Machine learning-based monitoring of complex reactive systems

by

Anjana Thimmaiah Puliyanda

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

Doctor of Philosophy

in

Process Control

Department of Chemical and Materials Engineering
University of Alberta

© Anjana Thimmaiah Puliyanda, 2022

Abstract

Processing of complex feedstocks for the production of value-added chemicals and fuels is industrially important. The lack of *a priori* knowledge of the innumerable species and the reaction pathways governing their conversion, has posed challenges to monitoring these processes. Although, data-driven models have been used, their lack of interpretability and an end-to-end modeling framework has limited the efficiency of diagnostic decisions in process monitoring. On the other hand, systems where the mechanistic knowledge of the species and their reactions are arrived at from first-principles simulations, face computational challenges in the deployment of such models for the process design. This thesis focuses on the following two aspects: (i) developing inferential machine learning models to enhance the interpretability of data-driven models, and (ii) developing predictive machine learning models to reduce the computational cost of first-principles simulations, in modeling chemical systems.

The first aspect of developing inferential machine learning models focuses on the identification of species, reaction pathways, and kinetic parameter estimation from spectroscopic data of the system, with application to the visbreaking of bitumen. Spectroscopic curve resolution methods that are structure-preserving, interpretable, and jointly parse data from multiple sensors, to extract latent features for species identification have been presented with an increasing degree of sophistication as follows:(i) self-modeling multivariate curve resolution (SMCR), (ii) joint non-negative matrix factorization (JNMF) as a data fusion analogue of SMCR where regularization constraints act like chemical information sieves to handle complementary, orthogonal and redundant features in the latent factorization of multi-sensor data and (iii)

joint non-negative tensor factorization (JNTF) as a structure-preserving higher order analogue of JNMF. Next, Bayesian structure learning among the extracted spectral features has been used to causally infer plausible reaction pathways that have been validated by domain knowledge. Finally, the latent factorization and causal inference models have been used as an engine to interpret the modes identified by training hidden semi-Markov models on spectra. This captures the time scales and dynamics of reaction mechanisms with changing temperatures, for the realtime monitoring of reactive systems purely from spectroscopic data. Projections of spectroscopic data onto the temporal mode of data collection via latent factorization, are interpreted as concentrations. Kinetic models constrained by physical laws and the reaction adjacency matrix deduced from the Bayesian network structure are implemented using chemical neural ODEs trained on the temporal concentrations. The prediction accuracy is seen to depend on the ability of latent factorization to handle process noise.

The second aspect of training predictive machine learning models, focuses on not only reducing the computational cost of the *ab initio* molecular dynamics (AIMD) simulations of chemical systems, but also the cost in itself of developing such models. This has been demonstrated with application to the transglycosylation of cellobiose, to assess whether or not the solvent molecules reorganize significantly in going from the reactant to the product configurations. A self-supervised 3D convolutional neural network autoencoder is trained to extract features from the reactant and product simulation trajectories, the probability distributions across the difference between which is used to assess if the solvent reorganization is significant. Cellobiose systems at lower temperatures are found to reorganize to a greater extent than those at higher temperatures, consistent with the decrease in the activation free energy barrier as temperature increases. Similarity between the reactant configuration features of other chemical systems with those extracted from that of the cellobiose systems, is then used as a basis to inform the extent of reorganization in the product profiles, without having to explicitly run AIMD simulations for the same.

Preface

The research presented in this thesis was carried out under the supervision of Dr Vinay Prasad and Dr Zukui Li. This research was primarily supported by funding from MITACS Globalink and Alberta Innovates (now InnoTech). The contributions and details of the chapters in this paper-based thesis are outlined as under.

A part of Chapter 1 of this thesis has been published as: A. Puliyanda, K. Srinivasan, K. Sivaramakrishnan, V. Prasad, A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems, *Digital Chemical Engineering* 2022, doi: 10.1016/j.dche.2021.100009. Anjana Puliyanda, Karthik Srinivasan, Kaushik Sivaramakrishnan and Vinay Prasad were involved in the conceptualization and writing the original draft of the manuscript.

Chapter 2 of this thesis is published as: K. Sivaramakrishnan, A. Puliyanda, A. de Klerk, V. Prasad, A data-driven approach to generate pseudo-reaction sequences for the thermal conversion of Athabasca bitumen, *React. Chem. Eng.* 2021, doi: 10.1039/D0RE00321B. Anjana Puliyanda and Kaushik Sivaramakrishnan were responsible for the methodology, implementation, formal analysis and manuscript composition. Arno de Klerk and Vinay Prasad were involved with conceptualization, methodology and writing the original draft of the manuscript.

Chapter 3 of this thesis is published as: A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, V. Prasad, Data fusion by joint nonnegative matrix factorization for hypothesizing pseudo-chemistry using Bayesian networks, *React. Chem. Eng.* 2020, doi: 10.1039/D0RE00147C. Anjana Puliyanda was responsible for the methodology, implementation, formal analysis and writing the original draft of the manuscript.

Kaushik Sivaramakrishnan was responsible for the formal analysis and manuscript composition. Zukui Li, Arno de Klerk and Vinay Prasad were responsible for the conceptualization, methodology and writing the original draft of the manuscript.

Chapter 4 of this thesis is published as: A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, and V. Prasad, Structure-preserving joint non-negative tensor factorization to identify reaction pathways using Bayesian networks, *J. Chem. Inf. Model.* 2021, doi: 10.1021/acs.jcim.1c00789. Anjana Puliyanda was responsible for the methodology, implementation, formal analysis and writing the original draft of the manuscript. Kaushik Sivaramakrishnan was responsible for formal analysis and manuscript composition. Zukui Li, Arno de Klerk and Vinay Prasad were involved in the conceptualization, methodology and writing the original draft of the manuscript.

Chapter 5 of this thesis has been submitted as: A. Puliyanda, Z. Li, V. Prasad, Real-time monitoring of reaction mechanisms using Hidden Semi-Markov Models for mode identification, *J. Process Control.* Anjana Puliyanda, Zukui Li and Vinay Prasad were responsible for the conceptualization, methodology, implementation and writing the original draft of the manuscript.

Chapter 6 of this thesis will be submitted as: A. Puliyanda, Z. Li, V. Prasad, Chemical reaction neural ODEs and latent factorization to deduce kinetic models from spectroscopic data. Anjana Puliyanda, Zukui Li and Vinay Prasad are responsible for the conceptualization, methodology, implementation and writing the original draft of the manuscript.

Chapter 7 of this thesis will be submitted as: A. Puliyanda, A.M.D. Padmanathan, S.H. Mushrif, V. Prasad, 3D Convolution neural network autoencoder for the prediction of solvent reorganization from MD simulation data. Anjana Puliyanda is responsible for the conceptualization, methodology, implementation and writing the original draft of the manuscript. Arul Mozhi Devan Padmanathan is responsible for supplying the simulation data and formal analysis. Samir Hemant Mushrif and Vinay Prasad are responsible for the conceptualization, methodology and manuscript composition.

To my wonderful parents, Sangeetha and Thimmaiah

Acknowledgements

I am immensely grateful to my academic supervisors, Dr Vinay Prasad and Dr Zukui Li for their constant guidance, support and encouragement throughout the course of my doctoral studies. Dr Prasad, your efforts in always drawing me to the bigger picture before focusing on the research gaps that our objectives sought to address, has brought invaluable perspective to my research pursuits. Dr Li, your attention to detail and prompt feedback on my progress has been vital in refining my research output. It has been enriching to imbibe from the research acumen of both my supervisors, in addition to taking a leaf out of their personal philosophies.

My sincere thanks to Dr Arno de Klerk, for being a part of the supervisory committee and taking interest in my work by providing recommendations after painstakingly examining my written drafts. I appreciate the enthusiastic discussions that stemmed from my collaboration with Dr Mushrif, highlighting the importance of extending the modeling philosophy of systems engineering across various length and time scales.

I acknowledge the generous financial support received from Alberta Innovates (now InnoTech), MITACS Globalink, the Chemical and Materials Engineering Department and the Faculty of Engineering at the University of Alberta.

I credit my former and current colleagues from the research groups of both my supervisors with the constructive working environment during my PhD. Thank you Dereje for being available for research discussions when I was starting out and for training me on the DeltaV interface so that I could smoothly conduct the control lab sessions; Khushaal, for discussions via email; Kaushik, for working closely with me and helping us reach mid-ground, despite our different working styles; Maryam,

Fereshteh, Ajay and Pedro, who shared adjacent desks with me in the DICE 3-225 office space, thank you for taking off your earphones to entertain my questions and for the occasional digressions from work with your lively banter on many a cold Edmonton day; Gokul, for being an equally wonderful friend and for the water cooler conversations that not only helped with hydration, but also with exchanging quick thoughts on training ML models; Farough and Hossein, for all things Hessian and optimization; Sanjula, for the memorable conference trips; Sushmitha, for my first week in Edmonton. It was a pleasure to get to know Karthik and Kiran, and interact with Kevin and Prince. I would also like to thank Arul from Dr Mushrif's group for our fruitful collaboration and long-drawn-out meetings; and all the members of the Computer Process Control (CPC) group who have supported me over the years.

My fledgling ideas about research in Chemical Engineering were shaped by my former teachers during my undergraduate years in NIT, Surathkal. Thanks to Dr Sai Dutta, for planting the allure of how multi-disciplinary the field is, Dr Hari Mahalingam and Dr Ashraf Ali for introducing me to process modeling and simulation.

Outside of work, I am grateful to have forged great relationships with my housemates, Shweta and Dhanvini. Dan, whom I have known since my NIT days was attuned to my idiosyncrasies and was readily available for layman feedback on dry runs of presentations. Deepa, would always be open to exploring new cafes, libraries, art galleries and ice cream shops. The Bhatts, the Somayajis, Shobita and Neeraj, have been gracious hosts in Edmonton. A special mention to Chinnappa uncle and the late Jaani aunty of Calgary, for making me feel at home, halfway across the globe.

Last, and definitely the most, I would like to thank my parents, back in India for their love, support, encouragement, patience, understanding and courage in sending their only child all the way to Canada. I am indebted to them for being happy, content and partaking in my journey of self-actualization.

Edmonton 2022

Anjana Puliyanda

Table of Contents

List of Tables

List of Figures

"Mathematical reasoning may be regarded rather schematically as the exercise of a combination of two facilities, which we may call intuition and ingenuity."

-Alan Turing

Chapter 1

Introduction

Process monitoring of chemically reactive systems is crucial for product quality control, for the optimization and control of the process itself to ensure plant safety [1], and to realize sustainable production objectives that are central to process intensification [2]. The challenges faced by the systems engineering approach to process monitoring of reactive systems via design, optimization and control broadly encompass

1. Model development of the complex reactive system, when the complete identification of the species and reactions may not be available. This has popularized the development of data-driven system inferential models [3], [4] based on the molecular-level information obtained from process integrated spectral analyzers using flow cells, quartz windows or immersion probes that are fast, noninvasive, non-destructive, inexpensive and do not require sample preparation [5], [6].

2. Model deployment, in the event mechanistic knowledge of the species and reactions are used for the first-principles simulation of chemical systems. Although the atomic simulations can potentially access length and time scales beyond the limit of experiments, they are found to be computationally intractable for systems with a large number of atoms [7]. This has prompted the training of

A portion of this chapter has been published as: A. Puliyanda, K. Srinivasan, K. Sivaramakrishnan, V. Prasad. A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems. *Digital Chemical Engineering* **2022**, *2*, 100009.

predictive machine learning models on the mechanistic simulations, as a computationally efficient surrogate that could then be deployed for advancements in drug design [8], computational chemistry in molecular and materials modeling [9], retrosynthesis and catalysis [7].

The thesis is directed to tackle the aforementioned challenges with regard to model development and deployment for chemically reactive systems, with the aim of furthering process intensification. This chapter provides an overview of the data-driven strategies used for species identification, reaction pathway determination and kinetic parameter estimation, before highlighting the knowledge gap that has motivated the development of such an end-to-end data-driven system inferential framework with application to the partial upgrading of bitumen.

This chapter also provides an overview of predictive machine learning models trained on first-principles molecular dynamics simulation data as a computationally efficient alternative. The cost of training such predictive models owing to the generation of target labels via sampling calculations from the simulation data has prompted the development of a self-supervised framework, where the density distribution of the extracted features informs label assignment for predictive models, with application to predicting the extent of solvent reorganization from the simulation trajectories of the reactive cellobiose and fructose systems, as demonstrated in this thesis.

1.1 System inferential modeling framework from process data of reactive systems

Bitumen being a complex reactive mixture is lacking in the exhaustive enumeration of its constituent species, let alone the reaction pathways governing their conversion. This poses a challenge to monitor composition changes in bitumen arising from reactions occurring during the partial upgrading of the complex feedstock; to obtain a pumpable product stream, in compliance with the North American standards of

pipeline transport (viscosity <350Cst at 7.5°C (winter) and density< $940 kg/m^3$ at 15°C and an olefinic content<1 decene equivalent) [10]. In such situations, reaction monitoring relies on experimental data of the system, obtained by integrating process analytical tools like spectroscopic sensors with reactors, for species detection and measuring analyte concentrations, to facilitate higher control of product composition and process intensification [11]. Real-time data from spectral measurement techniques *viz.* Fourier transform infrared (FTIR), Raman, Ultraviolet–visible (UV-vis), Nuclear magnetic resonance (NMR) and Mass spectrometry (MS) are used for species identification followed by monitoring changes to the structural and elemental composition during the conversion of reactants to products from which mechanisms are deduced for reaction optimization and process design [12]. The use of multiple spectral analyzers is seen to increase the confidence of the kinetic model estimates, making process analytical tools popular in industry to monitor chemical processes [13].

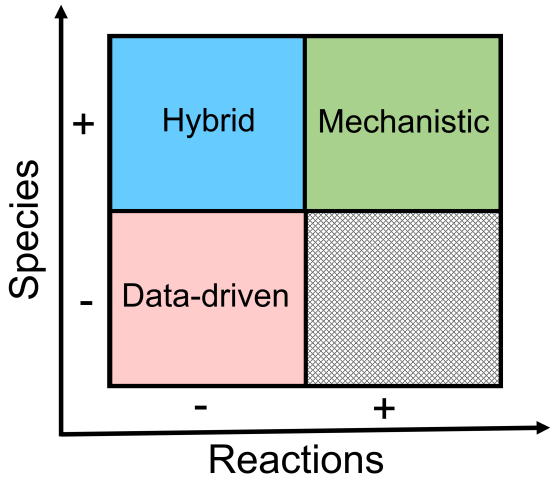


Figure 1.1: Classification of reaction systems based on degree of knowledge of species and reactions, and approaches for pathway determination.

At this point, it is useful to define the various classes of reaction monitoring and pathway determination problems [14]. For very well-understood systems, where full knowledge of participating species and reactions is available, the pathway determina-

tion problem is already solved. The monitoring problem, given the highly developed understanding of the system, can focus on the (real-time) estimation of kinetics and the full compositional profile of the products of the reaction scheme. When species but not (all) reactions are known, pathway determination may be attempted using heuristic approaches relying on the chemical expertise of humans, or through automated approaches. In this case, the monitoring problem typically focuses on estimation of conversion, apparent reaction order and apparent kinetics in the absence of pathway determination. If neither species nor reactions are known well (as in the analysis of petroleum residues or their resulting hydrocarbon fractions), then the estimation of conversion (which is often poorly defined) is usually all that is possible in the absence of (data-driven) methods for the identification of species and then of reaction pathways. In this case, too, the determination of reaction pathways can enable more sophisticated analyses for monitoring, and the ability to track the compositional profile of the products. Figure 1.1 presents the classification of systems based on the degree of knowledge of species and reactions ($-$ implying a lack of knowledge and $+$ implying knowledge of species/reactions), and the corresponding typical approach for pathway determination. A description of the methods to automate the reaction pathways and for the online monitoring of complex chemical systems via data-driven approaches, follows from Section 1.1.1 to Section 1.1.3.

1.1.1 Species identification

Multivariate statistical process monitoring (MSPM) for the design, analysis and control of systems lacking in *a priori* knowledge [15] has been popularized in pharma, food and biotechnology industries [16], with the increased data collection from analyzers that are built into a manufacturing process to relay molecular level information [6]. Spectroscopic and chromatographic analyzers (Fluorescence, Visual, Near infrared (NIR), Infrared (IR), Raman, proton nuclear magnetic resonance (^{1}H-NMR)) [17] [18], acquire process data that are high dimensional, non-causal, non-full rank,

noisy and have missing values. This has resulted in choosing MSPM methods to build mathematical system inferential models in the latent variable space for process monitoring [3][4] by developing statistical techniques like calibration, multivariate analysis and curve resolution [19]. This helps in obtaining compositions and chemical signatures of species from measurements (chromatographic and spectroscopic), as a first step in the data-driven monitoring of complex reactive systems.

Multivariate calibration refers to the process of relating, correlating, or modeling analyte concentration or the measured value of a physical or chemical property to a measured response [20]. Partial least-squares (PLS) regression is popular in multivariate calibration as it modifies relations between sets of the observed variables by a small number of latent variables that maximize covariance in predictor and response space (not directly observed or measured) by incorporating regression and dimension reduction techniques [21]. Multivariate analysis techniques like hierarchical clustering analysis (HCA) was applied on emissions from materials like polymers to identify spectral groupings [22], that are matched to compound classes in standard libraries [23] in a bid to automate species identification from spectra of complex mixtures.

Curve resolution in spectral data is a factor analytical decomposition that works by resolving the data into concentration and spectral profiles either bilinearly or multilinearly using the self-modeling multivariate curve resolutionalternating least-squares (SMCR-ALS) [24] [25] and the parallel factor analysis (PARAFAC) [26] [27] models, respectively. The initial estimates for the decision variables can be obtained by evolving factor analysis (EFA) on a row-wise augmented data matrix both in the forward and backward direction [28] if the data has an intrinsic order or by use of a global search technique in the feasible space using particle swarm optimization (PSO) [29]. The concentration and spectral profiles are subject to physically meaningful constraints like non-negativity, closure, unimodality to obtain a unique decomposition free from rotational and intensity ambiguities [30]. The PARAFAC model is inherently free of these ambiguities and is a unique decomposition as it preserves the

multi-linear structure of data and the inter-modal latent factor interactions while resolving the data into independent factor matrices [31], unlike the less restricted Tucker decomposition that decomposes data by singular value decomposition (SVD) into orthogonal factor matrices [32] .These decomposition methods depend on a parameter called the rank or the number of pseudo-components which capture most of the variance in the data. In orthogonal decomposition it is determined using emprical metrics [33] based on principles of SVD into principal uncorrleated directions; while in higher order PARAFAC decompositions methods like core consistency [34] and split half anlysis are used for rank determination based on the principles of PARAFAC being a restricted Tucker model [32]. Hence, it can be seen that resolution methods are applied to complex mixtures and aim 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 hyphenated analytical techniques employed without prior knowledge about the system; to then enable further chemical interpretation and understanding of reaction pathways [35].

1.1.2 Reaction pathway identification

Once the species have been identified, deciphering the reaction mechanisms underlying complex reactive systems is a pre-requisite to develop kinetic models to devise online monitoring strategies [36]. If the rules governing species conversion are known, then reaction pathway identification can be automated by first translating chemical knowledge into machine-level representations as follows [37]: (i) encoding chemical species using molecular descriptors of reaction cores using SMILES/SMARTS strings, Morgan fingerprints, or by using matrix representations of species as molecular graphs via edge/vertex adjacency matrices, (ii) encoding reaction rules as templates by representing a reaction as a difference in the fingerprints between the product and reactant representations. Then an automatic reaction network generator consisting of a generation algorithm, successively applies these rules to the species until a termination

criteria is met, as with the Rule Input Network Generator (RING) [38]. Manually encoding reaction rules is not only cumbersome but is also restrictive when it comes to discovery of novel mechanisms. Hence, there has been the development of algorithms that automatically extract reaction templates from databases by atom-atom mapping (AAM) between representations of reactants and products to develop transformation rules based on the identification of the reactive center about which there has been structural or bond changes [39]. In the event AAM is time-consuming to automate template learning from databases, there is evidence of using template-free approaches for predicting [40] and discovering reaction mechanisms [41].

Machine learning (ML) models are seen to improve the generalizability of automating reaction mechanisms in comparison to rule-based methods, by supplementing the aforesaid approaches with neural network machinery as universal function approximators of non-linear reaction dynamics by learning lower dimensional embeddings of input chemical data representations that are passed through non-linear activation functions to output predictions for the control and monitoring of chemical systems [42]. ML frameworks have been widely used for reaction prediction by candidate ranking. Given the reactants, reaction templates have been used to arrive at candidate products, a distribution over which is learned using a neural network for multi-class classification with a *softmax* activation in the output layer to identify the most probable products [43]. Reaction rules search for particular structural motifs in reactants prior to applying the transformation, hence are lacking in their ability to look at a molecule as a whole to check for the presence of conflicting motifs that may hinder transformation because of which given the reactants, a deep neural network was used to learn the distribution over the reaction rules to pick the most probable one [44]. Candidate ranking of rules and products has been deployed in a two-step framework for reaction prediction [45] where first, the concatenated fingerprint of the SMARTS representation of reactants and reagents is input to a neural network to learn a probability distribution across 17 types of reactions. Second, the transfor-

mation rules corresponding the most probable reaction is then applied to the input to obtain a distribution across the candidate products. A similar two-step approach has been used to realize a template-free approach to reaction prediction that involves the prediction of reactivity [46]. First, molecular graph representations of reactants and reagents are input to a graph convolutional neural network (GCNN) that ranks the likelihood of the enumerated products by the pairwise interaction of reactive sites, followed by another GCNN that learns a distribution across the Weishfehler Lehman Difference Networks (WLDN) representation of reactant-product reactivity. Reaction fingerprints obtained as the difference between molecular graphs of the reactants and products is used in the quantitative prediction of activation energy (a reaction property) using deep learning as a data-driven approach to leverage massive datasets to rank products based on whether they are energetically feasible [47]. It is seen that reaction prediction is either rule-based or template-free which involves predicting reactivity centres that however does not account for stereo-chemistry in the reacting species [48]. This calls for reaction prediction using an encoder-decoder architecture comprising 2 recurrent neural networks/ long short term memory units (RNNs/LSTMs) as using in machine translation models *seq2seq*, except that the language being translated here is the SMILES representation of reactants and products, wherein the semantics of translation synthetically deciphers the rules of the underlying reaction transformations [49]. The encoder RNN acts to classify the reaction, while the decoder RNN synthetically determines the appropriate transformations to result in the product as a SMILES output, which may run the risk of being spurious. Similar such machine translation architectures have been used to identify the electron source and sink, given the reactants, followed by proposing elementary reactions that are ranked using Siamese neural networks trained by shared weights to ultimately chain them to obtain overall reaction pathways [50].

1.1.3 Online monitoring and kinetic parameter estimation

Deduction of the species and the reaction mechanisms governing their conversions from process data is used as a basis for developing kinetic models *viz.* differential equations, Markov processes and state space representations using law of mass action kinetics, S-system or polynomial models [51]. These kinetic models are characterized by structure (species inter-conversion as reaction pathways) and parameters (rate constants, reaction orders, stoichiometric coefficients). The parameters are learned by fitting the model to experimental data by using Bayesian analysis, Monte Carlo sampling or evolutionary algorithms ([51], [52]) wherein sometimes, in the absence of prior knowledge of network topology, the structure is learned by virtue of parameter estimation [53]. These approaches can broadly be classified as [54] (a) simultaneous, where reaction pathways are learned from data by virtue of kinetic parameter estimation or (b) incremental, where given the reaction pathways, time series concentration data is used for kinetic parameter estimation either via a rate-based differentiation approach or an extent-based integration approach [55].

Simultaneous approach to kinetic modelling. The S-system formalism has been used as a tool to reverse engineer reaction networks from time series concentration data using a co-evolutionary algorithm to learn structure and parameters simultaneously by representing the non-linear system dynamics as a product of power law functions, whereby network topology is characterized by the power law parameters that give the cause-effect relationship among the species [56]. However, it does not guarantee a unique solution and fails to scale well for systems with large number of species, as the parameters scale quadratically ([56], [57]). Hence, simpler dynamical models that are linear in parameters are used to describe reaction rates based on the law of mass action as a linear combination of weighted polynomial basis functions [58]. The basis functions represent elementary reactions and are determined by model reduction in going from general to specific basis followed by least squares optimization

approach to estimate parameters by regressing against temporal process data [59]. Noisy process data could lead to multiple reaction networks giving rise to the same dynamics that are inferred from temporal concentration data. This is known as the *fundamental dogma of chemical kinetics* [60] because of which the structural identifiability of the network and its parameters is of significance in the distinguishability of optimization-based solutions, and is attempted to be achieved using model reduction to identify core reactions as to eliminate redundant terms [58].

The solution multiplicity and the reliance on user input to design the non-linear library of basis functions to represent rates (eg. reactive SINDy), can be mitigated by designing physically constrained data-driven models for kinetic structure and parameter inference [61]. Universal function approximators like neural networks that ordinarily lack interpretability in the function mappings between inputs and outputs, are architecturally designed to incorporate the law of mass action and the Arrhenius law to represent the non-linearity of reaction rates wherein the weights and biases correspond to the kinetic parameters [61]. In such hybrid approaches, first principles are built into data-driven neural network models such that all the uninterpretability goes into the feature identification and parameter estimation, compromising on obtaining causally interpretable features. This is overcome by the use of genetic algorithms to learn interpretable features that define the functional form of the non-linearities describing reaction rates, wherein a heuristic algorithm guides a population of species each associated with a vector of function transformations representative of explainable features in a combinatorial space of pre-defined functions [62]. This is followed by statistically estimating the parameters of the linear combinations of the features (or function transformations) extracted by the genetic algorithm, using OLS/LASSO regression to learn an interpretable causal map of fundamental reaction mechanisms using a purely data-driven approach in the absence of *a priori* mechanistic knowledge. Another data-driven approach for structure inference from trajectory data is to model the reaction system as a continuous time Markov chain that not only captures

reaction stochasticity but also learns kinetic parameters given by law of mass action using the maximum likelihood estimation [63]. Stochastic block model has been used as a statistical tool in conjunction with mechanistic knowledge of reaction pathways, to predict chemical reactions from species data to determine which of the mechanistic pathways can be reliably inferred from the noisy process data [64].

Incremental approach to kinetic modelling. The progress of reactions for optimization and control is characterized using the concept of extents, which is based on the principle of mass balance for example in an open homogeneous reaction system the reactants entering the reactor either convert into products in the reactor (extent of reaction), remain unconverted in the reactor (extent of inlet flow) or leave the reactor unconverted (extent of outlet flow) [65]. For a heterogeneous or a gas-liquid reaction system, the mass balance will be satisfied by incorporating an additional term for the extent of mass transfer between the phases [66]. The reaction rates can be independently deduced from the extent of reaction which is not just a pure function of species concentration or reaction variants (unless in a homogeneous batch reactor where reaction rates give true extents of reaction) due to the additional dependence on the flow variants, mass transfer variants and the invariant terms, because of which the species vector is transformed into a low-dimensional manifold of states to infer extent of reaction from concentration data [65],[66]. Alternatively, tendency models have been widely used for batch reactor optimization where the identified stoichiometry and kinetic models for a set of enumerated reactions are fit to a batch of data followed by optimization and model update over the subsequent batch of data in an iterative process over time [67]. Tendency models are a parsimonious approach to approximate the kinetics of complex reaction systems [68], without prior mechanistic knowledge of the system to predict the dynamic reaction tendencies in transient batch operations [69]. However the extent-based approach of directly inferring reaction rates from species concentration data is not only agnostic to the canonical expressions for reaction kinetics or mass transfer but also generalizes well across different reac-

tor configurations, is already a reduced model as redundant states are eliminated prior to identification, facilitates estimation of unmeasured species concentration by reconciling the measured concentrations and inlet flowrates with the variant states transformed as extents and finally integration of extent of reactions is a conducive approach to obtain model predicted concentrations that are fit to the process data for kinetic parameter estimation thereby overcoming the susceptibility to noise and sparsity while time differencing the measured concentrations for the same, as with the rate-based approaches [54],[55].

1.2 Predictive modeling from mechanistic simulations of reactive systems

When both the species and mechanisms governing their conversion are known, first-principles (*ab initio*) atomistic simulations are seen to provide insights into length and time scales, otherwise limited by experiments. High throughput molecular dynamics (MD) simulations are also seen to surmount difficulties in acquiring large amounts of consistent experimental data, in developing predictive models of a system [70]. However, these simulations involve computationally expensive quantum mechanical (QM) calculations using wave function theory, on-the-fly electron structure calculations via density function theory (DFT) and potential energy surface (PES) methods, making them intractable for systems with large number of atoms and longer time scales [7]. This has popularized the use of predictive machine learning to draw inferences from copious amounts of mechanistic simulation data, so that it can serve as a computationally efficient surrogate of the same, in the property prediction of molecular systems [71]. ML surrogates of these mechanistic models have not only improved their use for screening the design space, leading to advancements in drug design [9], computational chemistry in molecular and materials modeling [72], retrosynthesis and catalysis [7], but also led to the development of hierarchical multi-scale models that assess the impact of molecular-level mechanisms on mechanical properties of the material at the

macroscale [73].

When it comes to chemical reactive systems, ML models capture the quantitative structure-property relationships by learning features from the simulation data that are associated with thermodynamic properties of the system. ML regression models *viz.* LASSO, random forest, gradient tree boosting and support vector regression, trained on fingerprints extracted from MD simulations are shown to predict solvation free energy and partition coefficients that have been experimentally validated [74]. Evidence of using deep learning architectures like convolution neural networks (CNNs) to extract spatiotemporal features from the MD simulations of interfacial water densities are used to predict hydration free energies that characterizes hydrophobicity that drives protein folding mechanisms [75]. 3D molecular features constructed from the atomic partial charges, average number of water contact points, the number of hydrogen bonds, their shapes and sizes, as extracted from MD data, have been used to train CNNs to predict free energies of the drug-protein binding [76]. Chemical behavior in fuel combustion has been modeled by training a ML model on chemical reactivity data from MD simulation to predict the component fractions [77]. Predictive ML models trained on expensive electron structure calculations, were seen to efficiently reproduce the infrared spectra characterizing peptide dynamics [78]. Aside from regression models, ML classifiers *viz.* Linear discriminant analysis (LDA), support vector classifier (SVC) are trained on the MD data of the decomposition of dioxetane, to extract features from the nuclear coordinates that correspond to either successful or frustrated dissociations [79], the time scales of which impact the chemexcitation yield. More complex reactive systems, for instance catalytic biomass conversion, involve water that dissolves biomass, and polar aprotic cosolvents to accelerate reaction rates. Maximizing the rate and yield of biomass conversion in such systems depends on an optimal solvent:cosolvent ratio, the high throughput screening of which is facilitated by training a predictive ML surrogate that is computationally efficient at generating descriptors from simulation data, against which the reaction

rates are regressed [80].

1.3 Motivation

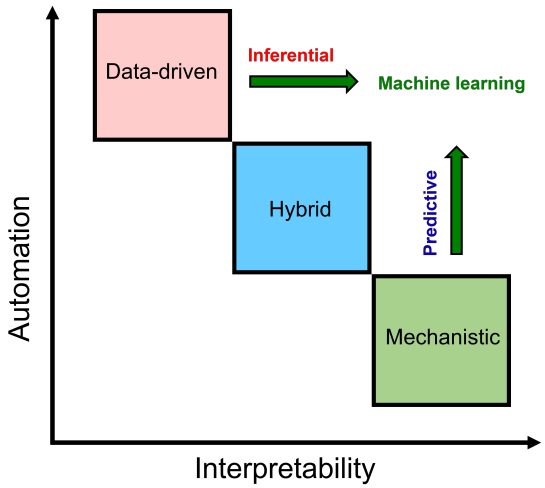


Figure 1.2: Machine learning has the potential to bridge the modeling tradeoff between automation and interpretability.

The work in this thesis is motivated by the challenges of using machine learning to (i) increase the interpretability of system inferential data-driven models, and (ii) to increase the automation capacity of prediction-based mechanistic models, for reactive chemical systems. The potential of machine learning to bridge the automation-interpretability tradeoff as shown in Figure 1.2, opens up avenues for a wide variety of modeling paradigms.

Interpretability challenges in end-to-end inferential models from experimental data. Based on the classification of chemical reactive systems (Figure 1.1), it can be seen that data-driven models are widely used to model complex reactive systems, where the prior knowledge of its constituent species and the underlying reaction pathways are obscure. Literature pertaining to the chronological identification of species, reaction pathways among them and then kinetic parameter estimation via data-driven system inferential models has been outlined in Section 1.1. However, an end-to-end machine learning framework for the same, is lacking.

The autonomy of data-driven models poses a challenge to their ability in explaining physical systems. Hybrid models that incorporate physical laws are seen to limit the autonomy of data-driven models that are constrained to be physically interpretable. For instance, the Beer Lambert's law in spectral curve resolution for species identification [5], [35] the law of mass action and Arrhenius law of temperature dependence [61] for simultaneous inference of reaction pathways by parameter estimation from temporal concentration data. Sometimes it may be difficult to obtain accurate measurements of non-equilibrium temporal concentration data of species [81]. Although, the projections of spectroscopic data onto the temporal mode of data collection gains interpretability as concentration (Beer's law), the simultaneous inference of reaction pathways and kinetic parameters in this case may not be reliable, owing to process noise [60]. This motivates the development of injecting interpretability into the data-driven approach for reaction inference, before using the reaction network structure as an additional constraint to guide kinetic parameter estimation from the aforementioned noisy concentration data. Preserving interpretability via physically meaningful constraints in going from the identification of species and reaction pathways, to finally kinetic parameter estimation, by solely relying on spectroscopic data of a reactive system, has not been established thus far.

Data obtained from different spectral analyzers contain multi-view information of the reactive chemical system. For instance, process Raman spectroscopy offers information pertaining to the molecular backbone as well as symmetrical non-polar groups, IR spectroscopy yields information pertaining to hydrogen bonding and asymmetric polar groups, and NMR spectrometry provides highly resolved information detailing specific proton environments [82]. Current spectroscopic curve resolution approaches, fall short of combining complementary information from multiple spectral sensors through data fusion architectures that retain semantic meaning of the sensor interrelationships by incorporating network regularization constraints [83], [84], to extract meaningful features for species identification.

Computational costs of predictive models from mechanistic simulations. Also, when it comes to first-principles mechanistic models of systems where both the species and conversion pathways are known *a priori*, literature points to the reduction of computational efforts by training predictive machine learning models as surrogates, as outlined in Section 1.2. The use of ML-derived insights from MD simulations to predict the mechanism, rate and yield of chemical systems as functions of its thermodynamic properties has been recognized as one of the six grand challenges of the 21st century [85]. However, these ML models are cost-effective only if the cost of training them and that of generating the simulation data to train on, is lesser than the cost of performing the first-principles calculations themselves. The training costs generally involve sampling calculations of the simulations to obtain *labels* against which the mapping of the features extracted from the mechanistic simulations, is learned [74],[75], [76], [79], [80]. This has motivated the development of self-supervised models, to extract features, the probability distribution across which is used as a basis for label assignment, when training ML models on the AIMD data of the transglycosylation of cellobiose to predict whether or not the solvent molecules reorganize significantly.

1.4 Thesis Objectives

Enhancing the interpretability of inferential machine learning models for reactive chemical systems can be facilitated by not only incorporating physical laws as constraints but also by jointly analyzing complementary molecular-level information from multiple spectroscopic sensors to limit solution ambiguity. This thesis leverages semantic meaning-based and structure-preserving data fusion architectures to extract meaningful spectroscopic features that act as a basis for species identification, reaction pathway identification via causal inference among the features, followed by estimating kinetics, without reliance on *a priori* knowledege of a reactive chemical system. A schematic representation of the end-to-end machine learning framework

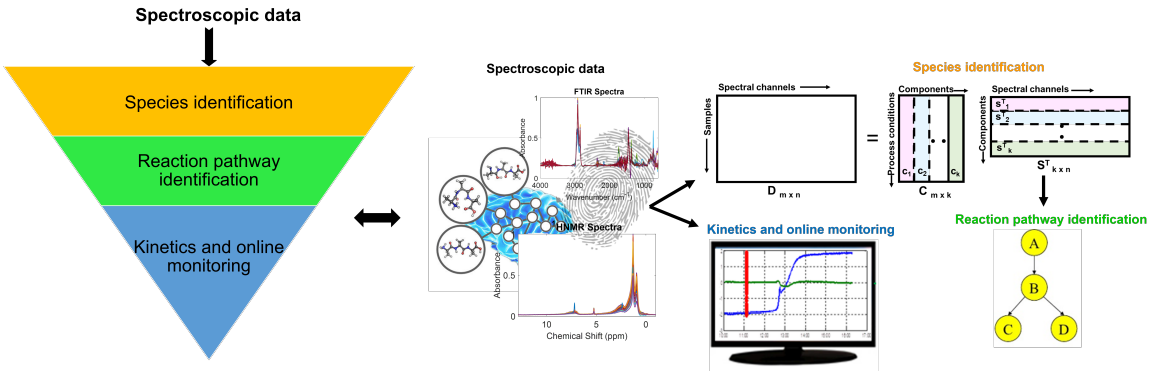


Figure 1.3: Schematic of the end-to-end inferential machine learning framework to model complex reactive systems in the absence of prior knowledge of its species or reaction pathways.

that has been developed for the same has been indicated in Figure 1.3. Following are the objectives that have been realized:

- Quantitative parameters computed from the FTIR multivariate curve resolution (MCR) of the thermal cracking of Athabasca bitumen have been used to propose plausible reaction pathways without prior knowledge of the system. Two solution engines *viz.* particle swarm optimization (PSO) and alternating least squares (ALS) have been devloped to solve the MCR objective at local operating temperatures and globally across all operating temperatures.

- MCR has been developed to jointly parse complementary information from multiple spectroscopic sensors by way of joint non-negative matrix factorization, while extracting features for species identification. Probabilistic causal structure inference among these features has then been used to hypothesize reaction pathways. This has been demonstrated with application to the partial upgrading of Cold Lake bitumen.

- Joint non-negative tensor factorization has been developed as a structure-preserving higher order analogue of the data fusion architecture in joint non-negative matrix factorization, before using causal structure inference among the extracted features to infer reaction pathways. The higher order latent factor decompo-

sition is shown to limit solution ambiguities even in the absence of additional redundancy penalizing constraints.

- A framework for using online spectroscopic data to monitor reaction dynamics with changing operating temperatures during the processing of complex feeds, has been developed. Hidden semi-Markov models are shown to facilitate dynamic mode identification of online spectra. The spectral mode segments are then interpreted as reaction mechanisms by using the earlier developed scheme of latent factor decomposition and causal structure inference. The dynamics of mode transitions are then demonstrated to reflect reaction mechanism dynamics in realtime.

- Projecting spectral data onto the temporal mode of data collection using the structure preserving tensor factorization is physically interpreted as the concentrations corresponding to the extracted spectral features. The reaction network structure causally inferred among the spectral features is used to additionally constrain a physical neural ODE architecture that has been demonstrated to fit kinetic models to the temporal concentrations.

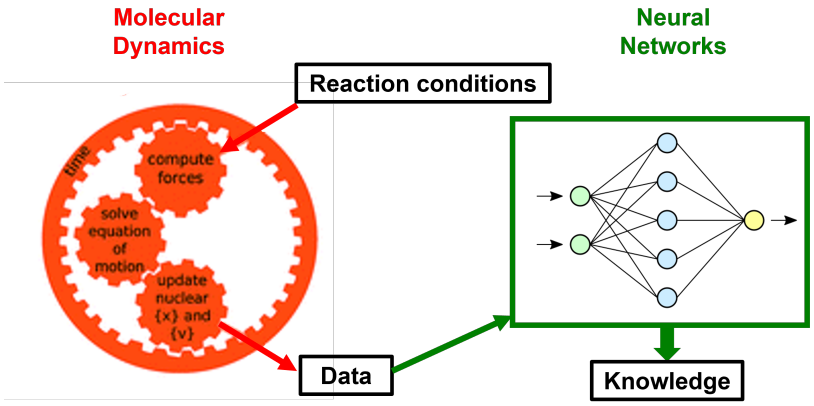


Figure 1.4: Predictive machine learning to limit computational costs of mechanistic simulations for reactive systems by extracting self-supervised insights

When it comes to reducing the computational costs of developing predictive models from *ab initio* mechanistic simulations, the cost of assigning labels in the train-

ing phase by expensive sampling calculations can be mitigated by the use of self-supervised neural network architectures to extract features from the MD trajectories. Also, the use of simple quadratic distance-based classifiers is seen to surmount the training costs of neural network classifiers [86]. A schematic of training ML models to extract actionable insights from mechanistic simulations has been shown in Figure 1.4. This thesis leverages insights from the probability distributions across the extracted features to inform the discrimination of whether solvent molecules reorganize significantly in reactive systems to realize the following objective:

- 3D CNN autoencoder has been implemented to extract features from the reactant and product configurations of the AIMD simulation data for the transglycosylation reaction of cellobiose. Probability distributions across the difference in features between the reactant and product configurations are used to assess whether or not the reorganization of solvent molecules is significant. Generalization of these insights enables associating features extracted from the reactant configurations of other systems to predict whether solvent molecules reorganize significantly in the product profiles, and if they are found not to a decision can be made to eliminate those molecules when running mechanistic simulations of the product configurations, saving computational effort.

1.5 Thesis Structure

This thesis is written in a paper-based format. Chapters 2 to 6 concern the development of interpretable machine learning models for the data-driven species and reaction pathway identification, followed by online monitoring and kinetic parameter estimation of reactive chemical systems using spectroscopic data, without reliance on prior knowledge of the systems. The hypothesized reaction pathways have been validated by domain knowledge. Chapter 7 deals with developing predictive machine learning models from AIMD simulations to reduce computational costs with application to

simulating reactions in cellobiose and fructose systems.

Chapter 2 investigates the effects of the solution engines (particle swarm optimization, alternating least squares), initialization techniques and heuristics in determining the number of latent components in the implementation of the self-modeling multivariate curve resolution algorithm to deconvolve FTIR data. Quantitative metrics deduced from the intensities of the absorption bands in the deconvolved spectral features of the latent components have been used to infer plausible reaction pathways in the thermal cracking of Athabasca bitumen.

Chapter 3 focuses on a data fusion framework that incorporates complementary spectral relationships from both the FTIR and ^{1}H-NMR spectroscopic data while penalizing redundancies among them via network regularization constraints. The algorithm has been developed to handle process data artefacts by imputing missing values, using a rotationally invariant norm for robustness to outliers and noise, and enforcing non-negativity constraints to ensure interpretation of the latent features in compliance with Beer's law. A metric-agnostic approach of using Bayesian structure learning for causal inference among the latent spectral features is demonstrated to hypothesize reaction pathways for the thermal cracking of Cold Lake bitumen.

Chapter 4 outlines the implementation of a structure-preserving data fusion framework as a higher order analogue of that demonstrated in Chapter 3. The FTIR and ^{1}H-NMR spectroscopic data collected across the process modes of temperature and residence time are decomposed via joint non-negative tensor factorization. The data projections onto the spectral channels specific to each sensor are interpreted as spectral features of the latent components, while the projections onto the temperature and time modes are interpreted as their corresponding concentrations. Reaction networks are hypothesized by structure learning among the latent spectral features. The higher order decomposition has limited solution ambiguity owing to its structure-preserving nature, obviating the need for network regularization constraints. However, a scalable method for parallelizing tensor factorization when using a robust norm to handle

outlier and noise has been proposed via grid tensor factorization.

Chapter 5 focuses on using tensor factorization and causal structure inference for species identification and reaction mechanism hypothesis, respectively, in the backend to facilitate the interpretation of the dynamically identified modes from the Hidden semi-Markov models (HSMM). HSMMs that explicitly model the duration distributions of the modes and their transition dynamics, using FTIR spectroscopic data collected across varying temperature conditions, has been developed for the realtime monitoring of reaction dynamics.

Chapter 6 utilizes the spectral projections onto the temporal mode of data collection obtained from non-negative tensor factorization of FTIR data, as concentrations to develop a kinetic model. A chemical reaction neural ODE that is structurally constrained by the law of mass action, the Arrhenius law of temperature dependence, and the adjacency matrix derived from the Bayesian network structure that has been causally inferred from the spectral features corresponding to the concentration profiles of species, has been used to learn kinetic models. This framework has been demonstrated on synthetically generated spectroscopic data from a known reaction template in the database, as it would be challengin to validate predictions in the absence of an exhaustive ground truth kinetic model for complex systems like bitumen or biomass.

Chapter 7 presents a self-supervised framework of feature extraction from the *ab initio* molecular dynamics simulations of the reactant and product configurations for the transglycosylation of cellobiose, using a 3D convolutional neural network autoencoder. The probability distribution across the difference between the features of the reactant and product configurations have been used assess whether or not solvent reorganization is significant. The proposed framework seeks to reduce the computational cost by eliminating solvent molecules when simulating the product configurations for those chemical systems, the features from the reactant configuration of which are most similar to the encoded features of the reactant cellobiose systems where the solvent molecules are found not to significantly reorganize in the product profiles.

Chapter 8 summarizes the key findings of the thesis and highlights avenues for future work.

Chapter 2

A data-driven approach to generate pseudo-reaction sequences for the thermal conversion of Athabasca bitumen

Abstract

This work focuses on the application of self-modeling multivariate curve resolution (SMCR) methods on the Fourier transform infrared (FTIR) spectra of the liquid products obtained from the thermal cracking of Athabasca bitumen in the temperature range of 300–420°C and reaction times ranging from 15 min to 27h. The objective was to develop a reaction pathway for the thermal cracking process from the SMCR methods and to identify key elements of the reaction chemistry that also affected physical properties like viscosity. An important aspect of this work was that minimum external chemical knowledge was used for the chemometric techniques. The SMCR method employed in our study was applied on both temperature-specific and augmented datasets considering all temperatures together to extract resolved concentration and spectral profiles using the alternating 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 vari-

This chapter has been published as: K. Sivaramakrishnan[‡], A. Puliyanda[‡], A. de Klerk, V. Prasad. A data-driven approach to generate pseudo-reaction sequences for the thermal conversion of Athabasca bitumen. *React. Chem. Eng.* **2021**, *6*, 3, 505-537.([‡] Equal contribution)

ance. The thermal conversion of Athabasca bitumen was shown to observe a series reaction sequence with methyl transfer dominant at lower temperatures and a greater extent of cracking at higher temperatures along with the formation of lighter products with a higher fraction of mono-substituted aromatics.

2.1 Introduction

Oil sands bitumen is a heavy residue feedstock of high density and viscosity, and this presents significant challenges in processing it to obtain hydrocarbon products. Two major issues are the difficulty with getting bitumen to flow at ground temperatures, and the high carbon to hydrogen ratio, which necessitates some form of upgrading before processing in a conventional refinery. Bitumen is customarily diluted with natural gas condensate/naphtha to improve the flow properties, but this is not ideal. An alternative that has been explored recently is the partial upgrading of bitumen, which aims to reduce the viscosity enough for the bitumen to flow easily without diluent. One of the techniques explored for partial upgrading is thermal conversion at relatively mild temperatures.[87] However, the chemistry behind the thermal conversion of oil sands 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 boiling point and solubility classification rather than individual chemical components due to the obvious difficulty in identifying the constituent species in bitumen.[88], [89] Though advances were made to identify the molecular structure and composition of the heavier components of bitumen like asphaltenes,[90] tracking changes in chemical structure during thermal treatment is a difficult task.

Given the difficulty of compositional analysis, there has been some research on using different process variables as indicators to track product composition. One way to achieve this was by setting up distributed monitoring networks to measure the process variables involved, but this was found to be expensive and inefficient.[91] However,

a central monitoring network that could measure the different process variables and eventually control the system by relying on data from the reaction progress would be a good approach moving forward.

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.[92],[93], [94],[95],[96], [97] 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), [98], [99], [100] proton nuclear magnetic resonance (^{1}H-NMR),[101] and electron spin resonance (ESR) [102], [103], [104] have been applied 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 noninvasive nature. [105], [106], [107] The inclusion of accessories like flow cells, quartz windows and immersion probes also facilitate faster characterization.[108] They also provide avenues for online monitoring of the system which is important if the goal is automation and control.[109] However, the challenge is that the data obtained is multi-dimensional and often represent overlapped spectra from a vast number of components.

Kinetic models provide an estimate of the probability of the occurrence of each constituent reaction through calculations of kinetic parameters like rate constant and activation energy. 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,[89] sample-to-sample variability in such

properties can also contribute to possible sources of error. [110], [111]

Chemometric techniques involving statistical approaches have been shown to be quite useful in tackling the challenges of higher dimensional data and overcoming the limitations of a kinetic model based on assumed lumping and relationships. [112], [113], [114], [115] 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. Chemometric techniques 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. The development of statistical models with limited reliance on prior knowledge is employed in the development of kinetic models for advancements in the control and monitoring aspects of reaction engineering in process systems.[116]

This work focuses on the use of chemometric techniques to identify species and reactions for the thermal conversion of Athabasca bitumen over the temperature range 300–420 °C based on FTIR spectra of the reaction products. Depending on the reaction times for which the bitumen sample was held at each temperature in a batch reactor, two regimes were considered 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.[117] Industrial visbreakers employ temperatures of 430–490°C, pressures in the range 0.3–2 MPa and the residence times in the order of seconds or minutes in a coil visbreaker (but much larger in a soaker-type visbreaker). [118], [119], [120] Typical conditions for industrial delayed coking are 480–510°C and higher residence times of upto 24 hours under low pressures of 0.6 MPa.[120]

The type of chemometric analysis conducted on the FTIR data was self-modeling multivariate curve resolution (referred to as SMCR or MCR) employing 2 different algorithms for extracting the final concentration and spectral profiles of a reduced num-

ber of components from the original bitumen. 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 spectrum to aid with the tracking of chemical changes with time and to get an insight into the specific kinds of reactions occurring during thermal conversion.

The key contribution of this study was to combine basic chemical knowledge of the system and the results of the chemometric techniques involving local and global models with both algorithms to propose a plausible reaction sequence for the thermal conversion of Athabasca bitumen. The consistency of the results from the global model was verified with that of the local models so as to verify the credibility of the developed reaction pathways. The results were compared with that obtained for Cold Lake bitumen [25] in terms of differences in reaction chemistry with the variation in chemical composition and possible relation to physical properties like viscosity. It should be noted that Athabasca bitumen resembles Cold Lake bitumen in overall chemical composition, but has a lower saturate content, higher asphaltene content and higher viscosity; and this study highlights the fact that different strategies would need to be employed for its partial upgrading.[110],[121], [122]

2.2 Origin of data

In this work, Athabasca bitumen was subjected to temperatures of 300 – 420°C 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.

Table 2.1: Experimental conditions of thermally processed samples used for data analysis.

Temperature	**Number of samples**	**Reaction times at each temperature (min)**
Feed	1	-
300°C	2	360, 1080
350°C	6	30, 60, 180, 240, 360, 480
380°C	6	120, 240, 360, 480, 1320, 1620
400°C	16	15, 30, 45, 60, 75, 90, 105, 120, 135, 150, 180, 210, 240, 360, 1170, 1440
420°C	4	360, 420, 480, 660

For information on the materials used, equipment and procedure and the analyses done on the feed and thermally converted products, the reader is referred to previous work conducted at 400°C on Athabasca bitumen. [111] The only difference from that work is that 4 additional temperatures between 300°C and 420°C have been explored as well and the corresponding reaction times are given in Table 2.1.

2.2.1 FTIR data

The dataset was obtained by analysis of 35 samples of liquid products that were collected from thermal conversion at 5 different temperatures, 300°C, 350°C, 380°C, 400°C and 420°C, and 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 2.1.

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 attenuated total reflectance (ATR) attachment employed and appeared as random peaks with arbitrarily high values of transmittance. The transmittance data was converted to absorbance units by

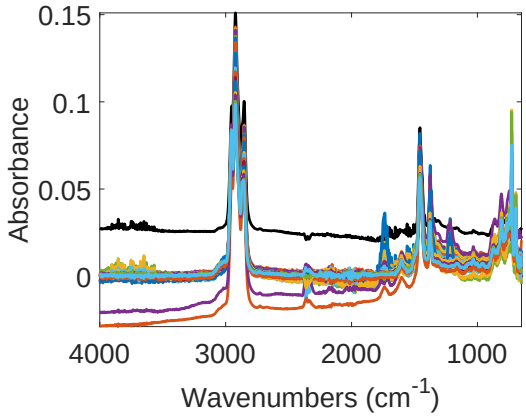


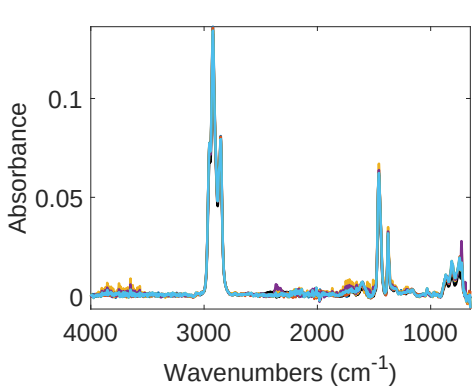
Figure 2.1: Raw FTIR absorbance spectra of 35 liquid products from thermal conversion of Athabasca bitumen at five different temperatures and reaction times before pre-processing.

their logarithmic relation, which is also related to Beer-Lambert's law. It is important to note that although the path length was the same for all the wavenumbers in one spectrum, it might vary between spectra. This raw absorbance data is shown in Figure 2.1.

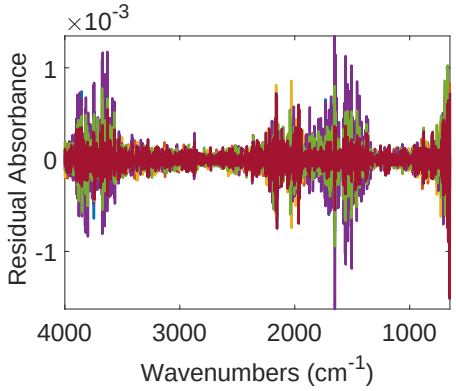
As can be seen from Figure 2.1, the region from 4000 – 3200 cm^{-1} mostly exhibits baseline intensities with no peaks worthy of chemical interpretation though the O-H and N-H groups present in bitumen (both free and hydrogen bonded) absorb in that region.

2.3 Methods and parameters used

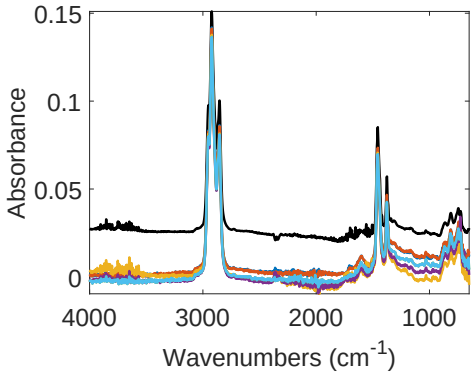
The theory behind the chemometric techniques used in this work, i.e. multivariate curve resolution employing 2 convergence algorithms is provided in this section. In addition, the software tools and the related functions along with the respective important parameters adopted in each step are highlighted. The reasoning behind the steps implemented during the mathematical analysis is also mentioned wherever appropriate.



(a) Baseline corrected and smoothed data



(b) residual after smoothing



(c) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 350°C.

Figure 2.2: Pre-processing the FTIR data.

2.3.1 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 and curve resolution. 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 2.1. This helped to identify whether there was a

difference in the reaction chemistry at low and high temperatures.

MATLAB R2017b (9.3.0) was used to carry out all the chemometric analysis in this study. The '*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 units, which is the default window size. Adjacent windows are located at a distance of 200 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 uses the well-known Savitzky-Golay (SG) filter, in which a least-squares two-degree polynomial is used for de-noising the spectra over every 5 samples (window size).[123] It is important to note that the normal SG filter requires 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. [124] 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 and curve deconvolution. This also made it irrelevant that there were differences in path length between the spectra due to the use of ATR. 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. [125], [126], [127] 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. [124]

All of these pre-processing steps including baseline correction, smoothing and stan-

dardization were applied to the FTIR data before proceeding with curve resolution. 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°C is shown in Figure 2.2. Similar results were obtained for other temperatures as well and is supplied in Appendix A.

The advantage of pre-processing the data is clearly visible in Figure 2.2a where the spectral features are more distinctly seen than in the raw data (Figure 22.2c). The stretching vibrations of the sp^3 hybridized C-H methylene groups that can belong to either the alkyl side chains or naphthenic rings occur at 2850 cm^{-1}, 2920 cm^{-1} while those of the 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 the 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 the ester-type and the 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 the 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 smaller number of pseudo-components that were representative of the change in properties of bitumen with time and develop a reaction sequence based on these results.

2.3.2 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, that can be decomposed in a bilinear or tri-linear fashion using SMCR or PARAFAC, respectively. [128], [129], [130], [131] SMCR is essentially a soft-modelling technique that utilizes factor analytical decomposition and invokes physically meaningful laws like Beer Lambert's law. [131] Beer's law is also a bilinear model that relates the absorbance of a light-irradiated species to its concentration and path length. Due to the pre-processing, only relationship within each spectrum remained, while that between spectra was lost due to normalization.

The curve deconvolution process involved three major steps: (i) data matrix decomposition as a means of exploratory analysis based on singular value decomposition 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) [132] by means of singular value decomposition (SVD); (ii) Elbergali's [133] recommendation based on the maximum ratio of derivatives of the second and third order of the Malinowski's [33] 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. Rank relaxation was carried out to ensure meaningful chemical information was not lost to noise in the approach where the indicator function is used to statistically determine the number of components. Fixed-size moving window evolving factor analysis (FSMW-EFA),[28] which is an iterative method, was used to obtain the initial estimates of concentration profiles in the MCR method since it was suggested as a better approach to distinguish concentration regions of a component from the noise as compared to the forward/backward EFA employed by Tefera et al. [25] Other contemporary techniques having similar goals to EFA are rank estimation-based methods like generalized rank annihilation method (GRAM),[97] window factor analysis (WFA),[134] and sub-window factor analysis (SFA)[135] 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 and 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.[25] 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. [136]

A key difference in the approach used in this work as compared to that of Tefera et al. [25] 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. [137] 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. [138] 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 this work.

Data decomposition

Let D be the data matrix that is composed of the FTIR spectra of 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:

$$D = CS^T + E \tag{2.1}$$

where C is an $m \times l$ matrix that contains the concentration of l components, S is a $n \times l$ 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 2.1. The procedure for finding the number of components is given in the sub-section 2.3.2 titled 'Determination of chemical rank of the system' under this main heading.

Multiset structures

If the experiments were analyzed with more than one characterization technique, Equation 2.1 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 A.1 and Equation A.1 in the Appendix. 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.

