G3. Thus, this can be a more feasible pathway for G4 → G3 than (33) to (41). The practicality of this pathway emerges from the fact that it did not require an aryl C-alkyl C bond to be broken, which could be quite difficult at these lower temperatures.

Interestingly, compound (41) can also be considered as belonging to G5 as this group has a mixture of wavenumbers that contain absorption bands related to aromatics and aliphatics (section 6.4.5). If this was true, G4 to G5 could be explained using the path from (33) to (41) and this relationship was the second strongest in network 1 (Figure 6.18) and third strongest in network 2 (Figure 6.19). It was interesting to see that G3 → G5 had a lower arc strength than other causal relationships explained above for network 1 but this route did not exist in network 2. Since compound (38) can be considered as belonging to G3, the pathway from (38) → (40) → (41) can be explanative of G3 → G5 involving G4 as well. It has to be noted that though (40) is readily formed from (38), it can be difficult for it to combine with (39) if the R contains branched moieties due to steric hindrance from the two aromatic groups on either side. Meanwhile, (39) can combine more easily with less sterically hindered radicals to yield the G5 class of molecules. The difficulty of G3 → G5 was captured better in network 2 more than in network 1. These type of recombination reactions (G4, G3 → G5, G4 → G3) involving free radicals can lead to an increase in chain length that results in the rise of the nCH_2/nCH_3 parameter that was seen to occur at 380 ºC, 400 ºC from S_2 to S_3 at later reaction times from the SMCR results (Table 6.9, Table 6.10). The pathway from compound (33) to (51) also signals a decrease in EOS which was found to occur at all temperatures except at 380 ºC in the SMCR results.

Hydrogen radicals can also add to (29) in a similar way as proposed to add to (34) to give (52) that is equivalent to (35) except that the R group is on the naphthene ring. Cracking of (52) at the indicated position (Figure 6.20) and subsequent hydrogen radical stabilization can yield alkenes like (55) along with naphthene-aromatics (54). The mixture of these molecules, especially the

alkene (55), can be another representation of G5 due to the high effective intensity of alkene for this cluster as determined in section 4.5. Thus, (29) to (55) can be seen as a pathway from G1 → G5 whose arc strength was lesser than G3 → G5, but definitely more than that involving G2 species (Figure 6.18). This pathway involves cracking and free-radical stabilization by hydrogen to produce lighter aliphatics and similar to the mechanism thought to occur at higher temperatures like 420 ºC (section 6.4.3.5).

Moreover, the free-radical (36), when stabilized by hydrogen addition, can crack at the favorable benzylic position to yield (57) and the cyclohexyl radical (58). On hydrogen addition, (58) gets converted to cyclohexane, which is an ideal representation of G2. Since (36) originates from G1, this pathway represents G1 → G2, which does not exist in either of the Bayesian networks calculated in section 6.4.6.1. The only family for G2 was G4 according to network 1, and this could be possible only through free-radical recombination of (39) and (58). The product (59) is a very good representation of a G2 group only if the R group has a terminal double bonded CH_2. On a parallel note, the removal of hydrogen from a cyclohexane is much more difficult than from a compound like (38) since the free-radical in (40) is stabilized by 2 aromatic rings whereas there is no group to stabilize the cyclohexyl radical. The Gibbs free energy of formation of cycloalkanes and naphthenes is also quite low due to the stable chair and boat conformations. [134] Hence, the formation of (58) back from cyclohexane is difficult, which makes G4 → G2 less probable. Furthermore, (59) can undergo further cracking at the existing alkyl group (R) if it is longer than 2 carbons. The resulting naphthenic molecule can be representative of G5 in the network, which gives the only possible pathway from G2 → G5. But since the former reaction is in itself less likely, the production of G5 compounds from G2 is even more less plausible due to the stability of stand-alone naphthenic molecules.

The entire reaction network is summarized in Table 6.14 including the major types of reactions occurring and their similarities to the ones proposed for the individual temperatures from the SMCR results. In view of the discussion above, it can be concluded that the derived reaction networks by all three algorithms made chemical sense, with the MMHC-produced network (network 2) being able to extract the key elements of the reaction pathway more accurately. The network produced by MMHC classified G2 as a separate entity, which was practical due to the stability of the naphthenes as discussed above. Also, G3 → G5 being less probable was attributed to steric hindrance in the bulky aromatic molecules. Finally, due to multiple pathways for the

formation of mono-substituted aromatics from the lighter aliphatic compounds, the stronger arc strength of G4 → G3 than G4 → G5 in the MMHC-produced network made more chemical sense. This is an example of a disconnected network (Figure 6.19) relating to the real chemical scenario more accurately than the continuous network (Figure 6.18). It was also seen that the Bayesian network structure corroborated the SMCR results fairly well and the plausible types of reactions discussed in section 6.4.3 as well, which was one of the main objectives of this work. It should still be kept in mind that these chemometric tools provide a means to arrive at certain hypotheses regarding the occurrence of different reactions that needed further experimental proof, which is out of scope for this chapter.

Table 6.14. Main reaction mechanisms associated with the reaction pathways derived from the Bayesian network structure.

Causal relationship	**Related compound transformations**	**Constituting reaction mechanisms**	**Temperature(s) of similar occurrence from SMCR**
G1 → G4	(29) → (33) (Figure 6.20)	Cracking, intramolecular H transfer	350 ºC 380 ºC
G4 → G3	(i) (42) → (49) (ii) (39) → (51) (Figure 6.21)	(i) H transfer to MNA; Intramolecular H transfer; Cracking, H abstraction; (ii) Free-radical recombination	380 ºC 400 ºC
G4 → G5	(33) → (41) (Figure 6.20)	Free-radical recombination	-
G3 → G5	(38) → (41) (Figure 6.20)	H abstraction; Free-radical recombination	-
G1 → G5	(29) → (55) (Figure 6.20)	H transfer to MNA; Cracking; H stabilization of free-radicals	420 ºC
G4 → G2	(i) (34) → (56) → (58); (ii) (39) + (58) → (59) (Figure 6.20)	(i) H transfer to MNA; Cracking, H stabilization; (ii) Free-radical recombination	

6.4.7 Global model for SMCR compared to Bayesian network

The global model incorporated datasets at all temperatures together, which would mean a column-wise augmented data matrix with the different sets of experiments analyzed by the same technique (FTIR). This data was analyzed with SMCR-ALS and the results were compared with that of the temperature-wise model and that of the Bayesian networks for two reasons: (i) to see the improvement in the resolution of the concentration and spectral profiles due to the larger number of samples; (ii) to verify whether the proposed reaction pathway from the Bayesian network fits the trends described by the quantitative parameters calculated from the resolved spectra in the global model. If the results of the global model were consistent and satisfactory, it could be useful in real-time control of the system. For example, in the case of thermal conversion of bitumen in a continuous process, it is most likely that the temperature would be continuously varied after a certain amount of time at each temperature. A local SMCR model would require more computational effort since it has to switch between every process condition from time to time, whereas a global model can skip this step and directly operate on the spectra of the liquid product.

Only a brief explanation of each part of the process, namely identification of chemical rank, estimating the initial concentration for the pseudo-components and the final resolved concentration and spectra are provided in this section. The number of components extracted for the global model were 3 since the ROD exhibited a maximum value of 22.3 when three principal factors were used. Figure D.19a in Appendix D shows the plot of ROD varying with the number of components for the 35 samples. The LOF as calculated from the residual obtained after performing SVD using 3 pseudo-components (equation 6.24) was 7.108 and was lower than when 2 components were used, as expected. The amount of variance explained was 99.49 % and there was not much difference to the R^2 when more than 3 components were added to the model.

Figure D.19b shows the plot for initial concentration estimates at all the process conditions considered in the global model. The overall trends in S_1 and S_3 appeared to be decreasing and increasing, respectively, with the trend for S_2 being similar to that in the SMCR models for the individual temperatures as well (Figure 6.5). Interestingly, S_1 showed a step-wise decrease in what seemed to happen at the start of every new dataset at each temperature that was augmented with

the previous one. This nature wasn't seen in the final resolved concentration profiles though the overall trends were the same as shown in Figure 6.22a. A fair adherence to the closure constraint was also seen with the sum of concentrations of the three pseudo-components for the datapoints 1-10 and 27-35 being between 1 and 1.2 while it reached a maximum of 1.6 for the 11th datapoint. This was an improvement over the ALS profiles in the local model where the summation reached 1.7 at two of the points for the 400 ºC dataset.

Figure 6.22 shows the concentration and spectral profiles resolved using ALS algorithm for the dataset considering all temperatures together. The quantitative derived parameters from the ALS-resolved spectral profiles for the global model is given in Table 6.15.

Table 6.15. Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number for the dataset comprising all experimental conditions.

Pseudo-component	**1**	**2**	**3**
n-CH_2/n-CH_3	1.72	1.22	1.27
Overall EOS	4.83	0.30	1.51
DOC (C=C stretch/C-H stretch wavenumber)	2.53	1.87	2.03
COC value	0.75	0.00	0.00

As seen from Table 6.15, $n$$CH_2$/$n$$CH_3$ showed an overall decrease that was similar to the trends from the SMCR results at all the individual temperatures except 300 ºC. There was a minimum at S_2 for this parameter, which then increased slightly to 1.27 for S_3. This trend seemed to capture the changes observed at higher temperatures (380 ºC, 400 ºC, 420 ºC) in the individual data sets from the local SMCR model and indicated that cracking occurred significantly with some amount of methyl transfer and free-radical recombination at larger reaction times. As shown in the reaction network derived based on the Bayesian results (Figure 6.20, Figure 6.21 and Table 6.14), hydrogen transfer to multi-nuclear aromatics also played an important role in reducing the -CH_2 content, which could lead to the observed decrease in the $n$$CH_2$/$n$$CH_3$ parameter. Another parameter that appeared to be consistent with the SMCR results for the local model was EOS, which showed a drastic decrease from S_1 to S_2 and then an increase from S_2 to S_3. The only temperature where EOS

increased overall was at 380 ºC, for which the related plausible reaction chemistry was explained in section 6.4.3.3. Also, the sharp reduction in EOS for S_2 was also observed at 400 ºC and explained by relaxing the number of components to 4 (section 6.4.3.4). Sivaramakrishnan et al. [24] (Chapter 5) reported an increase in the mono-substituted aromatic content during the thermal cracking of Athabasca bitumen conducted at 400 ºC and the increase in the absorption intensity at 727 cm^{-1} for S_2 and S_3 (Figure 6.22c) was also supportive of this observation. The multiple ways of obtaining mono-substituted aromatics from higher substituted aromatics were also discussed in the previous section on the proposed reaction network.

Although thought to be a relatively unreliable parameter, the overall decrease in DOC could have resulted from a higher increase in the number of aromatic hydrogens as compared to the C=C. This decrease in the condensation extent could also mean a conversion of higher substituted aromatics to lower substituted aromatics, which was indicated in the reaction network proposed (Figure 6.20 and Figure 6.21). Lastly, the resolution of the spectral profiles (Figure 6.22b-e) were good with no noisy spectra as was obtained in the 1800 – 1500 cm^{-1} in the 300 ºC and 350 ºC ALS-optimized profiles due to the presence of limited number of datapoints. Though there were no issues with the resolution quality of the spectral profiles at other individual temperatures and ALS-PSO method was also useful in reducing the noise, it is always useful to obtain a larger number of datapoints for analysis of a particular dataset when chemometric tools like curve resolution are used.

Since the global SMCR-ALS model considered all experimental conditions together, only generalized comments could be made with regards to chemical changes during thermal conversion, though most changes were able to be captured. This illustrates the importance of investigating the spectra at each temperature separately, though the global model is better applicable for online monitoring of continuous processes.

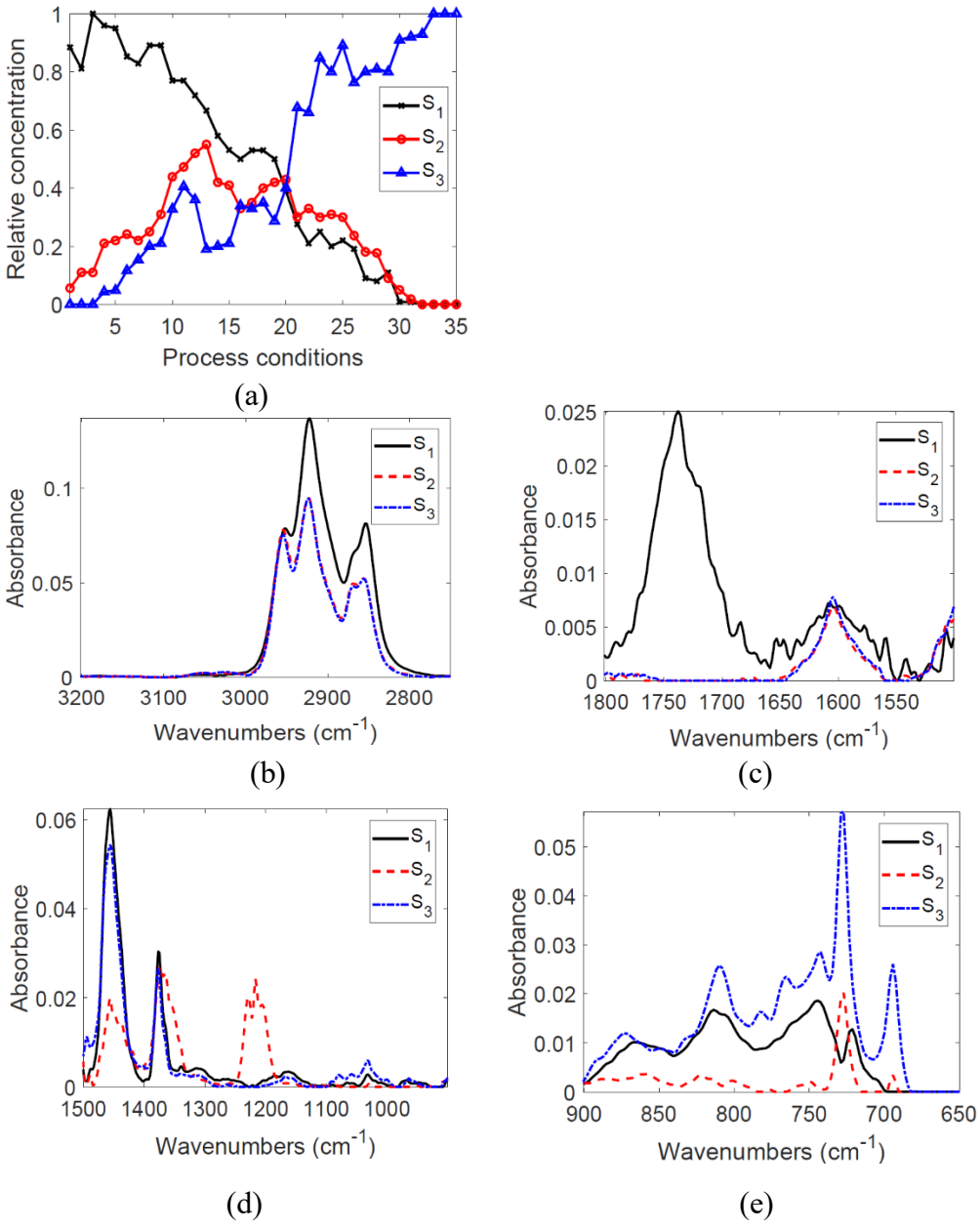


Figure 6.22. Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at temperatures in the range 300 – 420 ºC (global model). The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; and resolved spectra for each pseudo-component shows as absorbance vs. wavenumber in the ranges: (b) 3200 – 2750 cm^{-1}; (c) 1800 – 1500 cm^{1}; (d) 1500 – 900 cm^{-1}; (e) 900 – 650 cm^{-1}.

6.5 Implications on low temperature partial upgrading

A number of comparisons were made between Athabasca and Cold Lake bitumen in terms of probable reaction chemistry during thermal conversion at conditions that were less severe than those employed in the industry. [37] From all the proposed mechanisms in this work, some relevant questions regarding their implications to partial upgrading remain. Which type of bitumen is more processable? Are there any particular preferences for certain products from thermal cracking? Do these products interact with the diluent added during pipeline transportation? This section attempts to discuss some of these aspects and it must be noted that these are partly speculative and not specific answers or solutions to the above questions.

The major goal of partial upgrading is to meet the pipeline specifications that hold restrictions on the viscosity, density and the olefin content. [135] At the same time, partial upgrading by thermal conversion processes like visbreaking does not eliminate the need for diluent completely and just decrease the requirement from ~35 %v/v for the raw bitumen to ~15%v/v for the low-temperature cracked bitumen. [32] It is important for partial upgraders to consider whether a commonly used diluent like naphtha has any chemical interaction with thermally cracked bitumen, which can further impact physical properties like viscosity in addition to the dilution effect. It has been highlighted in this work that the formation of lighter products in both aliphatics (excluding gases) and lower-substituted aromatics from thermal conversion can serve as an internal diluent to decrease the viscosity. However, free-radical recombination possibly occurred at longer reaction times but did not have a major influence on the viscosity as it continuously decreased at all temperatures for Athabasca bitumen. Since the diluent is added after thermal conversion, the conditions are not expected to be conducive for a chemical reaction between the naphtha molecules and bitumen. Naphtha, by itself, was not shown to contain any free radicals [136] but thermally cracked bitumen does. Molecules like the aliphatic alkene free radical (26) in Figure 6.14 and the benzylic free-radical (57) in Figure 6.20 are stabilized by resonance but can definitely interact with double bonded naphtha molecules and cause polymerization at the pipeline temperature.

To compare the ease of processability of the two types of bitumen from just the results of this work and that of Tefera et al. [29] on Cold Lake bitumen is difficult. Being processable can imply easier reduction of viscosity at less severe upgrading conditions so that a lower amount of diluent can be used. The composition of thermally cracked bitumen plays a key role both in mid-stream

and downstream processing. The thermal conversion of Athabasca bitumen was shown to reduce the viscosity monotonically at all reaction times while that of Cold Lake bitumen experienced an increase in viscosity at later reaction times. This might tempt us to state that Athabasca bitumen is more processable but it is not only the final viscosity that is important. Lighter aliphatic and mono-substituted aromatic molecules like (33), (55) in Figure 6.20, (28) in Figure 6.14 and (51) in Figure 6.21 would not cause much problems but heavier compounds similar to (19) in Figure 6.12, (54), (38) and (41) in Figure 6.20 that include condensed aromatics with 2 or more rings, and aromatics with bulkier groups attached can cause clogging or even form a separate phase and precipitate even due to minute changes in pipeline pressures. Polynuclear aromatic compounds formed from ring closure reactions like the one shown in Figure 11 in Tefera et al. [29] can also have a deleterious effect for partial upgrading. Optimizing the process conditions with knowledge of reaction chemistry can help in monitoring the composition of thermally cracked bitumen and mitigating the undesired effects during pipeline transport. Chemometric tools like those investigated in this work help in developing deeper insights into possible product composition with minimal input information, and thus can aid in choosing the appropriate process condition for partial upgrading.

6.6 Conclusions

This work dealt with the application of chemometric tools on the FTIR spectra of liquid products obtained during thermal conversion of Athabasca bitumen in the temperature range of 300 – 420 ºC. The objective was to develop a credible reaction network for the thermal cracking process based on the results of the statistical approaches while using minimum prior chemical knowledge of the system. Similar studies were conducted previously on Cold Lake bitumen and the differences in chemometric results and subsequent chemical interpretation on the reaction chemistry of the two types of bitumen were also highlighted in this work. The chemometric methods utilized in this work comprised of two parts: (i) Self-modeling multivariate curve resolution (SMCR) using two optimization algorithms namely ALS and PSO embedded in ALS applied on 5 datasets at temperatures of 300 ºC, 350 ºC, 380 ºC, 400 ºC, 420 ºC separately (local model) as well as augmenting them together (global model); (ii) Bayesian hierarchical clustering (BHC) on the preprocessed FTIR spectra followed by Bayesian network learning to reveal the

underlying causal relationships between the clustered groups from BHC. The key conclusions are summarized as follows:

a. For both the local model and the global SMCR models, three pseudo-components were found to be representative of the bitumen reaction mixture as obtained from the maximum of ratio of derivatives (ROD) and second derivative (SD) values after performing singular value decomposition (SVD) on the respective matrices. More than 99.5 % of the variance was explained by 3 components.
b. Initial concentration estimates for SMCR-ALS and SMCR-ALS-PSO in the local models were obtained by FSMW-EFA and indicated a reaction pathway among the pseudo-components as $S_1 \rightarrow S_2 \rightarrow S_3$.
c. As suggested in previous works, the ALS-PSO approach was better than the ALS method for the local model in a number of ways. The spectral and concentration profiles from the former showed improvements as compared to their ALS-optimized counterparts in terms of noise reduction and stricter adherence to the closure constraint. In addition, the residuals and LOF were slightly lower for the final resolved data when ALS-PSO was employed. The R^2 values were marginally higher to the third decimal for the ALS-PSO algorithm and it took half the number of iterations to converge to the final solution as compared to the ALS method.
d. The majority of the trends in the final resolved spectra-derived quantitative parameters like the intensity ratio of methylene and methyl groups (nCH$_2$/nCH$_3$ - indicating alkyl chain length or naphthenic content), extent of aromatic substitution (EOS), degree of condensation (DOC) and complex oxygenate content (COC – indicating estimate of ester and anhydride-type molecules) were identical for both ALS and ALS-PSO methods in the SMCR approach, though the absolute values were slightly different. This was also reflective of the same interpretation of the underlying reaction chemistry irrespective of the resolution algorithm used.
e. $S_1 \rightarrow S_2 \rightarrow S_3$ was the common conversion pathway among the three pseudo-components in the final concentration profiles from both algorithms for the local model. At all temperatures except at 300 ºC, S_2 existed during the intermediate reaction times while S_3 was the dominant component at larger reaction times.

f. No major chemical changes in Athabasca bitumen occurred at 300 ºC as compared to Cold Lake, where viscosity decreased despite increasing alkyl chain length and increased degree of aromatic condensation. The lower reactivity of Athabasca bitumen was believed to be due to the lower saturate content in the feed that could manifest as shorter chain length of the alkyl substituents. Methyl transfer was suggested to have occurred at 300 ºC. The cracking extent was higher at 350 ºC accompanied by methane gas formation and less methyl transfer. Since EOS showed an overall decrease, a pathway was devised that included cracking, hydrogen abstraction and disproportionation to form mono-substituted aromatics without affecting di-substituted aromatic content. In contrast, EOS increased for Cold Lake cracking at 340 ºC, indicating ring-closure reactions. At 380 ºC, EOS increased for S_3 with respect to the feed (S_1) but the chain length parameter and DOC decreased. It was speculated that existing naphthene-aromatic compounds were converted to a complete aromatic molecule through hydrogen disproportionation and this facilitated cracking of the side chains at later times. It was noted that literature had mixed opinions on the occurrence of hydrogen disproportionation at temperatures < 400 ºC. Significant differences in chemical changes were observed at 400 ºC for both types of bitumen. The formation of lighter products resulted in a continuous decrease in viscosity for Athabasca bitumen and was reflected in the larger overall decrease in all quantitative parameters. Due to an extreme decrease in EOS (confirmed by relaxing number of components), the possibility of scission of Ar-C-alkyl C bond was explored to form mono-substituted aromatics. Free-radical recombination was an important reaction at larger reaction times but the rate of recombination and condensation reactions were thought to be higher in Cold Lake than in Athabasca at 400 ºC due to differing trends in viscosity. Severe cracking was seen to occur at 420 ºC for Athabasca bitumen with stabilization of side chains by hydrogen.

g. BHC yielded 5 clusters that could be classified into *o*-, *m*-, *p*-di- or tri-substituted aromatics, naphthenic, monosubstituted aromatics, aliphatics and a mixture of aromatic, aliphatic alkene and heteroatomic compounds. Tabu search and hill climbing algorithms for Bayesian learning provided the same underlying network structure but MMHC identified the second cluster (G2) as a separate group with no connection and a reduced number of families for the fifth cluster. G1 → G4 → G3 was the strongest connection in all three networks and the chemical interpretation was that higher substituted aromatics

cracked to lighter aliphatics and a pathway was also devised to form mono-substituted aromatics without breaking the strong Ar-C-alkyl C bond. Free-radical recombination, stabilization by hydrogen radicals and hydrogen transfer (intra and intermolecular to and from aromatics) were key elements of the reaction network and were consistent with the proposed mechanisms from SMCR local models. Interestingly, the position of the alkyl substituent on the starting compound was also significant in determining the nature of the products. The network produced by MMHC was determined to be more accurate in representing the real thermal conversion chemistry. The reaction network for Cold Lake bitumen conversion also indicated the formation of lighter aliphatic products but the local viscosity maxima attained at some temperatures could not be explained. The accompanying increase in condensed aromatic content that did not precipitate as coke could have been responsible for this, but this needs to be quantified.

h. The global SMCR model showed the value of having more samples in the dataset and the change in quantitative parameters was in good agreement with the proposed reaction network from the Bayesian methods. The potential for online monitoring of a continuous thermal cracker was also discussed.

i. The run-time for the SMCR methods was much lesser than the Bayesian methods which would make the SMCR techniques more suitable for implementation in automation and control. But at the same time, the Bayesian network makes it easier to derive a reaction pathway and thus, it is recommended that the use of both approaches together for analyzing the reactions of complex mixtures like bitumen.

In conclusion, both chemometric strategies used in this study gave consistent results and enabled us to propose a reaction network for the thermal conversion of Athabasca bitumen and compare it with that of Cold Lake bitumen comprehensively. However, it should be noted that the proposed reaction chemistry is based on the results of the statistical methods and further experiments using the appropriate model compounds are needed to prove their occurrence in real time. The chemometric methods can also serve as hypotheses generators for experimental procedures involving complex molecules.

Lastly, though this study implemented the suggestions of previous works by applying PSO embedded with ALS, and checking the correlation of the results from Bayesian network learning

and curve resolution methods, certain drawbacks need to be highlighted. The number of samples in the temperature-wise datasets were limited and it was seen that the global model improved the resolution to some extent. PSO parameters like swarm size and lower and upper bounds can be varied and the model can be extended to lower temperatures while incorporating other characterization methods like ^{1}H-NMR, ESR and chromatography as well. Also, since the original data matrix is three directional consisting of time, temperature and spectral channels, a tri-linear decomposition method like PARAFAC combined with ALS optimization can be used. The advantage of PARAFAC over the traditional PCA is that it calculates all the contributing components simultaneously rather than in a step-wise manner and improves signal-to-noise ratio.

6.7 References

(1) Belgrave, J. D. M.; Moore, R. G.; Ursenbach, M. G.; Bennion, D. W. A Comprehensive Approach to In-Situ Combustion Modeling. *SPE Adv. Technol. Ser.* **1993**, *1*, 98–107.

(2) Shu, W. R.; Venkatesan, V. N. Kinetics of Thermal Visbreaking of a Cold Lake Bitumen. *J. Can. Pet. Technol.* **1984**, *23*, 60–64.

(3) Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs. *Energy Fuels* **2017**, *31*, 13509–13518.

(4) Fortuna, L.; Graziani, S.; Rizzo, A.; Xibilia, M. G. *Soft Sensors for Monitoring and Control of Industrial Processes*; Springer Science & Business Media, 2007.

(5) Fischer, H. R.; Cernescu, A. Relation of Chemical Composition to Asphalt Microstructure - Details and Properties of Micro-Structures in Bitumen as Seen by Thermal and Friction Force Microscopy and by Scanning near-Filed Optical Microscopy. *Fuel* **2015**, *153*, 628–633.

(6) Hou, Y.; Wang, L.; Wang, D.; Guo, M.; Liu, P.; Yu, J. Characterization of Bitumen Micro-Mechanical Behaviors Using AFM, Phase Dynamics Theory and MD Simulation. *Materials (Basel).* **2017**, *10*, 1–16.

(7) Schuler, B.; Meyer, G.; Peña, D.; Mullins, O. C.; Gross, L. Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy. *J. Am. Chem. Soc.* **2015**, *137*, 9870–9876.

(8) Long, J.; Xu, Z.; Masliyah, J. H. Single Molecule Force Spectroscopy of Asphaltene Aggregates. *Langmuir* **2007**, *23*, 6182–6190.

(9) Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403–7429.

(10) Kumar, N.; Bansal, A.; Sarma, G. S.; Rawal, R. K. Chemometrics Tools Used in Analytical Chemistry: An Overview. *Talanta* **2014**, *123*, 186–199.

(11) Yoon, S.; Bhatt, S. D.; Lee, W.; Lee, H. Y.; Jeong, S. Y.; Baeg, J. O.; Lee, C. W. Separation and Characterization of Bitumen from Athabasca Oil Sand. *Korean J. Chem. Eng.* **2009**, *26*, 64–71.

(12) Adebiyi, F. M.; Thoss, V. Spectroscopic Characterization of Asphaltene Fraction of Nigerian Bitumen. *Pet. Sci. Technol.* **2015**, *33*, 245–255.

(13) Petersen, J. C. An Infrared Study of Hydrogen Bonding in Asphalt. *Fuel* **1967**, *46*, 295–305.

(14) Varanda, C.; Portugal, I.; Ribeiro, J.; Silva, C. M.; Silva, A. M. S. NMR Spectroscopy in Bitumen Characterization. In *Analytical Characterization Methods for Crude Oiland Related Products*; John Wiley & Sons Ltd: Chichester, UK, 2017; pp 141–161.

(15) Niizuma, S.; Steele, C. T.; Gunning, H. E.; Strausz, O. P. Electron Spin Resonance Study of Free Radicals in Athabasca Asphaltene. *Fuel* **1977**, *56*, 249–256.

(16) Yen, T. F.; Erdman, J. G.; Saraceno, A. J. Investigation of the Nature of Free Radicals in Petroleum Asphaltenes and Related Substances by Electron Spin Resonance. *Anal. Chem.* **1962**, *34*, 694–700.

(17) Ding, G.; Hou, Y.; Peng, J.; Shen, Y.; Jiang, M.; Bai, G. On-Line near-Infrared Spectroscopy Optimizing and Monitoring Biotransformation Process of γ-Aminobutyric Acid. *J. Pharm. Anal.* **2016**, *6*, 171–178.

(18) Blanco, M.; Villarroya, I. NIR Spectroscopy: A Rapid-Response Analytical Tool. *TrAC - Trends Anal. Chem.* **2002**, *21*, 240–250.

(19) Alves, J. C. L.; Henriques, C. B.; Poppi, R. J. Determination of Diesel Quality Parameters Using Support Vector Regression and near Infrared Spectroscopy for an In-Line Blending Optimizer System. *Fuel* **2012**, *97*, 710–717.

(20) Skibsted, E.; Engelsen, S. B. Spectroscopy for Process Analytical Technology (PAT). In *Encyclopedia of Spectroscopy and Spectrometry*; Elsevier, 2010; pp 2651–2661.

(21) Garrido, M.; Rius, F. X.; Larrechi, M. S. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) Applied to Spectroscopic Data from Monitoring Chemical Reactions Processes. *Anal. Bioanal. Chem.* **2008**, *390*, 2059–2066.

(22) Kapadia, P. R.; Kallos, M. S.; Gates, I. D. A New Kinetic Model for Pyrolysis of Athabasca Bitumen. *Can. J. Chem. Eng.* **2013**, *91*, 889–901.

(23) Selucky, M. L.; Chu, Y.; Ruo, T. C. S.; Strausz, O. P. Chemical Composition of Cold Lake Bitumen. *Fuel* **1978**, *57*, 9–16.

(24) Sivaramakrishnan, K.; De Klerk, A.; Prasad, V. Viscosity of Canadian Oilsands Bitumen and Its Modification by Thermal Conversion. In *Chemistry Solutions to Challenges in the Petroleum Industry*; Rahimi, P., Ovalles, C., Zhang, Y., Adams, J. J., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2019; Vol. 1320, pp 115–199.

(25) Willet, P. *Similarity and Clustering in Chemical Information*; Wiley: New York, 1987.

(26) Wang, Y.P.; Zou, Y.R.; Shi, J.T.; Shi, J. Review of the Chemometrics Application in Oil-Oil and Oil-Source Rock Correlations. *J. Nat. Gas Geosci.* **2018**, *3*, 217–232.

(27) Van de Vijver, R.; Devocht, B. R.; Van Geem, K. M.; Thybaut, J. W.; Marin, G. B. Challenges and Opportunities for Molecule-Based Management of Chemical Processes. *Curr. Opin. Chem. Eng.* **2016**, *13*, 142–149.

(28) Dellamorte, J. C.; Barteau, M. A.; Lauterbach, J. Opportunities for Catalyst Discovery and Development: Integrating Surface Science and Theory with High Throughput Methods. *Surf. Sci.* **2009**, *603*, 1770–1775.

(29) Tefera, D. T.; Agrawal, A.; Yañez Jaramillo, L. M.; De Klerk, A.; Prasad, V. Self-Modeling Multivariate Curve Resolution Model for Online Monitoring of Bitumen Conversion Using Infrared Spectroscopy. *Ind. Eng. Chem. Res.* **2017**, *56*, 10756–10769.

(30) Tefera, D. T.; Yañez Jaramillo, L. M.; Ranjan, R.; Li, C.; De Klerk, A.; Prasad, V. A Bayesian Learning Approach to Modeling Pseudoreaction Networks for Complex Reacting Systems: Application to the Mild Visbreaking of Bitumen. *Ind. Eng. Chem. Res.* **2017**, *56*, 1961–1970.

(31) Yañez Jaramillo, L. M.; De Klerk, A. Partial Upgrading of Bitumen by Thermal Conversion at 150-300 °C. *Energy Fuels* **2018**, *32*, 3299–3311.

(32) Zachariah, A.; De Klerk, A. Thermal Conversion Regimes for Oilsands Bitumen. *Energy Fuels* **2016**, *30*, 239–248.

(33) Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A. Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C. *Energy Fuels* **2014**, *28*, 5014–5022.

(34) Selucky, M. L.; Chu, Y.; Ruo, T.; Strausz, O. P. Chemical Composition of Athabasca Bitumen. *Fuel* **1977**, *56*, 369–381.

(35) McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. Heavy Petroleum Composition. 4. Asphaltene Compositional Space. *Energy Fuels* **2013**, *27*, 1257–1267.

(36) Speight, J. G. *The Chemistry and Technology of Petroleum*; Marcel Dekker: New York, 1991.

(37) Biasca, F. E.; Dickenson, R. L.; Chang, E.; Johnson, H. E.; Bailey, R. T.; Simbeck, D. R. Future Technology In Heavy Oil Processing. In *Upgrading heavy crude oils and residue to transportation fuel: Technology, Economics, and Outlook*; SFA Pacific Inc.: Palo Alto, CA, 2009.

(38) Tokarska, A. Investigations on the Processing of Oil Vacuum Residue and Its Mixtures with Coal and Coal Tars: Part 1. Primary Conversion of Crude Materials. *Fuel* **1996**, *75*, 1094–1100.

(39) Jolliffe, I. T. *Principal Component Analysis*; Springer: Berlin, Heidelberg, 2011.

(40) Elbergali, A.; Nygren, J.; Kubista, M. Automated Procedure to Predict the Number of Components in Spectroscopic Data. *Doktorsavhandlingar vid Chalmers Tek. Hogsk.* **1999**, *379*, 143–158.

(41) Malinowski, E. R. Determination of the Number of Factors and the Experimental Error in a Data Matrix. *Anal. Chem.* **1977**, *49*, 612–617.

(42) Keller, H. R.; Massart, D. L. Evolving Factor Analysis. *Chemom. Intell. Lab. Syst.* **1992**, *12*, 209–224.

(43) Malinowski, E. R. Window Factor Analysis: Theoretical Derivation and Application to Flow Injection Analysis Data. *J. Chemom.* **1992**, *6*, 29–40.

(44) Manne, R.; Shen, H.; Liang, Y. Subwindow Factor Analysis. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 171–176.

(45) Rubinstein, R. Y.; Kroese, D. P. *Simulation and the Monte Carlo Method*; John Wiley & Sons, 2016.

(46) Eberhart; Shi, Y. Particle Swarm Optimization: Developments, Applications and Resources. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat.*

No.01TH8546); IEEE: Seoul, 2001; Vol. 1, pp 81–86.

(47) Shinzawa, H.; Jiang, J.H.; Iwahashi, M.; Ozaki, Y. Robust Curve Fitting Method for Optical Spectra by Least Median Squares (LMedS) Estimator with Particle Swarm Optimization (PSO). *Anal. Sci.* **2007**, *23*, 781–785.

(48) Tadesse, M. G.; Sha, N.; Vannucci, M. Bayesian Variable Selection in Clustering High-Dimensional Data. *J. Am. Stat. Assoc.* **2005**, *100*, 602–617.

(49) Nia, V. P. Fast High-Dimensional Bayesian Classification and Clustering, EPFL, Lausanne, Switzerland, 2009.

(50) Zou, C.; Feng, J. Granger Causality vs. Dynamic Bayesian Network Inference: A Comparative Study. *BMC Bioinformatics* **2009**, *10*, 122.

(51) Selman, B.; Gomes, C. P. Hill-Climbing Search. In *Encyclopedia of Cognitive Science*; John Wiley & Sons, Ltd.: New York, 2006.

(52) Bai, X.; Padman, R. Tabu Search Enhanced Markov Blanket Classifier for High Dimensional Data Sets. In *The Next Wave in Computing, Optimization, and Decision Technologies*; Springer: Berlin, 2005; pp 337–354.

(53) Tsamardinos, I.; Brown, L. E.; Aliferis, C. F. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Mach. Learn.* **2006**, *65*, 31–78.

(54) Zachariah, A.; Wang, L.; Yang, S.; Prasad, V.; De Klerk, A. Suppression of Coke Formation during Bitumen Pyrolysis. *Energy Fuels* **2013**, *27*, 3061–3070.

(55) Beer's Law https://teaching.shu.ac.uk/hwb/chemistry/tutorials/molspec/beers1.htm.

(56) Press, W. H.; Teukolsky, S. A. Savitzky-Golay Smoothing Filters. *Comput. Phys.* **1990**, *4*, 669–672.

(57) van den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genomics* **2006**, *7*, 142.

(58) Eriksson, L. *Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*; Umetrics: Sweden, 1999.

(59) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J. C.; Jellema, R. H. Fusion of Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2005**, *77*, 6729–6736.

(60) Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon,

J. C.; Nicholson, J. K. Improved Analysis of Multivariate Data by Variable Stability Scaling: Application to NMR-Based Metabolic Profiling. *Anal. Chim. Acta* **2003**, *490*, 265–276.

(61) Bro, R. PARAFAC. Tutorial and Applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171.

(62) Duponchel, L.; Elmi-Rayaleh, W.; Ruckebusch, C.; Huvenne, J. P. Multivariate Curve Resolution Methods in Imaging Spectroscopy: Influence of Extraction Methods and Instrumental Perturbations. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2057–2067.

(63) Everitt, B. S.; Kroonenberg, P. M. *Three-Mode Principal Component Analysis: Theory and Applications*; DSWO Press: Leiden, The Netherlands, 1983.

(64) de Juan, A.; Jaumot, J.; Tauler, R. Multivariate Curve Resolution (MCR). Solving the Mixture Analysis Problem. *Anal. Methods* **2014**, *6*, 4964–4976.

(65) de Juan, A.; Rutan, S. C.; Tauler, R. Two-Way Data Analysis: Multivariate Curve Resolution-Iterative Resolution Methods. In *Comprehensive Chemometrics*; Elsevier: Oxford, 2009; pp 325–344.

(66) Sánchez, F. C.; Toft, J.; van den Bogaert, B.; Massart, D. L. Orthogonal Projection Approach Applied to Peak Purity Assessment. *Anal. Chem.* **1996**, *68*, 79–85.

(67) Sánchez, F. C.; van den Bogaert, B.; Rutan, S.; Massart, D. L. Multivariate Peak Purity Approaches. *Chemom. Intell. Lab. Syst.* **1996**, *34*, 139–171.

(68) Jiang, J. H.; Sasic, S.; Yu, R.Q.; Ozaki, Y. Resolution of Two-Way Data from Spectroscopic Monitoring of Reaction or Process Systems by Parallel Vector Analysis (PVA) and Window Factor Analysis (WFA): Inspection of the Effect of Mass Balance, Methods and Simulations. *J. Chemom.* **2003**, *17*, 186–197.

(69) Vandeginste, B. G. M.; Derks, W.; Kateman, G. Multicomponent Self-Modelling Curve Resolution in High-Performance Liquid Chromatography by Iterative Target Transformation Analysis. *Anal. Chim. Acta* **1985**, *173*, 253–264.

(70) Brown, S. D.; Tauler, R.; Walczak, B. *Comprehensive Chemometrics*; Elsevier: Oxford, 2009.

(71) De Leeuw, J. Block-Relaxation Algorithms in Statistics. In *Information Systems and Data Analysis*; Springer: Berlin, 1994; pp 308–324.

(72) Jiang, J. H.; Liang, Y.; Ozaki, Y. Principles and Methodologies in Self-Modeling Curve

Resolution. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 1–12.

(73) Shen, H.; Liang, Y.; Kvalheim, O. M.; Manne, R. Determination of Chemical Rank of Two-Way Data from Mixtures Using Subspace Comparisons. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 49–59.

(74) Hao, S. L.; Shao, L. M. Determining the Number of Principal Factors by Eigenvector Comparison of the Original Bi-Linear Data Matrix and the One Reconstructed from Key Variables. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 17–23.

(75) Malinowski, E. R. *Factor Analysis in Chemistry*; Wiley: Hoboken, NJ, 2002.

(76) Mandel, J. Use of the Singular Value Decomposition in Regression Analysis. *Am. Stat.* **1982**, *36*, 15–24.

(77) Wasim, M.; Brereton, R. G. Determination of the Number of Significant Components in Liquid Chromatography Nuclear Magnetic Resonance Spectroscopy. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 133–151.

(78) Windig, W.; Guilment, J. Interactive Self-Modeling Mixture Analysis. *Anal. Chem.* **1991**, *63*, 1425–1432.

(79) Malinowski, E. R. Obtaining the Key Set of Typical Vectors by Factor Analysis and Subsequent Isolation of Component Spectra. *Anal. Chim. Acta* **1982**, *134*, 129–137.

(80) Shinzawa, H.; Jiang, J.-H.; Iwahashi, M.; Noda, I.; Ozaki, Y. Self-Modeling Curve Resolution (SMCR) by Particle Swarm Optimization (PSO). *Anal. Chim. Acta* **2007**, *595*, 275–281.

(81) Workman, J. J.; Mobley, P. R.; Kowalski, B. R.; Bro, R. Review of Chemometrics Applied to Spectroscopy: 1985-95, Part I. *Appl. Spectrosc. Rev.* **1996**, *31*, 73–124.

(82) Tauler, R.; Maeder, M. Two-Way Data Analysis: Multivariate Curve Resolution – Error in Curve Resolution. In *Comprehensive Chemometrics*; Elsevier: Oxford, 2009; pp 345–363.

(83) Tauler, R.; Smilde, A.; Kowalski, B. Selectivity, Local Rank, Three-Way Data Analysis and Ambiguity in Multivariate Curve Resolution. *J. Chemom.* **1995**, *9*, 31–58.

(84) Sanchez, E.; Kowalski, B. R. Generalized Rank Annihilation Factor Analysis. *Anal. Chem.* **1986**, *58*, 496–499.

(85) Antunes, M. C.; J. Simão, J. E.; Duarte, A. C.; Tauler, R. Multivariate Curve Resolution of Overlapping Voltammetric Peaks: Quantitative Analysis of Binary and Quaternary Metal Mixtures. *Analyst* **2002**, *127*, 809–817.

(86) Byrd, R. H.; Gilbert, J. C.; Nocedal, J. A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming. *Math. Program.* **2000**, *89*, 149–185.

(87) Yen, J.; Liao, J. C.; Bogju Lee; Randolph, D. A Hybrid Approach to Modeling Metabolic Systems Using a Genetic Algorithm and Simplex Method. *IEEE Trans. Syst. Man Cybern. Part B* **1998**, *28*, 173–191.

(88) Gen, M.; Cheng, R. *Genetic Algorithms and Engineering Design*; John Wiley & Sons: New York, 1997.

(89) Bansal, J. C.; Singh, P. K.; Saraswat, M.; Verma, A.; Jadon, S. S.; Abraham, A. Inertia Weight Strategies in Particle Swarm Optimization. In *2011 Third World Congress on Nature and Biologically Inspired Computing*; IEEE: Salamanca, Spain, 2011; pp 633–640.

(90) Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In *IEEE International of first Conference on Neural Networks*; IEEE: Perth, 1995; pp 1942–1948.

(91) Shi, Y.; Eberhart, R. A Modified Particle Swarm Optimizer. In *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*; IEEE: Anchorage, AK, 2002; pp 69–73.

(92) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 2000.

(93) Kang, E. Y.; Ye, C.; Shpitser, I.; Eskin, E. Detecting the Presence and Absence of Causal Relationships between Expression of Yeast Genes with Very Few Samples. *J. Comput. Biol.* **2010**, *17*, 533–546.

(94) Triantafillou, S.; Lagani, V.; Heinze-Deml, C.; Schmidt, A.; Tegner, J.; Tsamardinos, I. Predicting Causal Relationships from Biological Data: Applying Automated Causal Discovery on Mass Cytometry Data of Human Immune Cells. *Sci. Rep.* **2017**, *7*, 1–11.

(95) Agrafiotis, D. K.; Bandyopadhyay, D.; Wegner, J. K.; Van Vlijmen, H. Recent Advances in Chemoinformatics. *J. Chem. Inf. Model.* **2007**, *47*, 1279–1293.

(96) Rangarajan, S.; Kaminski, T.; Van Wyk, E.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Algorithms of RING. *Comput. Chem. Eng.* **2014**, *64*, 124–137.

(97) Uusitalo, L. Advantages and Challenges of Bayesian Networks in Environmental Modelling. *Ecol. Modell.* **2007**, *203*, 312–318.

(98) Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.* **2000**, *7*, 601–620.

(99) Freedman, D.; Humphreys, P. Are There Algorithms That Discover Causal Structure? *Synthese* **1999**, *121*, 29–54.

(100) Lian, H. Sparse Bayesian Hierarchical Modeling of High-Dimensional Clustering Problems. *J. Multivar. Anal.* **2010**, *101*, 1728–1737.

(101) Heller, K. A.; Ghahramani, Z. Bayesian Hierarchical Clustering. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*; ACM Press: New York, New York, USA, 2005; pp 297–304.

(102) Heard, N. A.; Holmes, C. C.; Stephens, D. A. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes. *J. Am. Stat. Assoc.* **2006**, *101*, 18–29.

(103) Sivaramakrishnan, K.; Nie, J.; de Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, *57*, 13156–13176.

(104) Lauritzen, S. L. The EM Algorithm for Graphical Association Models with Missing Data. *Comput. Stat. Data Anal.* **1995**, *19*, 191–201. https://doi.org/10.1016/0167-9473(93)E0056-A.

(105) Berzan, C. An Exploration of Structure Learning in Bayesian Networks. PhD Thesis, Tufts University, 2012.

(106) Xu, Z.; Srihari, S. N. Bayesian Network Structure Learning Using Causality. In *2014 22nd International Conference on Pattern Recognition*; IEEE: Stockholm, Sweden, 2014; pp 3546–3551.

(107) Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques-Adaptive Computation and Machine Learning*; The MIT Press: Cambridge, MA, 2009.

(108) Leray, P.; Francois, O. *BNT Structure Learning Package: Documentation and Experiments*; 2006.

(109) Strausz, O. P.; Lown, E. M. *The Chemistry of Alberta Oil Sands, Bitumens and Heavy Oils*; Alberta Energy Research Institute: Calgary, AB, 2003.

(110) Craddock, P. R.; Le Doan, T. Van; Bake, K.; Polyakov, M.; Charsky, A. M.; Pomerantz, A. E. Evolution of Kerogen and Bitumen during Thermal Maturation via Semi-Open Pyrolysis Investigated by Infrared Spectroscopy. *Energy Fuels* **2015**, *29*, 2197–2210.

(111) Potts, R. H. Carboxylic Acids (Manufacture). In *Kirk-Othmer Encyclopedia of Chemical*

Technology; Wiley: New York, 1978; pp 835–845.

(112) Gray, M. R. *Upgrading Oilsands Bitumen and Heavy Oil*; The University of Alberta Press: Edmonton, Canada, 2015.

(113) Cronauer, D. C.; Jewell, D. M.; Shah, Y. T.; Modi, R. J. Mechanism and Kinetics of Selected Hydrogen Transfer Reactions Typical of Coal Liquefaction. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 153–162.

(114) Steacie, E. W. R. *Atomic and Free Radical Reactions. Vol. II*, 2nd ed.; Reinhold: New York, 1954.

(115) Billmers, R.; Griffith, L. L.; Stein, S. E. Hydrogen Transfer between Anthracene Structures. *J. Phys. Chem.* **1986**, *90*, 517–523.

(116) Arends, I. W. C. E.; Mulder, P. Study of Hydrogen Shuttling Reactions in Anthracene/9,10-Dihydroanthracene−Naphthyl-X Mixtures. *Energy Fuels* **1996**, *10*, 235–242.

(117) Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds.*; John Wiley & Sons: New York, 2014.

(118) Blanksby, S. J.; Ellison, G. B. Bond Dissociation Energies of Organic Molecules. *Acc. Chem. Res.* **2003**, *36*, 255–263.

(119) Jha, K. N.; Montgomery, D. S.; Strausz, O. P. Chemical Composition of Gases in Alberta Bitumens and in Low-Temperature Thermolysis of Oil Sand Asphaltenes and Maltenes. In *Oil Sand and Oil Shale Chemistry*; Strausz, O. P., Lown, E. M., Eds.; Verlag Chemie: New York, 1978; pp 33–54.

(120) Grzybowski, M.; Skonieczny, K.; Butenschön, H.; Gryko, D. T. Comparison of Oxidative Aromatic Coupling and the Scholl Reaction. *Angew. Chemie Int. Ed.* **2013**, *52*, 9900–9930.

(121) Khorasheh, F.; Gray, M. R. High-Pressure Thermal Cracking of n-Hexadecane in Aromatic Solvents. *Ind. Eng. Chem. Res.* **1993**, *32*, 1864–1876.

(122) Evdokimov, I. N. The Importance of Asphaltene Content in Petroleum II—Multi-Peak Viscosity Correlations. *Pet. Sci. Technol.* **2010**, *28*, 920–924.

(123) Bazyleva, A. B.; Hasan, M. A.; Fulem, M.; Becerra, M.; Shaw, J. M. Bitumen and Heavy Oil Rheological Properties: Reconciliation with Viscosity Measurements. *J. Chem. Eng. Data* **2010**, *55*, 1389–1397.

(124) Wiehe, I. A. A Phase-Separation Kinetic Model for Coke Formation. *Ind. Eng. Chem. Res.* **1993**, *32*, 2447–2454.

(125) Blanchard, C. M.; Gray, M. R. Free Radical Chain Reactions of Bitumen Residue. *ACS Div. Fuel Chem. Prepr.* **1997**, *42*, 137–141.

(126) Phillips, C. R.; Haidar, N. I.; Poon, Y. C. Kinetic Models for the Thermal Cracking of Athabasca Bitumen. *Fuel* **1985**, *64*, 678–691.

(127) Millour, J. P.; Moore, R. G.; Bennion, D. W.; Ursenbach, M. G.; Gie, D. N. A Simple Implicit Model for Thermal Cracking of Crude Oils. In *SPE Annual Technical Conference and Exhibition*; Society of Petroleum Engineers: Las Vegas, 1985; pp 1–15.

(128) Henderson, J. H.; Weber, L. Physical Upgrading of Heavy Crude Oils by the Application of Heat. *J. Can. Pet. Technol.* **1965**, *4*, 206–212.

(129) Wiehe, I. A. *Process Chemistry of Petroleum Macromolecules*; CRC Press: Boca Raton, FL, 2008.

(130) Noda, I. Determination of Two-Dimensional Correlation Spectra Using the Hilbert Transform. *Appl. Spectrosc.* **2000**, *54*, 994–999.

(131) Colthup, N.; Daly, L.; Wiberley, S. *Introduction to Infrared and Raman Spectroscopy*, 2nd Edn.; Academic Press: Boston, 1990.

(132) Beauchamp, J. Spectroscopy Tables https://www.cpp.edu/~psbeauchamp/pdf/spec_ir_nmr_spectra_tables.pdf (accessed Oct 12, 2018).

(133) Stagni, A.; Cuoci, A.; Frassoldati, A.; Faravelli, T.; Ranzi, E. Lumping and Reduction of Detailed Kinetic Schemes: An Effective Coupling. *Ind. Eng. Chem. Res.* **2013**, *53*, 9004–9016.

(134) Lide, D. R. *CRC Handbook of Chemistry and Physics*; CRC Press, 2004.

(135) de Klerk, A.; Gray, M. R.; Zerpa, N. Unconventional Oil and Gas. Oilsands. In *Future Energy: Improved, Sustainable and Clean Options for our Planet*; Elsevier: Amsterdam, 2013; pp 95–116.

(136) Uzcátegui, G.; Fong, S. Y.; De Klerk, A. Cracked Naphtha Reactivity: Effect of Free Radical Reactions. *Energy Fuels* **2018**, *32*, 5812–5823.

7. CONCLUSIONS

7.1 Introduction

With the ever-increasing demand for petrochemicals, there is a need for automating certain production processes for which the key requirement is online-monitoring. It should be noted that with the advancement of technology in the analytical characterization techniques employed, the amount of data generated is also immense and multi-dimensional. Understanding the chemical changes that occur during the ensuing reactions is another challenge due to molecular complexity of the reacting mixture. Chemometrics was identified as a useful tool to deal with huge volumes of multi-dimensional and noisy data and when integrated with the production process, it can result in efficient online monitoring and automation. The quantitative results from chemometric methods also help in deducing the reaction chemistry of the system in focus. Thus, both the quantitative and qualitative aspects of chemometrics are significant, i.e. the quantitative aspect deals with simplifying complex data and qualitative aspect deals with knowledge of the reaction chemistry.

Keeping this in mind, three particular systems of industrial significance were investigated in this study. The molecular complexity of both the feeds and reacting systems increased from the first to the third system and these were: (i) acid-catalyzed oligomerization of an aliphatic alkene, propylene; (ii) non-catalytic oxidation of a naphthene aromatic, tetralin; (iii) thermal conversion of a complex mixture of aromatic and non-aromatic hydrocarbons and heteroatoms, bitumen. The challenge in the first system was tracking the composition of the isomeric hydrocarbon products without the use of a kinetic or a reactor model as they were already explored in literature. Oxygen availability was known to control product selectivity in tetralin oxidation but identifying a single parameter that was responsible for the availability of oxygen in the liquid phase was critical for the up-scaling the design of flow reactors where mass transfer and hydrodynamics were important. The major challenge in the third system was non-monotonic changes in viscosity of bitumen as observed in previous works and explaining these observations through a fundamental understanding of the key features of the conversion chemistry of Athabasca bitumen was required from the perspective of partial upgrading.

The main conclusions of this entire study are provided in the next section.

7.2 Significant conclusions and findings

A review of the various chemometric tools suggested the usefulness of unsupervised and supervised classification and regression techniques like dimension reduction (PCA), pattern recognition (HCA), calibration and regression (SVM, PLSR, *i*-PLSR, simple and multiple linear regression), and resolution methods in modeling linear and nonlinear relationships in different disciplines of analytical science. Applications in some disciplines include material science, environmental and food science, medicine, biochemistry and petroleum.

In the first phase of this study, it was found that LS-SVM method employing the RBF kernel was the most suitable for predicting product concentrations in propylene oligomerization among other methods (PLSR, *i*-PLSR and LS-SVM employing a polynomial kernel). It was able to capture the nonlinearity in the spectrum-output property and reaction condition-output property relationships as well. It showed the ability to model different datasets by displaying the least tendency to overfit, the best learning efficiency when the number of samples was varied and the lowest run-time as compared to the other approaches mentioned above. The combination with PCA and HCA assisted in recognizing important groups of product streams which correlated with the wavenumber regions identified by i-PLSR as well. The identified clusters and wavenumber regions were roughly associated with the carbon numbers in the products and indications to the formation of large number of isomers in the C_6-C_{10} segment were evident from the plots of predicted concentrations.

The key learning from the first phase of this study was that machine learning techniques like support vector machine regression can be used to track the product composition in a moderately complex process effectively.

The second phase of this study was concerned with delineating the relative importance of each parameter related to either mass transfer or hydrodynamics of a two-phase tetralin oxidation system in a microfluidic reactor at 150 ºC and constant pressure of 90 kPa. In this process, multicollinearity among the input variables was identified through bivariate correlations and other diagnostics like correlation index and VIF and was seen to be a major problem. Among gas-liquid interfacial area, length of oxygen gas bubble and superficial velocity, which predicted product selectivity and tetralin conversion well, gas-liquid interfacial area was seen to be the best predictor

in explaining the output variance alone. This was achieved through two-variable MLR models and inspecting changes in regression coefficient estimates, their standard errors combined with statistical significance tests, partial correlations and standardized coefficients. It was also possible to devise a statistical pathway to predict selectivity and conversion from the only user-manipulated variable of tetralin injection flowrate to facilitate online monitoring of the process and save run-time.

The key implication of this study was that in order to build industrial-sized reactors like loop reactors for hydrocarbon oxidation to produce important petrochemicals, the geometry of the reactor should be such that only the gas-liquid interfacial area needs to be varied to control product selectivity.

Thermal conversion of bitumen was explored in the final two phases of this work, where it was necessary to explain non-monotonic viscosity trends as observed in certain previous studies on Cold Lake bitumen. In the third phase, it was found that experimental procedures after thermal reaction like extracting with poly-halogenated solvents like methylene chloride and rheological conditions of viscosity measurements impacted bitumen viscosity. Use of methylene chloride caused additional hydrogen bonding involving the hydrogen atoms from CH_2Cl_2 and heteroatoms in bitumen as detected through peak shifts to lower wavenumbers in FTIR spectra, which potentially increased viscosity by an order of magnitude. The micro-structure under shear was thought of moving from breakdown of the elastic gel-type structure to a viscous sol-type structure with the elastic nature re-occurring at high shear rates to maintain a constant viscosity in the second Newtonian region explained by power law.

During thermal conversion, the formation of lighter products and de-aggregation of aggregated material were seen as possible reasons for continuous viscosity decrease as observed for Athabasca bitumen in this work. It was also theorized that under the assumption of the colloidal model, the solvated layer of resins around asphaltenes could disrupt, thus exposing the asphaltenes for further aggregation but the hydrodynamic resistance provided by maltenes could have prevented an increase in viscosity at later reaction times, as opposed to Cold Lake bitumen. No proof of this argument was provided but was speculated based on experimental results.

Once the effect of post-reaction procedures on viscosity were studied, it was important to understand the key reactions taking place at different temperatures lower than that employed in industrial partial upgrading. The last phase of this study employed curve resolution (SMCR) and

Bayesian hierarchical clustering and structure learning on FTIR spectra to assist with deriving temperature-wise reaction mechanisms based on which a reaction network was proposed. The network pointed toward formation of lighter aliphatics and decrease in extent of substitution for the aromatics at most temperatures and reflected product properties identified from other experimental characterizations in the third phase of this study. It was also seen that PSO was better than ALS as the optimization technique for the final resolution in SMCR in terms of convergence speed, noise reduction and adherence to closure constraints. A global SMCR model consisting of all temperatures together showed the potential for monitoring a continuous thermal cracker for bitumen.

The key contribution of the final phase of this work was the ability of advanced quantitative chemometric methods to reveal underlying conversion chemistry with minimal external chemical knowledge. Overall, both experimental and chemometric tools were explored on different processes of importance to the petrochemical industry to develop an understanding of the fundamentals of the respective reaction chemistries and the chemometric tools showed potential for online monitoring of the processes as well.

7.3 Recommended future work

- A major limitation in the studies on propylene oligomerization and tetralin oxidation was the smaller number of samples available for construction of the regression models. Learning efficiency of the regression techniques were tested by varying the sample sizes within the 25 sets of samples but having more data at the same temperature (for example 346 ºC, 383 ºC) by varying the flowrates between the existing 172 and 440 mL/min would increase the predicting power of the models. This would reduce extrapolating the model for validation with respect to the flowrates even though temperature was shown to be having a greater effect on the product distribution. In the tetralin oxidation system, though it was not practical to conduct experiments at flowrates below 1 µL/min and greater than 15 µL/min as already explained in section 4.4.4.3 in Chapter 4, data could be collected at flowrates between 1 – 7 and used in the calibration set. This could improve the prediction accuracy of the linear regression models and further validate the applicability of significance tests through P-plots.

- At present, the interpretation of reaction chemistry from the chemometric results requires some human intervention, which can consume some time in the context of a complex process like thermal conversion of bitumen. Developing a database of possible model compounds and mapping the outputs of the chemometric models to these compounds through separate supervised classification methods can be a rigorous exercise in itself. This process can produce starting compounds from which further mechanisms can be developed.
- Most of data for chemometric investigations conducted as a part of this study were based on FTIR spectra. It could be useful if data from other characterizations like ESR, ^{1}H-NMR and chromatographic techniques were considered as inputs separately or fused along the frequency axis to carry out the data analysis. The data is available for the bitumen system as shown in Chapter 5. This can be further helpful in extracting knowledge through developing master equations and kinetic models for the bitumen system and can also be extended to other complex mixtures like biomass, slurries and colloidal systems.

7.4 Presentations and Publications

A list of the publications and presentations in conferences related to the work developed in the current research project is presented in the following:

- Sivaramakrishnan, K.; Puliyanda, A.; Tefera, D. T.; Ganesh, A.; Thirumalaivasan, S.; Prasad, V. Perspective on the Impact of Process Systems Engineering on Reaction Engineering. *Ind. Eng. Chem. Res.* Article ASAP, April 4th **2019**. Most parts of Chapter 2 were published in the section 'Chemometrics' in this review article.

- Sivaramakrishnan, K.; Nie, J.; De Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, 57, 13156 – 13176.
  (Presented at 4th Alberta Nano Research Symposium, 2017, Calgary, AB, Canada).

- Siddiquee, M. N.; Sivaramakrishnan, K.; Wu, Y.; De Klerk, A.; Nazemifard, N. A statistical approach dealing with multicollinearity among predictors in microfluidic reactor

operation to control liquid-phase oxidation selectivity. *Reaction Chemistry & Engineering* **2018**, 3, 972 – 990.

- Sivaramakrishnan, K.; de Klerk, A.; Prasad, V. Viscosity of Canadian Oilsands Bitumen and Its Modification by Thermal Conversion. In *Chemistry Solutions to Challenges in the Petroleum Industry*; Rahimi, P., Ovalles, C., Zhang, Y., Adams, J. J., Eds.; ACS Symposium Series; American Chemical Society: Washington, DC, 2019; Vol. 1320, pp 115–199.
  (Presented at the 255th ACS National Meeting, 2018, New Orleans, LA, US).

- Sivaramakrishnan, K.; Puliyanda, A.; De Klerk, A.; Prasad, V. A data-driven approach to generate pseudo-reaction networks for thermal conversion of Athabasca bitumen using chemometrics.
  (Submitted to Reaction Chemistry & Engineering).

REFERENCES

- Kumar, N.; Bansal, A.; Sarma, G. S.; Rawal, R. K. Chemometrics Tools Used in Analytical Chemistry: An Overview. *Talanta* **2014**, *123*, 186–199.
- Liang, Y.Z.; Xie, P.; Chan, K. Quality Control of Herbal Medicines. *J. Chromatogr. B* **2004**, *812*, 53–70.
- Willet, P. *Similarity and Clustering in Chemical Information*; Wiley: New York, 1987.
- Van de Vijver, R.; Devocht, B. R.; Van Geem, K. M.; Thybaut, J. W.; Marin, G. B. Challenges and Opportunities for Molecule-Based Management of Chemical Processes. *Curr. Opin. Chem. Eng.* **2016**, *13*, 142–149.
- Speight, J. G. Refining Heavy Oil and Extra-Heavy Oil. In *Heavy and Extra-heavy Oil Upgrading Technologies*; Gulf Professional Publishing: Boston, 2013.
- Belov, G. P.; Matkovsky, P. E. Processes for the Production of Higher Linear α-Olefins. *Pet. Chem.* **2010**, *50*, 283–289.
- Cai, T. Studies of a New Alkene Oligomerization Catalyst Derived from Nickel Sulfate. *Catal. Today* **1999**, *51*, 153–160.
- Meisel, S. L.; Weisz, P. B. Hydrocarbon Conversion and Synthesis Over ZSM-5 Catalysts. In *Advances in Catalytic Chemistry II Symposium*; Salt Lake City, Utah, 1982.
- Mohammed, A. A.; Fateen, S.-E. K.; Ahmed, T. S.; Moustafa, T. M. A Kinetic Model for Ethylene Oligomerization Using Zirconium/Aluminum- and Nickel/Zinc-Based Catalyst Systems in a Batch Reactor. *Appl. Petrochemical Res.* **2014**, *4*, 287–295.
- Xu, C.; Al Shoaibi, A. S.; Wang, C.; Carstensen, H.-H.; Dean, A. M. Kinetic Modeling of Ethane Pyrolysis at High Conversion. *J. Phys. Chem. A* **2011**, *115*, 10470–10490.
- Garg, R. K.; Krishnan, V. V.; Srivastava, V. K. Prediction of Concentration and Temperature Profiles for Non-Isothermal Ethane Cracking in a Pipe Reactor. *Korean J. Chem. Eng.* **2006**, *23*, 531–539.
- Suresh, A. K.; Sharma, M. M.; Sridhar, T. Engineering Aspects of Industrial Liquid-Phase Air Oxidation of Hydrocarbons. *Ind. Eng. Chem. Res.* **2000**, *39*, 3958–3997.
- Della Pina, C.; Falletta, E.; Rossi, M. Liquid Phase Oxidation of Organic Compounds by Supported Metal-Based Catalysts with a Focus on Gold. In *Liquid Phase Oxidation via Heterogeneous Catalysis: Organic Synthesis and Industrial Applications*; Wiley: Hoboken,

NJ, 2013; pp 221–262.

- Emanuel, N. M.; Denisov, E.; Maizus, Z. K. *Liquid-Phase Oxidation of Hydrocarbons*; Plenum Press: New York, 1967.
- Hill, C. L.; Kholdeeva, O. A. Selective Liquid Phase Oxidations in the Presence of Supported Polyoxometalates. In *Liquid Phase Oxidation via Heterogeneous Catalysis: Organic Synthesis and Industrial Applications*; Wiley: Hoboken, NJ, 2013; pp 263–319.
- The Future of Petrochemicals; *International Energy Agency* 2018.
- Gray, M. R. *Upgrading Oilsands Bitumen and Heavy Oil*; The University of Alberta Press: Edmonton, Canada, 2015.
- De Klerk, A.; Gray, M. R.; Zerpa, N. Unconventional Oil and Gas. Oilsands. In *Future Energy: Improved, Sustainable and Clean Options for our Planet*; Elsevier: Amsterdam, 2013; pp 95–116.
- Siddiquee, M. N.; de Klerk, A.; Nazemifard, N. Application of Microfluidics to Control Product Selectivity during Non-Catalytic Oxidation of Naphthenic-Aromatic Hydrocarbons. *React. Chem. Eng.* **2016**, *1*, 418–435.
- Zachariah, A.; De Klerk, A. Thermal Conversion Regimes for Oilsands Bitumen. *Energy Fuels* **2016**, *30*, 239–248.
- Yañez Jaramillo, L. M.; De Klerk, A. Partial Upgrading of Bitumen by Thermal Conversion at 150-300 °C. *Energy Fuels* **2018**, *32*, 3299–3311.
- Wang, L.; Zachariah, A.; Yang, S.; Prasad, V.; de Klerk, A. Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C. *Energy Fuels* **2014**, *28*, 5014–5022.
- Tefera, D. T.; Agrawal, A.; Yañez Jaramillo, L. M.; De Klerk, A.; Prasad, V. Self-Modeling Multivariate Curve Resolution Model for Online Monitoring of Bitumen Conversion Using Infrared Spectroscopy. *Ind. Eng. Chem. Res.* **2017**, *56*, 10756–10769.
- Rácz, A.; Bajusz, D.; Héberger, K. Chemometrics in Analytical Chemistry. In *Applied Chemoinformatics*; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2018; pp 471–499.
- Farrusseng, D.; Baumes, L.; Mirodatos, C. Data Management for Combinatorial Heterogeneous Catalysis: Methodology and Development of Advanced Tools. In *High-Throughput Analysis*; Springer: Boston, MA, 2003; pp 551–579.

- Caruthers, J. M.; Lauterbach, J. A.; Thomson, K. T.; Venkatasubramanian, V.; Snively, C. M.; Bhan, A.; Katare, S.; Oskarsdottir, G. Catalyst Design: Knowledge Extraction from High-Throughput Experimentation. *J. Catal.* **2003**, *216*, 98–109.
- Medford, A. J.; Kunz, M. R.; Ewing, S. M.; Borders, T.; Fushimi, R. Extracting Knowledge from Data through Catalysis Informatics. *ACS Catal.* **2018**, *8*, 7403–7429.
- Szymańska, E.; Gerretzen, J.; Engel, J.; Geurts, B.; Blanchet, L.; Buydens, L. M. Chemometrics and Qualitative Analysis Have a Vibrant Relationship. *TrAC Trends Anal. Chem.* **2015**, *69*, 34–51.
- Larson-Hall, J. Levels of Measurement of Variables in a Guide to Doing Statistics. In *Second language research using SPSS*; Taylor & Francis: New York, 2010; pp 33–35.
- Osborne, B. G. Near-Infrared Spectroscopy in Food Analysis. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: Chichester, UK, 2000.
- Igne, B.; Ciurczak, E. W. *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy*; CRC Press: Boca Raton, 2014.
- Ertaş, M.; Hakki Alma, M. Pyrolysis of Laurel (Laurus Nobilis L.) Extraction Residues in a Fixed-Bed Reactor: Characterization of Bio-Oil and Bio-Char. *J. Anal. Appl. Pyrolysis* **2010**, *88*, 22–29.
- Guo, X.; Wang, S.; Guo, Z.; Liu, Q.; Luo, Z.; Cen, K. Pyrolysis Characteristics of Bio-Oil Fractions Separated by Molecular Distillation. *Appl. Energy* **2010**, *87*, 2892–2898.
- Owen, T. *Fundamentals of UV-Visible Spectroscopy: A Primer*; Hewlett Packard, 1996.
- Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, D. L. *Spectrometric Identification of Organic Compounds.*; John Wiley & Sons: New York, 2014.
- Mathiesen, M. D.; Lubeck, A. J. Improving Accuracy in the Determination of Aromatics in Gasoline by Gas Chromatography-Mass Spectrometry. *J. Chromatogr. Sci.* **1998**, *36*, 449–456.
- Wang, Y.P.; Zou, Y.R.; Shi, J.T.; Shi, J. Review of the Chemometrics Application in Oil-Oil and Oil-Source Rock Correlations. *J. Nat. Gas Geosci.* **2018**, *3*, 217–232.
- Dellamorte, J. C.; Barteau, M. A.; Lauterbach, J. Opportunities for Catalyst Discovery and Development: Integrating Surface Science and Theory with High Throughput Methods. *Surf. Sci.* **2009**, *603*, 1770–1775.

- Carter, S.; Fisher, A.; Garcia, R.; Gibson, B.; Lancaster, S.; Marshall, J.; Whiteside, I. Atomic Spectrometry Update. Review of Advances in the Analysis of Metals, Chemicals and Functional Materials. *J. Anal. At. Spectrom.* **2015**, *30*, 2249–2294.
- Xie, L.; Ying, Y.; Ying, T.; Yu, H.; Fu, X. Discrimination of Transgenic Tomatoes Based on Visible/near-Infrared Spectra. *Anal. Chim. Acta* **2007**, *584*, 379–384.
- Kowalkowski, T.; Zbytniewski, R.; Szpejna, J.; Buszewski, B. Application of Chemometrics in River Water Classification. *Water Res.* **2006**, *40*, 744–752.
- Lankmayr, E.; Mocak, J.; Serdt, K.; Balla, B.; Wenzl, T.; Bandoniene, D.; Gfrerer, M.; Wagner, S. Chemometrical Classification of Pumpkin Seed Oils Using UV–Vis, NIR and FTIR Spectra. *J. Biochem. Biophys. Methods* **2004**, *61*, 95–106.
- Danezis, G. P.; Tsagkaris, A. S.; Camin, F.; Brusic, V.; Georgiou, C. A. Food Authentication: Techniques, Trends & Emerging Approaches. *TrAC Trends Anal. Chem.* **2016**, *85*, 123–132.
- Munck, L.; Nørgaard, L.; Engelsen, S. B.; Bro, R.; Andersson, C. A. Chemometrics in Food Science—a Demonstration of the Feasibility of a Highly Exploratory, Inductive Evaluation Strategy of Fundamental Scientific Significance. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 31–60.
- Nunes, C. A.; Alvarenga, V. O.; de Souza Sant'Ana, A.; Santos, J. S.; Granato, D. The Use of Statistical Software in Food Science and Technology: Advantages, Limitations and Misuses. *Food Res. Int.* **2015**, *75*, 270–280.
- Martin, Y. C.; Hackbarth, J. J. Theoretical Model-Based Equations for the Linear Free Energy Relations of the Biological Activity of Ionizable Substances. 1. Equilibrium-Controlled Potency. *J. Med. Chem.* **1976**, *19*, 1033–1039.
- Telnaes, N.; Dahl, B. Oil-Oil Correlation Using Multivariate Techniques. *Org. Geochem.* **1986**, *10*, 425–432.
- Øygard, K.; Grahl-Nielsen, O.; Ulvøen, S. Oil/Oil Correlation by Aid of Chemometrics. *Org. Geochem.* **1984**, *6*, 561–567.
- Sivaramakrishnan, K.; Nie, J.; de Klerk, A.; Prasad, V. Least Squares-Support Vector Regression for Determining Product Concentrations in Acid-Catalyzed Propylene Oligomerization. *Ind. Eng. Chem. Res.* **2018**, *57*, 13156–13176.
- Siddiquee, M. N.; Sivaramakrishnan, K.; Wu, Y.; de Klerk, A.; Nazemifard, N. A Statistical Approach Dealing with Multicollinearity among Predictors in Microfluidic Reactor Operation to Control Liquid-Phase Oxidation Selectivity. *React. Chem. Eng.* **2018**, *3*, 972–990.

- Kiralj, R.; Ferreira, M. M. C. The Past, Present, and Future of Chemometrics Worldwide: Some Etymological, Linguistic, and Bibliometric Investigations. *J. Chemom.* **2006**, *20*, 247–272.
- Martens, H.; Næs, T. Multivariate Calibration. In *Chemometrics*; Springer Netherlands: Dordrecht, 1984; pp 147–156.
- Sjöström, M.; Wold, S.; Lindberg, W.; Persson, J.-Å.; Martens, H. A Multivariate Calibration Problem in Analytical Chemistry Solved by Partial Least-Squares Models in Latent Variables. *Anal. Chim. Acta* **1983**, *150*, 61–70.
- Haykin, S. S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall: Upper Saddle River, N.J., 1998.
- Press, W. H.; Teukolsky, S. A. Savitzky-Golay Smoothing Filters. *Comput. Phys.* **1990**, *4*, 669–672.
- Afseth, N. K.; Kohler, A. Extended Multiplicative Signal Correction in Vibrational Spectroscopy, a Tutorial. *Chemom. Intell. Lab. Syst.* **2012**, *117*, 92–99.
- Kvalheim, O. M.; Brakstad, F.; Liang, Y. Preprocessing of Analytical Profiles in the Presence of Homoscedastic or Heteroscedastic Noise. *Anal. Chem.* **1994**, *66*, 43–51.
- Van Den Berg, R. A.; Hoefsloot, H. C.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. Centering, Scaling, and Transformations: Improving the Biological Information Content of Metabolomics Data. *BMC Genomics* **2006**, *7*, 142.
- Eriksson, L. *Introduction to Multi- and Megavariate Data Analysis Using Projection Methods (PCA & PLS)*; Umetrics: Sweden, 1999.
- Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J. C.; Jellema, R. H. Fusion of Mass Spectrometry-Based Metabolomics Data. *Anal. Chem.* **2005**, *77*, 6729–6736.
- Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Improved Analysis of Multivariate Data by Variable Stability Scaling: Application to NMR-Based Metabolic Profiling. *Anal. Chim. Acta* **2003**, *490*, 265–276.
- Dobos, L.; Abonyi, J. On-Line Detection of Homogeneous Operation Ranges by Dynamic Principal Component Analysis Based Time-Series Segmentation. *Chem. Eng. Sci.* **2012**, *75*, 96–105.
- Lavine, B. K.; Davidson, C. E.; Moores, A. J. Genetic Algorithms for Spectral Pattern Recognition. *Vib. Spectrosc.* **2002**, *28*, 83–95.

- Smith, G. L.; Mlynczak, P. E.; Potter, G. L. A Technique Using Principal Component Analysis to Compare Seasonal Cycles of Earth Radiation from CERES and Model Computations. *J. Geophys. Res. Atmos.* **2012**, *117*.
- Mandel, J. Use of the Singular Value Decomposition in Regression Analysis. *Am. Stat.* **1982**, *36*, 15–24.
- Camacho, J.; Picó, J.; Ferrer, A. Data Understanding with PCA: Structural and Variance Information Plots. *Chemom. Intell. Lab. Syst.* **2010**, *100*, 48–56.
- Jolliffe, I. T. Mathematical and Statistical Properties of Population Principal Components. In *Principal Component Analysis*; Springer Series in Statistics; Springer-Verlag: New York, 2002.
- Shao, Y.; He, Y. Visible/Near Infrared Spectroscopy and Chemometrics for the Prediction of Trace Element (Fe and Zn) Levels in Rice Leaf. *Sensors* **2013**, *13*, 1872–1883.
- Yeganeh, B.; Motlagh, M. S. P.; Rashidi, Y.; Kamalan, H. Prediction of CO Concentrations Based on a Hybrid Partial Least Square and Support Vector Machine Model. *Atmos. Environ.* **2012**, *55*, 357–365.
- Granato, D.; Putnik, P.; Kovačević, D. B.; Santos, J. S.; Calado, V.; Rocha, R. S.; Cruz, A. G. Da; Jarvis, B.; Rodionova, O. Y.; Pomerantsev, A. Trends in Chemometrics: Food Authentication, Microbiology, and Effects of Processing. *Compr. Rev. Food Sci. Food Saf.* **2018**, *17*, 663–677.
- Wold, S.; Sjöström, M. Chemometrics, Present and Future Success. *Chemom. Intell. Lab. Syst.* **1998**, *44*, 3–14.
- Zimmermann, B. Characterization of Pollen by Vibrational Spectroscopy. *Appl. Spectrosc.* **2010**, *64*, 1364–1373.
- Mok, P. Y.; Huang, H. Q.; Kwok, Y. L.; Au, J. S. A Robust Adaptive Clustering Analysis Method for Automatic Identification of Clusters. *Pattern Recognit.* **2012**, *45*, 3017–3033.
- Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An Efficient K-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892.
- Borg, I.; Groenen, P. J. F. *Modern Multidimensional Scaling*; Springer: New York, 2005.
- Shepard, R. N. The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function. I. *Psychometrika* **1962**, *27*, 125–140.

- Kruskal, J. B. Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis. *Psychometrika* **1964**, *29*, 1–27.
- Wang, Y. P.; Zhang, F.; Zou, Y. R.; Zhan, Z. W.; Peng, P. Chemometrics Reveals Oil Sources in the Fangzheng Fault Depression, NE China. *Org. Geochem.* **2016**, *102*, 1–13.
- Aqualog Software User's Guide for Version 3.6. Horiba Scientific 2012.
- Bro, R. PARAFAC. Tutorial and Applications. *Chemom. Intell. Lab. Syst.* **1997**, *38*, 149–171.
- Marx, B. D.; Eilers, P. H. C.; Li, B. Multidimensional Single-Index Signal Regression. *Chemom. Intell. Lab. Syst.* **2011**, *109*, 120–130.
- ten Berge, J. M. *Least Squares Optimization in Multivariate Analysis*; DSWO Press: Leiden, 1993.
- Tax, D.; Duin, R. Outlier Detection Using Classifier Instability. *Lect. notes Comput. Sci.* **1998**, *1451*, 593–601.
- Berrueta, L. A.; Alonso-Salces, R. M.; Héberger, K. Supervised Pattern Recognition in Food Analysis. *J. Chromatogr. A* **2007**, *1158*, 196–214.
- Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: San Diego, 1990.
- Subramanian, J.; Simon, R. Overfitting in Prediction Models – Is It a Problem Only in High Dimensions? *Contemp. Clin. Trials* **2013**, *36*, 636–641.
- Ståhle, L.; Wold, S. Partial Least Squares Analysis with Cross-Validation for the Two-Class Problem: A Monte Carlo Study. *J. Chemom.* **1987**, *1*, 185–196.
- Cebi, N.; Durak, M. Z.; Toker, O. S.; Sagdic, O.; Arici, M. An Evaluation of Fourier Transforms Infrared Spectroscopy Method for the Classification and Discrimination of Bovine, Porcine and Fish Gelatins. *Food Chem.* **2016**, *190*, 1109–1115.
- Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J. *Classification And Regression Trees*; Routledge: New York, 2017.
- Kvalheim, O. M. Latent-Structure Decompositions (Projections) of Multivariate Data. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 283–290.
- Christie, O. H. J.; Rácz, A.; Elek, J.; Héberger, K. Classification and Unscrambling a Class-inside-Class Situation by Object Target Rotation: Hungarian Silver Coins of the Árpád Dynasty, AD 997-1301. *J. Chemom.* **2014**, *28*, 287–292.