Algorithms to solve SMCR

Either non-iterative or iterative algorithms are used to solve Equation 2.1. [139] Non-iterative methods like heuristic evolving latent projections (HELP), WFA, orthogonal projection analysis (OPA),[140], [141] and parallel vector analysis (PVA) [142] 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), [143] 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.[144] 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 the 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.[145] 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 A.5 and Equation A.6 in the Appendix). 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 (given in the sub-section 2.3.2 titled 'Obtaining initial estimates' in this heading).[146] 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 the 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 with a tolerance of 0.001 for the convergence.

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 the Results and Discussion section.

Determination of the 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.[147] 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. [148] In the case of bitumen, the composition is quite complex and the number of components constituting bitumen and further taking part in the 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 [149] 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) routine that is sometimes used for PCA where the

loading vector is normalized,[150] 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. [33]

$$R(l) = D - USV^T(l) \tag{2.2}$$

$$RE(E) = RSD(l) = \sqrt{\frac{\sum_{j=l+1}^{n} R_j^2}{m(n-1)}} \tag{2.3}$$

$$IND(l) = \frac{RE(l)}{(n-l)^2} \tag{2.4}$$

$$IE(l) = \sqrt{\frac{l}{n}} RE(l) \tag{2.5}$$

$$SD(l) = \log\left[IND(l)\right] - 2\log\left[IND(l-1)\right] + \log\left[IND(l-2)\right] \tag{2.6}$$

$$ROD(l) = \frac{IND(l-2) - IND(l-1)}{IND(l-1) - IND(l)} \tag{2.7}$$

In terms of the 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 2.3). [133] 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 2.4). 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 2.5). 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. [133] 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 [133] 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 2.6). Apart from these, the ratio of derivatives (ROD) criterion was discovered to be the best indicator for the rank determination (Equation 2.7). 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.[151]

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 2.2), followed by its standard deviation considering the residual matrix as a single column vector of elements (Equation 2.3). 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 Equations 2.2-2.7.

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 the 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 the Results and Discussion section under the sub-heading 'ALS-optimized C, S profiles and spectra-derived quantitative parameters'.

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.[131] 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 2.1), EFA was thought to be quite suitable for initial profile estimation. However, methods like simple-to-use self-modeling analysis (SIMPLISMA),[152] key set factor analysis (KSFA)[153] 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 [28] as the noise EVs are also constant due to the fixed size of each window on which SVD is applied. Tefera et al.[25] 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°C 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. [154] compared the use of PSO against the use of EFA to obtain the initial estimate of concentration profiles in their work on the 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 the subsection 2.3.2 titled 'PSO and its use as a hybrid technique in this study'.

Limitations: Ambiguities

The uniqueness of solutions from the SMCR-ALS method is crucial for its reliability. However, the MCR solution potentially suffers from 3 types of ambiguity: intensity, permutation and rotational.[155] 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 A.3 in the Appendix. If concentration is optimized first in the ALS scheme, it is normalized before every iteration to help reduce the 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 the intensity or rotational ambiguities in terms of further chemical interpretation. The most commonly present ambiguity in MCR results is the rotational ambiguity, which is specified in Equation A.4 in the Appendix. 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. [156]

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. [131] 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:

$$f_{i,\min} = \min \frac{||c_i s_i^T||}{||CS^T||} \tag{2.8}$$

$$f_{i,\max} = \max \frac{||c_i s_i^T||}{||CS^T||} \tag{2.9}$$

The lower the value of $f_{i,\max} - f_{i,\min}$, the lower the extent of the 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 checked for deviations. However, this kind of testing was out of scope for our work.

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 sys-

tem. Their other function is to suppress ambiguities that arise in the solutions. [157] 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. [158]

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 the calibration samples. This is implemented using the correlation constraint. [159] 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.

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, 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.[25] 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. [160], [161] 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 (previous section on 'Obtaining initial estimates').

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.[162] 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 $x_i = (x_{i1}, x_{i2}, \cdots, x_{iD})$ and its velocity in different dimensions is represented by $v_i = (v_{i1}, v_{i2}, \cdots, v_{iD})$. While performing the search, the value of the objective function is calculated wherever the particles move and is compared with the previous values. If the current value is the least or maximum (according to the nature of the objective function) among all the previous values in the path of that particle, then it is the personal best position of that particle ($p_i = (p_{i1}, p_{i2}, \cdots, p_{iD})$). Similarly, a global best position ($p_g = (p_{g1}, p_{g2}, \cdots, p_{gD})$) 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:

$$v_{id}^{\text{new}} = w.v_{id}^{\text{old}} + c_1 r_1 (p_{id} - x_{id}^{\text{old}}) + c_2 r_2 (p_{gd} - x_{id}^{\text{old}}) \tag{2.10}$$

$$x_{id}^{\text{new}} = x_{id}^{\text{old}} + \mu v_{id}^{\text{new}} \tag{2.11}$$

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 2.10 and Equation 2.11 are usually chosen from experience but Bansal et al. [163] reviewed 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 [164] did not include w to update the particle velocity and instead, the velocity was capped at a maximum value. As an improvement, an 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. [165], [166] 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 A.1 in the Appendix.

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 the Appendix. The default number of iterations

for the PSO used is 600.

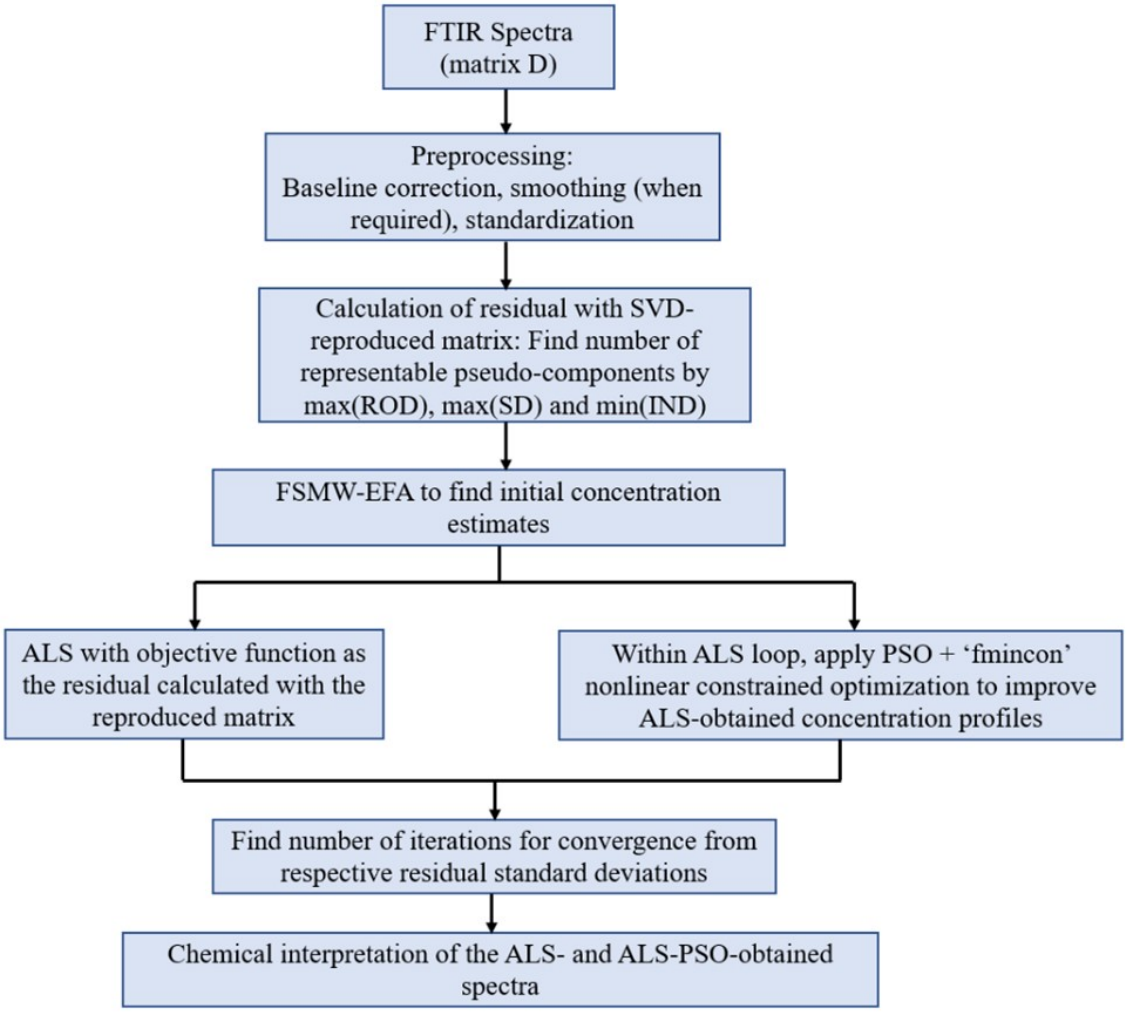


Figure 2.3: Sequence of steps followed in this work for chemometric analysis of the FTIR spectra through curve resolution.

Process flow followed in SMCR-ALS-PSO method

Fig. 2.3 provides the work flow followed in this work regarding the SMCR methods applied on the FTIR spectra of the liquid products from the thermal conversion of Athabasca bitumen. The reasoning for the use of the different steps employed is already highlighted in previous sections while describing each method. We would like to highlight that there are possible alternatives to two aspects of the solution procedure: instead of using the empirical approach of Malinowski to identify chemical

rank, a range of potential chemical ranks could be evaluated using the Akaike or Bayesian information criterion [167] as a basis; and global optimization approaches [168], [169] can be explored.

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 and (ii) final resolution of C and S profiles, both given in the first and third headings of the Results and Discussion section in this work. The performance of the SMCR techniques is validated with two measures: (i) the lack of fit (LOF) and (ii) 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 2.14. 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 2.15.

$$SSE = \sum r_i^2 \tag{2.12}$$

$$SSM = \sum D_i^2 \tag{2.13}$$

$$LOF = 100\sqrt{\frac{SSE}{SSM}} \tag{2.14}$$

$$R^2 = 100\left(1 - \frac{SSE}{SSM}\right) \tag{2.15}$$

where r_i is each element of the residual matrix calculated by subtracting the factor-reproduced matrix from the original data matrix, D.

Summary of methodological framework

We present a visual interpretation of the methods used in our analysis in Figure 2.4. The FTIR spectra are preprocessed using baseline correction, smoothing and normal-

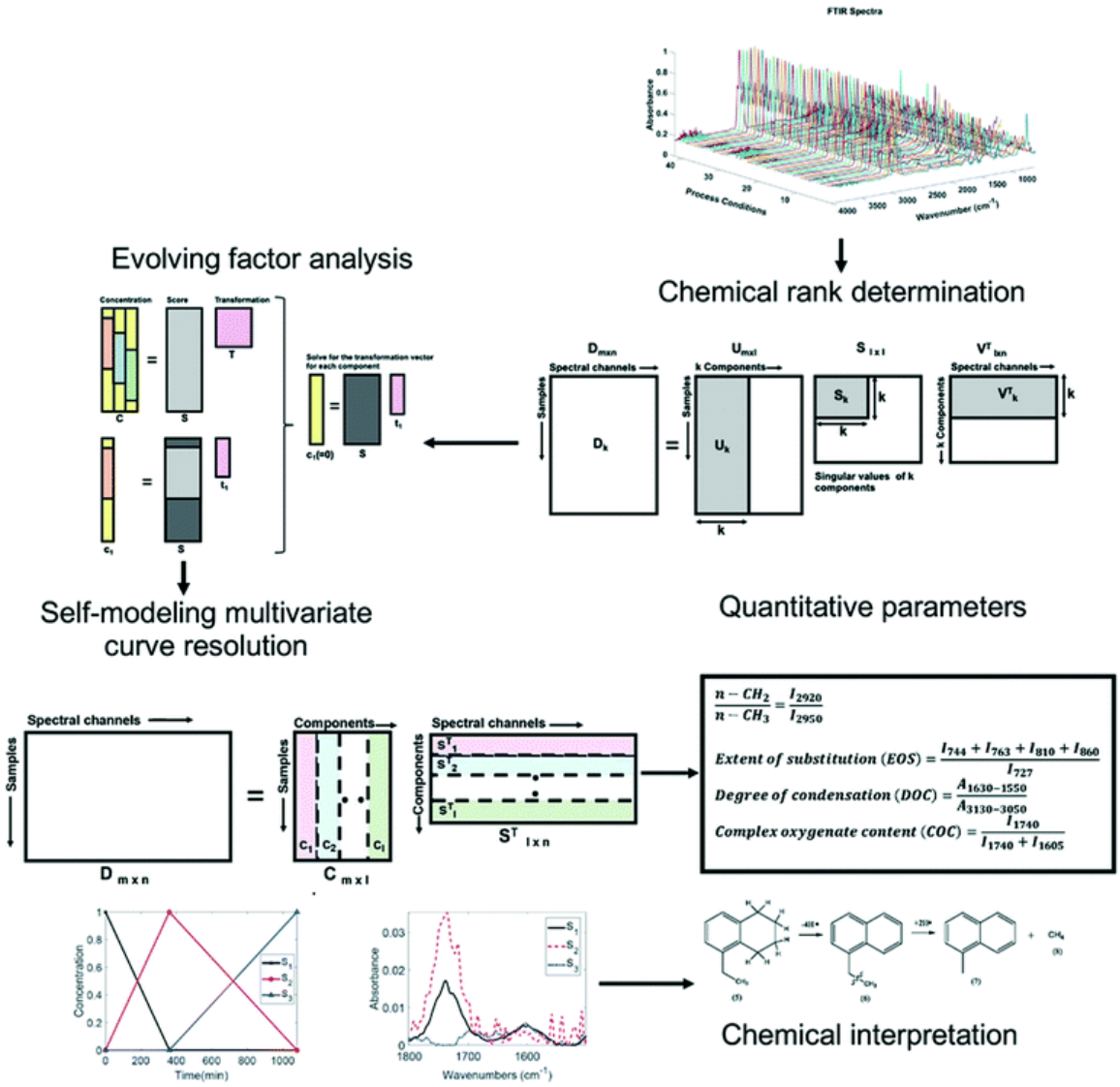


Figure 2.4: Visual representation of the transformation from spectra to pseudo-components and subsequent chemical interpretation.

ization. This is followed by chemical rank determination (identifying the number of pseudo-components), which is accomplished using singular value decomposition, calculating the indicator functions from Equations 2.4 -2.7 and identifying the number of components for which they achieve a maximum (ROD, Equation 2.7) or minimum (IND, Equation 2.4). Next, the initial estimates for the pseudo-component concentration profiles are obtained using fixed size moving window evolving factor analysis, which is an iterative unidirectional (i.e. using only forward SVD) method that decom-

poses a concentration matrix using a transformation and a corresponding score. Next, the SMCR-based concentration and spectroscopic profiles for the pseudo-components are obtained using these initial profiles as a starting point. This is accomplished using an alternating least squares approach, and then applying particle swarm optimization and '*fmincon*' (a constrained nonlinear optimizer) to get a better approach to a global optimum and remove ambiguities, and to apply non-negativity constraints, respectively. Once the spectral and concentration profiles are obtained, they are used along with quantitative performance indicators (described in a later section) for interpretation of the reactions occurring among the pseudo-components.

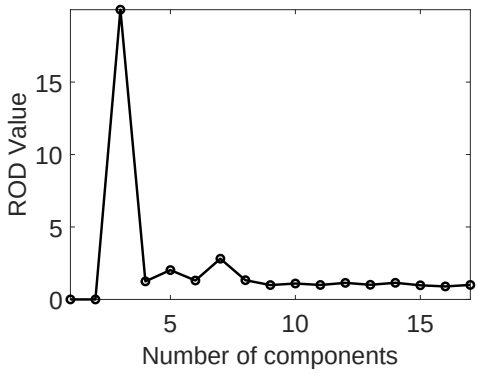
2.4 Results and Discussion

First, we provide results for local (i.e. valid for a specific temperature) models. These are obtained using both the ALS and the ALS–PSO-'*fmincon*' algorithms for SMCR. We then present the results from a global SMCR model for all temperatures considered together.

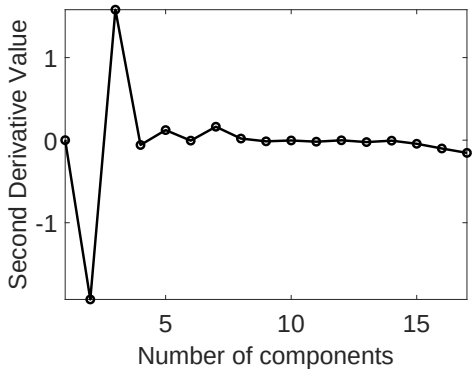
2.4.1 Rank of each sub-matrix

As mentioned earlier, ROD and SD were the indicators used to identify the number of chemical (pseudo-)components to be considered. Fig. 2.5 depicts the ROD, SD and the residual after performing SVD with the optimum number of components on the data obtained at 400°C. The plot of cumulative contribution of each component to the overall variance is also shown. These plots at other temperatures are provided in the Appendix.

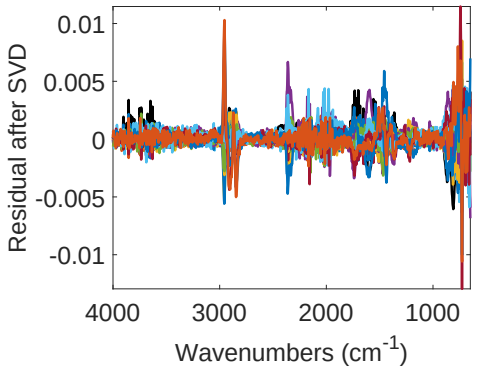
The x-axes in Figure 2.5a, 2.5b and 2.5c have 17 components/eigenvalues emerging from 16 experimentally obtained datapoints plus one for the feed. It can be seen that ROD has a maximum value of 20 when 3 components are used. This is consistent with the maximum for the second derivative of the IND function as shown in Figure 2.5b. Further support for a 3-component SMCR model comes from the residual



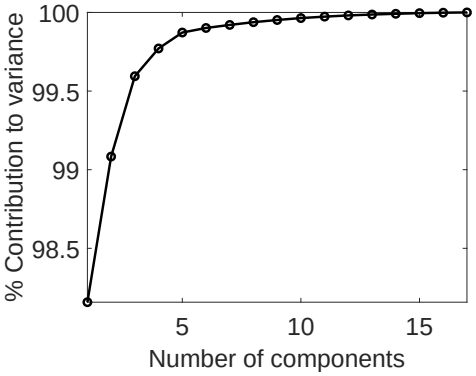
(a) ROD with respect to each component



(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.

Figure 2.5: Chemical rank determination for the data at 400°C

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 A.6a and Figure A.6b in the Appendix give the residual plots after conducting SVD with 2 and 4 components, respectively, for the 400°C 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 A.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 2.5d. The LOF (defined in the previous section), also decreased from 9.57 to 4.78 when 2 to 4 components were employed for the 400°C 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 A.2 in the Appendix.

Fig. A.7–A.10 in the Appendix show that ROD is maximum when 3 components are chosen for the SMCR models at the other temperatures of 300°C, 350°C, 380°C and 420°C as well.

It is to be noted that it would have been difficult to choose among 3, 4 or 5 components from the scree plot (representing the variance contribution with respect to number of components) (Fig.2.5d) alone since the inclusion of more than 2 components could explain ¿99% of the total variance. 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. Hence, it was decided to proceed with 3 pseudo-components for the rest of the curve resolution process, and this is consistent with the number of pseudo-components used to describe the thermal conversion of Cold Lake bitumen as well. [25] The first pseudo-component was chosen to be the one most likely to be representative of 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 2.4.3 titled 'ALS-optimized C, S profiles and spectra-derived quantitative parameters' in Results and discussion). 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.

2.4.2 Initial concentration estimates

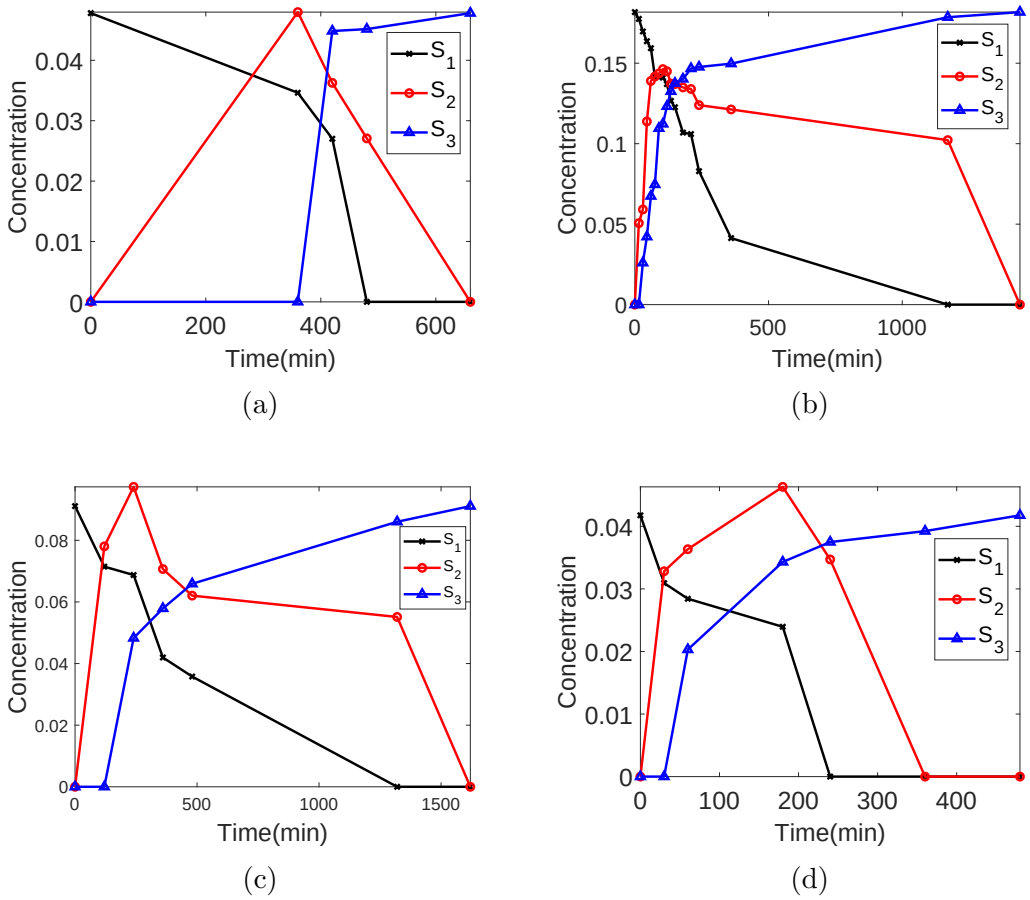


Figure 2.6: Plots of initial estimates of change in concentration of the three pseudo-components with process flow (reaction time in min) at the following temperatures: (a) 420 °C; (b) 400 °C; (c) 380 °C; (d) 350 °C

As detailed in the Methods and parameters used section, FSMW-EFA was used to obtain the initial estimates for the concentration profiles at each temperature. Figure 2.6 shows the initial concentration estimates obtained through this method at each temperature except 300 °C, which is shown in Figure A.11 in the Appendix. 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 2.6 and Figure A.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°C, 400°C, 380°C and 350°C 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) increased 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 (Fig. Figure 2.6a- 2.6d). Since there were only 3 points at 300°C, S_1 disappeared at the intermediate reaction time where only S_2 existed and S_3 remained at the final reaction time.

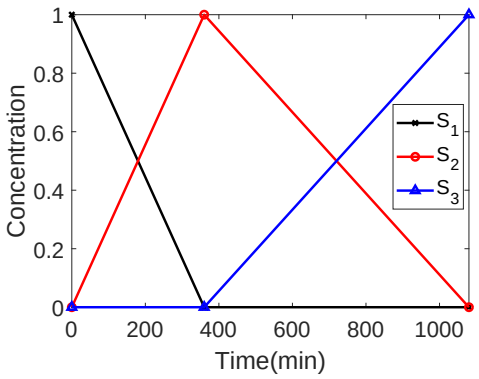
These observations suggest that S_2 is representative of an intermediate product and S_3 would be representative of the final product at all temperatures as indicated by the initial estimates. Any further interpretation based on these initial concentration profiles should be done with caution. We therefore moved onto 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.

2.4.3 ALS-optimized C, S profiles and spectra-derived quantitative parameters

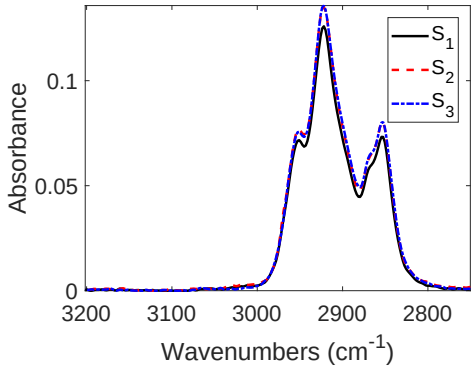
Results and analysis at 300 °C

Figure 2.7 shows the ALS-optimized concentration and spectral profiles for the pseudo-components when the reaction was conducted at 300°C. 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 A.12a of the Appendix.

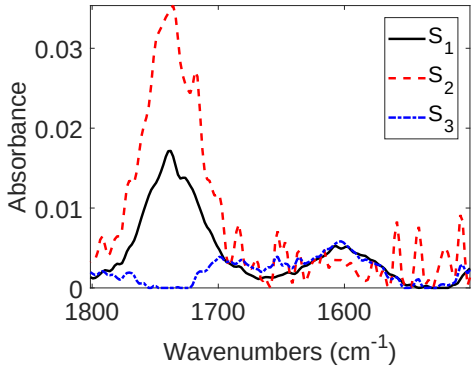
The pattern of the concentration profiles (Figure 2.7a) can be understood from the fact that the 300°C dataset consisted of only 3 data points including the feed. The initial estimates (Figure A.11 in the Appendix) that were not normalized, had a similar pattern and the conversion among the pseudo-components appeared to follow



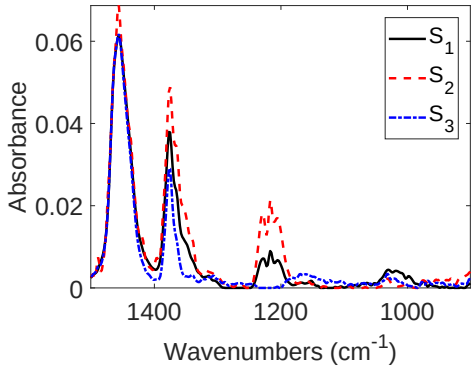
(a) Concentration vs. reaction time for the three pseudo-components



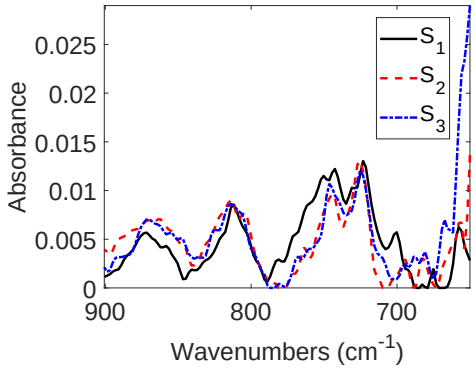
(b) Resolved spectra of the pseudo-components in the range 3200 – 2750 cm^{-1}



(c) Resolved spectra of the pseudo-components in the range 1800 – 1500 cm^{-1}



(d) Resolved spectra of the pseudo-components in the range 1500– 900 cm^{-1}



(e) Resolved spectra of the pseudo-components in the range 900 – 650 cm^{-1}

Figure 2.7: Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 300°C.

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 A.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 is representative of 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.

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:

$$n - \mathrm{CH_2}/n - \mathrm{CH_3} = \frac{I_{2920}}{I_{2950}} \tag{2.16}$$

$$\text{Overall extent of aromatic substitution (EOS)} = \frac{I_{744} + I_{763} + I_{810} + I_{860}}{I_{727}} \tag{2.17}$$

$$\text{Degree of condensation (DOC)} = \frac{A_{1630-1550}}{A_{3130-3050}} \tag{2.18}$$

$$\text{Complex oxygenate content (COC)} = \frac{I_{1740}}{I_{1740} + I_{1605}} \tag{2.19}$$

where I is the intensity at the respective wavenumber shown in subscript and A is the area under the regions in the wavenumber range specified as the subscript.

The $n - \mathrm{CH_2}/n - \mathrm{CH_3}$ 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 2.16) can also indicate the presence of

non-aromatic cyclic rings. [170] DOC was incorporated in a modified form from the work by Tefera et al.[25] 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 2.18). 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. [171] 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.

To mitigate these shortcomings, EOS was introduced in this work (Equation 2.17) 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 2.19). 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. [172] 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. [87] 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.

Table 2.2: Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 300°C.

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

The change in the above-mentioned quantitative parameters across S_1, S_2 and S_3 is summarized in Table 2.2. It can be seen from Table 2.2 that thermal conversion did not alter the ratio of the methylene groups to the 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 the cyclic moieties nor was there much formation of additional cycloalkanes due to thermal cracking at 300°C. 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 2.7b.

Cronauer et al. [173] suggested that hydrogen transfer from potential donors like naphthene-aromatics or benzylic carbon centres to multinuclear aromatics (MNA) would be difficult at temperatures lower than 300°C. At the same time, there have been studies that reported the occurrence of hydrogen transfer to MNAs at 150°C. [174], [175]

Tefera et al.[25] observed that the results and chemical interpretation for the SMCR-ALS resolved spectra for pyrolysis of Cold Lake bitumen at 300°C were similar to that at 150°C. nCH2/nCH3 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 2.2). 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°C, despite contrasting views in the literature as highlighted in the previous paragraph.

At 150°C, 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. [176] For complex mixtures, the change in viscosity and density cannot necessarily be directly related to the molecular weight when the composition varies. 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°C despite spectral features in the ALS-resolved profiles for Cold Lake bitumen remaining similar to those at 150°C. Though boiling point distributions for the products are not known, the density of most products obtained at 300°C 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. [111] as well.

Shifting our attention back to Athabasca bitumen in our work, the viscosity of the pyrolyzed products at 300°C 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.[177] 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 2.7d) 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. [177] This suggests that some methyl transfer may have occurred to an extent that led to an increase in the -CH2 content and was reflected in the mild increase of nCH2/nCH3

(Table 2.2). In fact, the contribution of methyl transfer to free radical reactions was recently demonstrated. [178] This process is illustrated in Figure 2.8.

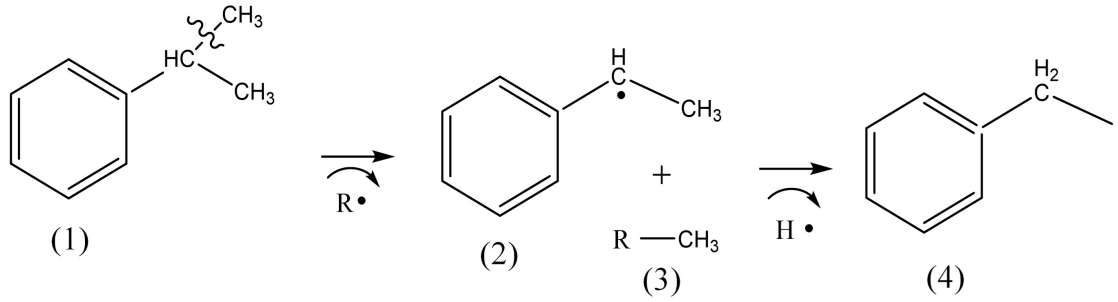


Figure 2.8: Methyl transfer from an isopropyl group attached to an aromatic (1) followed by hydrogen abstraction from the matrix leading to increased CH2 content (compound (4)).

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 2.2) 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 2.2) 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.[25] The increase in EOS for Cold Lake spectra could have been a result of the increase in DOC due to intramolecular ring closure reactions as suggested by Tefera et al. [25] On the other hand, the slight decrease in EOS despite the near-stable DOC for Athabasca was thought-provoking since both dealt with the aromatic part of the product. 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 2.7e shows that there was no significant change in the intensity at 727 cm^{-1} that indicated no additional formation of mono-substituted aromatics in S_3.

On the whole, considering the minimal changes in DOC and nCH2/nCH3 parameters and the little to no gas production, it can be concluded that the extent of cracking at 300 ^{o}C was indeed low. This could also be a reason for the reduced effect on vis-

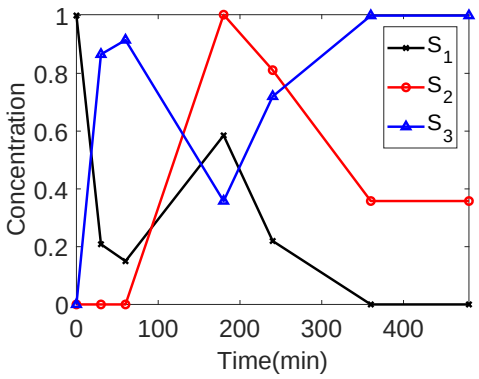
cosity at 300°C (which stayed nearly constant at the feed viscosity of 36 Pa.s even after 1080 min of reaction time). Selucky et al. [110] reported that the amount of saturates in Cold Lake bitumen is more than in Athabasca bitumen, making it more susceptible to cracking at lower temperature. This is probably another reason for the chemical changes being more significant in Cold Lake bitumen than in Athabasca bitumen when subjected to pyrolysis at 300°C.

Results and analysis at 350°C

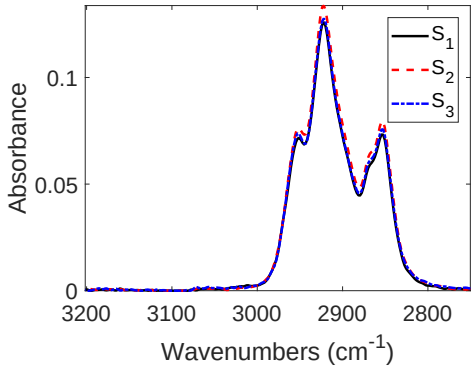
The ALS-resolved concentration and spectral profiles for the FTIR dataset at 350°C thermal conversion is given in Figure 2.9. The residual plot when the ALS-reproduced profiles are subtracted from the original matrix (D) is given in Figure A.12b in the Appendix.

The concentration profiles of the pseudo-components (Figure 2.9a) suggest that the reaction pathway is not as straightforward as was the case with 300°C. 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. In contrast to the initial concentration estimates (Figure 2.6d) 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 representative of the feed and S_3 as representative of the final product was considered as a fair assumption. The spectra-derived quantitative parameters for the pseudo-components at 350°C dataset are compiled in Table 2.3.

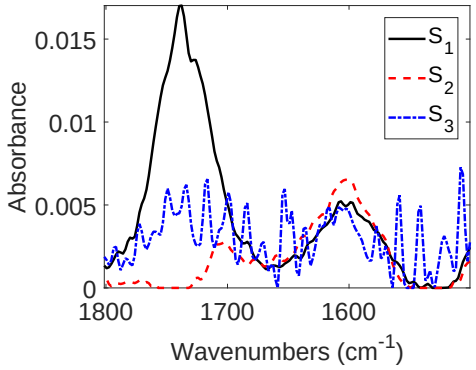
It was interesting to note that the ALS-resolved concentration profiles for that of Athabasca bitumen (Figure 2.9a) were similar to that of Cold Lake at 340°C



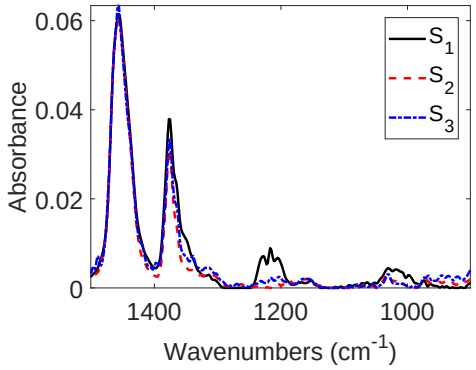
(a) Concentration vs. reaction time for the three pseudo-components



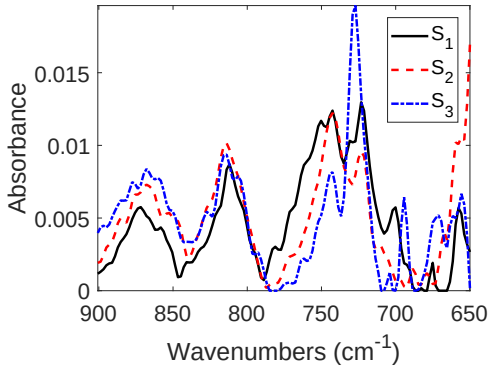
(b) Resolved spectra of the pseudo-components in the range 3200 – 2750 cm^{-1}



(c) Resolved spectra of the pseudo-components in the range 1800 – 1500 cm^{-1}



(d) Resolved spectra of the pseudo-components in the range 1500– 900 cm^{-1}



(e) Resolved spectra of the pseudo-components in the range 900 – 650 cm^{-1}

Figure 2.9: Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 350°C.

Table 2.3: Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 350°C.

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

[117] where the third pseudo-component started rising from the start when the first component began to decrease and both the second and the 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°C 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°C for Athabasca bitumen was 3-4 %wt. after 480 min and this was more than that produced at 300°C (previous section). A small but continuous decrease in nCH2/nCH3 (Table 2.3) possibly meant that there was a reduction of the length of aliphatic chains attached to the cyclic structures during cracking. At the same time, it could also mean a decrease in -CH2 content by the conversion of naphthenes to aromatics through hydrogen-disproportionation. However, the extents of these changes were 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 2.9b. The release of some gas indicated a certain extent of cracking.

In comparison, ALS-resolution of the spectra for Cold Lake bitumen at 340°C (ref.[25]) showed an increase in nCH2/nCH3 similar to the trends at lower temperatures, which was attributed to the presence of hydrogen transfer to multinuclear aromatics.

In addition, the C-H deformation band at 1380cm^{-1} was split for S_1 and S_2 but not for S_3 (Figure 2.9d). This potentially indicated that carbons having more than

one -CH3 attached to them initially have only one methyl group at longer reaction times. This could occur through methyl transfer, as was indicated at 300°C, 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°C as determined from GC-FID.[179] This was also supported by the work of Jha et al. [180] where they found that methane was the major product in the gaseous phase when Athabasca bitumen was thermally reacted at 300°C. On the other hand, if methyl transfer alone took place, it would have led to an increase in the CH2 content as noticed in the case of Cold Lake bitumen at 340°C and 360°C, while the opposite was observed with Athabasca bitumen with a decrease in the nCH2/nCH3 parameter at 350°C (Table 2.3). [25] On the whole, it can be said that methane formation happened to a larger degree than methyl transfer at 350°C 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 2.3) while the opposite was seen to happen with Cold Lake bitumen at 340°C, where it increased gradually from 0.18 to 0.39. [115] Despite this increase, they suggested that ring closure reactions through intra-aromatic coupling were suppressed. [181] 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°C decreased sharply from 3.14 to 1.30 (S_1 to S_3) while that of Cold Lake bitumen at 340°C 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 727cm^{-1} for S_3 as compared to the feed and S_2 (Figure 2.9e). 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.

Considering the variation in all the parameters at 350°C for Athabasca bitumen, a plausible reaction sequence is illustrated in Figure 2.10.

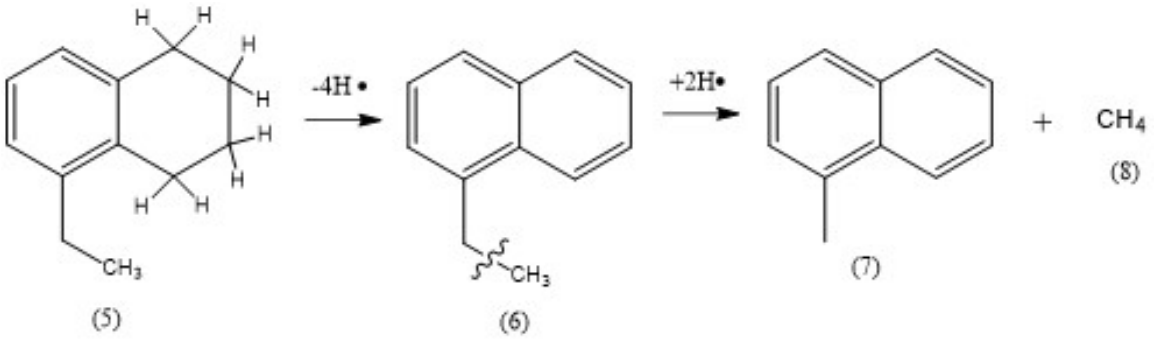


Figure 2.10: Sequence of reactions speculated to be occurring at 350°C based on SMCR results.

When compound (5) is converted to (6) through the loss of hydrogen, the -CH2 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 nCH2/nCH3 parameter (Table 2.3) 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 2.17). To verify this, another parameter that represents the ratio of tri-substituted, meta-, para-disubstituted aromatic C-H bends at 810, 860cm^{-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 Fig. 2.10 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.

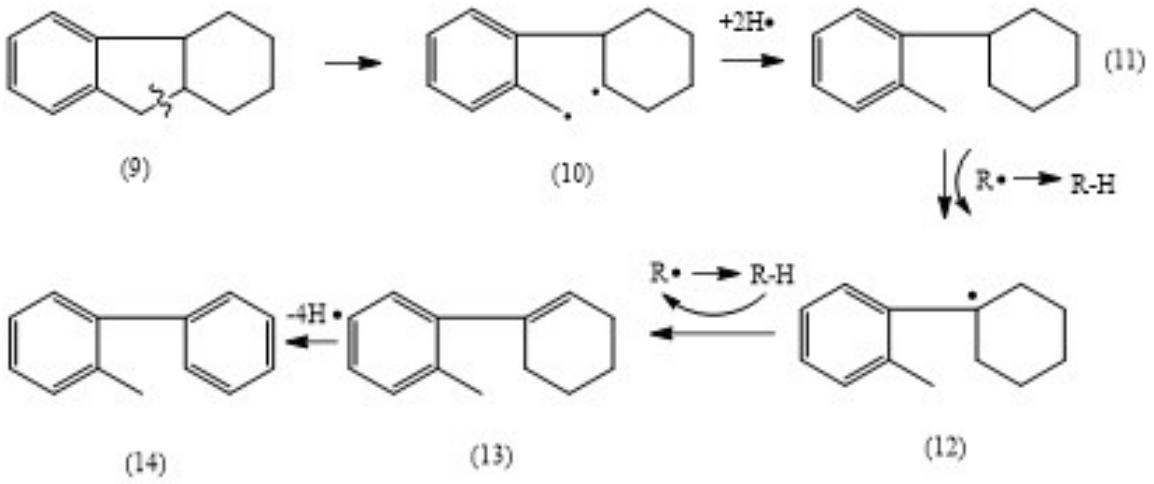


Figure 2.11: Pathway showing the increase in mono-substituted aromatic content from a naphthene, keeping the di-substituted content constant. The bond dissociation energy (BDE) for homolytic cleavage of the indicated bonds is also shown in kJ/mol.

The question remains whether the decrease in EOS for Athabasca bitumen at 350°C 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°C, which is equivalent to abstracting a hydrogen from methane to produce methyl radical, which is quite difficult under these conditions. [177] However, there is evidence for scission of the stronger Ar-C-H bond in benzene when it was used to study the cracking of nC16 paraffin in the temperature range of 398 – 450°C at pressures of 13 MPa. [180] Though cracking of the side chain was postulated to occur (Figure 2.10), the scission of the aryl C-alkyl C did not seem feasible at 350°C. Figure 2.11 depicts a reaction chemistry that can lead to a decrease in EOS without the cleavage of the aryl C-alkyl C bond and keeps 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 2.11 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). Direct conversion from (10) to (13) is also possible by intramolecular 1,4-hydrogen transfer.

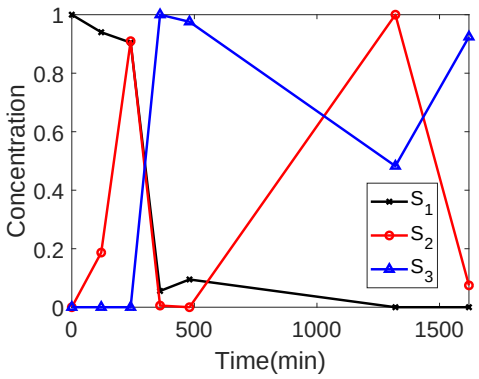
Little weightage should be given to the COC calculation as the resolved spectrum of S_3 was noisy (Figure 2.9c) and not representative of the spectra arising from real compound classes. The intensity at 1740 cm^{-1} appeared to decrease from S_1 to S_2 signifying the conversion of ester-type and anhydride-type compounds, but no interpretation can be assigned to S_3.

Results and analysis at 380°C

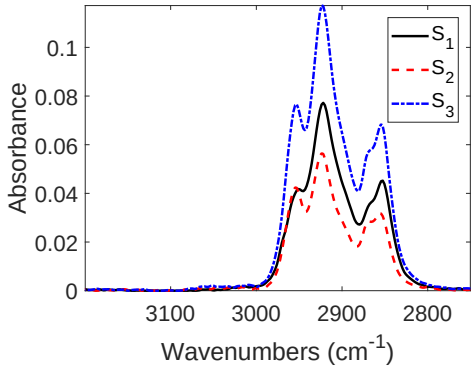
The final resolved concentration and spectral profiles for the FTIR spectra of the liquid products at 380°C are given in Figure 2.12. The residual plot when the matrix reproduced from the ALS-optimized profiles was subtracted from the original data matrix is given in Figure A.12c in the Appendix.

The initial estimates were good starting points for the dataset at this temperature (Figure 2.6c) as the final concentration profiles seemed to follow a similar trend (Figure 2.12a). 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°C from the nature of the concentration profiles of the pseudo-components. Table 2.4 provides the values of the derived quantitative parameters from the resolved spectra at 380°C.

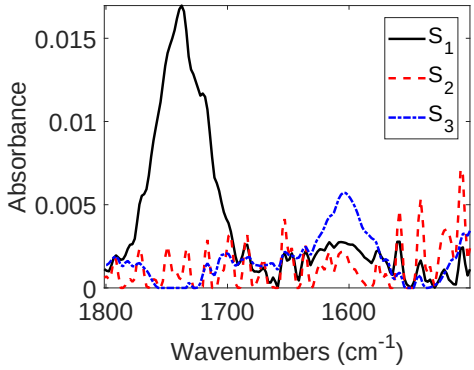
Significant changes can be observed in all the parameters as compared to the previous temperatures (Table 2.2 and Table 2.3). Overall, nCH2/nCH3 decreased from 1.80 (S_1) to 1.50 (S_3) with a minimum of 1.28 for S_2. The resolved spectra in the range 2750 – 3000 cm^{-1} reflected this trend and were different from the ones at



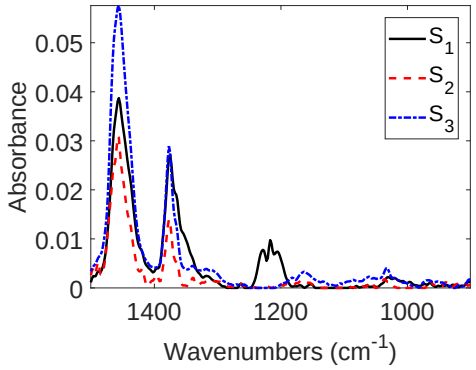
(a) Concentration vs. reaction time for the three pseudo-components



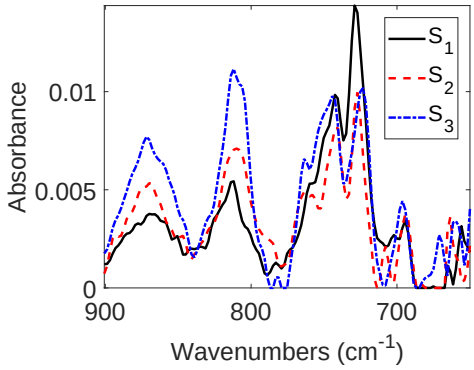
(b) Resolved spectra of the pseudo-components in the range 3200 – 2750 cm^{-1}



(c) Resolved spectra of the pseudo-components in the range 1800 – 1500 cm^{-1}



(d) Resolved spectra of the pseudo-components in the range 1500– 900 cm^{-1}



(e) Resolved spectra of the pseudo-components in the range 900 – 650 cm^{-1}

Figure 2.12: Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 380°C.

Table 2.4: Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 380°C.

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

300°C and 350°C where there was a minimal change in intensity in this wavenumber range. More gas was released during cracking at 380°C 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 nCH2/nCH3 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. [117] This was slightly different from that for Cold Lake bitumen conversion at 360°C where the nCH2/nCH3 increased from 1.82 to 1.89 for S_1 to S_2 but then decreased to 1.87 for the third pseudo-component.[176] This could indicate the onset of cracking at larger reaction times at 360°C 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 -CH3 content.

Surprisingly, DOC and EOS showed opposite trends for Athabasca bitumen conversion at 380°C in this work (Table 2.4). 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 nCH2/nCH3. This reaction sequence is illustrated in Figure 2.13.

The naphthenic ring ((15) in Figure 2.13) becomes more stable when it loses 4 hy-

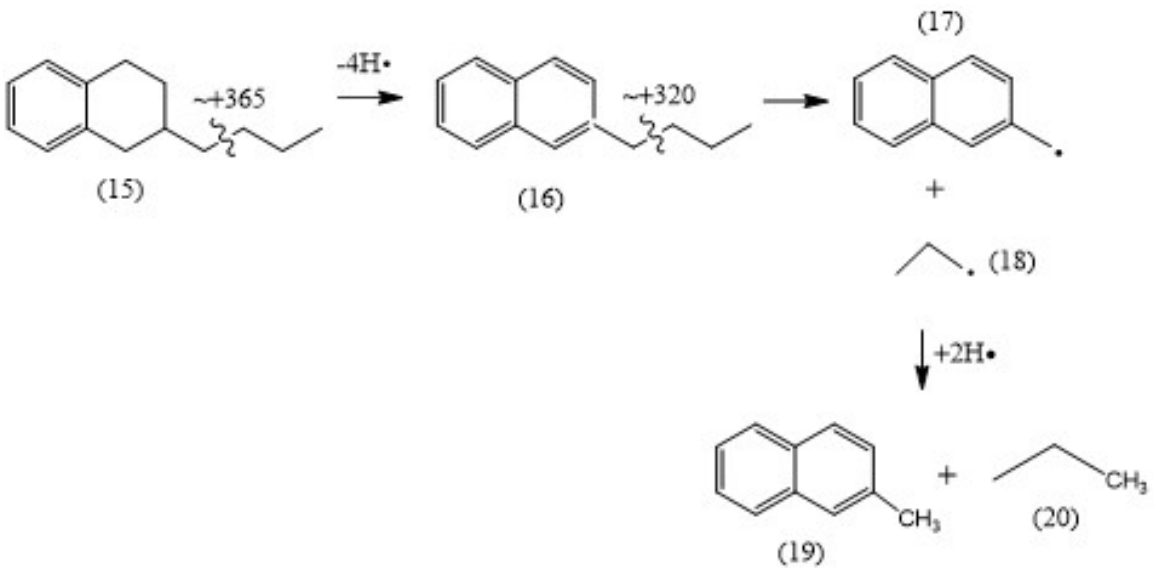


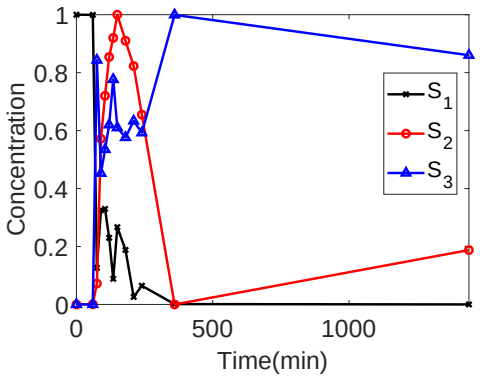
Figure 2.13: Proposed mechanism corresponding to the changes in derived quantitative parameters observed at 380°C. The energies for homolytic bond cleavage of the C-C bonds in (15) and (16) are given in kJ/mol.

drogens 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 2.13) less demanding as the BDE for its homolytic cleavage is much smaller than it was for (15) as indicated in Figure 2.13. 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, -CH3 content has increased and chain length decreased that manifests as a decrease in nCH2/nCH3 (Table 2.4). Due to the formation of an adjacent aromatic ring to the benzene ((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 2.18).

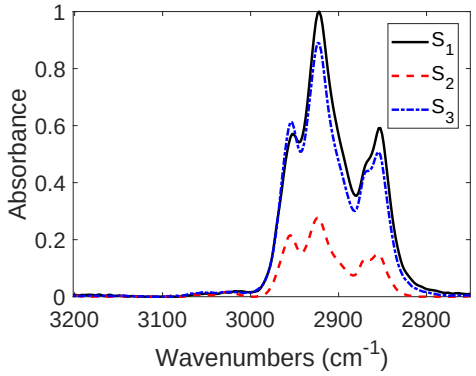
Also, when compared to (15), the intensities at 810 cm^{-1} and 860 cm^{-1} for (19) would definitely be higher (as observed in Figure 2.12e) due to m- and p- substitution with respect to the methyl substituent for the second aromatic ring on the right. If the mono-substituted aromatic content is considered constant, this would cause the

EOS to increase from the way it is calculated in this work (equation 2.17). Also, it can be seen from Figure 2.12e 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 but remained stable for S_2 and S_3. This further could contribute to the increase in EOS and corroborates the chemistry proposed in Fig. 2.13, though conclusive experimental proof is not available.

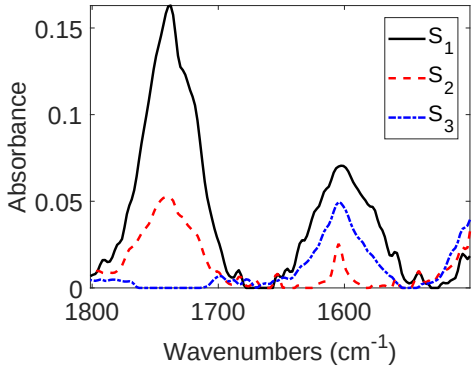
The absorption bands at 1380 cm^{-1} for all three components appeared to be split, indicating that not much methyl transfer had happened. Still, methane gas was found to be dominant in the gaseous products and might have been produced from the cracking of side chains as shown in Figure 2.10 for 350°C. On the other hand, DOC did not vary over the three pseudo-components for Cold Lake bitumen conversion at 360°C 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°C, 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.[110] Cold Lake bitumen was shown to exhibit a steeper viscosity decrease by Wang et al., [117] 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 nCH2/nCH3 at 380°C for Athabasca bitumen (Table 2.4), viscosity should also have increased at some point; but 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. [111], [182], [183]



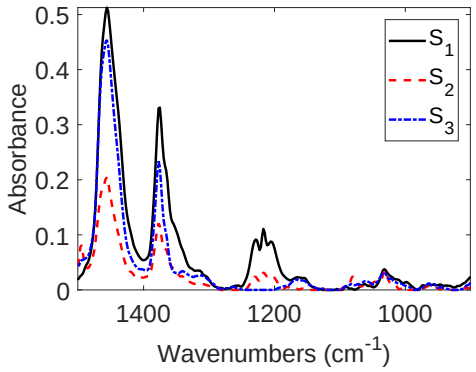
(a) Concentration vs. reaction time for the three pseudo-components



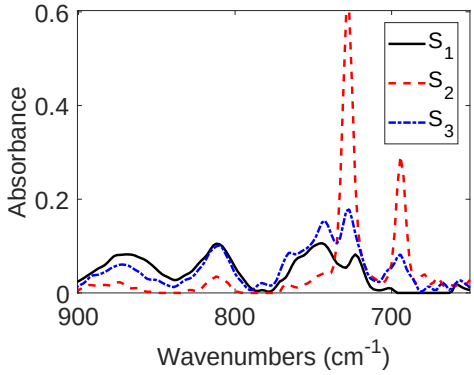
(b) Resolved spectra of the pseudo-components in the range 3200 – 2750 cm^{-1}



(c) Resolved spectra of the pseudo-components in the range 1800 – 1500 cm^{-1}



(d) Resolved spectra of the pseudo-components in the range 1500– 900 cm^{-1}



(e) Resolved spectra of the pseudo-components in the range 900 – 650 cm^{-1}

Figure 2.14: Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 400°C.

Results and analysis at 400°C

Among all temperatures, the most data was recorded for this dataset. Data for different types of characterization like 1H-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°C was documented in the study by Sivaramakrishnan et al. [111] 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°C dataset is provided in Figure 2.14 and the residual plots when the ALS-reproduced matrix is subtracted from the original data matrix is given in Figure A.12d in the Appendix.

The resolved concentration profiles for the three pseudo-components (Figure 2.14a) resembled those from the initial estimates (Figure 2.6b) for most parts, though S_1 experienced a steeper decrease and vanished at 360 min. It can also be seen that the concentration profiles deviated from the closure condition in the middle 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 (sub-section 2.3.2 titled 'Implementation of constraints'). This could be possibly because the intermediate times as mentioned above are where all the 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°C. [111] 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°C. Table 2.5 gives the values of the derived quantitative parameters from the ALS-resolved spectra for the 400°C dataset.

Table 2.5: Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 400°C.

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 nCH2/nCH3 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 2.5). 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 2.14a) so the conversion of $S_1 \rightarrow S_2$ was considered as important as $S_1 \rightarrow S_3$ at 400°C.

Gas production increased to 16 %wt. over 1440 min 24 and this combined with the overall decrease in the CH2 content or increase in the CH3 content indicated that cracking was taking place significantly. Interestingly, curve resolution on the FTIR spectra of cracked products of Cold Lake bitumen at 400°C determined that nCH2/nCH3 reduced from the first to the third pseudo-component but reached a maximum for the second component.[25] 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 nCH2/nCH3. In the case of Athabasca bitumen, the trend in nCH2/nCH3 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.

Inspection of the resolved spectral profiles at 1380 cm^{-1} (Figure 2.14d) provided evidence of methyl transfer happening at 400°C. 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 -CH2 content with methyl content being a constant (as was illustrated in Figure 2.8). This could have led to the observed increase in nCH2/nCH3 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 2.15) can also lead to an increase in -CH2 content, which is explained later in this section.

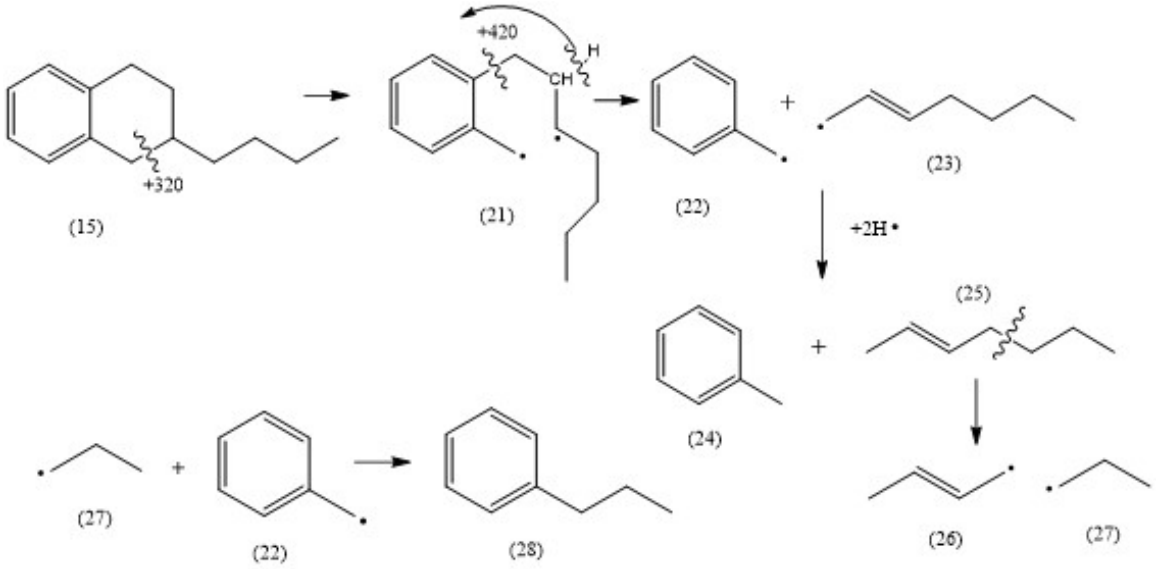


Figure 2.15: Plausible type of reaction happening at 400°C where cracking of the weaker benzylic C-tertiary C (in compound (15)) followed by intramolecular hydrogen transfer and hydrogen abstraction to yield the mono-substituted aromatic (compound (24)) and the conjugated free radical (23). This can crack further to give lighter aliphatic products. Possibility of free-radical recombination to form compound (28) is also shown.

DOC decreased dramatically from 2.81 to 1.77 in going from S_1 to S_3 with a minimum of 0.58 for S_2 (Table 2.5). This was another observation that differed from Cold Lake bitumen conversion at 400°C, 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.,[25] 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.[110] 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 required the coke to be hydrogen deficient in order to maintain the H/C ratio of the liquid products.

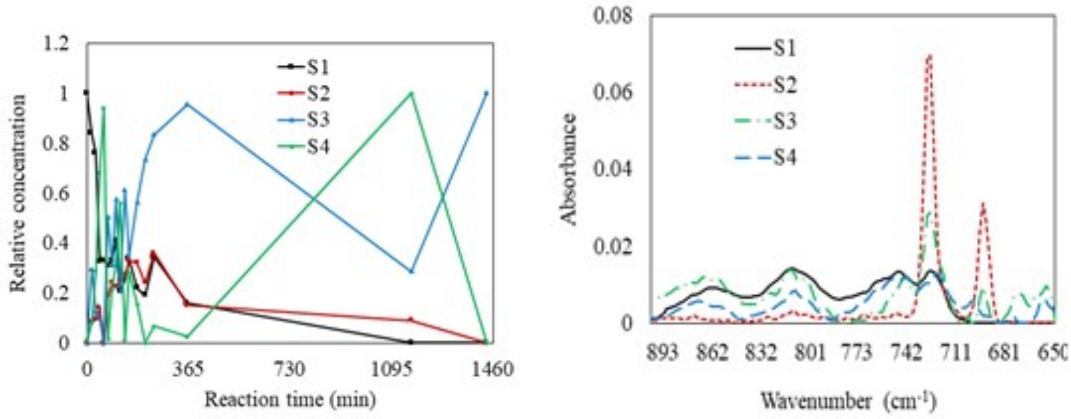
If Athabasca bitumen behaved similar to Cold Lake bitumen, its DOC should have increased slightly owing to the slight decrease in H/C. The fundamental difference between H/C and DOC was that H/C related to the whole liquid product while DOC conformed to the aromatic region only. The minimal change in H/C also found support in the work by Wiehe,[184] where in the model for coke formation from pyrolysis of Cold Lake bitumen at 400°C, he suggested that asphaltenes reached a constant ratio H/C once a second phase called mesophase (that eventually leads to coke) started 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°C. EOS decreased from 3.20 to a surprisingly low value of 0.16 in going 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 2.14e 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 reaction sequences: (i) as shown in Figure 2.11, 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 [185] 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 2.15. A previous study [180] 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°C. [180], [185] The possibility of further cracking of the 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.

The same compound (15) can follow different paths during cracking depending on the temperature and its structure (compare Figure 2.13 and Figure 2.15). 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 2.13). 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°C than at 380°C, conversion of the naphthene to an aromatic is not required to aid in the cleavage of the C-C bond in the C4 substituent as was proposed to occur at 380°C (Figure 2.13). 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 nCH2/nCH3 from S_2 to S_3 (Table 2.5). 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.



(a) Concentration profiles

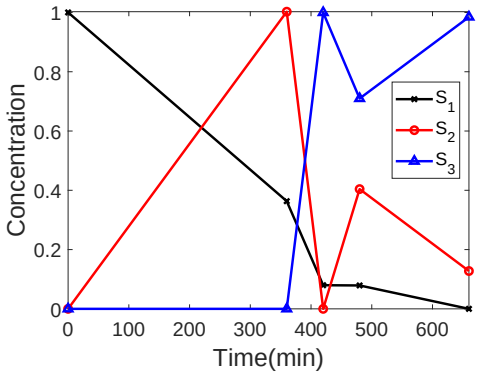
(b) Spectral profiles in the region 900 – 650 cm^{-1}

Figure 2.16: Curve resolution applied on the 400°C dataset using 4 pseudo-components.

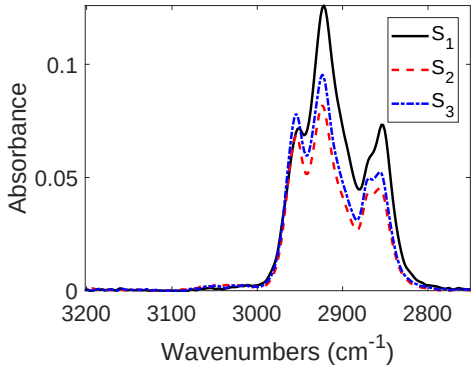
Since the value of EOS for S_2 was quite low, an internal test was done to check the validity of this observation. This was done by relaxing the number of components from 3 to 4 and inspecting the resolved concentration and spectral profiles in the 900–650 cm1 region. The concentration profiles with 4 components are noisier during the lower reaction times as shown in Fig. 2.16a. The spectral profiles (Fig. 2.16b) 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. There was no benefit in relaxing the number of components.

Results and analysis at 420°C

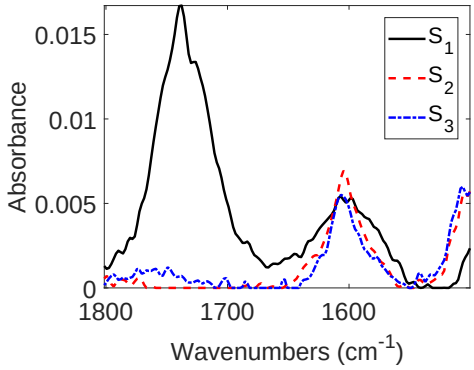
First, it should be noted that this temperature is higher than the lower visbreaking region as defined by Wang et al., [117] which was considered to be 400°C and lower.



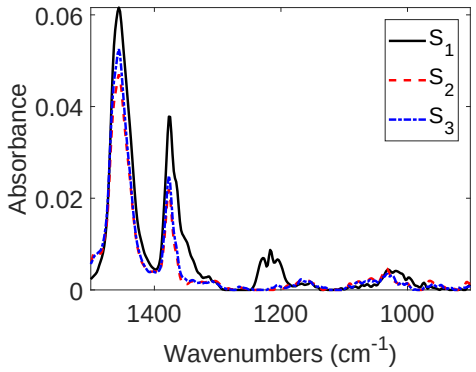
(a) Concentration vs. reaction time for the three pseudo-components



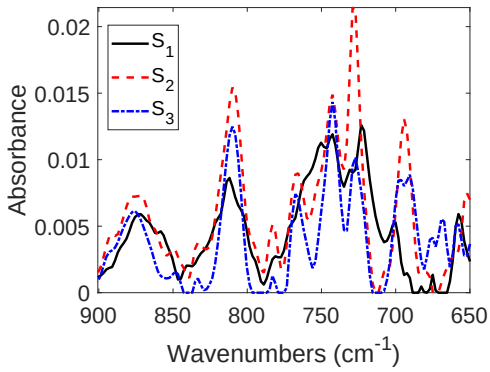
(b) Resolved spectra of the pseudo-components in the range 3200 – 2750 cm^{-1}



(c) Resolved spectra of the pseudo-components in the range 1800 – 1500 cm^{-1}



(d) Resolved spectra of the pseudo-components in the range 1500– 900 cm^{-1}



(e) Resolved spectra of the pseudo-components in the range 900 – 650 cm^{-1}

Figure 2.17: Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 420°C.

Since the lower temperatures were investigated previously, it was decided to operate close to industrial conditions that function between 430 – 490°C. [118] As expected, a large amount of coke was formed (17 %wt. at 420°C and 660 min compared to 12 %wt. at 400°C and 1440 min) with the liquid yield being less than that at lower temperatures. The resolved concentration and spectral profiles obtained through the ALS-optimization at 420°C are provided in Figure 2.17a and the residual plot is given in Figure A.12e in the Appendix.

There were similarities in the initial concentration estimates (Figure 2.6a) and the final resolved profiles (Figure 2.17a) 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 2.6 provides the values of the spectra-derived quantitative parameters for the three pseudo-components at 420°C.

Table 2.6: Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number at 420°C.

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

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. nCH2/nCH3 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 2.13), thus preventing them from further recombination to yield longer molecules such as (28) in Figure 2.15.

DOC decreased continuously from 2.82 for S_1 to 1.33 for S_3 (Table 2.6) 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 2.6). This value was lower than that observed at 400°C, suggesting the formation of more monosubstituted aromatics through either mechanism as proposed for 400°C in the previous section. The absorption bands for C-H deformation at 1380 cm^{-1} (Figure 2.17d) 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. [135] A clear increase in mono-substituted aromatic content (Figure 2.17e) meant that side reactions such as intra-aromatic ring closing were suppressed. Overall, it can be said that cracking was quite prominent at 420°C.

2.4.4 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 the ALS-resolved spectra, values in LOF and R^2 and the speed of convergence are discussed in this section.

Results at 300 °C

All concentration and spectral profiles for the PSO-optimized method are provided in the Appendix. Fig. A.13 depicts the concentration and spectral profiles along with the ALS–PSO residual when SMCR was conducted on the 300°C dataset.

In comparing the concentration profiles of ALS-PSO (Figure A.13a) and ALS (Figure 2.7a), 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 2.7c) but was well resolved in ALS-PSO (Figure A.13d). The residual plot for ALS-optimized profiles (Figure A.12a) extended only in the negative direction which seemed unusual. But in the case of ALS-PSO-derived profiles (Figure A.17b), the residual had both positive and negative values and was more symmetric than the ALS profiles alone. Table 2.7 shows the ALS-PSO spectra-derived quantitative parameters for S_1, S_2 and S_3 for the resolution performed on the 300°C dataset.

Although the absolute values of the parameters in Table 2.7 were slightly different compared to the ones derived from the ALS-resolved spectra (Table 2.2), the overall

Table 2.7: Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 300°C.

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

trends were the same, leading to the same interpretation of the chemical changes and types of reactions.

Results at 350°C

Figure A.14 in the Appendix 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 the ALS-PSO optimized profiles (Figure A.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 the 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).

Table 2.8: Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 350°C.

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

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 A.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 2.8. 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 2.8) in the PSO case whereas there was a maximum at S_2 for the ALS case (Table 2.3). 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 2.8.

Results at 380°C

The results of the ALS-PSO analysis on the FTIR spectra at 380°C along with the residual is provided in Figure A.15 in the Appendix. 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 A.15a). The concentration of S_3 showed a monotonic rise from the start which was different from that in the ALS profiles (Figure 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.

Table 2.9: Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 380°C.

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

Similar to other temperatures, the noisy region for S_2 in the 1800 – 1500 cm^{-1} region of the ALS profiles (Figure 2.12c) was mitigated by the addition of the PSO method as can be seen from Figure A.15d without altering the COC value too much (Table 2.9). The values of the spectra-derived parameters for the 380°C dataset are given in Table 2.9.

It was interesting to see that the trend in the nCH2/nCH3 was similar to that of profiles obtained through the 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 -CH3 stretch was more than -CH2 for this pseudo-component, which resulted in a decrease in the chain length parameter (Table 2.9). The trends in other parameters like EOS, DOC and COC were the same as in the ALS profiles, resulting in the same chemical interpretation. The persistence of band splitting for all three pseudo-components at 1380 cm^1 (Fig.A.15e) provides more credibility to the proposition of minimal methyl transfer occurring at 380°C.

Results at 400°C

The concentration and spectral profiles for the 400°C dataset resolved by the ALS-PSO optimization along with the residual plot are given in Figure A.16 in the Appendix. The relative concentrations of S_1, S_2 and S_3 followed similar paths for both the ALS and ALS-PSO-optimized profiles as seen in Figure 2.14a and Figure A.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 (sub-section 2.4.3 titled 'Results and analysis at 400°C'). 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 2.10. 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 727cm^{-1} for S_2 and S_3 (Figure A.16f) were similar to that of the ALS-optimized profile (Figure 2.14e). The spectral bands

Table 2.10: Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 400°C.

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

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 the corresponding section in the ALS-optimized profiles.

Results at 420°C

The results of ALS-PSO optimization at 420°C are given in Figure A.17 in the Appendix. The concentration of S_1 was similar to the profiles obtained from both the ALS (Figure 2.17a) and ALS-PSO methods (Figure A.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.

Table 2.11: Change in the ALS-PSO-resolved spectra-derived quantitative parameters with pseudo-component number at 420°C.

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

The residual appeared to fluctuate in a lower range for PSO ($\pm$0.003 in Figure A.17b) while it reached a maximum of 0.008 for the ALS-optimized profile (Figure

A.12e). This led to a lower LOF and R^2 by the slightest of margins, which is shown in the next section. The spectra-derived quantitative parameters for the ALS-PSO profiles at 420°C are summarized in Table 2.11.

At both 400°C and 420°C, ALS profiles were well resolved in all regions including the noise-prone 1800 – 1500cm^{-1} region (Figure A.17d) and reflects the decrease in COC value for the 420°C dataset which corroborates with the ALS profiles as well (Table 2.6). The absorption bands at 1380cm^{-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 the corresponding section on ALS-optimized profiles. While trends in nCH2/nCH3 and DOC were similar to the ALS-optimized profiles (Table 2.6), the value of EOS for S_3 slightly decreased from S_2 for the ALS-PSO-optimized profiles (Table 2.11) 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.

2.4.5 Comparison of ALS and ALS-PSO methods

Table 2.12: LOF and R^2 values for the dataset at each temperature when ALS and ALS-PSO were employed as the final optimization approach.

*Metric	350°C		380 °C		400°C		420°C	
	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 metrics at 300°C are not shown due to the lesser number of datapoints.

ALS and ALS– PSO are compared in terms of LOF, R^2 and the rate of convergence. The values of LOF and R^2 as calculated from the residual plots for each method are given in Table 2.12.

Tefera et al. [25] reported that the ALS algorithm applied on Cold Lake bitumen [186] converged in about 10 iterations at all temperatures and Shinzawa et al. [154] 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 the global phase angle proposed by Noda [187] and this considered the time sequence of events and the effect of any external perturbation. In our work, we can see from Table 2.12 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°C, 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.

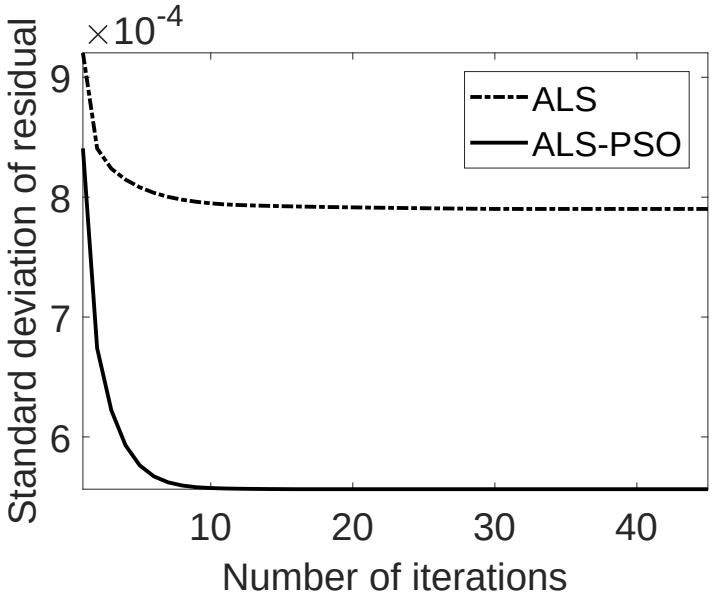


Figure 2.18: Rate of convergence in terms of standard deviation of residual vs. number of iterations for ALS and ALS-PSO algorithms used in MCR in this work.

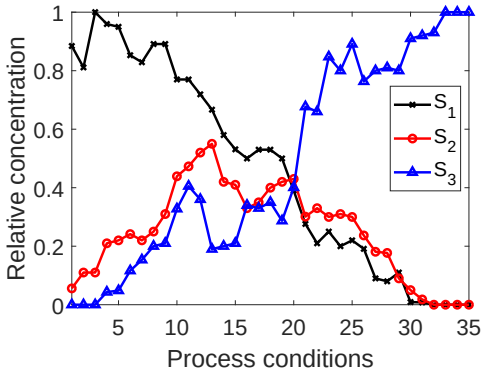
The rate of convergence for both the methods is depicted in Figure 2.18 and shown for 420°C and it was similar for the other temperatures considered. ALS converged in about 30 iterations and the standard deviation of the residual plotted on the y-axis is

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.

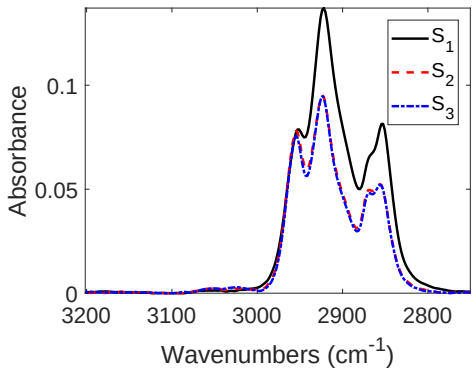
2.4.6 Global model for SMCR

In addition to the temperature-specific local models, a global model was developed that incorporated datasets at all temperatures together. The results were compared with that of the temperature-wise model to see the improvement in the resolution of the concentration and spectral profiles due to the larger number of samples and also to see whether the reaction sequences described previously continue to hold. 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.

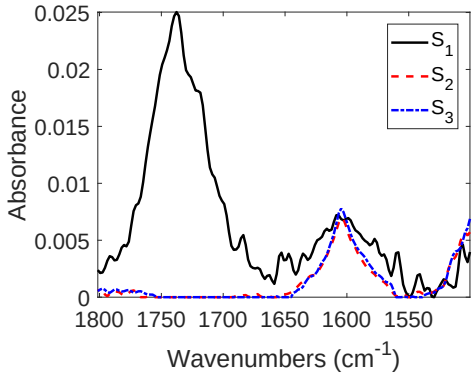
The number of components extracted for the global model were 3 since the ROD exhibited a maximum value of 22.3 when three principal factors were used. Fig. A.19a in the Appendix shows the plot of ROD varying with the number of components for the 35 samples. The LOF as calculated from the residual obtained after performing SVD using 3 pseudocomponents (eqn 2.14) 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 in the R^2 when more than 3 components were added to the model.



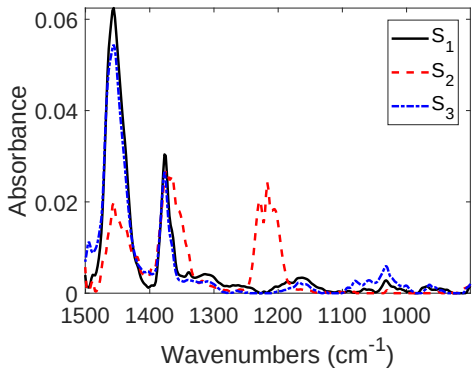
(a) Concentration vs. reaction time for the three pseudo-components



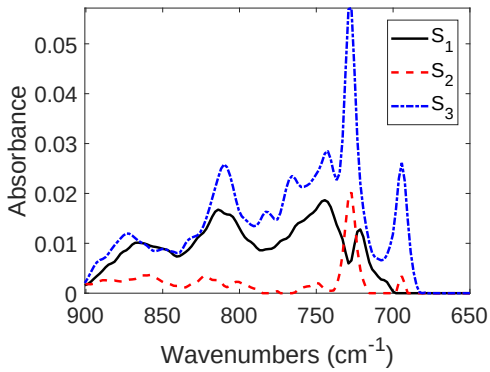
(b) Resolved spectra of the pseudo-components in the range 3200 – 2750 cm^{-1}



(c) Resolved spectra of the pseudo-components in the range 1800 – 1500 cm^{-1}



(d) Resolved spectra of the pseudo-components in the range 1500– 900 cm^{-1}



(e) Resolved spectra of the pseudo-components in the range 900 – 650 cm^{-1}

Figure 2.19: . Results of SMCR-ALS applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at temperatures in the range 300 – 420 °C (global model).

Figure A.19b shows the plot for the 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 2.6). 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 was not seen in the final resolved concentration profiles though the overall trends were the same as shown in Figure 2.19a. 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°C dataset.

Table 2.13: Change in the ALS-resolved spectra-derived quantitative parameters with pseudo-component number for the dataset comprising all experimental conditions.

Metric	350°C		380°C		400°C		420°C	
	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

Figure 2.19 shows the concentration and spectral profiles resolved using the 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 2.13.

As seen from Table 2.13, nCH2/nCH3 showed an overall decrease that was similar to the trends from the SMCR results at all the individual temperatures except 300°C. 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 °C, 400°C, 420°C) 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. 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°C, for which the related plausible reaction chemistry was explained in the sub-section 2.4.3 titled 'Results and analysis at 380°C'. Also, the sharp reduction in EOS for S_2 was also observed at 400°C and explained by relaxing the number of components to 4 (sub-section 2.4.3 titled 'Results and analysis at 400°C'). Sivaramakrishnan et al. [111] reported an increase in the mono-substituted aromatic content during the thermal cracking of Athabasca bitumen conducted at 400°C and the increase in the absorption intensity at 727 cm^{-1} for S_2 and S_3 (Figure 2.19c) was also supportive of this observation.

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 the higher substituted aromatics to lower substituted aromatics, which was indicated in the reaction sequence proposed from the local models at 420°C. Lastly, the resolution of the spectral profiles (Figure 2.19b-Figure 2.19e) were good with no noisy spectra as was obtained in the 1800 – 1500cm^{-1} in the 300°C and 350°C 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 exemplifies the importance of investigating the spectra at each temperature separately, though the global model is better applicable for online monitoring of continuous processes,

since monitoring can be accomplished by tracking changes in kinetics related to the single global mechanism applied over the entire range of operating conditions.

Our studies have covered a range of local models and the global model, with the tracking of many quantitative parameters and chemical interpretation and the postulation of reaction mechanisms based on the trends in these parameters. In Table 2.14, we summarize the trends in the parameters and the associated chemical interpretation. Note that, as mentioned earlier, the n-CH2/n-CH3 parameter is a measure of the average aliphatic chain length and an indication of the presence of non-aromatic cyclic rings. DOC and EOS primarily shed light on the aromatic nature of the components, and COC, which focuses on the C=O stretching frequencies for ester and anhydride-type carbonyl compounds and their changes, can indicate the conversion of complex oxygenates.

2.5 Conclusions

This work dealt with the application of chemometric tools on the FTIR spectra of liquid products obtained during the thermal conversion of Athabasca bitumen in the temperature range of 300–420°C. The objective was to develop reaction sequences for the thermal cracking process based on the results of the statistical approaches while using minimum prior chemical knowledge of the system. Some differences in chemometric results and subsequent chemical interpretation of the reaction chemistry of Athabasca and Cold Lake bitumen were also highlighted in this work.

In terms of methodology, both local and global SMCR models showed similar behaviour, and were able to represent the reaction system with three pseudo-components. The global SMCR model showed the value of having more samples in the dataset and was in good agreement with the proposed reaction sequences from the local models. As suggested in previous works, the ALS–PSO approach was found to be superior to the ALS method in terms of noise reduction, stricter adherence to the closure constraint and quicker convergence. Both methods predicted the same trends in the final

Table 2.14: Summary of ALS–PSO-resolved spectra-derived quantitative parameters and chemical interpretation for local and global models

	Local model (300°C)	**Local model (350°C)**	**Local model (380°C)**	**Local model (400°C)**	**Local model (420°C)**	**Global model**
n-CH2/n-CH3	Constant across S_1, S_2, S_3	Slight decrease from S_1 to S_2 to S_3	Decrease from S_1 to S_2, slight increase to S_3	Decrease from S_1 to S_2, slight increase to S_3	Decrease from S_1 to S_2/ S_3	Decrease from S_1 to S_2/ S_3
Overall EOS	Decrease from S_1/S_2 to S_3	Slight increase from S_1 to S_2 but overall sharp decrease to S_3	Decrease from S_1 to S_2 but overall increase to S_3	Decrease from S_1 to S_3, minimum at S_2	Decrease from S_1 to S_2/S_3	Decrease from S_1 to S_3, minimum at S_2
DOC	Decrease from S_1 to S_2/S_3	Increase from S_1 to S_2 and decrease to S_3	Decrease from S_1 to S_2/S_3	Decrease from S_1 to S_2, slight increase to S_3	Decrease from S_1 to S_2 to S_3	Decrease from S_1 to S_2, slight increase to S_3
COC	Non-zero for S_1	Significant non-zero value for S_1	Decrease from S_1 to S_2 to S_3	Comparable non-zero values for S_1, S_2	Non-zero for S_1	Non-zero for S_1
Chemical interpretation	Low extent of cracking, some methyl transfer	Some cracking, conversion of naphthenes to aromatics through hydrogen disproportionation, increased formation of o-disubstituted aromatics	Cracking and free-radical recombination at higher reaction times, hydrogen disproportionation from naphthene aromatics having transferable hydrogen and alkyl side chains	Cracking, intra-molecular hydrogen transfer and free-radical recombination at higher reaction times, formation of mono-substituted aromatics	Cracking, formation of mono-substituted aromatics, minimal methyl transfer, breakage of Ar–C–alkyl–C bond	Cracking with methyl transfer and free-radical recombination, hydrogen disproportionation from naphthene aromatics having transferable hydrogen and alkyl side chains,conversion of higher to lower substituted aromatics

resolved spectra-derived quantitative parameters.

In terms of the conversion chemistry, the pathway proposed among the three pseudo-components was $S_1 \rightarrow S_2 \rightarrow S_3$. As the temperature increased from 300°C to 420°C, it was seen that the extent of cracking increased and lighter products with a higher fraction of mono-substituted aromatics were formed. Methyl transfer was suggested to be dominant at the lower temperatures with minimal cracking. In contrast, for Cold Lake bitumen, ring-closure reactions were seen to be occurring at median temperatures of 340°C. At 380°C, it was speculated that conversion of naphthene rings to aromatics facilitated cracking of the side chains at later times. At 400°C, the possibility of scission of Ar–C–alkyl C bond was also seen to form mono-substituted aromatics. Free-radical recombination was an important reaction at larger reaction times, but the rate of the recombination and condensation reactions were thought to be higher in Cold Lake bitumen than in Athabasca bitumen at 400°C due to differing trends in viscosity. Severe cracking was seen to occur at 420°C for Athabasca bitumen with stabilization of side chains by hydrogen. 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 practice. The chemometric framework can also serve as a hypothesis generator for experimental procedures involving complex molecules.

Chapter 3

Data Fusion by Joint Non-negative Matrix Factorization for Hypothesizing Pseudo-chemistry Using Bayesian Networks

Abstract

Inferring the reaction pathways underlying the processing of complex feeds, using noisy data from spectral sensors that may contain information regarding molecular mechanisms, is challenging. This is tackled by a two-step approach for the partial upgrading of Cold Lake bitumen: first, joint non-negative matrix factorization (JNMF) is used as a data fusion algorithm to extract pseudocomponent spectra by combining complementary information about the reacting environment from Fourier transform infrared (FTIR) and proton nuclear magnetic resonance (^{1}H-NMR) spectroscopic sensors. Second, a probabilistic inferential model that hypothesizes reaction mechanisms among the identified pseudocomponent spectra is constructed using Bayesian networks that encode directed acyclic causal pathways among the nodes of the random variables (pseudocomponent spectra). The JNMF algorithm has been developed to handle process data artefacts by imputing missing data, using a rotationally invariant norm for robustness to outliers and noise, and enforcing the non-negativity

This chapter has been published as: A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, V. Prasad. Data fusion by joint non-negative matrix factorization for hypothesizing pseudo-chemistry using Bayesian networks. *React. Chem. Eng.* **2020**, *5*, 9, 1719-1737.

constraint to ensure physical interpretability in compliance with Beer's law for spectral data. The projected optimal gradient approach developed to solve the JNMF objective converges within fewer iterations at the specified tolerance as compared to the multiplicative update rules (MUR). Solution ambiguity in JNMF is limited by incorporating graph regularization terms: (a) Inter-sensor co-regularization that penalizes redundancy in the pseudocomponent spectra across spectral sensors (b) Intra-spectral manifold regularization that penalizes overfitting of the pseudocomponent spectra from each sensor by penalizing redundant peaks within a spectrum. Weighting the intra-spectral regularization term that minimizes similarly correlated peaks across spectral channels of a sensor to zero, is seen to result in chemically meaningful pseudocomponent spectra, given that different organic compounds share similar properties with respect to their hydrocarbon structure. Hence, the preferential weighting of regularizers is shown to act as a chemical information sieve by controlling the peaks that appear in the pseudocomponent spectra and thereby enabling the proposal of different reaction mechanisms, based on the similarity metric used to model the graph structure.

3.1 Introduction

Process integrated spectral analyzers that use flow cells, quartz windows or immersion probes are popularly used to obtain molecular-level information as they are fast, non-invasive, non-destructive, inexpensive and do not require sample preparation [188]. The process data from spectral analyzers are high dimensional, non-causal, non-full rank, noisy and may have missing values [189, 190]. Non-negative matrix factorization (NMF) has been used as a workhorse in signal and data analytics to extract latent features by the deconvolution of such low fidelity spectroscopic process data [191]. Existing multivariate curve resolution(MCR) algorithms on spectroscopic data [5, 25, 131] which constrain the factors to be non-negative in order to be physically interpretable by Beer Lambert's law are analogous to NMF; however, these lack the

ability to jointly analyze multiple spectral measurements as a way of incorporating complementary information. This works seeks to develop Joint Non-negative Matrix Factorization (JNMF) as a data-fusion analogue of curve resolution algorithms to obtain non-negative latent factors of spectral features across multiple sensors weighted additively in parts by the shared basis factor (interpreted as concentration). The components of the latent factors containing spectral features (pseudocomponent spectrum) are represented as nodes among which causal maps are learned using structure learning in Bayesian networks to generate reaction pathway hypotheses. The number of components in the latent factors arising from JNMF, is determined using the mathematical rank, based on singluar value decomposition (SVD) [192].

Previous work on generating reaction hypotheses employs encoding prior knowledge of the species and reaction rules as knowledge graphs [38, 41]. In the absence of prior knowledge of the species, hierarchical clustering of the spectral data to obtain clusters of spectral channels with similar absorbances, has been supplemented with domain knowledge to identify the classes of compounds[193], before learning causal pathways among them. Yet, the approach relies on prior knowledge of the number of clusters (representing species). Although MCR recovers the number of species and their corresponding latent spectra and concentrations, without *a priori* knowledge of the reaction system, spectral resolution has been formulated using a mixed integer non-linear programming approach [194]. This is owing to the rotational and intensity ambiguities in MCR [195]. Our approach proposes to develop JNMF as a data-fusion algorithm, with graph regularization and constrains the latent factors to limit solution ambiguity. JNMF in tandem with probabilistic graphical models, reduces the reliance on prior knowledge of the reaction system, while developing inferential models to generate reaction hypotheses. This framework is preliminary to the development of kinetic models that could be used to control the composition of complex mixtures[194], facilitating advances in reaction engineering using principles of process systems engineering [111]. In this work, data from Fourier Transform

Infrared (FTIR) spectroscopy and Proton Nuclear Magnetic Resonance (^{1}H-NMR) spectroscopy of the products from the thermal conversion of Cold Lake bitumen [196, 197], are mined to develop reaction pathways using machine learning tools.

3.1.1 Detailed background

NMF identifies latent factors to a level of limited ambiguities thereby increasing interpretability as compared to alternate factorization methods like singular value decomposition (SVD) and independent component analysis (ICA) that are based on orthogonal and independent factor decompositions that are unconstrained [191]. As an additive parts-based representation due to the non-negativity of latent factors, NMF has been used in the linear unmixing of spectroscopic data [198], soft clustering [199] and topic discovery in unlabeled datasets and has found wide usage in cancer subtype detection, blind source separation, text mining and image recognition [200]. Commonly used spectroscopic techniques produce multi-dimensional datasets providing multi-view information of the chemical samples being processed. In the realm of physically meaningful spectral interpretations based on Beer Lambert's law, these multi-dimensional datasets can be viewed as a linear mixing of weights (interpreted as concentrations) and reduced number of basis factors (interpreted as spectra of pseudocomponents); the linear unmixing of which is done using NMF [198]. As a blind source separation technique, it has been used to estimate spectra of reactant mixtures, whereby the inverse problem in chemical reactions which involves the profiling of spectra of reaction intermediates in the absence of pure component spectra is overcome [201]. NMF has also shown to be a promising tool to resolve mixed chemical signals of complex mixtures with a high degree of overlap as compared to traditional iterative and non-iterative spectral curve resolution techniques [202]. There is also evidence of using NMF for the time resolution of Raman spectra to fit kinetic models from information of the molecular structure of individual species encoded in the spectra of the latent factors [203].

Constrained NMF has a non-convex objective leading to a local optimum that leads to non-unique solutions, which is overcome by incorporating smoothness and sparsity constraints along with non-negativity for the case of hyperspectral unmixing [204]. An improvement in the uniqueness and performance of a typically unsupervised NMF algorithm used for clustering was observed by incorporating graph regularization to embed prior knowledge of geometrical structure and local invariance of the features in going from a higher to lower dimensional space [205]. Commonly used spectroscopic techniques produce multi-dimensional datasets providing multi-view information of the chemical samples being processed [82]. Our approach proposes to use multi-view spectral sensor data to incorporate graph regularization to limit solution ambiguity [195] while jointly factorizing data from multiple spectral sensors to obtain latent factors constrained to be non-negative. The graph regularization term which has been used in the semi-supervised NMF approach points to the following [206]: a) manifold regularization so that the low dimensional space (latent factor) of each view has similar geometrical structure to the high dimensional space of that view, b) co-regularization for the geometric similarity of the latent factors across views with the high-dimenisonal spaces across views. The geometric structure in graph regularization can be determined using metrics like 0-1 weighting, heat kernel weighting or dot product weighting, which (for a normalized quantity) provides the cosine similarity metric [207]. In this work, the dot product weighting of normalized spectral absorbances from each sensor and across both sensors is used as a metric to encode intraview similarity among spectral channels from each of the sensors and interview similarity among spectral channels across both sensors in going from raw data space to latent factor space. NMF with regularization has been applied to semi-supervised clustering [205, 207], with a recent extension of jointly using multiple measurements from similar sensors for the same [206]. Hence, our approach proposes NMF to integrate multi-view information, i.e. JNMF of spectral data from different or dissimilar sensors, while incorporating knowledge of the views as regularization terms to limit

ambiguity in spectral deconvolution.

Heterogeneous data mined by data fusion based on simultaneous matrix factorization to reveal the hidden underlying representations has been implemented in computational biology [208, 209]. Fusion of heterogenous data that are complementary, such as genomics and proteomics data, is seen to increase the predictive performance and robustness of models [209]. Data fusion methods can be broadly classified based on the stage at which the fusion is performed [208, 210]: (a) Early Fusion: Sequential concatenation of data by neglecting the modularity, (b) Late fusion: Fusing the prediction model results obtained from each data source separately (It is not trivial to retrace the separate contribution of sources when the final model is used for inference), and (c) Intermediate fusion: Fusion propagated by features of each independent data source, [83, 84] making the structure of the predictive model robust. A popular algorithm to implement intermediate fusion is constrained simultaneous matrix factorization [208], which is tantamount to multi-view JNMF.

Recently, JNMF was used for data fusion of multi-view gene interaction network data with sparse penalty regularization constraints [83] . Adaptive JNMF, with different user-defined weights for the NMF of data modalities in each view, was used for the fusion of genomics and proteomics data to build clinical predictive models [209]. Diverse-JNMF was used to penalize redundancy in the fusion of multi-view data by using an orthogonality regularizer between the multi-view basis factors [211]. Weighted-NMF, where missing values are imputed by zero, [212] and Robust-NMF, where the objective function is based on minimizing the L_{21} loss function to robustly deal with outliers and noise [213], are other variants of NMF that are proposed to be extended to multi-view data for JNMF in this work. The L_{21} norm is a row-wise rotationally invariant L_1 norm, as it is computed by adding the row-wise L_2 norms. Hence, the use of such a norm for the loss function is seen to diminish the influence of noise and outliers.

Evidence of stage-based fusion of FTIR, ^{1}H-NMR and Raman spectroscopic data

having resulted in better crude characterization [82], has motivated us to develop a more robust intermediate fusion algorithm for integrating multi-view spectral data to build models for hypothesis generation of chemical pathways. Since NMF has the advantage of being an interpretable factor decomposition method, utilizes optimization based matrix computation routines for its solution and has a scalable formulation for large-scale problems, this work focuses on using it as a semi-supervised technique for the soft clustering of multiview spectral data into basis factors of the underlying latent objects weighted by a common parts-based matrix across all views. The main contributions of this paper are as follows:

1. JNMF is implemented as a data fusion algorithm for factorizing FTIR and ^{1}H-NMR spectral data that is robust to outliers, handles missing values, favors sparse latent factors and incorporates graph regularization to limit solution ambiguity by penalizing redundancy and overfitting of latent factors from different spectral sensors.

2. A projected optimal step gradient-based algorithm is developed to solve the JNMF formulation. It has sound convergence properties in comparison with other typically used iterative update rules like multiplicative update rules (MUR) and alternating least squares (ALS) [214, 215].

3. The latent structure information obtained from JNMF is used to build probabilistic graphical models using Bayesian networks to hypothesize the chemical pathways among the components of the latent factors, which in the physical sense correspond to chemically similar compound signatures, i.e. pseudocomponents, and mathematically correspond to the rank of the matrix from a spectral data view.

The paper is structured as follows: Section 3.2 outlines the proposed framework which includes formulation of the JNMF objective (Section 3.2.1), ascertaining the

number of species using rank determination (Section 3.2.2), algorithms to solve the JNMF objective (Section 3.2.3) and structure learning using Bayesian networks (Section 3.2.4). Section 5.4 discusses the the resulting hypothesized reaction pathways. Section 3.3.7 discusses incorporating correlation among spectral channels within and across spectral sensors as regularization terms in JNMF and its impact on the hypothesized pathways. Finally, Section 3.4 summarizes the work presented in the paper and highlights avenues for future work.

3.2 Methods

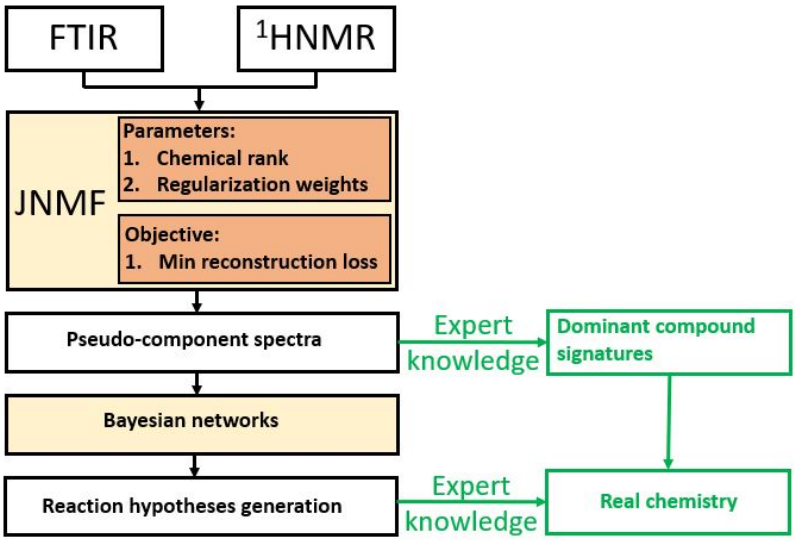


Figure 3.1: Schematic representation of the methods used to generate reaction hypotheses from spectral data and map it to real chemistry.

This work focuses on using JNMF as a data fusion projection-based method to jointly extract information from more than one spectral sensor to obtain latent feature representations which are referred to as the pseudocomponent spectra. The pseudocomponent spectra contain key molecular signatures of the dominant underlying compound classes, which are identified using expert knowledge. Hypotheses about the transitional pathways among the identified compound classes arise from learning the causal maps among the pseudocomponent spectra using Bayesian net-

works. These hypotheses then act as a basis to build a map to real chemistry using domain knowledge. A flowsheet that contextualizes the methods in the grand scheme of using spectral data to generate reaction hypotheses that are mapped to real chemistry using domain knowledge is shown in Figure 3.1.

3.2.1 Formulation of the objective function for JNMF

The constrained joint bilinear decomposition of the data blocks X_i according to the physically meaningful Beer Lambert law results in the commonly held matrix W, which is the concentration of pseudocomponents across the process conditions, and H_i, which is the absorbance spectra of the pseudocomponents. The number of pseudocomponents is determined as the minumum of the rank of individual data matrices using empirical measures of rank determination.

$$\min_{W,H_1,H_2\geq 0} F(W,H_1,H_2) = \sum_{i=1,2} ||X_i - WH_i||_F^2 \tag{3.1}$$

Equation 3.1 is tantamount to the minimization of the following trace:

$$\min_{W,H_1,H_2\geq 0} F(W,H_1,H_2) = Tr\big[\sum_{i=1,2}(X_i - WH_i)(X_i - WH_i)^T\big] \tag{3.2}$$

The simplest formulation of JNMF is shown by minimizing the objective function shown in Eqn. 3.1. The use of an L_2 norm in the objective function minimization is not robust to outliers as the outliers may drive $F(W,H_1,H_2)$ to undesirably large values. Hence, it is suggested to reformulate the objective function to minimize the L_{21} norm of the error, so that it is robust to outliers. The L_{21} norm of a matrix $A_{m\times n}$ is:

$$||A||_{21} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{m} A_{ji}^2} \tag{3.3}$$

Hence, Eqn. 3.1 translates to

$$\min_{W,H_1,H_2\geq 0} F(W,H_1,H_2) = \sum_{i=1,2} ||X_i - WH_i||_{21} \tag{3.4}$$

Just like Eqn.3.2, Eqn.3.4 is equivalent to the minimization of a trace that is scaled by a diagonal matrix $D(X = X_i - WH_i) = D_i$, which is associated with each term and is defined as:

$$D(X) = \frac{I_{n\times n}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \text{for any } X_{m\times n} \tag{3.5}$$

Hence, the minimization of the L_{21} norm is equivalent to minimizing the following scaled traces:

$$\min_{W,H_1,H_2\geq 0} F(W,H_1,H_2) = Tr\big[\sum_{i=1,2} (X_i - WH_i)D_i(X_i - WH_i)^T\big] \tag{3.6}$$

Additionally, the missing values in X_i are dealt with by imputing the matrices with corresponding weighting matrices P_i that weight the missing entries to zero and the rest to ones. Network regularization constraints that capture the relationship across the two blocks and within each of the blocks themselves by way of regulating the factorization using the underlying structure of the matrix cross-covariance and auto-covariance that are incorporated, besides having terms that regulate the sparsity of the decision variables W, H_1 and H_2.

Let $X_1\ _{n\times p_1}$ and $X_2\ _{n\times p_2}$ denote the blocks of FTIR and ^{1}H-NMR measurements, respectively. The autocovariances R_F and R_H are calculated as follows, where J is the notation for a matrix of ones:

$$R_F = \frac{\big(X_1 - \frac{J_{n\times n}X_1}{n}\big)^T\big(X_1 - \frac{J_{n\times n}X_1}{n}\big)}{n} \tag{3.7}$$

$$R_H = \frac{\big(X_2 - \frac{J_{n\times n}X_2}{n}\big)^T\big(X_2 - \frac{J_{n\times n}X_2}{n}\big)}{n} \tag{3.8}$$

The cross-covariance R_{FH} between the two blocks is

$$R_{FH} = \frac{\left(X_1 - \frac{J_{n\times n}X_1}{n}\right)^T \left(X_2 - \frac{J_{n\times n}X_2}{n}\right)}{n} \tag{3.9}$$

Accounting for this in the scheme of JNMF leads to the minimization of the following objective, making it a Weighted Robust JNMF with network regularization and sparsity constraints:

$$\begin{aligned} \min_{W,H_1,H_2\geq 0} F(W,H_1,H_2) &= \sum_{i=1,2} P_i||X_i - WH_i||_{21} + \alpha||H_1 R_{FH} H_2^T||_{21} \\ &+ \beta||H_1 R_F H_1^T + H_2 R_H H_2^T||_{21} + \gamma||W||_{21} + \lambda||H_1||_{21} + \lambda||H_2||_{21} \end{aligned} \tag{3.10}$$

Reformulating the above equation in terms of the minimization of the trace of scaled L_{21} norms gives

$$\begin{aligned} \min_{W,H_1,H_2\geq 0} F(W,H_1,H_2) &= Tr\big[\sum_{i=1,2} P_i(X_i - WH_i)D_i(P_i(X_i - WH_i))^T \\ &+ \alpha(H_1 R_{FH} H_2^T)D_3(H_1 R_{FH} H_2^T)^T \\ &+ \beta\{(H_1 R_F H_1^T)D_4(H_1 R_F H_1^T)^T + (H_2 R_H H_2^T)D_5(H_2 R_H H_2^T)^T\} \\ &+ \gamma W D_6 W^T + \lambda\{H_1 D_7 H_1^T + H_2 D_8 H_2^T\}\big] \end{aligned} \tag{3.11}$$

where $D_{i=1,2\cdots 8}$ corresponds to the diagonal scaling matrix evaluated using Eqn.C.5 for each term in the objective function.

3.2.2 Rank determination

The determination of the number of components/sources whose spectral signatures are mixed in proportion to their concentrations resulting in the measured spectral absorbances is a crucial step in MCR [216]. In an ideal scenario, where each chemical component makes a noise-free contribution to the data matrix, the number of principal factors equals the chemical rank (r). However, the practical determination is difficult due to the co-existence of instrumental factors and experimental noise [148]. The multitude of methods that have been developed to determine the number of principal factors can be broadly classified into : empirical, mathematical and statistical

methods [148]. A number of empirical and statistical methods have been reviewed [217], both for the prediction of rank when there is no prior assumption of noise, and for rank predictions when prior assumptions of noise are made using empirical indicator functions that are combined with a NIPALS (non-linear iterative partial least squares) routine to automate the prediction process. PCA was used to reduce the measured data to contain information relevant to the system, based on [33], where it was proved that the error associated with a dataset stems from the extracted error (contained within the minor PC dimensions $r+1, r+2, \cdots m$) and the imbedded error (contained in the r PCs) and can never completely be removed from the data. Decomposition of data (D) into an orthonormal basis set where $D_{m\times n}$ contains m recorded spectra as rows, each digitized into n points, results in $T_{m\times k}$ and $P^T_{n\times k}$: the score and loading matrix, respectively. The complete set of scores and loadings, i.e. with $k = n$, captures both the system variation and experimental noise. Hence, a number of metrics based on PCA [217–219] are used to separate the ($k = r < n$) eigenvectors that account for the systematic variations (imbedded error + variation in data) from those corresponding to noise (extracted error) in the leftover PCs; of which the Ratio of Derivatives based on Malinowski's indicator function is used to determine rank in this work.

$$D = T_{m\times r}P^T_{n\times r} + E_{m\times n} = \hat{D} + E_{m\times n} \tag{3.12}$$

Equation 3.13 indicates the computation of the eigenvalues of the k^{th} principal component as the sum of squares of the scores. Equation 3.14 computes the residual standard deviation (RSD) in terms of the eigenvalues of the remaining principal components normalized by their degree of freedom, which is then used to compute the imbedded error (IE) in Eqn. 3.15 in terms of which the Malinowski's indicator function is defined (Eqn. 3.16). The indicator function is used to define an empirical metric called the Ratio of Derivatives (ROD) indicated by Eqn. 3.17, an extrema in whose profile at a certain k number of components helps determine the rank.

$$
\begin{aligned}
q &= \min(n, m) \\
\mathrm{EV}_k &= g_k = \sum_{i=1}^{m} t_{ki}^2, \text{where k=}1, 2 \cdots q \qquad (3.13) \\
\mathrm{RSD(k)} &= \frac{k}{n}\sqrt{\frac{\sum_{j=k+1}^{q} g_j}{m(q-1)}} \qquad (3.14) \\
\mathrm{IE(k)} &= \sqrt{\frac{k}{n}\mathrm{RSD(k)}} \qquad (3.15) \\
\mathrm{IND(k)} &= \frac{n\ \mathrm{IE(k)}^2}{k(q-k)^2} \qquad (3.16) \\
\mathrm{ROD(k)} &= \frac{\text{IND(k-1)-IND(k)}}{\text{IND(k)-IND(k+1)}} \qquad (3.17)
\end{aligned}
$$

3.2.3 Algorithms to solve the Joint Non-negative Matrix Factorization Formulation

Multiplicative Update Rule

The multiplicative update rule algorithm [220] is a popular method for non-negative matrix factorization to find useful basis information of non-negative data by minimizing the Euclidean distance between approximate and true values, subject to constraints. Though it was shown that the non-convex objective function is non-decreasing [220], it was later proved that the decision variables converge to a stationary point with a slight modification [215], a condition that is vital to guarantee the local minima.

This works focuses on the extension of the modified algorithm of the MUR to obtain solutions to the optimization problem of Eqn. 3.1. The gradients of the function are computed as follows:

$$\nabla F_W = \sum_{i=1,2} P_i(WH_i - X_i)D_i H_i^T + \gamma W D_6 \tag{3.18}$$

$$\nabla F_{H_1} = W^T P_1(WH_1 - X_1)D_1 + \alpha H_1 R_{FH} H_2^T D_3 H_2 R_{FH}^T + \beta H_1 R_F H_1^T D_4 H_1 R_F + \lambda H_1 D_7 \tag{3.19}$$

$$\nabla F_{H_2} = W^T P_2(WH_2 - X_2)D_2 + \alpha H_1 R_{FH} H_2^T D_3 H_1 R_{FH} + \beta H_2 R_H H_2^T D_5 H_2 R_H + \lambda H_2 D_8 \tag{3.20}$$

The gradients given in Eqn. 3.18 are used to compute the modified step-sizes that are used in the MUR updates as outlined in Algorithm 1:

Algorithm 1 (Multiplicative Update Rule) ***Input****: Initialize decision variables* $\{ W^0, H_i^0\}$

Output*: W, H_i that solves* $\min_{W,H_i} F(W, H_i) \forall i = 1, 2$ *to specified tolerance*

Data*: Spectral data matrices* X_i

While $|\frac{F^{k+1} - F^k}{F^{k+1} - F^0}| \geq 10^{-6}$

Direction of descent: Compute $\nabla_W F(W^k, H_i^k), \nabla_{H_i} F(W^k, H_i^k)$

Compute Modified step size:

$$\eta_W = \frac{\bar{W}}{\sum_{i=1,2} \bar{W} H_i H_i^T + \delta} \qquad \bar{W} = \begin{cases} W & \nabla F_W(W,H_1,H_2) \geq 0; \\ max(W,\sigma) & \nabla F_W(W,H_1,H_2) < 0. \end{cases}$$

$$\eta_{H_i} = \frac{\bar{H}_i}{W^T W \bar{H}_i + \delta} \qquad \bar{H}_i = \begin{cases} H_i & \nabla F_{H_i}(W,H_1,H_2) \geq 0; \\ max(H_i,\sigma) & \nabla F_{H_i}(W,H_1,H_2) < 0. \end{cases}$$

Update:

$$\begin{aligned} W^{k+1} &= W^k - \eta_W^k \nabla F_W^k \\ H_i^{k+1} &= H_i^k - \eta_{H_i}^k \nabla F_{H_i}^k \end{aligned}$$

Since NMF is a non-convex optimization problem, the quality of the solution depends on the initialization of the factor matrices [221]. Either NMF with random initializatons is run a number of times and the best run is selected based on the criterion of lowest Frobenius residual error, i.e. the least value of the objective function to ensure robust and reproducible NMF results, or SVD- based initializations are used [222]. For JNMF using MUR on the normalized FTIR and ^{1}H-NMR data, the Non-negative Double Singular Value Decomposition (NNDSVD) technique is used to initialize the decision variables.

Projected Optimal Gradient approach to JNMF

Solving bound-constrained optimization problems using projected gradients with respect to NMF has been investigated in the literature [223]. Projected gradient methods are shown to converge faster than MUR and have sounder optimization properties [214]. This work seeks to improve the typical project gradient algorithm by using an optimal step size which is updated on each iteration, with application to joint matrix factorization.

The objective function to be minimized is given in Eqn. 3.11. The computation of the gradients given in Eqn. 3.18 is used in the projected optimal gradient algorithm:

Algorithm 2 (Projected Optimal Gradient Algorithm) ***Input****: Initialize decision variables* x=$\{\ W^0, H_i^0\}$

Output*:* x *that solves* $\min_x f(x)$ *to specified tolerance*

Data*: Spectral data matrices* X_i

While*:* $|\frac{f^{k+1}-f^k}{f^{k+1}-f^0}| \geq 10^{-6}$

Direction of descent: $-\nabla_x f(x^k)$

Optimal step size: $\eta_x^k = \arg\min_\eta f(x - \eta\nabla_x f)$

Update: $x^{k+1} = x^k - \eta_x^k \nabla_x f(x^k)$

Projection: $x^{k+1} = max(x^{k+1}, \epsilon)$, *where* $\epsilon = 10^{-6}$

3.2.4 Bayesian Networks

Generating causal structures from the experimental data in the framework of pairwise conditional independence tests between the process variables [224] or causal Bayesian networks constructed on a Markov assumption has facilitated the advanced reasoning of molecular processes in systems biology [225, 226]. Translating the information from chemometric models into knowledge of real chemistry using databases or expert systems that map properties and activities deduced from algorithms to predict candidate molecular structures forms the basis of QSAR/QSPR (Quantitative Structure Activity Relationship/ Quantitative Structure Property Relationship) [227]. The current state-of-the-art methods for the reaction network representation largely rely on encoding prior knowledge and the use of databases : manually encoding reaction rules that generate reaction networks among reactants and products using the Rule Input Network Generator (RING) algorithm, [38] or High Throughput Reaction Prediction (HTRP), which uses link predictions for binary reactions in multi-modal graphs whose feasibility is evaluated using filters implemented by using the Tanimoto similarity score of fingerprints at each reaction node [41]. Hence, there is an imperative need to develop data-driven causal methods [228] to generate hypotheses on reaction networks of complex chemical processes, in the absence of prior knowledge of the process composition or physicochemical/kinetic models, to develop an understanding of molecular-level mechanisms. An attempt in this direction has been made in the current work that focuses on using a data fusion algorithm to extract spectral signatures of the underlying pseudocomponents from heterogeneous measurements before using causal methods to hypothesize their reaction chemistry.

Bayesian networks are a mathematically coherent framework for encoding causal relations as probabilistic graphical models in complex systems [229] as they are robust to handling missing data, combining data with domain knowledge and avoid overfitting [230], thereby resulting in good prediction accuracy in high dimensional

space with fewer samples. They can also be used to build models from noisy experimental data [225]. Bayesian networks consist of nodes of random variables with directed acyclic links among them in accordance with the conditional Markov assumption [231]. Beliefs about the values of random variables are described as probability distributions.

Working with Bayesian networks is a two-step process:

1. Learning the structure of the network that encodes the conditional independence of the random variables represented by the nodes using either score-based or constraint-based methods so that the joint factorization of the directed acyclic graph (DAG) can be expressed as:

$$P(X_1, X_2 \cdots X_N) = \prod_{i=1}^{N} P(X_i | Pa_{X_i}) \tag{3.21}$$

2. Estimation of the parameters of the joint factorization of the conditional probability distribution model of the Bayesian networks from data using an EM algorithm [232] by iteratively calculating the MLE for each of the parameters.

Most analysis tools 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 [193, 225]. In this work, a data fusion algorithm of joint non-negative matrix factorization has been used to identify the spectral signatures of the pseudocomponents. The random variables that comprise absorbance intensities across both measurement techniques of similar compounds are assumed to follow a multinomial distribution with a Dirichlet conjugate prior, as is typically the case with experimental data that is noisy or incomplete [229, 233].

Assume $D = \{X_1, X_2 \cdots X_N\}$ comprises multinominal data which has the following distribution:

$$P(X_i = x|\theta) = \theta_i, \text{where } i = 1, 2 \cdots N \tag{3.22}$$

Equation 3.22 can be used to construct the likelihood function $P(D_k|\theta)$ as a product of the probabilities of the mutually independent random variables that encompass the data. Here, the parameters are $\theta = \{\theta_2, \theta_3 \cdots \theta_N\}$, where $\theta_1 = 1 - \sum \theta_i$ are the parameters that correspond to the physical probabilities of the random variables that have a Dirichlet distribution

$$P(\theta|\beta) = \mathrm{Dir}(\theta|\beta_1, \beta_2 \cdots \beta_N) = \frac{\Gamma(\sum_{i=1}^{N} \beta_i)}{\prod_{i=1}^{N} \Gamma(\beta_i)} \prod_{i=1}^{N} \theta_i^{\beta_i - 1} \tag{3.23}$$

β_i represents a non-negative vector of scaling coefficients referred to as the hyper-parameters of the distribution.

The next step involves algorithms to learn a structure among the nodes (random variables) that encode a probabilistic causal map between clusters of similar wavenumbers that represent a class of pseudocomponents and their chemical interactions during the partial upgrading process, facilitating the generation of reaction hypotheses.