- Wold, M. S. SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In *Chemometrics theory and application, American Chemical Society Symposium Series*; American Chemical Society: Washington, D. C., 1977; pp 243–282.
- Gondim, C. D. S.; Junqueira, R. G.; Souza, S. V. C. D.; Ruis´anchez, I.; Callao, M. P. Detection of Several Common Adulterants in Raw Milk by MID-Infrared Spectroscopy and One-Class and Multi-Class Multivariate Strategies. *Food Chem.* **2017**, *230*, 68–75.
- Gurbanov, R.; Gozen, A. G.; Severcan, F. Rapid Classification of Heavy Metal-Exposed Freshwater Bacteria by Infrared Spectroscopy Coupled with Chemometrics Using Supervised Method. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2018**, *189*, 282–290.
- Kanik, E. A.; Orekici Temel, G.; Erdogan, S.; Ersoz Kaya, I. Affected States Soft Independent Modeling by Class Analogy from the Relation Between Independent Variables, Number of Independent Variables and Sample Size. *Balkan Med. J.* **2013**, *30*, 28–32.
- Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics: A Practical Guide*; Wiley & Sons: New York, 1998.
- P´erez-Caballero, G.; Andrade, J. M.; Olmos, P.; Molina, Y.; Jim´enez, I.; Dur´an, J. J.; Miguel-Cruz, F. Authentication of Tequilas Using Pattern Recognition and Supervised Classification. *TrAC Trends Anal. Chem.* **2017**, *94*, 117–129.
- Niazi, A.; Goodarzi, M.; Yazdanipour, A. A Comparative Study between Least-Squares Support Vector Machines and Partial Least Squares in Simultaneous Spectrophotometric Determination of Cypermethrin, Permethrin and Tetramethrin. *J. Braz. Chem. Soc.* **2008**, *19*, 536–542.
- Asuero, A. G.; Sayago, A.; González, A. G. The Correlation Coefficient: An Overview. *Crit. Rev. Anal. Chem.* **2006**, *36*, 41–59.
- Koljonen, J.; Nordling, T. E. M.; Alander, J. T. A Review of Genetic Algorithms in near Infrared Spectroscopy and Chemometrics: Past and Future. *J. Near Infrared Spectrosc.* **2008**, *16*, 189–197.
- Hoerl, A. E.; Kennard, R. W. Ridge Regression:Biased Estimation for Nonorthogonal Problems. *Technometrics* **1970**, *12*, 55–67.
- Kalivas, J. H. Overview of Two-Norm (L2) and One-Norm (L1) Tikhonov Regularization Variants for Full Wavelength or Sparse Spectral Multivariate Calibration Models or Maintenance. *J. Chemom.* **2012**, *26*, 218–230.

- Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. Interval Partial Least-Squares Regression (IPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54*, 413–419.
- Geladi, P.; Kowalski, B. R. An Example of 2-Block Predictive Partial Least-Squares Regression with Simulated Data. *Anal. Chim. Acta* **1986**, *185*, 19–32.
- Frank, I. E.; Friedman, J. H. A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **1993**, *35*, 109–135.
- Balabin, R. M.; Lomakina, E. I. Support Vector Machine Regression (SVR/LS-SVM)—an Alternative to Neural Networks (ANN) for Analytical Chemistry? Comparison of Nonlinear Methods on near Infrared (NIR) Spectroscopy Data. *Analyst* **2011**, *136*, 1703.
- Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 2000.
- Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.*; MIT Press: Cambridge, MA, 2002.
- Chauchard, F.; Cogdill, R.; Roussel, S.; Roger, J. M.; Bellon-Maurel, V. Application of LS-SVM to Non-Linear Phenomena in NIR Spectroscopy: Development of a Robust and Portable Sensor for Acidity Prediction in Grapes. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 141–150.
- Miller, A. *Subset Selection in Regression*; Chapman and Hall: London, 1990.
- Norgaard, L.; Leardi, R. *IToolbox Manual*; 2005.
- Rajalahti, T.; Arneberg, R.; Berven, F. S.; Myhr, K.-M.; Ulvik, R. J.; Kvalheim, O. M. Biomarker Discovery in Mass Spectral Profiles by Means of Selectivity Ratio Plot. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35–48.
- Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: River Edge, NJ, 2002.
- Wahba, G. Numerical Methods. In *Spline Models for Observational Data*; Society for Industrial and Applied Mathematics: Philadelphia, Pennsylvania, 1990; pp 135–143.
- Efron, B.; Tibshirani, R. *An Introduction to the Bootstrap*; Chapman and Hall: New York, 1993.
- Filzmoser, P.; Liebmann, B.; Varmuza, K. Repeated Double Cross Validation. *J. Chemom.* **2009**, *23*, 160–171.

- Cui, C.; Fearn, T. Comparison of Partial Least Squares Regression, Least Squares Support Vector Machines, and Gaussian Process Regression for a near Infrared Calibration. *J. Near Infrared Spectrosc.* **2017**, *25*, 5–14.
- de Juan, A.; Tauler, R. Chemometrics Applied to Unravel Multicomponent Processes and Mixtures: Revisiting Latest Trends in Multivariate Resolution. *Anal. Chim. Acta* **2003**, *500*, 195–210.
- Ho, C. N.; Christian, G. D.; Davidson, E. R. Application of the Method of Rank Annihilation to Quantitative Analyses of Multicomponent Fluorescence Data from the Video Fluorometer. *Anal. Chem.* **1978**, *50*, 1108–1113.
- Maeder, M. Evolving Factor Analysis for the Resolution of Overlapping Chromatographic Peaks. *Anal. Chem.* **1987**, *59*, 527–530.
- Tauler, R.; Marques, I.; Casassas, E. Multivariate Curve Resolution Applied to Three-way Trilinear Data: Study of a Spectrofluorimetric Acid–Base Titration of Salicylic Acid at Three Excitation Wavelengths. *J. Chemom.* **1998**, *12*, 55–75.
- Sánchez, F. C.; Toft, J.; van den Bogaert, B.; Massart, D. L. Orthogonal Projection Approach Applied to Peak Purity Assessment. *Anal. Chem.* **1996**, *68*, 79–85.
- Keller, H. R.; Massart, D. L. Peak Purity Control in Liquid Chromatography with Photodiode-Array Detection by a Fixed Size Moving Window Evolving Factor Analysis. *Anal. Chim. Acta* **1991**, *246*, 379–390.
- Christiansen, T. F.; Busch, J. E.; Krogh, S. C. Successive Determinations of Calcium and Magnesium in Drinking Water by Complexometric, Potentiometric Digital Titration to Two Equivalence Points. *Anal. Chem.* **1976**, *48*, 1051–1056.
- Wu, A. H. B.; Malmstadt, H. V. Versatile Microcomputer-Controlled Titrator. *Anal. Chem.* **1978**, *50*, 2090–2096.
- Ruzicka, J.; Hansen, E. H. *Flow Injection Analysis*; Wiley: New York, 1981.
- Betteridge, D.; Cheng, W. C.; Dagless, E. L.; David, P.; Goad, T. B.; Deans, D. R.; Newton, D. A.; Pierce, T. B. An Automated Viscometer Based on High-Precision Flow Injection Analysis. Part II. Measurement of Viscosity and Diffusion Coefficients. *Analyst* **1983**, *108*, 17.
- Stieg, S.; Nieman, T. A. Application of a Microcomputer Controlled Chemiluminescence Research Instrument to the Simultaneous Determination of Cobalt (II) and Silver (I) by Gallic Acid Chemiluminescence. *Anal. Chem.* **1980**, *52*, 800–804.

- Sly, T. J.; Betteridge, D.; Wibberley, D.; Porter, D. G. An Improved Flow-through Phototransducer. *J. Automat. Chem.* **1982**, *4*, 186–189.
- Belchamber, R. M.; Betteridge, D.; Chow, Y. T.; Sly, T. J.; Wade, A. P. The Application of Computers in Chemometrics and Analytical Chemistry. *Anal. Chim. Acta* **1983**, *150*, 115–128.
- Marec, A.; Thomas, J.-H.; El Guerjouma, R. Damage Characterization of Polymer-Based Composite Materials: Multivariable Analysis and Wavelet Transform for Clustering Acoustic Emission Data. *Mech. Syst. Signal Process.* **2008**, *22*, 1441–1464.
- Moran, M. G.; Kowalski, B. R. Image Analysis. *Anal. Chem.* **1979**, *51*, 776A-788A.
- Gordon, R.; Herman, G. T.; Johnson, S. A. Image Reconstruction from Projections. *Sci. Am.* **1975**, *233*, 56–71.
- Martin, Y. C.; Hackbarth, J. J. Examples of the Application of Nonlinear Regression Analysis to Chemical Data. In *Chemometrics: Theory and Application ACS Symposium Series1*; American Chemical Society: Washington, D. C., 1977; pp 153–164.
- Albert, A.; Serjeant, E. P. *The Determination of Ionization Constants*; Chapman and Hall: London, 1971.
- Eide, I.; Neverdal, G.; Thorvaldsen, B.; Grung, B.; Kvalheim, O. M. Toxicological Evaluation of Complex Mixtures by Pattern Recognition: Correlating Chemical Fingerprints to Mutagenicity. *Environ. Health Perspect.* **2002**, *110*, 985–988.
- Eide, I.; Zahlsen, K. A Novel Method for Chemical Fingerprinting of Oil and Petroleum Products Based on Electrospray Mass Spectrometry and Chemometrics. *Energy Fuels* **2005**, *19*, 964–967.
- Qian, K.; Rodgers, R. P.; Hendrickson, C. L.; Emmett, M. R.; Marshall, A. G. Reading Chemical Fine Print: Resolution and Identification of 3000 Nitrogen-Containing Aromatic Compounds from a Single Electrospray Ionization Fourier Transform Ion Cyclotron Resonance Mass Spectrum of Heavy Petroleum Crude Oil. *Energy Fuels* **2001**, *15*, 492–498.
- Wang, Y. P.; Zhang, F.; Zou, Y. R.; Sun, J. N.; Lin, X. H.; Liang, T. Oil Source and Charge in the Wuerxun Depression, Hailar Basin, Northeast China: A Chemometric Study. *Mar. Pet. Geol.* **2018**, *89*, 665–686.
- Mudge, S. Aspects of Hydrocarbon Fingerprinting Using PLS—New Data From Prince William Sound. *Environ. Forensics* **2002**, *3*, 323–329.

- Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. Chromatographic Alignment by Warping and Dynamic Programming as a Pre-Processing Tool for PARAFAC Modelling of Liquid Chromatography–Mass Spectrometry Data. *J. Chromatogr. A* **2002**, *961*, 237–244.
- Comas, E.; Gimeno, R. A.; Ferré, J.; Marcé, R. M.; Borrull, F.; Rius, F. X. Quantification from Highly Drifted and Overlapped Chromatographic Peaks Using Second-Order Calibration Methods. *J. Chromatogr. A* **2004**, *1035*, 195–202.
- Ho, C. N.; Christian, G. D.; Davidson, E. R. Application of the Method of Rank Annihilation to Fluorescent Multicomponent Mixtures of Polynuclear Aromatic Hydrocarbons. *Anal. Chem.* **1980**, *52*, 1071–1079.
- De Braekeleer, K.; de Juan, A.; Massart, D. L. Purity Assessment and Resolution of Tetracycline Hydrochloride Samples Analysed Using High-Performance Liquid Chromatography with Diode Array Detection. *J. Chromatogr. A* **1999**, *832*, 67–86.
- Salau, J. S.; Honing, M.; Tauler, R.; Barceló, D. Resolution and Quantitative Determination of Coeluted Pesticide Mixtures in Liquid Chromatography–Thermospray Mass Spectrometry by Multivariate Curve Resolution. *J. Chromatogr. A* **1998**, *795*, 3–12.
- Dinç, E.; Baydan, E.; Kanbur, M.; Onur, F. Spectrophotometric Multicomponent Determination of Sunset Yellow, Tartrazine and Allura Red in Soft Drink Powder by Double Divisor-Ratio Spectra Derivative, Inverse Least-Squares and Principal Component Regression Methods. *Talanta* **2002**, *58*, 579–594.
- Chen, Y.; Zhu, S. B.; Xie, M. Y.; Nie, S. P.; Liu, W.; Li, C.; Gong, X. F.; Wang, Y. X. Quality Control and Original Discrimination of Ganoderma Lucidum Based on High-Performance Liquid Chromatographic Fingerprints and Combined Chemometrics Methods. *Anal. Chim. Acta* **2008**, *623*, 146–156.
- Rohman, A.; Che Man, Y. B. Application of Fourier Transform Infrared (FT-IR) Spectroscopy Combined with Chemometrics for Authentication of Cod-Liver Oil. *Vib. Spectrosc.* **2011**, *55*, 141–145.
- Eaton, G. R.; Eaton, S. S.; Barr, D. P.; Weber, R. T. *Quantitative EPR*; Springer: Vienna, 2010.
- Koskinen, J. R.; Kowalski, B. R. Analysis of the Electron Spin Resonance of Spin Labels Using Chemometric Methods. In *Chemometrics: Theory and Application ACS Symposium Series*; American Chemical Society: Washington, D. C., 1977; pp 117–126.

- Nascimento, K. S. do; Gasparotto Sattler, J. A.; Lauer Macedo, L. F.; Serna González, C. V.; Pereira de Melo, I. L.; da Silva Araújo, E.; Granato, D.; Sattler, A.; de Almeida-Muradian, L. B. Phenolic Compounds, Antioxidant Capacity and Physicochemical Properties of Brazilian Apis Mellifera Honeys. *LWT* **2018**, *91*, 85–94.
- Grunert, T.; Stephan, R.; Ehling-Schulz, M.; Johler, S. Fourier Transform Infrared Spectroscopy Enables Rapid Differentiation of Fresh and Frozen/Thawed Chicken. *Food Control* **2016**, *60*, 361–364.
- D'Archivio, A. A.; Giannitto, A.; Maggi, M. A.; Ruggieri, F. Geographical Classification of Italian Saffron (Crocus Sativus L.) Based on Chemical Constituents Determined by High-Performance Liquid-Chromatography and by Using Linear Discriminant Analysis. *Food Chem.* **2016**, *212*, 110–116.
- Muñiz-Valencia, R.; Jurado, J. M.; Ceballos-Magaña, S. G.; Alcázar, Á.; Hernández-Díaz, J. Characterization of Mexican Coffee According to Mineral Contents by Means of Multilayer Perceptrons Artificial Neural Networks. *J. Food Compos. Anal.* **2014**, *34*, 7–11.
- O'Farrell, M.; Lewis, E.; Flanagan, C.; Lyons, W.; Jackman, N. Comparison of K-NN and Neural Network Methods in the Classification of Spectral Data from an Optical Fibre-Based Sensor System Used for Quality Control in the Food Industry. *Sensors Actuators B Chem.* **2005**, *111–112*, 354–362.
- Kuuliala, L.; Abatih, E.; Ioannidis, A.-G.; Vanderroost, M.; De Meulenaer, B.; Ragaert, P.; Devlieghere, F. Multivariate Statistical Analysis for the Identification of Potential Seafood Spoilage Indicators. *Food Control* **2018**, *84*, 49–60.
- Olsson, D. M.; Nelson, L. S. The Nelder-Mead Simplex Procedure for Function Minimization. *Technometrics* **1975**, *17*, 45–51.
- Xavier-De-Souza, S.; Suykens, J. A. K.; Vandewalle, J.; Bolle, D. Coupled Simulated Annealing. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **2010**, *40*, 320–335.
- Routh, M. W.; Swartz, P. A.; Denton, M. B. Performance of the Super Modified Simplex. *Anal. Chem.* **1977**, *49*, 1422–1428.
- Ryan, P. B.; Barr, R. L.; Todd, H. D. Simplex Techniques for Nonlinear Optimization. *Anal. Chem.* **1980**, *52*, 1460–1467.
- Borgen, O. S.; Kowalski, B. R. An Extension of the Multivariate Component-Resolution Method to Three Components. *Anal. Chim. Acta* **1985**, *174*, 1–26.

- Shmygelska, A.; Hoos, H. H. An Ant Colony Optimisation Algorithm for the 2D and 3D Hydrophobic Polar Protein Folding Problem. *BMC Bioinformatics* **2005**, *6*, 1–22.
- Shinzawa, H.; Jiang, J.-H.; Iwahashi, M.; Noda, I.; Ozaki, Y. Self-Modeling Curve Resolution (SMCR) by Particle Swarm Optimization (PSO). *Anal. Chim. Acta* **2007**, *595*, 275–281.
- Rousseeuw, P. J. Least Median of Squares Regression. *J. Am. Stat. Assoc.* **2012**, *79*, 871–880.
- Shinzawa, H.; Jiang, J.-H.; Iwahashi, M.; Ozaki, Y. Robust Curve Fitting Method for Optical Spectra by Least Median Squares (LMedS) Estimator with Particle Swarm Optimization (PSO). *Anal. Sci.* **2007**, *23*, 781–785.
- Rangarajan, S.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Description of RING. *Comput. Chem. Eng.* **2012**, *45*, 114–123.
- Ipatieff, V.; Corson, B. B.; Egloff, G. Polymerization, a New Source of Gasoline. *Ind. Eng. Chem. Res.* **1935**, *27*, 1077–1081.
- Tabak, S. A.; Krambeck, F. J.; Garwood, W. E. Conversion of Propylene and Butylene over ZSM-5 Catalyst. *AIChE J.* **1986**, *32*, 1526–1531.
- Quann, R. J.; Green, L. A.; Tabak, S. A.; Krambeck, F. J. Chemistry of Olefin Oligomerization over ZSM-5 Catalyst. *Ind. Eng. Chem. Res.* **1988**, *27*, 565–570.
- Garwood, W. E. Conversion of C2-C10 to Higher Olefins over Synthetic Zeolite ZSM-5. In *Intrazeolite Chemistry*; American Chemical Society: Washington, D. C., 1983; pp 383–396.
- Pines, H. Acid-Catalyzed Reactions. In *The Chemistry of Catalytic Hydrocarbon Conversions*; Academic Press: New York, 1981; pp 1–122.
- Nicholas, C. P. Applications of Light Olefin Oligomerization to the Production of Fuels and Chemicals. *Appl. Catal. A Gen.* **2017**, *543*, 82–97.
- Osborne, B. G. Near-Infrared Spectroscopy in Food Analysis. In *Encyclopedia of Analytical Chemistry*; John Wiley & Sons: Chichester, UK, 2000.
- Igne, B.; Ciurczak, E. W. *Pharmaceutical and Medical Applications of Near-Infrared Spectroscopy*; CRC Press: Boca Raton, 2014.
- Van Den Berg, J. P.; Wolthuizen, J. P.; van Hooff, J. H. C. Reaction of Small Olefins on Zeolite H-ZSM-5. A Thermogravimetric Study at Low and Intermediate Temperatures. *J. Catal.* **1983**, *80*, 139–144.
- Wold, S.; Sjöström, M.; Eriksson, L. PLS-Regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.

- Frank, I. E. A Nonlinear PLS Model. *Chemom. Intell. Lab. Syst.* **1990**, *8*, 109–119.
- Wold, S. Nonlinear Partial Least Squares Modelling II. Spline Inner Relation. *Chemom. Intell. Lab. Syst.* **1992**, *14*, 71–84.
- Haykin, S. S. *Neural Networks: A Comprehensive Foundation*; Prentice Hall: Upper Saddle River, N.J., 1998.
- Cortes, C.; Vapnik, V. Support Vector Networks. *Mach. Learn.* **1995**, *20*, 273–297.
- Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*; Cambridge University Press: Cambridge, 2000.
- Thissen, U.; Pepers, M.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. Comparing Support Vector Machines to PLS for Spectral Regression Applications. *Chemom. Intell. Lab. Syst.* **2004**, *73*, 169–179.
- Gunn, S. R. *Support Vector Machines for Classification and Regression*; School of Electronics and Computer Science, University of Southampton, 1998.
- Gibson, N. P.; Aigrain, S.; Roberts, S.; Evans, T. M.; Osborne, M.; Pont, F. A Gaussian Process Framework for Modelling Instrumental Systematics: Application to Transmission Spectroscopy. *Mon. Not. R. Astron. Soc.* **2012**, *419*, 2683–2694.
- Tefera, D. T.; Yañez Jaramillo, L. M.; Ranjan, R.; Li, C.; De Klerk, A.; Prasad, V. A Bayesian Learning Approach to Modeling Pseudoreaction Networks for Complex Reacting Systems: Application to the Mild Visbreaking of Bitumen. *Ind. Eng. Chem. Res.* **2017**, *56*, 1961–1970.
- Tefera, D. T.; de Klerk, A.; Prasad, V. A Bayesian Learning and Data Mining Approach to Reaction System Identification: Application to Biomass Conversion. In *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*; IEEE, 2017; pp 178–183.
- Kaufman, L.; Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley Series in Probability and Statistics; John Wiley & Sons, Inc.: Hoboken, NJ, 1990.
- Ge, Z.; Song, Z.; Ding, S. X.; Huang, B. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access* **2017**, *5*, 20590–20616.
- Lavine, B. K.; Nuguru, K.; Mirjankar, N.; Workman, J. Development of Carboxylic Acid Search Prefilters for Spectral Library Matching. *Microchem. J.* **2012**, *103*, 21–36.

- Smith, G. L.; Mlynczak, P. E.; Potter, G. L. A Technique Using Principal Component Analysis to Compare Seasonal Cycles of Earth Radiation from CERES and Model Computations. *J. Geophys. Res. Atmos.* **2012**, *117*.
- Fiser, K.; Sieger, T.; Schumich, A.; Irving, J.; Dworzak, M. N.; Vormoor, J. MRD Monitoring of Childhood ALL Using Hierarchical Clustering and Support Vector Machine Learning of Complex Multi-Parameter Flow Cytometry Data. *Blood* **2008**, *112*, 1508.
- Fan, C. Y.; Fan, P. S.; Chan, T. Y.; Chang, S. H. Using Hybrid Data Mining and Machine Learning Clustering Analysis to Predict the Turnover Rate for Technology Professionals. *Expert Syst. Appl.* **2012**, *39*, 8844–8851.
- Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. An Efficient K-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 881–892.
- Balabin, R. M.; Safieva, R. Z.; Lomakina, E. I. Gasoline Classification Using near Infrared (NIR) Spectroscopy Data: Comparison of Multivariate Techniques. *Anal. Chim. Acta* **2010**, *671*, 27–35.
- Fernández Pierna, J. A.; Baeten, V.; Renier, A. M.; Cogdill, R. P.; Dardenne, P. Combination of Support Vector Machines (SVM) and near-Infrared (NIR) Imaging Spectroscopy for the Detection of Meat and Bone Meal (MBM) in Compound Feeds. *J. Chemom.* **2004**, *18*, 341–349.
- Pelckmans, K.; Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; Lukas, L.; Hamers, B.; De Moor, B.; Vandewalle, J. LS-SVMlab Toolbox User's Guide. *Pattern Recognit. Lett.* **2003**, *24*, 659–675.
- Norgaard, L.; Leardi, R. *IToolbox Manual*; FOSS, Data & Algorithms in R&D, 2005.
- Colthup, N.; Daly, L.; Wiberley, S. *Introduction to Infrared and Raman Spectroscopy*, 2nd Edn.; Academic Press: Boston, 1990.
- Lazraq, A.; Cléroux, R.; Gauchi, J. P. Selecting Both Latent and Explanatory Variables in the PLS1 Regression Model. *Chemom. Intell. Lab. Syst.* **2003**, *66*, 117–126.
- Snee, R. D. Validation of Regression Models: Methods and Examples. *Technometrics* **1977**, *19*, 415–428.
- Thissen, U.; Üstün, B.; Melssen, W. J.; Buydens, L. M. C. Multivariate Calibration with Least-Squares Support Vector Machines. *Anal. Chem.* **2004**, *76*, 3099–3105.

- Den Hollander, M. A.; Wissink, M.; Makkee, M.; Moulijn, J. A. Gasoline Conversion: Reactivity towards Cracking with Equilibrated FCC and ZSM-5 Catalysts. *Appl. Catal. A Gen.* **2002**, *223*, 85–102.
- Stasinopoulos, M. D.; Rigby, R. A.; Heller, G. Z.; Voudouris, V.; De Bastiani, F. *Flexible Regression and Smoothing*; CRC Press: Boca Raton, FL, 2017.
- Krier, C.; Rossi, F.; François, D.; Verleysen, M. A Data-Driven Functional Projection Approach for the Selection of Feature Ranges in Spectra with ICA or Cluster Analysis. *Chemom. Intell. Lab. Syst.* **2008**, *91*, 43–53.
- Centi, G.; Cavani, F.; Trifiro, F. *Selective Oxidation by Heterogenous Catalysis*; Kluwer/Plenum: New York, 2001.
- Clerici, M. G.; Kholdeeva, O. A. *Liquid Phase Oxidation via Heterogeneous Catalysis*; John Wiley & Sons, Inc.: Hoboken, NJ, 2013.
- Hermans, I.; Peeters, J.; Jacobs, P. A. Autoxidation of Hydrocarbons: From Chemistry to Catalysis. *Top. Catal.* **2008**, *50*, 124–132.
- De Klerk, A. Continuous-Mode Thermal Oxidation of Fischer-Tropsch Waxes. *Ind. Eng. Chem. Res.* **2003**, *42*, 6545–6548.
- Govindan, V.; Suresh, A. K. Modeling Liquid-Phase Cyclohexane Oxidation. *Ind. Eng. Chem. Res.* **2007**, *46*, 6891–6898.
- Siddiquee, M. N.; de Klerk, A. In Situ Measurement of Liquid Phase Oxygen during Oxidation. *Ind. Eng. Chem. Res.* **2016**, *55*, 6607–6618.
- Gemoets, H. P. L.; Su, Y.; Shang, M.; Hessel, V.; Luque, R.; Noël, T. Liquid Phase Oxidation Chemistry in Continuous-Flow Microreactors. *Chem. Soc. Rev.* **2016**, *45*, 83–117.
- Sobieszuk, P.; Aubin, J.; Pohorecki, R. Hydrodynamics and Mass Transfer in Gas-Liquid Flows in Microreactors. *Chem. Eng. Technol.* **2012**, *35*, 1346–1358.
- Neuenschwander, U.; Jensen, K. F. Olefin Autoxidation in Flow. *Ind. Eng. Chem. Res.* **2014**, *53*, 601–608.
- Vanoye, L.; Aloui, A.; Pablos, M.; Philippe, R.; Percheron, A.; Favre-Reguillon, A.; Bellefon, C. De. A Safe and Efficient Flow Oxidation of Aldehydes with O 2. *Org. Lett.* **2013**, *15*, 5978–5981.
- Hamano, M.; Nagy, K. D.; Jensen, K. F. Continuous Flow Metal-Free Oxidation of Picolines Using Air. *Chem. Commun.* **2012**, *48*, 2086–2088.

- Marre, S.; Adamo, A.; Basak, S.; Aymonier, C.; Jensen, K. F. Design and Packaging of Microreactors for High Pressure and High Temperature Applications. *Ind. Eng. Chem. Res.* **2010**, *49*, 11310–11320.
- Jevtic, R.; Ramachandran, P. A.; Dudukovic, M. P. Capillary Reactor for Cyclohexane Oxidation with Oxygen. *Chem. Eng. Res. Des.* **2010**, *88*, 255–262.
- Leclerc, A.; Alam, M.; Schweich, D.; Pouteau, P.; Bellefon, C. De. Gas – Liquid Selective Oxidations with Oxygen under Explosive Conditions in a Micro-Structured Reactor. *Lab Chip* **2008**, *8*, 814–817.
- Fischer, J.; Lange, T.; Boehling, R.; Rehfinger, A.; Klemm, E. Uncatalyzed Selective Oxidation of Liquid Cyclohexane with Air in a Microcapillary Reactor. *Chem. Eng. Sci.* **2010**, *65*, 4866–4872.
- Günther, A.; Khan, S. A.; Thalmann, M.; Trachsel, F.; Jensen, K. F. Transport and Reaction in Microscale Segmented Gas–Liquid Flow. *Lab Chip* **2004**, *4*, 278–286.
- Triplett, K. A.; Ghiaasiaan, S. M.; Abdel-Khalik, S. I.; Sadowski, D. L. Gas–Liquid Two-Phase Flow in Microchannels Part I: Two-Phase Flow Patterns. *Int. J. Multiph. Flow* **1999**, *25*, 377–394.
- Yao, C.; Dong, Z.; Zhao, Y.; Chen, G. An Online Method to Measure Mass Transfer of Slug Flow in a Microchannel. *Chem. Eng. Sci.* **2014**, *112*, 15–24.
- Yue, J.; Luo, L.; Gonthier, Y.; Chen, G.; Yuan, Q. An Experimental Study of Air – Water Taylor Flow and Mass Transfer inside Square Microchannels. *Chem. Eng. Sci.* **2009**, *64*, 3697–3708.
- Fries, D. M.; Trachsel, F.; Rohr, P. R. Von. Segmented Gas – Liquid Flow Characterization in Rectangular Microchannels. *Int. J. Multiph. Flow* **2008**, *34*, 1108–1118.
- Yun, J.; Lei, Q.; Zhang, S.; Shen, S.; Yao, K. Slug Flow Characteristics of Gas-Miscible Liquids in a Rectangular Microchannel with Cross and T-Shaped Junctions. *Chem. Eng. Sci.* **2010**, *65*, 5256–5263.
- Yan, X.; Xiaogang, S. Simple Linear Regression. In *Linear Regression Analysis: Theory and Computing*; World Scientific, 2009; pp 9–39.
- Freedman, D. *Statistical Models: Theory and Practice*; Cambridge University Press: Cambridge, 2009.

- Lai, T. L.; Robbins, H.; Wei, C. Z. Strong Consistency of Least Squares Estimates in Multiple Regression II. *J. Multivar. Anal.* **1979**, *9*, 343–361.
- Lange, K. L.; Little, R. J. A.; Taylor, J. M. G. Robust Statistical Modeling Using the t Distribution. *J. Am. Stat. Assoc.* **1989**, *84*, 881–896.
- Adeboye, N. O.; Fagoyinbo, I. S.; Olatayo, T. O. Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. *J. Math.* **2014**, *10*, 16–20.
- Brannick, M. *Regression with Two Independent Variables*; University of South Florida.
- Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons, Inc.: New York, 1980.
- Midi, H.; Sarkar, S. K.; Rana, S. Collinearity Diagnostics of Binary Logistic Regression Model. *J. Interdiscip. Math.* **2010**, *13*, 253–267.
- Johnston, J. *Econometric Methods*, 3rd ed.; McGraw-Hill: New York, 1984.
- Mela, C. F.; Kopalle, P. K. The Impact of Collinearity on Regression Analysis: The Asymmetric Effect of Negative and Positive Correlations. *Appl. Econ.* **2002**, *34*, 667–677.
- Siddiquee, M. N.; De Klerk, A. Hydrocarbon Addition Reactions during Low-Temperature Autoxidation of Oilsands Bitumen. *Energy Fuels* **2014**, *28*, 6848–6859.
- Montgomery, D. C.; Peck, E. A.; Vining, G. G. *Introduction to Linear Regression Analysis*, 5th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, 2012.
- Vatcheva, K. P.; MinJae, L.; McCormick, J. B.; Rahbar, M. H. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiol. Open Access* **2016**, *6*, 227.
- Kowalski, B. R. *Chemometrics*; Springer Netherlands: Dordrecht, 1984.
- Thode, H. C. *Testing for Normality*; Marcel Dekker: New York, 2002.
- Zaiontz, C. Pearson's Correlation Table http://www.real-statistics.com/statistics-tables/pearsons-correlation-table/ (accessed May 7, 2018).
- Dinov, I. D. Statistics Online Computational Resource (SOCR). *UCLA Statistics, Neurology, LONI* 2016.
- Green, P. E.; Tull, D. S.; Albaum, G. S. *Research for Marketing Decisions*, 5th ed.; Prentice Hall: Englewood Cliffs,NJ, 1988.
- Berry, W.; Feldman, S. *Multiple Regression in Practice*; SAGE Publications, Inc.: Thousand Oaks, CA, 1985. https://doi.org/10.4135/9781412985208.

- Tull, D. S.; Hawkins, D. I. *Marketing Research: Measurement and Method*; Maxwell Macmillan International: New York, 1993.
- Lehmann, D. R.; Gupta, S.; Steckel, J. H. *Marketing Research*; Addison-Weasley: Reading, MA, 1998.
- Field, A. P. *Discovering Statistics Using SPSS*; SAGE Publications, Inc.: London, 2005.
- Stewart, G. W. Collinearity and Least Squares Regression. *Stat. Sci.* **1987**, *2*, 68–84.
- Stitt, E. H. Alternative Multiphase Reactors for Fine Chemicals: A World beyond Stirred Tanks? *Chem. Eng. J.* **2002**, *90*, 47–60.
- Havelka, P.; Linek, V.; Sinkule, J.; Zahradník, J.; Fialová, M. Hydrodynamic and Mass Transfer Characteristics of Ejector Loop Reactors. *Chem. Eng. Sci.* **2000**, *55*, 535–549.
- Piegorsch, W. W. *Tables of P-Values for t- and Chi-Square Reference Distributions*; University of South Carolina, Columbia, 2002.
- Kutner, M.; Nachtsheim, C.; Neter, J. *Applied Linear Statistical Models*; McGraw-Hill: Irwin, 2004.
- Cramers, P. H. M. R.; Beenackers, A. A. C. M. Influence of the Ejector Configuration, Scale and the Gas Density on the Mass Transfer Characteristics of Gas-Liquid Ejectors. *Chem. Eng. J.* **2001**, *82*, 131–141.
- Snead, D. R.; Jamison, T. F. A Three-Minute Synthesis and Purification of Ibuprofen: Pushing the Limits of Continuous-Flow Processing. *Angew. Chemie - Int. Ed.* **2015**, *54*, 983–987.
- Macchietto, S.; Hewitt, G. F.; Coletti, F.; Crittenden, B. D.; Dugwell, D. R.; Galindo, A.; Jackson, G.; Kandiyoti, R.; Kazarian, S. G.; Luckham, P. F.; Matar, O. K.; Millan-Agorio, M.; Müller, E. A.; Paterson, W.; Pugh, S. J.; Richardson, S. M.; Wilson, D. I. Fouling in Crude Oil Preheat Trains: A Systematic Solution to an Old Problem. *Heat Transf. Eng.* **2011**, *32*, 197–215.
- Derakhshesh, M. Asphaltene Aggregation and Fouling Behavior. Ph.D. Thesis, University of Alberta, September 2012.
- Strausz, O. P.; Lown, E. M. *The Chemistry of Alberta Oil Sands, Bitumens and Heavy Oils*; Alberta Energy Research Institute: Calgary, 2003.
- Mullins, O. C.; Sabbah, H.; Eyssautier, J.; Pomerantz, A. E.; Barré, L.; Andrews, A. B.; Ruiz-Morales, Y.; Mostowfi, F.; McFarlane, R.; Goual, L.; Lepkowicz, R.; Cooper, T.; Orbulescu,

J.; Leblanc, R. M.; Edwards, J.; Zare, R. N. Advances in Asphaltene Science and the Yen-Mullins Model. *Energy Fuels* **2012**, *26*, 3986–4003.

- Mullins, O. C.; Sheu, E. Y.; Hammami, A.; Marshall, A. G. *Asphaltenes, Heavy Oils, and Petroleomics*; Springer: New York, 2006.
- Brons, G.; Yu, J. M. Solvent Deasphalting Effects on Whole Cold Lake Bitumen. *Energy Fuels* **1995**, *9*, 641–647.
- Hart, A. A Review of Technologies for Transporting Heavy Crude Oil and Bitumen via Pipelines. *J. Pet. Explor. Prod. Technol.* **2014**, *4*, 327–336.
- From Diluted Bitumen to Synthetic Crude: Upgrading Explained https://www.oilsandsmagazine.com/technical/bitumen-upgrading (accessed Jan 5, 2019).
- Maples, R. E. *Petroleum Refinery Process Economics*; PennWell: Tulsa, OK, 1993.
- Rana, M. S.; Sámano, V.; Ancheyta, J.; Diaz, J. A. I. A Review of Recent Advances on Process Technologies for Upgrading of Heavy Oils and Residua. *Fuel* **2007**, *86*, 1216–1231.
- Speight, J. G. *The Chemistry and Technology of Petroleum*; Marcel Dekker: New York, 1991.
- Biasca, F. E.; Dickenson, R. L.; Chang, E.; Johnson, H. E.; Bailey, R. T.; Simbeck, D. R. Future Technology In Heavy Oil Processing. In *Upgrading heavy crude oils and residue to transportation fuel: Technology, Economics, and Outlook*; SFA Pacific Inc.: Palo Alto, CA, 2009.
- Raseev, S. *Thermal and Catalytic Processes in Petroleum Refining*; Marcel Dekker: New York, 2003.
- Tokarska, A. Investigations on the Processing Vacuum Residue and Its Mixtures and Coal Tars of Oil with Coal of Crude Materials. *Fuel* **1996**, *75*, 1094–1100.
- Mcketta Jr, J. J. *Petroleum Processing Handbook*; Marcel Dekker: New York, 1992.
- Shu, W. R.; Venkatesan, V. N. Kinetics of Thermal Visbreaking of a Cold Lake Bitumen. *J. Can. Pet. Technol.* **1984**, *23*, 60–64.\
- Henderson, J. H.; Weber, L. Physical Upgrading of Heavy Crude Oils by the Application of Heat. *J. Can. Pet. Technol.* **1965**, *4*, 206–212.
- Castillo, J.; De Klerk, A. Visbreaking of Deasphalted Oil from Bitumen at 280–400 °C. *Energy Fuels* **2018**, *33*, 159–175.

- Dutta, R. P.; McCaffrey, W. C.; Gray, M. R.; Muehlenbachs, K. Thermal Cracking of Athabasca Bitumen: Influence of Steam on Reaction Chemistry. *Energy Fuels* **2000**, *14*, 671–676.
- Prado, G. H. C.; Riya; Hyrve, M.; De Klerk, A. Role of Metal Halides in Coke Formation during Bitumen Upgrading. *Fuel* **2018**, *211*, 775–782.
- Vilcáez, J.; Watanabe, M.; Watanabe, N.; Kishita, A.; Adschiri, T. Hydrothermal Extractive Upgrading of Bitumen without Coke Formation. *Fuel* **2012**, *102*, 379–385.
- Junaid, A. S. M.; Wang, W.; Street, C.; Rahman, M.; McCaffrey, W.; Kuznicki, S. Upgrading of Athabasca Oilsands Bitumen by Low Temperature Natural Zeolite Cracking. *10AIChE - 2010 AIChE Annu. Meet. Conf. Proc.* **2010**, *4*.
- Joshi, J. B.; Pandit, A. B.; Kataria, K. L.; Kulkarni, R. P.; Sawarkar, A. N.; Tandon, D.; Ram, Y.; Kumar, M. M. Petroleum Residue Upgradation via Visbreaking: A Review. *Ind. Eng. Chem. Res.* **2008**, *47*, 8960–8988.
- Washimi, K.; Limmer, H. New Unit to Thermal Crack Resid. *Hydrocarb. Process.* **1989**, *68*, 69.
- Pauling, L. *The Chemical Bond: A Brief Introduction to Modern Structural Chemistry*; Cornell University Press: Ithaca, NY, 1967.
- Smith, J. W. Hydrogen-Bonding and Complex-Forming Properties. In *The Carbon-Halogen Bond (1973)*; John Wiley & Sons, Ltd.: Chichester, UK, 1973; pp 265–300.
- Prado, G. H. C.; De Klerk, A. Halogenation of Oilsands Bitumen, Maltenes, and Asphaltenes. *Energy Fuels* **2014**, *28*, 4458–4468.
- Lesueur, D. The Colloidal Structure of Bitumen: Consequences on the Rheology and on the Mechanisms of Bitumen Modification. *Adv. Colloid Interface Sci.* **2009**, *145*, 42–82.
- Dealy, J. M. Rheological Properties of Oil Sand Bitumens. *Can. J. Chem. Eng.* **1979**, *57*, 677–683.
- Niizuma, S.; Steele, C. T.; Gunning, H. E.; Strausz, O. P. Electron Spin Resonance Study of Free Radicals in Athabasca Asphaltene. *Fuel* **1977**, *56*, 249–256.
- Yen, T. F.; Erdman, J. G.; Saraceno, A. J. Investigation of the Nature of Free Radicals in Petroleum Asphaltenes and Related Substances by Electron Spin Resonance. *Anal. Chem.* **1962**, *34*, 694–700.