There are three approaches to learning the structure among the nodes [234] : 1) Constraint-based : relying on pairwise conditional independence tests between the random variables [235], which are typically unreliable in high dimensional variable space, 2) Score-based : involving the use of a scoring function that evaluates how well a structure represents the data, which happens to be NP hard as Bayesian networks with N nodes have $2^{\mathrm{O(n)}^2}$ possible structures, making it intractable to score all of them to choose the best one; hence, greedy search methods are used to find the best structure[193, 234], and 3) Bayesian model averaging based: involving the use of an ensemble of possible structures from both the constraint and score-based techniques and then averaging out the prediction of the ensemble, instead of just relying on one best structure [234]. Constraint-based methods are sensitive to individual failure and compromise fit and scalability in the presence of noise [193], which is present

in experimental data, making score-based methods the preferred choice for structure learning in this work.

The most commonly used score is the Bayesian Information Criterion (BIC), which is the posterior probability of a structure given the data, penalized by the the dimension of the structure to favor sparser networks [225, 236]. The BIC is written in terms of the log likelihood of the data, given the structure, which in turn can be expressed in terms of the mutual information and entropy pursuant to the connections among the nodes.

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)} \tag{3.24}$$

Equation 3.24 (Bayes theorem) can then be used in conjunction with the prior distribution to evaluate the log-likelihood function for a certain graph structure which is computed as given in Eqn. 3.26, in terms of the mutual information (Eqn. 3.27) and entropy (Eqn. 3.28). The log likelihood is penalized by the dimension of the graph structure to facilitate sparse networks in the computation of the BIC in Eqn. 3.25

$$\begin{aligned}
\mathrm{BIC(G,X)} &= \mathrm{LL(G,X)} - \frac{\log \mathrm{m}}{2}\,\mathrm{Dim(G)} && (3.25)\\
\mathrm{LL(G,X)} &= \mathrm{m}\sum_i \mathrm{I(X}_i;\mathrm{Pa}_{\mathrm{X}_i}) - \mathrm{m}\sum_i \mathrm{H(X}_i) && (3.26)\\
\mathrm{I(X}_i;\mathrm{Pa}_{\mathrm{X}_i}) &= \sum_{\mathrm{X}_i}\sum_{\mathrm{Pa}_{\mathrm{X}_i}} \mathrm{P(X}_i,\mathrm{Pa}_{\mathrm{X}_i})\,\log\frac{\mathrm{P(X}_i,\mathrm{Pa}_{\mathrm{X}_i})}{\mathrm{P(X}_i)\mathrm{P(Pa}_{\mathrm{X}_i})} && (3.27)\\
\mathrm{H(X}_i) &= -\sum_{\mathrm{X}_i} \mathrm{P(\ X}_i)\,\log(\mathrm{P(X}_i)) && (3.28)
\end{aligned}$$

The algorithms that are typically used to find suitable structural connections to maximize the BIC include Hill climbing [237], Tabu search [238], Maximum minimum hill climbing [239, 240] , simulated annealing and genetic algorithms [234]. Hill climbing involves making locally optimum search iterations with random restarts,

while Tabu search is primarily the same except that it involves penalties on the reversal and repetition of selected moves. MMHC, on the other hand, is a hybrid algorithm that combines constraint-based and score-based methods. Since there is no guarantee that any one of these algorithms by themselves could give 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 is reinforced when more than one of these algorithms return an identical result.

In addition to learning the graph structure among nodes, the strength of the connections calculated based on mutual information (Eqn. 3.27), can be used to generate explanations for reasoning in Bayesian networks. The arc weights are computed as the link strength (LS) along the directional edge between two nodes as the mutual information between the nodes conditioned on the joint distribution of all the other parent nodes:

$$\mathrm{LS}(X \rightarrow Y) = I(X, Y|\mathrm{Pa}_{Y-\{X\}}) \tag{3.29}$$

Eqn. 3.29 can be interpreted as the decrease in the uncertainty of a random variable (node) given its parent, conditioned on the joint probability of all its other parent nodes.

3.3 Results and Discussion

3.3.1 Origin of datasets

The spectroscopic data used in this study were obtained from the experimental investigation of the low temperature thermal cracking of Cold Lake bitumen [196, 197]. Samples of bitumen from the Cold Lake region in Alberta were thermally converted with varying durations of reaction time between 0-8 hours, spanning a range of temperatures between 150°C and 400°C in pressurized micro-batch reactors flushed with nitrogen. The liquid products obtained after conversion were employed to obtain

spectral measurements. Of relevance to the interpretation of the spectral data is that some products may have contained residual solvent, methylene chloride (CH_2Cl_2).

Fourier transform infrared (FTIR) spectroscopic analysis were carried out in an ABB MB3000 equipped with a MIRacle™ Reflection Attenuated Total Reflectance (ATR) diamond crystal plate and pressure clamp. The infrared spectrometer used a deuterated triglycine sulfate (DTGS) detector. The spectra were obtained at a resolution of 2 cm^{-1} as the average of 120 scans over the spectral region 4000-600 cm^{-1}. ^{1}H-NMR spectra were obtained in a Nanalysis 60 MHz NMReady - 60 spectrometer. The equipment was pre-calibrated with deuterated chloroform. For the analysis, 0.15 g of the sample were dissolved in 0.7 μL deuterated chloroform and placed in NMR tubes. The ^{1}H-NMR analyses were performed using the following conditions: 0-12 ppm; number of scans for sample: 32; 14.7 seconds was the average scan time and 4096 points were recorded per scan. A total of 42 FTIR and 32 ^{1}H-NMR spectra were collected, in addition to the measurement at 20°C and 0 min reaction time that was used for the purpose of baseline correction; these have been reported in Table B.1.

3.3.2 Treatment of datasets

The data obtained from the spectral sensors as outlined above have been used to evaluate the performance of JNMF as a data fusion algorithm to generate latent representations among which Bayesian networks are constructed to arrive at chemically meaningful reaction hypotheses. It is assumed that the samples are reacted over increasing intervals of residence times between 1 and 8 hrs at 7 different temperatures (Table B.1), leading to a total of 56 process conditions. The lack of both spectral measurements at any of these process conditions is treated as missing data that is imputed to zeros in the JNMF framework. The JNMF objective as given in Eqn. 3.11 includes parameters α and β; these are weights assigned to co-regularization and manifold regularization terms, respectively, while γ, λ are weights assigned for the sparsity of the shared latent factor W and the unshared latent factors H_1 and

H_2, respectively. These weights are not known *a priori* and a parametric study on these values is performed with the objective of tuning these values to obtain the least reconstruction error computed as follows:

$$E(W, H_1, H_2) = \sum_{i=1,2} P_i * ||X_i - WH_i||_{21} \tag{3.30}$$

It is worthwhile to provide some intuition regarding the regularization terms used in the objective function. The latent factor matrix $H_1 = [h_{1,1}^T; h_{1,2}^T; \cdots h_{1,R}^T] \in \mathscr{R}^{R \times C_F}$ consists of pseudocomponent signatures in the FTIR space, where C_F is the number of wavenumber channels in FTIR, R is the rank and $h_{1,r} \in \mathscr{R}^{C_F \times 1}$ is the FTIR spectrum of the r^{th} pseudocomponent. Similarly, latent factor matrix $H_2 = [h_{2,1}^T; h_{2,2}^T; \cdots h_{2,R}^T] \in \mathscr{R}^{R \times C_H}$ consists of pseudocomponent signatures in the ^{1}H-NMR space. The cosine similarity matrix between absorbances across wavenumbers and chemical shifts is given by $R_{FH} = [r_F^1 r_F^2 \cdots r_F^{C_H}] \in \mathscr{R}^{C_F \times C_H}$, where $r_F^i \in \mathscr{R}^{C_F \times 1}$ is the cross-correlation of absorbances across all FTIR channels with the ith ^{1}H-NMR channel. A qualitative colormap of the cross-covariance among the spectral channels of the two measurement sensors is given in Figure 3.2e. It is believed that using either cross-covariance or cross-correlation would capture structurally identical regions of spectral similarity, as highlighted in Figure 3.2f, implying that they differ only by numerical scaling. In the notation used, the first index in the subscript indicates the spectral sensor, the second index indicates the pseudocomponent and the superscript is the spectral channel. With this in mind, the co-regularization term $H_1 R_{FH} H_2^T \in \mathscr{R}^{R \times R}$ can be written as follows:

$$\begin{bmatrix} \sum_{i=1}^{C_H} h_{1,1}^T r_F^i h_{2,1}^i & \sum_{i=1}^{C_H} h_{1,1}^T r_F^i h_{2,2}^i & \cdots\cdots & \sum_{i=1}^{C_H} h_{1,1}^T r_F^i h_{2,R}^i \\ \sum_{i=1}^{C_H} h_{1,2}^T r_F^i h_{2,1}^i & \ddots & & \\ \vdots & & \ddots & \\ \sum_{i=1}^{C_H} h_{1,R}^T r_F^i h_{2,1}^i & \cdots\cdots & \cdots\cdots & \sum_{i=1}^{C_H} h_{1,R}^T r_F^i h_{2,R}^i \end{bmatrix} \tag{3.31}$$

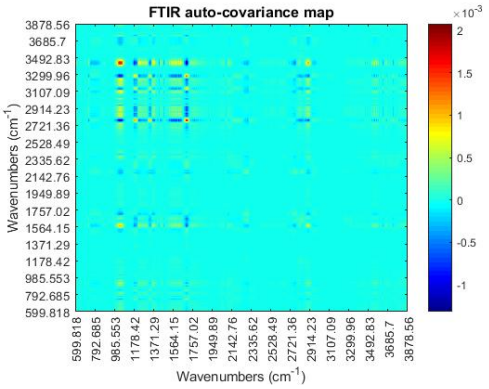
So, the trace of this co-regularization term that is incorporated into the objective given in Eqn. 3.10 is the measure of the similarity between the FTIR and ^{1}H-NMR spectra of a pseudocomponent, which is sought to be minimized for all R pseudocomponents. Using the same intuition, minimizing the trace of the manifold regularization terms $H_1 R_F H_1^T$and$H_2 R_H H_2^T$ involves reducing the similarity in absorbances across spectral channels for the R pseudocomponent spectra in FTIR and ^{1}H-NMR space, respectively. Using the same notation as above, the similarity among absorbances across wavenumbers from the FTIR sensor is given by $R_F = [r_F^1 r_F^2 \cdots r_F^{C_F}] \in \mathscr{R}^{C_F \times C_F}$, where $r_F^i \in \mathscr{R}^{C_F \times 1}$ is the auto-correlation of absorbances across all the wavenumbers with the ith wavenumber. Hence the manifold regularization term $H_1 R_F H_2^T \in \mathscr{R}^{R \times R}$ can be written as follows and can be similarly deduced for $H_2 R_H H_2^T$:

$$\begin{bmatrix} \sum_{i=1}^{C_F} h_{1,1}^T r_F^i h_{1,1}^i & \sum_{i=1}^{C_F} h_{1,1}^T r_F^i h_{1,2}^i \cdots\cdots & \sum_{i=1}^{C_F} h_{1,1}^T r_F^i h_{1,R}^i \\ \sum_{i=1}^{C_F} h_{1,2}^T r_F^i h_{1,1}^i & \ddots\ddots\ddots\ddots\ddots\ddots\ddots & \\ \vdots & \ddots\ddots\ddots\ddots\ddots & \\ \sum_{i=1}^{C_F} h_{1,R}^T r_F^i h_{1,1}^i & \cdots\cdots\cdots\cdots\cdots & \sum_{i=1}^{C_F} h_{1,R}^T r_F^i h_{1,R}^i \end{bmatrix} \tag{3.32}$$

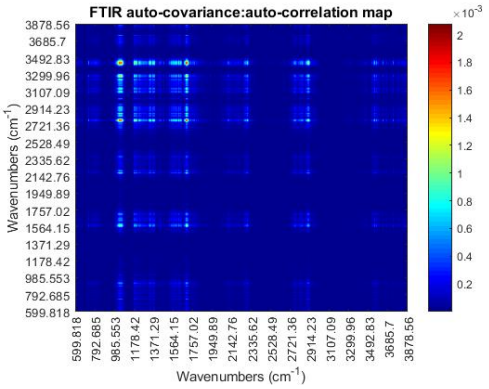
Qualitative colormaps of wavenumber regions that are similar in the FTIR data are indicated in Figure 3.2a and Figure 3.2b, while those for ^{1}H-NMR are given in Figure 3.2c and Figure 3.2d. The inter-spectral similarities among wavenumbers and chemical shifts, across the FTIR and ^{1}H-NMR spectra, are indicated in Figure 3.2e and Figure 3.2f.

The FTIR and ^{1}H-NMR spectra at intermediate residence times of reaction are chosen to compute the above correlation matrices using Eqn. 3.7- Eqn. 3.9. The selected FTIR spectra (150°C-306 min, 200°C-246 min,250°C-246 min, 300°C-306 min) and ^{1}H-NMR spectra (150°C-5 min, 200°C-5 min, 250°C-5 min, 300°C-5 min) are used for the computation.

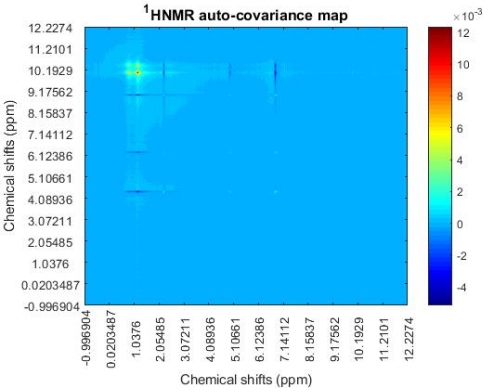
The highly correlated wavenumber pairs from the FTIR heatmap, i.e. correlation



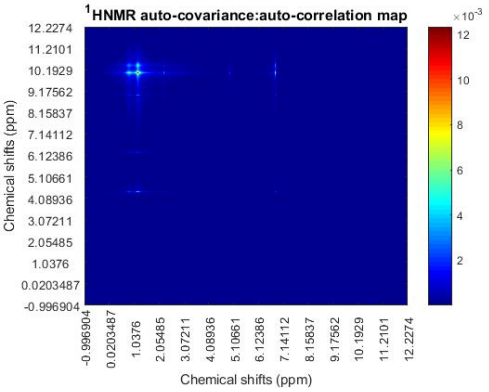
(a) Auto-covariance map among FTIR wavenumbers



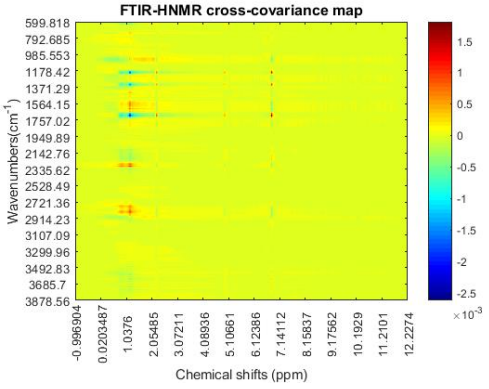
(b) Ratio of auto-covariance: autocorrelation map for FTIR wavenumbers



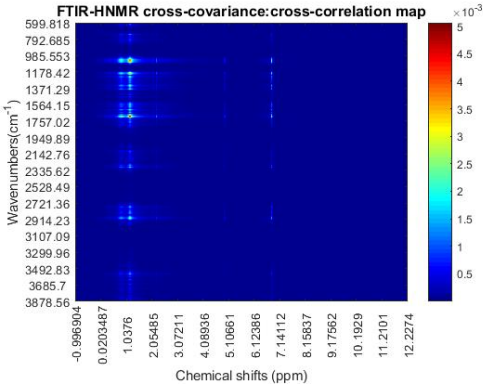
(c) Auto-covariance map among ^{1}H-NMR chemical shifts



(d) Ratio of auto-covariance: autocorrelation map for ^{1}H-NMR chemical shifts



(e) Cross-covariance map among FTIR-^{1}HNMR



(f) Ratio of cross-covariance:cross-correlation map among FTIR-^{1}HNMR

Figure 3.2: Auto-covariance and cross-covariance matrices used to penalize redundancy.

> 0.9 are given in Table B.2 of Appendix B. Some of the key correlations are:

1. 764 and 1159 cm^{-1} indicating meta di-substituted aromatic esters
2. 831 and 1458 cm^{-1} are likely disubstituted aromatics with sp^3 CH bend
3. 3117 and 3194 cm^{-1} are aromatic sp^2 CH stretch
4. 705 and 2962 cm^{-1} showing aromatic ring bending coexists with a terminal methyl group
5. 783 and 3058 cm^{-1} suggested that ortho di-substituted aromatics coexist with alkenes
6. 686 and 1757 cm^{-1} indicated aromatic esters and/or with aromatic anhydrides

The highly correlated pairs of chemical shifts from the ^{1}H-NMR heatmap i.e. correlation > 0.9 are given in Table B.4 of the Appendix. Some of the key correlations are:

1. 0.79 with 0.87 ppm, 0.79 with 0.95 ppm: self-correlated terminal R-CH_3 methyl groups
2. 0.91 with 1.28 ppm : R-CH_2-CH_3
3. 1.81 with 6.16 ppm, 2.25 with 6.28 ppm : ortho di-substituted alkene aromatics
4. 1.12 with 7.59 ppm : aromatics with aliphatic side chains
5. 1.97 with 7.75 ppm, 1.93 with 7.79 ppm, 6.61 with 7.79 ppm, 7.67 with 7.83 ppm, 7.83 with 7.91 ppm : self correlated aromatics

The highly correlated pairs of wavenumbers and chemical shifts from the FTIR and ^{1}H-NMR heatmap, i.e. correlation > 0.7 are given in Table B.6 of the Appendix. Some of the key observed correlations are:

1. 1595 cm^{-1} with 2.29 ppm: Benzylic proton exists with an aromatic C=C stretch
2. 2883 cm^{-1} with 2.29 ppm: Benzylic proton with a methylene group
3. 889 cm^{-1} with 2.3 ppm: meta di-substituted aromatic with benzylic proton
4. 2362 cm^{-1} with 0.956 ppm: thiol/methyl with terminal methyl
5. 2864 cm^{-1} with 1.32 ppm: both represent methylene CH_2 stretch in the middle of an aliphatic chain
6. 1236 cm^{-1} with 5.269 ppm: phenylic alcohol or CH from methylene chloride exists with the solvent peak
7. 1379 cm^{-1} with 5.2 ppm: sp^3 CH bend/ aliphatic CH peak/ methylene content of bitumen which correlates to methylene chloride used as a solvent
8. 1224 cm^{-1} with 7.26 ppm: phenols correlate with aromatics
9. 746 cm^{-1} with 7.42 ppm: ortho di-substituted aromatics correlate with aromatics
10. 2887 cm^{-1} with 6.5 ppm: aromatics with side chains
11. 2925 cm^{-1} with 6.69 ppm: methylene CH
12. 2941 cm^{-1} with 7.06 ppm: benzylic aromatic ie terminal CH_3 with aromatics
13. 865 cm^{-1} with 7.58 ppm: para di-substituted aromatics
14. 746 cm^{-1} with 7.75 ppm: ortho di-substituted aromatics

The rank-based metric based on SVD provides the number of components accounting for most of the variance in the data, which is used as a parameter in the regularized JNMF decomposition. Depending on the weight assigned to the regularizers, they act as a sieve by rendering the latent components with structural similarities from the original space, facilitating meaningful decomposition. The JNMF

objective given in Eqn. 3.11 uses co-regularization to limit the similarity between different spectral sensors of a pseudocomponent and the manifold regularization term to limit the similarity among spectral channels of each sensor for a pseudocomponent. An extensive parametric study was performed to ascertain a reasonable weighting of the regularizers in JNMF while not only seeking to minimize the reconstruction loss and limit overfitting but to also result in chemically meaningful latent spectra. For the co-regularization weight (α) spanning a varied order of magnitude $[0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000] \in \mathscr{W}$, the values of the other regularization weights $(\beta, \gamma, \lambda) \in \mathscr{W}$ resulting in the least reconstruction errors are presented in Table 3.1:

Table 3.1: Parameter values that yield least reconstruction error*

α	β	γ	λ
0	0	10^{-2}	10^{-2}
10^{-3}	0	0	10^{-1}
10^{-2}	0	10^{-1}	10^{-2}
10^{-1}	0	10^{-3}	10^{-1}
1	0	0	1
10	0	10^{-1}	0
10^{2}	0	10^{-2}	1
10^{3}	0	10^{-3}	10

*α and β are weights of the inter-sensor co-regularization and the intra-spectral manifold regularization terms respectively, while γ and λ are weights of the sparsity regularization terms for the shared concentration (W) and the unshared pseudo-component spectra (H_i) respectively.

It can be seen that for any combination of parameter weights, the least reconstruction error is obtained for $\beta = 0$, implying that minimizing similarly correlated absorbance peaks among wavenumbers/ chemical shifts does not result in chemically meaningful pseudocomponent spectra. This makes chemical sense, because different classes of organic compounds would still share common properties related to hydrocarbon structure. For example, a phenol, alkyl aromatic and aromatic ester would all have spectral data that include aromatic C-H.

A tuning curve that depicts the effect of α on reconstruction error based on the above table is shown in Figure 3.3. It can be seen from the figure that the reconstruction error has the least value for $\beta = 0$, $\alpha = 0$, which is a trivial case of deconvolution

without including the regularizers, leading to overfitting that is undesirable. Regularizers are necessary at the expense of reconstruction error to result in chemically meaningful deconvolution, to limit overfitting and to improve the uniqueness of NMF as discussed earlier [204].

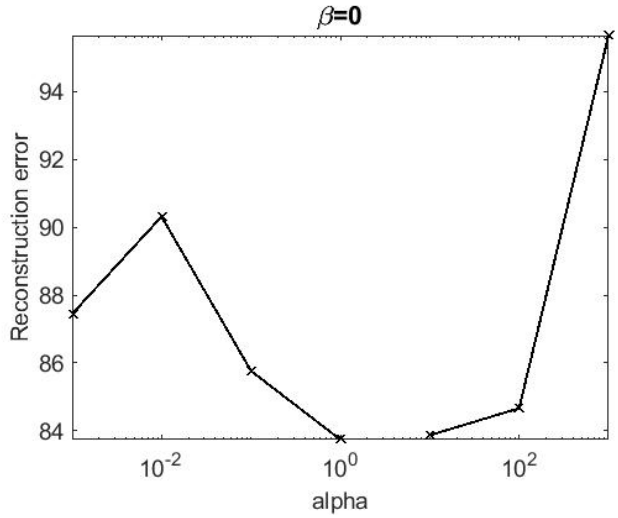


Figure 3.3: Least reconstruction errors over different α values

Keeping this tradeoff in mind, the chosen set of parameters is highlighted in Table 3.1. For values of $\alpha \geq 10$ the reconstruction error increases, so it is undesirable, and it is preferred to pick $\alpha \leq 1$, where the other regularizers are also preferentially weighted and the reconstruction error is not too high. The results of the parametric studies for the chosen α are discussed in Section 3.3.4, while the rest are included in the SI.

3.3.3 Comparison of convergence i.e. Multiplicative Update Rules (MUR) vs Projected Optimal Gradient (POptGrad) Algorithm

A comparison in the performance of the Multiplicative Update Rule (Algorithm 1) and the Projected Optimal Gradient (Algorithm 2), in solving the JNMF objective as given in Eqn. 3.11 using the regularization weights chosen from Table 3.1 on the basis of the tuning curve in Figure 3.3, is illustrated in this section.

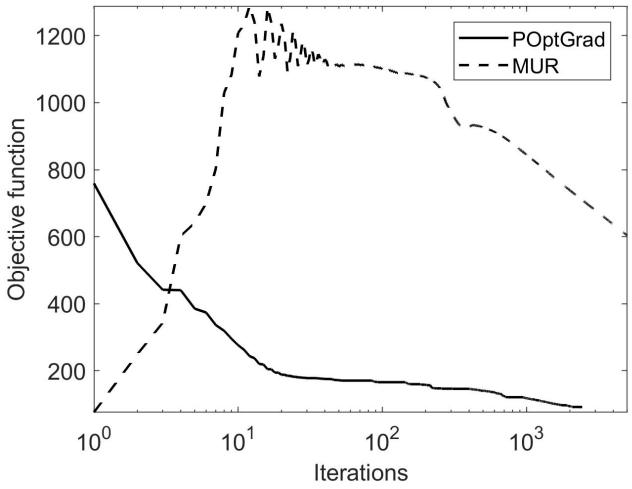


Figure 3.4: Comparison of the convergence of MUR vs Projected Optimal Gradient algorithms

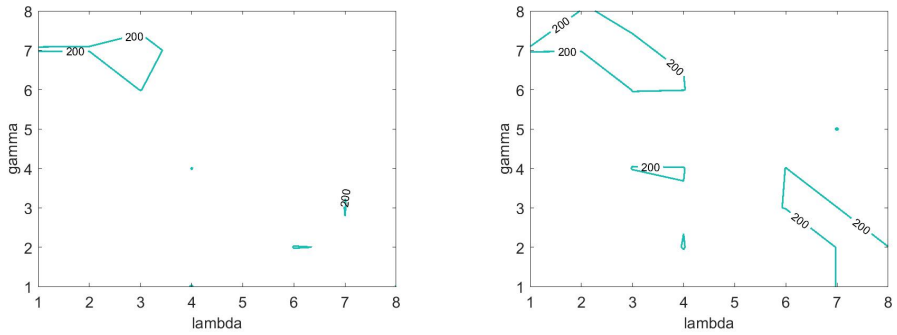
It can be seen from Figure 3.4 that the projected optimal gradient approach to solving the JNMF objective converges faster as compared to the MUR which does not converge to a tolerance of 10^{-6} within 5000 iterations. Since the JNMF objective function is non-convex , it it crucial to have an algorithm that has sound convergence properties that guarantees the stationarity of the local minima indicated by the decision variables. MUR was found to be lacking in the above. The variants of the projected gradient approach stem from the step size, which could be either set to be constant, obtained by using the Armijo rule, or explicity optimizing for the step size in each iteration [214]. The latter, though expensive to optimize, is shown to be an improved variant of the projected gradient approach that is directly applied to NMF with a proof of convergence [214]. Hence, it was extended to be applied directly to solve the JNMF objective in this paper, thereby hastening the convergence of the algorithm in comparison to MUR, which does not converge at the specified tolerance.

3.3.4 Spectral profiles and pseudo-reaction hypotheses based on regularized JNMF

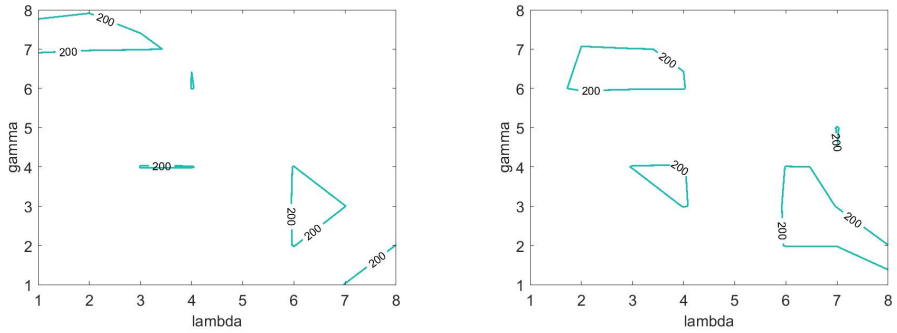
This section includes a detailed discussion on the pseudo-spectral profiles obtained from the JNMF algorithm, using the values of $\alpha = 10^{-1}, \beta = 0, \gamma = 10^{-3}, \lambda = 10^{-1}$ for the weights on the basis of the tuning curve. The isocontours for low reconstruction error for the value of $\alpha = 10^{-1}$ across weights of the other parameters $(\beta, \gamma, \lambda) \in [0, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000]$, is indicated in Figure 3.5. The contours in Figure 3.5e indicate low reconstruction error that is achieved across many more combinations of sparsity weights (γ, λ) for $\beta = 0$, of which $\gamma = 10^{-3}, \lambda = 10^{-1}$ result in the least reconstruction loss. The pseudo-component spectra and the resulting causal pathways over a range of $\alpha \in \mathscr{W}$, for specific sparsity weights at $\beta = 0$ (from the contour plots), is outlined in the Appendix between sections B.3 and B.9.

The Bayesian networks in Figure 3.7 that are constructed with the pseudo-spectra represented in Figure 3.6b as the nodes, have provided interpretable chemistry using expert knowledge, which is discussed in detail. The pseudocomponent concentrations over hourly increasing periods of residence times between 0 and 8 hours, across the 7 temperatures is indicated over the resulting 56 process conditions in Figure 3.6a.

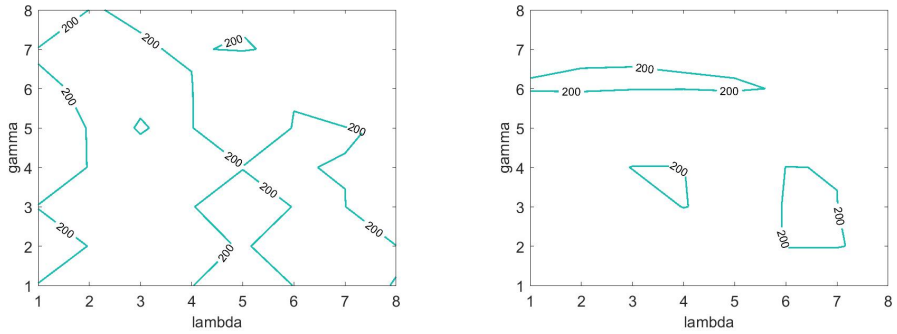
From the spectral signatures given in Figure 3.6b, it can be seen that strong absorption peaks at 2950 cm^{-1}, 2920 cm^{-1}, 2850 cm^{-1} indicating sp^3 C-H stretching are present in all profiles, alongside peaks at 1380 cm^{-1}, 1450 cm^{-1} typical of C-H bending vibrations. Besides these peaks that are common to all PCs, it can be seen in PC_1 contains peaks at 740 cm^{-1}, corresponding to aromatic hydrogen that is coupled with a peak at 1583 cm^{-1} corresponding to an inherently weak C=C stretch, indicating that PC_1 primarily constitutes ortho-substituted aromatics. Likewise for PC_2, a peak at 808 cm^{-1} indicates that it is comprised of meta-substituted aromatics, and PC_3 has a peak at 1720 cm^{-1} coupled with one at 1219 cm^{-1}, indicating that it is composed prominently of esters. This information is used to construct candidate molecules representative of each PC (as given in Figure C.10) and present a plausi-



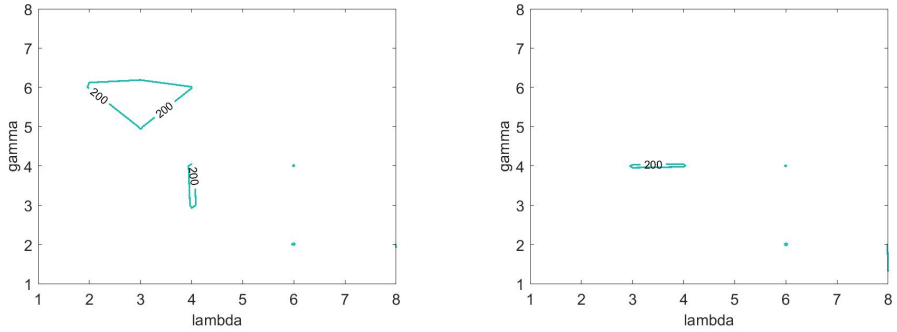
(a) $\beta = 10^{-3}$ (b) $\beta = 10^{-2}$



(c) $\beta = 10^{-1}$ (d) $\beta = 10^{0}$

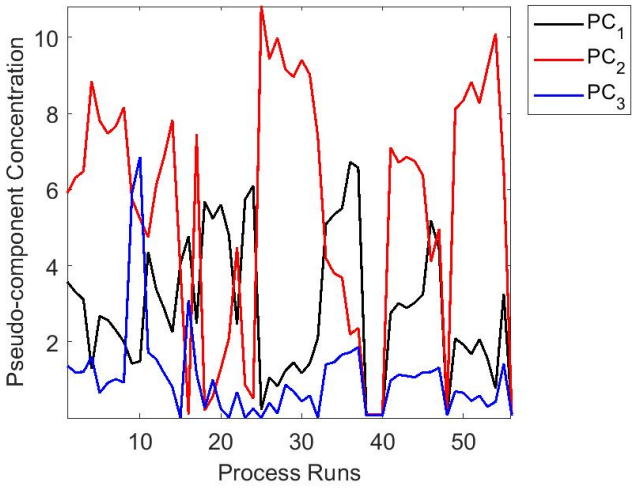


(e) $\beta = 0$ (f) $\beta = 10^{1}$

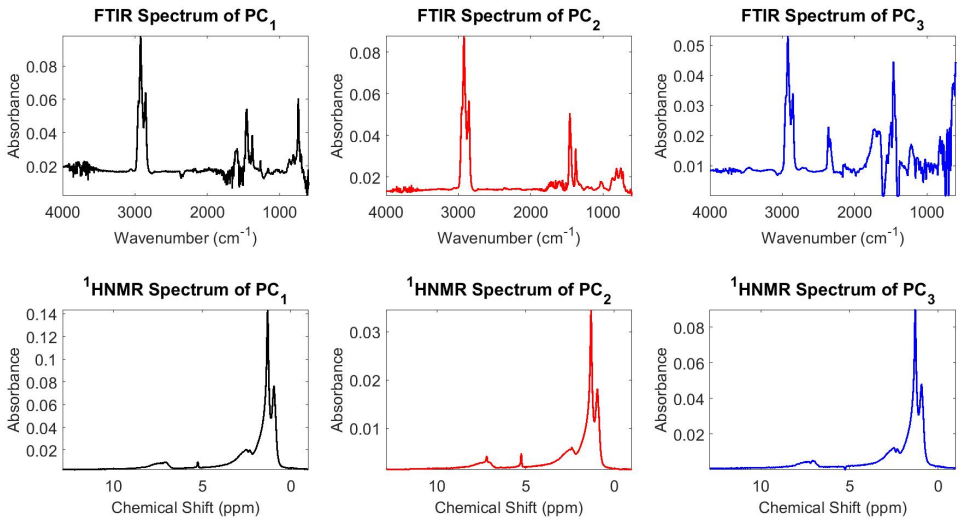


(g) $\beta = 10^{2}$ (h) $\beta = 10^{3}$

Figure 3.5: Isocontours for reconstruction error $E \leq 200$ for $\alpha = 10^{-1}$.



(a) Concentration profiles



(b) Pseudo-component spectra

Figure 3.6: JNMF profiles for $\alpha = 10^{-1}, \beta = 0, \gamma = 10^{-3}, \lambda = 10^{-1}$.

ble interpretation of the reaction chemistry (i.e. pathways) encoded in the Bayesian network of Figure 3.7.

Bitumen has an abundance of free radicals [241]. Hence, given an ortho-substituted aromatic compound (1 in Figure C.10) [242], there is a transfer of free radical hydrogens from the naphthenic ring to other species present in the bitumen. Hydrogen transfer is known to occur even at 150 °C [242], the lowest temperature for which data was included in this study. The transfer of hydrogen leads to the formation of

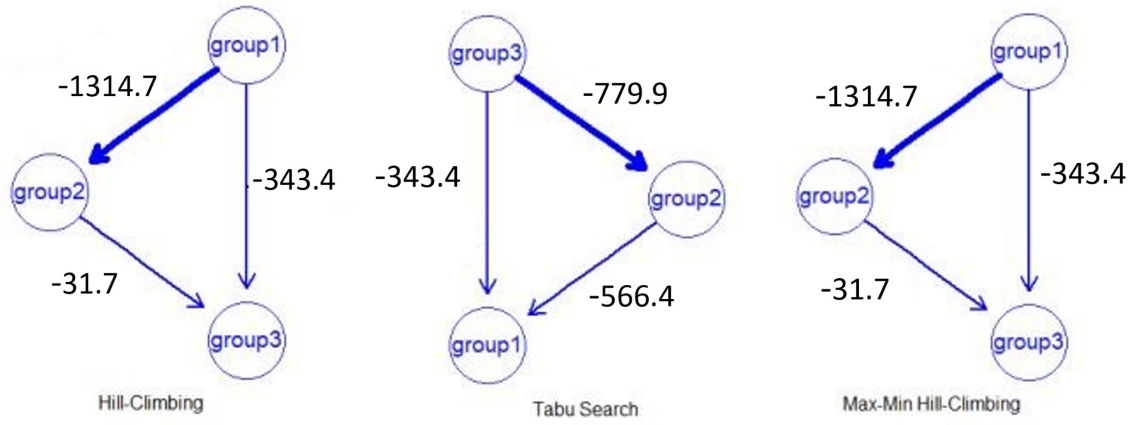


Figure 3.7: Bayesian networks constructed from the PC spectra

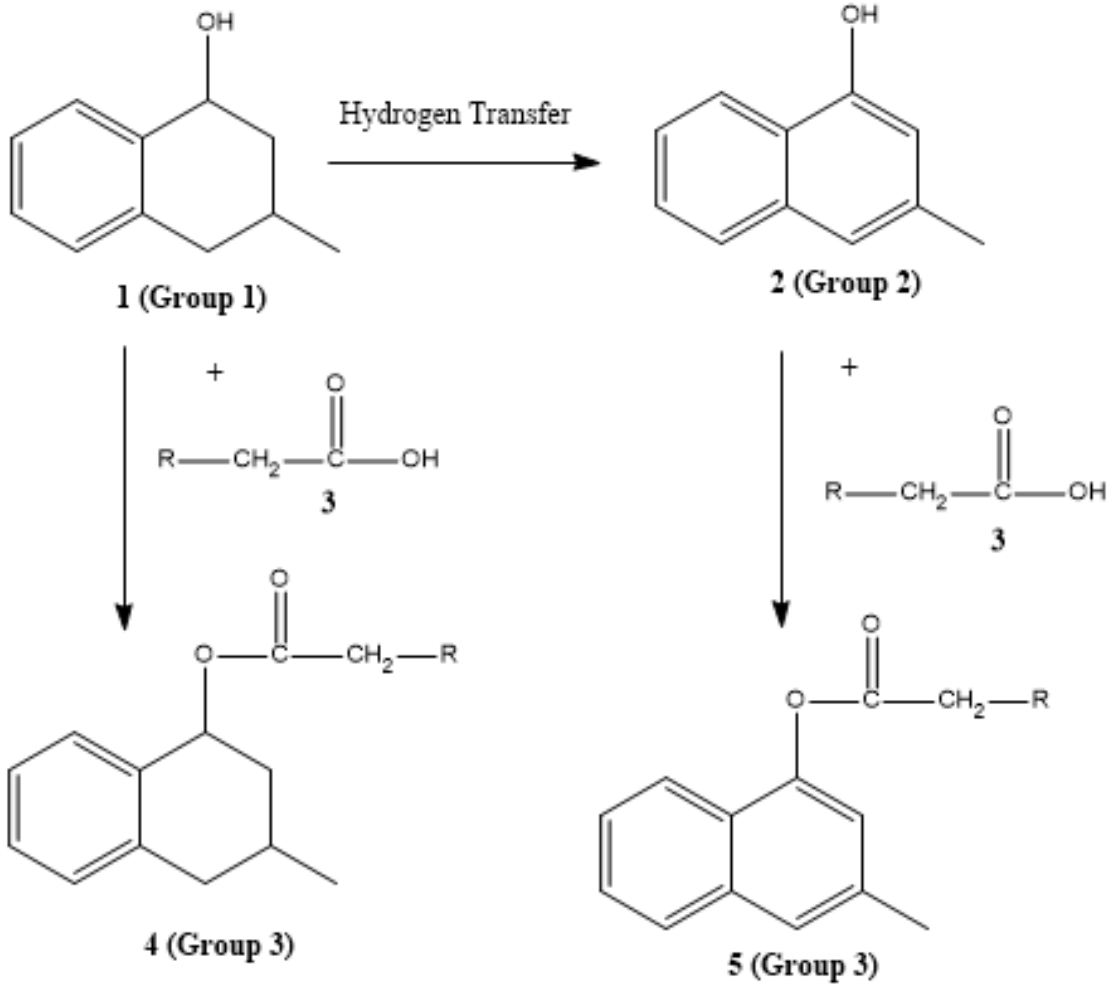


Figure 3.8: Reaction pathways hypothesized from pseudocomponent signatures using domain knowledge.

an aromatic from the naphthenic ring (2 in Figure C.10). In this specific example, the naphthenic ring had meta-substitution (2 in Figure C.10), which was subsequently reflected in the meta-substitution of the aromatic formed by hydrogen transfer. The appearance of aromatic meta-substitution could therefore be explained in terms of hydrogen transfer rather than isomerization. The -OH groups then combine with acids (3 in Figure C.10) to eliminate water, resulting in the formation of esters as a final product (4 and 5 in Figure C.10). This is an equilibrium limited reaction. Under thermal conversion conditions with sufficiently high temperature to vaporize most of the water, the formation of the ester is favored. Phenols are more acidic than naphthenic alcohols, which would have some impact on the esterification reaction, but otherwise the esterification reaction is not affected by the ring to which the alcohol is attached.

3.3.5 Impact of JNMF rank relaxation on spectral deconvolution and pseudo-chemistry

This section is based on using the earlier set of regularization weights in the JNMF algorithm on spectral deconvolution with a higher value of rank. In Section 3.3.4, the number of pseudocomponents is obtained using the statistical notion of 'rank' which has been determined using empirical metrics as discussed in Section 3.2.2, which could have possibly resulted in the loss of a meaningful chemical signal to what was statistically considered as noise. This has been investigated by repeating the JNMF of Section 3.3.4 but successively increasing the rank. It was found that a unit increase in rank resulted in interpretable causal maps from which meaningful chemical pathways were predicted using expert knowledge. A further increase in the rank did not yield chemically meaningful pseudocomponent spectra as it led to noise being assessed as a chemical signature.

The pseudocomponent concentrations across the 56 process conditions of increasing temperature and residence times is given in Figure 3.9, with the corresponding

pseudocomponent spectra indicated in Figure 3.10.

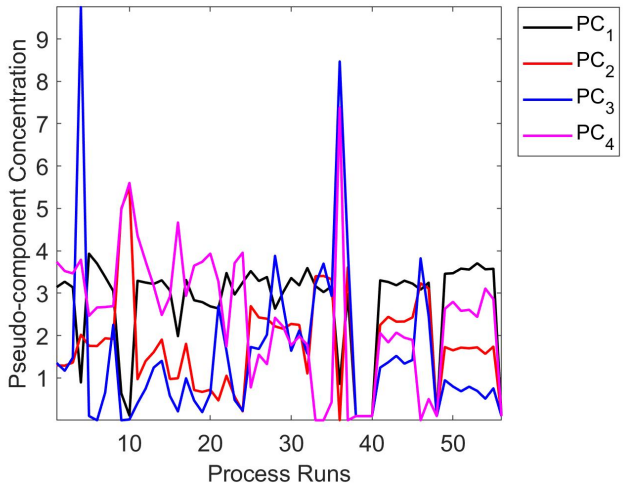


Figure 3.9: Concentration profiles of the pseudocomponents across all the process runs

It can be seen from the pseudocomponent profiles in Figure 3.10 that strong absorption peaks at 2950 cm^{-1}, 2920 cm^{-1}, 2850 cm^{-1} pointing to the sp^3 C-H stretch are present alongside peaks at 1380 cm^{-1}, 1450 cm^{-1} that are typical of C-H bending vibrations in all the PCs. Also, PC_1 has a high intensity peak at 742 cm^{-1}, strongly indicating the presence of ortho substituents, along with weaker absorption peaks at 808 cm^{-1}, 864 cm^{-1} that point to meta and para substituted aromatics. Hence, we may infer that PC_1 mainly consists of substituted aromatics.

PC_2 has absorption peaks at 1657 cm^{-1}, 817 cm^{-1} that correspond to the C=C stretch and C-H bend of olefins. Hence, we may assume that PC_2 is primarily olefinic. PC_3 is seen to have a high intensity peaks for orthogonal substitutions at 740 cm^{-1} that runs alongside a peak at 1607 cm^{-1} corresponding to the aromatic C=C stretch, indicating that this pseudocomponent may contain naphthene aromatic compounds or ortho-substituted aromatics. PC_4 exhibits a C=O stretch at 1705 cm^{-1} along with a peak at 1219 cm^{-1} indicative of a C-O stretch, thus enabling us to deduce that this class of compounds is comprised largely of esters or anhydrides.

Figure 3.11 shows the Bayesian networks obtained using the the pseudocomponent

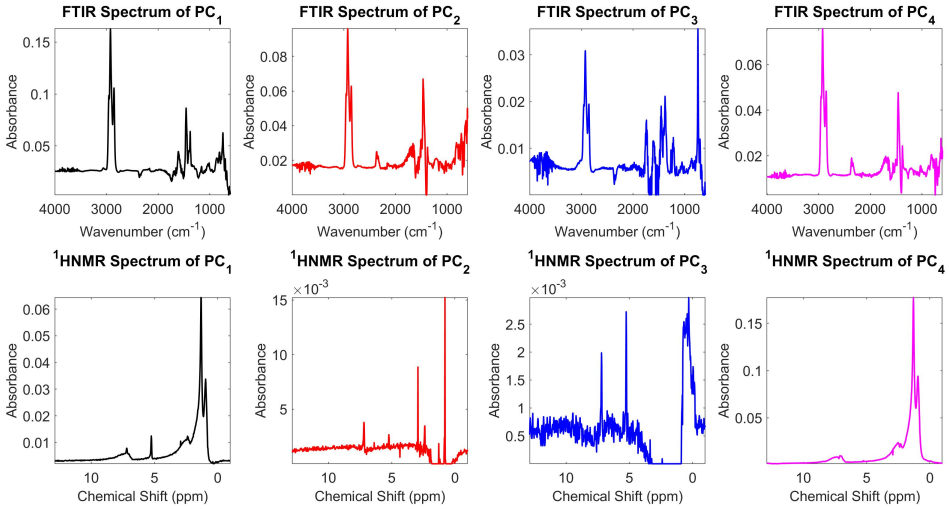


Figure 3.10: Pseudo-component spectra obtained by relaxing the JNMF rank using optimal regularization weights as obtained from the tuning curve.

signatures from Figure 3.10 as nodes.

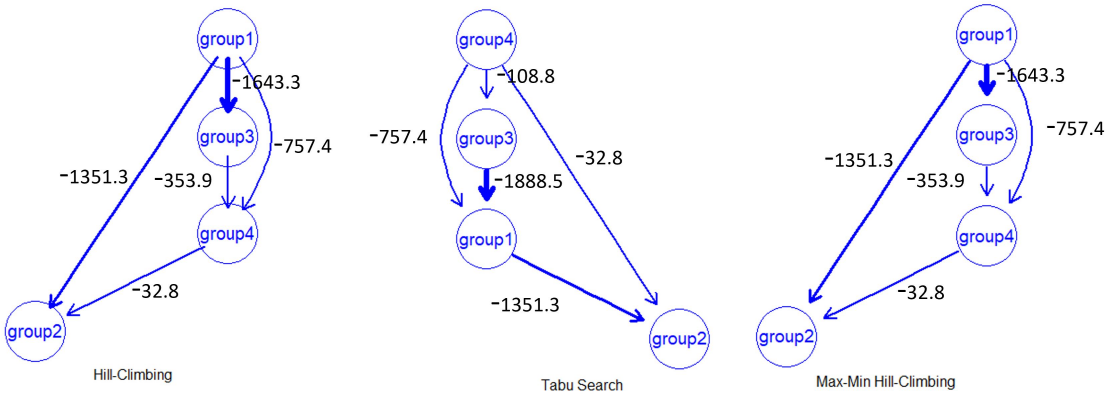


Figure 3.11: Bayesian networks constructed from the PC spectra obtained with rank relaxation.

The Bayesian network of Figure 3.11 is consistent with the reaction chemistry described in Figure C.11.

As in Figure C.10, we use the same type of candidate molecule to represent PC_1 (substituted aromatics while discussing chemical pathways shown in Figure C.11). It can be seen that substituted aromatics thermally crack to form olefins (7 in Figure C.11), or their -OH groups could also combine with acids (3 in Figure C.11) present

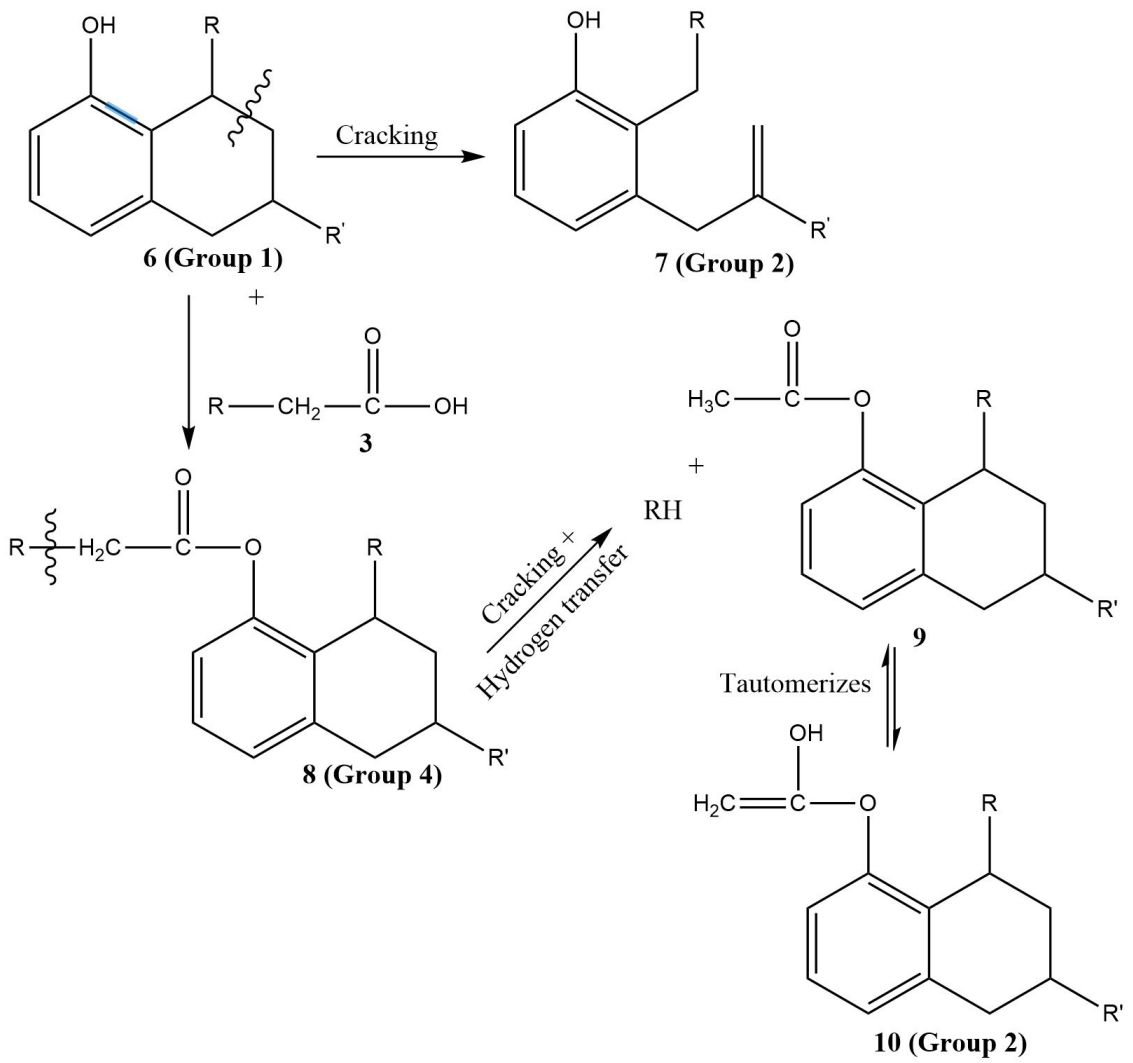


Figure 3.12: Reaction pathways to indicate conversion paths among substituted aromatics, anhydrides and olefins.

to give esters (8 in Figure C.11) that could undergo further cracking to produce a compound (9 in Figure C.11) that undergoes keto-enol tautomerization to produce olefinic compounds (10 in Figure C.11). *Note that the terminology 'group' in Figure C.11 and subsequent figures corresponds to the PC.*

We use a different candidate molecular structure (11 in Figure C.12) to represent PC_1 when the substituent to the aromatic ring is naphthenic. Due to the free radical transfer mechanism [242], a hydrogen disproportionation reaction leads to the formation of a naphthene aromatic-like compound (12 in Figure C.12). The -OH group

attached to the aromatic benzene ring could subsequently react with hydrocarbon acids (3 in Figure C.12) to form esters (13 in Figure C.12).

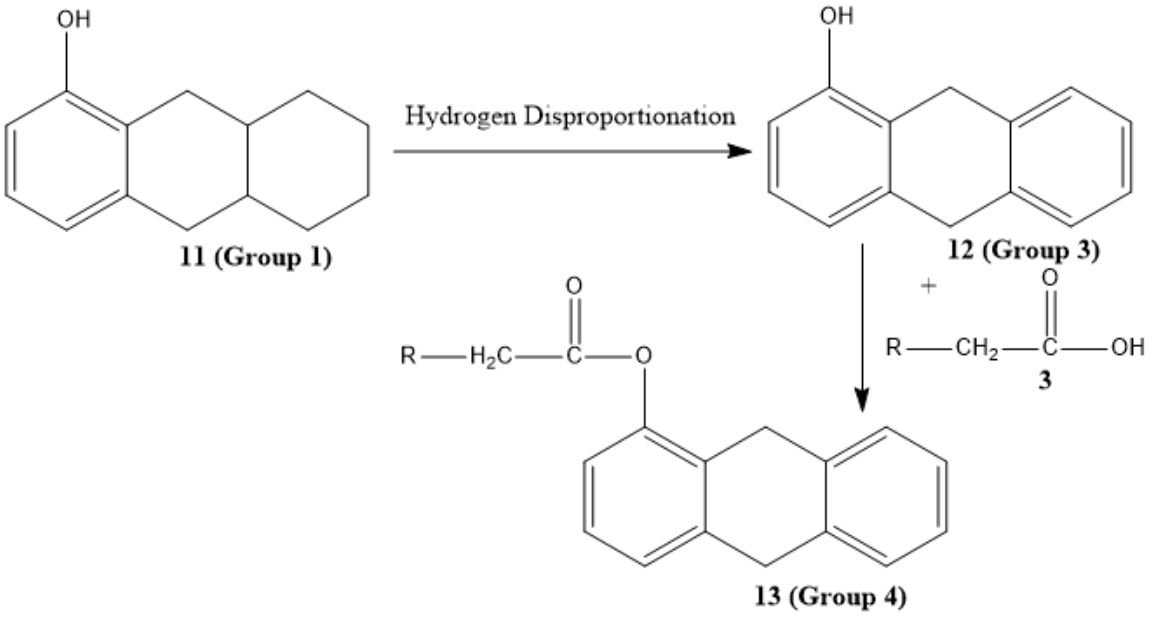


Figure 3.13: Reaction pathways for conversion paths among substituted aromatics, naphthene aromatics and anhydrides.

Since the path going from substituted aromatics to naphthene aromatics has the strongest arc strength (Figure 3.11), another candidate molecule is used to represent PC_1 to highlight the possibility that meta and para substituted aromatics could convert to ortho-substituted aromatics that show similar peaks as a naphthene ring attached to an aromatic (Figure C.13).

This is consistent with our analysis of representative compounds of each pseudo-component upon analyzing their signatures, where we asserted that PC_3 indicates the presence of ortho substitutions, which could either be a napthenic ring as shown in Figure C.12 or be straight chain substituents in the ortho positions as shown in Figure C.13.

It should be emphasized that the hypothetical molecular structures shown in Figures C.11-C.13 are plausible substructures of species commonly found in bitumen [243]. Although it is unlikely that the specific molecules shown would represent the

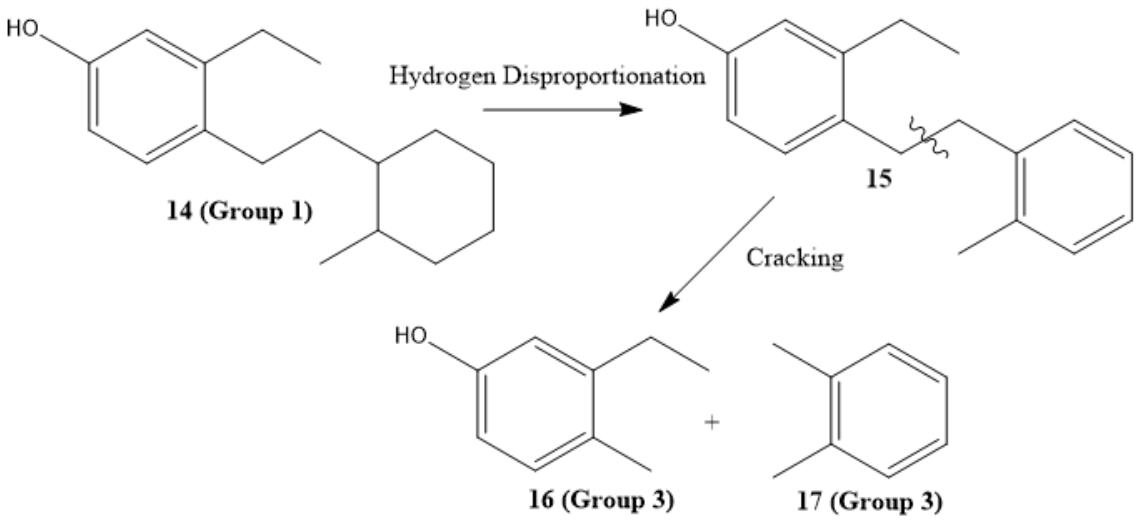


Figure 3.14: Reaction pathways for conversion of meta, para substituted aromatics to ortho substituted aromatics.

spectral data, the substructures present as part of many different and heavier compounds have the spectral features that make the reaction chemistry related to the spectroscopic data plausible.

3.3.6 Spectral profiles and chemical pathways using JNMF with orthogonally weighted manifold regularization

The manifold regularization terms used earlier ($H_1 R_F H_1^T, H_2 R_H H_2^T$) incorporated an auto-correlation term that weights similarly correlated spectral channels. In this section, we investigate spectral profiles and reaction pathways stemming from the JNMF pseudocomponent signatures with a modified manifold regularization term where the spectral channels are asserted to be uncorrelated among themselves ($H_1 I_F H_1^T, H_2 I_H H_2^T$), i.e. absorbance at each wavenumber/ chemical shift is perfectly similar to itself and does not correlate with values across other wavenumbers/chemical shifts; this implies orthogonality among spectral channels. This translates to replacing $R_F = I_F, R_H = I_H$ which simplifies the manifold regularization term in Eqn. 3.32 to the following diagonal matrix, minimizing the trace of which is equivalent to favoring sparser representations, because of which JNMF is performed with $\gamma = 0, \lambda = 0$.

$$\begin{bmatrix} h_{1,1}^T h_{1,1} & 0 \cdots\cdots\cdots 0 \\ 0 & \ddots \\ \vdots & \ddots \\ \vdots & \ddots \\ 0 \cdots\cdots\cdots\cdots & h_{1,R}^T h_{1,R} \end{bmatrix}$$

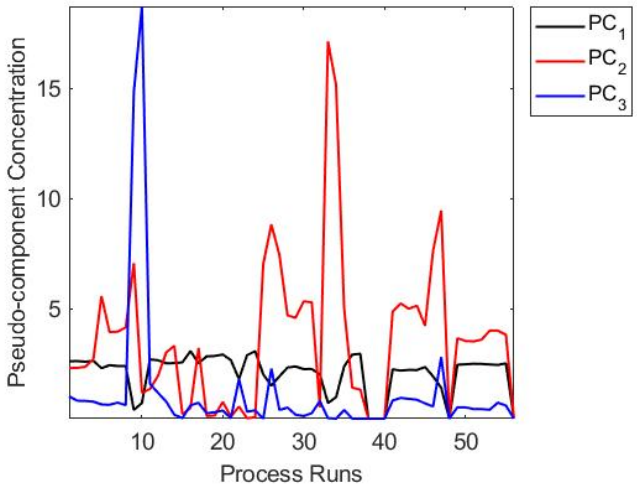
The co-regularization term continues to use the cross-correlation among wavenumbers and chemical shifts as a weight of similarity. The sparsity regularizers are not considered in the objective given in Eqn. 3.33, as the use of an identity matrix in manifold regularization is equivalent to forcing sparsity in the pseudo-spectra.

$$\begin{aligned} \min_{W,H_1,H_2 \geq 0} F(W,H_1,H_2) &= \sum_{i=1,2} P_i * ||X_i - WH_i||_{21} \\ &+\alpha||H_1 R_{FH} H_2^T||_{21} + \beta||H_1 I_F H_1^T + H_2 I_H H_2^T||_{21} \end{aligned} \tag{3.33}$$

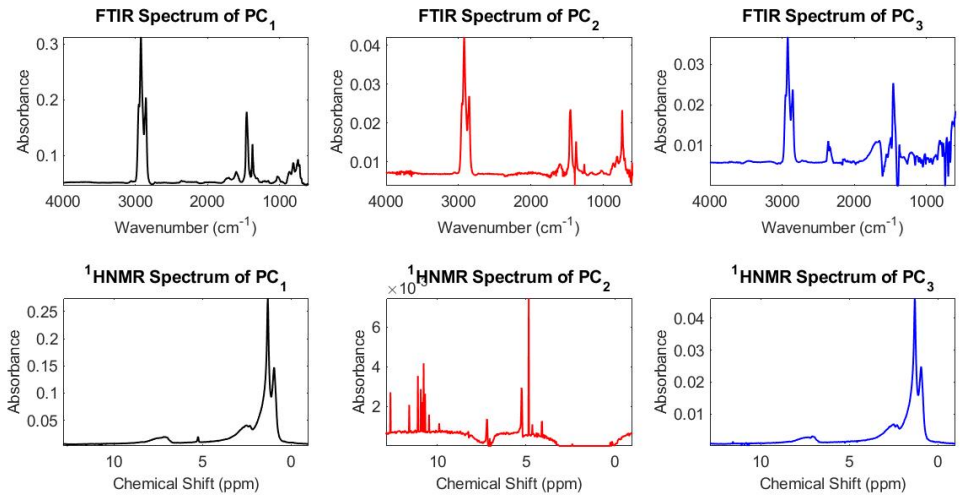
The pseudocomponent concentrations across the 56 process conditions of increasing temperature and residence times is given in Figure 3.15a, with the corresponding pseudocomponent spectra indicated in Figure 3.15b.

It can be seen from the pseudocomponent profiles in Figure 3.15b that strong absorption peaks at 2950 cm^{-1}, 2920 cm^{-1}, 2850 cm^{-1} pointing to the sp^3 C-H stretch are present alongside peaks at 1380 cm^{-1}, 1450 cm^{-1} typical of C-H bending vibrations in all the PCs. Besides these peaks, it can be seen that PC_1 has prominent peaks at 740 cm^{-1}, 1724 cm^{-1} and 1603 cm^{-1} that correspond to aromatic C-H bending vibrations with ortho substitutions, leading us to conclude that ortho substituted phenyl esters are the most appropriate model compounds for PC_1. PC_2 has a strong peak at 740 cm^{-1} indicating that it is mainly comprised of ortho substituted aromatics or naphthene aromatic compound structures. PC_3 has peaks at 1610 cm^{-1} and 1203 cm^{-1} that correspond to the C=C aromatic stretch and the phenolic C-O stretch indicating that this pseudocomponent encompasses phenols and aliphatics.

The Bayesian networks that have been constructed using the pseudo-spectra from Figure 3.15b using the heuristic score search methods outlined earlier are given in Figure 3.16.



(a) Concentration profiles of the pseudocomponents



(b) Pseudo-component spectra

Figure 3.15: Pseudocomponent spectral profiles with orthogonally weighted manifold regularization, $\alpha = 10^{-2}, \beta = 1$

An ortho-substituted phenyl ester (18 in Figure 3.17) is chosen as a candidate compound for PC_1 in the hypothesis of plausible reaction pathways corresponding to the Bayesian networks of Figure 3.16; the pathways of which are shown in Figure 3.17.

Since the arc strengths of the reactions paths between $PC_1 \rightarrow PC_2$ and $PC_1 \rightarrow PC_3$ are relatively stronger (Figure 3.16), another model compound, i.e. ortho susb-

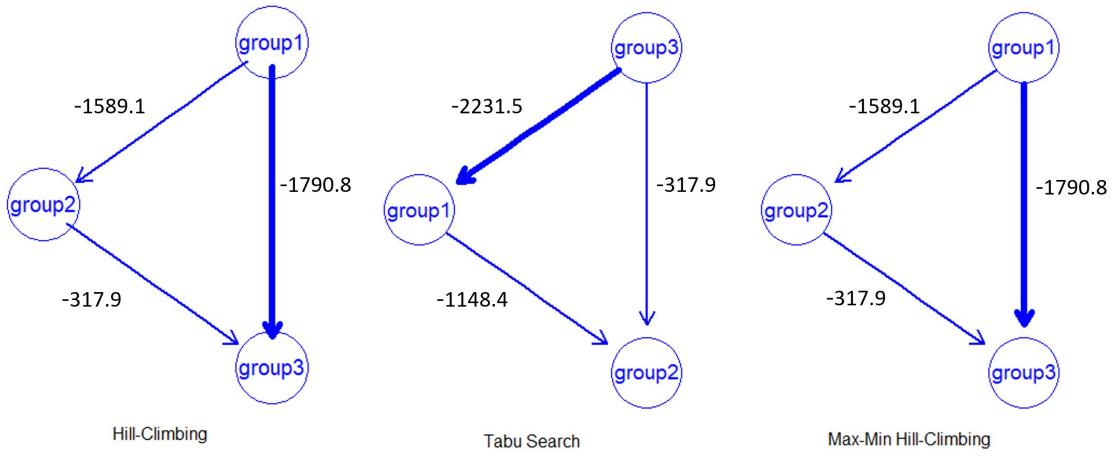


Figure 3.16: Bayesian networks constructed from the PC spectra.

stituted phenyl carboxylate (26 in Figure 3.18) is used to represent PC_1, the cracking of which produces ortho-susbtituted aromatics (28 in Figure 3.18), followed by the hydrolysis of the by-product, phenyl carboxylate (27 in Figure 3.18), to give phenols (20 in Figure 3.18) and carboxylic acids (29 in Figure 3.18). This alternate hypothesis is depicted in Figure 3.18.

3.3.7 Discussion: Impact of the correlation-based regularization in JNMF on the hypothesized reaction mechanisms

The correlations that are found to exist within the FTIR wavenumbers and ^{1}H-NMR chemical shifts themselves are just another tool to support the possible existence of the molecules in the proposed reaction chemistries in Sections 3.3.4, 3.3.5 and 3.3.6. Since FTIR indicates only the presence of functional groups and ^{1}H-NMR indicates the presence of different types of protons and not the actual molecules, only representative compounds for each pseudocomponent can be deciphered or proposed. Each pseudocomponent has its own spectral profile calculated using the JNMF methods and these intra-spectral and inter-sensor correlations give further evidence of the presence of a certain type of compound class. In general, FTIR-FTIR correlations are more useful than NMR-NMR correlations because the type of aromatic substitution

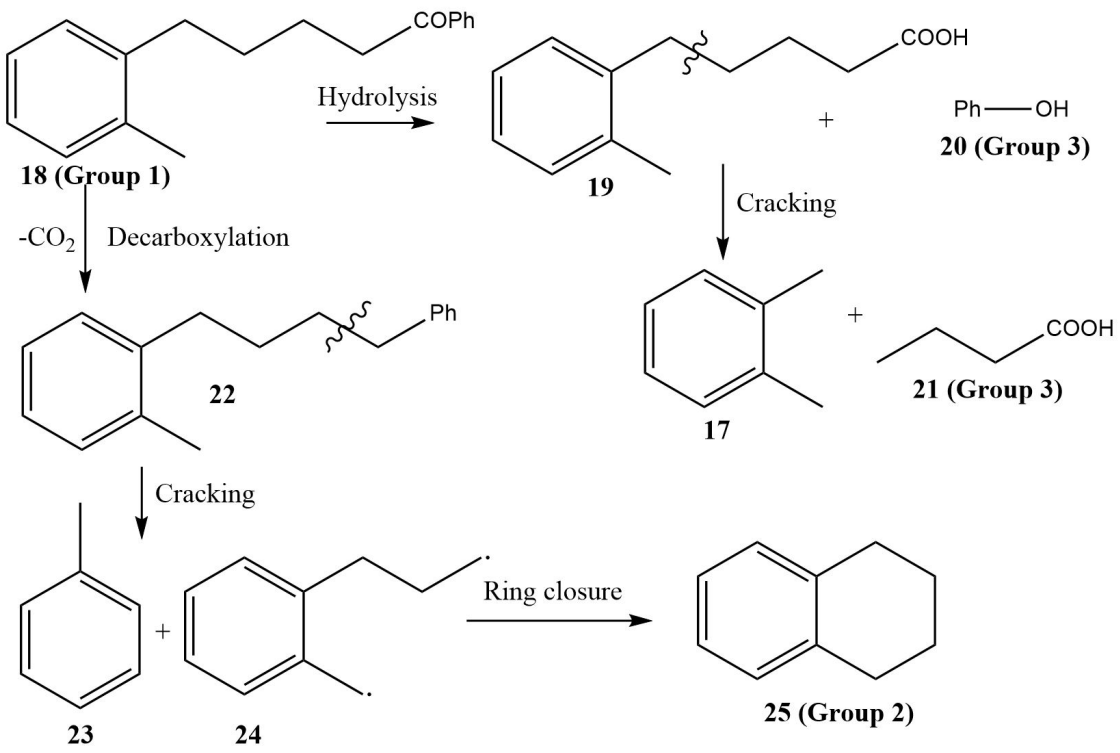


Figure 3.17: Reaction pathway hypothesis under orthogonal manifold regularization.

is more pronounced in FTIR rather than in ^{1}H-NMR where it is just an overlapped peak over a range of chemical shifts. The sensitivity of the instrument used was insufficient to identify ortho-, meta- or para substituted aromatics. Another drawback of the NMR-NMR correlations was that hydroxyl and phenolic protons could not be indicated separately, whereas they could be identified in FTIR-NMR inter-correlations.

First, let us consider FTIR-FTIR correlations as outlined in Section 3.3.2. It is important to note that the aromatic bending vibrations for C-H bonds also overlap with C-H bending absorptions for olefins. But since the concentration of olefins is lesser than that of aromatics, we focus on the aromatic vibrations in the 690 - 900 cm^{-1} region. Meta di-substituted aromatic esters are indicated by the correlation between wavenumbers at 763 and 1159 cm^{-1}, and this is considered as a representative molecule belonging to group 3 in the reaction chemistry proposed for the first type of analysis (Section 3.3.4 and Figure C.10). The next correlation indicates sp^3 C-H

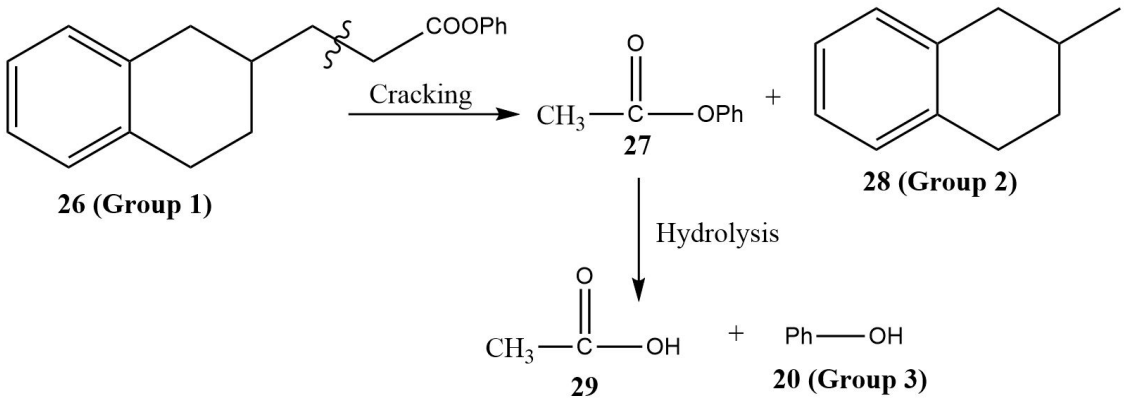


Figure 3.18: Alternate reaction hypothesis under orthogonal manifold regularization.

bend along with aromatic sp^2 C-H bend, and points towards aromatics with aliphatic side chains. These types of molecular sub-structures are very common in bitumen [243] and its cracked products and are presented in Figure C.10, Figure C.11, Figure C.12, Figure C.13 and Figure 3.17. Aromatic sp^2 C-H stretches are weak but their presence further indicates evidence of aromatics involved in the reactions. Toluene-like species are indicated by correlation between aromatic ring bending and terminal methyl groups and are shown to exist in group 2 of Figure C.10 where one of the aromatic substituents is a methyl group. The presence of ortho disubstituted aromatics along with olefinic side chains is clearly indicated in group 2 in Figure C.11. Bitumen does not contain much olefins, but they can be formed through cracking as it is a free radical mechanism. However, ^{1}H-NMR evidence suggests that the concentration of olefins is still low in cracked products, probably due to hydrogen transfer in the liquid phase. Finally, the correlation between aromatic ring bending and ester/anhydride-type carbonyl stretches are indicative of the presence of aromatic esters that are final products in the first two sets of reaction networks (Figure C.10 and Figure C.11, Figure C.12) and the starting molecule in the third reaction network proposed in this work (Figure 3.17).

The ^{1}H-NMR correlations between the chemical shifts, as identified in Section

3.3.2, provide less information on their own than the FTIR correlations because there is more overlap in the aromatic region. Though the nature of substitution cannot be deciphered clearly, the correlation between 1.12 and 7.59 ppm indicates the presence of substituted aromatics. The presence of methyl and methylene groups in aliphatic side chains are also indicated by the first two correlations. Another interesting correlation is the one between 1.97 and 7.75 ppm, which points towards the existence of a benzylic proton. Most of the species in the proposed reaction chemistry have a benzylic proton which makes this correlation important.

Perhaps the most support for the hypothetical molecules or sub-structures involved in the proposed reaction networks in this work is provided by the correlations between FTIR wavenumbers and ^{1}H-NMR chemical shifts. If the R group is longer than 1 carbon in Figure C.11, the aromatic ester and phenolic olefin shown in Figure C.11 are supported by the second correlation indicating a minimum of 1 benzylic proton with a methylene group. The existence of meta substituted aromatics was already supported by FTIR-FTIR correlations and is further backed by the correlation with the benzylic proton shift in the NMR. Aliphatic side chains are an important component of bitumen cracking feedstock as the phenyl carboxylate shown in Figure 3.18 and the presence of mid-chain methylene groups is indicated by the correlation of -CH2- stretch in FTIR around 2864 cm^{-1} and the methylene proton shift in NMR. They are also supported by the 10^{th} cross-correlation where NMR gives the shifts for aromatic protons and FTIR indicates the methylene C-H stretches. For the formation of esters as final products (as in sections 3.3.4 and 3.3.5), phenolic compounds are probable starting materials (Figures C.11 and C.12) and phenols can also be formed through ester decomposition (Figure 3.17). The correlation between C-O stretch at 1224 cm^{-1} and aromatic shift in NMR clearly supports the involvement of phenols in bitumen thermal chemistry. ortho disubstituted aromatics that mostly manifest themselves in the form of naphthene aromatics are probably the most abundant in bitumen feed and consequently become an important constituent of the reaction network. Their pres-

ence is indicated by multiple correlations between the aromatic region in the NMR and C-H bend for ortho disubstituted aromatics in FTIR (correlations 9 and 14).

Certain correlations identified are not captured in the reaction chemistry like the para-disubstituted aromatics (13^{th} correlation) and the thiol group correlated with terminal methyl (4^{th} correlation). Nevertheless, the overall significance of the correlations in the proposed reaction networks deduced with the help of Bayesian methods is profound.

3.4 Conclusions

Robust weighted JNMF with graph regularization has been demonstrated as a semantic meaning-based framework for information fusion from multiple spectral sensors with an application to extracting pseudocomponent spectra by curve resolution. The semantic meaning arises from using regularization as a similarity metric to model the graph structure within and across spectral sensors, while limiting overfitting and solution ambiguities of JNMF, alongside yielding chemically meaningful pseudocomponent spectra with minimal reconstruction loss. The parametric study of the regularization weights have revealed the intra-spectral regularization term (β) to be weighted by zero to obtain chemically meaningful pseudocomponent spectra. This is corroborated by the fact that different organic compounds share similar properties with respect to their hydrocarbon structure. Hence, minimizing similarly correlated peaks across spectral channels of a sensor, results in pseudocomponent spectra that poorly reconstruct the original data and lack chemical meaning.

The projected optimal gradient algorithm which has been developed to solve the JNMF objective, is seen to converge within fewer iterations at the specified tolerance as compared to the multiplicative update rule algorithm. The resulting pseudocomponent spectra represent the latent features extracted in the domain of each of the spectral sensors, among which directed acyclic causal pathways are learned using Bayesian structure learning. This probabilistic approach to reaction hypotheses gen-

eration has been validated by domain knowledge. The reaction hypotheses are seen to depend on the parameters of JNMF : chemical rank and regularization weights. The number of pseudocomponents is empirically determined using the notion of chemical rank and determines the number of nodes in a Bayesian network. The weights assigned to the regularization terms impact the peaks in the pseudocomponent spectra, as does the value of chemical rank.

The Bayesian networks constructed with optimal weights of the regularization terms and the empirically determined rank, were seen to hypothesize mechanisms involving the conversion of substituted aromatics to esters and anhydrides. However, for the same optimal weights, heuristically relaxing rank to keep in check the truncation of chemical information to noise, revealed additional pathways of substituted aromatics, esters and anhydrides further decarboxylating to give olefins as the final product. Additionally, it is shown that different reaction hypotheses are generated when the similarity metric used in the manifold regularization term was changed to an identity matrix, where ortho susbtituted phenyl esters through hydrolysis and cracking produce phenols and aliphatics as the end products. This indicates that the parameters of JNMF regulate the peaks that appear in the pseudo-spectra across multiple sensors, that are later represented as nodes of random variables among which the structure of Bayesian networks is learned to probabilistically hypothesize reaction mechanisms among the nodes. The demonstrated use of statistical methods for latent feature extraction followed by causal structure inference has lead to the deployment of data-driven system inferential models to demystify the hitherto unknown chemical reaction pathways; proving vital for the real-time process monitoring and control of complex reacting mixtures.

Chapter 4

Structure-preserving joint non-negative tensor factorization to identify reaction pathways using Bayesian networks

Abstract

Extracting meaningful information from spectroscopic data is key to species identification, as a first step to monitoring chemical reactions in unknown complex mixtures. Spectroscopic data collected over multiple process modes (temperature, residence time) from different sensors (Fourier Transform Infrared (FTIR), Proton Nuclear Magnetic Resonance (^{1}H-NMR)) comprises hidden complementary information of the underlying chemical system. This work proposes an approach to jointly capture these hidden patterns in a structure-preserving and interpretable manner using coupled non-negative tensor factorization to achieve uniqueness in decomposition. Projections onto the modes of spectral channels, specific to each sensor, are interpreted as pseudo-component-component spectra, while projections onto the shared process modes are the corresponding pseudo-component concentrations across temperature and residence times. Causal structure inference among these pseudo-component spectra (using Bayesian networks) is then used to identify plausible reaction pathways

This chapter has been published as: A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, V. Prasad. Structure-Preserving Joint Non-negative Tensor Factorization to Identify Reaction Pathways Using Bayesian Networks. *J. Chem. Inf. Model.* **2021**, *61*, 12, 5747-5762.

among the identified species representing each pseudo-component. Tensor decomposition of the FTIR data enables the development of reaction sequences based on the identified functional groups, while that of the ^{1}H-NMR by itself is lacking in mechanism development as it solely reveals the proton environments in a pseudo-component. However, jointly parsing spectra from both the sensors is seen to capture complementary information, wherein insights into the proton environment from ^{1}H-NMR disambiguates pseudo-components that have similar FTIR peaks. A scalable method of parallelizing tensor decomposition to handle high dimensional modes in process data by using grid tensor factorization, while being robust to process data artefacts like outliers, noise and missing data, has been developed.

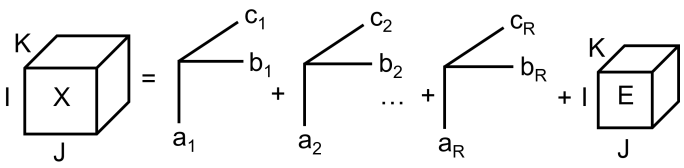
4.1 Introduction

The extensive use of sensors to relay analytical measurements collected as multi-way/multi-modal data is prevalent in neuroscience [244], signal processing [245], chemometrics [32], social network analysis [246], metabolomics, text mining and computer vision [247], because of which tensor decompositions are an imperative tool for exploratory analysis involving factor analytical decompositions as it captures the inter modal interactions among the latent factors across modes. In metabolomics, heterogeneous data from different sensors like Nuclear Magnetic Resonance (NMR), Liquid chromatography–mass spectrometry (LC-MS) and Fluorescence spectroscopy (FS) have been jointly analyzed in terms of shared and unshared factors in the framework of structure-revealing data fusion so that complementary information about biomarkers from different sensors are fused to obtain physically interpretable latent factors corresponding to the biomarker, facilitating disease characterization. [248] Coupled decomposition of a tensor and matrix, through a common shared factor is shown to be structure-preserving and unique as compared to separate decomposition [249]. This work seeks to implement data fusion through a structure-preserving framework of joint tensor decomposition, to simultaneously analyze process data during the partial

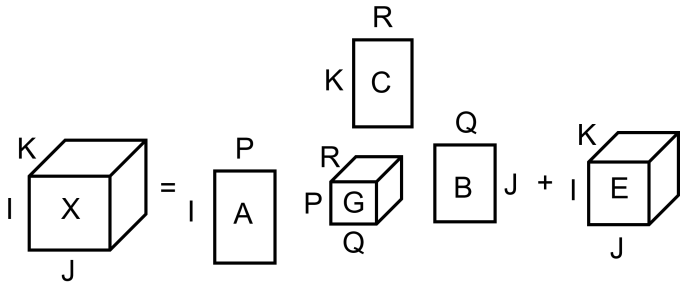
upgrading of Cold Lake bitumen from 2 spectral sensors : FTIR, ^{1}H-NMR, coupled by the shared factors of the temperature and residence time process modes. The multi-modal structure of the tensor data from both sensors is exploited by their low rank representation to yield latent factors across the unshared mode of spectral channels, physically interpreted as pseudo-component spectra for each of the sensors in accordance with Beer's law [25]. The interactions among the pseudo-components are proposed to be encoded by a causal framework in an attempt to build an inferential model to hypothesize the underlying pseudo-reaction chemistry during the upgrading process.[193],[250]

Advancement in reaction engineering using the principles of process systems engineering suggests using data fusion algorithms as soft sensors to obtain interpretable latent factors by imposing physically meaningful constraints.[111] This spans 2 broad areas of chemometrics which are preliminary to state and parameter estimation and control of product composition :1. curve resolution (chemical signatures) 2. calibration (compositions).[32] There have been attempts to deduce underlying reactions from pseudo-component spectra obtained using curve resolution algorithms [25],[5]; however, it is lacking in its ability to incorporate information from mutliple spectral measurements into the curve resolution framework. Also, the use of Bayesian clustering to develop groups of wavenumbers having similar absorbances as nodes to build causal maps that hypothesize reaction paths[193] is limited by prior knowledge of the number of clusters. Hence, it is proposed in this work to jointly factorize multi-modal data from spectral measurements, wherein the number of components in the latent factors is determined using the mathematical notion of 'rank', followed by which causal models are built using the latent factors as the nodes to represent the underlying chemistry. Multi-modal data also called tensors are denoted by $Z \in R^{I_1 \times I_2 \times \cdots I_N}$, where N is the number of modes and I_n is the dimension of the n^{th} mode, where $n \in \{1, 2 \cdots N\}$.

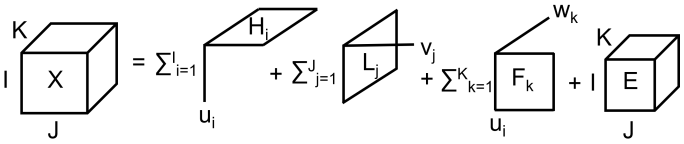
Some of the tensor decomposition formulations for a 3 mode tensor X shown



(a) PARAFAC



(b) Tucker



(c) Slice oriented decomposition

Figure 4.1: Types of tensor decomposition illustrated for a 3 mode tensor X

in Figure 4.1 are as follows: (a) Parallel factor analysis (PARAFAC), equivalently known as Canonical decomposition (CANDECOMP)/ Canonical polyadic decomposition (CPD),[251] where a tensor is described as a sum of rank-1 tensors (b) Tucker decomposition [252] where the tensor is expressed as sum of outer products of different rank factor matrices in each mode weighted by a hypercube, and more recently (c) Slice oriented decomposition[253] where a tensor is represented as the sum of an outer product of column vector of the n^{th} mode factor matrix with tensor hyper-slices of the remaining modes. This decomposition, implemented on a 3-way tensor of electroencephalogram (EEG) data, is seen to be robust to outliers and captures patterns in the slices. [253]

PARAFAC is a restricted Tucker model that can further be interpreted as a restricted principal component analysis (PCA) model on the unfolded multi-modal

data, with increasing degree of freedom over successive models, and hence the tendency to fit more noise as model complexity increases [32]. Thereby, the parsimonious PARAFAC/CPD is preferred as it has the least model complexity decomposing the tensor into independent factor matrices and is free of rotational ambiguities, implying that it is a unique decomposition where the factor matrices are subject to trivial permutation and scaling ambiguities but is robust to noise [31]. Kruskal [254, 255] has shown CPD to be unique and hence capable of representing the underlying generative phenomena should the following sufficient and necessary conditions be satisfied: $\sum_{n=1}^{N} K_{U^n} \geq 2R + (N-1)$ for sufficiency, *i.e.* sum of the ranks of all mode matrices from the decomposition must be at least greater than a function of the tensor rank and number of modes. The necessary conditions are a) $\min_{1,\cdots N} \text{rank}(U^{(1)} \odot \cdots \odot U^{(n-1)} \odot U^{(n+1)} \odot \cdots \odot U^{(N)}) = R$ *i.e.* across all modes, the minimum value of the rank of the column-wise Kronecker products of mode matrices excluding that in consideration must give the tensor rank b) $\min_{1\cdots N}(\prod_{m=1,m\neq n}^{N} \text{rank}(U^{(M)})) \geq R$ *i.e.* across all modes, the minimum value of the product of ranks of the matrix modes except that in consideration must at least be greater than the tensor rank.

Tucker decomposition, which is considered a higher order analogue of SVD [255] handles degeneracy of factor matrices by enforcing orthogonality, which in an independent decomposition as with PARAFAC can be tackled by constraining the factors [32]. Rotational ambiguities can be limited by incorporating constraints and sparsity regularizations however, most of these need prior knowledge [31].

An attempt to limit solution ambiguity by jointly factorizing unfolded data in terms of shared and unshared factors using non-negative matrix factorization with graph regularization and sparsity constraints across the latent factors, to arrive at interpretable factors [250] required prior knowledge of correlations within the FTIR and ^{1}H-NMR spectra as well as across the 2 sensor measurements. The FTIR and ^{1}H-NMR datasets for partial upgrading of Cold Lake bitumen comprise absorbances

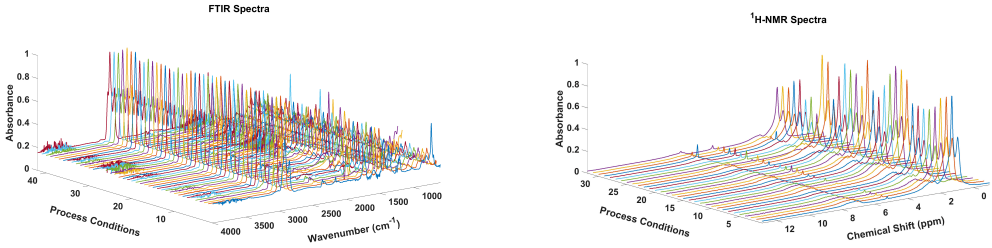
recorded over wavenumbers (mode H_1) and chemical shifts (mode H_2), respectively. The different conditions of temperature (mode A) and residence times (mode B) of processing the Cold lake bitumen samples in the visbreaker at which both the sensors record spectra, are considered as shared latent factors/mode matrices between the sensors. For the 3-way FTIR and ^{1}H-NMR tensors, the third mode is that of the spectral channels (wavenumbers and chemical shifts), which are unshared by the two process data tensors.

This work is motivated to develop an algorithm that jointly factorizes multi-modal process data tensors in terms of their shared and unshared latent factors, which are constrained to be physically interpretable, by enforcing non-negativity of the mode matrices in compliance with Beer Lambert's law [25]. This results in a unique decomposition where the absorbance projected onto temperatures and residence times are interpreted as pseudo-component concentrations across temperatures and residence time, respectively, while the absorbances projected across the unshared spectral channels *viz.* wavenumbers and chemical shifts are interpreted as the FTIR and ^{1}H-NMR pseudo-component spectra respectively.

Joint tensor decomposition has been asserted as a step towards unique decomposition [249] and robust components [256]. Consequently, non-negative multiple tensor factorization, [257] where sparsity in the target tensor is compensated for by simultaneously factorizing multiple auxilliary tensors that provide multiview information, has been used to obtain spectral features among which causal structures are learned using Bayesian networks to generate reaction hypotheses underlying complex systems. The development of JNTF as a unique structure-preserving data fusion framework to extract complementary spectral features is the main contribution of this work. The framework also brings flexibility by robustness to noise and outliers, handles missing values by imputation and enables parallelization of the scalable gradient-based optimization approach to solving tensor decomposition through subtensors using grid tensor factorization (GTF) for high dimensional tensor modes.

4.2 Description of datasets

The FTIR and ^{1}H-NMR spectroscopic datasets [196], [197] of the low-temperature thermal cracking of Cold Lake bitumen over increasing duration of residence times between 0-8 hours over different temperatures in the range of 150°C-400°C are shown in the Figure 5.1. The combination of temperature and residence time conditions are lumped together as process conditions at which the absorbances across the spectral channels are measured using the spectroscopic sensors. The samples of bitumen are reacted in a pressurized micro-batch reactor flushed with nitrogen, and the liquid products after conversion are separated by solvent extraction using methylene chloride (CH_2Cl_2) prior to obtaining the spectral measurements.



(a) FTIR dataset (b) ^{1}H-NMR dataset

Figure 4.2: Data from the spectroscopic sensors used for experimental investigation in this study

FTIR spectroscopic analysis is carried out in an ABB MB3000 equipped with a MIRacle™ Reflection Attenuated Total Reflectance (ATR) diamond crystal plate and pressure clamp. The infrared spectrometer used a deuterated triglycine sulfate (DTGS) detector. The spectra were obtained at a resolution of 2 cm^{-1} as the average of 120 scans over the spectral region 4000-600 cm^{-1}. ^{1}H-NMR spectra were obtained in a Nanalysis 60 MHz NMReady - 60 spectrometer. The equipment was pre-calibrated with deuterated chloroform. For the analysis, 0.15 g of the sample were dissolved in 0.7 μL deuterated chloroform and placed in NMR tubes. The ^{1}H-NMR analyses were performed using the following conditions: 0-12 ppm; number of scans for sample: 32; 14.7 seconds was the average scan time and 4096 points were recorded

per scan. A total of 42 FTIR and 32 ^{1}H-NMR spectra were collected, in addition to the measurement at 20°C and 0 min reaction time that was used for the purpose of baseline correction; these have been reported in Table C.1.

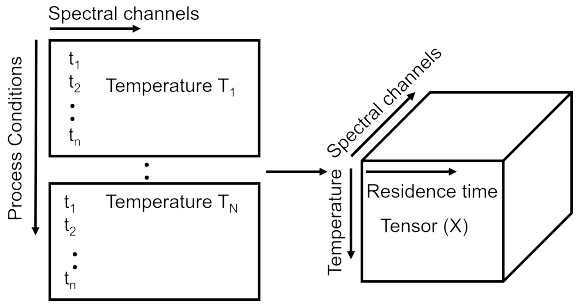


Figure 4.3: Structure-preserving tensor arrangement of spectroscopic data

In the prior work conducted by us using this dataset, the FTIR and ^{1}H-NMR data were cast as matrices that we jointly factorized by projecting the absorbance onto the shared basis of lumped process conditions and the unshared basis of the spectral channels [250]. This method relied on prior knowledge to incorporate graph regularization and sparsity constraints to not only limit solution ambiguity but also encourage the structural interpretability of the solutions. In this paper, the need of such constraints is obviated by casting the spectroscopic data as tensors, as shown in Figure 4.3, that are structure-preserving and unique in decomposition.

4.3 Methods

This section briefly outlines the multi-linear basis of jointly mining multi-sensor spectroscopy data that have an inherent structure that standard flat-view matrix decompositions are unable to exploit. JNTF is used to obtain more general hidden latent factors in the feature space of the process modes *temperature, residence times, spectral channels*. The number of components in the latent factor space is determined using the multi-linear notion of 'tensor rank'. The latent factor projections in the feature space of spectral channels of the sensors gain interpretability as the pseudo-component spectra from which the dominant compounds classes underlying the complex feed are

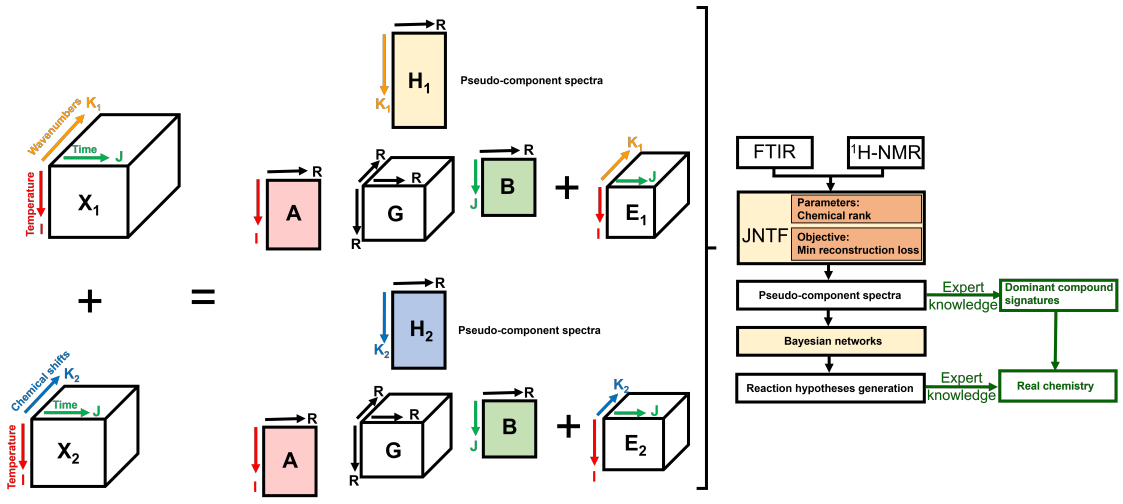


Figure 4.4: Outline of methods used in joint tensor factorization of multi-view spectral data to generate reaction hypotheses

deduced using expert knowledge. These compound signatures are then represented as nodes among which the causal paths are learned using Bayesian structure learning, as a way of generating reaction hypotheses [250]. The hypotheses are then validated using domain knowledge to infer the chemical interactions among the identified compound classes. The broad scheme of going from spectroscopic data to knowledge of chemical interactions agnostic to prior mechanistic insights of complex chemical systems is shown in Figure 4.4.

Notations: Tensors are denoted as $\mathcal{X}, \mathcal{Y}, \cdots$. Matrices, vectors and scalars are denoted by bold uppercase, bold lowercase, and lowercase respectively, like **A**,**a** and a. Element (i, j, k) of a tensor $\mathcal{A} \in \mathscr{R}^{I \times J \times K}$ is symbolized as a_{ijk}, element (i, j) of a matrix $\mathbf{A} \in \mathscr{R}^{I \times J}$ as a_{ij} and the i^{th} entry of a vector $\mathbf{a} \in \mathscr{R}^{I}$ as a_i. Moreover, $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product, $\mathbf{A} \odot \mathbf{B}$ is the Khatri-Rao or the column-wise Kronecker product, $\mathbf{A} * \mathbf{B}$ is the element-wise Hadamard product, $\mathbf{a} \circ \mathbf{b}$ is the outer product of vectors.

4.3.1 Rank determination of the tensor

A tensor is a multi-dimensional array such that an N^{th} order tensor is an element of the tensor product of N vector spaces, each having its own co-ordinate system and must not be confused with tensors in physics and engineering (such as stress tensors), also referred to as tensor fields [255]. The spectral data is viewed as a 3 way array, with temperature, residence time and spectral channels as its modes; hence making it favorable to be analyzed in a tensorial framework that captures the intermodal interactions during the decomposition.

The rank of the tensor that gains interpretability as the number of pseudo-components is the necessary number of rank-1 tensors for the lower dimensional representation of multi-way data, the determination of which is non-deterministic in polynomial time (NP hard) [258]. The tensor rank is determined by fitting PARAFAC as a restricted Tucker model and estimating the core consistency of the Tucker core [259]. Core consistency is a metric used to compare the super-diagonal elements of the restricted Tucker core with that of the core from the actual Tucker decomposition to automatically determine the rank (model complexity) without *a priori* assumptions regarding residuals [34],[260]. The tradeoff between lack of fit and core consistency for noisy data as rank R increases, where there is a sharp drop in core consistency at a point where noise is fit, is used as an indicator of overfactoring while choosing R [258]. Under-specifying rank may cause independent chemical responses to be mixed into a one component, while overfactoring could fit to noise; hence, multiple other validation techniques can also be used such as randomness of residuals (mainly white noise), visualizing spectral loadings and split-half analysis[261]. Split half analysis [262] exploits the uniqueness of CPD and involves fitting models for a given rank R on a portion of the data and testing to see if identical latent factors are obtained when the same model is fit to the test data.

Tensor decompositions can be considered as higher order extensions of matrix SVD

and broadly fall into 2 categories [255] : CANDECOMP [251]/PARAFAC [26]which represents a tensor as a sum of rank 1 tensors and the Tucker decomposition [252], which is a higher order PCA.

For an $N^(\text{th})$ order tensor$\mathcal{Z} \in \Re^{I_1 \times I_2 \times \cdots \times I_N}$

CPD:

$$\mathcal{Z} = \sum_{r=1}^{R} u_r^{(1)} \circ u_r^{(2)} \circ \cdots \circ u_r^{(N)} + E \tag{4.1}$$

where $\mathbf{U}^{(n)} = [\mathbf{u}_1^{(n)}, \mathbf{u}_2^{(n)}, \cdots, \mathbf{u}_R^{(n)}] \in \Re^{I_n \times R}$ denotes a component matrix for mode n; R being the rank of the CP decomposition. In tensor matrix form, eqn 4.1 can be written as

$$\mathcal{Z} = I \times_1 \mathbf{U}^{(1)} \times_2 \mathbf{U}^{(2)} \times_3 \cdots \times_N \mathbf{U}^{(N)} + E \tag{4.2}$$

where I is an identity hypercube.

Tucker Decomposition:

$$\mathcal{Z} = \sum_{r_1=1}^{R_1} \sum_{r_2=1}^{R_2} \cdots \sum_{r_N=1}^{R_N} g_{r_1\, r_2 \cdots r_N} \mathbf{a}_{r_1}^{(1)} \circ \mathbf{a}_{r_2}^{(2)} \circ \cdots \circ \mathbf{a}_{r_N}^{(N)} + E \tag{4.3}$$

where $\mathcal{Z}_{r_1\, r_2 \cdots r_N} = \mathbf{a}_{r_1}^{(1)} \circ \mathbf{a}_{r_2}^{(2)} \circ \cdots \circ \mathbf{a}_{r_N}^{(N)}$ is a rank 1 tensor and $\mathcal{Z}$ is a summation of $R_1 \times R_2 \times \cdots \times R_N$ rank 1 tensors with a Tucker core $G \in \Re^{R_1 \times R_2 \times \cdots \times R_N}$.

In matrix form, eqn 4.3 can be written as follows:

$$Z = G \times_1 A^{(1)} \times_2 A^{(2)} \times_3 \cdots \times_N A^{(N)} + E \tag{4.4}$$

where $A^{(n)} = [a_1^{(n)}, a_2^{(n)}, \cdots, a_{R_n}^{(n)}] \in \Re^{I_n \times R_n}$ denotes a component matrix in the $N^(\text{th})$ mode.

The main difference between CPD and Tucker decomposition is that in CPD the number of components across the modes is invariant; while it is not so in the Tucker

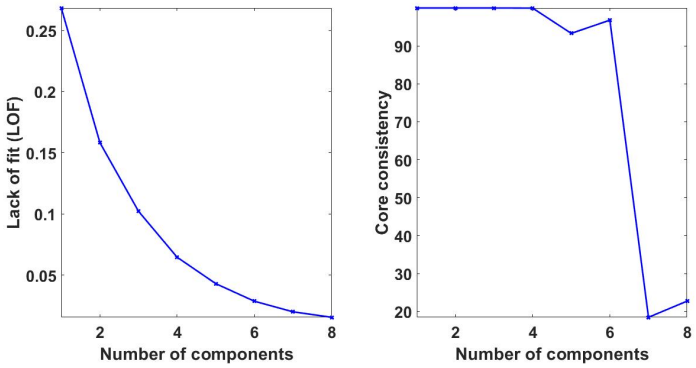
model. [244] Therefore, CPD, in a sense, can be seen as a restricted Tucker model with equal number of components in each mode. This leads to an important diagnostic called the core consistency diagnostic, which is used to determine the number of components in CPD/PARAFAC models. Fitting PARAFAC models with an arbitrary number of components and then casting them into Tucker models results in an identity hypercube, G, if the right number of components are used. [259] The PARAFAC tensor when matricized in the $n^(\text{th})$ mode can be written as a restricted Tucker model as follows:

$$Z_{(n)} = U^{(n)} T^{R \times (R * R \cdots N-1 \text{times})} (U^{(1)} \otimes U^{(2)} \otimes \cdots \otimes U^{(n-1)} \otimes U^{(n+1)} \otimes \cdots \otimes U^{(N)})^T \quad (4.5)$$

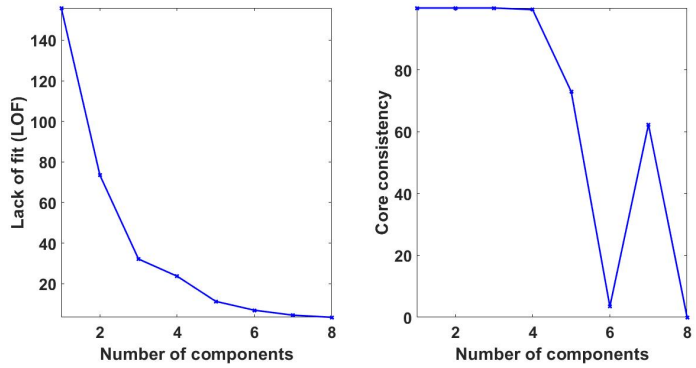
This can be then used to calculated the core consistency as follows

$$\text{Core Consistency} = 100\Big(\frac{1 - ||G - T||_F^2}{||T||_F^2}\Big) \quad (4.6)$$

It can be seen from Figure 4.5 that the tensorial blocks have 4 components that represent the underlying pseudo-component classes. This is used as a basis for CPD using an optimization framework that is an improvement over the Alternating Least Squares (ALS) approach [26],[251] which was not reliable as it was not guaranteed to converge to a stationary point. Hence, a gradient-based optimization technique was devised for simultaneously solving over all factor matrices [258]. In theory, the Tucker decomposition does not possess unique solutions even though it is subjected to the permutation and variance indeterminacies. [245, 255] In practice, when additional assumptions are introduced on the different modes, the Tucker decomposition can be unique. [263] This is because Tucker decomposition is based on the orthogonal premise of decomposition, while CPD is based on decomposing into independent individual components; which accounts for the uniqueness of the decomposition despite its scaling and permutation indeterminacies. [32]



(a) FTIR rank determination



(b) ^{1}HNMR rank determination

Figure 4.5: Lack of fit (LOF) and core consistency plots for FTIR ^{1}H-NMR tensor blocks.

4.3.2 JNTF objective function

Joint non-negative tensor factorization (JNTF), which is considered as a higher order analogue of joint non-negative matrix factorization (JNMF) [250] is a structure preserving latent variable model for the additive (due to non-negativity constraints) parts-based combination of basis factors. Non-negative tensor factorization (NTF) has been implemented similarly using cost functions as with non-negative matrix factorization (NMF): the least square error under the assumption of homoscedastic Gaussian noise; and Kullback-Leibler divergence [215] under the assumption of Poisson noise.

The NTF objective function is formulated to minimize the least squares error between the data tensor and the projections onto its mode matrices that are constrained

to be non-negative. This can be extended to jointly factorize multiple tensors that are coupled through shared mode matrices that are held common in their factorization. The number of components in each mode is the tensor rank (R). The process modes of temperature and residence time that are common to the sensors are denoted by matrices $A \in \mathscr{R}^{I \times R}$ and $B \in \mathscr{R}^{J \times R}$, respectively. The dimensions I and J of the process modes denote the number of temperature and residence time points, respectively, at which the spectra are measured. The unshared spectral channel modes of the sensors are denoted by the mode matrix $H_i \in \mathscr{R}^{K_i \times R}$, where K_i is the dimension of the spectral channels over which absorbance is measured for the FTIR and ^{1}H-NMR spectral sensors, as shown in Figure 4.4. It is to be pointed out that both sensor measurements are not available for all combinations of the process conditions outlined in Table C.1. The missing measurements are accounted for, by using a weighting tensor W^i that imputes them by zero in the multi-linear decomposition. The least squares objective function to be minimized for JNTF is given in Eqn 4.7.

$$\min_{\mathbf{A},\mathbf{B},\mathbf{H}_i \geq 0} F_W(A, B, H_i) = \sum_{i=1,2} ||W^i * (Z^i - [[A, B, H_i]])||^2 \tag{4.7}$$

The use of the Multiplicative Update Rules (MUR) [264] to solve for the factor matrices is seen not to guarantee the stationarity of the limit points because of which it is undesirable for finding the optima of non-convex NTF based objective functions. Alternatively, the ALS method, wherein the gradients of the objective function with respect to the latent factor matrices are used to develop updating schemes for the factor matrices in a round robin fashion, has slow convergence, is not accurate in overfactoring [258] and is not guaranteed to converge to a stationary point [265]. The ALS convergence can be hastened by the use of sparsity regularization, weight matrices in the objective and projected gradients [265]; however, the solution is still suboptimal [266]. Non-negative least squares (NLS) was deployed as an accurate alternative to solve the NTF objective [258] but proved to be computationally slow due

to Jacobian calculations, leading to a gradient-based optimization method that solves for all the factor matrices simultaneously, is accurate, has the same computational overhead as ALS and is tractable for large tensors. The gradients of the objective function are given by Eqns 4.8-4.10, for which $X^i = W^i * [[A, B, H_i]]$ and $Y^i = W^i * Z^i$.

$$\nabla_A F_W = \sum_i 2(X^i_{(1)} - Y^i_{(1)})(H_i \odot B) \tag{4.8}$$

$$\nabla_B F_W = \sum_i 2(X^i_{(2)} - Y^i_{(2)})(H_i \odot A) \tag{4.9}$$

$$\nabla_{H_i} F_W = 2(X^i_{(3)} - Y^i_{(3)})(B \odot A) \tag{4.10}$$

The above problem is solved using the LBFGSB solver of the Poblano optimization toolbox, developed by Sandia Laboratories on Matlab [267]. The matrices are initialized based on the average values of the individual tensorial decompositions : solving NTF by $\min_{A_i, B_i, H_i \geq 0} ||W^i * (Z^i - [[A_i, B_i, H_i]])||^2$ for $i = 1, 2$ separately, followed by initializing : $A^0 \leftarrow \frac{1}{2} \sum_{i=1,2} A_i$, $B^0 \leftarrow \frac{1}{2} \sum_{i=1,2} B_i$, $H^0_i \leftarrow H_i$ for solving the routine in eqn 4.7.

4.3.3 Robust formulation of JNTF

The earlier formulation of tensor factorization is based on the assumption of noise being independently and identically distributed (*iid*) Gaussian. For a 3 way tensor Z of size $I_1 \times I_2 \times I_3$, a rank R CP decomposition gives the matrix components in each mode : $A \in R^{I_1 \times R}$, $B \in R^{I_2 \times R}$, $C \in R^{I_3 \times R}$ by solving the least squares objective outlined in eqn 4.7, which is the same as the result of maximizing the log likelihood function resulting from the assumption of Gaussian noise [268]. This can be illustrated as follows:

$$z_{i_1 i_2 i_3} = \sum_{r=1}^{R} a_{i_1\, r} b_{i_2\, r} c_{i_3\, r} + \epsilon_{i_1 i_2 i_3} \text{where } \epsilon_{i_1 i_2 i_3} \sim N(0, \sigma^2) \tag{4.11}$$

Consequently, $z_{i_1 i_2 i_3} \sim N(\sum_{r=1}^{R} a_{i_1\, r} b_{i_2\, r} c_{i_3\, r}, \sigma^2)$, from which it follows that the prob-

ability of the independent tensorial elements conditioned on the decision variables of CPD can be expressed as

$$P(z_{i_1 i_2 i_3} | \sum_{r=1}^{R} a_{i_1\ r} b_{i_2\ r} c_{i_3\ r}) \sim \exp\left(- \frac{||z_{i_1 i_2 i_3} - \sum_{r=1}^{R} a_{i_1\ r} b_{i_2\ r} c_{i_3\ r}||}{2\sigma^2} \right)^2 \tag{4.12}$$

Finally, the expression for the log likelihood of the N (here, =3) mode data can be written as

$$\log \prod_{i_1 i_2 i_3} P(Z|[A,B,C]) = -\frac{1}{2\sigma^2} \sum_{i_1 i_2 i_3} ||z_{i_1 i_2 i_3} - \sum_{r=1}^{R} w_{i_1 i_2 i_3} * a_{i_1\ r} b_{i_2\ r} c_{i_3\ r}||^2 \tag{4.13}$$

Maximizing the above data log-likelihood is tantamount to minimizing the least squares objective given in eqn 4.7, the only difference being the imputation of missing measurements with a weighting tensor. It can be seen that using the least squares objective in tensor factorization makes CP decomposition sensitive to non-Gaussian noise. Besides, the presence of outliers tends to dominate a squared objective function [269]. Hence, it is imperative to re-formulate the factorization objective so that it is robust to handle non-Gaussian noise and outliers, and this is typically achieved by using the L_1 norm in the tensor factorization objective [268]:

$$L_1(A,B,C) = \sum_{i_1 i_2 i_3} \sqrt{z_{i_1 i_2 i_3} - \sum_{r=1}^{R} a_{i_1\ r} b_{i_2\ r} c_{i_3\ r}} \tag{4.14}$$

The least absolute error has been used as a robust alternative to least squares when the process errors are additive (*iid*) Laplacian, which is more heavily tailed than Gaussian, making it better suited to model noise and outliers [270]. A majorization-minorization approach of solving the above non-convex objective by breaking it into convex sub-problems in terms of each of the decision variables that are updated while the others are fixed, followed by a round robin scheme of solving the sub-problems

to update the other decision variables in their turn, has been outlined in other works [268]. Each sub-problem is in effect the rank R approximation of the mode-n matricized tensor, formulated as a weighted l_1 regression problem. The underlying philosophy of this method is similar to the Canonical Polyadic ALS (CPALS) approach using a different norm for the objective function, and as discussed earlier, CPALS has not demonstrated guaranteed convergence to a stationary point [26],[251]. It is desired to develop a framework in which all the factor matrices can be solved for simultaneously using one of the many gradient-based optimization solvers [258]. The robust formulation of the objective for NMF using the L_{21} norm has been outlined in Kong, et al.[213] Using this as a starting point, we seek to extend the concept of using this robust norm to solve for factor decomposition in each of the matricized tensor modes. Combining the individual sub-problems into a single objective function using the L_{21} norm alongside the imputation of missing measurements gives the following objective function for the weighted robust non-negative tensor factorization:

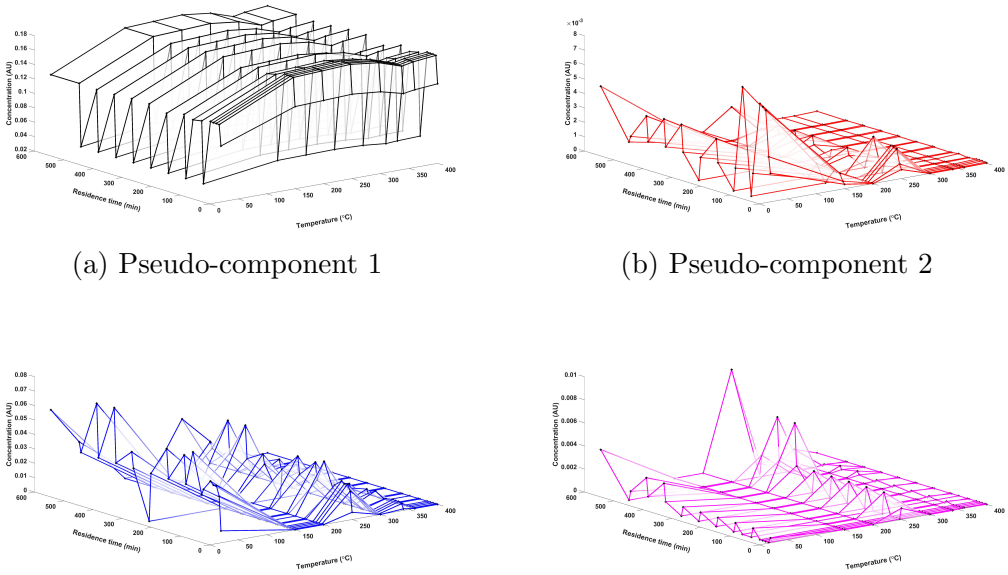
$$\begin{aligned}\min_{A,B,C\geq 0} F(A,B,C) = &||W_{(1)} * [Z_{(1)} - A(C \odot B)^T]||_{21} + ||W_{(2)} * [Z_{(2)} - B(C \odot A)^T]||_{21} \\ &+ ||W_{(3)} * [Z_{(3)} - C(B \odot A)^T]||_{21}\end{aligned} \tag{4.15}$$

Gradient computation for the NTF objective involves the column-wise Kronecker product of modal factors, making its computation intractable due to huge memory requirements for high dimensional tensor modes [271], which is overcome by grid tensor factorization (GTF) which breaks the tensor into subtensors. The subtensors are factorized independently using CPD in parallel before integrating results for the whole tensor to estimate factors in each mode [272]. Solving Eqn 4.15 for all the mode matrices simultaneously using gradient-based optimization, by parallelizing the tensor decomposition using sub-tensors has been outlined in Section C.2.

4.4 Results and Discussion

This section presents findings from the individual NTF decomposition of data from the FTIR and ^{1}H-NMR spectral sensors, in contrast to their coupled analysis. These tensor decompositions have been found to satisfy the necessary and sufficient conditions for uniqueness as outlined by Kruskal [254],[255], and are free of solution ambiguities. The discussion of the tensor decomposition and subsequent interpretation of the Bayesian networks constructed from the pseudo-component spectra are provided together with the results for each case so as to have an easier interpretation. The reaction pathways hypothesized from the Bayesian networks have been validated against literature pertaining to conversion chemistry in bitumen that has been investigated using quantitative metrics reflecting composition changes of model compounds, representative of the complex reactive system [273], [250], [193], [25]. It must be noted that the pseudo-component signatures from the tensor decomposition does not point to a single molecular structure, but a class of compounds. Suitable model compounds with structures representative of the pseudo-component spectra have been used to indicate plausible conversion pathways in line with the Bayesian networks. The merit of the framework lies in the structure-preserving data fusion from different spectroscopic sensors to develop reaction hypotheses among the identified pseudo-components, without prior knowledge of the reactive system in terms of either its species or underlying conversion pathways.

The results presented are based on the robust formulation of NTF using the L_{21} norm in the objective function, to facilitate handling non-Gaussian noise in the sensor data during the decomposition. The results from repeating the analysis using the squared norm in the objective (Frobenius *i.e.*L_2), under the assumption of Gaussian noise have been included in the Appendix between Section C.5.1 and Section C.5.3.



(a) Pseudo-component 1

(b) Pseudo-component 2

(c) Pseudo-component 3

(d) Pseudo-component 4

Figure 4.6: Concentrations of the pseudo-components across the reaction space of the FTIR spectra

4.4.1 Individual analysis of FTIR data

Figures 4.6a, 4.6b, 4.6c and 4.6d provide the concentration profiles across the reaction space of temperature and residence times, for the 4 pseudo-components obtained through rank determination using LOF and core consistency metrics as described in Section 4.3.1. The tensor decomposition of the normalized FTIR data accounts for the inter-modal interactions among temperatures, residence times and spectral channels (the 3 modes) while projecting FTIR data onto each of the modes. However, to derive physical meaning from the pseudo-component concentrations that are impacted by the coupling between reaction temperatures and residence times, a surface plot of the profiles across the said modes have been outlined in the earlier figures, while Figure 4.7 gives the extracted spectral profiles obtained by projection onto the FTIR spectral channels for the 4 pseudo-components. The infrared spectra of all four principal components have realistic absorption bands, although PC_4 has a noticeable noise component.