- Blanchard, C. M.; Gray, M. R. Free Radical Chain Reactions of Bitumen Residue. *ACS Div. Fuel Chem. Prepr.* **1997**, *42*, 137–141.
- Trukhan, S. N.; Yudanov, V. F.; Gabrienko, A. A.; Subramani, V.; Kazarian, S. G.; Martyanov, O. N. In Situ Electron Spin Resonance Study of Molecular Dynamics of Asphaltenes at Elevated Temperature and Pressure. *Energy Fuels* **2014**, *28*, 6315–6321.
- Gabrienko, A. A.; Lai, C. H.; Kazarian, S. G. In Situ Chemical Imaging of Asphaltene Precipitation from Crude Oil Induced by *n*-Heptane. *Energy Fuels* **2014**, *28*, 964–971.
- Zachariah, A.; Wang, L.; Yang, S.; Prasad, V.; De Klerk, A. Suppression of Coke Formation during Bitumen Pyrolysis. *Energy Fuels* **2013**, *27*, 3061–3070.
- Babu, D. R.; Cormack, D. E. Effect of Low-Temperature Oxidation on the Composition of Athabasca Bitumen. *Fuel* **1984**, *63*, 858–861.
- Petersen, J. C.; Harnsberger, P. M.; Robertson, R. E. Factors Affecting the Kinetics and Mechanisms of Asphalt Oxidation and the Relative Effects of Oxidation Products on Age Hardening. *ACS Div. Fuel Chem. Prepr.* **1996**, *41*, 1232–1241.
- Ancheyta, J.; Centeno, G.; Trejo, F.; Marroquín, G.; García, J. A.; Tenorio, E.; Torres, A. Extraction and Characterization of Asphaltenes from Different Crude Oils and Solvents. *Energy Fuels* **2002**, *16*, 1121–1127.
- Strausz, O. P.; Peng, P.; Murgich, J. About the Colloidal Nature of Asphaltenes and the MW of Covalent Monomeric Units. *Energy Fuels* **2002**, *16*, 809–822.
- Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs. *Energy Fuels* **2017**, *31*, 13509–13518.
- Speight, J. G. *Fouling in Refineries*; Elsevier: Amsterdam, 2015.
- Turuga, A. S. S. Effect of Solvent Deasphalting Process on the Properties of Deasphalted Oil and Asphaltenes from Bitumen. MSc. Thesis, University of Alberta, 2017.
- Wang, J.; Buckley, J. *Standard Procedure for Separating Asphaltenes from Crude Oils*; New Mexico, 2002.
- Luo, P.; Wang, X.; Gu, Y. Characterization of Asphaltenes Precipitated with Three Light Alkanes under Different Experimental Conditions. *Fluid Phase Equilib.* **2010**, *291*, 103–110.
- ASTM D189-06 (Reapproved 2014). *Standard Test Method for Conradson Carbon Residue of Petroleum Products*; ASTM: West Conshohocken, PA, 2014.

- Noel, F. Alternative to the Conradson Carbon Residue Test. *Fuel* **1984**, *63*, 931–934.
- ASTM D7169-16. Standard Test Method for Boiling Point Distribution of Samples with Residues Such as Crude Oils and Atmospheric and Vacuum Residues by High Temperature Gas Chromatography. ASTM: West Conshohocken, PA 2016.
- Bielski, B. H. J.; Gebicki, J. M. *Atlas of Electron Spin Resonance Spectra*; Academic Press: New York, 1967.
- Cui, Q.; Nakabayashi, K.; Ma, X.; Miyawaki, J.; Ideta, K.; Tennichi, Y.; Ueda, M.; Al-Mutairi, A.; Marafi, A. M. J.; Park, J. II; Yoon, S. H.; Mochida, I. Studying Rotational Mobility of V=O Complexes in Atmospheric Residues and Their Resins and Asphaltenes by Electron Spin Resonance. *Energy Fuels* **2017**, *31*, 4748–4757.
- Malhotra, V. M.; Buckmaster, H. A. 9 and 34 GHz EPR Study of the Free Radicals in Various Asphaltenes: Statistical Correlation of the g-Values with Heteroatom Content. *Org. Geochem.* **1985**, *8*, 235–239.
- Khulbe, K. C.; Mann, R. S.; Lamarche, G.; Lamarche, A. M. Electron Spin Resonance Study of the Thermal Decomposition of Solvent Extracted Athabasca Tar Sand Bitumen. *Fuel Process. Technol.* **1992**, *31*, 91–103.
- Beauchamp, J. Spectroscopy Tables https://www.cpp.edu/~psbeauchamp/pdf/spec_ir_nmr_spectra_tables.pdf (accessed Oct 12, 2018).
- Saoula, S.; Soudani, K.; Haddadi, S.; Munoz, M. E.; Santamaria, A. Analysis of the Rheological Behavior of Aging Bitumen and Predicting the Risk of Permanent Deformation of Asphalt. *Mater. Sci. Appl.* **2013**, *4*, 312–318.
- Rek, V.; Barjaktarovi, Z. M. Dynamic Mechanical Behavior of Polymer Modified Bitumen. *Mater. Res. Innov.* **2002**, *6*, 39–43.
- Abivin, P.; Taylor, S. D.; Freed, D. Thermal Behavior and Viscoelasticity of Heavy Oils. *Energy Fuels* **2012**, *26*, 3448–3461.
- Behzadfar, E.; Hatzikiriakos, S. G. Rheology of Bitumen: Effects of Temperature, Pressure, CO_2 Concentration and Shear Rate. *Fuel* **2014**, *116*, 578–587.
- McNab, J. G.; Smith, P. V. J.; Betts, R. L. Evolution of Petroleum. *Ind. Eng. Chem.* **1952**, *44*, 2556–2563.

- Schramm, L. L.; Kwak, J. C. T. Rheological Properties of an Athabasca Bitumen and Some Bituminous Mixtures and Dispersions. *J. Can. Pet. Technol.* **1988**, *27*, 26–35.
- Wallace, D.; Henry, D. A Correlation for Correcting the Viscosity of Solvent-Extracted Bitumen to a Solvent-Free Basis. *AOSTRA J. Res.* **1987**, *3*, 239–248.
- Masliyah, J. H.; Xu, Z.; Czarnecki, J. A. *Handbook on Theory and Practice of Bitumen Recovery from Athabasca Oil Sands*; Kingsley Knowledge Publishing, 2011.
- Selucky, M. L.; Chu, Y.; Ruo, T.; Strausz, O. P. Chemical Composition of Athabasca Bitumen. *Fuel* **1977**, *56*, 369–381.
- Selucky, M. L.; Chu, Y.; Ruo, T. C. S.; Strausz, O. P. Chemical Composition of Cold Lake Bitumen. *Fuel* **1978**, *57*, 9–16.
- McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. Heavy Petroleum Composition. 4. Asphaltene Compositional Space. *Energy Fuels* **2013**, *27*, 1257–1267.
- Poveda, J. C.; Molina, D.; Martínez, H.; Florez, O.; Campillo, B. Molecular Changes in Asphaltenes within H_2 Plasma. *Energy Fuels* **2014**, *28*, 735–744.
- Pomerantz, A. E.; Hammond, M. R.; Morrow, A. L.; Mullins, O. C.; Zare, R. N. Two-Step Laser Mass Spectrometry of Asphaltenes. *J. Am. Chem. Soc.* **2008**, *130*, 7216–7217.
- Friend, J. N.; Hargreaves, W. D. Viscosity and the Hydrogen Bond. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **1945**, *36*, 731–756.
- Smith, D. F.; Schaub, T. M.; Kim, S.; Rodgers, R. P.; Rahimi, P.; Teclemariam, A.; Marshall, A. G. Characterization of Acidic Species in Athabasca Bitumen and Bitumen Heavy Vacuum Gas Oil by Negative-Ion ESI FT-ICR MS with and without Acid-Ion Exchange Resin Prefractionation. *Energy Fuels* **2008**, *22*, 2372–2378.
- Erdman, J. G.; Dickie, J. P. Mild Thermal Alteration of Asphaltic 702. *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **1964**, *9*, 69–79.
- Ball, M. W. Development of the Athabaska Oil Sands. *Can. Inst. Min. Metall.* **1941**, *44*, 58–91.
- Steacie, E. W. R. *Atomic and Free Radical Reactions*, 2nd ed.; Reinhold: New York, 1954.
- Jha, K. N.; Montgomery, D. S.; Strausz, O. P. Chemical Composition of Gases in Alberta Bitumens and in Low-Temperature Thermolysis of Oil Sand Asphaltenes and Maltenes. In *Oil Sand and Oil Shale Chemistry*; Strausz, O. P., Lown, E. M., Eds.; Verlag Chemie: New York, 1978; pp 33–54.

- Dunn, G. E.; Thimm, H. F.; Mohanty, R. K. Kinetics and Mechanism of Decarboxylation of Some Pyridinecarboxylic Acids in Aqueous Solution. III. 3-Hydroxy- and 3-Aminopyridine-2-Carboxylic Acids. *Can. J. Chem.* **1979**, *57*, 1098–1104.
- Montoya Sánchez, N.; De Klerk, A. Oxidative Ring-Opening of Aromatics: Decomposition of Biphenyl Carboxylic Acids and Zinc Biphenyl Carboxylates. *Energy Fuels* **2015**, *29*, 7910–7922.
- Sachanan, A. N. *Conversion of Petroleum. Production of Motor Fuels by Thermal and Catalytic Processes*, 2nd ed.; Reinhold: New York, 1948.
- Govindhakannan, J.; Khulbe, C. Effect of Pressure on Delayed Coking of Athabasca Bitumen. *Prepr. Am. Chem. Soc. Div. Pet. Chem.* **2010**, *55*, 44–45.
- Carbognani, L.; Lubkowitz, J.; Gonzalez, M. F.; Pereira-Almao, P. High Temperature Simulated Distillation of Athabasca Vacuum Residue Fractions. Bimodal Distributions and Evidence for Secondary "on-Column" Cracking of Heavy Hydrocarbons. *Energy Fuels* **2007**, *21*, 2831–2839.
- Lee, J. M.; Baker, J. J.; Rolle, J. G.; Llerena, R. Comparison of Fuel Properties of Petroleum Cokes and Coals Used in Power Generation. *Abstr. Pap. Am. Chem. Soc.* **1999**, 80–89.
- Mohapatra, D. P.; Kirpalani, D. M. Bitumen Heavy Oil Upgrading by Cavitation Processing: Effect on Asphaltene Separation, Rheology, and Metal Content. *Appl. Petrochemical Res.* **2016**, *6*, 107–115.
- Phillips, C. R.; Haidar, N. I.; Poon, Y. C. Kinetic Models for the Thermal Cracking of Athabasca Bitumen. *Fuel* **1985**, *64*, 678–691.
- Millour, J. P.; Moore, R. G.; Bennion, D. W.; Ursenbach, M. G.; Gie, D. N. A Simple Implicit Model for Thermal Cracking of Crude Oils. In *SPE Annual Technical Conference and Exhibition*; Las Vegas, 1985; pp 1–15.
- Brauch, R.; Fainberg, V.; Kalchouck, H.; Hetsroni, G. Correlations between Properties of Various Feedstocks and Products of Visbreaking. *Fuel Sci. Technol. Int.* **1996**, *14*, 753–765.
- Hasan, M. A.; Shaw, J. M. Rheology of Reconstituted Crude Oils: Artifacts and Asphaltenes. *Energy Fuels* **2010**, *24*, 6417–6427.
- Mehrotra, A. K.; Eastick, R. R.; Svrcek, W. Y. Viscosity of Cold Lake Bitumen and Its Fractions. *Can. J. Chem. Eng.* **1989**, *67*, 1004–1009.

- Kapadia, P. R.; Kallos, M. S.; Gates, I. D. A New Kinetic Model for Pyrolysis of Athabasca Bitumen. *Can. J. Chem. Eng.* **2013**, *91*, 889–901.
- Poutsma, M. L. Free-Radical Thermolysis and Hydrogenolysis of Model Hydrocarbons Relevant to Processing of Coal. *Energy Fuels* **1990**, *4*, 113–131.
- Hou, Y.; Wang, L.; Wang, D.; Guo, M.; Liu, P.; Yu, J. Characterization of Bitumen Micro-Mechanical Behaviors Using AFM, Phase Dynamics Theory and MD Simulation. *Materials (Basel).* **2017**, *10*, 1–16.
- Fischer, H.; Stadler, H.; Erine, N. Quantitative Temperature-Depending Mapping of Mechanical Properties of Bitumen at the Nanoscale Using the AFM Operated with PeakForce Tapping TM Mode. *J. Microsc.* **2013**, *250*, 210–217.
- Yu, X.; Zaumanis, M.; Dos Santos, S.; Poulikakos, L. D. Rheological, Microscopic, and Chemical Characterization of the Rejuvenating Effect on Asphalt Binders. *Fuel* **2014**, *135*, 162–171.
- Guo, M.; Huang, Y.; Wang, L.; Yu, J.; Hou, Y. Using Atomic Force Microscopy and Molecular Dynamics Simulation to Investigate the Asphalt Micro Properties. *Int. J. Pavement Res. Technol.* **2018**, *11*, 321–326.
- Schmets, A.; Kringos, N.; Pauli, T.; Redelius, P.; Scarpas, T. On the Existence of Wax-Induced Phase Separation in Bitumen. *Int. J. Pavement Eng.* **2010**, *11*, 555–563.
- Fischer, H. R.; Cernescu, A. Relation of Chemical Composition to Asphalt Microstructure - Details and Properties of Micro-Structures in Bitumen as Seen by Thermal and Friction Force Microscopy and by Scanning near-Filed Optical Microscopy. *Fuel* **2015**, *153*, 628–633.
- Grover Allen, R.; Little, D. N.; Bhasin, A.; Glover, C. J. The Effects of Chemical Composition on Asphalt Microstructure and Their Association to Pavement Performance. *International Journal of Pavement Engineering*. Taylor & Francis 2014, pp 9–22.
- Schuler, B.; Meyer, G.; Peña, D.; Mullins, O. C.; Gross, L. Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy. *J. Am. Chem. Soc.* **2015**, *137*, 9870–9876.
- Mullins, O. C. The Modified Yen Model. *Energy Fuels* **2010**, *24*, 2179–2207.
- Zhang, L.; Xie, L.; Shi, C.; Huang, J.; Liu, Q.; Zeng, H. Mechanistic Understanding of Asphaltene Surface Interactions in Aqueous Media. *Energy Fuels* **2017**, *31*, 3348–3357.

- Uzcátegui, G.; Fong, S. Y.; De Klerk, A. Cracked Naphtha Reactivity: Effect of Free Radical Reactions. *Energy Fuels* **2018**, *32*, 5812–5823.
- Sabbah, H.; Morrow, A. L.; Pomerantz, A. E.; Zare, R. N. Evidence for Island Structures as the Dominant Architecture of Asphaltenes. *Energy Fuels* **2011**, *25*, 1597–1604.
- Wang, W.; Taylor, C.; Hu, H.; Humphries, K. L.; Jaini, A.; Kitimet, M.; Scott, T.; Stewart, Z.; Ulep, K. J.; Houck, S.; Luxon, A.; Zhang, B.; Miller, B.; Parish, C. A.; Pomerantz, A. E.; Mullins, O. C.; Zare, R. N. Nanoaggregates of Diverse Asphaltenes by Mass Spectrometry and Molecular Dynamics. *Energy Fuels* **2017**, *31*, 9140–9151.
- Schuler, B.; Fatayer, S.; Meyer, G.; Rogel, E.; Moir, M.; Zhang, Y.; Harper, M. R.; Pomerantz, A. E.; Bake, K. D.; Witt, M.; Peña, D.; Kushnerick, J. D.; Mullins, O. C.; Ovalles, C.; Van Den Berg, F. G. A.; Gross, L. Heavy Oil Based Mixtures of Different Origins and Treatments Studied by Atomic Force Microscopy. *Energy Fuels* **2017**, *31*, 6856–6861.
- Zhang, H. T.; Li, R.; Yang, Z.; Yin, C. X.; Gray, M. R.; Bohne, C. Evaluating Steady-State and Time-Resolved Fluorescence as a Tool to Study the Behavior of Asphaltene in Toluene. *Photochem. Photobiol. Sci.* **2014**, *13*, 917–928.
- Tanaka, R.; Sato, E.; Hunt, J. E.; Winans, R. E.; Sato, S.; Takanohashi, T. Characterization of Asphaltene Aggregates Using X-Ray Diffraction and Small-Angle X-Ray Scattering. *Energy Fuels* **2004**, *18*, 1118–1125.
- Hosseini-Dastgerdi, Z.; Tabatabaei-Nejad, S. A. R.; Khodapanah, E.; Sahraei, E. A Comprehensive Study on Mechanism of Formation and Techniques to Diagnose Asphaltene Structure; Molecular and Aggregates: A Review. *Asia-Pacific J. Chem. Eng.* **2015**, *10*, 1–14.
- Gawrys, K. L.; Blankenship, G. A.; Kilpatrick, P. K. Solvent Entrainment in and Flocculation of Asphaltenic Aggregates Probed by Small-Angle Neutron Scattering. *Langmuir* **2006**, *22*, 4487–4497.
- Rueda-Velásquez, R. I.; Freund, H.; Qian, K.; Olmstead, W. N.; Gray, M. R. Characterization of Asphaltene Building Blocks by Cracking under Favorable Hydrogenation Conditions. *Energy Fuels* **2013**, *27*, 1817–1829.
- Karimi, A.; Qian, K.; Olmstead, W. N.; Freund, H.; Yung, C.; Gray, M. R. Quantitative Evidence for Bridged Structures in Asphaltenes by Thin Film Pyrolysis. *Energy Fuels* **2011**, *25*, 3581–3589.

- Chacón-Patiño, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 2: Selective Separation Method That Reveals Fractions Enriched in Island and Archipelago Structural Motifs by Mass Spectrometry. *Energy Fuels* **2018**, *32*, 314–328.
- Chacón-Patino, M. L.; Rowland, S. M.; Rodgers, R. P. Advances in Asphaltene Petroleomics. Part 3. Dominance of Island or Archipelago Structural Motif Is Sample Dependent. *Energy Fuels* **2018**, *32*, 9106–9120.
- Podgorski, D. C.; Corilo, Y. E.; Nyadong, L.; Lobodin, V. V.; Bythell, B. J.; Robbins, W. K.; McKenna, A. M.; Marshall, A. G.; Rodgers, R. P. Heavy Petroleum Composition. 5. Compositional and Structural Continuum of Petroleum Revealed. *Energy Fuels* **2013**, *27*, 1268–1276.
- Zhao, H. Y.; Cao, Y.; Sit, S. P.; Lineberry, Q.; Pan, W. P. Thermal Characteristics of Bitumen Pyrolysis. *J. Therm. Anal. Calorim.* **2012**, *107*, 541–547.
- Taş, S.; Yürüm, Y. Co-Firing of Biomass with Coals: Part 2. Thermogravimetric Kinetic Analysis of Co-Combustion of Fir (Abies Bornmulleriana) Wood with Beypazari Lignite. *J. Therm. Anal. Calorim.* **2011**, *107*, 293–298.
- Shao, J.; Yan, R.; Chen, H.; Wang, B.; Lee, D. H.; Liang, D. T. Pyrolysis Characteristics and Kinetics of Sewage Sludge by Thermogravimetry Fourier Transform Infrared Analysis. *Energy Fuels* **2008**, *22*, 38–45.
- Boucher, J. L.; Wang, I.-H.; Martinez, D. F. Elimination Chemistry in Asphalt. *Prepr. Pap. − Am. Chem. Soc., Div. Pet. Chem.* **1990**, *35*, 550–555.
- Petersen, J. C. An Infrared Study of Hydrogen Bonding in Asphalt. *Fuel* **1967**, *46*, 295–305.
- Allen, F. H.; Wood, P. A.; Galek, P. T. A. Role of Chloroform and Dichloromethane Solvent Molecules in Crystal Packing: An Interaction Propensity Study. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2013**, *69*, 379–388.
- Ding, X.; Tuikka, M.; Haukk, M. Halogen Bonding in Crystal Engineering. In *Recent Advances in Crystallography*; InTech, 2012.
- Calkins, W. H. The Chemistry of Sulfur in Coal - A Historical Perspective. *Prepr. Fuel Chem. Div., Am. Chem. Soc.* **1993**, *32* (2), 358–368.
- Brammer, L.; Bruton, E. A.; Sherwood, P. Understanding the Behavior of Halogens as Hydrogen Bond Acceptors. *Cryst. Growth Des.* **2001**, *1*, 277–290.

- Blyler, L. L.; Haas, T. W. The Influence of Intermolecular Hydrogen Bonding on the Flow Behavior of Polymer Melts. *J. Appl. Polym. Sci.* **1969**, *13*, 2721–2733.
- Bellamy, L. J.; Hallam, H. E.; William, R. J. Infra-Red Spectra and Solvent Effects. Part 1.—X—H Stretching Frequencies. *Trans. Faraday Soc.* **1958**, *54*, 1120–1127.
- Koyo, A. I. D. A. Raman Spectra of Liquid Ammonia Solutions. VI : Liquid Ammonia and a Few Alkyl Chlorides. *Sci. reports Res. Institutes, Tohoku Univ. Ser. A, Physics, Chem. Metall.* **1957**, *9*, 458–466.
- Earp, D. P.; Glasstone, S. Dielectric Polarisation and Molecular-Compound Formation in Solution. Parts I and II. *J. Chem. Soc.* **1935**, 1709–1723.
- Vogtle, F.; Weber, E. Crown Ethers-Complexes and Selectivity. In *Crown Ethers and Analogs*; Weber, E., Toner, J. L., Goldberg, I., Vgtle, F., Laidler, D. A., Stoddart, J. F., Bartsch, R. A., Liotta, C. L., Eds.; John Wiley & Sons, Inc.: Chichester, UK, 1989; pp 207–304.
- McBride, M. T.; Luo, T. J. M.; Tayhas R Palmore, G. Hydrogen-Bonding Interactions in Crystalline Solids of Cyclic Thioureas. *Cryst. Growth Des.* **2001**, *1*, 39–46.
- Davey, R. J.; Blagden, N.; Righini, S.; Alison, H.; Quayle, M. J.; Fuller, S. Crystal Polymorphism as a Probe for Molecular Self-Assembly during Nucleation from Solutions: The Case of 2,6-Dihydroxybenzoic Acid. *Cryst. Growth Des.* **2001**, *1*, 59–65.
- Thallapally, P. K.; Nangia, A. A Cambridge Structural Database Analysis of the C–H···Cl Interaction: C–H···Cl− and C–H···Cl–M Often Behave as Hydrogen Bonds but C–H···Cl–C Is Generally a van Der Waals Interaction. *CrystEngComm* **2001**, *3*, 114–119.
- Murai, H.; Torres, M.; Strausz, O. P. Quintet State Triplet-Triplet Radical Pairs. 2. *J. Am. Chem. Soc.* **1980**, *102*, 7390–7391.
- Soos, Z. G.; Hughes, R. C. Paramagnetic Susceptibilities and Temperature-Dependent Excitation Energies in Linear Organic Crystals. *J. Chem. Phys.* **1967**, *46*, 253–259.
- Rudnick, L. R.; Audeh, C. A. Thermal Chemistry of a Utah Tar Sand: An Electron Spin Resonance (ESR) Characterization. *Ind. Eng. Chem. Res.* **1988**, *27*, 1366–1369.
- Behzadfar, E.; Hatzikiriakos, S. G. Viscoelastic Properties and Constitutive Modelling of Bitumen. *Fuel* **2013**, *108*, 391–399.
- Sunthar, P. Polymer Rheology. In *Rheology of Complex Fluids*; Springer: New York, 2010; pp 171–191.

- Little, D. N.; Prapnnachari, S.; Letton, A.; Kim, Y. R. *Investigation of the Microstructural Mechanism of Relaxation and Fracture Healing in Asphalt*; College Station, Texas, 1993.
- Silbert, L. E.; Melrose, J. R. The Rheology and Microstructure of Concentrated, Aggregated Colloids. *J. Rheol. (N. Y. N. Y).* **1999**, *43*, 673–700.
- Saal, R. N. J.; Labout, J. W. A. Rheological Properties of Asphaltic Bitumen. *J. Phys. Chem.* **1940**, *44*, 149–165.
- Dwiggins, C. W. A Small Angle X-Ray Scattering Study of the Colloidal Nature of Petroleum. *J. Phys. Chem.* **1965**, *69*, 3500–3506.
- Hasan, S. W.; Ghannam, M. T.; Esmail, N. Heavy Crude Oil Viscosity Reduction and Rheology for Pipeline Transportation. *Fuel* **2010**, *89*, 1095–1100.
- Mouazen, M.; Poulesquen, A.; Vergnes, B. Correlation between Thermal and Rheological Studies to Characterize the Behavior of Bitumen. *Rheol. Acta* **2011**, *50*, 169–178.
- Schabron, J. F.; Rovani, J. F. J. Method for Determining Asphaltene Stability of a Hydrocarbon-Containing Material. US 8,367,425 B1, 2013.
- Yamakawa, H. *Modern Theory of Polymer Solutions*; Harper & Row: New York, 1971.
- Iwao, T. *Polymer Solutions: An Introduction to Physical Properties*; Wiley: New York, 2002.
- Gawrys, K. L.; Blankenship, G. A.; Kilpatrick, P. K. On the Distribution of Chemical Properties and Aggregation of Solubility Fractions in Asphaltenes. *Energy Fuels* **2006**, *20*, 705–714.
- Anderson, D. A.; Christensen, D. W.; Bahia, H. U.; Dongre, R.; Sharma, M. G.; Antle, C. E. *Binder Characterization and Evaluation Volume 3: Physical Characterization*; Washington, D. C., 1994.
- Spiecker, P. M.; Gawrys, K. L.; Kilpatrick, P. K. Aggregation and Solubility Behavior of Asphaltenes and Their Subfractions. *J. Colloid Interface Sci.* **2003**, *267*, 178–193.
- Lesueur, D. Letter to the Editor: On the Thermorheological Complexity and Relaxation Modes of Asphalt Cements. *J. Rheol. (N. Y. N. Y).* **1999**, *43*, 1701–1704.
- Koschmieder, E. L. *Bénard Cells and Taylor Vortices*; Cambridge University Press: Cambridge, UK, 1993.
- Bird, R. B.; Stewart, W. E.; Lightfoot, E. N. *Transport Phenomena*; John Wiley & Sons, 2001.
- Belgrave, J. D. M.; Moore, R. G.; Ursenbach, M. G.; Bennion, D. W. A Comprehensive Approach to In-Situ Combustion Modeling. *SPE Adv. Technol. Ser.* **1993**, *1*, 98–107.

- Fortuna, L.; Graziani, S.; Rizzo, A.; Xibilia, M. G. *Soft Sensors for Monitoring and Control of Industrial Processes*; Springer Science & Business Media, 2007.
- Long, J.; Xu, Z.; Masliyah, J. H. Single Molecule Force Spectroscopy of Asphaltene Aggregates. *Langmuir* **2007**, *23*, 6182–6190.
- Yoon, S.; Bhatt, S. D.; Lee, W.; Lee, H. Y.; Jeong, S. Y.; Baeg, J. O.; Lee, C. W. Separation and Characterization of Bitumen from Athabasca Oil Sand. *Korean J. Chem. Eng.* **2009**, *26*, 64–71.
- Adebiyi, F. M.; Thoss, V. Spectroscopic Characterization of Asphaltene Fraction of Nigerian Bitumen. *Pet. Sci. Technol.* **2015**, *33*, 245–255.
- Varanda, C.; Portugal, I.; Ribeiro, J.; Silva, C. M.; Silva, A. M. S. NMR Spectroscopy in Bitumen Characterization. In *Analytical Characterization Methods for Crude Oiland Related Products*; John Wiley & Sons Ltd: Chichester, UK, 2017; pp 141–161.
- Ding, G.; Hou, Y.; Peng, J.; Shen, Y.; Jiang, M.; Bai, G. On-Line near-Infrared Spectroscopy Optimizing and Monitoring Biotransformation Process of γ-Aminobutyric Acid. *J. Pharm. Anal.* **2016**, *6*, 171–178.
- Blanco, M.; Villarroya, I. NIR Spectroscopy: A Rapid-Response Analytical Tool. *TrAC - Trends Anal. Chem.* **2002**, *21*, 240–250.
- Alves, J. C. L.; Henriques, C. B.; Poppi, R. J. Determination of Diesel Quality Parameters Using Support Vector Regression and near Infrared Spectroscopy for an In-Line Blending Optimizer System. *Fuel* **2012**, *97*, 710–717.
- Skibsted, E.; Engelsen, S. B. Spectroscopy for Process Analytical Technology (PAT). In *Encyclopedia of Spectroscopy and Spectrometry*; Elsevier, 2010; pp 2651–2661.
- Garrido, M.; Rius, F. X.; Larrechi, M. S. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) Applied to Spectroscopic Data from Monitoring Chemical Reactions Processes. *Anal. Bioanal. Chem.* **2008**, *390*, 2059–2066.
- Sivaramakrishnan, K.; De Klerk, A.; Prasad, V. Viscosity of Canadian Oilsands Bitumen and Its Modification by Thermal Conversion. In *Chemistry Solutions to Challenges in the Petroleum Industry (In Press)*; American Chemical Society, 2019.
- Jolliffe, I. T. *Principal Component Analysis*; Springer: Berlin, Heidelberg, 2011.

- Elbergali, A.; Nygren, J.; Kubista, M. Automated Procedure to Predict the Number of Components in Spectroscopic Data. *Doktorsavhandlingar vid Chalmers Tek. Hogsk.* **1999**, *379*, 143–158.
- Malinowski, E. R. Determination of the Number of Factors and the Experimental Error in a Data Matrix. *Anal. Chem.* **1977**, *49*, 612–617.
- Keller, H. R.; Massart, D. L. Evolving Factor Analysis. *Chemom. Intell. Lab. Syst.* **1992**, *12*, 209–224.
- Malinowski, E. R. Window Factor Analysis: Theoretical Derivation and Application to Flow Injection Analysis Data. *J. Chemom.* **1992**, *6*, 29–40.
- Manne, R.; Shen, H.; Liang, Y. Subwindow Factor Analysis. *Chemom. Intell. Lab. Syst.* **1999**, *45*, 171–176.
- Rubinstein, R. Y.; Kroese, D. P. *Simulation and the Monte Carlo Method*; John Wiley & Sons, 2016.
- Eberhart; Shi, Y. Particle Swarm Optimization: Developments, Applications and Resources. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*; IEEE: Seoul, 2001; Vol. 1, pp 81–86.
- Tadesse, M. G.; Sha, N.; Vannucci, M. Bayesian Variable Selection in Clustering High-Dimensional Data. *J. Am. Stat. Assoc.* **2005**, *100*, 602–617.
- Nia, V. P. Fast High-Dimensional Bayesian Classification and Clustering, EPFL, Lausanne, Switzerland, 2009.
- Zou, C.; Feng, J. Granger Causality vs. Dynamic Bayesian Network Inference: A Comparative Study. *BMC Bioinformatics* **2009**, *10*, 122.
- Selman, B.; Gomes, C. P. Hill-Climbing Search. In *Encyclopedia of Cognitive Science*; John Wiley & Sons, Ltd.: New York, 2006.
- Bai, X.; Padman, R. Tabu Search Enhanced Markov Blanket Classifier for High Dimensional Data Sets. In *The Next Wave in Computing, Optimization, and Decision Technologies*; Springer: Berlin, 2005; pp 337–354.
- Tsamardinos, I.; Brown, L. E.; Aliferis, C. F. The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm. *Mach. Learn.* **2006**, *65*, 31–78.
- Beer's Law https://teaching.shu.ac.uk/hwb/chemistry/tutorials/molspec/beers1.htm.

- Keun, H. C.; Ebbels, T. M. D.; Antti, H.; Bollard, M. E.; Beckonert, O.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Improved Analysis of Multivariate Data by Variable Stability Scaling: Application to NMR-Based Metabolic Profiling. *Anal. Chim. Acta* **2003**, *490*, 265–276.
- Duponchel, L.; Elmi-Rayaleh, W.; Ruckebusch, C.; Huvenne, J. P. Multivariate Curve Resolution Methods in Imaging Spectroscopy: Influence of Extraction Methods and Instrumental Perturbations. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2057–2067.
- Everitt, B. S.; Kroonenberg, P. M. *Three-Mode Principal Component Analysis: Theory and Applications*; DSWO Press: Leiden, The Netherlands, 1983.
- de Juan, A.; Jaumot, J.; Tauler, R. Multivariate Curve Resolution (MCR). Solving the Mixture Analysis Problem. *Anal. Methods* **2014**, *6*, 4964–4976.
- de Juan, A.; Rutan, S. C.; Tauler, R. Two-Way Data Analysis: Multivariate Curve Resolution-Iterative Resolution Methods. In *Comprehensive Chemometrics*; Elsevier: Oxford, 2009; pp 325–344.
- Sánchez, F. C.; Toft, J.; van den Bogaert, B.; Massart, D. L. Orthogonal Projection Approach Applied to Peak Purity Assessment. *Anal. Chem.* **1996**, *68*, 79–85.
- Sánchez, F. C.; van den Bogaert, B.; Rutan, S.; Massart, D. L. Multivariate Peak Purity Approaches. *Chemom. Intell. Lab. Syst.* **1996**, *34*, 139–171.
- Jiang, J. H.; Sasic, S.; Yu, R.Q.; Ozaki, Y. Resolution of Two-Way Data from Spectroscopic Monitoring of Reaction or Process Systems by Parallel Vector Analysis (PVA) and Window Factor Analysis (WFA): Inspection of the Effect of Mass Balance, Methods and Simulations. *J. Chemom.* **2003**, *17*, 186–197.
- Vandeginste, B. G. M.; Derks, W.; Kateman, G. Multicomponent Self-Modelling Curve Resolution in High-Performance Liquid Chromatography by Iterative Target Transformation Analysis. *Anal. Chim. Acta* **1985**, *173*, 253–264.
- Brown, S. D.; Tauler, R.; Walczak, B. *Comprehensive Chemometrics*; Elsevier: Oxford, 2009.
- De Leeuw, J. Block-Relaxation Algorithms in Statistics. In *Information Systems and Data Analysis*; Springer: Berlin, 1994; pp 308–324.
- Jiang, J. H.; Liang, Y.; Ozaki, Y. Principles and Methodologies in Self-Modeling Curve Resolution. *Chemom. Intell. Lab. Syst.* **2004**, *71*, 1–12.

- Shen, H.; Liang, Y.; Kvalheim, O. M.; Manne, R. Determination of Chemical Rank of Two-Way Data from Mixtures Using Subspace Comparisons. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 49–59.
- Hao, S. L.; Shao, L. M. Determining the Number of Principal Factors by Eigenvector Comparison of the Original Bi-Linear Data Matrix and the One Reconstructed from Key Variables. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 17–23.
- Malinowski, E. R. *Factor Analysis in Chemistry*; Wiley: Hoboken, NJ, 2002.
- Wasim, M.; Brereton, R. G. Determination of the Number of Significant Components in Liquid Chromatography Nuclear Magnetic Resonance Spectroscopy. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 133–151.
- Windig, W.; Guilment, J. Interactive Self-Modeling Mixture Analysis. *Anal. Chem.* **1991**, *63*, 1425–1432.
- Malinowski, E. R. Obtaining the Key Set of Typical Vectors by Factor Analysis and Subsequent Isolation of Component Spectra. *Anal. Chim. Acta* **1982**, *134*, 129–137.
- Workman, J. J.; Mobley, P. R.; Kowalski, B. R.; Bro, R. Review of Chemometrics Applied to Spectroscopy: 1985-95, Part I. *Appl. Spectrosc. Rev.* **1996**, *31*, 73–124.
- Tauler, R.; Maeder, M. Two-Way Data Analysis: Multivariate Curve Resolution – Error in Curve Resolution. In *Comprehensive Chemometrics*; Elsevier: Oxford, 2009; pp 345–363.
- Tauler, R.; Smilde, A.; Kowalski, B. Selectivity, Local Rank, Three-Way Data Analysis and Ambiguity in Multivariate Curve Resolution. *J. Chemom.* **1995**, *9*, 31–58.
- Sanchez, E.; Kowalski, B. R. Generalized Rank Annihilation Factor Analysis. *Anal. Chem.* **1986**, *58*, 496–499.
- Antunes, M. C.; J. Simão, J. E.; Duarte, A. C.; Tauler, R. Multivariate Curve Resolution of Overlapping Voltammetric Peaks: Quantitative Analysis of Binary and Quaternary Metal Mixtures. *Analyst* **2002**, *127*, 809–817.
- Byrd, R. H.; Gilbert, J. C.; Nocedal, J. A Trust Region Method Based on Interior Point Techniques for Nonlinear Programming. *Math. Program.* **2000**, *89*, 149–185.
- Yen, J.; Liao, J. C.; Bogju Lee; Randolph, D. A Hybrid Approach to Modeling Metabolic Systems Using a Genetic Algorithm and Simplex Method. *IEEE Trans. Syst. Man Cybern. Part B* **1998**, *28*, 173–191.

- Gen, M.; Cheng, R. *Genetic Algorithms and Engineering Design*; John Wiley & Sons: New York, 1997.
- Bansal, J. C.; Singh, P. K.; Saraswat, M.; Verma, A.; Jadon, S. S.; Abraham, A. Inertia Weight Strategies in Particle Swarm Optimization. In *2011 Third World Congress on Nature and Biologically Inspired Computing*; IEEE: Salamanca, Spain, 2011; pp 633–640.
- Kennedy, J.; Eberhart, R. Particle Swarm Optimization. In *IEEE International of first Conference on Neural Networks*; IEEE: Perth, 1995; pp 1942–1948.
- Shi, Y.; Eberhart, R. A Modified Particle Swarm Optimizer. In *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*; IEEE: Anchorage, AK, 2002; pp 69–73.
- Kang, E. Y.; Ye, C.; Shpitser, I.; Eskin, E. Detecting the Presence and Absence of Causal Relationships between Expression of Yeast Genes with Very Few Samples. *J. Comput. Biol.* **2010**, *17*, 533–546.
- Triantafillou, S.; Lagani, V.; Heinze-Deml, C.; Schmidt, A.; Tegner, J.; Tsamardinos, I. Predicting Causal Relationships from Biological Data: Applying Automated Causal Discovery on Mass Cytometry Data of Human Immune Cells. *Sci. Rep.* **2017**, *7*, 1–11.
- Agrafiotis, D. K.; Bandyopadhyay, D.; Wegner, J. K.; Van Vlijmen, H. Recent Advances in Chemoinformatics. *J. Chem. Inf. Model.* **2007**, *47*, 1279–1293.
- Rangarajan, S.; Kaminski, T.; Van Wyk, E.; Bhan, A.; Daoutidis, P. Language-Oriented Rule-Based Reaction Network Generation and Analysis: Algorithms of RING. *Comput. Chem. Eng.* **2014**, *64*, 124–137.
- Uusitalo, L. Advantages and Challenges of Bayesian Networks in Environmental Modelling. *Ecol. Modell.* **2007**, *203*, 312–318.
- Friedman, N.; Linial, M.; Nachman, I.; Pe'er, D. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.* **2000**, *7*, 601–620.
- Freedman, D.; Humphreys, P. Are There Algorithms That Discover Causal Structure? *Synthese* **1999**, *121*, 29–54.
- Lian, H. Sparse Bayesian Hierarchical Modeling of High-Dimensional Clustering Problems. *J. Multivar. Anal.* **2010**, *101*, 1728–1737.

- Heller, K. A.; Ghahramani, Z. Bayesian Hierarchical Clustering. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*; ACM Press: New York, New York, USA, 2005; pp 297–304.
- Heard, N. A.; Holmes, C. C.; Stephens, D. A. A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes. *J. Am. Stat. Assoc.* **2006**, *101*, 18–29.
- Lauritzen, S. L. The EM Algorithm for Graphical Association Models with Missing Data. *Comput. Stat. Data Anal.* **1995**, *19*, 191–201. https://doi.org/10.1016/0167-9473(93)E0056-A.
- Berzan, C. An Exploration of Structure Learning in Bayesian Networks. PhD Thesis, Tufts University, 2012.
- Xu, Z.; Srihari, S. N. Bayesian Network Structure Learning Using Causality. In *2014 22nd International Conference on Pattern Recognition*; IEEE: Stockholm, Sweden, 2014; pp 3546–3551.
- Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques-Adaptive Computation and Machine Learning*; The MIT Press: Cambridge, MA, 2009.
- Leray, P.; Francois, O. *BNT Structure Learning Package: Documentation and Experiments*; 2006.
- Craddock, P. R.; Le Doan, T. Van; Bake, K.; Polyakov, M.; Charsky, A. M.; Pomerantz, A. E. Evolution of Kerogen and Bitumen during Thermal Maturation via Semi-Open Pyrolysis Investigated by Infrared Spectroscopy. *Energy Fuels* **2015**, *29*, 2197–2210.
- Potts, R. H. Carboxylic Acids (Manufacture). In *Kirk-Othmer Encyclopedia of Chemical Technology*; Wiley: New York, 1978; pp 835–845.
- Cronauer, D. C.; Jewell, D. M.; Shah, Y. T.; Modi, R. J. Mechanism and Kinetics of Selected Hydrogen Transfer Reactions Typical of Coal Liquefaction. *Ind. Eng. Chem. Fundam.* **1979**, *18*, 153–162.
- Billmers, R.; Griffith, L. L.; Stein, S. E. Hydrogen Transfer between Anthracene Structures. *J. Phys. Chem.* **1986**, *90*, 517–523.
- Arends, I. W. C. E.; Mulder, P. Study of Hydrogen Shuttling Reactions in Anthracene/9,10-Dihydroanthracene−Naphthyl-X Mixtures. *Energy Fuels* **1996**, *10*, 235–242.
- Blanksby, S. J.; Ellison, G. B. Bond Dissociation Energies of Organic Molecules. *Acc. Chem. Res.* **2003**, *36*, 255–263.

- Grzybowski, M.; Skonieczny, K.; Butenschön, H.; Gryko, D. T. Comparison of Oxidative Aromatic Coupling and the Scholl Reaction. *Angew. Chemie Int. Ed.* **2013**, *52*, 9900–9930.
- Khorasheh, F.; Gray, M. R. High-Pressure Thermal Cracking of n-Hexadecane in Aromatic Solvents. *Ind. Eng. Chem. Res.* **1993**, *32*, 1864–1876.
- Evdokimov, I. N. The Importance of Asphaltene Content in Petroleum II—Multi-Peak Viscosity Correlations. *Pet. Sci. Technol.* **2010**, *28*, 920–924.
- Bazyleva, A. B.; Hasan, M. A.; Fulem, M.; Becerra, M.; Shaw, J. M. Bitumen and Heavy Oil Rheological Properties: Reconciliation with Viscosity Measurements. *J. Chem. Eng. Data* **2010**, *55*, 1389–1397.
- Wiehe, I. A. A Phase-Separation Kinetic Model for Coke Formation. *Ind. Eng. Chem. Res.* **1993**, *32*, 2447–2454.
- Wiehe, I. A. *Process Chemistry of Petroleum Macromolecules*; CRC Press: Boca Raton, FL, 2008.
- Noda, I. Determination of Two-Dimensional Correlation Spectra Using the Hilbert Transform. *Appl. Spectrosc.* **2000**, *54*, 994–999.
- Stagni, A.; Cuoci, A.; Frassoldati, A.; Faravelli, T.; Ranzi, E. Lumping and Reduction of Detailed Kinetic Schemes: An Effective Coupling. *Ind. Eng. Chem. Res.* **2013**, *53*, 9004–9016.
- Lide, D. R. *CRC Handbook of Chemistry and Physics*; CRC Press, 2004.
- Kuhn, H. W.; Tucker, A. . Nonlinear Programming. In *Traces and Emergence of Nonlinear Programming*; Giorgi, G., Kjeldsen, T. H., Eds.; Springer: Basel, Switzerland, 2014; pp 247–258.
- Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.
- Doraiswamy, L. K. *Chemical Reaction Engineering : Beyond the Fundamentals*; Taylor & Francis: Boca Raton, 2014.
- Daoud, J. I. Multicollinearity and Regression Analysis. *J. Phys. Conf. Ser.* **2017**, *949*, 012009.
- O'Brien, R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual. Quant.* **2007**, *41*, 673–690.
- Freund, R. J.; Wilson, W. J. *Regression Analysis: Statistical Modeling of a Response Variable*; Academic Press: Amsterdam, 2006.
- Dietz, W. A. Response Factors for Gas Chromatographic Analyses. *J. Chromatogr. Sci.* **1967**, *5*, 68–71.

- Yieru, H.; Qingyu, O.; Weile, Y. Characteristics of Flame Ionization Detection for the Quantitative Analysis of Complex Organic Mixtures. *Anal. Chem.* **1990**, *62*, 2063–2064.
- Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Qualitative Structure-Property Relationships Treatment. *Anal. Chem.* **1994**, *66*, 1799–1807.
- Diaz, M.; Vega, A.; Coca, J. Correlation for the Estimation of Gas-Liquid Diffusivity. *Chem. Eng. Commun.* **1987**, *52*, 271–281.
- Eberhart, R. C.; Yuhui Shi. Tracking and Optimizing Dynamic Systems with Particle Swarms. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*; IEEE: Seoul, 2002; pp 94–100.
- Xin, J.; Chen, G.; Hai, Y. A Particle Swarm Optimizer with Multi-Stage Linearly-Decreasing Inertia Weight. In *2009 International Joint Conference on Computational Sciences and Optimization*; IEEE: Sanya, China, 2009; pp 505–508.
- Arumugam, M. S.; Rao, M. V. C. On the Performance of the Particle Swarm Optimization Algorithm with Various Inertia Weight Variants for Computing Optimal Control of a Class of Hybrid Systems. *Discret. Dyn. Nat. Soc.* **2006**, *2006*, 1–17.
- Moré, J. J.; Sorensen, D. C. Computing a Trust Region Step. *SIAM J. Sci. Stat. Comput.* **1983**, *4*, 553–572.
- Steihaug, T. The Conjugate Gradient Method and Trust Regions in Large Scale Optimization. *SIAM J. Numer. Anal.* **1983**, *20*, 626–637.
- Byrd, R. H.; Schnabel, R. B.; Shultz, G. A. Approximate Solution of the Trust Region Problem by Minimization over Two-Dimensional Subspaces. *Math. Program.* **1988**, *40*, 247–263.
- Coleman, T. F.; Verma, A. A Preconditioned Conjugate Gradient Approach to Linear Equality Constrained Minimization. *Comput. Optim. Appl.* **2001**, *20*, 61–72.
- Wu, H. C. The Karush–Kuhn–Tucker Optimality Conditions in Multiobjective Programming Problems with Interval-Valued Objective Functions. *Eur. J. Oper. Res.* **2008**, *196*, 49–60.
- Fletcher, R. *Practical Methods of Optimization*; John Wiley & Sons: New York, 1987.
- Schittkowski, K. NLPQL: A Fortran Subroutine Solving Constrained Nonlinear Programming Problems. *Ann. Oper. Res.* **1986**, *5*, 485–500.
- Han, S. P. A Globally Convergent Method for Nonlinear Programming. *J. Optim. Theory Appl.* **1977**, *22*, 297–309.

- Powell, M. J. A Fast Algorithm for Nonlinearly Constrained Optimization Calculations. In *Numerical Analysis*; Springer: Berlin, Heidelberg, 1978; pp 144–157.
- Waltz, R. A.; Morales, J. L.; Nocedal, J.; Orban, D. An Interior Algorithm for Nonlinear Optimization That Combines Line Search and Trust Region Steps. *Math. Program.* **2006**, *107*, 391–408.

APPENDIXES

APPENDIX A.

Supporting Information of Least Squares-Support Vector Regression for determining product concentrations in acid catalyzed propylene oligomerization

A.1 Experimental: Materials

The different materials utilized for this work along with their respective purities are described in the manuscript. Table A.1 provides specific properties of the catalyst as given by the supplier.

Table A.1. Properties of ZSM-5 catalyst used in the reaction as characterized by supplier.

Property	Value/Characteristic
SiO_2/Al_2O_3 molar ratio	50
Na_2O weight %	0.05
Surface area, m^2/g	425

A.2 Experimental setup

The details of the flow reactor and the procedure are given in the manuscript. A complete view of the experimental setup is shown in Figure A.1.

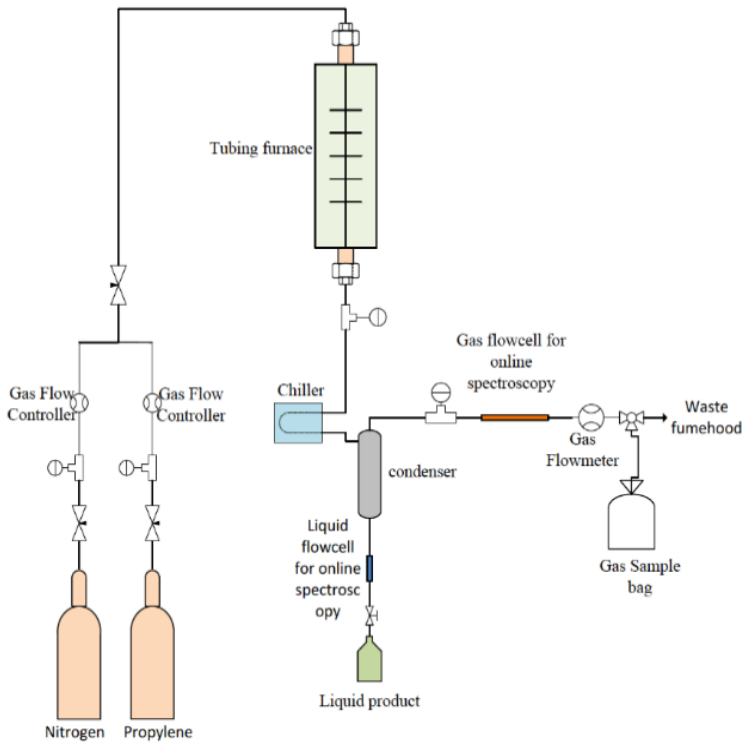


Figure A.1. Schematic of the complete experimental setup.

A.3 Experimental: Analyses

Information on the spectral parameters used for obtaining FTIR spectra are provided in the manuscript. Details of the gas chromatographic analyses are given in Table A.2.

Table A.2. Summary of analysis parameters for gas chromatography.

	Gas samples		Liquid samples
Injection volume	~1 mL		0.2 µL
Packed Purged GC Inlet	Heater Temperature	200 ^{0}C	250 ^{0}C
	Total gas flow	28 mL/min	25 mL/min
	Oven Temperature procedure	Step 1: Hold at 70 ^{0}C for 7 min Step 2: 70 – 250 ^{0}C at 10 ^{0}C/min Step 3: Hold at 250 ^{0}C for 2 min Step 4: 250 – 70 ^{0}C at -30 ^{0}C/min Step 5: Hold at 70 ^{0}C for 7 min	Step 1: 40 – 150 ^{0}C at 5 ^{0}C/min Step 2: Hold at 150 ^{0}C for 5 min
	Size	10 m × 200 µm × 0.5 µm	50 m × 200 µm × 0.5 µm
GC Column	Type	Agilent PLOT Alumina	Agilent 19091S-001
	Heater Temperature	250 ^{0}C	300 ^{0}C
Carrier gas	Helium		Helium

A.4 Experimental: Data available

The particulars of the different reaction conditions at which the experiments were conducted are given in Table A.3. Furthermore, the full set of hydrocarbon product concentrations ($C_1 – C_{10}$) for each reaction condition is given in Table A.4.

Table A.3. Reaction conditions with reactant conversion and residence times for the 25 sets of experiments.

Experiment No.	Temperature (T) (°C) [a, b]	Feed rate ($F_{m_{RT}}$) (mol/h) [b]	Residence times ($t_{R_{reaction}}$) (s)	Propylene Conversion
1	346	0.426	0.32	0.92
2	383	0.426	0.31	0.90
3	405	0.426	0.30	0.89
4	433	0.426	0.28	0.85
5	456	0.426	0.28	0.82
6	367	0.575	0.23	0.91
7	389	0.575	0.22	0.89
8	414	0.575	0.22	0.85
9	437	0.575	0.21	0.83
10	462	0.575	0.20	0.80
11	376	0.714	0.18	0.89
12	398	0.714	0.18	0.87
13	420	0.714	0.17	0.84
14	443	0.714	0.17	0.81
15	467	0.714	0.16	0.77
16	383	0.852	0.15	0.87
17	405	0.852	0.15	0.85
18	425	0.852	0.14	0.82
19	450	0.852	0.14	0.79
20	471	0.852	0.13	0.76
21	398	1.090	0.12	0.84
22	414	1.090	0.11	0.81
23	433	1.090	0.11	0.81
24	456	1.090	0.11	0.76
25	477	1.090	0.10	0.72

[a] Temperature at the catalyst zone/bed of the reactor measured by a thermocouple; [b] Parameters set/controlled by the user.

Table A.4. Hydrocarbon product concentrations (in % mol) obtained from GC analyses after propylene oligomerization at different reaction conditions.

Exp. No.	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
1	0.002	0.32	8.26	19.70	26.23	21.47	10.18	9.86	3.47	0.51
2	0.005	0.60	10.25	21.74	26.73	21.01	9.35	8.22	1.97	0.12
3	0.008	0.77	11.75	21.03	26.06	21.66	8.90	7.67	1.94	0.21
4	0.014	1.11	16.08	20.90	22.88	21.14	8.51	7.30	1.87	0.20
5	0.025	1.38	19.16	19.46	20.48	22.65	8.23	6.93	1.51	0.17
6	0.003	0.45	10.10	21.01	26.39	20.93	9.56	8.73	2.58	0.24
7	0.005	0.58	11.84	21.49	26.20	21.12	9.13	7.73	1.76	0.13
8	0.008	0.79	15.22	21.53	24.21	20.92	8.47	6.97	1.70	0.19
9	0.013	0.99	17.78	19.80	22.03	22.61	8.32	6.80	1.50	0.15
10	0.022	1.22	20.84	17.91	18.93	25.29	7.91	6.49	1.27	0.13
11	0.004	0.50	11.83	21.63	25.93	20.74	9.19	7.97	1.97	0.22
12	0.005	0.61	13.92	21.39	25.23	21.07	8.76	7.21	1.69	0.04
13	0.008	0.76	16.80	20.27	23.31	22.42	8.22	6.54	1.48	0.18
14	0.013	0.93	19.49	18.76	21.04	24.10	7.99	6.30	1.29	0.08
15	0.021	1.13	23.58	16.89	17.72	26.02	7.58	5.84	1.11	0.10
16	0.004	0.51	13.54	21.53	25.43	20.89	8.88	7.41	1.70	0.12
17	0.006	0.61	15.40	20.58	24.53	22.09	8.55	6.57	1.58	0.08
18	0.008	0.73	18.44	19.32	22.37	23.25	8.07	6.40	1.27	0.15
19	0.012	0.84	21.97	17.13	19.21	25.92	6.02	6.02	1.06	0.06
20	0.020	1.08	24.47	15.66	16.69	27.97	7.37	5.60	1.04	0.09
21	0.005	0.61	16.82	20.81	24.17	21.56	8.48	6.10	1.44	0.00
22	0.006	0.64	19.09	19.69	22.51	22.72	8.13	5.90	1.22	0.09
23	0.006	0.59	19.29	18.09	21.65	26.37	7.77	5.33	0.85	0.06
24	0.013	0.82	24.50	16.00	17.82	27.77	7.16	5.18	0.69	0.05
25	0.020	1.04	28.67	14.39	14.67	28.14	6.83	5.23	0.95	0.05

A.5 MDA theory: Principal component analysis (PCA)

Consider a m x n data matrix X, which represents the data set after mean-centering and standardization. Dimension m represents the number of samples and dimension n the number of variables. The eigenvectors of the covariance matrix $R = X^T X$ (since X is mean-centered and R is n x n) represent the principal component loadings, i.e. the new set of orthogonal directions, and the eigenvalues indicate the extent of variance captured. The total population variance is given by the sum of the eigenvalues of R, and the proportion of the total variance explained by each principal component is given by the ratio of each eigenvalue to the sum of all the eigenvalues. The principal component scores, which are the projections of the data matrix onto the loading vectors, represent the contribution of each variable on to the principal components, i.e. transformed data on to the new space which can be represented in equation A.1 as follows:

$$T = XP \qquad \text{Equation A.1}$$

where the columns of P (n x k) represent the loadings, and the columns of T (m x k) represent the scores. Since the number of variables is truncated to $k < n$, dimensional reduction is achieved.

A.6 MDA theory: Least squares – support vector machine (LS-SVM) regression

The transformation of the original data into a higher-dimensional space is enabled by means of a kernel function and is shown in Figure A.2.

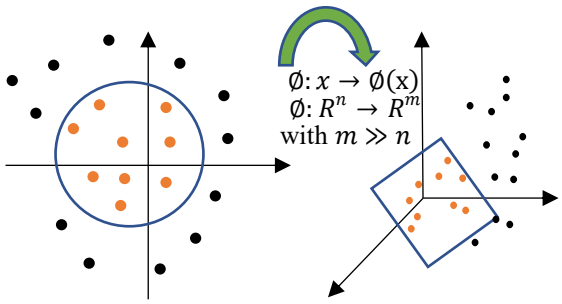


Figure A.2. Separation of training data set points in original feature space (circular curve) as compared to the transformed feature space (2-D plane).

The solution to equation 3.3 in Chapter 3 is obtained by solving the Karush-Kuhn-Tucker conditions arising from the first partial derivative condition of $L(w, b, e; \alpha)$ with respect to w, b, e and α. [1] Eliminating w and α results in the following matrix (1_n is the $n \times 1$ matrix of ones, I is the $n \times n$ identity matrix):

$$\begin{bmatrix} K + \frac{1}{\gamma} I & 1_n \\ 1_n^T & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ b \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}$$

Equation A.2

The calculation results in an inner product of the transformations which is expressed as the kernel function K in equation A.2, as long as Mercer's conditions are met. [2]

A.7 Modeling methods: Software tools and other operations for pre-processing

In the case of combination 1 and 2 (Table 3.2 in Chapter 3), temperature and molar flow rates were supplied as raw data to the LS-SVM based regression model. The lumped concentrations were calculated purely by addition of the respective concentrations in the 4 streams as classified by PCA for each experimental observation in the calibration dataset, before feeding to the regression models. For combinations 3 and 4 (Table 3.2 in Chapter 3), the smoothed, baseline corrected spectra were supplied as input to the LS-SVM tool which inherently performed standardization before constructing the calibration model. Furthermore, the clustered intensities for a particular observation in combination 5 and 6 were calculated by taking the square root of the sum squared intensities corresponding to the wavenumber indices belonging to the respective clusters formed by HCA (Table 3.4 and Table 3.5 in Chapter 3). The most significant PC loadings from PCA were given as input to HCA for cluster formation. These recalculated clustered intensities were then used as inputs to construct the LS-SVM and PLS based calibration models and during prediction for the validation dataset.

Lastly, while using interval-PLSR in combinations 3 and 4, the intensities were standardized before being input into the model in order to maintain consistency in preprocessing methods across the various models.

A.8 Modeling methods: Model tuning and optimization

The following is a brief description of the Venetian blinds method of cross validation used in *i*-PLS. The 15 calibration samples in the training set are split into 5 segments with the first segment comprising of sample numbers 1, 6, 11; the second segment comprising of samples 2, 7, 12 and so on. Each of these segments act as the test set with the remaining samples used for model building. An averaged RMSE value over the five segments is reported for each LV.

SIMPLEX is a multi-dimensional, unconstrained nonlinear optimization to find a function-minima and does not require any gradient information. It is based on the Nelder-Mead algorithm. [3] The optimized values and plausible interpretations of the hyper-parameters found for each output variable ($C_1 - C_{10}$ concentrations for combinations 1, 3, 5 and 6 (or) the four lumped concentrations for combinations 2 and 4) are reported and discussed in section 3.5 of Chapter 3.

A.9 Modeling methods: Model performance evaluation

The model performance measures utilized in this work for evaluation of the predicted results from the calibration models are given in Table A.5.

Table A.5. Statistical parameters used in this work to assess model predictive performance. y_i is the experimental output measure, y_i^* is the predicted output value, y_{mean} is the average of the measured output observations, n is the number of points in the corresponding validation set.

Evaluation parameter	Calculation formula
Root Mean Squared Error (RMSE)	$\left(\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - y_i^*\rvert^2\right)^{\frac{1}{2}}$
Relative Absolute Error (RAE)	$\frac{\sum_{i=1}^{n}\lvert(y_i - y_i^*)\rvert}{\sum_{i=1}^{n}\lvert(y_i - y_{mean})\rvert}$
Coefficient of determination (R^2)	$1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - y_{mean})^2}$

A.10 Results and Discussion: Principal component analysis (PCA)

For case (2) where PCA was applied to FTIR spectra with the spectral channels as the variables and 15 samples in the calibration set as observations, a plot of the loadings for the first two principal components is given in Figure A.3.

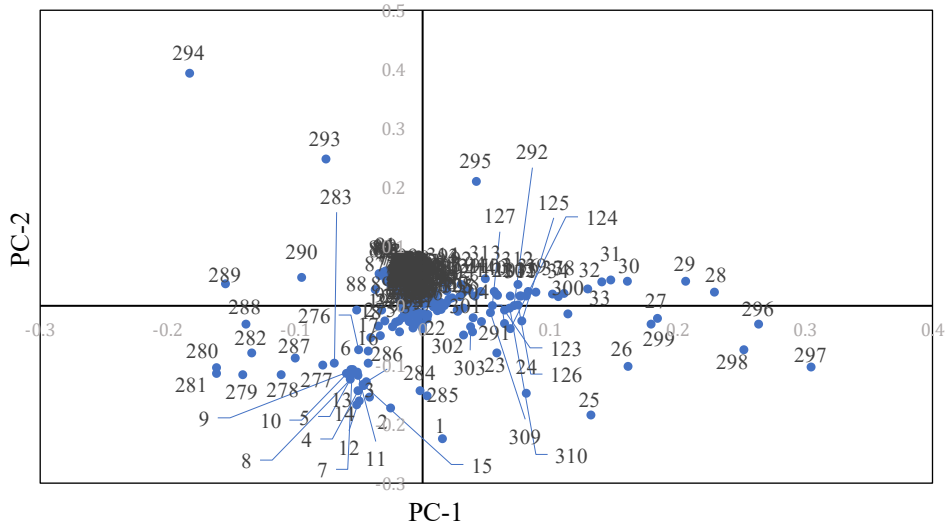


Figure A.3. Loadings plot of PC-1 vs. PC-2 for PCA applied on FTIR spectra with the 350 spectral channels as the variables and 15 samples as observations.

A.11 Combination 1: Inlet operating conditions as predictors and entire product composition as outcomes

The parameters for all regression models with the operating conditions (reaction temperature and propylene flowrate) as inputs and the entire product composition as outputs is given in Table A.6.

Table A.6. Tuned hyper parameters for the LS-SVM and the optimal LV for PLSR calibration models corresponding to combination 1 (Table 3.6 in Chapter 3) for outlet concentrations of each carbon number. LV is the optimum number of latent variables for the PLSR, t is the constant term in the polynomial kernel (equation 3.4 in Chapter 3), d is the degree of the polynomial, γ is the regularization parameter in the LS-SVM cost function (equation 3.2 in Chapter 3) and σ^2 is the square of the width of the Gaussian RBF kernel (equation 3.5 in Chapter 3).

C No	Model [a]	LV [b]	t	d	γ	σ^2
C_1	PLS	2 (92.5)	-	-	-	-
	Poly	-	3	5	19	-
	RBF	-	-	-	48	6
C_2	PLS	2 (97.5)	-	-	-	-
	Poly	-	5	6	2	-
	RBF	-	-	-	53	40
C_3	PLS	2 (99.1)	-	-	-	-
	Poly	-	306	3	.96	-
	RBF	-	-	-	35	41
C_4	PLS	2 (96)	-	-	-	-
	Poly	-	.71	4	3	-
	RBF	-	-	-	73	18
C_5	PLS	2 (95)	-	-	-	-
	Poly	-	1.4	6	1	-
	RBF	-	-	-	54	5
C_6	PLS	2 (87)	-	-	-	-

	Poly	-	1.3	5	2	-
	RBF	-	-	-	108	15
	PLS	1 (77)	-	-	-	-
C_7	Poly	-	15	6	31	-
	RBF	-	-	-	123	9
	PLS	1 (94)	-	-	-	-
C_8	Poly	-	830	5	6	-
	RBF	-	-	-	97	9
	PLS	2 (90)	-	-	-	-
C_9	Poly	-	1.2	6	10	-
	RBF	-	-	-	135	5
	PLS	1 (39)	-	-	-	-
C_{10}	Poly	-	2.5	6	22	-
	RBF	-	-	-	332	4

[a] Polynomial-based LS-SVM methods abbreviated as Poly in table and LS-SVM-Poly in text; RBF-based LS-SVM methods abbreviated as RBF in table and LS-SVM-RBF in text; [b] Numbers in brackets indicate the % variance explained by the optimal LV.

The results of the regression for combination 1 organized experiment-wise for the validation set are shown in Table A.7. The interpretation of the results in this table are provided in section 3.5.3.3 in Chapter 3.

Table A.7. Results of combination 1: Validation set samples for the entire product concentration distribution arranged experiment-wise.

Exp No.	PLSR			LS-SVR - Polynomial kernel			LS-SVR - RBF kernel		
	R_P^2	RMSEP	RAE_p	R_P^2	RMSEP	RAE_p	R_P^2	RMSEP	RAE_p
1	.9237	2.4983	.2398	.9740	1.4593	.1278	.9970	.4919	.0476
2	.9969	.5285	.0428	.9995	.2142	.0188	.9995	.2111	.0185
3	.9952	.6447	.0549	.9957	.6100	.0506	.9960	.5884	.0488
4	.9939	.6952	.0624	.9986	.3152	.0218	.9995	.1903	.0165
5	.9949	.6401	.0543	.9983	.3655	.0270	.9988	.3067	.0245
21	.9936	.7512	.0598	.9974	.4736	.0358	.9985	.3646	.0302
22	.9956	.6242	.0473	.9965	.5539	.0428	.9976	.4605	.0363
23	.9914	.9075	.0683	.9895	.9989	.0803	.9928	.8313	.0661
24	.9969	.5600	.0446	.9956	.6734	.0479	.9984	.4006	.0264
25	.9919	.9488	.0685	.9891	1.1005	.0733	.9945	.7822	.0541

A.12 Combination 2: Inlet reaction conditions as predictors and lumped product composition as outcomes

The results for the application of the all regression models with the operating conditions (reaction temperature and propylene flowrate) as inputs and the lumped product concentrations as the outputs is shown in Table A.8.

Table A.8. Tuned hyper parameters for the LS-SVM and the optimal LV for PLSR calibration models corresponding to combination 2 (Table 3.7 in Chapter 3) for the lumped product concentrations. LV is the optimum number of latent variables for the PLSR, t is the constant term in the polynomial kernel (equation 3.4 in Chapter 3), d is the degree of the polynomial, γ is the regularization parameter in the LS-SVM cost function (equation 3.2 in Chapter 3) and σ^2 is the square of the width of the Gaussian RBF kernel (equation 3.5 in Chapter 3).

S.No [a]	Model [b]	LV [c]	t	d	γ	σ^2
1	PLS	2 (97.5)	-	-	-	-
	Poly	-	20	4	9	-
	RBF	-	-	-	22	7
2	PLS	2 (97.2)	-	-	-	-
	Poly	-	12	4	13	-
	RBF	-	-	-	18	4
3	PLS	2 (92.5)	-	-	-	-
	Poly	-	2	6	7	-
	RBF	-	-	-	96	4
4	PLS	2 (93.5)	-	-	-	-
	Poly	-	1	7	.23	-
	RBF	-	-	-	102	1

[a] 1 – Light stream, 2 – Dimer stream, 3 – Mid stream, 4 – Heavy stream; [b] Polynomial-based LS-SVM methods abbreviated as Poly in table and LS-SVM-Poly in text; [c] Numbers in the brackets show % variance explained by the chosen number of LV.

A.13 Combination 3: FTIR full spectrum intensities as predictors and full product composition as outcomes

Results of the *i*-PLS regression applied on combination 3 is given in Table A.9.

Table A.9. Summary of the number of factors in PLS and optimized intervals in *i*-PLS for combination 3.

Carbon No.	Global optimized LV [a]	Global minimum RMSECV	Best performing region(s) (cm^{-1})	Optimal LV of interval	Interval minimum RMSECV	Possible functional group
C_1	7 (98.3)	.0058	1535–1396	5	.0047	sp^3 C-H bend
C_2	7 (95.7)	.21	1535–1396	5	.17	sp^3 C-H bend
C_3	10 (99.9)	3.60	1535–1396	5	2.43	sp^3 C-H bend
C_4	9 (99.9)	2.30	1535–1396	5	1.34	sp^3 C-H bend
C_5	9 (99.5)	3.05	1535–1396	5	2.22	sp^3 C-H bend
C_6	10 (99.9)	2.75	1535–1396 & 3139–3008	5 for both	1.42	sp^3 C-H bend, Ar C=C stretch & Ar sp^2 C-H stretch
C_7	1 (40.6)	.92	1396 – 1535 & 840 - 979	4 for both	.72	sp^3 C-H bend, Ar C=C- stretch, Ar sp^2 C-H bend & alkene sp^2 C-H bend
C_8	10 (99.8)	.71	840 – 979 & 3008 - 3139	4 & 3	.60	Ar sp^2 C-H bend, alkene sp^2 C-H bend & Ar sp^2 C-H stretch
C_9	11 (99.8)	.29	-	-	-	-
C_{10}	10 (97.8)	.0620	-	-	-	-

[a] Numbers in brackets indicate the % variance explained by the LV chosen for the global PLS.

The tuned hyper parameters along with the RMSECV corresponding to the calibration models developed for combination 3 is given in Table A.10.

Table A.10. Tuned hyper parameters for the LS-SVM calibration models and the RMSECV for combination 3 corresponding to the entire outlet product concentrations (Table 3.8 in Chapter 3). t is the constant term in the polynomial kernel (equation 3.4 in Chapter 3), d is the degree of the polynomial, γ is the regularization parameter in the LS-SVM cost function (equation 3.2 in Chapter 3) and σ^2 is the square of the width of the Gaussian RBF kernel (equation 3.5 in Chapter 3).

C No	Model [a]	t	d	γ	σ^2	RMSECV
C_1	PLS	-	-	-	-	.0058
	i-PLS	-	-	-	-	.0047
	Poly	45	3	21	-	.0069
	RBF	-	-	312	5	.0026
C_2	PLS	-	-	-	-	.21
	i-PLS	-	-	-	-	.17
	Poly	32	5	17	-	.28
	RBF	-	-	723	32	.11
C_3	PLS	-	-	-	-	3.60
	i-PLS	-	-	-	-	2.43
	Poly	36	4	56	-	.46
	RBF	-	-	351	24	.33
C_4	PLS	-	-	-	-	2.30
	i-PLS	-	-	-	-	1.34
	Poly	42	3	33	-	1.11
	RBF	-	-	661	23	.95
C_5	PLS	-	-	-	-	3.05
	i-PLS	-	-	-	-	2.22
	Poly	4	6	31	-	1.46
	RBF	-	-	820	21	1.03
C_6	PLS	-	-	-	-	2.75
	i-PLS	-	-	-	-	1.42
	Poly	11	5	11	-	1.38
	RBF	-	-	334	5	1.04

C_7	PLS	-	-	-	-	.92
	i-PLS	-	-	-	-	.72
	Poly	21	6	6	-	.67
	RBF	-	-	1012	4	.54
C_8	PLS	-	-	-	-	.71
	i-PLS	-	-	-	-	.60
	Poly	33	6	17	-	.57
	RBF	-	-	1493	5	.44
C_9	PLS	-	-	-	-	.29
	i-PLS	-	-	-	-	-
	Poly	13	6	43	-	.27
	RBF	-	-	776	7	.18
C_{10}	PLS	-	-	-	-	.0620
	i-PLS	-	-	-	-	-
	Poly	27	6	196	-	.0648
	RBF	-	-	1932	5	.0511

[a] Polynomial-based LS-SVM abbreviated as Poly in the table and as LS-SVM-Poly in the text. RBF-based LS-SVM abbreviated as RBF in the table and as LS-SVM-RBF in the text.

The experiment-wise performances of the calibration models for this combination are reported in Table A.11. The interpretations are discussed in section 3.5.5.3 of Chapter 3.

Table A.11. Results of combination 3: For each experiment in the validation set.

Exp No.	PLSR			LS-SVR Polynomial kernel			LS-SVR RBF kernel		
	R_P^2	RMSEP	RAE_p	R_P^2	RMSEP	RAE_p	R_P^2	RMSEP	RAE_p
1	.9545	1.9283	.1803	.9960	.5727	.0493	.9990	.2877	.0287
2	.9834	1.2170	.1122	.9967	.5381	.0461	.9978	.4387	.0298
3	.9926	.8007	.0651	.9973	.4826	.0342	.9980	.4378	.0323
4	.9874	1.0019	.0804	.9814	1.2174	.0953	.9960	.5644	.0374
5	.9722	1.4907	.0973	.9934	.7212	.0573	.9975	.4441	.0332
21	.9842	1.1781	.0949	.9942	.7155	.0551	.9971	.5069	.0433
22	.9854	1.1305	.0861	.9951	.6579	.0531	.9986	.3174	.0264
23	.9868	1.1229	.0868	.9903	.9611	.0520	.9962	.5998	.0408
24	.9950	.7134	.0536	.9945	.7479	.0541	.9972	.5307	.0401
25	.9897	1.0702	.0841	.9871	1.1972	.0854	.9968	.5945	.0515

A.14 References

(1) Kuhn, H. W.; Tucker, A. . Nonlinear Programming. In *Traces and Emergence of Nonlinear Programming*; Giorgi, G., Kjeldsen, T. H., Eds.; Springer: Basel, Switzerland, 2014; pp 247–258.

(2) Vapnik, V. N. *Statistical Learning Theory*; Wiley: New York, 1998.

(3) Olsson, D. M.; Nelson, L. S. The Nelder-Mead Simplex Procedure for Function Minimization. *Technometrics* **1975**, *17*, 45–51.

APPENDIX B.

Supporting Information of A statistical approach dealing with multicollinearity among predictors in microfluidic reactor operation to control liquid-phase oxidation selectivity

B.1 Experimental Microfluidic reactor

The microfluidic experimental setup consisted of a microfluidic reactor (Dolomite Microfluidics, Charlestown, MA, USA), syringe pump (KDS-210, KD Scientific, USA), oxygen and nitrogen gas cylinders (Praxair Inc., Edmonton, Canada), pressure transducer (Swagelok, Canada), gas flow meter (Swagelok, Canada), pressure bomb (Swagelok, Canada) and backpressure regulator (Swagelok, Canada), Heidolph MR Hei-Standard hot plate (Model: 505-20000-01-2, Heldolph Instruments, Germany), a surface mounted thermocouple (Model: CO 1, Cement-on Thermocouple, Omega Engineering, Inc., USA), a Flea3FL3-U3-13E4M camera (Point Grey Research Inc., Canada), a Fiber–Lite lamp (Model: 3100, Dolan-Jenner Industries, Inc., USA) and PTFE tubing, 1/16″ OD x 0.8 mm ID (Dolomite Microfluidics, Charlestown, MA, USA).

B.2 Experimental: Gas Chromatograph

An Agilent 7890A GC-FID equipped with DB-5 MS column 30 m × 0.25 mm × 0.25 µm column was used for quantitative analysis. The injector temperature of the GC was 250 °C and the split ratio was 10:1. Helium was used as a carrier gas, which was flowed through the column at a constant flowrate of 2 mL/min during the experiments. Oven temperature was varied throughout the experiments. Initially, the oven temperature was 75 ºC, which was kept constant for 0.5 minutes and then temperature was raised from 75 ºC to 325 ºC at a rate of 20 ºC/min, and finally, the temperature was kept constant at 325 ºC for 5 minutes. HPLC grade chloroform was used for sample preparation and hexachlorobenzene was used as an internal standard.

The oxidation products of tetralin were identified by using an Agilent 7820A GC coupled with an Agilent 5977E mass spectrometer. The products were separated on an HP-5 30 m × 0.25 mm × 0.25 µm column which have identical separation characteristics of DB-5 MS column used in GC-

FID. The temperature programs of both the GC-FID analyses and GC-MS analyses were the same. Oxidation products were classified as primary (alcohol and ketones of tetralin), secondary (products containing more than one ketone and/or alcohol functional groups) and addition products (products containing at least a dimer having different functional groups). GC-MS spectrum of commercially available alcohol and ketone of tetralin were used to identify the primary oxidation products whereas secondary and addition products were identified using GC-MS spectrum and NIST library. Details of product identification are available from a previous study. [1]

B.3 Experimental: Calculations

Different hydrodynamic parameters and mass transfer coefficients were calculated from the images captured during experiments in microfluidic reactor as elaborated in our previous paper. [2]

a) a (gas liquid interfacial area per unit liquid slug volume) was calculated from the dimension of the rectangular channel reactor ($h \times w$) and image analysis of gas bubbles and liquid slugs.

Surface area of gas bubble:

$$S_G = 2\left(wL_{G,actual} + hL_{G,actual}\right) + 4\pi((w+h)/4)^2 \qquad \text{Equation B.1}$$

$$L_{G,actual} = L_G - (w+h)/2 \qquad \text{Equation B.2}$$

Volume of liquid slug:

$$V_L = whL_S + wh[(w+h)/2] - (4/3)\pi[(w+h)/4]^3 \qquad \text{Equation B.3}$$

Gas liquid interfacial area per unit liquid slug volume:

$$a = S_G/V_L \qquad \text{Equation B.4}$$

Here, S_G is the surface of the gas bubble, L_G and L_S are the lengths of the gas bubble and liquid slug respectively, and w and h are the width and depth of the reactor channel, respectively.

Approximated radius of the cap of liquid slug:

$$r_{cap} = (w+h)/4 \qquad \text{Equation B.5}$$

Since geometry formed by the two liquid caps is not a complete sphere, the approximation was made.

b) U_L (superficial liquid slug velocity) and U_G (gas bubble velocity) were calculated from the distance travelled by the slug and bubble in a particular time. Two phase superficial velocity (U_{TP}) was calculated as follows:

$$U_{TP} = \varepsilon_G U_G + (1 - \varepsilon_G) U_L \quad \text{Equation B.6}$$

Here, the volume fraction of gas bubble (ε_G) was calculated as:

$$\varepsilon_G = \frac{V_G}{V_G + V_L} \quad \text{Equation B.7}$$

V_L is the volume of liquid slug was calculated according to equation B.3.

V_G is the volume of gas bubble, given as:

$$V_G = whL_{G,actual} + (4/3)\pi((w+h)/4)^2 \quad \text{Equation B.8}$$

c) Average residence time:

The two-phase superficial velocity (U_{TP}) was divided by the reactor length to calculate the average residence time.

d) δ (liquid film thickness surrounding a gas bubble) by using the correlations provided by Yun et al. [3] for a rectangular microchannel reactor as follows:

$$\frac{\delta_{\max}}{D_h} = 0.39\, We^{0.09} \quad \text{Equation B.9}$$

$$\frac{\delta_{\min}}{D_h} = 0.02\, We^{0.62} \quad \text{Equation B.10}$$

Here, Weber number,

$$We = \frac{D_h U_T^2 \rho_l}{\sigma_l} \quad \text{Equation B.11}$$

Hydraulic diameter of the channel (m),

$$D_h = 2[wh/(w+h)] \quad \text{Equation B.12}$$

$\delta_{\max}$ and $\delta_{\min}$ are the maximum and minimum thicknesses of the liquid film (m), respectively.

U_{TP} (m/s) is the two-phase superficial gas velocity, ρ_l is the density of liquid and σ_l is the surface tension of liquid (N/m).

w and h are the width and depth of the reactor channel, respectively.

e) $k_L a$ (volumetric mass transfer coefficient, s^{-1}) was calculated from k_L and a. Film theory was applied to calculate k_L [4] as follows:

$$k_L = \frac{D_A}{\delta} \qquad \text{Equation B.13}$$

Here, D_A is the diffusivity of oxygen in tetralin, δ is the thickness of liquid film surrounding the oxygen bubble.

$k_{L(max)}a$ and $k_{L(min)}a$ were based on the equation B.9 and equation B.10, respectively.

Calculation of conversion and product selectivity from GC analysis:

f) Product selectivity was obtained from the relative peak area of the products as follows:

$$\text{Product selectivity (\%)} = \frac{\text{relative peak area of specific product}}{\text{sum of relative peak area of all the products}} \text{ x } 100 \qquad \text{Equation B.14}$$

Ketone-to-alcohol selectivity in primary oxidation products was calculated by dividing ketone selectivity and alcohol selectivity.

g) Tetralin conversion was calculated by using GC-FID response factor (detailed calculation of conversion and response factor is given in section B.13 and section B.12, respectively). In the selectivity calculations, response factors for products were not used due to the diversity of oxidation products.