Figure 4.8 provides the Bayesian networks indicating causal relationships between the pseudo-component groups obtained from Hill climbing (HC), Tabu search and Maximum minimum hill climbing (MMHC) structure learning algorithms. It should be noted that the groups in the Bayesian network (Figure 4.8) are the same as pseudo-components (PCs) in the extracted spectral profiles from tensor decomposition so both these terms will be used interchangeably in this paper. The projection of absorbance onto the modes of temperature and residence time, rendering their interpretation as concentrations along those modes, could further be used to develop kinetic models for the process and is explored in our upcoming works. In this work, we shall use these concentrations for qualitative corroboration with the Bayesian networks learned from the pseudo-component spectra. The results of using tensor decomposition of synthetically generated FTIR data across a range of operating temperatures and residence times is seen to provide more continuous concentration surfaces, as presented in Section C.3.

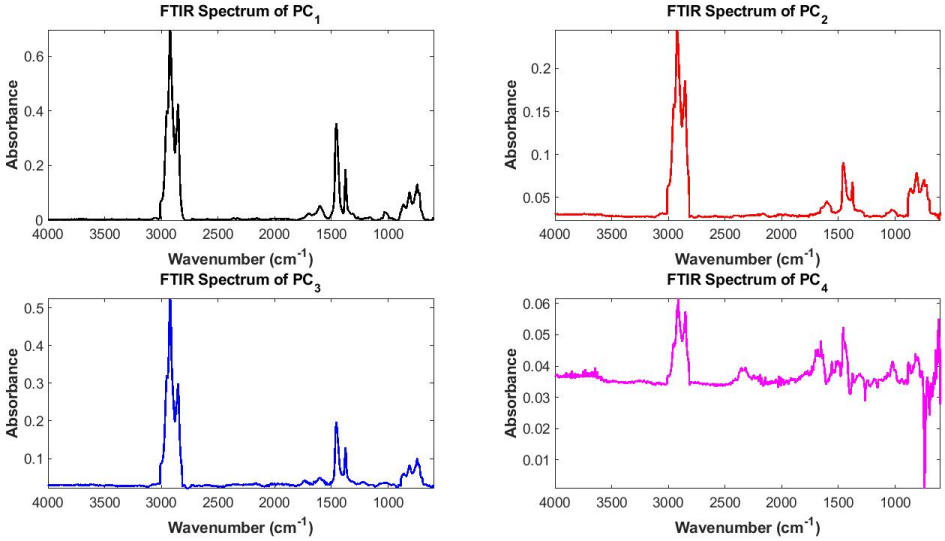


Figure 4.7: Spectra of pseudo-components from FTIR tensor decomposition

For the robust decomposition of FTIR data, hill climbing and MMHC produced similar networks. Hence, these network structures shall be used as a basis for generating reaction hypotheses. These networks corroborate the qualitative trends in the

concentration profiles, where it can be seen that the concentration of PC_1 is much higher than the rest based on the Figure 4.6, indicating that it represents a class of reacting species. This is reflected in the Bayesian network structure, where G1 converting to G3 has the highest arc strength (Figure 4.8). Correspondingly, PC_3 is seen to exist in higher concentration than PC_2 and PC_4 at most temperatures and reaction times (Figure 4.6). The networks from the Bayesian structure learning indicate that G2 tends to be one of the final products while G3 and G4 are intermediate products in the sequence. A sharper increase in PC_4 is noticed in the concentration-time profiles at longer reaction times across intermediate temperatures, while PC_2 is widely present at lower temperatures.

Before delving into the conversion chemistry, it is important to identify the major functional groups present in each pseudo-component and a representative compound for each group in the Bayesian network developed for the robust formulation of FTIR data. It is to be noted that the baseline of the extracted profile of PC_4 is noisier than the other 3 profiles and this has been kept in mind while identifying the characteristic functional groups. All the PCs show the characteristic aliphatic sp^3 C-H stretches at 2850 and 2920 cm^{-1} for methylene (CH_2) groups and at 2950 cm^{-1} for methyl groups, respectively. The bending frequencies for sp^3 C-H bonds are seen at 1380 and 1450 cm^{-1} for every pseudo-component and the intensity of the methylene stretch is more than the methyl stretch, indicating the presence of side chains as well as naphthene rings as CH_2 groups could be a constituent of both. This is consistent with the composition of bitumen, which on a molar basis has a heteroatom-to-carbon ratio of around 0.03, but a hydrogen-to-carbon ratio of around 1.5 [243]. Other distinct stretches of all the groups are given in Table 4.1.

Table 4.1: Absorption regions for all groups in robust FTIR formulation.

Wavenumber (cm^{-1})	Functional group	Vibration type	PCs/groups present
1597 1653	C=C aromatic C=C	Stretch	PC1, PC2, PC3

Table 4.1 continued from previous page

1701	C=O of carboxylic acid	Stretch	PC1, PC4
1172 – 1203	Acyl, phenolic C-O	Stretch	PC1, PC3 (mild)
1018	C-O of aliphatics	Stretch	All 4 PCs
862	sp^2 C-H in p-substituted aromatics	Bend	All 4 PCs
810	sp^2 C-H in m-substituted aromatics	Bend	All 4 PCs (higher in PC2)
740	sp^2 C-H in o-substituted aromatics	Bend	All 4 PCs but higher for PC1 and PC3
723	sp^2 C-H in mono-substituted aromatics	Bend	Clearer in PC2 – as a shoulder with 740 cm-1
1740	C=O in esters/anhydrides	Stretch	PC3, PC2 (mild)

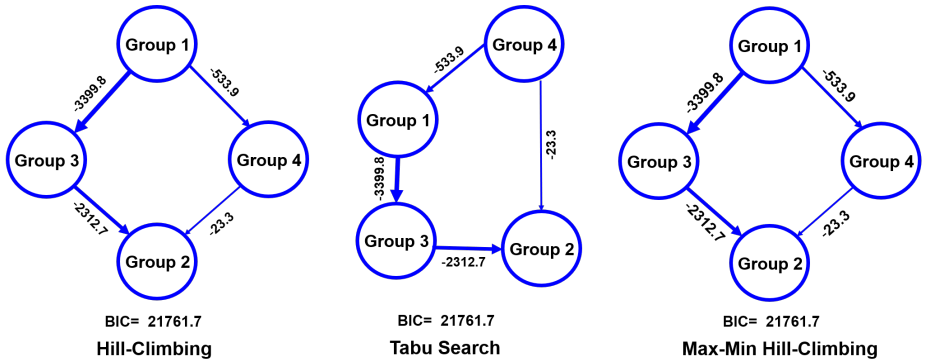


Figure 4.8: Bayesian networks from the unique FTIR pseudo-component spectra

Here, we will be accounting for the 4 pathways shown in Figure 4.8 as predicted by the Bayesian network in order of their probability of occurrence. For PC_2, absorption at 1700 cm^{-1} indicated the presence of carboxylic acid and its co-existence with C-O acylic group at 1175 cm^{-1} confirmed this observation. Presence of aliphatic alcohol was also marked by absorption at 1018 cm^{-1}. The sp^2 C-H bends at 740 cm^{-1} for o-aromatics were seen to be of maximum intensity. The representative compounds for each group are shown in Figure 4.9, Figure 4.10, Figure 4.11 and Figure 4.12 that also depict the proposed reaction pathways based on the results of Bayesian networks from robust formulation of FTIR data.

Compound (1) is a representative molecule for G1 since it has a carboxylic acid, aliphatic alcohol in the naphthene ring, a side chain and an aromatic ring that is

substituted in o-, m- and p- positions. These substitution patterns are not reflected in the representation shown in Figure 4.9. The chemical structure of G3 species is not much different than G1 but can comprise a phenolic and an ester group, which can be obtained by condensation of the middle ring to become an aromatic (compound (2)) or esterification of the COOH group by combining with an alcohol (compound (3)), respectively. Compounds with carboxylic acid and phenolic functional groups were identified in bitumen and the cracking products from bitumen.[274],[275] Ester formation by the reaction of carboxylic acids and alcohols can take place and benefits from vaporization of the co-produced water, as this is an equilibrium limited conversion pathway.[276] In the specific example shown in Figure 4.9, the probability of the end ring turning into aromatic is lower than that of the middle ring , since only the middle ring benefits from having benzylic hydrogens that on transfer will yield resonance stabilized benzylic radicals. Hydrogen transfer reactions in general,[277] as well as hydrogen transfer reaction in bitumen specifically,[278] have been observed over the temperature range of the data in this study. These reactions lead to hydrogen disproportionation between the molecules that could either lead to the net decrease in hydrogen by the conversion of cycloalkane structures to aromatic rings or the saturation of aromatic structures This reaction sequence could explain some of the characteristics for the conversion of group 1 to group 3 species.

Compounds (2) and (3) had ortho, meta and para substitutions but another striking feature of the G3 species is the highest intensity for ortho-substituted aromatics. So, for proposing a plausible reaction sequence the second-most probable pathway from G3 to G2, an ortho-substituted aromatic with an ester group (compound (4)) has been considered as representative of group 3 that has a similar backbone as the compounds (2) and (3). This is shown in Figure 4.10. Four hydrogens are lost from the middle ring of compound (4) to convert it into an aromatic, which facilitates cracking of the third ring and its side chain due to lowering of C-C bond energy. However, this is not an essential step and the cracking pattern to yield an ortho-

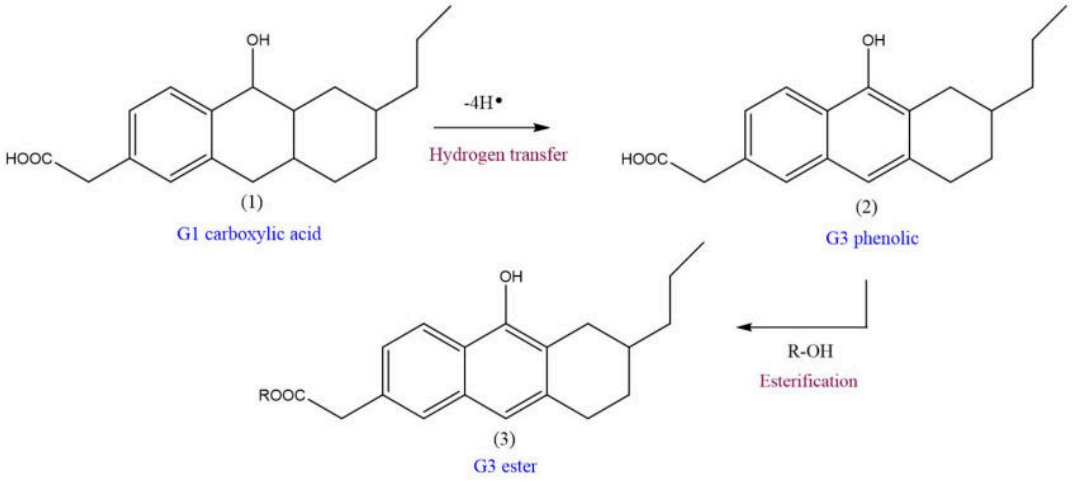


Figure 4.9: Proposed reaction pathway of G1 (group 1) to G3 (group 3) conversion.

substituted aromatic can be obtained from cracking of the middle ring. This results in the formation of compound (6), which is straight-chain olefin and representative of group 2 and can undergo further hydrogen transfer to yield a stable conjugated diene (7). Thermal cracking of longer alkyl chains or cycloalkane rings followed by free radical stabilization via hydrogen transfer, is shown to result in lighter aliphatic and olefinic compounds.[279] As shown in Figure 4.10, it is more likely that hydrogen transfer from the central ring precedes cracking, since cracking to form a resonance stabilized benzylic free radical is more favorable.[280] Meanwhile, the ester group in compound (4) can undergo hydrolysis to yield the alcohol that turned into a phenolic group and forms compound (5), which is also meta- substituted to illustrate the development of the high intensity of C-H bends for a meta-substituted aromatic at $\sim$ 810 cm^{-1}. It is to be noted that this also corresponds with the observation that the ortho-substitution is still present in group 2 as indicated from the extracted FTIR spectra (Figure 4.7). The reaction sequence indicated in Figure 4.10 implies hydrolysis of the ester, which as mentioned before, is a reversible reaction. However, the same reaction sequence without hydrolysis or the formation of a phenol is possible by the direct thermal decomposition of the ester. Direct thermal decomposition of the ester

leads to the elimination of a carboxylic acid with concomitant formation of C=C in the naphthenic ring.[281] Of relevance to the reaction pathway is that both thermal cracking and thermal decomposition of ester groups can be sources of alkenes.

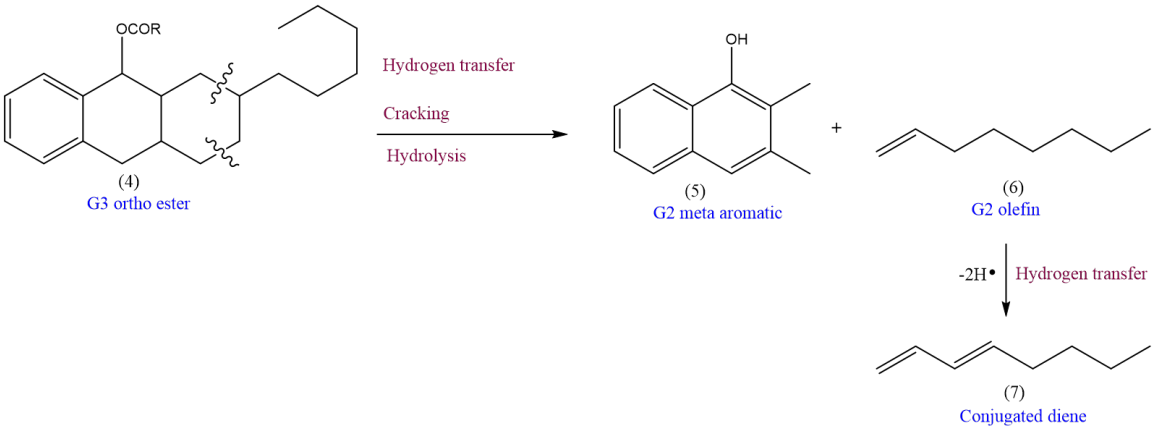


Figure 4.10: Proposed reaction pathway of group 3 to group 2 conversion.

The third-most probable pathway is the conversion of group 1 to group 4 with the third highest arc strength in the causal structure. A plausible sequence is shown in Figure 4.11. Though the functional groups are the same in compounds (1) and (8), the carboxylic acid group is attached to the third naphthene ring in (8) instead of the aromatic ring as in (1). Group 4 species can be represented by compounds (9) and (10) which are formed by cracking of (8) at the indicated positions, followed by decarboxylation of the olefin. Group 4 species shows absorption for aliphatic alcohol between 1013-1100 cm^{-1} and also contains C=C stretch for alkene and aromatic groups at 1653 and 1600 cm^{-1} , respectively.

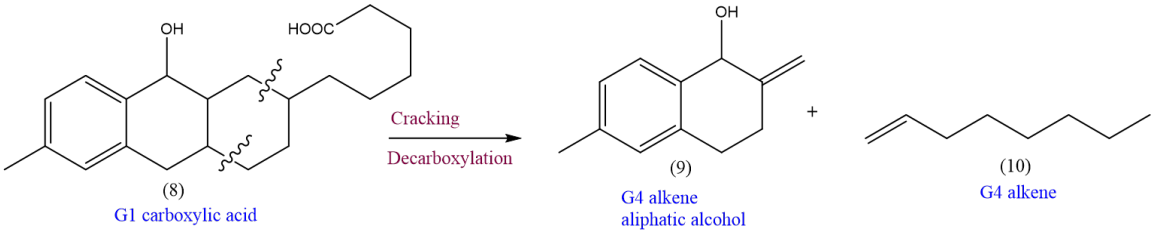


Figure 4.11: Proposed reaction pathway of group 1 to group 4 conversion.

Finally, the least plausible, but the pathway with the least arc strength according

to the Bayesian network structure is the conversion of group 4 species to group 2 moiety where it can be said that the alkene C=C stretch was more clearly seen in group 4 than in group 2 though the extracted FTIR spectrum of PC_4 was noisier (Figure 4.7). There are 2 possible pathways for cracking that will yield to relatively stable products. In both cases, cracking occurs at the $\alpha - \beta$ C-C bond as indicated, but this can occur with or without hydrogen transfer from the naphthenic ring. If the naphthenic ring with a double bond loses 2 more hydrogens via hydrogen transfer, it becomes an aromatic and yields compound (12). On the other hand, if cracking occurs before hydrogen transfer, then a conjugated alkene-aromatic (14) is formed. The side alkyl chain gives the olefin (13) as the other cracked product and (13) and (14) are representative of group 2 species as well. Since, group 4 and group 2 species are structurally and functionally similar, this conversion can be considered the least probable.

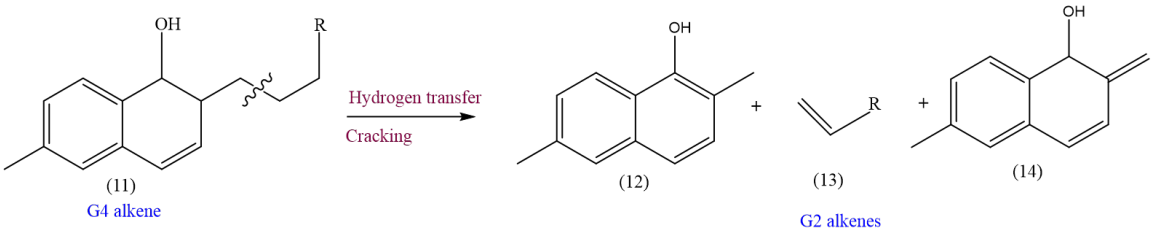


Figure 4.12: Proposed reaction pathway for group 4 to group 2 conversion.

In conclusion, it was possible to construct plausible reaction sequences that represented the networks shown in Figure 4.8. The functional groups were representative of those identified for the pseudo-compound in the PC_1-PC_4 FTIR spectra.

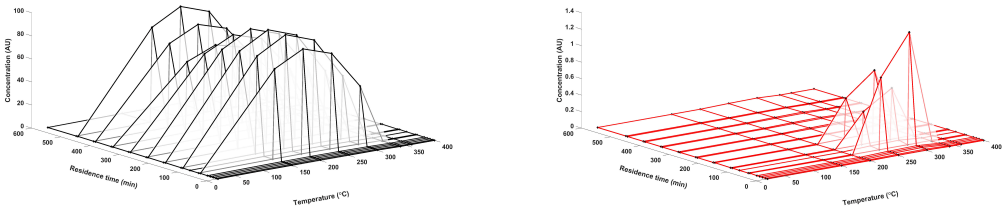
4.4.2 Coupled analysis of FTIR and ^{1}H-NMR data

The ^{1}H-NMR spectra provide information about the proton environment in a compound. Peaks in the range 1.5-2.5 ppm point to aliphatic hydrogens, in the 4-6 ppm point to olefinic hydrogens and the 7-9 ppm range of chemical shifts comprise overlapping peaks of aromatic hydrogens. The results from the individual tensor de-

composition of the said spectra are outlined in Section C.4. It is seen that at lower temperatures and residence times of reaction, pseudo-components with an aliphatic character are predominant, which at higher temperatures result in compounds with pronounced aromatic and olefinic character. Knowledge of only the proton environment has hindered the hypothesis of a detailed conversion chemistry solely based on the ^{1}H-NMR pseudo-component spectra from tensor decomposition. It is believed that information about the proton environment from the ^{1}H-NMR spectra could complement that of the functional groups from FTIR, when jointly analyzed, facilitating the disambiguation of pseudo-components that share similar FTIR peaks.

Interestingly, when FTIR and ^{1}H-NMR data are fused in the input to tensor decomposition, the characteristic peaks in the extracted ^{1}H-NMR profiles are not altered much from the profiles obtained when the ^{1}H-NMR data was considered separately as discussed in section C.4. However, the FTIR pseudo-component profiles from the joint analysis differ considerably from that of Section 4.4.1. Figures 4.13a to 4.13d provide the concentration profiles of the pseudo-components in the space of the temperature and residence time modes, for the joint robust formulation, and appear to differ only in terms of the scaling while having identical trends, as reported in Figure C.4. Figure 4.14 gives the jointly extracted FTIR and ^{1}H-NMR profiles for the 4 pseudo-components and Figure 4.15 shows the Bayesian networks obtained through the 3 greedy search algorithms with the 4 PCs as the nodes. The HC and MMHC score search methods return identical network structure and concur with the qualitative insights obtained from the concentration profiles of Figure 4.13. The reader is kindly referred to the discussion in section 4.4.1 and Table 4.1 for the characteristic peaks and the corresponding functional groups present.

It must be noted that despite PARAFAC resulting in a unique decomposition, yet the factor matrices are subject to trivial permutation [31]. As a result, the spectrum of PC_1 from the joint decomposition points to a noisy baseline as given by Figure 4.14, in comparison to the decomposition of only the FTIR data, where PC_4 was seen



(a) Pseudo-component 1

(b) Pseudo-component 2

(c) Pseudo-component 3

(d) Pseudo-component 4

Figure 4.13: Concentrations of the pseudo-components across the reaction space from the joint decomposition of FTIR and ^{1}H-NMR spectra

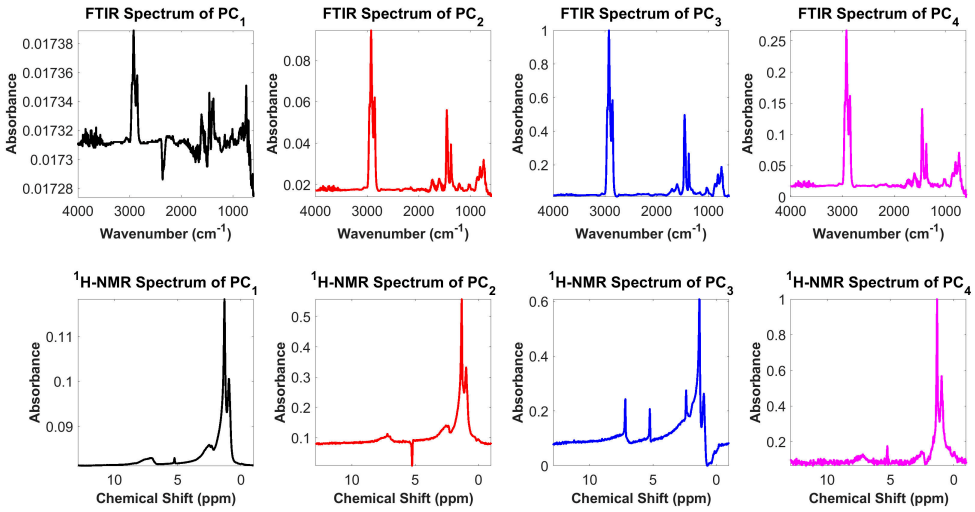


Figure 4.14: Spectra of pseudo-components from joint tensor decomposition

to be noisy (Figure 4.7). Although the FTIR spectrum of PC_1 is a bit noisy, we can clearly see a dominant peak at 740 cm^{-1} indicating the presence of ortho-substituted aromatics. A peak at 1155 cm^{-1} and 1011 cm^{-1} point to the C-O stretching of an aliphatic alcohol, the peak at 1610 cm^{-1} indicates C=C stretching and the inverted

peak at 2359 cm^{-1} corresponds to the O=C=O stretch of carbon dioxide. PC_2 has a peak at 1030 cm^{-1} that indicates the C-O stretch of alcohols, followed by peaks at 1215 cm^{-1} and 1736 cm^{-1} that correspond to the C=O stretch of esters and the peak at 1607 cm^{-1} pointing to the C=C olefinic stretch. The spectrum of PC_3 has a peak at 1030 cm^{-1} indicating C-O stretch of alcohols, a peak at 1603 cm^{-1} that points to the olefinic stretch and a peak at 1700 cm^{-1} indicating a carboxylic acid functional group stretch. The FTIR spectrum of PC_4 is very similar to that of PC_3. However, the ^{1}H-NMR spectra of PC_3 indicates a dominant olefinic behavior in the region between 4-6ppm, as compared to that of PC_4, implying that it comprises condensed aromatics.

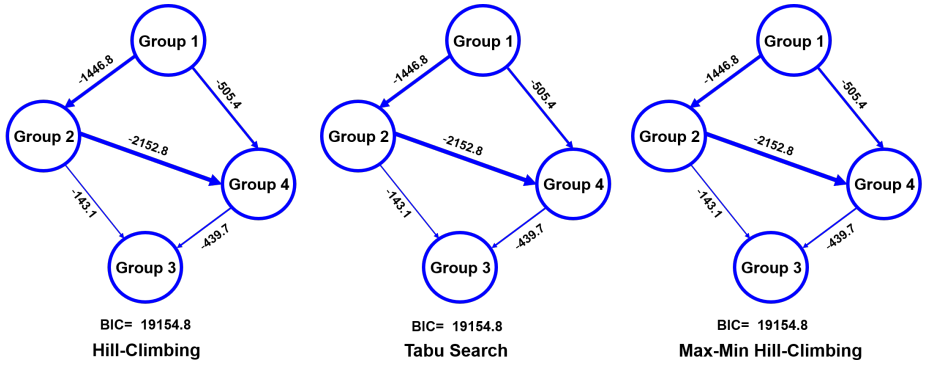


Figure 4.15: Bayesian networks from the unique joint pseudo-component spectra

Having identified the representative compounds of each pseudo-component from the spectra in Figure 4.14, the network structure among them resulting from the Bayesian networks of Figure 4.15 facilitate the development of a plausible reaction sequence in Figure 4.16. It indicates that PC_1, which comprises orthosubstituted or naphtenaromatics with alcohol and carbonyl stretches in the substituent functional groups, undergoes hydrolysis and decarboxylation to result in the formation of esters (PC_2) and carboxylic acids (PC_4). The esters in PC_2 could further hydrolyse to produce carboxylic acids (PC_4) that upon cracking and hydrogen transfer could result in the final products (PC_3), as indicated by the reactions in Figure 4.16.

Reaction mechanisms from the individual decomposition of the FTIR data, as

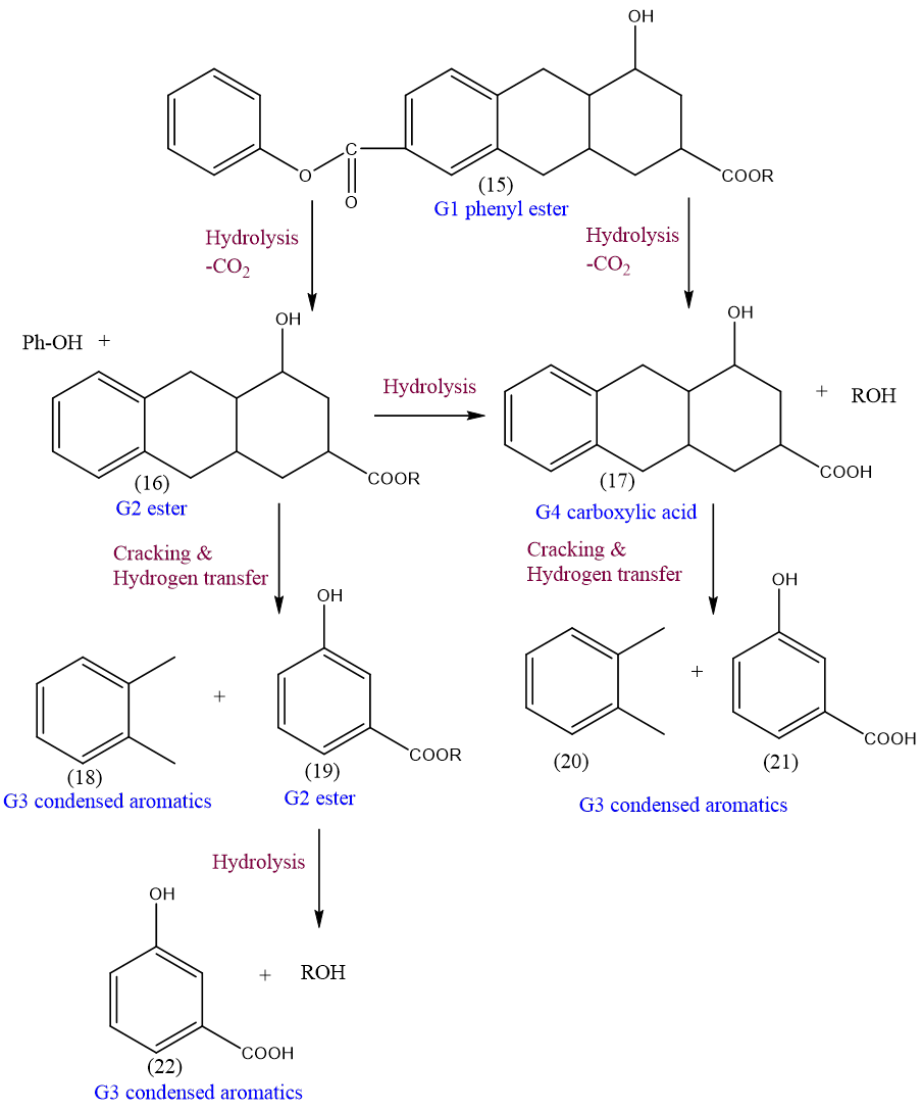


Figure 4.16: Proposed reaction hypothesis from jointly analyzing FTIR and ^{1}H-NMR data

shown in Figure 4.9, reveal that aromatics with substituent side chains of carboxylic acid and alcohol functional groups upon hydrogen transfer and cracking produce condensed aromatics and olefins on one hand, while also undergoing esterification on the other. The concentration profiles in Figure 4.13 support the sequence suggested by the Bayesian networks as they point to the increase in concentration of esters, olefins and condensed aromatics at higher temperatures and longer residence times. The importance of capturing complementary information about the functional groups from

FTIR data and the proton environment from the ^{1}H-NMR spectra by joint tensor decomposition has revealed that pseudo-components with similar FTIR peaks can be discriminated based on their NMR profiles, which indicates one to be the condensed aromatic by-product of the other owing to more pronounced aromatic and olefinic hydrogen peaks as opposed to its parent pseudo-component. Hence, the plausible reaction sequence that could be suggested based on the joint analysis in Figure 4.16 not only indicates the decarboxylation and hydrolysis of aromatics with carbonyl substituents to give esters, but also captures the further cracking and hydrolysis of esters to result in condensed carboxylic acids and aliphatic alcohols. It has thus been demonstrated that the data fusion framework of combining complementary information from spectral sensors captures additional arcs that could be interpreted as plausible reaction pathways, when probabilistic graphical models are used for structure learning among the spectral features.

The data fusion tensor decomposition has been implemented using a L_{21} norm for robustness to noise and outliers. Upon implementing the joint tensor decomposition using the L_2 norm under the assumption of the Gaussian noise in the process data, different reaction hypotheses have been arrived at, as outlined in Section C.5.3. The substituted aromatics upon cracking are seen to produce condensed alcohols that undergo esterification followed by thermal cracking of the esters to give condensed aromatics as the final product. It is worthwhile to note that both the robust and the Gaussian joint tensor decompositions capture similar reaction sequences of the cracking and esterification of the substituted aromatics, and further hydrolysis and cracking of the esters to result in condensed products; however, an additional decarboxylation pathway has been inferred from the robust formulation.

In our earlier work, where joint non-negative matrix factorization had been demonstrated as a data fusion algorithm to extract pseudo-component spectra, the number of components had been empirically determined using the notion of chemical rank [250]. The initially obtained reaction hypotheses were seen to indicate the cracking

and hydrogen transfer in substituted aromatics, as well as esterification to produce condensed aromatics and esters. However, heuristically relaxing rank to keep in check the truncation of chemical information was shown to reveal additional plausible pathways of esters further cracking and tautomerizing to produce condensed aromatics with an olefinic stretch. However, in this paper, the use of multi-linear based core consistency for rank determination, and the higher order tensor decompositions that are structure preserving and robust to process data noise are found to limit such a loss of chemical information.

4.5 Conclusions

This work has demonstrated tensor decomposition as a structure-preserving approach of jointly parsing spectral measurements from multiple sensors. The spectra are recorded over varying conditions of the two process modes of temperature and residence time, and the spectral channel mode of wavenumbers and chemical shifts for the FTIR and ^{1}H-NMR sensors, respectively. The structural information is preserved by capturing the inter-modal interactions while projecting the spectroscopic data onto each of the modes, resulting in a unique decomposition. This obviates the need to incorporate prior knowledge-based regularization constraints to limit the rotational and intensity ambiguities. The non-negativity constraints on the latent factors facilitates physical interpretation in accordance with Beer's law, owing to which the projections onto the process modes have been interpreted as concentrations of the pseudo-components across the varying temperatures and residence times, while the latent factor from projection onto the mode of the spectral channels have been interpreted as the FTIR and ^{1}H-NMR pseudo-component spectra. The number of pseudo-components or the number of components in each of the latent factor modes have been determined using the multi-linear metric called the core consistency diagnostic. A causal structure among the pseudo-component spectra is learned using Bayesian structure learning to infer plausible reaction hypotheses. The spectral peaks

that indicate representative compounds of the pseudo-components and the qualitative insights into their concentrations over temperature and residence time in conjunction with the inferred network structure has been central to the data-driven development of illustrative plausible reaction pathways for the complex reactive system, which are validated using domain knowledge.

Jointly decomposing FTIR and ^{1}H-NMR data is seen to capture complementary information by way of disambiguating pseudo-components that share similar FTIR peaks owing to common functional groups, but differ in terms of the proton environment as revealed by the corresponding ^{1}H-NMR spectra of the pseudo-components. The spectral features from the fused analysis have revealed additional structural paths in the Bayesian networks that point to added conversion pathways. These tensor decompositions have been implemented to handle process data artefacts like missing observations by imputations, non-Gaussian noise by formulating the objective to minimize a robust norm and have been parallelized to handle large amount of process data by dividing the tensors into sub-tensors prior to grid tensor factorization. The fused tensor decomposition with the robust norm is seen to compare well with the Gaussian norm, except for additional peaks being captured in the pseudo-component spectra of the robust case, pointing to additional conversion paths when interpreted using the Bayesian networks. The joint tensor decomposition is also seen to limit the loss of chemical information, owing to its higher order structure-preserving nature as compared to matrix-based data fusion techniques, where heuristically relaxing the chemical rank was shown to capture additional reaction pathways.

In the context of automating the discovery of reaction mechanisms in complex reactive mixtures, it is crucial to identify the species and the reaction pathways among them in the absence of prior knowledge of the system. The present study demonstrates a data-driven approach of species identification by obtaining unique pseudo-component spectra from tensor decompositions that represent molecular candidates; followed by using probabilistic graphical models to learn a causal structure

among the latent representation of molecular candidates as a way of hypothesizing reaction mechanisms. The semantic descriptions of the chemical mechanisms inferred from the Bayesian networks, along with the concentration trends across temperature and residence times conditions could be used in the future for diagnostic decisions in automation and control.

Chapter 5

Real-time monitoring of reaction mechanisms from spectroscopic data using hidden semi-Markov models for mode identification

Abstract

In this work, we present a framework for process monitoring focusing on the dynamics of reaction mechanisms based purely on online spectroscopic data. This is accomplished by developing an explicit duration hidden semi-Markov model (HSMM) that is used to monitor changes in reaction mechanisms with changing temperatures in a complex reacting system by dynamically identifying groups of spectroscopic samples that belong to a mode, and the mode duration. An expectation maximization algorithm is used for parameter re-estimation, and Viterbi state decoding is used to identify the most likely sequence of hidden states that may have generated the observation sequence. The reaction mechanism associated with samples of a mode is then deduced by extracting latent features among spectra of the mode and learning a probabilistic graphical structure among the features using Bayesian networks, which represent a network or mechanism of hypothesized reactions. The technique is demonstrated on case studies related to the partial upgrading of bitumen using thermochemical conversion based on the acquisition of Fourier transform infrared spectroscopic data; this system is complex enough that prior information on both

species and reactions is unavailable. Both offline and online monitoring are implemented, and the technique provides monitoring of the multi-modal process and, at the same time, provides insight into the chemistry specific to each mode, which makes it useful both for process control and fundamental studies into process chemistry.

5.1 Introduction

Spectroscopic data containing molecular-level information offers significant promise to decipher the underlying mechanisms in complex reactive systems through data-driven machine learning methods [25], [250], [282]. Monitoring such complex systems is key to automating supervisory control that is vital to process safety and product quality [283]. Process monitoring approaches can broadly be categorized as [284],[285],[286]: (a) model-based techniques using mechanistic models of the system, (b) knowledge-based expert systems relying on the accumulation of past experience, and (c) data-based systems that are not limited by the inability to fundamentally characterize a system, as is the case with the former approaches. In-situ detection and the elimination of sample post-processing have made spectroscopic sensors a low-cost option for online monitoring [287], reaction trajectory optimization to maximize product yield [288], and in designing reactive processes at the Pareto front of environmental and economic objectives [289]. Central to the shift of process systems engineering (PSE) principles from model-based methods for optimization, control and monitoring[290]; are multivariate statistical process monitoring (MSPM) models developed using in-line spectral data to monitor reactive systems [291],[292]. Spectral data have been supplemented by mechanistic models for bioprocess monitoring [293]; and by estimates from MSPM models acting as digital twins/soft sensors to limit uncertainty in the measured data, while estimating target variables for model-based control [294]. Although mechanistic models are favorable in monitoring and control due to their interpretability, there exists a significant knowledge gap owing to the complexity in developing such models for reactive systems as bitumen [14], being investigated in

this work. MSPM models have been used in conjunction with latent feature extraction, curve resolution [273] and calibration models on spectral data [295] to facilitate interpretation by correlating the data to physically meaningful quantities like concentration or particle size distribution [296], [297] to enable making inferences about reaction engineering systems [111] even in the absence of first principles models.

The use of MSPM to extract knowledge from process data for decision support has been the primary goal of data mining in process analytics [298]. The MSPM test statistics used for fault detection assume linear correlation among process variables, Gaussian distributions, sample independence and stationarity [299], [300]. Developments in process monitoring to handle non-Gaussian behaviour using latent variable models and mixture models, and non-linearity among the process variables is tackled using kernel methods, while adaptive and moving window techniques are used for non-stationarity in process data [301]. The use of robust feature extraction techniques in latent variable models handles the outliers in noisy high dimensional process measurements[302]. Hidden Markov models (HMM) not only overcome the need to used modifications to MSPM models that follow simplifying assumptions but is also inherently robust to process noise and measurement uncertainties, besides accounting for the temporal dependence in multi-modal data arising from setpoint change of process variables due to different product specifications or changes in operating conditions [303], [304]. In this paper, online Fourier transform infrared (FTIR) spectroscopic data collected across varying temperatures and residence times during the partial upgrading of Cold Lake bitumen is used to train an explicit duration HMM. The model handles uncertainty, captures the time scales and dynamics of the process to ultimately slice the spectral data in time-points via dynamic mode identification. Interpretability in the absence of a mechanistic model is facilitated by inferring reaction mechanisms from probabilistic structure learning among latent spectral features extracted by factor decomposition of the spectra associated with the identified modes [250]. Thereby, the proposed methodology is seen to achieve the real-time monitor-

ing of reaction mechanisms associated with changing process conditions in complex systems that lack mechanistic models.

5.1.1 Detailed background

Literature pertaining to mode identification in process data, followed by developing local models on the data of the identified modes for process monitoring has been reviewed in this section. Further, literature pointing to achieve the same when using HMMs and its extended capabilities in modeling time-scales through duration distributions has also been presented.

Strategies for mode identification when process data statistics change with operating conditions in multi-modal processes involve static and dynamic methods. The global approach of clustering statistically similar observation data from a mode into a cluster [305],[306] paves the way for static mode identification as it assumes modes are independent. However, dynamic mode transitions have been captured by the use of Gaussian mixture models (GMM) to cluster process data into modes, followed by the Bayesian estimation of the state transition matrix based on the mode membership of the data history, that is updated upon receiving new data [307]. Dynamics of a process have also been captured by developing robust adaptive local models such as a Partial Least Squares (PLS) model for each mode followed by tracking mode transitions in a moving window of data by comparing the similarity of its PLS model with that of the local modes for online identification [210]. The similarity of process correlation in temporal slice of data is used to segment the sequence in condition-driven analytics, followed by using slow feature analysis to track the dynamics along the condition mode with changing process conditions[308].In the context of tracking mode transitions, the mathematically rich structure of HMMs that captures stochasticity in temporal dynamics using a doubly-embedded model structure of observation data being emitted probabilistically by hidden states representing the otherwise obscure source characteristics, has been historically used in speech recognition [309]. The

type of system being modeled by HMMs enables interpretability of the hidden states as regimes in financial markets [310], machine degradation states in condition-based monitoring [311], operating modes characterized by ranges of analyte concentration in pharmaceutical production [312], and user preference when used for change point detection to generate sequential recommendations [313]. In the event that process shifts are driven by standard operating procedures, HMMs with mode reachability constraints in state decoding via the Viterbi algorithm adds interpretability to the mode dynamics and also reduce model complexity [314], [315].

HMMs are trained offline followed by using the model to infer in real-time the optimum dynamic sequence of modes using the Viterbi algorithm on a moving window of data, instead of identifying modes independently for the samples using the maximum *a posteriori* probability [316]. MSPM models like mixture principal component analysis (PCA), PLS [317], kernel PCA [318] and independent component analysis (ICA) [319] are then fit locally to the data from each of the modes, followed by kernel density estimation to determine the threshold for the MSPM test statistics *viz.* the Hotelling's T^2 and squared prediction error (SPE) for the detection of faults in each mode. The combined approaches of using MSPM for fault detection in the localized modes identified by HMMs, process knowledge structure encoded as Bayesian networks for mode diagnosis [320], or the use of models like self-organizing feature maps [321], PCA, ICA for feature extraction as a dimension reduction pre-processing step is seen to increase the sensitivity of fault detection algorithms by reducing false alarms and even the computational load of HMMs [322]. A probability ratio strategy based on the premise that probability of an observation generated by its own mode is far greater than the rest if the mode is stable, rather than if it is transitional when two or more modes have comparable probabilities of generating the observation, has been used prior to state decoding in the stable modes [323]. Subsequently, localized fault detection among the identified states of the stable modes using information theoretic-based novel process monitoring indices like the Mahalanobis distance to capture local

information, the negative log-likelihood that encompasses global information or the weighted combination of both is seen to outperform HMM-PCA, HMM-ICA in monitoring performance [324].

Typically, in a HMM, the state duration or the number of observation samples emitted while in a state inherently follows a geometric distribution [309] and does not take into account the dependence of a given mode and its duration on that of the previous mode. This lacuna while modeling processes that have modes with an explicit sojourn time like in condition-based monitoring for the predictive maintenance of machines to estimate the remaining useful life of machines [325] or pattern recognition for change point detection in genome sequences of continuous segments [326], is overcome by using a generalized HMM also called a Hidden Semi-Markov Model (HSMM). The ability to model the state as a *complex* entity comprising not only the state but also its duration, resulting in a transition being described as a function of the previous state and its duration, is seen to make the HSMM a superior modeling choice. Evidence of diagnosing process operating conditions by modeling transition probabilities as a function of a mode indicating scheduling variable is seen to account for the asymmetric temporal transitions of the different modes [327]. However, accounting for the mode and its duration as a complex state in HSMMs, facilitates tracking the probabilities of both mode transitions and duration while modeling temporal state dynamics [328]. It comes at the expense of model complexity, which can be handled either by using efficient algorithms for parameter estimation using incremental mapping over the conventional Expectation Maximization (EM) Baum Welch algorithm [329], or by leaning towards parsimonious models by incorporating constraints [314], [315] and relaxing assumptions of the complex state in a HSMM [328]. The simplest and least computationally complex HSMM is the explicit duration model that assumes transition to be independent of the previous state and duration to be conditioned upon the current state only [328]. HSMM with duration explicitly modeled using a non-parametric distribution has resulted in better mode

localization before using PCA for fault detection in each of the modes [330]. However, the model has many parameters to estimate and requires the *a priori* specification of the maximum duration in each state. In this work, an explicit duration HSMM with the duration modeled by a parametric Poisson distribution is proposed to limit the parametric complexity of a typical HSMM when used for the purpose of mode localization by capturing both the mode shifts and duration probabilities during the time-varying system of a complex reaction mixture. The main contributions of this paper are as follows:

1. Addresses limitations of HMMs in monitoring reactive systems using spectral data [331] by developing HSMMs that characterize both the dynamics and time-scales of a process.

2. Interprets the identified HSMM modes by constructing probabilistic graphical models for structure learning among latent features of data in localized observation segments of online spectra [250].

The rest of the paper is structured as follows: Section 5.2 describes the data generating process. Section 5.3 has a detailed description of the model, comprising Section 5.3.1 that outlines the distributions, assumptions and boundary conditions of the HSMM, Section 5.3.2 describes the Expectation-Maximization algorithm for parameter estimation of the HSMM, and Section 5.3.3 presents the use of the Viterbi algorithm for mode identification using the HSMM. Section 5.4 presents the findings of the mechanisms inferred in the locally identified HSMM modes for 2 cases of an operating temperature signal, and consists of Section 5.4.1 that discusses model complexity, Section 5.4.2 that presents results for a decreasing temperature signal, Section 5.4.3 that presents the findings for a randomized temperature signal, Section 5.4.4 that investigates the impact of increasing the model complexity on the findings reported in Section 5.4.3, and Section 5.4.5 that discusses the results of online monitoring with real-time data for the cases of both the temperature signals. Fi-

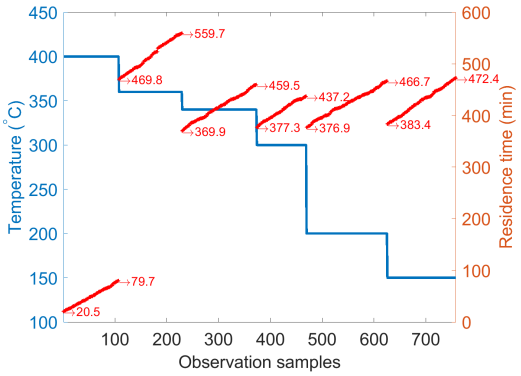
nally, Section 5.5 summarizes the work in the paper and highlights future avenues for research.

5.2 Problem description

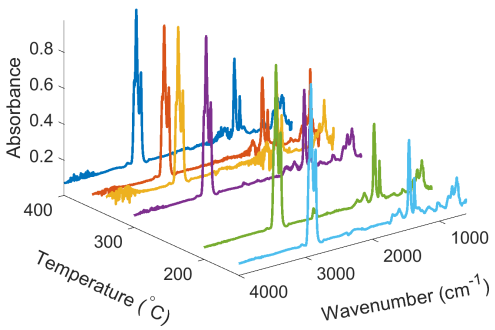
This work seeks to use an explicit duration HSMM model to monitor reaction mechanisms with changing temperatures by dynamically identifying groups of samples that belong to a mode. The reaction mechanism associated with samples of a mode is then deduced by extracting latent features among spectra of the mode and learning a probabilistic graphical structure among the features using Bayesian networks, as a way of hypothesizing reactions [250]. The proposed framework uses a sequence of spectroscopic observations to infer the time scales and dynamics of the reaction mechanisms associated with the identified modes, by way of the state duration and state transition probabilities of the HSMM, respectively.

The FTIR spectroscopic data [196], [197] of the products of low-temperature thermal cracking of Cold Lake bitumen over residence times at each temperature in the range of 150°C-400°C are shown in Figure 5.1. A total of 42 FTIR spectra were collected in that work, in addition to a measurement at 20°C and 0 min reaction time that was used for the purpose of baseline correction; these have been reported in Table D.1. The tabulated spectra are measured at fewer conditions of residence time at each of the reacting temperatures, prompting us to generate synthetic data that bears semblance to a continuous reaction process.

We have generated realistic synthetic spectra by random linear interpolation [332] of spectra at each temperature such that successive spectra differ by a sampling interval of ~ 10 seconds. This is followed by randomly sampling spectra over residence times in the range 60 to 90 min at half hour intervals, at each temperature as shown in Figure 5.1a. FTIR spectra at an intermediate reaction time for each of the temperatures have been shown alongside in Figure 5.1b, for the purpose of illustration. Each spectrum is an observation sample whose absorption intensities across the spec-



(a) Residence times associated with temperature



(b) Illustrative FTIR spectra

Figure 5.1: Spectral data over randomly sampled residence times at decreasing temperatures

tral channels changes during the visbreaking process across varying temperatures, thereby pointing to different reaction mechanisms. The synthetic spectral datasets that are used to develop the HSMM models to identify the modes and their state durations, have been published on the following link: https://github.com/Anjana-T-Puliyanda/HSMM-for-realtime-reaction-mechanism-monitoring. The physical interpretation of the identified modes as reaction mechanisms, and their durations as time scales of the mechanisms, is facilitated by developing causal maps among the latent factors extracted from the spectra of each mode using multivariate factor decomposition models. [250], [273]

5.3 Methods

This section outlines the model used to implement the explicit duration HSMM, the use of information theory-based metrics to limit model complexity, the EM algorithm for parameter estimation from the process data and finally the Viterbi algorithm for deducing the globally optimal dynamic sequence of states underlying the observation data for the given trained HSMM. The states are then structurally characterized using the associated spectra to deduce the reaction mechanism of the process mode

[250]. State, in the context of a HSMM for our system, refers to a complex state that encompasses not just the notion of the reaction mechanism characterizing the spectra emitted in a particular operating mode but also the duration for which it remains in that mode.

5.3.1 Model description

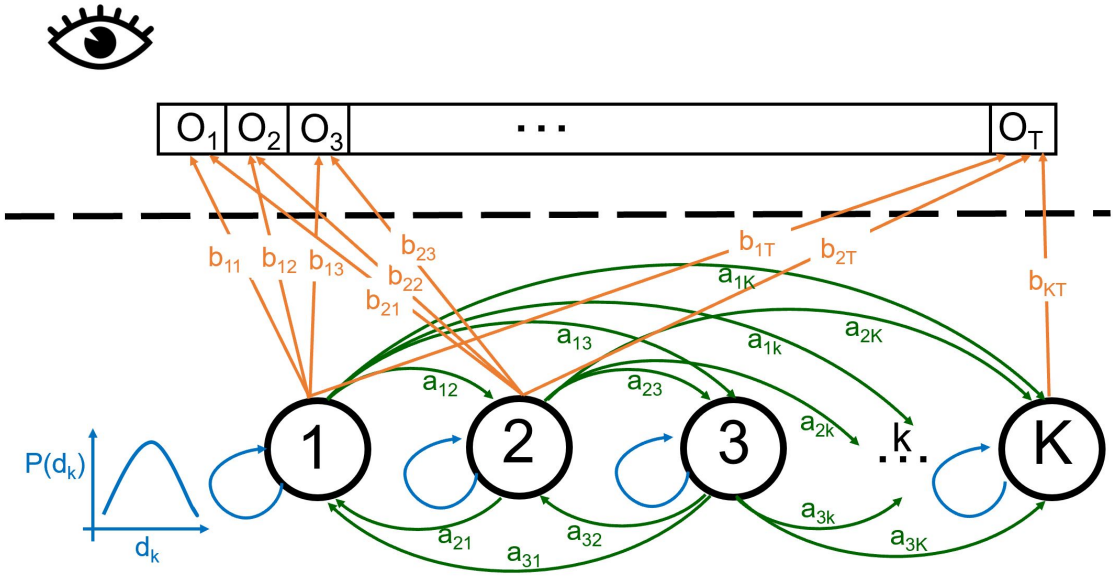


Figure 5.2: Schematic representation of the explicit duration HSMM as a doubly embedded stochastic process

The HSMM is a doubly embedded process that captures the dynamic transitions among the states that persist over durations [309],[333], stochastically generating the given observation sequence as illustrated in Figure 5.2. The observation sequence of given length T is denoted as $O_{1:T} = \{O_1, O_2, \cdots O_T\}$, where each observation $O_t \in \mathscr{R}^d$ is a multivariate data sample. The discrete sequence of hidden states is given by $S_{1:T} = \{(i_1, d_1), (i_2, d_2), \cdots (i_N, d_N)\}$, where the state $i_n \in \mathbb{S} = \{1, 2, \cdots K\}$ corresponds to the chemical mechanisms associated with changing operating conditions belonging to one of the K modes, and its duration $d_n \in \mathbb{D} = \{1, 2, \cdots T\}$ is a random variable such that $\sum_{n=1}^{N} d_n = T$. Since the number of observations a state

can emit is discrete, the mode duration is explicitly modeled using a Poisson distribution in this work, as has been done for modeling state duration distributions for automatic speech recognition [334]. The probability of the i_n mode having a duration d_n given by $P(S_{t+1:t+d_n} = i_n)$, where $t = \sum_{l=0}^{n-1} d_l \, \forall \, n \in [1, N]$ is given in Equation 5.1, which is parametric in λ_{i_n}, which is the average duration of the i_n mode.

$$P_{i_n}(d_n) = e^{-\lambda_{i_n}} \frac{\lambda_{i_n}^{d_n}}{d_n!} \tag{5.1}$$

Generally, the first and the last state can exist before and after $t = 1$ and $t = T$ respectively, if the process ranges from $(-\infty, +\infty)$ [309]. However, we assume that the first state begins at t=1 and the last state ends at t=T. Hence, the initial state (i_0, d_0), is defined using a state distribution given in Equation 5.2, under the assumption that its duration is zero, i.e., $d_0 = 0$ as this is prior to obtaining the first sample.

$$\pi_{i_0} = P(S_t = i_0) \qquad \forall \qquad t \leq 0 \qquad \text{S.T.} \qquad \sum_{i_0 \in \mathbb{S}} \pi_{i_0} = 1 \tag{5.2}$$

The probability of state transitions $(i_{n-1}, d_{n-1}) \rightarrow (i_n, d_n) \, \forall n \in [1, N]$, $\forall i_{n-1}, i_n \in \mathbb{S}$, $\forall d_{n-1}, d_n \in \mathbb{D}$ is denoted by $a_{(i_{n-1},d_{n-1})(i_n,d_n)} = P(S_{t+1:t+d_n} = i_n | S_{t-d_{n-1}+1:t} = i_{n-1})$ such that $i_n \neq i_{n-1}$, as the state i_{n-1} ending at time t cannot transition to itself at time t+1, because the state durations have been modeled explicitly. Equation 5.3 indicates that the transition probability can be simplified under the assumption of an explicit duration HSMM such that the transition is independent of the duration of the previous state and the duration of the present state is conditioned only upon itself [328].

$$a_{(i_{n-1},d_{n-1})(i_n,d_n)} = a_{i_{n-1},i_n} P_{i_n}(d_n) \quad \forall i_{n-1}, i_n \in \mathbb{S}, \forall d_n \in \mathbb{D} \quad \text{S.T.} \sum_{i_n \neq i_{n-1}} \sum_{d_n} a_{(i_{n-1},d_{n-1})(i_n,d_n)} = 1 \tag{5.3}$$

The probability of emitting d_n observations while in mode i_n, is modeled using a mixture of Gaussian distributions as the observations are continuous multivari-

ate spectra [335], [336]. The emission probability of the observations is denoted by $b_{i_n,d_n}(O_{t+1:t+d_n}) = P(O_{t+1:t+d_n}|S_{t+1:t+d_n} = i_n)$. The observations are assumed to be conditionally independent given the mode, leading to an expression for the emission probability distribution as given in Equation 5.4. The number of mixture components are denoted by M, while the $C_{i_n m}$ are the mixing weights of each of the multivariate Gaussian distributions denoted by $\mathscr{N}$.

$$b_{i_n,d_n}(O_{t+1:t+d_n}) = \prod_{\tau=t+1}^{t+d_n} b_{i_n}(O_\tau) = \prod_{\tau=1}^{d_n} \sum_{m=1}^{M} C_{i_n m} \mathscr{N}(O_\tau, \mu_{i_n m}, \Sigma_{i_n m}) \qquad \forall i_n \in \mathbb{S}, d_n \in \mathbb{D}, M \geq 1$$
$$\text{S.T.} \qquad C_{i_n m} \geq 0, \qquad \sum_{m=1}^{M} C_{i_n m} = 1 \tag{5.4}$$

5.3.2 EM algorithm for parameter re-estimation

It can be seen from Section 5.3.1 that the complete specification of the given HSMM model involves specifying the length of the observation sequence T, the dimension of each observation $O_t \in \mathscr{R}^d$, the number of hidden states K and the mixture components M in the Gaussian distribution of the emission probabilities. The parameters of the resulting model can be outlined as follows:

- The initial state distribution π comprising K-1 parameters.
- The average duration for state duration distribution λ comprising K parameters.
- The state transition probability matrix $A \in \mathscr{R}^{K \times K}$ where self-transitions are not allowed $a_{ij} = 0 \, \forall i = j$ as state duration is explicitly modeled, thereby comprising K(K-2) parameters.
- The emission distribution characterized by the mixing coefficient matrix $C \in \mathscr{R}^{K \times M}$ comprising K(M-1) parameters, the mean $\mu \in \mathscr{R}^{K \times d}$ and covariance $\Sigma \in \mathscr{R}^{K \times 1}$ of the multivariate Gaussian with KdM and KM parameters respectively.

All of the above enumerated model parameters can be collectively represented as $\Theta = (\pi, \lambda, A, C, \mu, \Sigma)$ for simplicity. These parameters are estimated by the iterative routine of maximizing likelihood through expectation maximization using the Baum Welch forward backward algorithm until the estimated relative error of the parameters falls below a certain threshold [337], [309], [329].

The joint probability of observing a sequence of samples $O_{1:T}$ and states $S_{1:T}$ given the model parameters is given by Equation 5.5, where $t = \sum_{l=1}^{n-1} d_l$

$$P(S_{1:T}O_{1:T}|\Theta) = \prod_{n=1}^{N} P(S_{1:T})P(O_{1:T}|S_{1:T}, \Theta) = \prod_{n=1}^{N} a_{(i_{n-1},j_n)} P_{j_n}(d'_n) b_{j_n d'_n}(O_{t+1:t+d'_n}) \tag{5.5}$$

The likelihood of observation sequence is obtained by marginalizing the joint probability across all possible values of the state sequence as given in Equation 5.6

$$P(O_{1:T}|\Theta) = \sum_{S_{1:T} \in \{\mathbb{S},\mathbb{D}\}} P(S_{1:T}O_{1:T}|\Theta) \tag{5.6}$$

The computational effort involved in maximizing the above likelihood for parameter estimation using EM is simplified using the forward-backward algorithm based on probabilities of the partial observation and state sequences. They are defined using [338]: (i) forward variable $\alpha_t(j, d) = P(S_{t-d+1:t}, O_{1:t}|\Theta)$, which is the joint probability of the mode j existing for a duration d upto the current time t and the partial observation sequence until the current time step, given the model parameters (ii) backward variable $\beta_t(j, d) = P(O_{t+1:T}|S_{t-d+1:t} = j, \Theta)$, which is the conditional probability of observing the sequence of samples from the next time step to the end, given the model parameters and the state j existing for a duration d upto the current time. Based on the assumptions of the Markov property that the current/future observations are dependent on the current state and independent of the previous observations, and the conditional independence of the observations, the forward and backward variables can be recursively computed as indicated in Equation 5.7 and Equation 5.8.

$$\alpha_t(j,d) = \sum_{i\neq j\in\mathbb{S},h\in\mathbb{D}} P(S_{t-d-h+1:t-d}=i, S_{t-d+1:t}=j, O_{1:t}|\Theta)$$
$$= \sum_{i\neq j\in\mathbb{S},h\in\mathbb{D}} \alpha_{t-d}(i,h)a_{ij}P_j(d)\prod_{\tau=t-d+1}^{t} b_j(O_\tau) \quad \forall t>0 \qquad \text{S.T.}\sum_{i\neq j}\alpha_t(i)a_{ij} = \begin{cases}\pi_j & \text{if } t=0,\\ 0 & \text{if } t<0.\end{cases} \tag{5.7}$$

$$\beta_t(j,d) = \sum_{i\neq j\in\mathbb{S},h\in\mathbb{D}} P(S_{t+1:t+h}=i, O_{t+1:T}|S_{t-d+1:t}=j,\Theta)$$
$$= \sum_{i\neq j\in\mathbb{S},h\in\mathbb{D}} a_{ji}P_i(h)\prod_{\tau=t+1}^{t+h} b_i(O_\tau).\beta_{t+h}(i,h) \quad \forall t<T \qquad \text{S.T.}\beta_t(i) = \begin{cases}1 & \text{if } t=T,\\ 0 & \text{if } t>T.\end{cases} \tag{5.8}$$

It must be noted that marginalizing the above forward-backward recursive probability definitions over the duration results in $\alpha_t(j)$ and $\beta_t(j)$ $\forall j \in \mathbb{S}$, which are used to compute (i) the Probability of being in mode i at time t and mode j at t+1 given the model and the partial observation sequence $\xi_t(i,j) = P(S_t = i, S_{t+1} = j|O_{1:t},\Theta) = \alpha_t(i)a_{ij}\sum_{d\in\mathbb{D}} P_j(d).\prod_{\tau=t+1}^{t+d} b_j(O_\tau).\beta_{t+d}(j)\ \forall i,j\in\mathbb{D}, t\in[1,T]$ (ii) the Probability of being in state i at time t given the model and the partial observation sequence $\gamma_t(i) = P(S_t = i|O_{1:t},\Theta) = \sum_{j=1}^{K}\xi_t(i,j)\ \forall i,j\in\mathbb{D}, t\in[1,T]$. Since the probability of being in a state at a certain time takes into account all M components of the Gaussian distribution for the emission probability of the observation at that instant, the probability of the m^{th} component's contribution alone is weighted as follows:

$$\gamma_t(i,m) = \gamma_t(i)\frac{C_{im}\mathcal{N}(O_t,\mu_{im},\Sigma_{im})}{\sum_{l=1}^{M} C_{il}\mathcal{N}(O_t,\mu_{il},\Sigma_{il})} \tag{5.9}$$

The parameters of the HSMM are then re-estimated using the above computed probabilities over multiple iterations of the Baum Welch algorithm using Equations 5.10- 5.15 [339], [328], [340]:

$$\bar{\pi}_i = \gamma_1(i) \tag{5.10}$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \sum_{j=1,\neq i}^{K} \xi_t(i,j)} \tag{5.11}$$

$$\bar{\mu}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m) O_t}{\sum_{t=1}^{T} \gamma_t(i,m)} \tag{5.12}$$

$$\bar{\Sigma}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)(O_t - \mu_{im})(O_t - \mu_{im}^T)}{\sum_{t=1}^{T} \gamma_t(i,m)} \tag{5.13}$$

$$\bar{C}_{im} = \frac{\sum_{t=1}^{T} \gamma_t(i,m)}{\sum_{t=1}^{T} \gamma_t(i)} \tag{5.14}$$

$$\bar{\lambda}_i = \frac{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i).(t_1 - t_0 + 1)}{\sum_{t_0=1}^{T} \sum_{t_1=t_0}^{T} \chi_{t_0,t_1}(i)} \tag{5.15}$$

where $$\chi_{t_0,t_1}(i) = \frac{\sum_{j=1,\neq i}^{K} \alpha_{t_0-1}(j) a_{ji} \prod_{\tau=t_0}^{t_1} b_i(O_\tau) P_i(t_1 - t_0 + 1) \beta_{t_1}(i)}{P(O_{1:T}|\Theta)}$$

5.3.3 Viterbi state decoding

Once the parameters of the HSMM have been learned by maximizing the expectation across the observation data sequence, the most likely sequence of hidden states that may have generated the observation samples is decoded using the Viterbi algorithm. The Viterbi algorithm is designed to find the global optimal sequence of states by maximizing the probability of the observation data sequence conditioned on all possible hidden state sequence combinations [309], [328]. The most likely partial state sequence that ends at time t in state j of duration d, $\delta_t(j,d)$ and the partial best path

that achieves the maximum probability $\Psi(t, j, d)$ is recorded by caching the ending time of the state, the state and its duration as backpointers for traceback [340]. The probability of the best partial state sequence is computed recursively whilst caching its path $\forall 1 \leq t \leq T, j \in \mathbb{S}, d \in \mathbb{D}$ as given in Equations 5.16-5.18. The previous state selected by $\delta_t(j, d)$, and its ending time is recorded in $\Psi(t, j, d)$, as given in Equation 5.18 where i^* is the previous state that has survived, h^* being its duration and (t-d) is its ending time.

$$\delta_t(j,d) = \max_{i \neq j \in \mathbb{S}, h \in \mathbb{D}} \delta_{t-d}(i,h) a_{ij} P_j(d) \prod_{\tau=t-d+1}^{t} b_j(O_\tau) \tag{5.16}$$

$$(\mathbb{S}^*, \mathbb{D}^*) = _{i \neq j \in \mathbb{S}, d \in \mathbb{D}} \ \delta_{t-d}(i,h) a_{ij} P_j(d) \prod_{\tau=t-d+1}^{t} b_j(O_\tau) \qquad \delta_t(j,d) = \begin{cases} \pi_j & \text{if } t = 0, \\ 0 & \text{if } t < 0. \end{cases} \tag{5.17}$$

$$\Psi(t,j,d) \equiv t - d, i^*, h^* \ \text{ where, } i^* \in \mathbb{S}^*, h^* \in \mathbb{D}^* \tag{5.18}$$

The partial probability computed recursively in Equation 5.16 differs from those calculated in the forward-backward algorithm, since it represents the probability of the most likely path to a state j at time t, and not the cumulative probability of all paths to the state. With this in mind, the state decoding is performed by looking at the whole sequence to estimate the most probable state sequence $(j_n^*, d_n^*), (j_{n-1}^*, d_{n-1}^*), \cdots, (j_1^*, d_1^*)$ by tracing back the cached pointers to decipher the globally optimal path as outlined in Equations 5.19-5.20.

$$(\mathbb{S}_1^*, \mathbb{D}_1^*) = _{j \in \mathbb{S}, d \in \mathbb{D}} \ \delta_{t_1}(j,d) \qquad \text{where, } \ t_1 = T \tag{5.19}$$

$$t_n, j_n^*, d_n^* = \Psi(t_{n-1}, j_{n-1}^*, d_{n-1}^*) \qquad \text{where, } \ j_{n-1}^* \in \mathbb{S}_{n-1}^*, d_{n-1}^* \in \mathbb{D}_{n-1}^* \text{ and,}$$
$$n = 2, 3, \cdots \text{until } \ t_n - d_n^* + 1 \leq 1 \tag{5.20}$$

For a detailed description of the methods used to develop reaction mechanisms for each mode that has been identified, we refer readers to our previous work [250].

5.4 Results and Discussion

In this section, we discuss the results obtained by the mode localization of spectral measurements collected across temperature sequences in the range of 150 °C-400 °C, with the process residing at each temperature for randomly sampled time durations in the range 60-90 minutes. All the spectra belonging to identical modes are then queried into a factor decomposition model to obtain latent spectral features, and structure learning in Bayesian networks is used as a graph theoretic approach of causally inferring reaction mechanisms from these spectral features. The hypothesized reaction mechanisms associated with each of the identified modes resembles the underlying temperature sequence. Results presented in Section 5.4.2 monitor reaction mechanisms over a decreasing temperature sequence, while Section 5.4.3 presents mechanisms of the identified local modes for a randomized temperature sequence using a HSMM model with 4 states, and Section 5.4.4 discusses the findings if a larger number of states were used instead. The decreasing temperature sequence is representative of 'optimal' recipes that are developed offline, and the randomized temperature sequence is representative, from the viewpoint of the HSMM modeling, of changes to temperature that may be made by a feedback controller.

5.4.1 Model complexity

It can be seen from Section 5.3.2 that the total number of parameters for a given HSMM model scale according the choice of the number of hidden states K and the number of mixture components M chosen to model the emission probability distribution. A parsimonious model with fewer parameters is generally preferred to minimize computational effort; hence, a choice of M=2 has been used to model the GMM of the emission probabilities in this work. However, order selection of the HSMM concerning the choice of the number of hidden states is done using the maximum likelihood approach based on information criteria like the Akaike information criteria (AIC)

and the Bayesian information criteria (BIC) [341],[337]. Equations 5.21-5.22 define the information criteria that are to be minimized in order to maximize the model likelihoods, while being used to determine the model order.

$$AIC = -2\log(P(O_{1:T}|\Theta)) + 2\mathrm{N}_{\mathrm{params}} \tag{5.21}$$

$$BIC = -2\log(P(O_{1:T}|\Theta)) + \mathrm{N}_{\mathrm{params}}\log(T) \tag{5.22}$$

The aforementioned approach for model order selection could lead to the selection of more states than required for a particular model, as additional states may capture fine-grained data structure, i.e., outliers, heterogenity, thereby demanding a pragmatic approach of using practical system knowledge besides these criteria for applications concerning interpretability of these states [342].

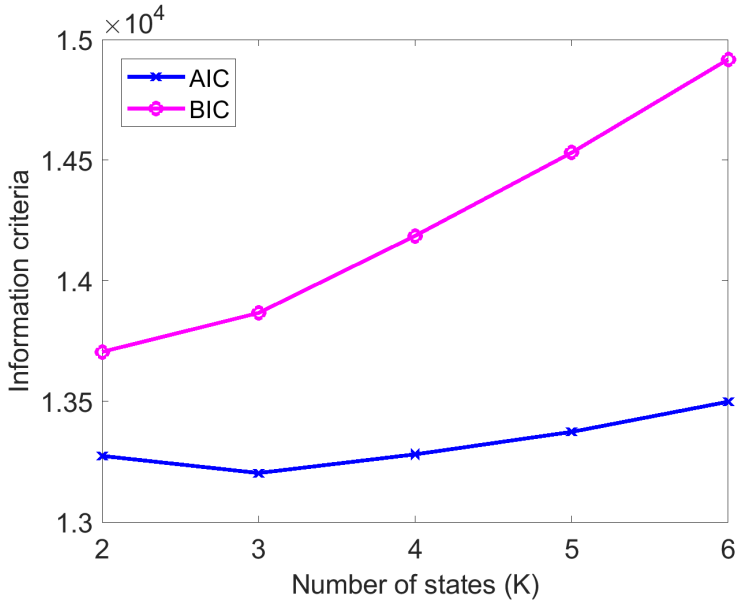


Figure 5.3: Model order selection based on information criteria to maximize model likelihood

In this work, the states are interpreted as the reaction mechanisms associated with the operating conditions during the partial upgrading of bitumen. These reaction mechanisms are inferred probabilistically using structure learning of Bayesian networks [250] among features extracted from the spectra of the associated states. The

Viterbi algorithm is used to decode the state sequences dynamically, as a way of tracking how the reaction mechanisms transition across varying temperature conditions. There are 6 different temperature conditions, because of which it is believed that the number of states may not exceed 6, even if there were distinct reaction mechanisms associated with each of the temperatures. The information theoretic criteria used to assess the model complexity that adequately fits the data differ in terms of how the free parameters of a model are penalized to limit overfitting. The penalty term in the BIC incorporates a prior notion of the data generating process in terms of its sequence length, in addition to the model parameters and runs the risk of underfitting, especially when the length of the observation sequence is large. However, the AIC aims at finding the best model when the data generating process is unknown.These information criteria obtained by varying the model order upto a maximum of 6 states have been reported in Figure 5.3, from which it can be seen that models with fewer states than the temperature conditions perform better, with the optimal being 3 states because the least AIC is achieved for a model with 3 states, while for the BIC only a marginal increase has been observed in going from from 2 to 3 states as opposed to further increasing the model complexity. However, simulations in Section 5.4.2 and Section 5.4.3 have been reported with the choice of an additional state to capture aberrations in the data structure as discussed earlier [342]. The performance of the HSMMs in tracking reaction mechanisms associated with a decreasing temperature sequence has been investigated in Section 5.4.2, while an investigation on the similar lines in the event of a randomized temperature sequence is outlined in Section 5.4.3. Further, Section 5.4.4 elucidates insights into tracking the reaction dynamics accompanying the aforementioned randomized temperature sequence with a fine grained HSMM comprising a larger number of states.

5.4.2 Decreasing temperature sequence

Randomly sampled spectra over a decreasing temperature sequence of 400°C, 360°C, 340°C, 300°C, 200°C and 150°C, at residence times of an hour or an hour and a half at each of the temperatures have been used to train a HSMM model. The model is trained with 4 states, in accordance with the AIC/BIC model complexity curves shown in Figure 5.3. Since the states are physically interpreted as the mechanisms associated with the underlying temperature sequence, it is believed that the maximum number of states can only be as many as the number of unique operating temperatures. Also, it is physically possible that similar mechanisms exist across certain operating temperatures, owing to which a fewer number of states have been considered in accordance with information theoretic criteria to limit the model complexity. The localized modes identified by the Viterbi state decoding of the input spectral data have been plotted against the underlying decreasing temperature sequence is shown in Figure 5.4.

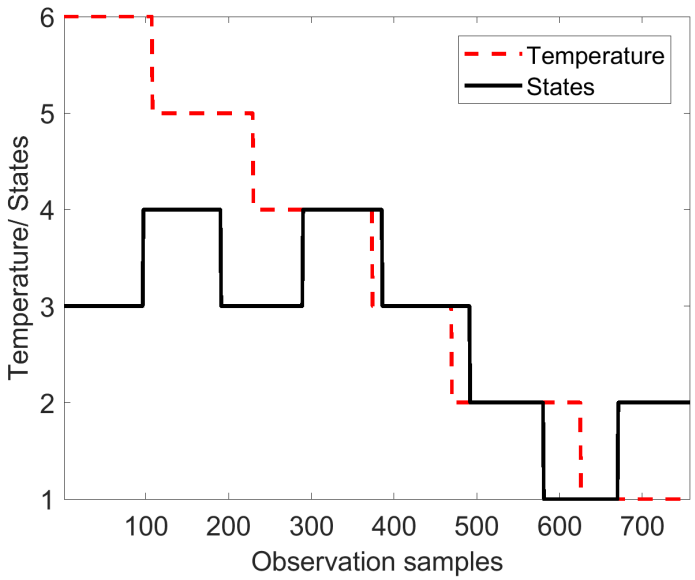
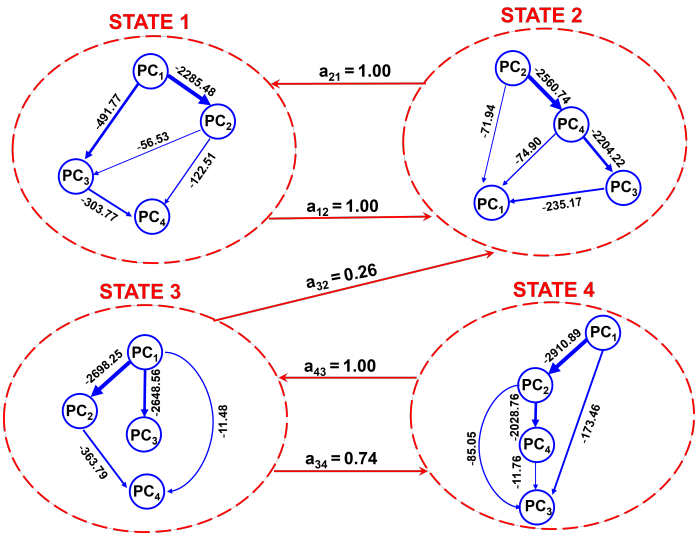


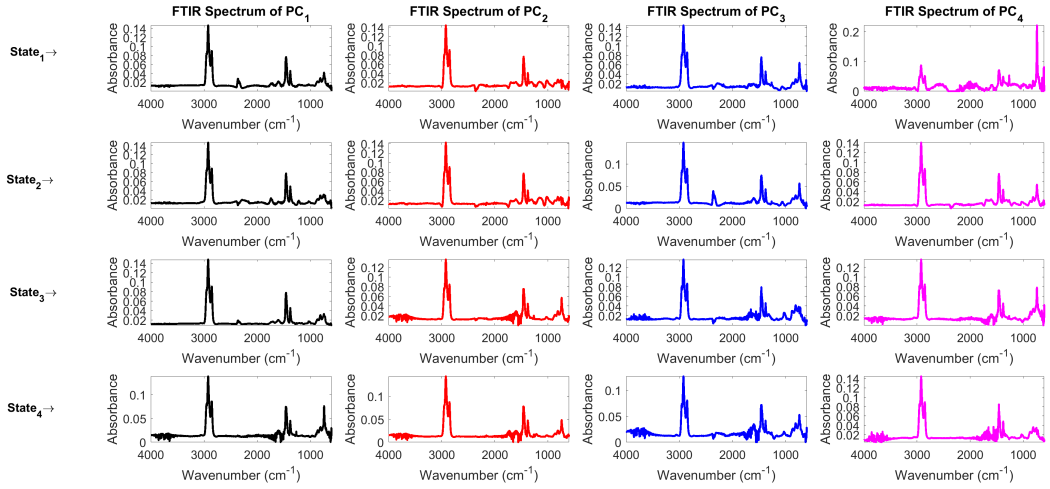
Figure 5.4: Mode identification of the decreasing temperature sequence

The temperature sequence has been scaled to lie in the range between 1 and 6, with 1 being the least temperature and 6 being the highest temperature. The du-

ration distributions of each of the modes is given in Figure D.1, and the posterior probabilities of the modes are given in Figure D.2. The reaction mechanisms inferred by Bayesian structure learning of all the spectra identified as a unique mode, and the latent spectral features of each mode are given in Figure 5.5a and Figure 5.5b, respectively.



(a) Reaction mechanisms of the modes with the state transition probabilities



(b) Pseudocomponent spectra associated with the modes

Figure 5.5: Reaction mechanisms associated with the pseudocomponent spectra of each state

It can be seen from Figure 5.4 and Figure D.1 that state 3 and state 4 represent

mechanisms associated with higher temperatures and persist over marginally longer duration as compared to that of state 2 and state 1, which represent lower temperature mechanisms. The spectral signatures of the pseudocomponents can provide indications of the nature of the reactions in the mechanisms. From the spectral signatures of the states given in Figure 5.5b, it can be seen that all the pseudocomponents exhibit strong absorption peaks at 2950 cm^{-1}, 2920 cm^{-1} and 2850 cm^{-1} indicating sp^3 C-H stretching, in addition to peaks at 1380 cm^{-1} and 1450 cm^{-1} that point to the C-H bending vibrations.

Besides these peaks, for state 3, PC_1 is seen to have peaks at 740 cm^{-1}, 810 cm^{-1} and 864 cm^{-1} indicating the presence of ortho, meta and para substituted aromatics, along with a peak at 1600 cm^{-1} corresponding to the C=C stretch and one at 1717 cm^{-1}, implying that it primarily constitutes esters. PC_2 has a sharper peak at 740 cm^{-1} that runs with a peak at 1607 cm^{-1}, pointing to naphthenearomatics, in addition to a very small peak at 1032 cm^{-1}, indicating the presence of aliphatic alcohols (C-O aliphatic stretch) or aliphatic straight chains with an alcohol functional group, attached as a substituent to the naphthearomatic ring. PC_3, aside from having peaks representing the substituted aromatics, has a peak at 1034 cm^{-1} (C-O aliphatic stretch) and 1600 cm^{-1} (C=C stretch), indicating that it may consist of condensed aliphatic alcohols and olefins. PC_4 also has peaks at 740 cm^{-1} and 1600 cm^{-1}, indicating that it primarily comprises naphthenearomatics. Having identified the compound classes belonging to each of the pseudocomponents, the reaction network associated with state 3 represents the hydrolysis of esters ($PC_1 \rightarrow PC_2$) to give naphthenearomatics with substituents of aliphatic straight chains with alcohol functional groups which upon cracking and hydrogen transfer results in olefins and substituted aromatics ($PC_2 \rightarrow PC_4$). The hydrolysis of esters, followed by thermal cracking and hydrogen transfer leads to the formation of condensed alcohols and olefins ($PC_1 \rightarrow PC_3$), while the direct decarboxylation and cracking of esters could also produce substituted aromatics and olefins ($PC_1 \rightarrow PC_4$).

Similarly, elucidating the compound classes in each of the pseudocomponents in State 4: PC_1 is seen to comprise naphthenearomatics and phenols owing to the peaks at 740 cm^{-1}, 1600 cm^{-1} and the C-O phenolic stretch at 1261 cm^{-1}. PC_2 has peaks at 740 cm^{-1}, 1600 cm^{-1} and 1730 cm^{-1}, indicating that it represents a class of o-substituted phenyl esters. PC_3 is also shown to contain peaks indicative of naphthenearomatics and the C-O aliphatic stretch of alcohols besides the C=C stretch at 1600 cm^{-1}, pointing to condensed alcohols. PC_4, in addition to comprising substituted aromatics, also comprises olefins and condensed alcohols. Hence, the reaction mechanism associated with state 4 corresponds to the reaction of phenols with carboxylic acid to give phenyl esters ($PC_1 \rightarrow PC_2$), the subsequent hydrolysis and cracking accompanied by free radical hydrogen transfer of which produces condensed alcohols and olefins ($PC_2 \rightarrow PC_4$ and PC_3). Thermal cracking of the saturated ring attached to the aromatic phenol ring could directly result in olefins, as well ($PC_1 \rightarrow PC_3$).

When it comes to characterizing the low temperature states, it can be seen that for state 2: PC_1 comprises phenols and esters, while PC_2 primarily constitutes condensed alcohols and phenols, PC_3 and PC_4 comprise naphthenearomatics, condensed alcohols and olefins. The reaction mechanism underlying state 2 can be hypothesized to include hydrogen transfer of condensed alcohols and phenols to produce naphthenearomatics, followed by cracking of side chain substituents to give condensed aliphatic alcohols ($PC_2 \rightarrow PC_4$ and PC_3). The alcohols and phenol functional groups present in PC_2, PC_4 and PC_3 potentially combines with carboxylic acids to yield esters, unconverted phenols and substituted aromatics (PC_2, PC_4 and $PC_3 \rightarrow PC_1$).

For the other low temperature state, i.e., state 1, PC_1 encompasses phenyl esters due to peaks at 740 cm^{-1}, 1600 cm^{-1} and 1738 cm^{-1}. PC_2 includes condensed alcohols and phenols, PC_3 has o-substituted phenols and alcohols, while PC_4 has naphthenearomatics and olefins. Hence, the reaction mechanism representing this state primarily focuses on hydrolysis and cracking of phenyl esters to give condensed alcohols and phenols ($PC_1 \rightarrow PC_2$), which further undergo cracking and hydrogen

disproportionation to give olefins and susbstituted aromatics ($PC_2 \rightarrow PC_3$ and PC_4). The phenyl esters could also hydrolyze to give phenols and alcohols ($PC_1 \rightarrow PC_3$), which on further cracking and hydrogen transfer result in olefins and naphthenearomatics ($PC_3 \rightarrow PC_4$).

It can be seen that the reaction mechanisms over decreasing temperatures point to the hydrolysis, cracking and decarboxylation of esters at higher temperatures to result in substituted aromatics, condensed alcohols and phenols. The alchols and phenols may reversibly combine with acids present, to form esters that again undergo hydrolysis and milder cracking at slightly lower temperatures. The decreasing severity of thermal cracking accompanied by the hydrolysis of esters followed by the reaction of the phenols and alcohols thus formed, to reversibly form esters again, leads to the formation of lighter aliphatics and olefins produced at the expense of substituted aromatics, condensed alchols and phenols.

5.4.3 Randomized temperature sequence using 4 states

This section investigates the mode localization performance of the HSMM, when spectroscopic data over a residence time of 1 -1.5 hours is randomly sampled at each of the 6 temperatures occurring in an arbitrary sequence. The model has been trained using 4 states; however, the identification of the reaction mechanisms (modes) associated with the temperature signal, as shown in Figure 5.6, is seen to visit only 3 states, as corroborated by the posterior probability of the states given in Figure 5.7, where state 3 is found to have the least posterior probability at all sample times because of which the process does not transit through this mode in the optimal Viterbi sequence. The duration distribution of the states in Figure D.3 indicates that state 1 and state 2, which are the frequently encountered reaction mechanisms, have a slightly higher model residence time, as compared to state 4 (which is the mechanism prevalent at 300 °C and 150 °C and early stages of 200 °C). The reaction mechanisms associated with each of the modes are given in Figure 5.8, with the corresponding

pseudocomponent spectra of each of the states in Figure D.4. The compound classes are enumerated on the basis of the pseudocomponent spectra of the states, and the reaction mechanisms associated with the corresponding states are hypothesized by using Bayesian networks. The pseudocomponent spectra are modeled as the random variables at the nodes among which the directed acyclic paths are learned via score search optimization routines, to arrive at a network structure that maximizes the Bayesian Information criteria.

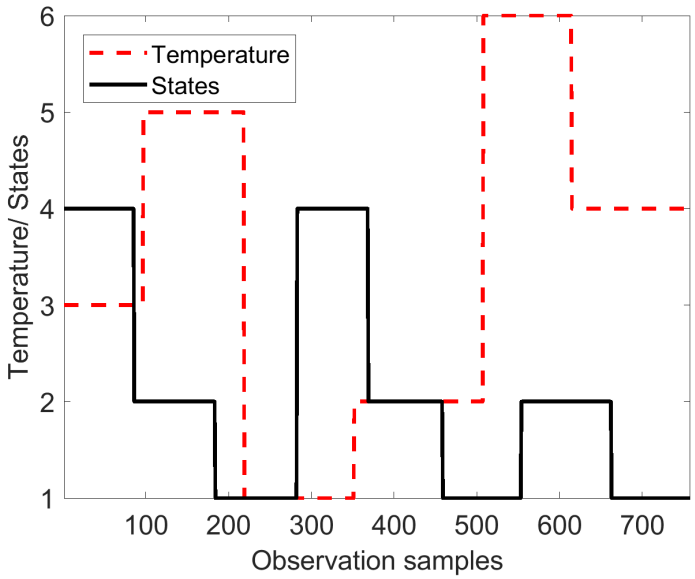


Figure 5.6: Mode identification of the randomized temperature sequence

Compound classes representative of the pseudocomponent spectra are enumerated on the same lines, as done in Section 5.4.2. The peaks at 2950 cm^{-1}, 2920 cm^{-1} and 2850 cm^{-1} are common across all the pseudocomponent spectra and point to the sp^3 C-H stretching, while the peaks at 1380 cm^{-1} and 1450 cm^{-1} indicate the C-H bending vibrations. In addition to these peaks, it is seen that for State 4, PC_1 has peaks at 740 cm^{-1}, 1014 cm^{-1}, 1171 cm^{-1} and 1600 cm^{-1}, indicating that it constitutes naphthenearomatics, condensed alcohols and phenols; PC_2 has peaks corresponding to -o,-m and -p substitutions that run along with peaks at 1050 cm^{-1} , 1260 cm^{-1} , 1600 cm^{-1} and 1750 cm^{-1}, pointing to a class of phenyl esters with

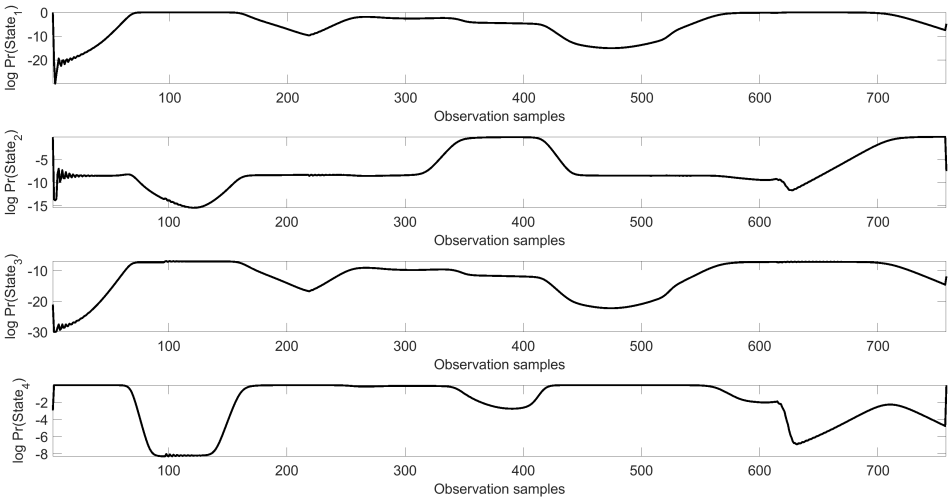


Figure 5.7: Posterior probabilities of the states

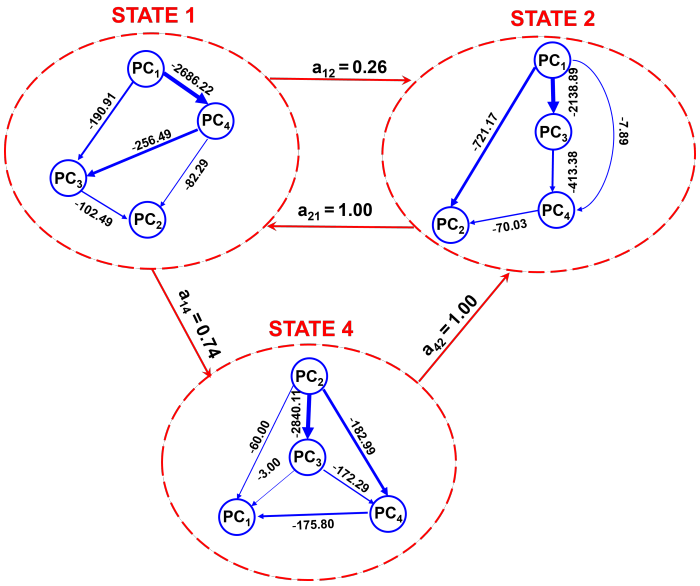


Figure 5.8: Reaction mechanisms of the modes with the state transition probabilities

substituents; PC_3 has substituted aromatic peaks and also peaks at 1040 cm^{-1}, 1217 cm^{-1} and 1740 cm^{-1}, indicating that it comprises phenols and carboxylic acids; and finally PC_4 is seen to comprise substituted aromatics and condensed alcohols owing to its peaks at 1020 cm^{-1} and 1600 cm^{-1}. From the reaction network structure for State 4, as shown in Figure 5.8, it can be hypothesized that phenyl esters hydrolyse to give

phenyl carboxylates ($PC_2 \rightarrow PC_3$). The phenyl esters ($PC_2 \rightarrow PC_1$ and PC_4) and phenyl carboxylates ($PC_3 \rightarrow PC_1$ and PC_4) may either undergo decarboxylation and subsequent cracking or even hydrolyse with further cracking to result in the formation of substituted aromatics, alcohols and phenols. Intermolecular hydrogen transfer can result in a more saturated class of products in going from $PC_4 \rightarrow PC_1$.

Similarly, analyzing the peaks of the pseudocomponents of State 2, it can be seen that PC_1 comprises substituted aromatics, PC_2 comprises naphthenearomatics owing to peaks at 740 cm^{-1} and 1600 cm^{-1}, PC_3 and PC_4 both consist of substituted aromatics and condensed alcohols owing to peaks at 1035 cm^{-1} and 1600 cm^{-1}. From the reaction network structure of state 2, we may infer that cracking of side chains in substituted aromatics followed by ring closure could lead to the formation of naphthenearomatics ($PC_1 \rightarrow PC_2$). Cracking and hydrogen transfer of substituted aromatics with -OH functional groups in the substituents could give rise to condensed alcohols ($PC_1 \rightarrow PC_3$ and PC_4). Condensed aromatics through intermolecular hydrogen transfer could give rise to naphthenearomatics ($PC_4 \rightarrow PC_2$).

In state 1, PC_1 is seen to comprise naphthenearomatics and alcohols due to peaks at 740 cm^{-1}, 1600 cm^{-1} and 1050 cm^{-1}; PC_2 consists of substituted aromatics, phenols (1165cm^{-1}) and alcohols (1038 cm^{-1}); PC_3 in addition to the peaks for alcohols and phenols also consists of naphthenearomatics due to peaks at 740 cm^{-1} and 1600 cm^{-1}; PC_4 is seen to have condensed substituted aromatics owing to the C=C stretch at 1600 cm^{-1} and also alcohols. From the reaction network structure learned among the pseudocomponent spectral features, we may hypothesize that naphthenearomatics undergo ring opening of the saturated ring by cracking, followed by intramolecular hydrogen transfer to produce condensed substituted aromatics ($PC_1 \rightarrow PC_4$, $PC_3 \rightarrow PC_2$). Intermolecular hydrogen transfer to the condensed aromatics could lead to the formation of more saturated products ($PC_4 \rightarrow PC_3$).The cracking could also be confined to the longer alkyl chains attached to the naphthenearomatics to produce lighter aliphatics ($PC_1 \rightarrow PC_3$, $PC_4 \rightarrow PC_2$).

Having identified the reaction mechanisms associated with each of the identified modes along the temperature signal, the state transitions of Figure 5.8 determined from the HSMM model point to the reaction dynamics. The decarboxylation and subsequent cracking of phenyl esters, or even the hydrolysis and subsequent decarboxylation and cracking of phenyl esters produce substituted aromatics, alcohols and phenols. This mechanism then transitions into one where further cracking of the substituted aromatics , accompanied by intermolecular hydrogen transfer leads to the formation of saturated lighter products formed at the expense of condensed aromatics. This reaction state then transitions to one where the saturated lighter aromatics could be subject to ring opening and intramolecular hydrogen transfer to produce aromatics with shorter condensed chains that could further crack to produce lighter aliphatics. The mechanism is also seen to reversibly transition back to ring closure and intermolecular hydrogen transfer to result in naphthenearomatics, or could also go down the path of hydrolysis and decarboxylation of esters, formed by the reversible combination of phenols and alcohols with acids present in the mixture.

5.4.4 Randomized temperature sequence using 6 states

The results obtained by training the HSMM model on the randomized temperature sequence data of Section 5.4.3 with 6 states have been presented in this section. The localized modes identified by the Viterbi state decoding as given in Figure 5.9 are seen to transition among 5 states. States 1,2 and 3 are seen to prevail over shorter duration as compared with the other states, as shown in Figure D.5. The posterior probabilities of the states as shown in Figure D.6 indicate states 5 and 6 to not only have the least probabilities over sampling times but also to be identical. It may be inferred that states 5 and 6 are identical, implying that a HSMM model with at most 5 states would suffice to capture the process dynamics. The pseudocomponent spectra obtained by the factor decomposition of the spectra associated with each of the states are given in Figure 5.10b. The Bayesian network structure learned from the pseudocomponent

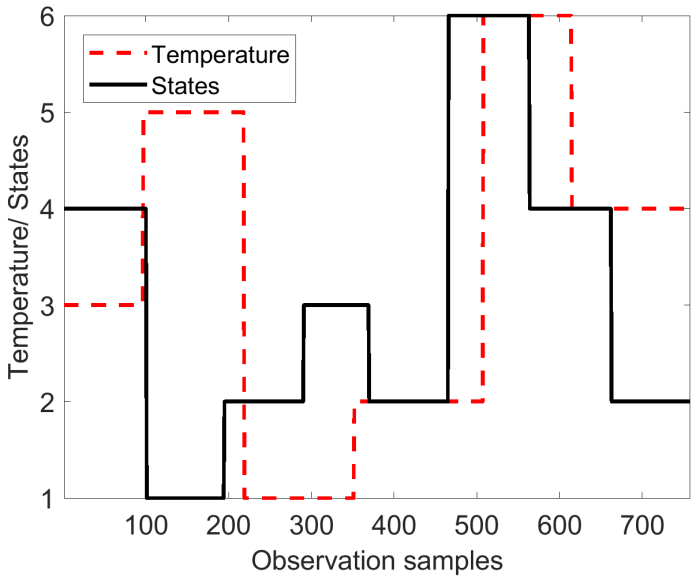
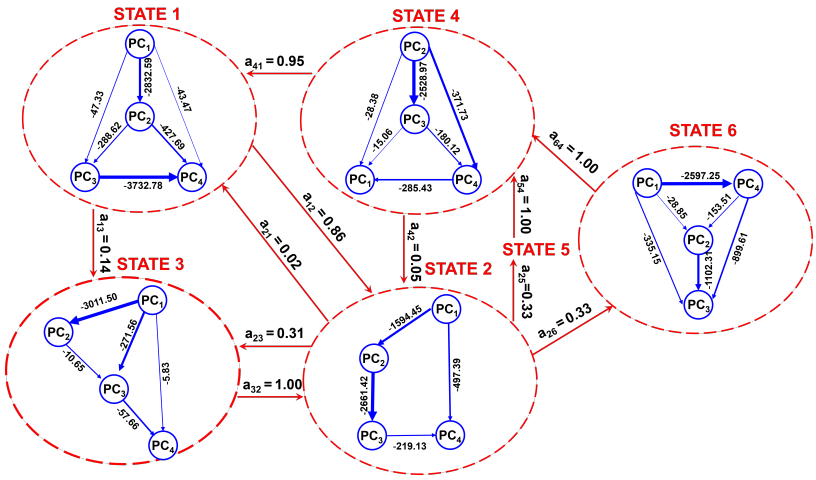


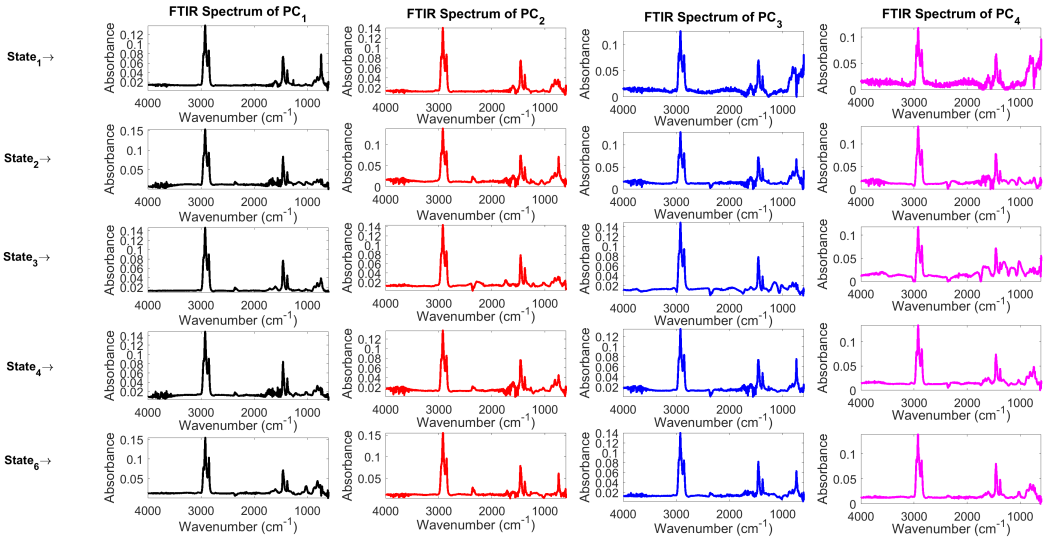
Figure 5.9: Mode identification of the randomized temperature sequence

spectra, used to hypothesize reaction mechanisms of the states visited by the process, is outlined in Figure 5.10a. The peaks in the pseudocomponent spectra have been interpreted on the similar lines as done in the previous sections. The peaks at 2950 cm^{-1}, 2920 cm^{-1} and 2850 cm^{-1} are common across all the pseudocomponent spectra and point to sp^3 C-H stretching, while the peaks at 1380 cm^{-1} and 1450 cm^{-1} indicate C-H bending vibrations.

In addition to these peaks, in state 4, PC_1 has peaks corresponding to substituted aromatics, and smaller peaks at 1030 cm^{-1} and 1217 cm^{-1} pointing to the C-O stretch of alcohols and phenols respectively, with a few noisy peaks around 1700 cm^{-1}, indicating that it mainly constitutes phenyl carboxylates. PC_2 consists of substituted aromatics and condensed phenols (peak at 1250 cm^{-1} for C-O phenolic stretch and 1600 cm^{-1} for C=C stretch), while PC_3 is representative of phenyl esters (peaks at 740 cm^{-1}, 1600 cm^{-1}, 1730 cm^{-1}) and PC_4 comprises substituted aromatics and condensed alcohols (C-O alcohol stretch at 1050cm^{-1} along with a peak at 1600 cm^{-1}). From the structure of the Bayesian networks, it can be seen that the state represents the reaction of condensed alcohols and phenols with carboxylic acids to give phenyl



(a) Reaction mechanisms of the modes with the state transition probabilities



(b) Pseudocomponent spectra associated with the modes

Figure 5.10: Reaction mechanisms associated with the pseudocomponent spectra of each state

esters ($PC_2 \rightarrow PC_3$) and phenyl carboxylates ($PC_2 \rightarrow PC_1$). The condensed phenols and alchols may also undergo thermal cracking of aliphatic alcohol substituents to produce aromatics and alcohols ($PC_2 \rightarrow PC_4$), which may recombine with acids present to form carboxylates ($PC_4 \rightarrow PC_1$). The phenyl esters may also hydrolyse to produce phenyl carboxylates ($PC_3 \rightarrow PC_1$) and further cracking upon hydrolysis may even produce aliphatic alcohols and condensed aromatics ($PC_3 \rightarrow PC_4$).

Analysis of the pseudocomponent spectra for State 1 indicates that PC_1 consti-

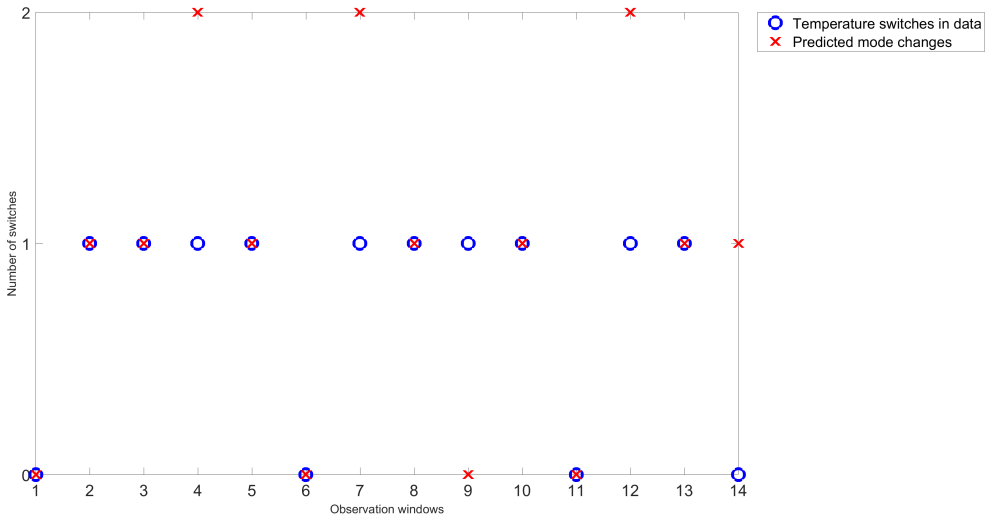
tutes naphthenearomatics and phenols due to peaks at 740 cm^{-1} corresponding to -o substitutions running along with a peak for the C=C stretch at 1600 cm^{-1} , and a C-O phenolic stretching peak at 1260 cm^{-1}. The spectra of PC_2, PC_3 and PC_4 are all characterized by peaks pointing to substituted aromatics and the C=C olefinic stretch at 1600 cm^{-1}. From the reaction network, it can be hypothesized that naphthenearomatics form condensed substituted aromatics by intermolecular hydrogen transfer ($PC_1 \rightarrow PC_2$, PC_3, PC_4), which further undergo cracking and intramolecular hydrogen transfer to shorter chained condensed aromatics ($PC_2 \rightarrow PC_3$, PC_4 and $PC_3 \rightarrow PC_4$).

For state 2, the pseudocomponent spectrum of PC_1 is seen to include peaks for substituted aromatics, weak peaks at 1030 cm^{-1} and 1190 cm^{-1} pointing to the C-O stretch of alcohols and phenols, respectively and a noisy peak in the 1600 cm^{-1} indicating the C=C olefinic stretch. Consequently, PC_1 is seen to represent condensed substituted aromatics with alcohol groups in the aliphatic side chains. Similarly, it can be seen that a sharp peak at 740 cm^{-1} (-o substituted aromatics) along with peaks at 1050 cm^{-1} (C-O alcohol stretch), 1261 cm^{-1} (C-O phenol stretch), 1630 cm^{-1} (C=C stretch) and 1750 cm^{-1} (C=O stretch), indicate -o substituted phenyl esters to be representative of PC_2. PC_3 is considered to be predominantly naphthenearomatic due to peaks at 740 cm^{-1} and 1600 cm^{-1}, while PC_4 represents condensed substituted aromatics, alcohols and phenols. It can be inferred from the reaction network that condensed phenols react with carboxylic acids to yield phenyl esters ($PC_1 \rightarrow PC_2$) that subsequently hydrolyse and crack to form naphthenearomatics by intermolecular hydrogen transfer ($PC_2 \rightarrow PC_3$). The naphthenearomatics may undergo ring opening and intramolecular hydrogen transfer to form condensed substituted aromatics ($PC_3 \rightarrow PC_4$). Cracking of aliphatic side chains with alcohol functional groups followed by hydrogen transfer in the aromatics of PC_1 may also result in condensed alcohols and aromatics ($PC_1 \rightarrow PC_4$).

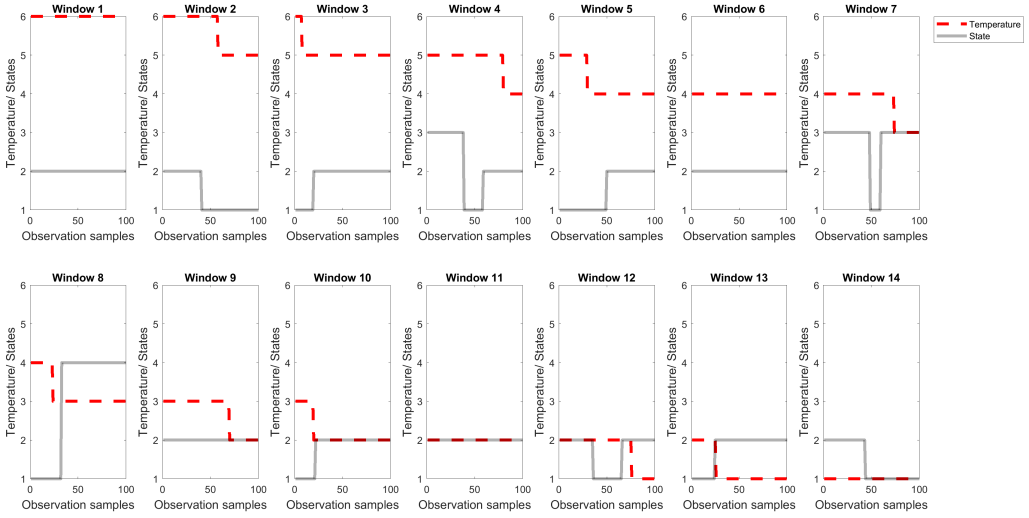
Analysis of the spectra in state 3 indicates PC_1 to comprise peaks pointing to

the presence of substituted aromatics alongside peaks at 1600 cm^{-1} and 1700 cm^{-1} representing the C=C stretch and the C=O stretch, respectively. Consequently, PC_1 could be hypothesized to consist of condensed substituted aromatic anhydrides. PC_2 is seen to have peaks at 1710 cm^{-1} (C=O stretch), 1217 cm^{-1} (C-O phenolic stretch), 1040 cm^{-1} (C-O alcohol stretch) besides the substituted aromatic peaks, implying that the representative compounds could include phenyl carboxylates. The spectra of PC_3 and PC_4 are seen to comprise condensed aromatics and phenols, owing to the peaks corresponding to the C-O stretch of alcohols and phenols along with the C=C olefinic stretch. The mechanism inferred from the reaction networks, points to the hydrolysis of aromatic anhydrides to produce phenyl carboxylates ($PC_1 \rightarrow PC_2$), which on further hydrolysis and cracking produces condensed aromatics and phenols ($PC_2 \rightarrow PC_3$,PC_4). The aromatic anhydrides may also directly undergo hydrolysis and subsequent thermal cracking to form to form condensed aromatics and phenols ($PC_1 \rightarrow PC_3$,PC_4).

Finally, analyzing the spectra of state 6, reveals that PC_1 and PC_4 comprise substituted aromatics and condensed alcohols (owing to peaks ar 1015cm^{-1} and 1600 cm^{-1}). PC_2 has a sharp peak at 740 cm^{-1} along with peaks at 1610 cm^{-1} and 1700 cm^{-1}, thereby consisting of phenyl esters, while PC_3 primarily constitutes naphthenearomatics due to a sharp peak at 740cm^{-1} besides a few noisy peaks in the 1600 cm^{-1} region. It can be inferred from the reaction networks that condensed alcohols combine with acids to form phenyl esters ($PC_1, PC_4 \rightarrow PC_2$), which upon decarboxylation and hydrogen disproportionation result in naphthenearomatics ($PC_2 \rightarrow PC_3$). The condensed aromatics with aliphatic alcohols attached as substitutents may also undergo side chain cracking and intermolecular hydrogen transfer to produce naphthenearomatics ($PC_1, PC_4 \rightarrow PC_3$).



(a) Number of mode changes vs. the number of temperature changes across windows

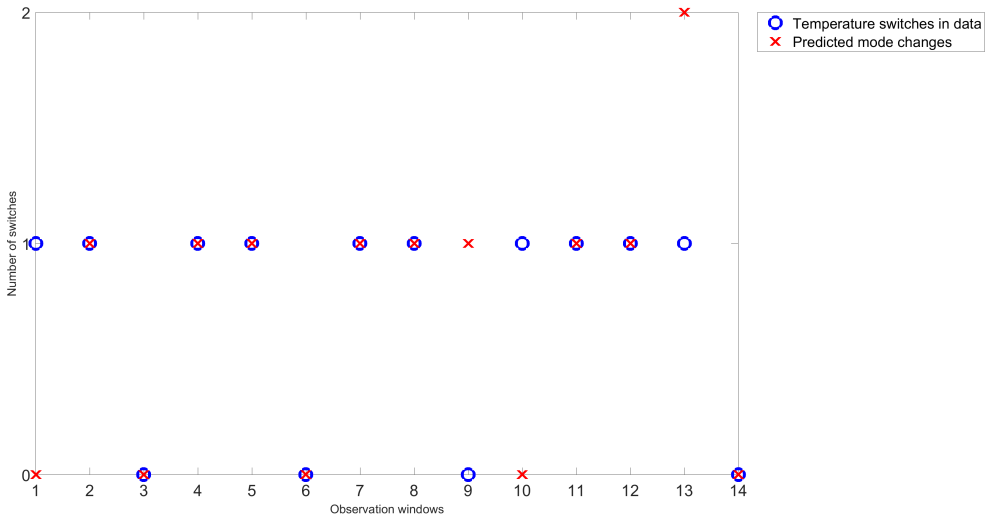


(b) Modes identification in sample windows

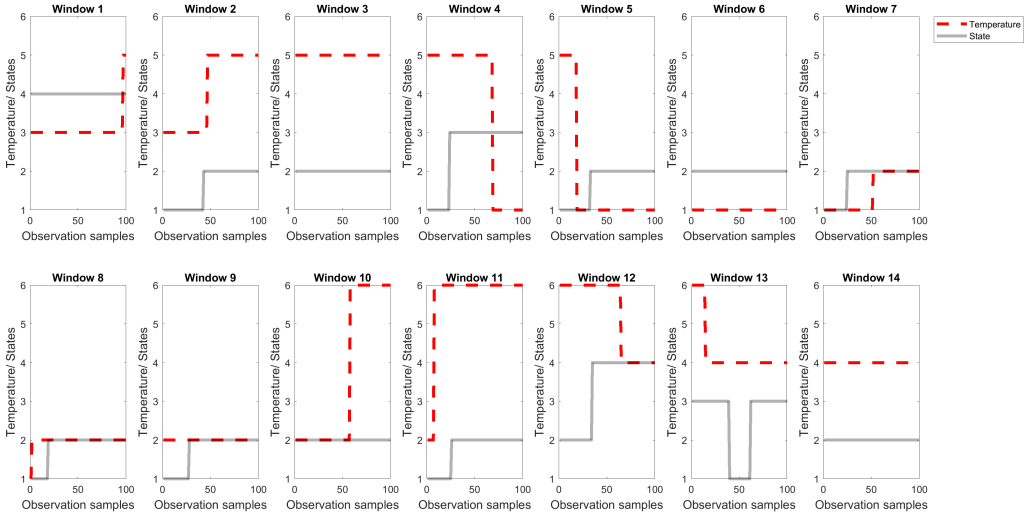
Figure 5.11: Online monitoring of reaction mechanisms by mode identification in a moving window of samples for the decreasing temperature sequence

5.4.5 Online monitoring

Non-stationarity in process data arising from a shift in operating conditions is typically handled by online monitoring using moving windows of process data [301]. Alternatively, modifications to the Viterbi algorithm for real-time state decoding by developing partial path hypotheses on successively expanding windows until conver-



(a) Number of mode changes vs. the number of temperature changes across windows



(b) Modes identification in sample windows

Figure 5.12: Online monitoring of reaction mechanisms by mode identification in a moving window of samples for the randomized temperature sequence trained using 4 states

gence is reached, overcomes the limitation of tracing back from the end of the sequence using backpointers [343]. While using HMMs, state decoding can be achieved based on different optimality criteria [309], the two most popular being posterior decoding, which gives the minimum error path by deciphering the most likely state at a given instant, and the maximum a posteriori probability (MAP) path obtained by Viterbi

state decoding, where a valid sequence of states that generates the observation sequence $O_{1:t}$ is obtained [344]. It must be noted that posterior decoding yields the most likely state at each sample time but does not guarantee a valid sequence of states as the states in consecutive positions may have zero transition probability.

For fault detection, HMMs have been trained offline to learn the model parameters, based on which process monitoring indices are evaluated, followed by using kernel density estimation to determine an index threshold for normal operation [324]. Online fault detection is then realized by calculating these indices based on posterior decoding for each new sample and comparing against the threshold. Recently, modeling the state durations using probability distribution matrices via training a HSMM for multi-modal processes has been used to determine mode affiliation by posterior decoding for improved fault detection [330]. However, monitoring just one sample at a time using posterior decoding for mode identification treats consecutive samples to be independent, fails to capture the process dynamics and is susceptible to noise, because of which the aforementioned fault detection framework has been modified to include a moving window of samples for online Viterbi state decoding [315]. Additionally, modeling the state duration using a probability matrix increases the number of parameters to be estimated and also depends on specifying the maximum amount of time spent in each state [333]. This has led to the use of parametric distributions to model state duration densities [340]. There is evidence of using a HMM model trained offline for online fault detection by determining if the variance of posterior probability over a moving window of samples exceeds a threshold to assess if the underlying process mode is known [322], followed by using either process monitoring indices [324] or the probability ratio strategy to identify [323] faults in stable and transient process operation. It is imperative to use a correct window size when using moving window approaches for online monitoring, as a large window could increase the computational load and a small window may not capture the essential process dynamics [316]. There is not a standard way to ascertain the correct window size.

In order to overcome some of the shortcomings discussed above, we use a HSMM model with the state duration modeled using the Poisson probability distribution, parametrized by the average duration of the state. The size of the moving window is ascertained based on the value of the average state duration of the HSMM model trained offline. It can be seen from Figure D.1, Figure D.3 and Figure D.5 that the average state duration is ~ 100. Consequently, a moving window size of 100 observation samples is chosen for online monitoring. A stride length of half the window size is chosen, as it is believed to be reasonable to assume that half the average state duration captures the time scale of the reaction dynamics in real-time monitoring. The parameters of the HSMM model trained offline are updated in each window of the streaming sample spectral data, and are used as an initial point for each subsequent window. The Viterbi algorithm used for the MAP state decoding takes into account the dependence of the sample window instead of considering each sample to be independent, as is the case with posterior decoding. The model parameters of each window are used to determine the optimal sequence of states in a given window, that physically correspond to a change in reaction mechanisms associated with the change in the operating temperature. The results of online monitoring using a HSMM model with 4 states using moving windows of observation data on the decreasing temperature sequence of data as in Section 5.4.2 is given in Figure 5.11, while Figure 5.12 are the results for the same, but using random temperature sequence data as that in Section 5.4.3. In Figure 5.11a and Figure 5.12a, the number of times a state change is registered in the Viterbi path is indicated as the predicted mode changes, and is compared against the actual temperature changes in each window. The differential mode vector of a sliding window of samples could then be used as an indicator to switch to a suitable diagnostic strategy, to facilitate monitoring multi-modal processes [317].

5.5 Conclusions

A framework for the process monitoring of reaction mechanism dynamics from spectroscopic data has been developed. The process data is deemed to be multi-modal, owing to the varying modes of temperatures and residence times across which the spectra are recorded. Hidden Markov models which capture the dynamic interactions in multi-modal data have been extended to include a distributional assumption on the duration of the process modes, by way of the explicit duration hidden semi-Markov model. The Viterbi state decoding based on the HSMM has enabled optimal mode localization by accounting for the sequential dependence of the FTIR observation data. Building a two-step statistical inferential model on the mode localized data has enabled each of the modes to be interpreted as a reaction mechanism hypotheses associated with the changing process operating temperature. The first step concerns the statistical factor decomposition of spectra constrained by Beer's law to obtain latent spectral features, while the second concerns causally inferring reaction hypotheses by a graph theoretic approach of Bayesian structure learning using the spectral features. The HSMM has been trained to monitor reaction dynamics by accounting for the probabilistic transition of a reaction hypotheses representing a mode to another, and also the duration probability distributions of the modes that physically translate to time scales of the mechanisms.