B.4 Assumptions in regression analysis: Linearity plots

Figure B.1 shows the scatter plots for both the outputs varying with the explanatory variables. By looking at this data that was obtained from our previous work, [2] it was realized that not all variations were entirely linear in nature, which was expected. This did not hinder the development of linear regression models as different powers of the variables were also checked for comparing the fits in both the calibration (section 4.3.5 in Chapter 4) and validation data sets (section 4.4.3 in Chapter 4) before arriving at the best fits for further analyses and confirmation.

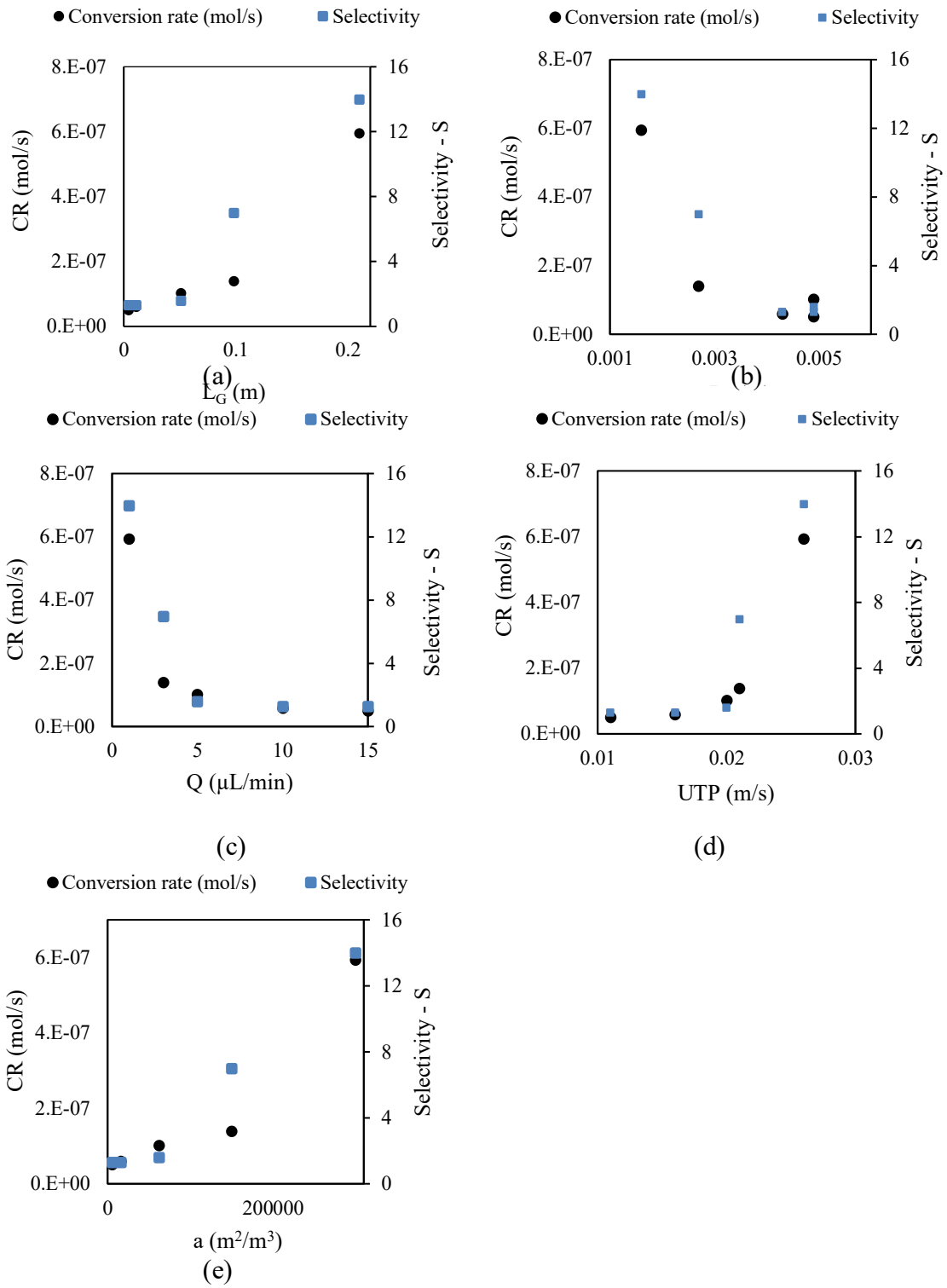


Figure B.1. Scatter plots for tetralin conversion rate and oxidation product selectivity versus: (a) length of gas bubble (L_G); (b) length of liquid slug (L_S), (c) tetralin flow rate (Q), (d) two-phase superficial velocity (U_{TP}), (e) gas-liquid interfacial area (a).

B.5 Relevant formulae and statistical calculations

(i) Pearson's correlation coefficient (r)

For a dependent variable defined as a vector $\mathbf{y} = \{y_i \in E, i = 1, 2, .., n)\}$ consisting of n observations (E is a one-dimensional vector space), the correlation with an explanatory variable $\mathbf{X} = \{x_i \in E, i = 1, 2, .., n)\}$ also comprising n observations is given in equation B.15 as:

$$r_{\mathbf{X-y}} = \frac{\sum_{i=1}^{n}(x_i - x_{mean})(y_i - y_{mean})}{\sqrt{\sum_{i=1}^{n}(x_i - x_{mean})^2}\sqrt{\sum_{i=1}^{n}(y_i - y_{mean})^2}} \qquad \text{Equation B.15}$$

where x_{mean} and y_{mean} are the average of all **X** and **y** observations, respectively.

Since the definition describes the correlation as the average of the product of mean-subtracted random variables, 'product-moment' is appended to the name. The square of the bivariate correlation coefficient gives the proportion of shared variance between the two variables (also same as r^2 for simple linear regression).

(ii) Significance of a correlation

Determination of the significance of a correlation involves formulation of a null (H_0: $\rho = 0$) and an alternative hypothesis (H_A: $\rho \neq 0$), where ρ is the population correlation coefficient between the two variables. It is assumed that the sampling distribution of the test statistic (t-statistic is used here) of the Pearson's correlation coefficient follows a t-distribution with 3 degrees of freedom (n - 2) under the null hypothesis in our case. The t-statistic in this case takes the form:

$$t = \frac{r - \rho}{\sqrt{\left(\frac{1-r^2}{n-2}\right)}} \qquad \text{Equation B.16}$$

However, since we do not have the population data, we can only estimate the population correlation in terms of the sample coefficients. The null hypothesis states that the population correlation coefficient is near to 0 while the alternative hypothesis is that it is significantly different from 0, implying a linear relationship between the variables in question. Finally, the p-value is the probability (simply, area under the t-distribution curve) of obtaining future values of the standardized form (t-statistic) of r at least as extreme as that observed value for the existing correlation in the direction of the H_A, still assuming that the null hypothesis was true. In other words, if it is quite low, there is lesser possibility that the correlation coefficient is near to zero, i.e. null can be rejected.

(iii) SLR and MLR models

The simple linear regression (SLR) models in Table 4.7 (S1 – S5) and Table 4.9 (SS1 – SS5) of Chapter 4 can be viewed as given by equation B.17:

$$\mathbf{y} = b_0 + b_1\mathbf{X} + \varepsilon \qquad \text{Equation B.17}$$

where b_0 is the intercept (constant) term and b_1 is the regression coefficient estimate (unstandardized) corresponding to the single explanatory variable **X**; ε is the normally distributed error with mean as 0 and variance σ^2 ($\varepsilon \sim \mathrm{N}(0, \sigma^2)$). In case of a quadratic fit, b_1 would be the coefficient estimate of the quadratic term and the linear term is excluded. Again, this is because of the limited number of observations available in the calibration set. When **X** and **y** are standardized, the intercept term disappears and the standardized coefficient equals the simple correlation coefficient between **X** and **y**.

The expression for multiple linear regression (MLR) with two variables, such models M1 – M4 (Table 4.7 in Chapter 4) and SM1 – SM4 (Table 4.9 in Chapter 4) is given in equation B.18 as:

$$\mathbf{y} = b_0 + b_1\mathbf{X_1} + b_2\mathbf{X_2} + \varepsilon \qquad \text{Equation B.18}$$

where b_0 is the intercept, b_1 and b_2 are the unstandardized slope estimates for the two explanatory variables $\mathbf{X_1}$ and $\mathbf{X_2}$ each consisting of n observations, respectively. Here again, ε is normally distributed as given by $\mathrm{N}(0, \sigma^2)$. The standardized coefficients in MLR do not equal the

respective simple correlation coefficients of the predictors with $\mathbf{y}$ as in SLR, but depend on both the simple correlations between $\mathbf{y}$ and $\mathbf{X}$ as well as on the correlation between $\mathbf{X_1}$ and $\mathbf{X_2}$.

The standard errors for b_1 and b_2 are given as:

$$SE_{b_1} = \sqrt{\frac{SSE}{(n-k-1)\left(\sum_{i=1}^{n}(X_{1i}-X_{1mean})^2\right)\left(1-r_{X_1-X_2}^2\right)}} \quad \text{Equation B.19}$$

$$SE_{b_2} = \sqrt{\frac{SSE}{(n-k-1)\left(\sum_{i=1}^{n}(X_{2i}-X_{2mean})^2\right)\left(1-r_{X_1-X_2}^2\right)}} \quad \text{Equation B.20}$$

where SSE is the sum of squared errors of the regression model and is directly proportional to the standard deviation of the residuals; k is the number of explanatory variables. SSE is defined as:

$$SSE = \sum_{i=1}^{n}(y_i - y_i^*)^2 \quad \text{Equation B.21}$$

and $r_{X_1-X_2}$ is the Pearson correlation coefficient between $\mathbf{X_1}$ and $\mathbf{X_2}$.

Like with the Pearson's correlation, the regression coefficient estimates for the MLR and SLR are tested for significance with the help of a t-statistic defined as:

$$t_{b_i} = \frac{b_i}{SE_{b_i}} \quad \text{Equation B.22}$$

where t_{b_i} is the t-statistic corresponding to the i^{th} coefficient estimate (b_i).

(iv) F-Statistic

This statistic is given by equation B.23 as:

$$F = \frac{MSM}{MSE} \quad \text{Equation B.23}$$

where MSM and MSE are the mean squared regression model and mean squared error defined as:

$$MSM = \frac{\sum_{i=1}^{n}\left(y_i^* - y_{mean}\right)^2}{k} \quad \text{Equation B.24}$$

$$MSE = \frac{\sum_{i=1}^{n}\left(y_i - y_i^*\right)^2}{n-k-1} \quad \text{Equation B.25}$$

where y_i^* is the predicted value of each observation in **y** through regression.

It follows an F-distribution with k and $n-k-1$ degrees of freedom (DF). The computed F-statistic for a regression model is compared to the critical value obtained from the F-table for 5% level of significance corresponding to (1, 3) DF for SLR and (2, 2) DF for MLR in our study.

The second type of F-statistic used for testing the significance of the addition of a second variable to an existing SLR model in improving the r^2 of the existing SLR model. Let R_{MLR}^2 be the coefficient of determination of the MLR model including the added second variable, r_{SLR}^2 be the coefficient of determination of the already existing SLR model, k_{MLR} and k_{SLR} be the number of predictors in the MLR and SLR respectively. Then, the F-statistic distributed with $(k_{MLR} - k_{SLR})$ and $(n - k_{MLR} - 1)$ DF is given by:

$$F_{inc} = \frac{(R_{MLR}^2 - R_{SLR}^2)/(k_{MLR} - k_{SLR})}{(1 - R_{MLR}^2)/(n - k_{MLR} - 1)} \qquad \text{Equation B.26}$$

B.6 RMSE and R^2

It is important to note that r^2 for SLR is also equal to the squared Pearson's correlation between input and output variable but R^2 in MLR is the proportion of variance explained in **y** due to both explanatory variables in our case. However, it can be calculated as the correlation between the predicted and experimental output values also. The RMSE and R^2 are given as:

$$RMSE = (MSE)^{\frac{1}{2}} = \left(\frac{1}{n-k-1}\sum_{i=1}^{n} |y_i - y_i^*|^2\right)^{\frac{1}{2}} \qquad \text{Equation B.27}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sum_{i=1}^{n}(y_i - y_{imean})^2} \qquad \text{Equation B.28}$$

B.7 Multicollinearity Diagnostics

Two types of diagnostics are used to confirm the presence of multicollinearity in the data and the degree of impact on regression. These are the following:

(i) Variance Inflation Factor (VIF) [5]

$$VIF = \frac{1}{1-R^2_{x_1 x_k}}$$ Equation B.29

where $R^2_{x_1 x_k}$ is the coefficient of determination when the variable x_1 is regressed on x_k, which represents the set of all other explanatory variables except x_1. A higher value of VIF indicates higher correlated variables. VIF would be 1 in a simple regression and higher in the case of multiple regression with collinear variables.

As with Pearson's correlation coefficient, there is a caution for false diagnosis of multicollinearity with VIF as well since there is no consensus on the threshold value. [6] Kutner et al. [7] suggest a minimum value of 10 while Vatcheva et al. [8] demonstrated that even a value of < 5 could be problematic. More than the absolute value, a change in VIF magnitude towards the higher side could provide evidence towards multicollinearity, which is what is pursued in this study by comparing multiple regression models with the simple regression counterparts as detailed in further sections in the manuscript. In addition, VIF can also be compared with $\frac{1}{1-R^2_{model}}$ to know whether the correlation between the regressors is stronger than the overall regression model. [9]

(ii) Eigen values (EV) and Condition Index (CI)

The sum of the eigenvalues of the correlation matrix (obtained through eigenvalue decomposition) will equal the number of explanatory variables in the system but the distribution of the eigenvalues across the dimensions of the matrix would point towards the presence or absence of linear dependencies. [10] If the variables are linearly independent, all eigenvalues will equal unity and in the case of correlated variables, certain dimensions would show eigenvalues close to 0. The latter situation indicates that the regression parameter estimates when regressed using these input variables would be very sensitive to changes in the data. Condition index (CI) helps in amplifying the unequal distribution of the eigenvalues and is given in equation B.30 as:

$$CI = \sqrt{\frac{\lambda_{max}}{\lambda_i}}$$ Equation B.30

where λ_{max} and λ_i are the maximum and the i^{th} eigen value respectively.

According to Midi et al., [11] if the CI falls below 15, then multicollinearity is not a serious concern. Johnston [12] proclaimed of inconsequential collinearity till $CI < 20$. Further, the detection process is also assisted by the variance decomposition proportions for each predictor, i.e. the proportion of variance for the regression coefficient estimates of each input variable that belongs to every dimension. Significantly correlated variables would have higher variance proportions concentrated on the same eigenvalue dimension. We have considered this aspect in our study also. Another diagnostic that has been reported in the literature but used on lesser number of occasions is the determinant of the correlation matrix, where a lower value indicates multicollinearity. It was not used in this study.

B.8 SLR models with S as outcome and residual plots

When the output of the SLR was selectivity, the prediction errors for the best fit regression models with L_G, U_{TP} and Q were larger than that for a by 107%, 229% and 243%, respectively. Plus, the averaged RMSEP for all the fits combined was the smallest for a (1.35) with the errors for U_{TP}, L_G, Q and L_S being 55%, 168%, 214% and 431% higher than a. Similar to the model with CR as the outcome, the significance of a and L_G terms were maximum with almost equivalent p-values of 0.002 and 0.003, respectively. In addition, the amount of variance in the outcome explained with respect to the calibration set by the best fits of the explanatory variables were in the order: a (0.98) > L_G (0.96) > U_{TP}^2 (0.79) > Q (0.58). Interestingly, r_C^2 for a linear fit with L_S as the explanatory variable was 0.93 but it was not considered for analysis because of the poor performance in predicting the validation set selectivity data. Lowest value of calibration error averaged for all the fits (including exponential) was obtained for a (RMSEC average = 1.63) while the corresponding values for L_G, L_S, U_{TP} and Q were higher than that for a by 12%, 15%, 83% and 188%, respectively.

Figure B.2 shows the residual vs. predicted value plots for the SLR models with selectivity and conversion rate as the output. It mainly helps in verification of the assumption of equal variance for validity of the regression model. Equal variance of the residuals essentially means that the observed experimental data points are spread out above and below the regression curve that represents the predicted values in a symmetric fashion and within the standard deviation of

the residuals (which is also used as the estimate of the standard error for the population). Further interpretation of the plots are summarized in section 4.4.4.4 of the manuscript.

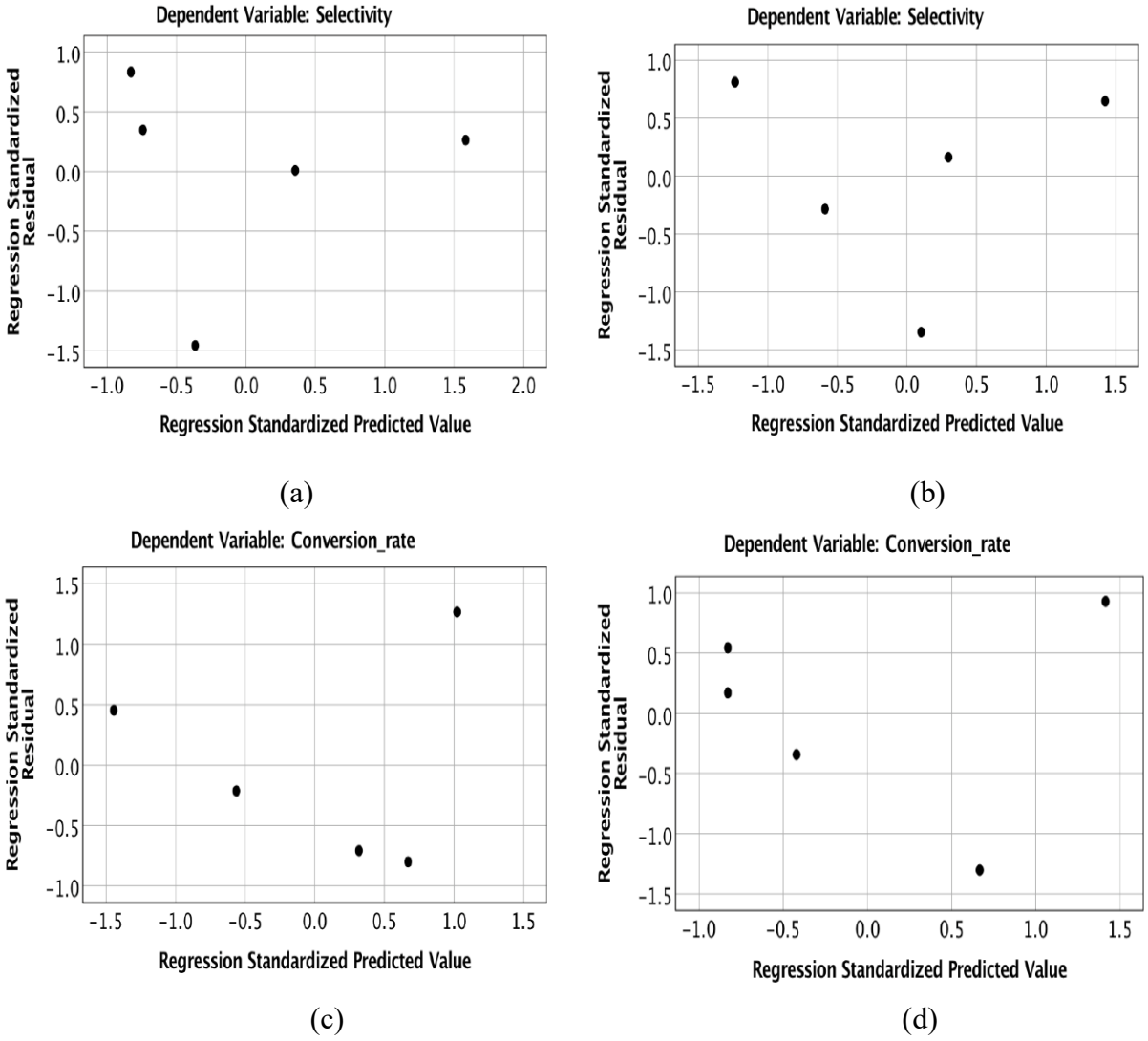


Figure B.2. Residual vs. predicted value plots in standardized forms for SLR of: (a) S on a; (b) S on U_{TP}^2; (c) CR on Q and (d) CR on L_S.

B.9 Comparison of MLR with SLR models

In this section, the regression coefficient estimates of the predictors, their standard errors and the significance of each term for the developed SLR models are compared with four MLR models involving gas-liquid interfacial area combined with the other 4 input variables one at a time, for both the outputs.

B.9.1 Effect of explanatory variables on tetralin conversion rate

Here, the analysis of the MLR models M3 and M4 (Table 4.7 in Chapter 4) involving a^2 & Q and a^2 & L_S in comparison with the respective SLR models are provided (Table B.1).

Table B.1. Parameter estimates, significances and output variances of SLR models with CR as the output and individual explanatory variables as the input. Compare the results with Table 4.7 in Chapter 4 that shows the parameter estimates for the MLR models.

Model [a]	Var. involved [b]	b_n [c]	SE [d]	Pr > $\lvert t_{crit} \rvert$ [e]	Std. coeff. *	R_C^2	F [f]	VIF	EV [g]	CI	Variance Proportions	
											b_0	T_1
S1	b_0 a^2 (T_1)	5.2E-8 5.9E-18	1.7E-8 4.1E-19	.056 .001	 .993	.986	210.7 (.001)	1	1.56 .438	1.0 1.9	.22 .78	.22 .78
S2	b_0 U_{TP}^3 (T_1)	-8.6E-8 .0343	7.9E-8 .0081	.356 .024	 .925	.856	17.87 (.024)	1	1.82 .176	1.0 3.2	.09 .91	.09 .91
S3	b_0 L_G (T_1)	-4.0E-9 2.6E-6	5.0E-8 4.7E-7	.941 .012	 .954	.910	30.25 (.012)	1	1.70 .296	1.0 2.4	.15 .85	.15 .85
S4	b_0 Q (T_1)	3.8E-7 -2.8E-8	1.4E-7 1.7E-8	.079 .203	 -.683	.467	2.63 (.203)	1	1.80 .199	1.0 3.0	.10 .90	.10 .90
S5	b_0 L_S (T_1)	6.8E-7 -1.3E-4	1.8E-7 4.7E-5	.034 .066	 -.853	.727	8.00 (.066)	1	1.94 .058	1.0 5.8	.03 .97	.03 .97

[a] Prefix 'S' corresponds to simple linear regression models; [b] b_0 represents the constant or intercept term in the regression equation. T_1 indicates the respective input variables used to specify the variance proportions associated with that variable; [c] Regression coefficient or slope estimates; [d] Standard errors

associated with the respective coefficient estimate; [e] p-value indicating the probability that the t-statistic for the term is greater than the critical value which depends on the degrees of freedom (DF) of the model. For SLR, DF = 3 & t_{crit} = 3.182; [f] F-statistic, distributed as F with 1 & 3 DF for SLR ($F_{critical}$ = 10.12). Statistic for the test of overall R^2 of the model. Value in bracket is the p-value for the significance of the overall model; [g] Eigenvalue for the respective dimension; * Standardized coefficient for SLR is the same as the zero-order Pearson's correlation coefficient between the variable and the output.

(i) Effect of gas-liquid interfacial area and injection flowrate of tetralin

All the tables referred to in this section and the next, belong to Chapter 4 so that it is not repeated at every point of referral. From Table 4.8, it can be recognized that a^2 and Q share the least correlation (-0.704) with each other compared to all other combinations of explanatory variables with a^2. This is supported by the VIF values from model M3, which was the lowest among the MLR models M1-M4 (Table 4.7). Though the VIF value of 2 was larger than that for the SLR models S1 and S4, the effect of collinearity on the MLR model was not as much a problem in this case as was with the previous 2 combinations. This was because model S4 itself could explain only 46.7% of the variance in CR since the Pearson's correlation between Q and CR was on the lower side (-0.683).

So, it was quite lucid that a^2 was more adept at influencing the output variation than Q and that the combined model involving both these predictors would not inflate the standard errors of the respective regression coefficients by a huge amount. This premise was evident from the results in Table 4.7 as the standard error for the regression coefficient estimate of a^2 increased only by 69 % (compare with 377 % and 231 % in the previous cases). The slope estimate was still significant in the MLR model with a p-value of 0.013. The presence of some extent of collinearity, between the predictors caused a sign reversal in the regression coefficient estimate of Q but surprisingly, the standard error of the coefficient estimate decreased by 73 %. The change in the sign of the slope estimate for Q rendered it insignificant (p-value = .818), though the overall model was still significant with F (2, 2) = 72.65 and p-value = 0.014.

There was a marginal increase of 0.011 and 0.203 units in the standardized coefficient of Q in M3 as compared to U_{TP}^3 in M2 and L_G in M1, respectively. It is more a consequence of a lesser correlation between a^2 and Q than the individual contribution of Q itself because the standardized

coefficient of a^2 was still much higher (1.014). A CI value of 5.49 indicated that collinearity was not a serious concern as the increase from the individual models (S1 and S3) was not as large as in the previous cases. Variance proportions of 71% and 90% provided further evidence of the diminished presence of collinearity in this case as compared to the previous combinations of predictors.

Examination of Table 4.8 for CR also reveals that the overall R^2 of model S1 is not significantly improved by adding Q (F (1, 2) = 0.00). On the other hand, the incremental effect of adding a^2 to model S4 (involving Q as the explanatory variable) was magnified with an F-statistic value of 74.14, which was much higher than $F_{critical}$ = 18.51. Moreover, the partial correlation between a^2 & CR while controlling for Q was 0.987, which implied that Q did not have a serious impact on the relationship between a^2 and CR, unlike the previous two situations. The correlation between Q and CR decreased by 73% (zero order correlation = 0.683; partial correlation = 0.182), while controlling for a^2. The effect of a^2 on the relationship between U_{TP}^3 & CR and L_G & CR was much greater as it lowered the correlations by 94% and 137%, respectively.

(ii) Effect of gas-liquid interfacial area and length of liquid slug

The first observation from Table 4.7 with a^2 and L_S as the predictors in model M4 was that the model produced the highest R_C^2 (0.997) with an overall significance of 0.003, which was the highest among other MLR models for CR explored in this study. It must also be noted that the pairwise correlation coefficient between a^2 and L_S (-0.905) was stronger than between a^2 & Q but weaker than the combinations of L_G and U_{TP}^3 with a^2. VIF showed a value of 6.00 for the variables involved and reflected this observation. Moreover, a major proportion of the variances in the regression coefficient estimates for a^2 (87%) and L_S (100%) was centered on the third eigenvalue dimension, thus indicating sufficient linear dependency between the regressors. The % variance proportion for the regression coefficient of a^2 in the third dimension was still lesser than in the case of M1 and M2.

Comparing M4 and S5, the sign of the slope estimate for L_S turned positive from being negative in S5, possibly due to the effect of collinearity. Interestingly, the standard error for this coefficient decreased by 72 % from S5, but still the lower bound value at the 95 % confidence interval was negative (-1.6E-5). Although the L_S term was insignificant, its p-value was lower than

that of L_G, U_{TP}^3 and Q in models M1, M2 and M3, respectively. On the other hand, the standard error for the coefficient estimate of a^2 hiked by a paltry 24 %, while the regression coefficient estimate increased by 23 %, thus rendering the term still highly significant (p-value = 0.005). The relative of importance of a^2 in influencing the outcome was further evident with the larger standardized coefficient (1.220) in comparison with that of L_S (0.251). It was intriguing to note that the standardized coefficient of L_S was greater than Q, U_{TP}^3 and L_G in M3, M2 and M1 models by 0.221, 0.232 and 0.424 units, respectively.

In addition, augmenting L_S to model S1 resulted in maximum value of F-statistic (7.33 –Table 4.8) compared to the other three MLR models, though it was still insignificant ($F_{critical}$ (1, 2) = 18.51). The addition of a^2 to L_S caused the F-statistic to swell up to 180.00, portending that the incremental effect of a^2 over L_S in improving the model R^2 was the maximum out of all MLR models. Notably, the partial correlation of L_S with CR, controlling for a^2 increased to 0.903 from the absolute zero order correlation of 0.853. Also, there was a petty increase of 0.002 units from the zero-order correlation between a^2 and CR to the partial correlation between them, while controlling for L_S.

B.9.2 Effect of input variables on oxidation product selectivity

Here, the analysis of the MLR models SM3 and SM4 (Table 4.9 in Chapter 4) involving a & Q and a & L_S in comparison with the respective SLR models are provided (Table B.2).

Table B.2. Parameter estimates, significances and output variances of SLR models with S as the output and individual explanatory variables as the input. Compare the results with Table 4.9 in Chapter 4 that shows the parameter estimates for the corresponding MLR models.

Model [a]	Var. involved [b]	b_n [c]	SE [d]	Pr > |t_{crit}| [e]	Std. coeff. *	R_C^2	F [f]	VIF	EV [g]	CI	Variance Proportions	
											b_0	T_1
SS1	b_0	.24	.61	.725					1.7	1.0	.15	.15
	a (T_1)	4E-5	4E-6	.002	.988	.977	125 (.002)	1	.30	2.4	.85	.85
SS2	b_0	.20	.79	.817					1.7	1.0	.15	.15
	L_G (T_1)	65	7	.003	.981	.856	71 (.003)	1	.30	2.4	.85	.85
SS3	b_0	-3.9	2.9	.275					1.9	1.0	.05	.05
	U_{TP}^2 (T_1)	2E4	7E3	.043	.890	.792	11 (.043)	1	.10	4.3	.95	.95
SS4	b_0	10	3.1	.048					1.8	1.0	.10	.10
	Q (T_1)	-.75	.37	.137	-.759	.576	4 (.137)	1	.20	3.0	.90	.90
SS5	b_0	19	2.2	.004					1.9	1.0	.03	.03
	L_S (T_1)	-3E3	562	.007	-.966	.934	42 (.007)	1	.06	5.8	.97	.97

[a] 'SS' corresponds to simple linear regression models with S as the output; [b] b_0 represents the constant or intercept term in the regression equation. T_1 indicates the respective input variable used to specify the variance proportions associated with that variable; [c] Regression coefficient or slope estimates; [d] Standard errors associated with the respective coefficient estimate; [e] p-value indicating the probability that the t-statistic for the term is greater than the critical value of t-stat, that depends on the degrees of freedom (DF) of the model. For SLR, DF = 3 & t_{crit} = 3.182; [f] F-statistic, distributed as F with 1 and 3 DF for SLR ($F_{critical}$ = 10.12). Statistic for the test of overall R^2 of the model. Value in bracket is the p-value for the significance of the overall model; [g] Eigenvalue for the respective dimension; * Standardized coefficient for SLR is the same as the zero-order Pearson's correlation coefficient between the variable and the output.

(i) Effect of gas-liquid interfacial area and injection flowrate of tetralin

Similar to the previous discussion, all references to tables in this section and the next belong to Chapter 4. Like in model M3 where the correlation between a^2 and Q was minimum among all other combinations, the simple correlation between a and Q was the least (-0.839) as compared to other correlations with a (Table 4.8). This was also reflected in the low value for VIF (3.37) and CI (8.32). Although the absolute value of CI was lower than the thumb rule cut off [11], it was ~3.5 times and ~2.5 times higher than in SS1 and SS4 models (Table 4.9 and Table B.2). The variance proportions for the regression coefficients of a (87 %) and Q (90 %) were more for the 3rd dimension of the eigenvalues but were lesser than that for SM1 and SM2, indicating moderate collinearity. The regression coefficient estimate of a did not experience any dramatic change in SM3 as compared to SS1, as a consequence of which the coefficient was significant in the MLR model (p-value = 0.008).

As with model M3, the slope estimate for Q became positive from being negative (in SS4) with a decrease in standard error by 72 %. The coefficient was still insignificant (compare with SS4), but the overall model SM3 was significant with a F (2, 2) = 145.9 and p-value of 0.007. Moreover, due to the low variance explained by Q in the output (r^2 = 0.576 for SS4), the addition of a to Q had a significant contribution in improving the r^2 (F (1, 2) = 119.14 – Table 4.8). On the other hand, Q did not produce a significant increase to the overall R^2 of SS1 when added to a as apparent from a much lower F-statistic value of 4.57 (F_{crit} = 18.51). Additionally, the partial correlation of a and S, while controlling for Q increased to 0.992 from the zero-order correlation of 0.988. Although the presence of a did not impact the relationship between Q and S in a negative way (value of 0.842), the partial correlation between Q and S was still lesser than that between a and S.

(ii) Effect of gas-liquid interfacial area and length of liquid slug

There was evidence for a strong linear dependency between a and L_S from the Pearson's simple correlation coefficient value of -0.945 (Table 4.8). This value was very similar to that between a and U_{TP}^2 but the direction of the correlation was opposite. It was interesting to note that

though the VIF (9.35 –Table 4.9) in model SM4 was almost equal to that of model SM2, the value of CI for SM4 was higher than that of SM2, which only indicated that collinearity could impact regression negatively. 93 % and 100 % of the variance proportions of a and L_S were focused on the third eigenvalue dimension, like in previous instances.

The presence of collinearity between the predictors did not affect the sign of regression coefficient estimate of L_S but increased its standard error by 67 % in SM4 as compared to SS5. It can be seen from Table 4.9 that both regression coefficients (for a and L_S) were insignificant in model SM4 but the overall model was still significant (p-value = 0.013, R^2 = 0.987). Thus, it made sense to interpret the slope estimates relative to each other because the overall regression was significant in all MLR models with S or CR as the output variable. It can be inferred that L_S affected the relationship between a and S more than other predictors because of the decreased standardized coefficient of a in SM4 (0.700) as compared with other MLR models. But this was much higher than that for L_S (-0.305), which signaled that S was more sensitive to changes in a than in L_S.

The partial correlations also tell a congruent story, giving a value of 0.894 for the relationship between a and S, while controlling for L_S and decreased from the zero-order value by 10%. On the contrary, a affected the link between L_S and S on a greater scale as the reduction from zero-order to partial correlation was 32% (-0.966 to -0.655). The improvement in r^2 of both models SS1 (0.977) and SS5 (0.934) to SM4 (0.987) caused by addition of L_S to a (F-stat = 1.54) and a to L_S (F-stat = 8.15), respectively were both insignificant (F_{crit} = 18.51). But on a relative measure, a contributed more significantly to improve the model r^2 due to the higher value for F-statistic (Table 4.8).

B.10 Physicochemical Properties of Tetralin and Oxygen

Physicochemical properties of tetralin and oxygen used for the calculation shown in the study are listed in Table B.3. Most of the properties are interpolated from the values obtained from the reported literatures.

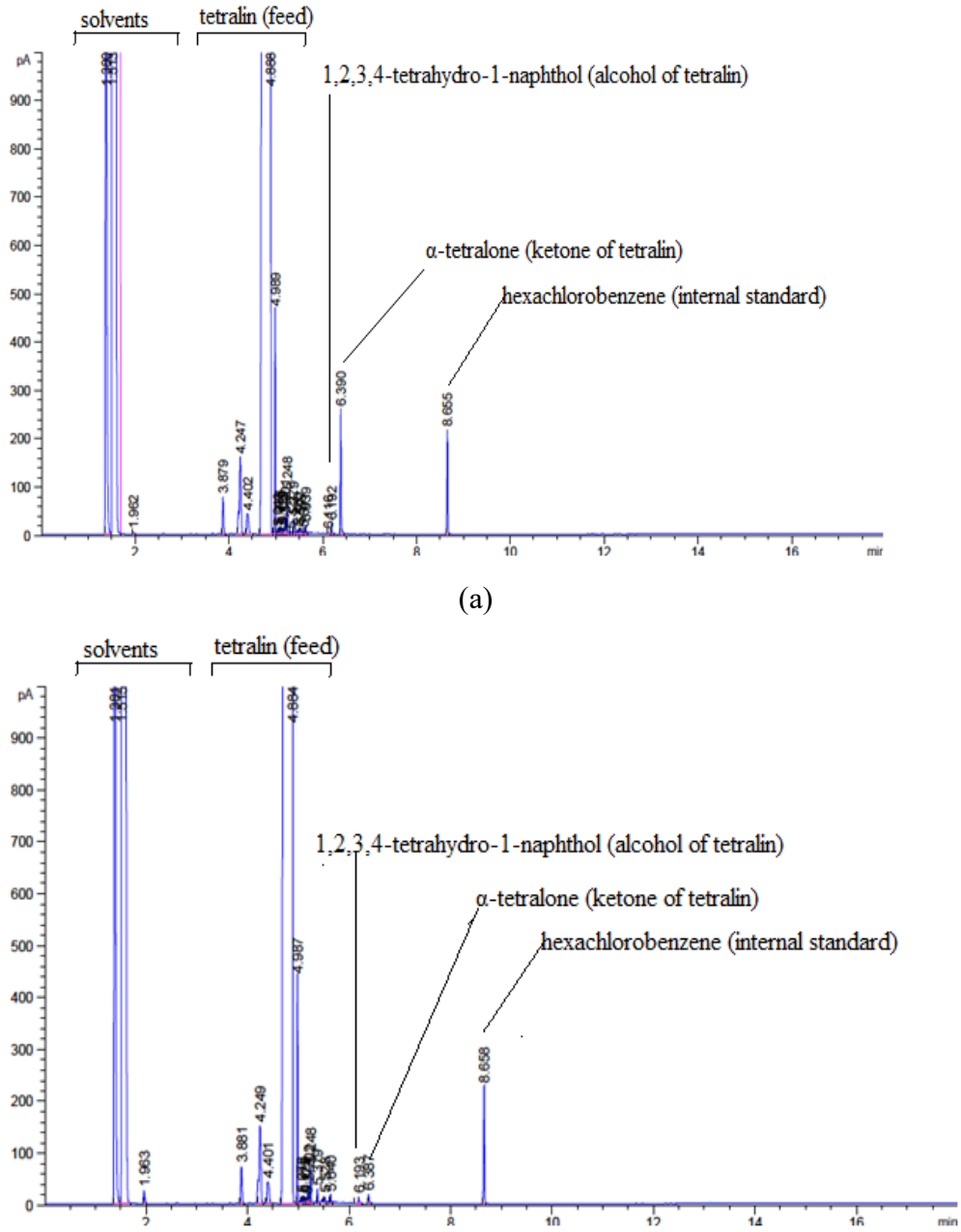
Table B.3. Physicochemical properties of tetralin and oxygen at different experimental conditions.

	Tetralin				Oxygen			
T (^{o}C)	**Density (kg/m^3)** [1]	**Surface tension** [a] **(N/m)** [4]	**Dynamic viscosity (Pa.s)** [1]	**Kinematic viscosity (m^2/s)** [1]	**Dynamic viscosity (Pa.s)** [3]	**Density** [b] **(kg/m^3)**	**Kinematic viscosity (m^2/s)** [3]	**D_A (m^2/s)** [2]
25	966	0.0351	1.17E-03	1.21E-06	2.15E-05	2.36E+00	9.08E-06	2.73E-09
120	887	0.0257	6.36E-04	7.17E-07	2.59E-05	1.79E+00	1.44E-05	2.14E-08
130	879	0.0248	5.84E-04	6.65E-07	2.64E-05	1.75E+00	1.51E-05	2.52E-08
140	871	0.0238	5.33E-04	6.13E-07	2.68E-05	1.71E+00	1.57E-05	2.93E-08
150	862	0.0228	4.83E-04	5.61E-07	2.73E-05	1.67E+00	1.64E-05	3.39E-08
160	854	0.0218	4.34E-04	5.08E-07	2.78E-05	1.63E+00	1.71E-05	3.90E-08

[a] with respect to air; [b] density of oxygen was calculated at experimental pressure using ideal gas law.

B.11 Product identification

GC-FID chromatograms of tetralin oxidized at 150 ºC in a microfluidic reactor are shown in Figure B.3 to illustrate the ketone-to-alcohol selectivity in primary oxidation product.



(a)

(b)

Figure B.3. GC-FID chromatogram of tetralin oxidized at 150 °C in a microfluidic reactor at gas-liquid interfacial area: (a) 3×10^5 m^2/m^3 (Series A: Table 4.2 in Chapter 4) and (b) 5×10^3 m^2/m^3 (Series E: Table 4.2 in Chapter 4).

B.12 Flame Ionization Detector (FID) Response Factors

The flame ionization detector (FID) has different responses to various organic compounds. So, it is required to calculate response factors for accurate quantification of oxidative conversion by GC-FID. The Dietz-method [13] was used to calculate the response factors:

$$Response\ factor\ (RF)_{Dietz} = \frac{(Area\ of\ compound)\times(Mass\ of\ standard)}{(Mass\ of\ compound)\times(Area\ of\ standard)} \qquad \text{Equation B.31}$$

Heptane was used as the standard and its response factor was 1.00. The calculated relative response factors are tabulated in Table B.4. The calculated relative response factors are very close the response factors reported in literature. [13–15] The FID response factors previously reported in literature are also listed in Table B.4 for comparison.

Table B.4. FID response factors of various compounds.

Compound Name	Retention Time (min)	Response Factor (RF)	Reported RF value
Heptane	1.72	1.00 ± 0.00	1.00 [13]
$CHCl_3$	1.52	0.09 ± 0.01	-
Hexachlorobenzene	8.67	0.32 ± 0.01	0.31 [14]
Tetralin	4.90	1.08 ± 0.01	1.02 [15]
1, 2, 3, 4 – tetrahydro-1-naphthol	6.35	0.82 ± 0.02	-
Alpha-tetralone	6.51	0.84 ± 0.01	0.80 [15]

B.13 Conversion Calculations

Conversion was calculated based on the tetralin disappearance and did not reflect the extent of oxidation. The percentage conversion was calculated as follows: [1]

$$W_i = \frac{A_i * W_{HCB}}{A_{HCB} * RRF_{i,HCB}} \qquad \text{Equation B.32}$$

where,

$$RRF_{i,HCB} = \frac{RF_i}{RF_{HCB}}$$ Equation B.33

This is the relative response factor of model compounds with respect to hexachlorobenzene (internal standard).

In equations B.32 and equation B.33,

RF_{HCB} = Response factor of hexachlorobenzene with respect to heptane

RF_i = Response factor of model compound with respect to heptane

W_i = Weight % of model compounds

W_{HCB} = Weight % of hexachlorobenzene

A_i = Peak area of model compounds

A_{HCB} = Peak area of hexachlorobenzene

For the conversion less than 1 % wt/wt, the tetralin conversion was calculated based on the formation of products. A conversion factor was calculated using the data obtained from oxidation of tetralin with air conducted in a semi-batch reactor (Table B.5). [15] Conversion factor was multiplied by sum of relative peak areas of products to get the conversion. Conversion factor was selected based on the sum of product area.

Table B.5. Conversion data for oxidation of tetralin with air at 150 ºC conducted in a semi-batch reactor. [15]

Time (h)	Conversion (% wt/wt)	Sum of oxidized products	Conversion factor
0.5	0.8	214.8	0.0035
1	1.1	643.3	0.0017
2	2.1	1128.1	0.0019
4	4.5	2922.5	0.0015
6	6.9	4628.7	0.0015

B.14 Diffusion Coefficient Calculation

Different correlations are available in literature to calculate the diffusivities in liquid. The correlation provided by Díaz et al. [16] can be used to calculate diffusivity of gases in liquid over wide temperature range. This correlation is used to calculate the diffusion coefficient of oxygen in tetralin (D_A) at 150 ºC. The correlation goes as follows:

$$(D_A)_T = 4.996 \times 10^3 (D_{AB})_{T=25^0C} e^{-\frac{2539}{T}} \qquad \text{Equation B.34}$$

where,

$$(D_A)_{T=25^0C} = 6.02 \times 10^{-5} \frac{V_B^{0.36}}{\mu_B^{0.61} V_A^{0.64}} \qquad \text{Equation B.35}$$

$(D_A)_T$ is the diffusion coefficient of oxygen in tetralin at given temperature in cm^2/s.

$(D_A)_{T=25\,°C}$ is the diffusion coefficient of oxygen in tetralin at 25 ºC in cm^2/s.

T is the absolute temperature (K) = 423 K.

μ_B is the viscosity of tetralin = 2 cp.

V_A is the molar volume of oxygen at the normal boiling point temperature ($cm^3/gmol$) = 27.9 $cm^3/gmol$.

V_B is the molar volume of tetralin at the normal boiling point temperature ($cm^3/gmol$) = 135.7 $cm^3/gmol$.

The values obtained at the different relevant temperatures were:

$(D_A)_{T=25\,°C}$ = 2.7 x 10^{-09} m^2/s

$(D_A)_{T=150\,°C}$ = 3.4 x 10^{-08} m^2/s

B.15 References

(1) Siddiquee, M. N.; De Klerk, A. Hydrocarbon Addition Reactions during Low-Temperature Autoxidation of Oilsands Bitumen. *Energy Fuels* **2014**, *28*, 6848–6859.

(2) Siddiquee, M. N.; de Klerk, A.; Nazemifard, N. Application of Microfluidics to Control Product Selectivity during Non-Catalytic Oxidation of Naphthenic-Aromatic

Hydrocarbons. *React. Chem. Eng.* **2016**, *1*, 418–435.

(3) Yun, J.; Lei, Q.; Zhang, S.; Shen, S.; Yao, K. Slug Flow Characteristics of Gas-Miscible Liquids in a Rectangular Microchannel with Cross and T-Shaped Junctions. *Chem. Eng. Sci.* **2010**, *65*, 5256–5263.

(4) Doraiswamy, L. K. *Chemical Reaction Engineering : Beyond the Fundamentals*; Taylor & Francis: Boca Raton, 2014.

(5) Daoud, J. I. Multicollinearity and Regression Analysis. *J. Phys. Conf. Ser.* **2017**, *949*, 012009.

(6) O'Brien, R. M. A Caution Regarding Rules of Thumb for Variance Inflation Factors. *Qual. Quant.* **2007**, *41*, 673–690.

(7) Kutner, M.; Nachtsheim, C.; Neter, J. *Applied Linear Statistical Models*; McGraw-Hill: Irwin, 2004.

(8) Vatcheva, K. P.; MinJae, L.; McCormick, J. B.; Rahbar, M. H. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiol. Open Access* **2016**, *6*, 227.

(9) Freund, R. J.; Wilson, W. J. *Regression Analysis: Statistical Modeling of a Response Variable*; Academic Press: Amsterdam, 2006.

(10) Belsley, D. A.; Kuh, E.; Welsch, R. E. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; John Wiley & Sons, Inc.: New York, 1980.

(11) Midi, H.; Sarkar, S. K.; Rana, S. Collinearity Diagnostics of Binary Logistic Regression Model. *J. Interdiscip. Math.* **2010**, *13*, 253–267.

(12) Johnston, J. *Econometric Methods*, 3rd ed.; McGraw-Hill: New York, 1984.

(13) Dietz, W. A. Response Factors for Gas Chromatographic Analyses. *J. Chromatogr. Sci.* **1967**, *5*, 68–71.

(14) Yieru, H.; Qingyu, O.; Weile, Y. Characteristics of Flame Ionization Detection for the Quantitative Analysis of Complex Organic Mixtures. *Anal. Chem.* **1990**, *62*, 2063–2064.

(15) Katritzky, A. R.; Ignatchenko, E. S.; Barcock, R. A.; Lobanov, V. S.; Karelson, M. Prediction of Gas Chromatographic Retention Times and Response Factors Using a General Qualitative Structure-Property Relationships Treatment. *Anal. Chem.* **1994**, *66*, 1799–1807.

(16) Diaz, M.; Vega, A.; Coca, J. Correlation for the Estimation of Gas-Liquid Diffusivity. *Chem. Eng. Commun.* **1987**, *52*, 271–281.

APPENDIX C.

Supporting Information of Viscosity of Canadian oilsands bitumen and its modification by thermal conversion

C.1 Calibration curve used for quantification of free radical spins in EPR

DPPH, containing a free radical on the nitrogen attached to the aromatic ring was used as a standard for calibration because its *g*-value (2.0037 $\pm$ 0.0002 – refer Ref. 57 in Chapter 5) is similar to that of an organic free radical. To build the calibration curve, DPPH solutions of concentrations ranging from 0.1 mM to 16 mM were prepared gravimetrically in toluene. The corresponding EPR spectra were obtained for each solution and the area under the absorption curves were calculated by double integration. The spin concentration of the unknown analyte is generally expressed in spins per gram of the solute. The molarity of DPPH solutions (M) and the corresponding number of spins of DPPH per mL of sample volume in the resonant cavity (N_{mL}) exposed to the magnetic field was calculated for each solution according to equation C.1 and equation C.2, respectively.

$$M\ (molarity\ in\ mM) = \frac{\left(\frac{Amount\ of\ DPPH\ taken\ in\ mg}{MW\ of\ DPPH\ in\ g}\right)}{\left(\frac{Volume\ of\ toluene\ in\ mL}{1000}\right)} * P \qquad \text{Equation C.1}$$

$$N_{mL} = M\ (in\ mM) * \frac{N_A}{10^6} \qquad \text{Equation C.2}$$

$$N_g = \frac{N_{mL}}{\rho\ in\ \frac{g}{mL}} \qquad \text{Equation C.3}$$

where P is the purity of DPPH (0.9), N_A is the avagadro number (6.023×10^{23} spins/mol of DPPH), ρ is the density of toluene (0.867 g/mL). The number of free radicals per molecule is one for DPPH.

Since the concentration of the solutions are low (in mM), the density of the sample inside the EPR tubes in the resonant cavity of the spectrometer can be approximated to be that of toluene's.

This was used so as to convert N_{mL} to the number of spins per gram of the solution (N_g – equation 3). N_g is plotted against the area under the EPR curves (double integrated intensities - DII) for the 13 samples to get the calibration curve as shown in Figure A1. It is to be noted that there are no genuine outliers in this plot and all points, from lower ($< 10^5$) to higher ($>10^6$) DI values lie within acceptable distance from the regressed straight line.

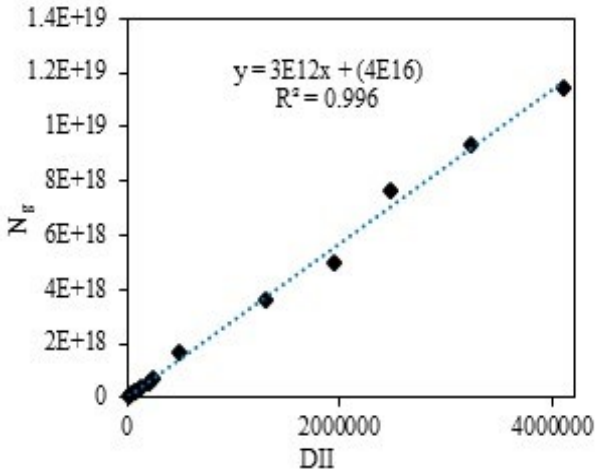


Figure C.1. Calibration curve for EPR spectra with DPPH in toluene as reference standard. The *g*-value for DPPH averaged across all data points is 2.0033 ± 0.0002.

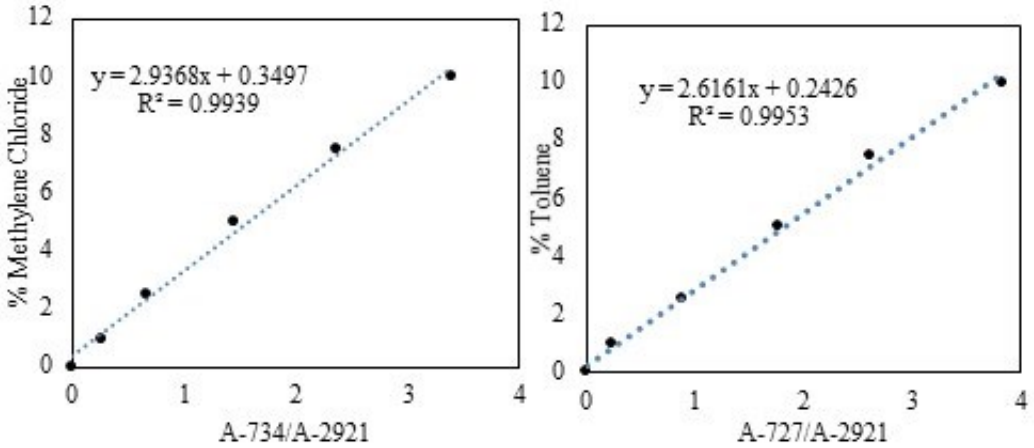
C.2 Estimation of remaining solvent in the thermally converted samples

Solutions of 1 %, 2.5 %, 5 %, 7.5 % and 10 %wt. concentrations of each thermally converted product were prepared in both methylene chloride and toluene separately. The FTIR spectra for each of the solutions were acquired. In the case of methylene chloride, a calibration curve for the known amount of solvent in the product against the ratio of the intensities of the peaks characteristic of C-Cl stretch and the dominant methylene group stretch was plotted for each thermally converted sample.

It is to be noted that in the spectrum of pure methylene chloride, the C-Cl stretch occurred as a doublet at 744 cm^{-1} (asymmetric stretch) and 717 cm^{-1} (symmetric) (Figure 5.12a) and the monoaromatic stretches of straight thermally converted products and bitumen itself occurred at 727 (strong) and 744 cm^{-1} (weak). Surprisingly, the methylene chloride-extracted product showed a broadened peak with a maximum at ~734 cm^{-1} (Figure 5.12c), which was taken as the hydrogen bonded C-Cl stretch, both for quantifying the amount of solvent present as well as for further

interpretation, as provided in the manuscript. The peak was found to broaden gradually with time as the spectra for the prepared samples for calibration (obtained immediately) showed a relatively sharp peak at 734 cm^{-1}. On the other hand, for toluene, the calibration curve was plotted with the known percent toluene in the sample against the ratio of the intensities for the aromatic C-H stretch at 727 cm^{-1} and that of the methylene group stretch at 2920 cm^{-1}. No such shift in frequency was observed in the case of toluene-extracted thermally converted products.

The ratios of the corresponding intensities (transmittance converted to absorption) in the unknown samples which were extracted with the respective solvents, were obtained and the corresponding amount of remaining solvent in them was calculated from the calibration equation. Example calibration curves for the thermally converted products at 15 min (for methylene chloride) and at 45 min (for toluene) are shown in Figure C.2. All curves for both the solvents were found to be linear.



(a) (b)

Figure C.2. Calibration curve for estimation of remaining (a) methylene chloride and (b) toluene in the thermally converted product obtained from thermal cracking at 400 ºC and 15 min and 45 min, respectively.

C.3 Torque generated in a Searle system for viscosity measurements

Since the torque was an important parameter in viscosity determination, it was worthwhile to look at the working equation for the measured torque in a viscometer setup. The fluid is located in the annular space in between the concentric cylinders representing the cup and the bob as shown

in Figure C.3.

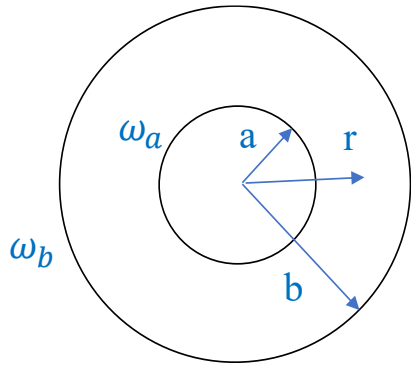


Figure C.3. Concentric cylinders representing the cup and the bob arrangement in a viscometer/rheometer. The bob has radius 'a' rotating with angular velocity ω_a and the cup has a radius 'b' rotating with angular velocity ω_b. In our case, $\omega_b = 0$.

In a Searle system, the bob is rotated by the motor, while the cup remains stationary. Conceptually, the torque required to turn the inner cylinder or the bob is a function of the viscosity of the annular fluid. The rheometer and the viscometer is operated in servo mode where the current generated in the motor is proportional to the torque required to rotate the inner cylinder. An important dimensionless number to be considered here is the Taylor number (Ta), which is a ratio of the centrifugal forces and the viscous forces. [152] The system is prone to vortex formation since the inner cylinder is rotated and care should be taken not to exceed the Taylor number. For a particular fluid element, the radial pressure gradient should balance the centripetal acceleration in order to guard against vortex formation, which can lead to inaccurate viscosity values.

Solving the Navier-Stokes equation, [153] the velocity profile of the annular fluid is obtained as inversely dependent on the radial distance in the case of inner cylinder rotation. The shear stress at the inner cylinder (τ_a) is calculated from the derived velocity profile and the torque is given as follows:

$$u = \frac{\omega_a a^2}{r} \qquad \text{Equation C.4}$$

$$\tau_a = -\frac{2\mu\omega_a b^2}{b^2 - a^2} \qquad \text{Equation C.5}$$

$$M_a = \tau_a A a = \frac{4\pi\mu a^2 b^2 \omega_a}{b^2 - a^2} \qquad \text{Equation C.6}$$

$$A = 2\pi aL$$ Equation C.7

A is the curved surface area of the inner cylinder of length L. It can be seen that the torque is a function of the viscosity of the annular fluid (μ) with other parameters being the dimensions of the cylinders and the angular velocity of rotation.

APPENDIX D.

Supporting Information of A data-driven approach to generate pseudo-reaction networks for thermal conversion of Athabasca bitumen

D.1 Pre-processed and residual data for temperatures of 420 ºC, 400 ºC, 380 ºC, 300 ºC

The FTIR spectra of liquid samples obtained after thermal conversion at 350 ºC after baseline correction and SG filtering are provided in Figure 6.2 in Chapter 6. The respective plots along with the residual obtained from smoothing and the raw data for the other 4 temperatures are given in Figure D.1, Figure D.2, Figure D.3 and Figure D.4.

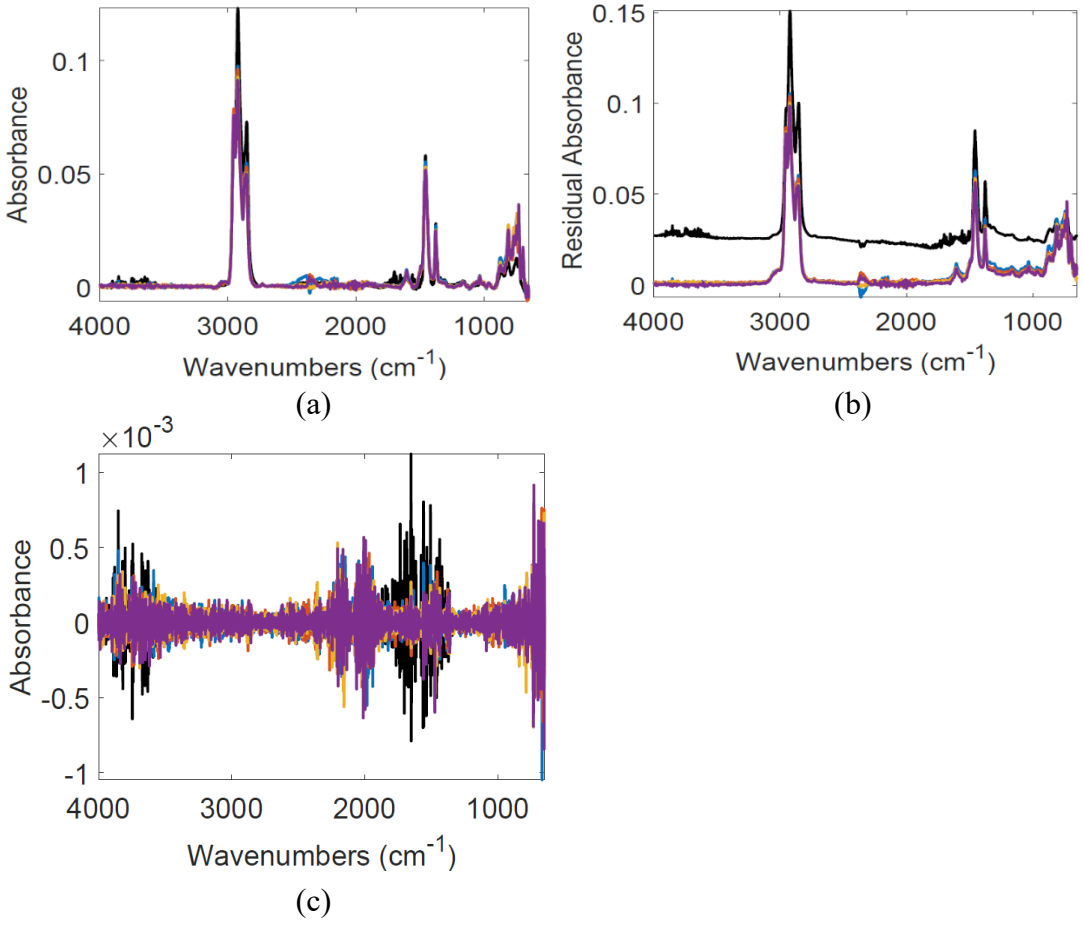
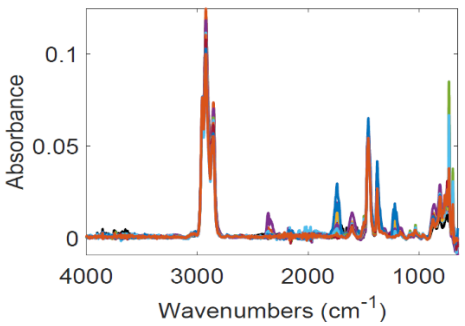
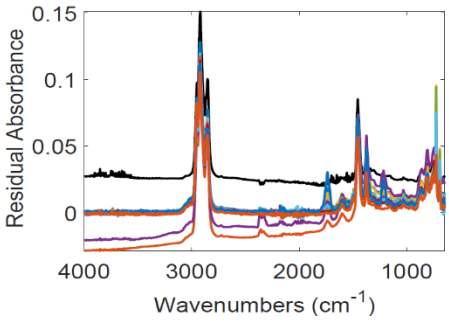


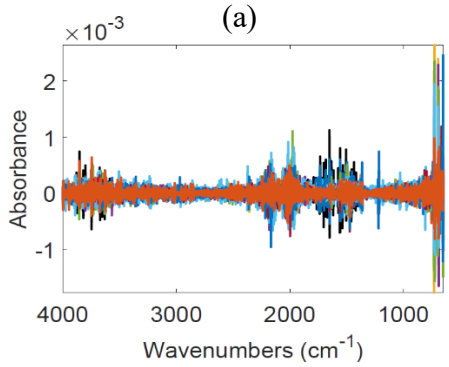
Figure D.1. Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 420 ºC; (c) residual after smoothing.



(a)

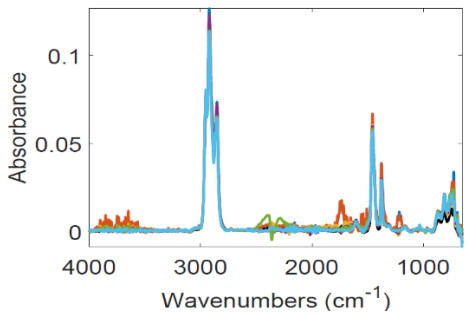


(b)

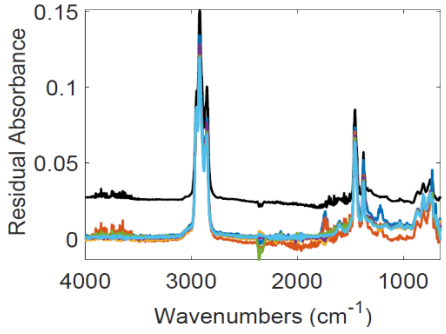


(c)

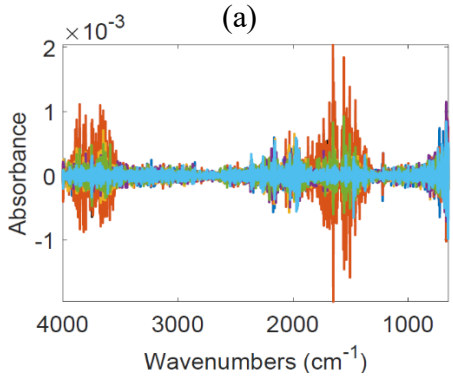
Figure D.2. Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 400 ºC; (c) residual after smoothing.



(a)

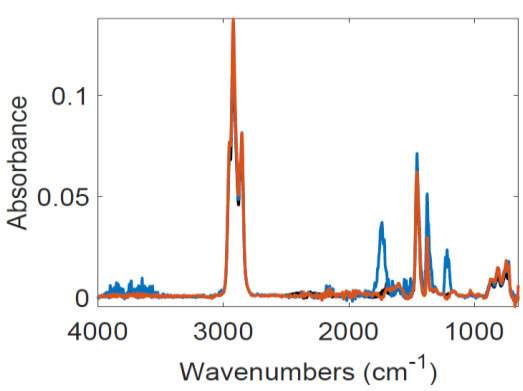


(b)

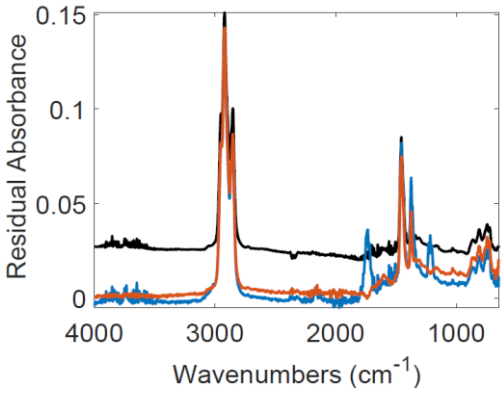


(c)

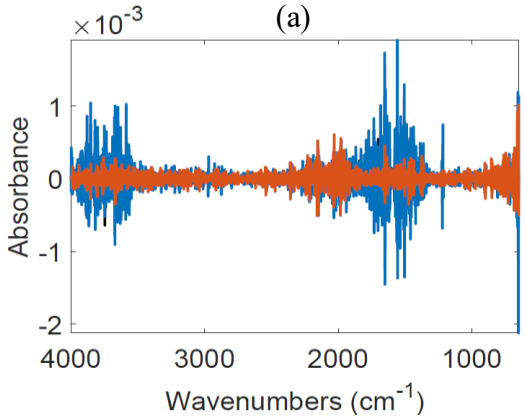
Figure D.3. Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 380 ºC; (c) residual after smoothing.



(a)



(b)



(c)

Figure D.4. Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 300 ºC; (c) residual after smoothing.

D.2 SMCR-ALS and SMCR-PSO methods and algorithms

To deal with some of the limitations of SMCR like rotational and intensity ambiguities, datasets from different runs and techniques are combined together into a single data matrix. A row-wise combination is performed when the same batch of experiments is monitored by different sets of techniques like FTIR, NMR, ESR, etc. The parent equation is illustrated in equation D.1. A column-wise matrix is obtained when multiple batches of experiments conducted at different experimental conditions are monitored by the same technique. This is given in equation D.2.

$$[D_1\ D_2\ D_3\ \dots D_n] = C[S_1\ S_2\ S_3\ \dots S_n]^T + [E_1\ E_2\ E_3\ \dots . E_n] \quad \text{Equation S.1}$$

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} S^T \quad \text{Equation S.2}$$

Intensity ambiguity is represented by:

$$D = (Ck)(S^T \tfrac{1}{k}) \quad \text{Equation S.3}$$

where k is a scalar.

Rotational ambiguity is given in equation S4 by:

$$D = (CT)(T^{-1}S^T) + E \quad \text{Equation S.4}$$

where T is a non-singular invertible matrix that multiplies with C and whose inverse multiplies with S. There are infinite possibilities for T in the absence of other constraints.

The ALS-optimization algorithm and the accompanying constraints is described in the manuscript. The respective equations of the alternative minimization of the Frobenius norm of the residual are given below:

$$\min_{S \geq 0}(\left|\left|D - CS^T\right|\right|^2) \quad \text{Equation S.5}$$

$$\min_{C \geq 0}(\left|\left|D - CS^T\right|\right|^2) \quad \text{Equation S.6}$$

Table D.1 gives some of the common strategies of choosing the inertia weight parameter for velocity updating in PSO.

Table D.1. Common strategies for inertia weight employed in the PSO literature.

Type of strategy	Remarks
Constant [1]	A value between 0.7 – 1 shows lower error but larger number of iterations for convergence
Random [2]	Increases convergence in early stages of PSO; Gives faster overall convergence
Linearly decreasing [3]	Decreasing values in the range 0.9 – 0.4 are employed but risk of local optimum exists; Gives low error
Global-local best inertia weight [4]	Falls in between constant and random inertia weight strategies; takes global and local best particle positions into consideration but gives large error

As mentioned in section 6.3.3.8 in Chapter 6, *'fmincon'* was used to further carry out a local search for the PSO-optimized concentration profiles inside the ALS loop. The next few paragraphs discuss two algorithms used by *'fmincon'* for the optimization process in further detail. These are the 'Sequential Quadratic Programming' algorithm and the 'Interior Point' algorithm. First, a nonlinear unconstrained minimization problem of a general nature is explained, followed by the algorithms for the constrained optimization.

(i) Unconstrained minimization:

Consider a scalar function $f(x)$ whose minimum point and the corresponding value needs to be found. Most algorithms are based on building trust regions around the neighborhood (N) for a simplified version q of f. [5] The trust region sub-problem is expressed in equation S7 as:

$$\min_s q(s), s \in N \qquad \text{Equation D.7}$$

where s is a sample step that assists in updating the present position if $f(x+s) < f(x)$.

The challenge is to define q and the trust region N. Expressing q in terms of the first two terms of the Taylor's expansion, the quadratic programming problem comes down to solving the equation:

$$\min \frac{1}{2} s^T H s + s^T g \text{ for } ||Ds|| \leq \Delta \qquad \text{Equation D.8}$$

In equation D.8, D is the diagonal scaling matrix, Δ is a positive tolerance level for the constraint and can be adjusted according to whether the updated value of f meets the inequality condition or not, H is the square matrix of second derivatives of f (Hessian) and g is the gradient of f. A number of approaches to solve this equation are given in the literature. [6,7] All these algorithms require rigorous calculations of eigenvalues, but it is easier to solve using the definition of a sub-space s that forms a boundary for the trust region. s is constructed in the 2-D space as a combination of the gradient direction (s_1) and the Newton direction (s_2), which is the solution to the following equation

$$H.s_2 = -g \qquad \text{Equation D.9}$$

The solution to equation D.9, which is a system of linear equations, is given by the preconditioned conjugate gradient (CG) method whose output direction, p is used to build the sub-space. The key step in solving unconstrained optimization problems is determining the 2-D sub-space. It is chosen such that global convergence is achieved through the steepest descent direction while local convergence is accomplished through the Newton step. Nonlinear least squares and linear least squares solutions also work on similar principles of trust regions and 2-D sub-space.

(ii) Constrained minimization:

Two common constraints for these kinds of problems are linear equality and box constraints. The linear equality constrained problems are solved considering an initial point that satisfies the equality $Ax_0 = b$, where A and b are known. A matrix system is created to calculate s and is elaborated by Coleman and Verma. [8] Box constraints consist of lower and upper bounds and a scaled Newton step evolving from the Karush-Kuhn-Tucker (KKT) conditions is considered to

find the sub-space for solving the problem. [9] The solution also comprises of a reflection step that delineates the step size.

(iii) Algorithms used by 'fmincon'

Active set algorithm:

This is a medium-scale algorithm where full matrices are generated and complex linear algebra is used to solve the constrained equations. They were based on the conversion of the constrained problem into an unconstrained one by the use of a penalty function. The KKT conditions are necessary and sufficient for optimality when both the objective function and the constraints are convex. The KKT conditions of the quadratic programming problem are given as:

$$\nabla f(x_s) + \sum_{i=1}^{m} \lambda_i . \nabla \mathrm{G_i}(x_s) = 0$$

$$\lambda_\mathrm{i} . \mathrm{G_i}(x_s) = 0 \text{ and } \lambda_i \geq 0 \qquad \text{Equation D.10}$$

where λ_i are the Lagrange multipliers that take positive values only and serve as a link between the objective and constraint functions. The solution revolves around finding the Lagrange multipliers for each data point.

Sequential Quadratic Programming (SQP) algorithm:

'*fmincon'* utilizes SQP methods frequently to solve the constrained optimization problems. The principle of SQP rests on creating quadratic programming sub-problems at each loop iteration. [10] It is analogous to the active-set algorithm explained in the previous section and instead of a Newton step used for the unconstrained optimization (equation D.9), a quasi-Newton updating procedure is used for dealing with the Hessian matrix (H). Detailed reviews of the method are available in various texts in the literature. [11,12]

The solution of the quadratic sub-problem is used to form a search direction for the variable x as:

$$x_{k+1} = x_k + \alpha_k d_k \qquad \text{Equation D.11}$$

Here, d_k is the search direction and α_k is the step length parameter obtained by line search. It helps the solution to progress toward the function minimum by decreasing the value of the objective function. Schittkowski [12] also opined that the advantage of utilizing the SQP method is that it makes the constrained optimization converge faster than an unconstrained problem due to a fixed search area and α_k. The SQP algorithm has 4 major steps:

a. Updating the Hessian (H_k) of the Lagrangian formulation

The Lagrangian formulation of the quadratic problem is given by the following equation:

$$L(\lambda, x) = f(x) + \sum \lambda_i . g_i(x) \qquad \text{Equation D.12}$$

A quasi-Newton approximation of $H(L(\lambda, x))$ is conducted at each iteration. In order to track the convergence path in MATLAB, the 'Display' option can be set to 'iter'. When this is done, messages such as 'Hessian modified', 'infeasible' are displayed that indicate that the extent of nonlinearity is high.

b. Solution of the QP sub-problem

The solution of this problem is executed by the active-set method described in the previous section. It is also called a projection method. This involves primarily two steps: estimating a feasible starting point and then generating a number of points that remain active throughout the iterations and subsequently converge to the final solution. The active points lead to the search direction (d_k in equation D.11) that is present on the boundaries of the given constraints. This search direction facilitates the calculation of the new point of x in the search space (equation D.11). d_k is usually obtained through a linear combination of a vector that is orthogonal to the active points.

Two directional choices are available for α_k during the line search procedure. One is the direct step along d_k that would lead to the optimum of $f(x)$ considering the active point set and thus, the solution of the QP sub-problem. If this does not occur, further iterations are required to reach the solution. The condition of positive Lagrange multipliers needs to be satisfied, otherwise the equality constraint is violated and the data point corresponding to this violation is removed from the algorithm.

c. Finding the starting point

This can be done by finding an x that satisfies the equality constraint in the QP sub-problem. A system of linear equations needs to be solved to obtain the initial point. The initial search direction can be obtained by substituting d_k for s in equation D.9.

d. Merit function and step length

A merit function proposed by Han [13] is used and a penalty parameter was introduced by Powell. [14] The merit function is similar to the Lagrangian function L but has more parameters. The penalty parameter distinguishes between constraints having smaller and larger gradients and penalizes the smaller gradients more. The step length parameter, as discussed before, reduces the merit function value.

From the implementation viewpoint, the algorithm in MATLAB allows for failed steps in the case of a bogus value for the objective function. During the running of the algorithm, lesser memory and time is consumed as compared to the active-set strategy though both are medium-scale algorithms. In addition, in the case of some nonlinear constraints being violated, SQP calculates a second order approximation for the constraints and proceeds with the iteration, though it sacrifices convergence speed.

Interior Point Algorithm:

This is the default algorithm adopted by MATLAB for the *'fmincon'* function. A detailed description of this method is given by Waltz et al. [15] and only the two important steps of the solution process are described in this section. The main objective function is split into constituent small-scale optimization problems given by equation D.13:

$$\min_{x,s} f(x,s) = \min f(x,s) - \mu \sum \ln (s_i) \qquad \text{Equation D.13}$$

where s_i are the slack variables and μ is a positive parameter that controls the barrier function $\sum \ln (s_i)$.

The purpose of the approximate problem is the conversion of inequality constraints to equality constraints to make it easier for problem solving. Equation D.13 can be solved by taking either of the following 2 steps: direct step or a CG step. The KKT conditions are applied to the QP and the obtained system of equations are tried to be solved by linear approximation. This is the first and

default step attempted by the algorithm. The CG step comes into play when the objective functions fails to remain convex at any iteration. In either case, a merit function that combines the objective function and the constraints is required to be decreased in value as much as possible. The algorithm can deal with constraint violations when a particular point x_j returns an unreal value for the constraint function. In this situation, the step length is modified to a shorter value and the iteration is continued.

In the direct step, matrix factorization gives information about the Hessian. If the Hessian is not positive definite, the algorithm attempts to solve the system of equations using the CG method. Similar to the unconstrained minimization, CG utilizes a trust region to create a sub-space for the solution to the QP problem. As with other cases, Lagrangian multipliers are obtained from solving KKT condition equations to obtain the solution for the interior point algorithm. Unlike SQP, interior point algorithm is a large-scale algorithm that does not store or generate full sized matrices and thus, lesser space is used and is the preferred approach for computer programming.

D.3 Modeling methods and parameters used: Bayesian networks

Figure D.5 shows the importance index for the first 1550 wavenumbers. The procedure behind the choice of these wavenumbers is given in section 6.3.4.1 in Chapter 6.

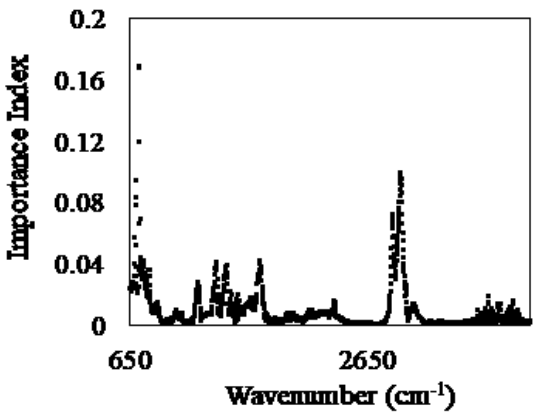


Figure D.5. Plot of importance index of the selected 1550 wavenumbers.

D.4 Results and discussion: Rank determination of each sub-matrix

Figure D.6 gives the plots of residuals obtained after performing SVD on the 400 ºC data set choosing 2 and 4 components while the manuscript gives the residual plot for SVD performed with optimal 3 components. The ROD, SD, residual after performing SVD with 3 components and the scree plots for data sets at the other 4 temperatures (300 ºC, 350 ºC, 380 ºC, 420 ºC) are given in Figure D.7, Figure D.8, Figure D.9 and Figure D.10 respectively.

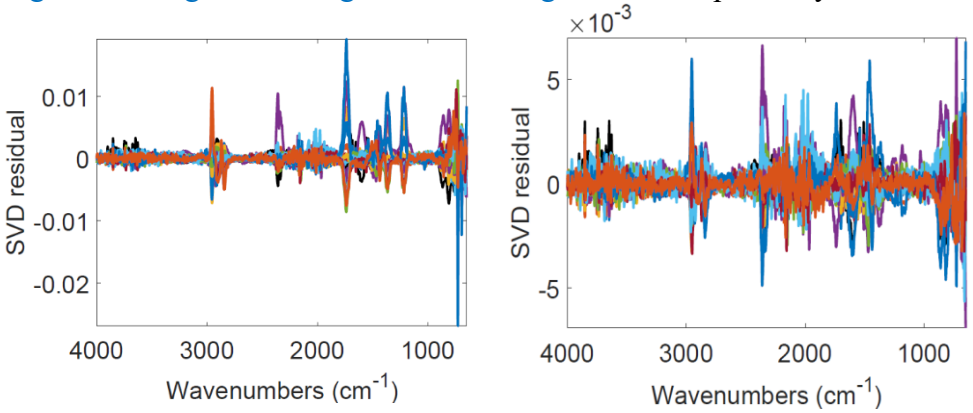


Figure D.6. Residuals obtained after performing SVD on the 400 ºC data set considering: (a) 2 components and (b) 4 components.

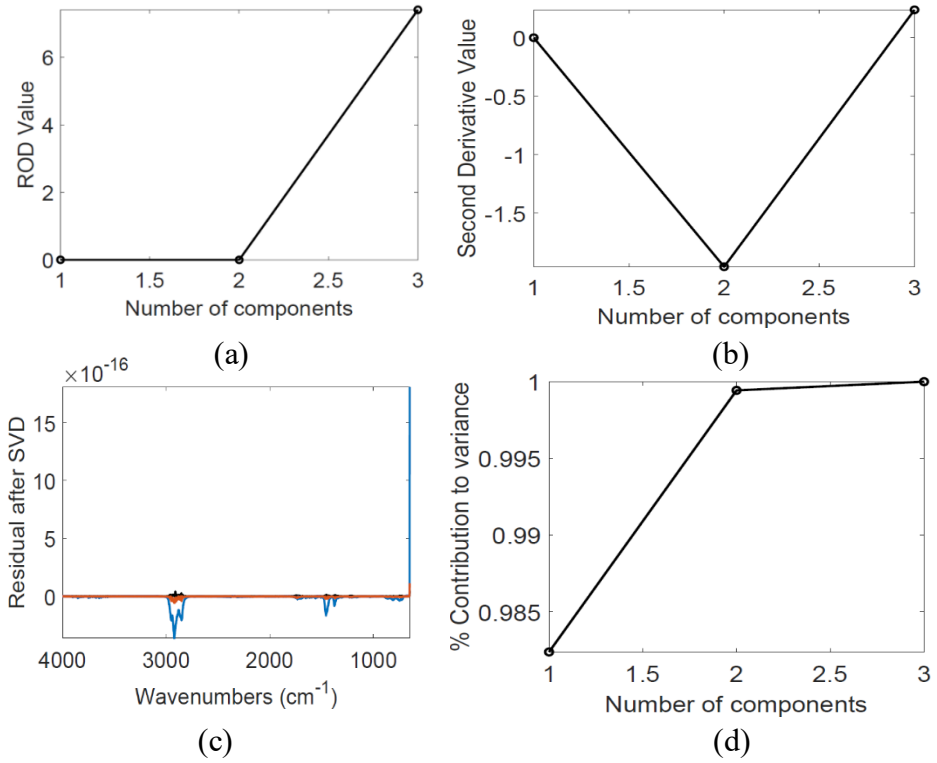


Figure D.7. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 300 ºC.

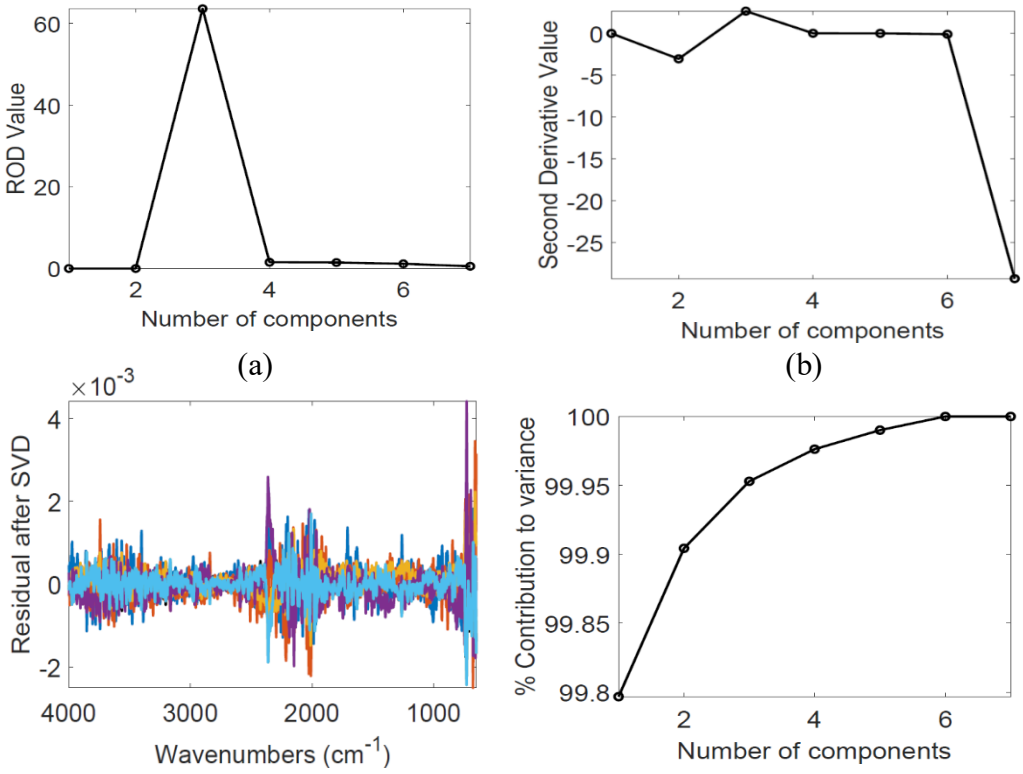


Figure D.8. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 350 ºC.

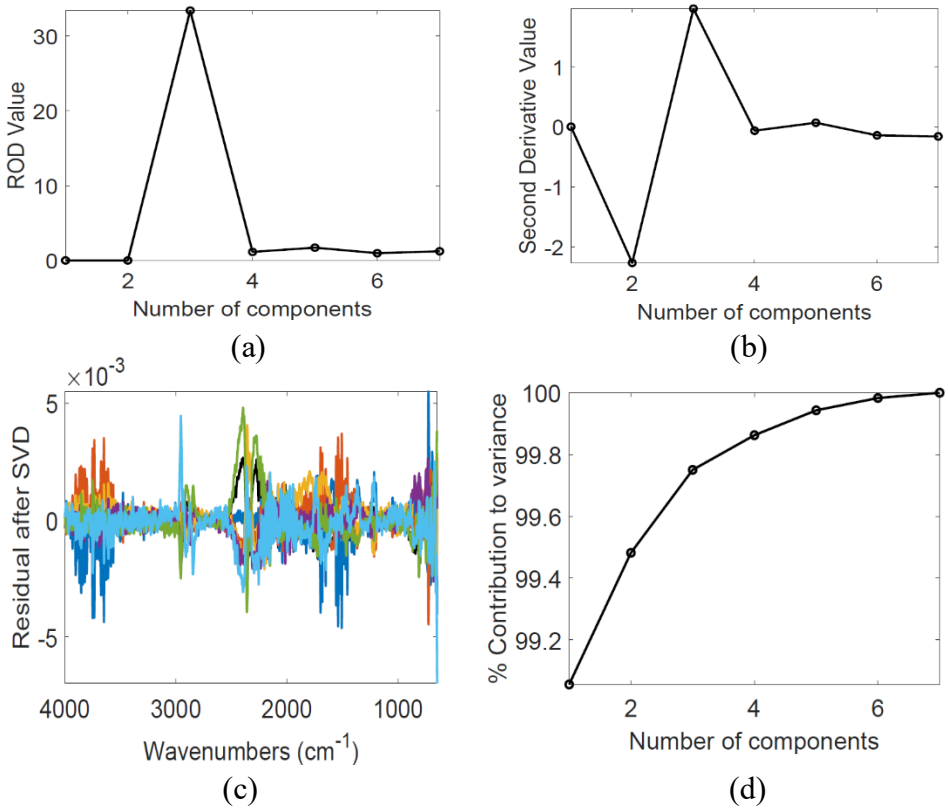


Figure D.9. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 380 ºC.

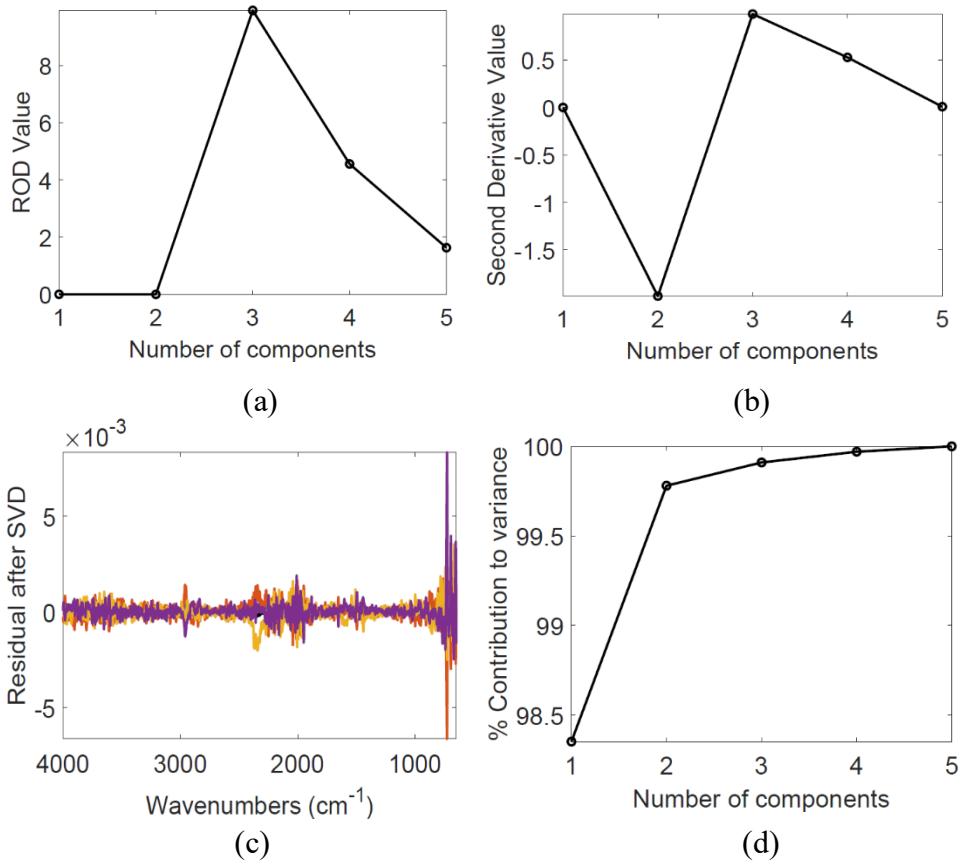


Figure D.10. Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each component in the system. These results correspond to data obtained at 420 ºC.

The values of performance indicators (LOF and R^2) for SVD with 2, 3 and 4 pseudo-components are given in Table D.2.

Table D.2. LOF and R^2 values (% contribution to variance) on reconstruction of the original matrix after performing SVD for the datasets at 300 ºC, 350 ºC, 380 ºC and 420 ºC.

	300 ºC		350 ºC			380 ºC			420 ºC		
No. of components	**2**	**3**	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**
LOF	2.38	8.27E-14	3.09	2.17	1.54	7.20	4.99	3.71	4.72	2.93	1.83
$\boldsymbol{R^2}$	99.94	100	99.90	99.95	99.97	99.48	99.75	99.86	99.78	99.91	99.96

D.5 Results and discussion: Initial concentration estimates

The initial estimates of concentration profiles at 300 ºC are given in Figure D.11.

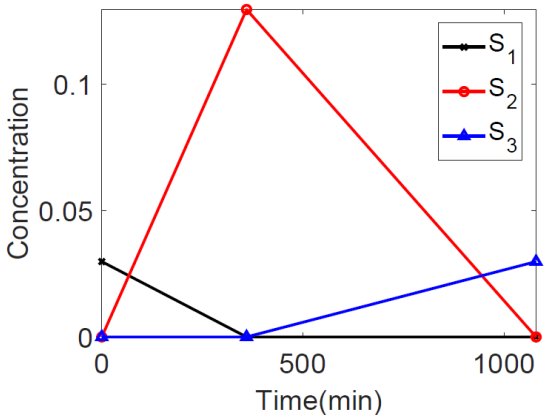


Figure D.11. Initial concentration estimates for S_1, S_2 and S_3 at 300 ºC.

D.6 Results and discussion: ALS-optimized profiles and spectra-derived quantitative parameters

The residuals obtained after subtracting the ALS-reproduced matrix from the original matrix for datasets at all temperatures are given in Figure D.12.

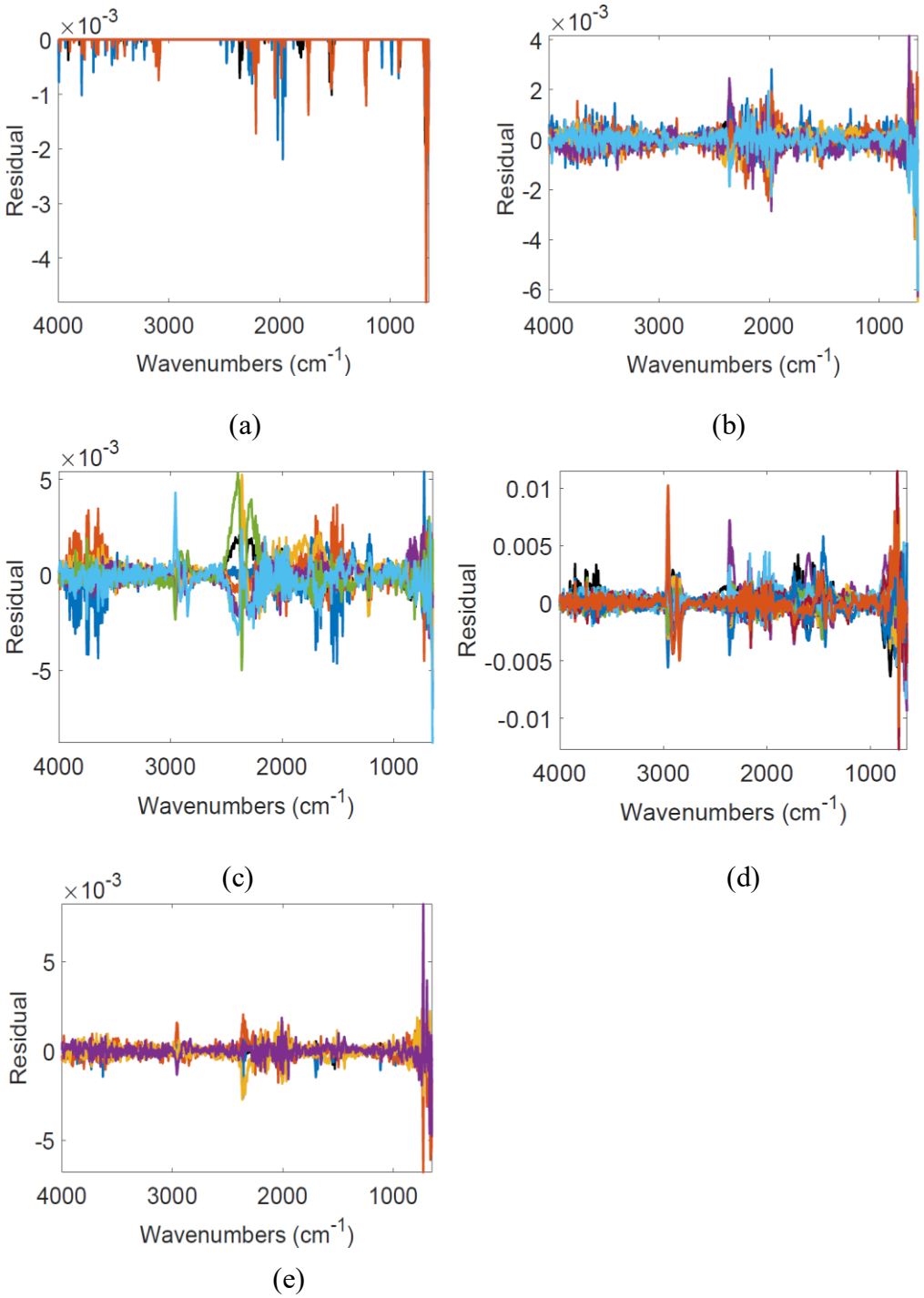
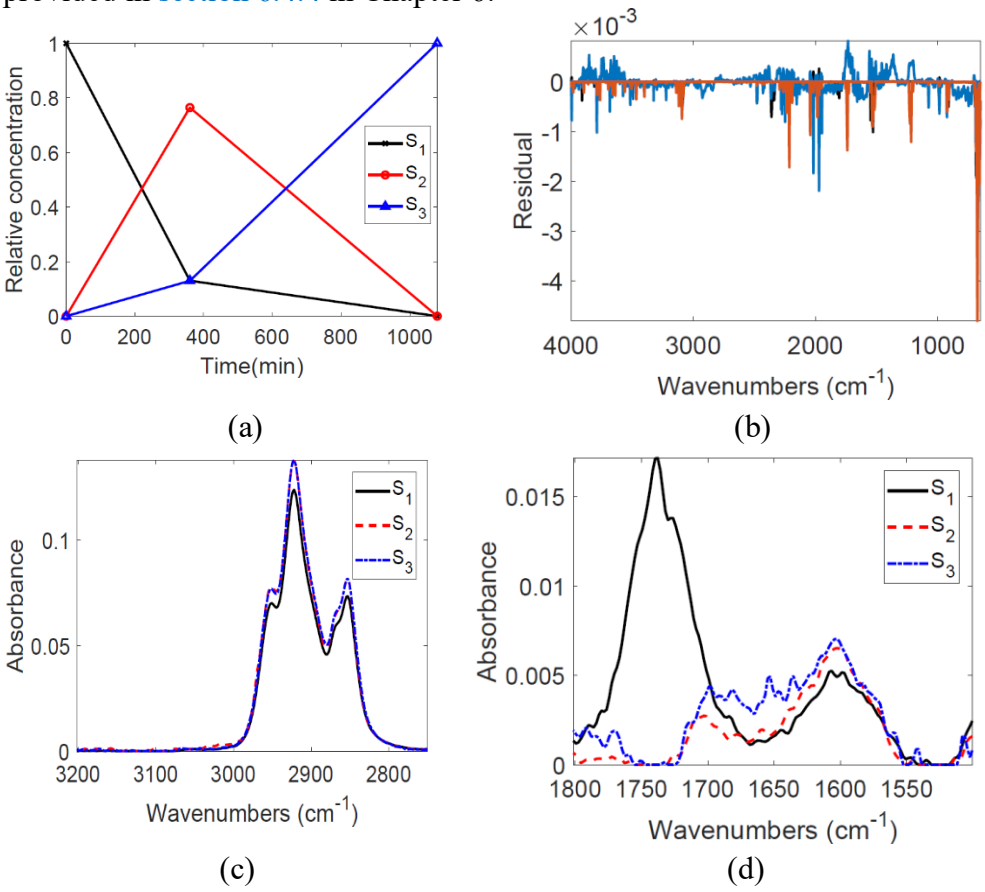


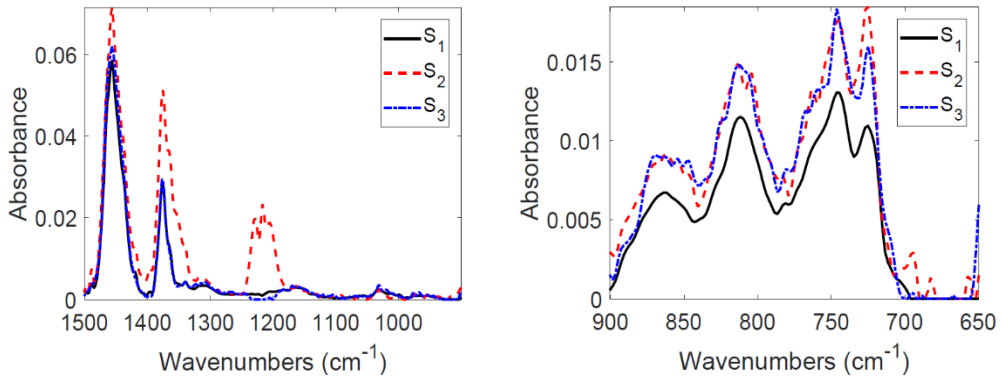
Figure D.12. ALS residuals for datasets obtained at: (a) 300 ºC; (b) 350 ºC; (c) 380 ºC; (d) 400 ºC; (e) 420 ºC.

D.7 Results and discussion: PSO-optimized concentration and spectral profiles

D.7.1 Results at 300 ºC

The concentration and spectral profiles when the ALS-PSO algorithm was used to resolve the FTIR spectra obtained at 300 ºC for Athabasca bitumen is given in Figure D.13. The residual when the reproduced matrix from the ALS-PSO-resolved profiles is subtracted from the original data matrix is also provided in this figure (Figure D.13b). Discussion on the differences of these profiles with respect to ALS-optimized results in terms of resolution quality and convergence speed is provided in section 6.4.4 in Chapter 6.





(e) (f)

Figure D.13. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 300 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

D.7.2 Results at 350 ºC

The concentration and spectral profiles when the ALS-PSO algorithm was used to resolve the FTIR spectra obtained at 350 ºC for Athabasca bitumen are given in Figure D.14. The residual when the reproduced matrix from the ALS-PSO-resolved profiles is subtracted from the original data matrix is also provided in this figure. Discussion on the differences of these profiles from the ALS-optimized results in terms of resolution quality and convergence speed is given in section 6.4.4 in Chapter 6.

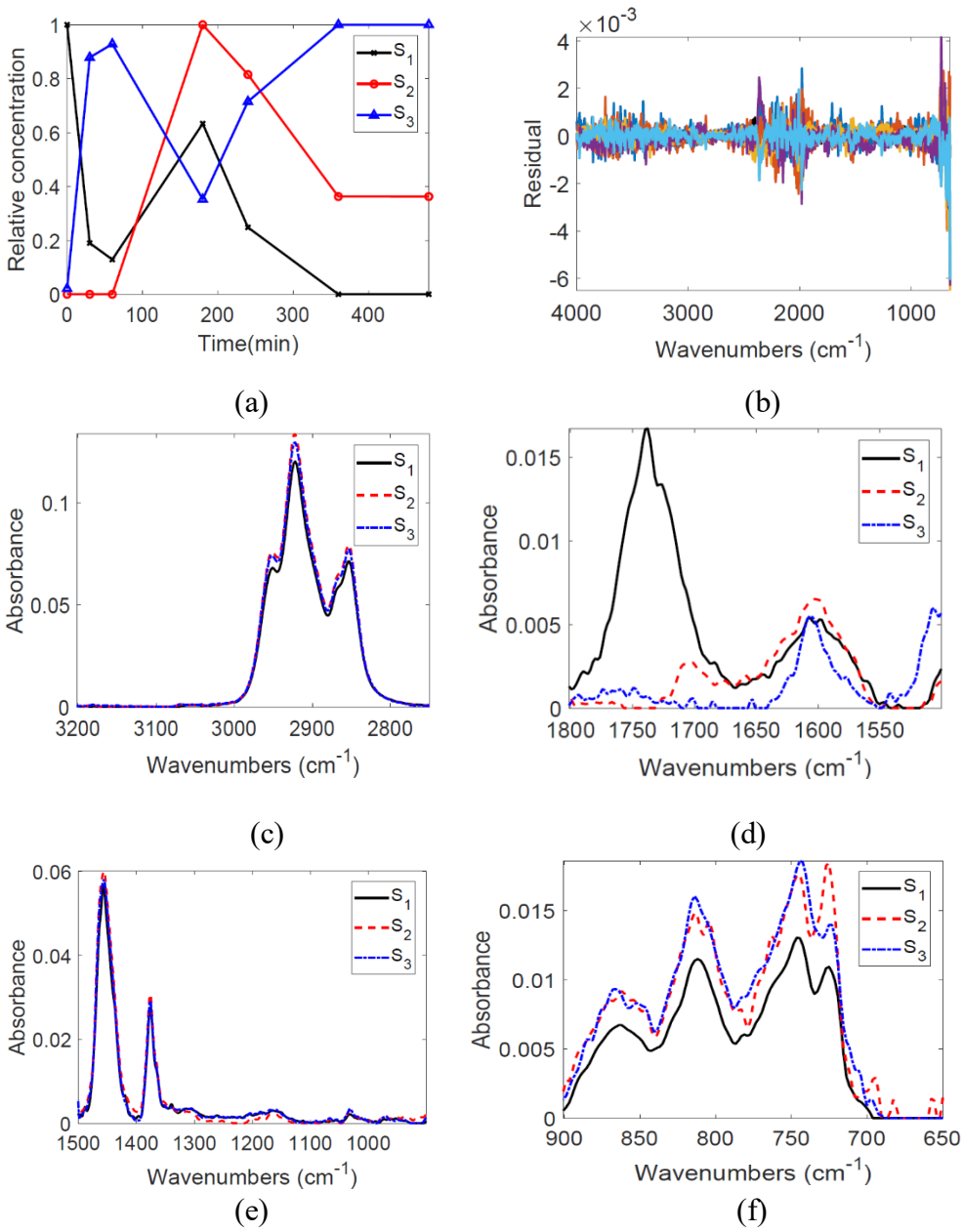


Figure D.14. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 350 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

D.7.3 Results at 380 ºC

Figure D.15 provides the ALS-PSO-resolved concentration and spectral profiles for the 380 ºC dataset. The residual plot when the reproduced matrix is subtracted from the original data matrix is also provided in the figure.

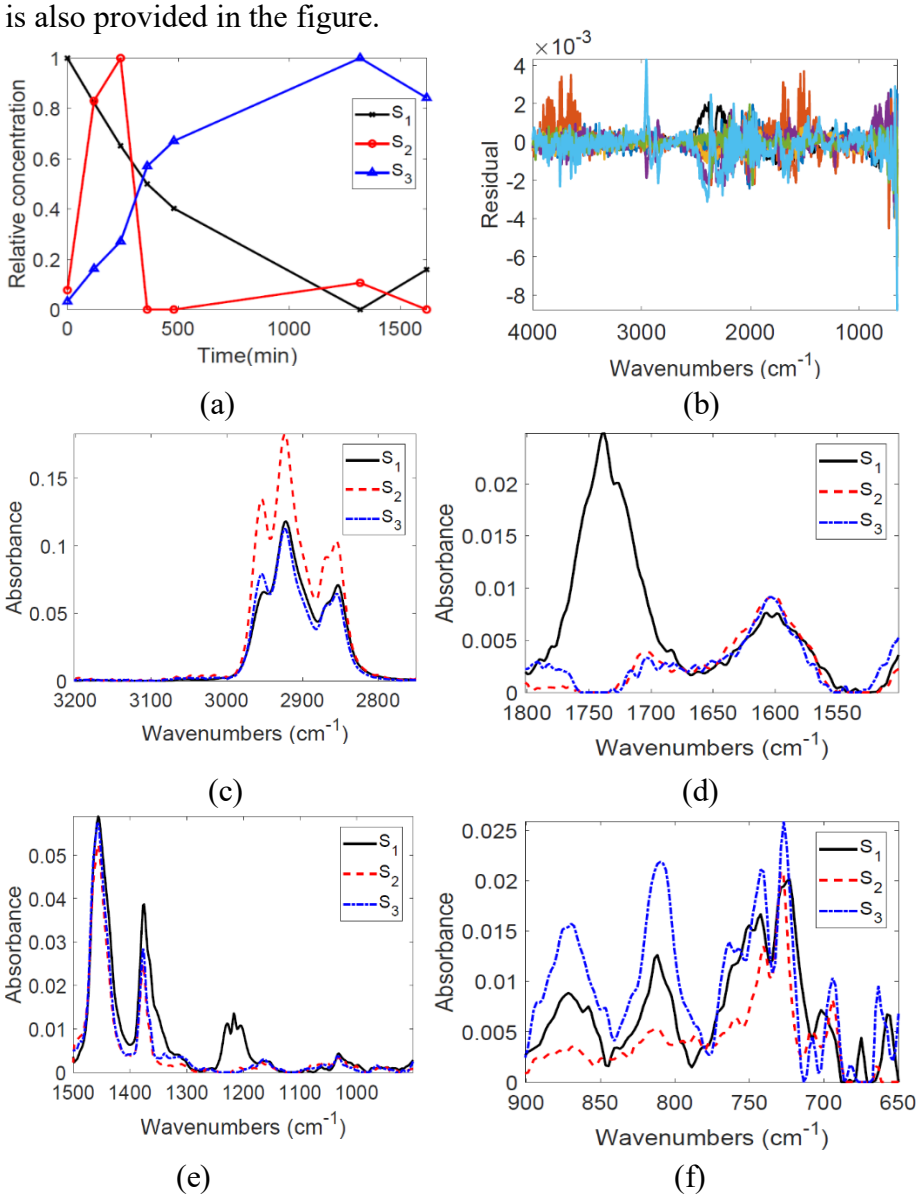


Figure D.15. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 380 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

D.7.4 Results at 400 ºC

Figure D.16 gives the ALS-PSO resolved final profiles for the dataset obtained at 400 ºC. The residual plot when the reproduced matrix is subtracted from the original data matrix is also provided in the figure.

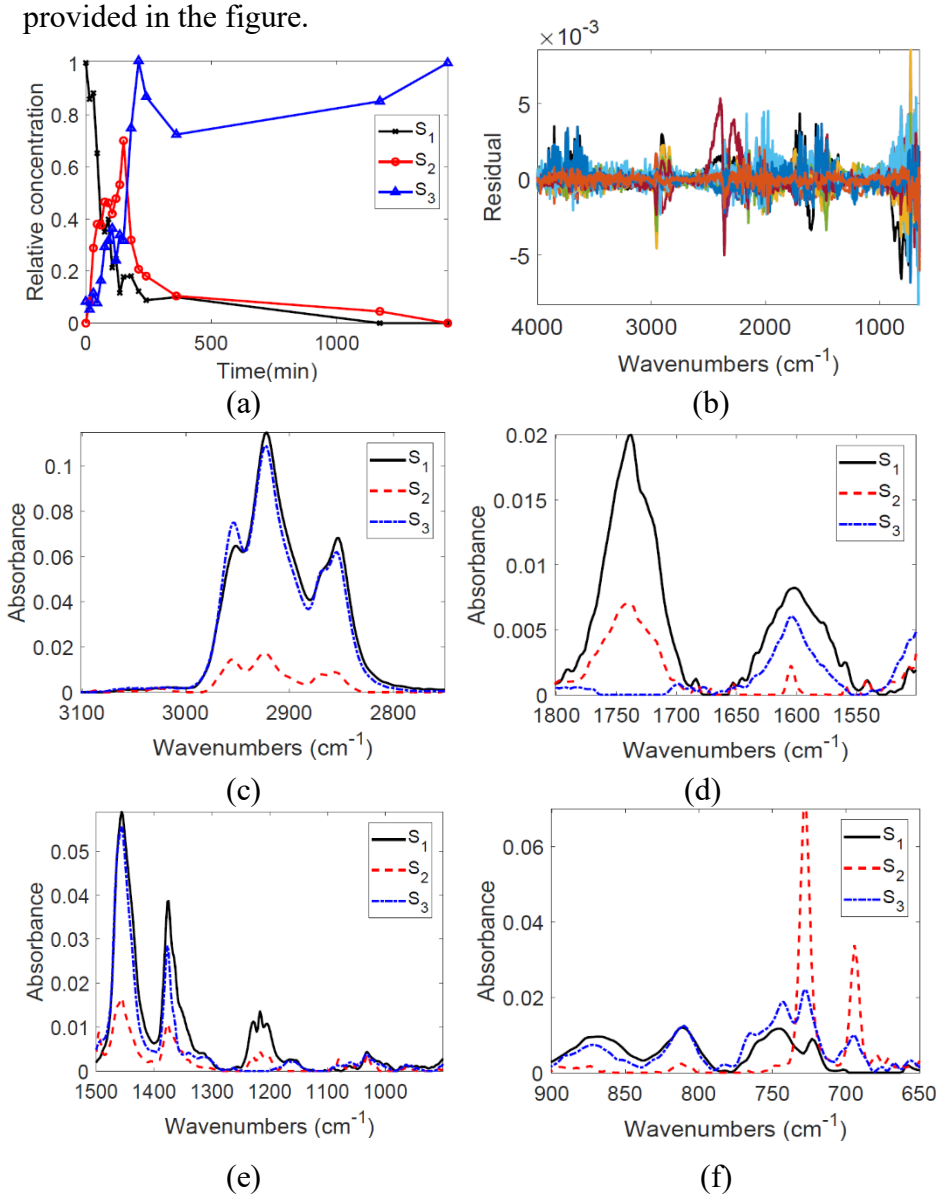


Figure D.16. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 400 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shows as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

D.7.5 Results at 420 ºC

Figure D.17 provides the concentration and spectral profiles for the ALS-PSO optimized profiles including the residual obtained when the reproduced data matrix is subtracted from the original matrix.

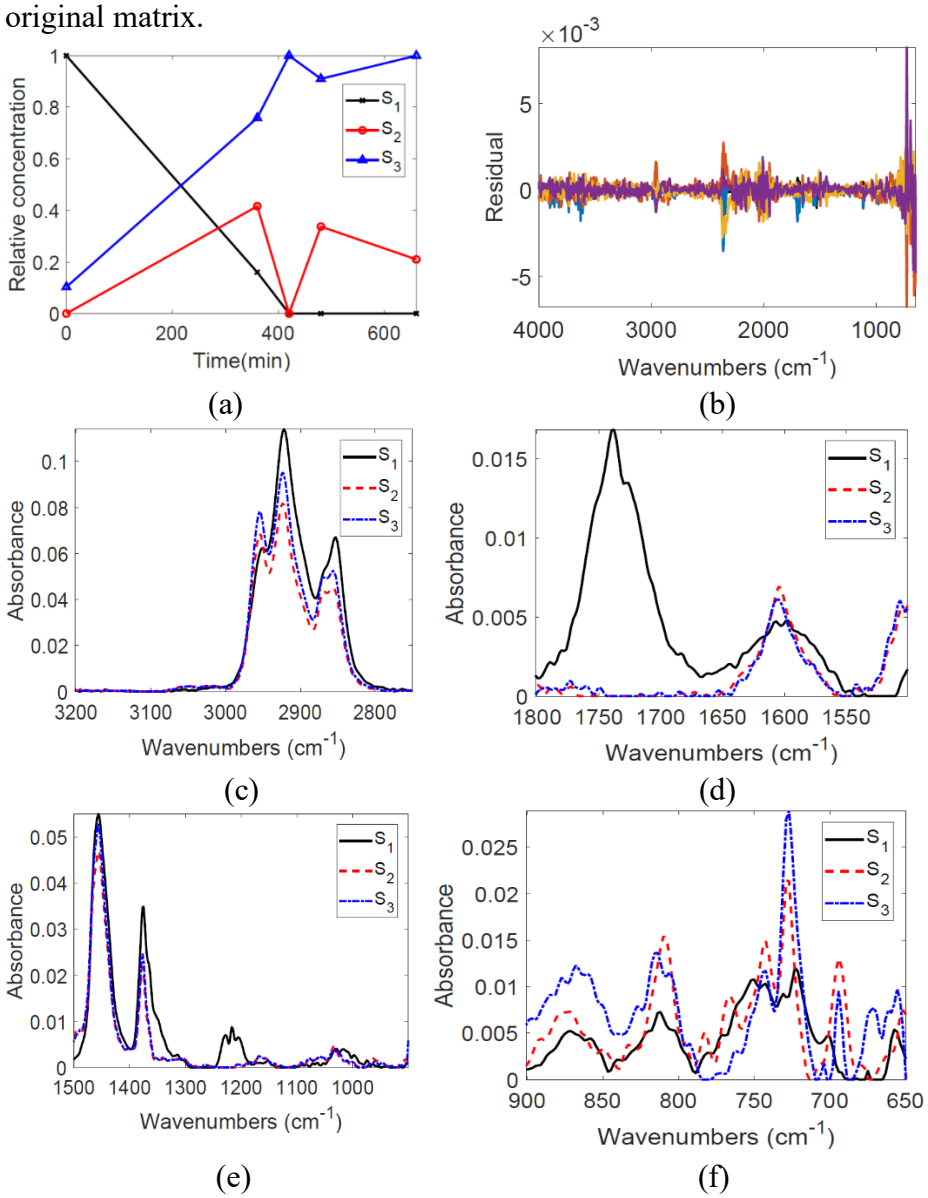


Figure D.17. Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 420 ºC. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shows as absorbance vs. wavenumber in the ranges: (c) 3200 – 2750 cm^{-1}; (d) 1800 – 1500 cm^{-1}; (e) 1500 – 900 cm^{-1}; (f) 900 – 650 cm^{-1}.

D.8 Results and discussion: BHC and associated chemical signatures relative to the others

The variation of effective intensity for each wavenumber in this cluster is shown in Figure D.18.

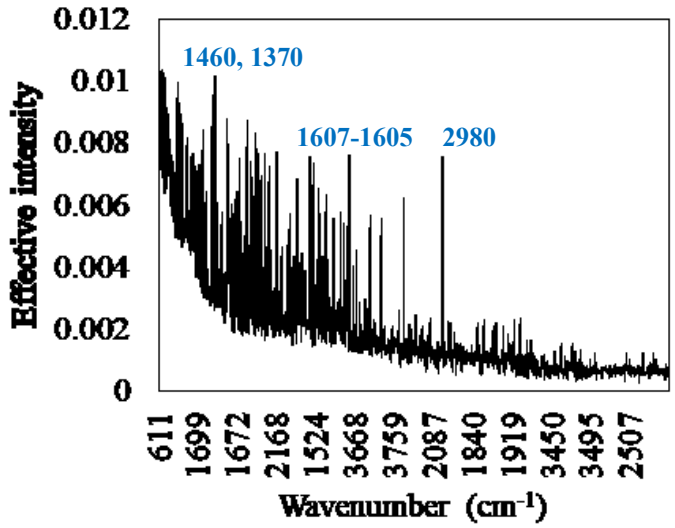
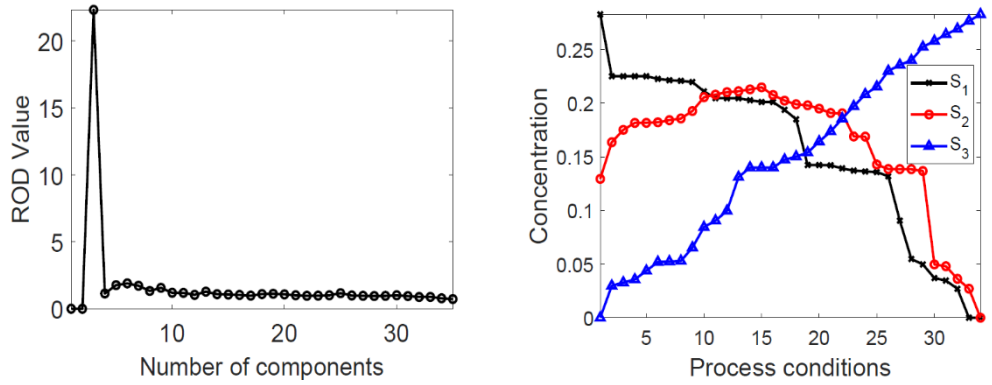


Figure D.18. Effective intensity for each wavenumber in the fifth cluster (Table 6.14 in Chapter 6). Some of the important peaks are indicated.

D.9 Results and discussion: ALS-optimized profiles for the global model

Figure D.19 provides the plots for the ROD and initial concentration estimates obtained through EFA for the 35 samples when the augmented matrix consisting of all temperatures and respective reaction times was used for SMCR analysis.



(a) (b)

Figure D.19. Plots of: (a) ROD vs. number of components and (b) initial estimates of concentration obtained through EFA for the 35 samples at various process conditions used in the SMCR-ALS global model.

D.10 References

(1) Shi, Y.; Eberhart, R. A Modified Particle Swarm Optimizer. In *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*; IEEE: Anchorage, AK, 2002; pp 69–73.

(2) Eberhart, R. C.; Yuhui Shi. Tracking and Optimizing Dynamic Systems with Particle Swarms. In *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*; IEEE: Seoul, 2002; pp 94–100.

(3) Xin, J.; Chen, G.; Hai, Y. A Particle Swarm Optimizer with Multi-Stage Linearly-Decreasing Inertia Weight. In *2009 International Joint Conference on Computational Sciences and Optimization*; IEEE: Sanya, China, 2009; pp 505–508.

(4) Arumugam, M. S.; Rao, M. V. C. On the Performance of the Particle Swarm Optimization Algorithm with Various Inertia Weight Variants for Computing Optimal Control of a Class of Hybrid Systems. *Discret. Dyn. Nat. Soc.* **2006**, *2006*, 1–17.

(5) Moré, J. J.; Sorensen, D. C. Computing a Trust Region Step. *SIAM J. Sci. Stat. Comput.* **1983**, *4*, 553–572.

(6) Steihaug, T. The Conjugate Gradient Method and Trust Regions in Large Scale Optimization. *SIAM J. Numer. Anal.* **1983**, *20*, 626–637.

(7) Byrd, R. H.; Schnabel, R. B.; Shultz, G. A. Approximate Solution of the Trust Region

Problem by Minimization over Two-Dimensional Subspaces. *Math. Program.* **1988**, *40*, 247–263.

(8) Coleman, T. F.; Verma, A. A Preconditioned Conjugate Gradient Approach to Linear Equality Constrained Minimization. *Comput. Optim. Appl.* **2001**, *20*, 61–72.

(9) Wu, H. C. The Karush–Kuhn–Tucker Optimality Conditions in Multiobjective Programming Problems with Interval-Valued Objective Functions. *Eur. J. Oper. Res.* **2008**, *196*, 49–60.

(10) Suykens, J. A. K.; Van Gestel, T.; De Brabanter, J.; De Moor, B.; Vandewalle, J. *Least Squares Support Vector Machines*; World Scientific: River Edge, NJ, 2002.

(11) Fletcher, R. *Practical Methods of Optimization*; John Wiley & Sons: New York, 1987.

(12) Schittkowski, K. NLPQL: A Fortran Subroutine Solving Constrained Nonlinear Programming Problems. *Ann. Oper. Res.* **1986**, *5*, 485–500.

(13) Han, S. P. A Globally Convergent Method for Nonlinear Programming. *J. Optim. Theory Appl.* **1977**, *22*, 297–309.

(14) Powell, M. J. A Fast Algorithm for Nonlinearly Constrained Optimization Calculations. In *Numerical Analysis*; Springer: Berlin, Heidelberg, 1978; pp 144–157.

(15) Waltz, R. A.; Morales, J. L.; Nocedal, J.; Orban, D. An Interior Algorithm for Nonlinear Optimization That Combines Line Search and Trust Region Steps. *Math. Program.* **2006**, *107*, 391–408.