It has been observed from the duration probability of the states that reaction mechanisms hypothesized at higher temperatures have longer average state durations than those at lower temperatures. The cyclical dynamics of reaction mechanisms accompanying the underlying temperature changes as discovered by the HSMM point to the hydrolysis and cracking of esters to give condensed substituted aromatics, alcohols and phenols; the intermolecular hydrogen transfer and cracking of condensed aromatics to give napthenearomatics and lighter aliphatic products; and finally the recombination of alcohols and phenols with acids to give esters, which subsequently

hydrolyse and crack again. The data-driven process monitoring framework has also been demonstrated to detect temporal patterns of changing mechanism hypotheses from online spectroscopic data across varying temperatures, using Viterbi state decoding on a moving window of sample data based on adaptively updated HSMM parameters. Validating the hypothesized reaction mechanism dynamics by mapping to domain knowledge demonstrates the importance of semantic descriptors used in tandem with data-driven inferential methods, presenting a step toward developing expert systems for monitoring unknown complex reacting systems.

Chapter 6

Chemical reaction neural ODEs and latent factorization to deduce kinetic models from spectroscopic data

Abstract

Kinetic model identification generally relies on accurate measurements of non-equilibrium temporal concentration of the reacting species, which in most cases is difficult to obtain. The lack of prior knowledge of species, poses an added challenge in developing kinetic models. This work demonstrates a framework wherein the latent factorization of realtime spectroscopic data tackles the aforementioned setbacks. The projection of the spectra onto the temporal mode of data collection is shown to gain interpretability as the time varying concentrations, whose direct measurements would otherwise be challenging. Also the latent spectral features corresponding to these temporal concentrations, characterizes the species in obscure chemical systems. Furthermore, the adjacency matrix deduced from the structure of reaction pathways hypothesized by causal structure inference among the latent spectral features is used to constrain the development of kinetic models using the temporal concentrations as inputs to a chemical reaction neural ODE. The incorporation of the law of mass action, the Arrhenius law of temperature dependence in addition to the structural constraints of the reaction network, are seen to enable the neural network recover kinetic models, even

in the presence of considerable noise that challenges the accuracy of spectroscopic deconvolution. The framework has been illustrated using synthetic spectroscopic data from a known reaction template in the database, due to the absence of a precise ground truth model for validation in complex systems like bitumen.

6.1 Introduction

Process intensification by rationalizing the design and optimization of processes involving the conversion of complex reactive feedstocks, depends on modeling the underlying kinetic framework [2]. Developing kinetic models requires mechanistic knowledge of the reactive species and the pathways detailing their conversion, following which the kinetic parameters are estimated from experimental data [345]. However, it is daunting to develop a kinetic framework for complex systems like bitumen/biomass that lack an exhaustive enumeration of the underlying species, let alone the reaction mechanisms underlying their conversion. This has prompted the integration of reactors with spectroscopic sensors that provide molecular-level information of the reactive mixtures [11],[12], which is then used as a basis for developing data-driven models for species identification and the generation of plausible reaction hypotheses [273]. Upon species identification, reaction pathways can be deduced by perceiving chemistry as a series of graph transformations in the space of all possible reactions [346], wherein a molecular fingerprint at the reactant node results in candidate fingerprints at the product nodes, a distribution across which is learned via neural networks to rank the candidates[45]. Statistical models like multivariate curve resolution have been extended to jointly resolve data from multiple spectroscopic sensors in compliance with Beer's law, so that the latent factor projections onto the spectral channels and the temporal mode of data collection are physically interpreted as the pseudo-spectra of the reactive species and their corresponding concentrations, respectively [250]. Domain knowledge is used to identify species from their pseudo-component spectra, while reaction pathways among them are devised

by Bayesian structure learning among the pseudo-component spectra. The present work seeks to develop a chemical reaction neural ODE constrained by the adjacency matrix derived from the Bayesian network structure, pertaining to the hypothesized reaction pathways, aside from incorporating the physical laws of mass action and the Arrhenius law of temperature dependence [347] to fit a kinetic model to the temporal concentration data from the latent projections. The proposed framework paves way for a system-agnostic approach of identifying species, hypothesizing reaction pathways among them and subsequently estimating kinetic models constrained by the reaction network adjacency matrix, purely from the spectroscopic data of the reactive system.

The general form of a kinetic model for the time evolution of n species is given by a function parametrized by the kinetic parameters (θ), for a vector $C(t) = [C_1(t), C_2(t), \cdots C_n(t)]^{\mathrm{T}}$

$$\frac{dC}{dt} = f_\theta(t, T, C_1(t), C_2(t), \cdots, C_n(t)) \tag{6.1}$$

The kinetic model function f described by ODEs, Markov processes and state space representations using the law of mass action kinetics, S-system or polynomial models [51] is characterized by a structure that is derived from the reaction pathways among the species and a set of parameters θ (rate constants, stoichiometric coefficients, orders). Estimating the parameters by fitting the model to experimental concentration data [348] is known as the inverse problem in chemical kinetics and could lead to multiple solutions resulting from the same reaction dynamics [349]. Attempts to use sparsity constraints are found not to be reliable in recovering unique solutions, thereby pushing for the incorporation of additional knowledge about the system [350]. Yet, in the absence of prior knowledge of the network topology, the structure is learned by virtue of kinetic parameter estimation [53] resulting in larger degrees of freedom that challenge a unique solution owing to the *fundamental dogma of chemical kinetics* [349]. Additionally, when it comes to the inverse problem, obtaining measurements of non-equilibrium temporal concentrations of the species is challenging [81]. Therefore, the present work seeks to use knowledge of the reaction mechanisms causally inferred

from spectral resolution to constrain the inverse kinetic model development using the temporal concentrations from the latent factorization. Also, the knowledge of physical laws such as mass action and Arrhenius temperature dependence encapsulated in a system of coupled ODEs indicated by f in Equation 6.1 are used to structurally constrain neural networks that are trained as function approximators of the true kinetic model. This is believed to be superior to cases where the reaction dynamics are modeled as a linear combination of weighted polynomial basis functions representing individual reactions [58], and its sparse variant with a curated library of vector-valued ansatz functions called 'reactive SINDy' [351], where the parameters are estimated by regressing against the temporal concentration data but lack interpretability in the context of the true kinetic model, as physical laws are not explicitly accounted for, and are limited in their function approximation ability as compared to neural networks [352]. We shall now proceed to review some works where neural networks have been used to model chemical kinetics.

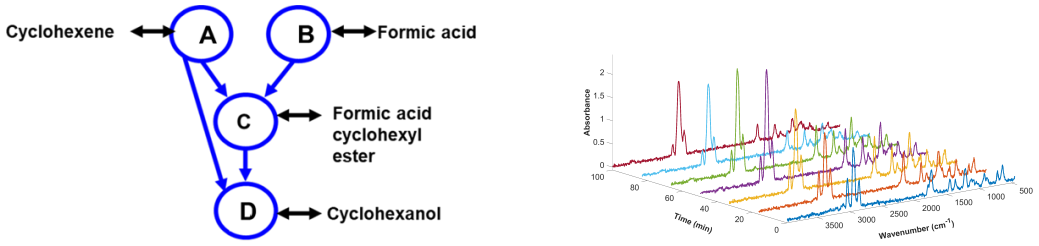
Solving kinetic models in multi-dimensional vector fields of reactive flow problems from direct numerical solution of stiff ODEs, owing to varied reaction time-scales, is seen to be computationally expensive and scales with the number of species [353]. Instead of using simplifying assumptions like quasi-steady state, neural networks have been used for thermokinetic modeling [354] by learning a functional mapping between the true kinetic model (encompassing all mechanisms and transport limitations) and the time evolution of species concentration [355]. Although these neural networks maps are computationally efficient in evaluating kinetic models, they come with a training overhead that requires data obtained either by solving first principle ODEs if the system is known, or from experimental data in the absence of prior knowledge of the system. The training data overhead can be reduced by using hybrid neural networks that are structured with prior knowledge of physical laws, besides improving the generalizability of the function approximation [356]. A Physics-informed neural networks used to model chemical kinetics [357], [358] by mapping a discrete space of

time points to species concentrations, encodes physical laws in its training by minimizing the residual loss between the species conversion rates obtained by automatic differentiation of the predicted concentrations, and the underlying physical ODEs. There is evidence of using experimental data from gas chromatography and heat flux calorimetry to train neural networks to fit kinetic models for complex reactions like esterification and heterogeneous liquid-liquid mononitration [359], and also from reaction colorimeter data for a heterogeneous oxidation process [360]. These outputs of neural network models that learn a mapping between the input species concentrations and the rates of the chemical state space modeled by the ODEs, when time integrated, are seen to diverge from the true species concentration profiles, thereby shifting focus to neural ODEs that integrate the outputs while training, leading to parameter gradients being backpropagated across the ODE solver, while minimizing the difference between the neural network predictions and the true ODE solution [361]. Neural ODEs have also shown promise in learning model dynamics from temporal data obtained from stiff ODEs that are prevalent in kinetic models of chemical and biological systems [362], and differ from physics-informed neural networks in that they can model irregular and incomplete sampled time series data.

Neural ODEs where physical laws are enforced as structural constraints have been used to autonomously infer reaction pathways from time series concentration data, by virtue of kinetic parameter estimation, but rely on grid search to optimize the number of reactions as hyperparameters [347]. There is evidence of using spectroscopic data to propose kinetic models by way of the Deep kinetic spectroscopy network (DeepSKAN) that uses convolution neural networks to obtain time resolved features from the spectra in the affine space of the data collection axes, namely, probe delay and wavenumbers [363]. The latent space of probe delay reveals velocity constants of the mechanisms underlying the photoinduced electronic excitation process, and is used to develop kinetic models but lacks prior knowledge of the potential reaction pathways. The present work seeks to bridge the gap in developing kinetic models

from spectroscopic data, where we develop chemical reaction neural ODEs to incorporate prior knowledge of reactions and network adjacency constraints deduced from the reaction pathways that are hypothesized by Bayesian structure learning among the pseudocomponent spectra. Multivariate curve resolution algorithms that project FTIR spectra onto the data collection axes using a fixed number of components, ascertained by the chemical 'rank', resulting in the pseudo-component spectra from the wavenumber axis, and their corresponding concentrations from the temporal axis, are used at the backend with Bayesian structure learning [273],[250] to provide the temporal concentration data, and reaction pathway constraints, respectively, for the training of chemical reaction neural ODEs that have been implemented using the *torchdiffeq* library on PyTorch [364], [365].

6.2 Description of datasets



(a) Reaction template from database

(b) Synthetic FTIR spectra

Figure 6.1: Synthetic data generation from the reaction network template

We seek to demonstrate our framework of deducing kinetics from spectroscopic deconvolution and causal inference, by choosing a model system from a database where the pure component spectra and the pathways among them are known *a priori*. Knowledge of the ground truth enables us to verify the predictions from our framework, which would otherwise be a non-trivial task for complex systems like bitumen/biomass where the ground truth concerning species enumeration, their reactions pathways and kinetics have not yet been ascertained exhaustively. Hence, in this work, synthetic spectroscopic data is generated from the pure component FTIR

profiles of species following a reaction template that has been obtained from the NIST database [366]. For a given system with N_S species and N_R reaction pathways from the database, the kinetic model constrained by the reaction network adjacency and following the law of mass action can be described by the following system of ODEs, where $n \in \{1, 2, \cdots N_S\}$ and $m \in \{1, 2, \cdots N_R\}$

$$\frac{dC_n}{dt} = \sum_{m=1}^{N_R} \mathbb{1}(Adj_{mn} = 1) K_m \prod_{n=1}^{N_S} C_n^{\mathrm{O}_n} - \sum_{m=1}^{N_R} \mathbb{1}(Adj_{mn} = -1) K_m \prod_{n=1}^{N_S} C_n^{\mathrm{O}_n} \tag{6.2}$$

The ODEs in Equation 6.2 are parametrized by the kinetic parameters *viz.* the order of the n^{th} species given by O_n and the rate constant of the m^{th} reaction pathway that are modeled to account for their temperature dependence in accordance with the Arrhenius law, as given below

$$K_m = K_{m0} e^{\frac{-E_a}{RT}} \tag{6.3}$$

The ODEs are also constrained by the adjacency matrix ($Adj \in \Re^{N_R \times N_S}$) using an indicator function ($\mathbb{1}$) as given in Equation 6.2. The adjacency matrix derives its structure from the reaction pathway network, where each row corresponds to a certain m^{th} reaction, and comprises entries -1 or 1 for each of the N_S species, indicating its participation in the said reaction, either as a reactant or product, respectively. A zero entry is used for species that are non-participating in the reaction.

The reaction template that has been chosen for this study is shown in Figure 6.1a, and is seen to have $N_S = 4$ species that are undergoing $N_R = 2$ reactions as follows

$$\begin{aligned} A + B &\xrightarrow{K_1} C \\ A + C &\xrightarrow{K_2} D \end{aligned} \tag{6.4}$$

For the above reaction template, the ODEs in Equation 6.2 are solved over a time interval $t \in [0, 100\text{min}]$ using a random choice of kinetic parameters and a multi-level pseudo-random temperature signal in the interval $T \in [200^\circ C, 400^\circ C]$ to perturb the

system dynamics via the rate constants, as modeled in Equation 6.3. The pure component spectra of the species are then weighted by the concentration profiles from the ODE solutions, followed by the addition of white Gaussian noise that mimics the effects of random processes while generating synthetic spectra over the time interval t [367]. Illustrative samples of the synthetic spectra at a select few points in the time interval are shown in Figure 6.1b. The framework and results of training a chemical neural ODE on the deconvolved concentration profiles from this synthetic spectral dataset, constrained by the causal structure inferred from among the pseudocomponent spectra, are described in the subsequent sections.

6.3 Methods

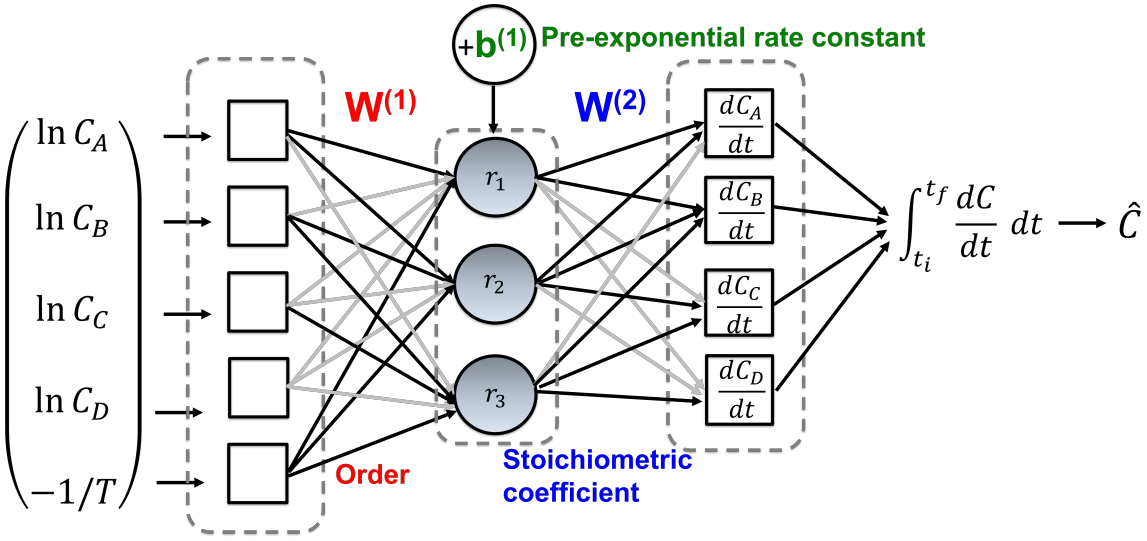


Figure 6.2: Schematic representation of the chemical reaction neural ODE

The synthetic spectroscopic data generated as outlined in Section 6.2 comprises absorbances recorded across time and the spectral channels (wavenumbers). Multivariate curve resolution is used to obtain latent projections of the absorbances across the time and wavenumber axes, as described in our earlier works [273]. The number of components in the latent space is determined using the mathematical notion of 'rank' that indicates the number of latent components that sufficiently capture the variance

of the data in the original space. Since, the latent factorization constrains the projections to be non-negative, the latent components can be interpreted as a chemical species, and their projections onto the axes of time and spectral channels gain interpretability as the concentrations and pseudo-component spectra, respectively. The pseudo-component spectra are then represented as random variables at the nodes, and are modeled using probability distributions to learn a directed acyclic graphical structure among the nodes via heuristic score-search methods in order to maximize the Bayesian Information Criteria (BIC) [250]. The adjacency matrix deduced from the structure of the Bayesian networks inferred from the pseudo-component spectra is used to constrain the development of kinetic models using their corresponding concentration profiles by training chemical neural ODEs, the underpinnings of which are described in this section.

Let us consider the following reaction involving 4 species

$$\nu_A A + \nu_B B \xrightarrow{k} \nu_C C + \nu_D D \tag{6.5}$$

The rate R, of this reaction can be represented in terms of the time rate of change of concentrations of the species $(\dot{C}_A, \dot{C}_B, \dot{C}_C, \dot{C}_D)$and their respective stoichiometric coefficients $(\nu_A, \nu_B, \nu_C, \nu_D)$ that indicate the number of moles of each of the species that participates in the reaction, as indicated from the balanced chemical equation of the reaction [368].

$$r = \frac{-1}{\nu_A}\frac{dC_A}{dt} = \frac{-1}{\nu_B}\frac{dC_B}{dt} = \frac{1}{\nu_C}\frac{dC_C}{dt} = \frac{1}{\nu_D}\frac{dC_D}{dt} \tag{6.6}$$

The kinetic rate expressions based on the law of mass action [369] is as follows

$$r = kC_A^a C_B^b \tag{6.7}$$

In Equation 6.7, k is the rate constant, while a,b are the reactant orders that indicate the degree to which the rate depends on the concentration of a specific reactant. The *orders* are neither related nor identical to the stoichiometric coefficients, with

the exception of elementary reactions. Since it is difficult to determine beforehand, whether or not a reaction is elementary, we would like to proceed by assuming that the orders and stoichiometric coefficients are not the same. Incorporating the temperature dependence of the rate constant as outlined in Equation 6.3, the rate expression in Equation 6.7 can be expressed as an exponential of the linear combination of the logarithm of the species concentrations, weighted by their orders, and that of the negative reciprocal of the temperature, weighted by the ratio of the activation energy and the universal gas constant (E_a/R) to which the logarithm of the pre-exponential rate constant (K_0) is added as a bias term.

$$r = \exp\left[\ln k_0 - \frac{E_a}{RT} + a \ln C_A + b \ln C_B\right] \tag{6.8}$$

Representing the rate in this manner enables the weights and biases to be interpreted as kinetic parameters, and makes the choice of the non-linear activation domain-informed, when neural networks are used as function approximators to learn the dynamics by mapping time series concentrations to reaction rates. Inspired from neurobiology, neural networks combine multiple inputs as their linear weighted sum translated by a bias term, the result of which is non-linearly transformed by the choice of an activation function to result in hidden features that are similarly combined to result in outputs that are trained to approximate any function to arbitrary precision [370]. Neural networks where the computed hidden features are re-used by similar weighted combination and non-linear activation to produce a hierarchy of hidden features over subsequent layers are said to be *deep*, whereas those with just one layer of hidden features are considered *shallow*. The number of hidden features in each layer, referred to as the neurons, and the number of layers themselves comprise the hyperparameters (network topology) and guide the precision of the neural network as a universal function approximator, parametrized by the weights and biases that are learned by gradient descent optimization (backpropagation *i.e.* the gradients of the loss function computed at the output with respect to the parameters are propagated

backwards through the successive layers) [371]. Deep neural networks comprise more hyperparameters than shallow neural networks, and thereby suffer from overfitting due to the model complexity that is sought to be handled by effective regularization of the parameters [372] and the reconciliation of domain knowledge into the network structure [351], [373]. These approaches to limit the overfitting and improve the generalizability of the neural networks also promote model interpretability and reduce the requirement of large amounts of training data.

In this work, we demonstrate the use of a shallow neural ODE, a schematic of which has been indicated in Figure 6.2. The neural network is seen to comprise i) an input layer consisting of the logarithm of the temporal concentration of species obtained from multivariate curve resolution of the synthetic spectra, and the negative of the reciprocal of time varying temperature. Let us denote the input data at time t by a vector $X_t = [\ln C_1(t), \ln C_2(t), \cdots \ln C_{N_S}(t), -1/T(t)]^T$, such that $X_t \in \Re^{(N_S+1)\times 1}$ is the temporal vector fed into the network. ii) a single hidden layer consisting of as many neurons as the number of reaction pathways. The features in the hidden layer are denoted by a vector $H_t \in \Re^{N_R\times 1}$ that consists of the reaction rates $H_t = [r_1(t), r_2(t), \cdots r_{N_R}(t)]^T$ iii) an output layer with as many nodes as the number of species, where each node corresponds to the predicted time rate of change of the species concentration, given by a vector $\hat{\dot{C}}_t \in \Re^{N_S\times 1}$ and iv) an ODESolve function to integrate the time rate of the species' concentration over an interval to result in predictions of their corresponding concentration profiles, in vector $\hat{C}_t \in \Re^{N_S\times 1}$ given by $\hat{C}_t = \left[\hat{C}_1(t), \hat{C}_2(t), \cdots \hat{C}_{N_S}(t)\right]^T$. The parameters of the network denoted by θ comprise the weights of the first two layers, denoted by $W^{(1)} \in \Re^{N_R\times(N_S+1)}$ and $W^{(2)} \in \Re^{N_S\times N_R}$, and the bias associated with the first layer, denoted by $b^{(1)} \in \Re^{N_R\times 1}$. The weights of the two layers are interpreted as the order and stoichiometric coefficients, respectively, while the bias points to the pre-exponential rate constants, as can be seen from Equation 6.7 and Equation 6.8. The weights of the network are regularized by the adjacency matrix $Adj \in \Re^{N_R\times N_S}$ as illustrated in the following

set of equations in the forward pass of the neural ODE, where $\mathbb{1}$ is the indicator function, while $\mathbf{1}$ is a notation for a vector of ones appended to the adjacency matrix to account for the temperature term in the input, aside from the logarithm of the species concentrations.

$$H_t = \exp\left[\left(W^{(1)} * \left[\mathbb{1}(Adj = -1)|\mathbf{1}^{N_R \times 1}\right]\right) X_t + b^{(1)}\right] \tag{6.9}$$

$$\hat{\dot{C}}_t = \left(W^{(2)} * \mathbb{1}(Adj \neq 0)^T\right) H_t \tag{6.10}$$

$$\begin{aligned} \hat{C}_t &= \hat{C}_{t-1} + \int_{t-1}^{t} \hat{\dot{C}}_{t-1}\, dt \\ &= \text{ODESolve}(\hat{C}_{t-1}, \hat{\dot{C}}_{t-1}, [t-1, t], \theta) \end{aligned} \tag{6.11}$$

The network is trained to not only reconcile the predicted concentration profiles with that obtained from the deconvolution of synthetic spectra, but also to minimize the difference between the predicted time rate of change of the species concentration and the numerically computed values from finite differences of the temporal concentrations from the spectral curve resolution across all time points, as indicated by the loss function given below. Additionally, sparsity among the weights is enforced via the adjacency matrix deduced from the Bayesian network structure, penalized by the regularization weight α. All of the weights *not* used in the forward pass computations, given in Equations 6.9-6.11 as constrained by the adjacency matrix, are forced towards sparsity.

$$\begin{aligned} L(\theta) = \sum_t \left(C_t - \hat{C}_t\right)^2 + \sum_t \left(\dot{C}_t - \hat{\dot{C}}_t\right)^2 + \alpha \left(W^{(1)} * \left[\mathbb{1}(Adj \neq -1)|\mathbf{0}^{N_R \times 1}\right] + \right. \\ \left. W^{(2)} * \mathbb{1}(Adj = 0)^T\right) \end{aligned} \tag{6.12}$$

Minimizing the loss function in Equation 6.12 involves solving the ODE in the forward pass, and the continuous backpropagation of the gradient requires solving the augmented ODE backwards in time [364] as it involves computing the derivatives of the ODE solution with respect to the network parameters. This has been implemented

using the adjoint sensitivity analysis by framing a set of auxiliary ODEs, the solution of which evaluates to provide the aforementioned derivatives, while training the neural ODE. The PyTorch library, *torchdiffeq* [364], [365] encapsulates code for the same, and has been used to train the neural ODE presented in this work.

6.4 Results and Discussion

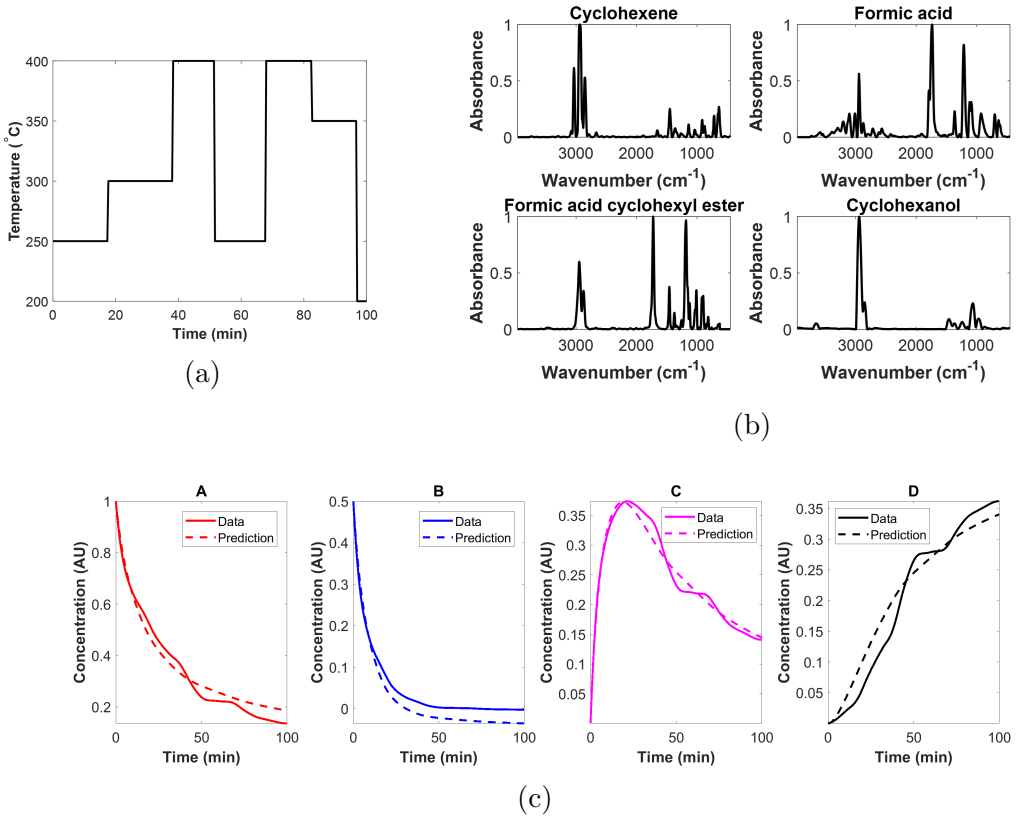


Figure 6.3: (a) Multi-level pseudo random temperature signal, (b) Pure component spectra from the database, (c) Predictions of the chemical reaction neural ODE compared against the temporal concentration data obtained by solving a known ODE system for kinetics.

A known reaction template from literature, for cyclohexanol production via the esterification of cyclohexene with formic acid and the subsequent hydration of formic acid cyclohexyl ester to form cyclohexanol, is considered [374]. The reaction template is outlined in Figure 6.1a. As an initial test to validate the capabilities of the proposed

chemical reaction neural ODE to accurately capture kinetics, temporal concentration profiles are obtained by solving the system of ODEs as outlined in Section 6.2. Baseline neural ODE predictions are tested on the data from solving the ODEs to ensure that kinetic models can be reliably deciphered in the event spectral deconvolution perfectly recovers the underlying pure component spectral profiles (Figure 6.3b) and their corresponding temporal concentrations, in an ideal case not limited by process and measurement noise. A multi-level pseudorandom temperature signal signal as shown in Figure 6.3a was used to perturb the the kinetic model of Equation 6.2. The kinetic mechanism is seen to comprise 4 species undergoing 2 reactions, as indicated by Equation 6.4. Random initializations were used for the concentrations of the reacting species (A and B), to obtain concentration profiles using Equation 6.2 which are supplied to the chemical neural ODE constrained by the following adjacency matrix deduced from the template structure:

$$Adj = \begin{bmatrix} -1 & -1 & 1 & 0 \\ -1 & 0 & -1 & 1 \end{bmatrix}$$

The concentration predictions from the neural ODE are compared against the profiles of the temporal concentrations recovered from solving the ODEs, as shown in Figure 6.3c. It can clearly be seen that constraints on the neural networks structure and parameters, prevents it from overfitting the data. Hence, in the future when the model is trained on synthetically generated noisy data, it is expected to run a low risk of fitting the noise.

On the above lines, we proceed to test the model performance in the presence of noise. Two cases, one with Gaussian white noise at a signal to noise ratio (SNR) of 35, and another at a SNR of 100 have been used for synthetic data generation. The impact of the noise threshold in data on the spectral curve resolution, the subsequent identification of species and inference of reaction pathways among the pseudo-component spectra, and thereafter the pathway constrained kinetic model identification using temporal projections of the resolved spectra via chemical neural ODEs, is investi-

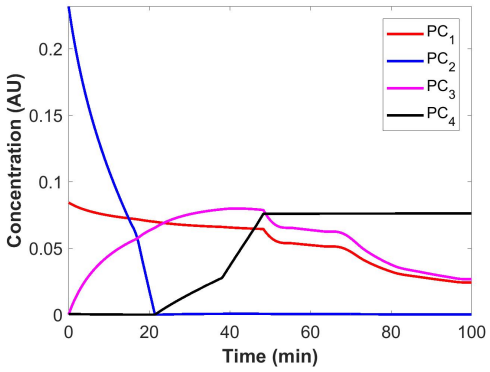
gated.

In the first case, white Gaussian noise at a signal to noise ratio of 35 is added to the synthetically generated data as described in Section 6.2, before it is supplied to spectral curve resolution. The curve resolution with a rank of 4 is seen not to perfectly recover the pure component spectra, as shown in the noise contaminated deconvolution results of Figure E.1 in Appendix E. The similarity of the recovered pseudo-component spectra (Figure E.1c) with the pure component spectra (Figure 6.3b) helps in identifying the species from the database template that the pseudo-components map to. It can be seen that arriving at perfectly resolved pseudo-component spectra is challenging in the presence of noise. Confounding patterns are observed in the resolved peaks of pseudo-component 4 (PC_4) and pseudo-component 2 (PC_2) that correspond to compounds B and C from the database (Figure 6.1a), respectively. Consequently, the causally inferred reaction network among the pseudo-components spectra (Figure E.1b) when compared with the reaction template structure (Figure 6.1a), points to the presence of an additional conversion pathway (A $\rightarrow$ B). This could largely be attributed to the fact that a directed edge, PC_3 (compound A) $\rightarrow$ PC_2 (compound C) in Figure E.1b) with the highest arc strength points to the conversion of compound A to compound C. In the event that peaks in PC_2, corresponding to compound C are confounded with PC_4, corresponding to compound B, there exists a fair chance of observing an additional directed arc from PC_3 (compound A) to PC_4 (compound B). The structure of the adjacency matrix in this case assumes the following form:

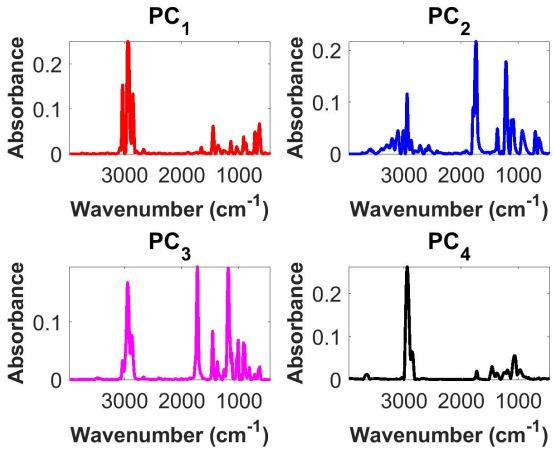
$$Adj = \begin{bmatrix} -1 & -1 & 1 & 0 \\ -1 & 0 & -1 & 1 \\ -1 & 1 & 0 & 0 \end{bmatrix}$$

The predictions of the chemical neural ODE with the above adjacency constraints are compared against the reconstructed data from integration of the smoothed time derivative of the noisy concentration profiles from spectral deconvolution as given in Figure E.2. The neural predictions are seen to capture trends in the noisy concentration profiles, without fitting the noise, except for the profiles of PC_2. Thereby,

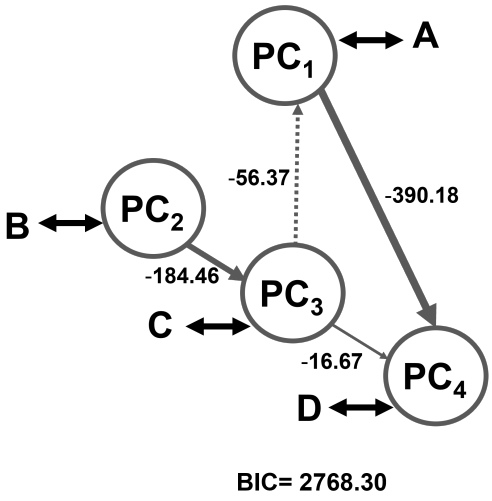
despite improper spectral deconvolution, it has been demonstrated that fairly reliable kinetic models for most of the identified species can be recovered, starting from noisy spectroscopic data.



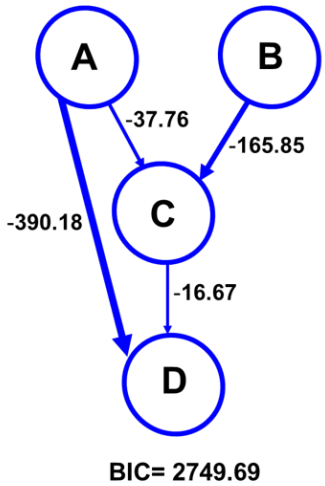
(a) Temporal concentrations



(b) Pseudo-component spectra



(c) Reaction network structure, arc strengths and score inferred from the pseudo-component spectra



(d) Reaction network arc strengths and score inferred from the reaction template

Figure 6.4: Spectral deconvolution and causal inference using noisy synthetic data at a signal to noise ratio of 100.

In the second case, white Gaussian noise at a signal to noise ratio of 100 is added during the synthetic data generation process. At relatively lower noise levels, the spectral curve resolution is seen to result in cleaner temporal concentration profiles

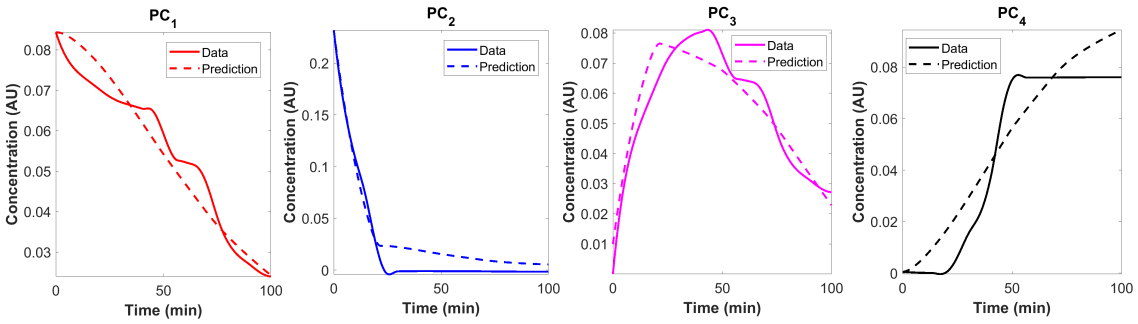


Figure 6.5: Comparison of the predictions from the chemical neural ODE against the reconstructed data from integration of the smoothed time derivative of temporal concentration obtained by the deconvolution of synthetic spectroscopic data, at a signal to noise ratio of 100.

(Figure 6.4a) and pseudo-component spectra (Figure 6.4b) where there are fewer confounding peaks in the deconvolved spectral profiles that are found to be increasingly comparable with the pure component spectra from the database (Figure 6.3b). The pseudo-components are mapped to the pure components based on the similarity between their spectra, followed by inferring reaction pathways among them by causal structure learning as shown in Figure 6.4c. The skeleton of the inferred network structure is exactly the same as the reaction template (Figure 6.1a), except for the reversal of the arc between the nodes of compound A and compound C. This is largely owing to the fact that greedy heuristic score search algorithms for causal structure inference by maximizing the Bayesian Information Criteria (BIC) are faced with a large number of locally optimal network structures [375]. This has been verified by computing the arc strengths and the BIC score shown in Figure 6.4d, given the directed edges among the compounds nodes from the reaction template network structure of Figure 6.1a. The arc strengths and the BIC score, given the network structure from the template are found to be comparable to those when the network structure is inferred by heuristic score-search algorithms as shown in Figure 6.4c. Hence, the reversal of the arc between nodes A and C, in comparison with the original template can be rationalized as occurring due to multiple local optima in the search space of feasible network structures during causal inference.

The issue of local optima in structure learning can be circumvented by preferentially weighting and even eliminating certain wavenumber absorption bands in the deconvolved pseudo-component spectra, as shown in Figure E.3. Four absorption band regions, *viz.* 786- 1310 cm^{-1}, 1570-1898 cm^{-1}, 2686- 3122 cm^{-1} and 3530-3806 cm^{-1} that are predominantly seen to exhibit convoluted peaks, as seen in Figure 6.4b, are chosen. The absorbances in these wavenumber bands are then preferentially weighted using a Gaussian filter that is centered in each of the bands, with a standard deviation of 200, with weights for the bands in the regions 1570-1898 cm^{-1} and 3530-3806 cm^{-1} being scaled by a factor of 10 times as compared to the two other bands, in order to obtain a clear distinction between the spectral profiles of formic acid and its derivatives (compounds B and C), and that of cyclohexene and its derivatives (compounds A and D), as seen in Figure E.3a. It can be seen that the arc strengths, score and network structure learned from the preferentially weighted pseudo-component spectra, as shown in Figure E.3b concur with those, given the reaction template structure, as outlined in Figure E.3c. Hence, it can be seen that the use of prior knowledge to preferentially weight certain absorption bands in the pseudo-component spectra, facilitates distinction of the identified species, to overcome the limitation of confounded peaks in the deconvolution. However, since the discussion in this chapter focuses on limiting the use of prior knowledge-based heuristics in the end-to-end modeling framework, proceeding further on these lines is out of the scope of the current work.

Therefore, the adjacency matrix, following from the causally inferred network structure (Figure 6.4c), in the absence of any prior knowledge-based preferential weighting heuristics of the pseudo-component spectra, is used to constrain the kinetic model identification as follows:

$$Adj = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 0 & -1 & 1 \end{bmatrix}$$

The predictions from the chemical neural ODE used to fit a kinetic model is compared

against the reconstructed data from integration of the smoothed time derivative of the temporal concentration projections from spectral resolution (Figure 6.4a) as shown in Figure 6.5. It can be seen that the neural kinetic model predictions very closely capture the trends in the resolved concentration profiles for all of the identified species, at a much lower noise threshold (as compared to the case where a SNR of 35 was used), despite being constrained by a network structure that slightly differs from the original reaction template.

6.5 Conclusions

We have presented a chemically constrained neural ODE to fit kinetic models to temporal concentration data. Latent factorization of spectroscopic data that results in projections onto the temporal mode of data collection and the spectral channels, gain interpretability as time varying concentrations and the associated pseudo-component spectra of the underlying species, respectively. This overcomes the difficulty in directly measuring species concentrations, more so in cases when the underlying species lack enumeration. The adjacency matrix deduced from the Bayesian networks learned by causal structure inference among the pseudo-component spectra is used to constrain the weights of the neural ODE that is also structured to incorporate the law of mass action and the Arrhenius law of temperature dependence, to achieve a two-fold purpose: (i) facilitate interpretability of the neural ODE model that learns the system kinetics, (ii) limit the tendency of the neural ODE to fit process noise that is ubiquitous when it comes to spectroscopic measurements. However, the accuracy of the causally inferred Bayesian network structure is seen to be limited at the level of uncertainty not only by way of the confounding peaks in two or more pseudo-component spectra, owing to improper constrained latent deconvolution in the presence of noise beyond a particular threshold, but also by way of multiple local optima faced by the heuristic structure learning score-search algorithms. Despite these limitations, this framework is shown to have the potential in reliably developing an end-to-end

modeling framework for species, reaction pathway and kinetic model identification of reactive systems without reliance on prior knowledge. Future work seeks to extend this framework fro complex hydrocarbon systems like bitumen and biomass.

Chapter 7

3D Convolution neural network autoencoder for the prediction of solvent reorganization from MD simulation data

Abstract

Assessing the impact of solvent reorganization in reactive chemical systems is vital in deciding whether or not computational effort in simulating the solvent molecules when running first principles *ab initio* molecular dynamics simulations for these mechanistic reactions, must be expended. A 3D CNN autoencoder is proposed to extract spatio-temporal features from the reactant trajectory simulations, followed by using a distance-based quadratic classifier to assess their *closeness* to features corresponding to the reactant trajectories of systems with strong solvent reorganization. To establish the ground truth of the extent of solvent reorganization in systems, the reactant and product simulations for the condensed phase pyrolytic decomposition of cellobiose was used. Kernel density estimation was used to analyze the probability distributions across the difference in the encoded features between the product and reactant trajectories of the cellobiose systems at different temperatures, as a way of establishing the ground truth. It was found that the cellobiose systems at lower temperatures (100 K, 500K) exhibited larger solvent reorganization, as opposed to the ones at 900 K and 1200 K. To make predictions to quantify solvent reorganization

in other systems other than cellobiose, the reactant trajectories of the aqueous phase acid catalyzed conversion of fructose to HMF in the presence of water as the solvent and DMSO as the cosolvent was chosen. The extent of solvent reorganization in the fructose systems was predicted to increase at first, followed by a linear decrease with increasing DMSO concentrations, and was found to be consistent with trends in the difference between the free energy surface minima.

7.1 Introduction

First-principles (*ab initio*) simulation of chemical systems at the atomistic level provides insights that experiments limited by costs or the feasibility of achieving physical conditions, fail to provide in practice. However, these simulations involve obtaining the energy and forces of a hypothetical configuration of atoms based on computationally expensive quantum mechanical (QM) calculations using wave function theory, on-the-fly electron structure calculations via density function theory (DFT) and potential energy surface (PES) methods, making them intractable for systems with large number of atoms and longer time scales [7]. This can be overcome either by sacrificing accuracy by coarse-graining atomic calculation methods using empiricism-based classical molecular mechanics, or by using machine learning to accelerate QM based (*ab initio* molecular dynamics (AIMD)) simulations [376]. The *acceleration* of QM simulations broadly encompasses (a) machine learning to estimate the PES as a function of atomic coordinates. The force fields computed as gradients of the PES speed up molecular dynamics simulations of chemical structures [377], (b) predictive machine learning to draw inferences from copious amounts of AIMD simulation data, so that it can serve as a computationally efficient surrogate of the same, in the property prediction of molecular systems [71], and (c) generative machine learning to learn probability distributions of molecular representations from AIMD data over macroscopic properties to advance computer aided molecular design [8]. This work seeks to develop a self-supervised machine learning model by way of using a 3D convolution

neural network (CNN) autoencoder for spatio-temporal feature extraction from the AIMD data of the reactant and product configurations, the difference between the features of which is fit to a probability distribution to assess the extent of molecular re-organization. This subsequently informs the development of a Mahalanobis distance-based classifier to predict the extent of solvent reorganization in newer systems by assessing the distance of its reactant features from the distribution of those encoded from systems where the reorganization extent has already been quantified.

The rest of this section will outline some works where ML has accelerated either QM or classical molecular mechanical (MM) simulations so as to access larger length and time scales that would have otherwise been impossible via experiments or simulations alone. This has led to advancements in drug design [9], computational chemistry in molecular and materials modeling [72], retrosynthesis and catalysis [7]. The discussion elaborates the broad areas listed earlier (except generative machine learning, which is still in its infancy [378]), and contextualizes our work and its novelty.

The first application of ML in chemistry was the use of neural networks, Gaussian process regression and kernel regression to extract features from the PES before fitting them to predict energy (machine learning potentials (MLP)) or force fields (machine learning force fields (MLFF)) from QM or DFT-based calculations [71]. MLPs like the Behler-Parrinello networks, or MLFFs trained on DFT calculations (ANInet) were trained to learn from the PES; however, there is evidence of learning from AIMD trajectories from which physically meaningful distance-based power series of atomic coordinates are extracted to be regressed against force fields from DFT [77]. MLPs as computationally tractable surrogates for *ab initio* calculations using neural network architectures to capture spatial atomic interactions via convolutions (SchNet) and physical symmetries via local frame coordinates (DeepPMD), are seen to speed up catalyst screening by efficiently computing reaction activation energies [379] and modeling solid-liquid interfaces [380] in heterogeneous catalysis. DeepPMD has demonstrated how ML can scale the accuracy of *ab initio* calcula-

tions in surface chemistry, from a system with 1000 atoms to that with 100 million atoms by determining the total potential energy as a sum in parts of that of the local atomic environments by using their extracted symmetrical spatio-temporal features, to assess whether or not solvent molecules dissociate at the interface [381]. Developing MLFFs via delta-learning schemes, whereby the loss function minimizes the difference between force fields calculated from ML and that from reference DFT calculations (Δ-NetFF) [382], is only as good as the reference data. To handle this, and also include quantum effects in the reference data, biased sub-sampling of configurations to increase the presence of the least accessible high-energy states from AIMD data[383], and the incorporation of spatio-temporal symmetries [384], has been used to train MLFFs. The high dimensionality of the data, typically 3N atomic coordinates for a system with N atoms, not only limits the use of ML to sample from QM/MM data to construct the PES, but also causes the computational cost of reference data from DFT calculations to scale cubically as the number of electrons in the system [385]. This has been surmounted by using deep neural networks (time-lagged autoencoders (TAE)) to project high dimensional atomic coordinates to a low dimensional space, by constraining the latent features (collective variables) to reproduce the conformational dynamics by maximizing time-lagged autocorrelation within the original space [345].

Deploying ML to derive interpretable insights from AIMD data by predicting the mechanism, rate and yield of chemical systems as functions of thermodynamic properties has been recognized as one of the six grand challenges of the 21st century [85]. Preserving chemical physical intuition by supplying physically meaningful data representations like molecular fingerprints, local environment descriptors or distance metrics [7] facilitates the ML model to recognize meaningful correlations between the system properties and the features extracted from data. ML regression models *viz.* LASSO, random forest, gradient tree boosting and support vector regression, trained on fingerprints extracted from MD simulations are shown to predict solvation free en-

ergy and partition coefficients that have been experimentally validated [74]. However, such ML frameworks perform poorly when they encounter an atomic configuration not present in the training data. Hence, an adaptive ML regression framework that uses a decision engine to query the similarity of fingerprints in the newer configurations to that in the training dataset, based on which the ML model is retrained on the fly using the simulation data of the newer configuration, has been shown to result in more reliable predictions [386]. Aside from regression models, ML classifiers *viz.* Linear discriminant analysis (LDA), support vector classifier (SVC) have been trained on the AIMD data of the decomposition of dioxetane, to extract features from the nuclear coordinates that correspond to either successful or frustrated dissociations [79]. Since the chemexcitation yield is impacted by the time scale of dissociation, it is beneficial to use ML to screen those systems that have successful dissociations before training models to predict dissociation times. Evidence of training Bayesian neural networks on atomic coordinates and velocities from AIMD simulations to predict dissociation times conditioned on a distribution of network weights and biases is seen to be robust to uncertainties and provides physical insights by revealing correlations between inputs and the output predictions via feature saliency maps [387]. The emphasis on retaining physical intuition like symmetry or translation invariance of the atomic configurations in the ML models has popularized the use of convolution neural networks (CNNs) that capture spatio-temporal patterns via parameter sharing, thereby making the predictive ML regression and classification models more efficient [388]. This demands physically informed representations like density grids of molecules to be parsed from simulation data. It can be seen that the density of water molecules stacked over time resulting in 3D grids, or the time-averaged density maps of water, resulting in 2D grids when fed into 3D CNNs and 2D CNNs respectively, efficiently captures spatio-temporal variation in water density while predicting the hydration free energy that rationalizes interfacial hydrophobicity, which drives protein folding mechanisms [75]. However, some reactive systems, for instance

catalytic biomass conversion, not only involve water that dissolves biomass, but also polar aprotic cosolvents to accelerate reaction rates. Maximizing the rate and yield of biomass conversion depends on an optimal solvent:cosolvent ratio, the high throughput screening of which is facilitated by training a ML surrogate that is faster, and eliminates the need of prior knowledge of the reactions involved in biomass conversion processes to generate descriptors computed from simulation data, against which the reaction rates are regressed [80]. In such cases, a 3D histogram that captures the density of water, the cosolvent and the reactant in discrete volume elements of the simulation box was used to generate voxel representations from classical MD data, across 3 separate channels of the molecular species, before being fed into a 3D CNN that is trained to predict reaction rates. This pre-trained model called SolventNet has also been used to predict kinetic solvent parameters which is then used to calculate the thermodynamic selectivity of the product, using just 2ns of classical MD trajectories, thereby speeding up the screening of solvent compositions [389].

From the above discussion, it can be deduced that ML models are effective in accelerating MD simulations only if the cost of training ML models and that of generating the training data (simulation data and labels) is lesser than the cost of explicitly performing first principles calculations. Generating labels to train predictive ML models like reaction rates [80], dissociation time [387] involves experiments or indirect sampling calculations from simulation data to obtain hydration free energies [75]. In this work, we seek to overcome the cost of assigning labels using experiments or sampling calculations by proposing a self-supervised 3D CNN autoencoder to extract spatio-temporal features from the reactant and product trajectories supplied from AIMD simulation. The probability distribution of the root mean square deviation (RMSD) between the features of the trajectories in the reacting cellobiose system is fit via kernel density estimation (KDE), and is used to assess the extent of solvent reorganization. A threshold set on the RMSD is used to assign labels to the features extracted from the initial snapshot. This framework is tested out for another sys-

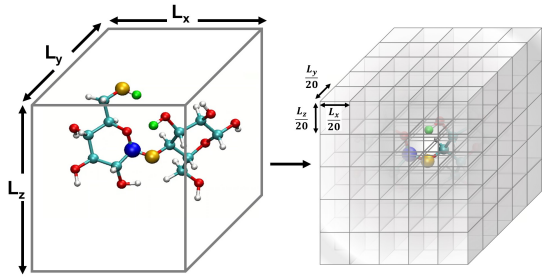
tem, wherein fructose converts to hydroxymethylfurfural (HMF), given water as the solvent and DMSO as the cosolvent. The distance of the encoded fructose features from the distributions of the encoded features of the cellobiose systems, is used to develop a simple quadratic classifier to assess the extents of solvent reorganization in fructose systems with varying solvent:cosolvent ratios, without reliance on simulating the product (HMF) profiles. The rationale behind the choice of this distance-based classifier was to avoid the training costs, as with neural network-based classifiers [86]. This framework has the potential to reduce the cost of MD simulations and that of training ML models by predicting the extent of solvent reorganization, as a basis to inform if the solvent molecules are to be considered while simulating the final product configuration or not.

7.2 Methods

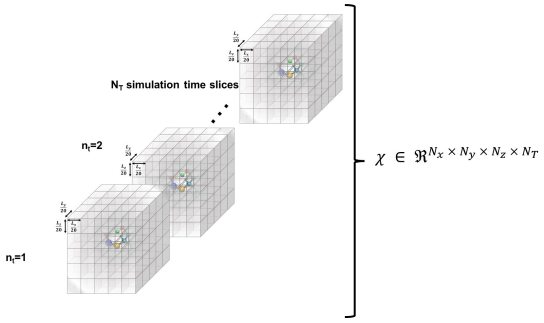
CNNs have been used to extract rotational and translation invariant features from a wide variety of data modalities that are represented as grids, *viz.* 1D grid of vectors for signals, sequences and language models, 2D grid of image pixel matrices stacked into red, green, blue channels, 3D grid of image voxel tensors that comprise snapshots of image pixel matrices along the time axis, as with video data, and higher order grids like 4D grids of image voxel tensors that are stacked into red, blue, green intensity channels [390]. Aside from traditionally using CNNs to process grids of image and video data, the recent past has witnessed their wider application to datasets from molecular simulations that can be expressed as grid data. There is evidence of using the coordinates of all the atoms in a protein structure, stacked in x,y,z channels corresponding to the axes, as an input to train a 1D CNN to learn features in the protein conformation [391]. However, in systems with an arbitrary number of atoms, the input data would have to be padded or truncated accordingly, thereby impacting the patterns the CNN learns. Hence, the representation of atomic coordinates or molecular features extracted from the MD simulations before being fed into the CNN

is vital. 3D molecular features constructed from the atomic partial charges, average number of water contact points, the number of hydrogen bonds, their shapes and sizes, as extracted from MD data, are deemed to be better than the traditional molecular fingerprints in training a CNN to predict free energies of drug-protein binding [76]. In some cases, image-based representations of electrostatic potential that retain spatial information, evaluated using a Gaussian kernel, given the atomic coordinates from simulation data, have been used to train CNNs to predict the energy of the atomic configurations, otherwise obtained from expensive DFT calculations [392]. In the present work, we seek to use a voxel representation of the atomic configurations to train the 3D CNN autoencoder, where the x, y, z atomic coordinates with respect to the size of the simulation box are discretized into volume elements bound by grids [393]. In order for the spatial location of the atoms to be invariant of the grid resolution, the density distribution of atoms in the voxels have widely been used as inputs [388]. Positional atomic densities from MD simulations have been supplied as x-y grids averaged across simulation time steps to train 2D CNNs, or as tensors where separate x-y grids generated for the water and hydrogen molecules have been stacked into channels along the simulation time steps to train 3D CNNs to predict interfacial hydrophobicity [75]. Similarly, positional densities of atoms in the x-y-z space recovered from classical MD simulations, averaged across simulation time steps, have been supplied as tensors stacked into channels grouped by the category of the molecules *viz.*, reactant, solvent and co-solvent, to train 3D CNNs to predict reaction rates[80]. This work differs from those efforts in that the positional densities of the atoms in the x-y-z space from AIMD coordinate trajectories are represented as voxels that are stacked across several simulation time steps, shown in Figure 7.1, to train a self-supervised 3D CNN autoencoder to extract spatio-temporal features.

The x-y-z atomic positions of cellobiose with respect to the simulation box of dimension $L_x \times L_y \times L_z$, is represented as a probability density distribution of the atoms existing in a certain discrete volume element of dimension $\frac{L_x}{N_x} \times \frac{L_y}{N_y} \times \frac{L_z}{N_z}$, where



(a) Voxel representation of atomic density distribution in the simulation box



(b) Voxels stacked across the channel of simulation time steps

Figure 7.1: Spatio-temporal representations of atomic coordinates from AIMD simulation data.

N_x, N_y, N_z (all chosen to be 20 in the present work) are the number of grid elements that the simulation box is discretized into along each axes, as illustrated in Figure 7.1a. Each simulation of the reactant and product configurations for the transglycosylation of cellobiose at four different temperatures (100K, 500K, 900K, 1200K) has been performed over 8ns using GROMACS [394], and the coordinate positions have been recorded every 1ps. The positional voxel density representations of the atomic coordinates are stacked across $N_T = (100\text{ps})$ simulation time steps as shown in Figure 7.1b, to generate T=80 spatiotemporal tensor samples $\mathcal{X} \in \Re^{N_x \times N_y \times N_z \times N_T}$, from the simulation trajectory of either the reactant or product, for a given system.

For the total number of N ($= 2T \times$ number of systems modeled) input tensor

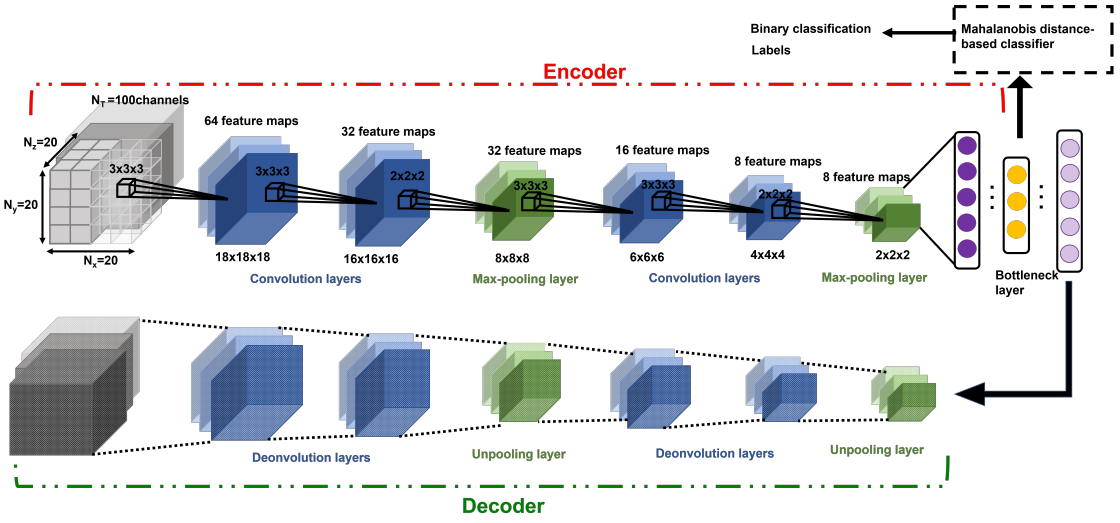


Figure 7.2: Architecture of the 3D CNN autoencoder-based classification model

samples from both the reactant and product trajectories of all systems, $\mathcal{X}^{(i)}$ for $i = \{1, 2, \cdots N\}$, a 3D CNN autoencoder is trained as a hierarchical model that uses a sequence of convolution, activation, pooling, flattening and fully connected layers in the encoder (E), before symmetrically unrolling the sequence in the decoder (D) to reconstruct the input as $\hat{\mathcal{X}}^{(i)} = f_{\mathrm{D}}(f_{\mathrm{E}}(\mathcal{X}^{(i)}, \theta^{E}), \theta^{D})$. The parameters of the encoder and decoder functions ($\theta = \{\theta^{E}, \theta^{D}\}$) of the self-supervised autoencoder network are learned by minimizing the following loss function

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\mathcal{X}^{(i)} - \hat{\mathcal{X}}^{(i)})^2 \tag{7.1}$$

The number of parameters in the 3D CNN autoencoder scales with the choice of hyperparameters that govern the network architecture in the hierarchy of operations. The convolution operation given by $\mathcal{C} \in \Re^{n_{cx} \times n_{cy} \times n_{cz} \times N_T \times q}$ comprises a 3D kernel of dimension $(n_{cx} \times n_{cy} \times n_{cz})$ that performs convolutions across a stride of s voxels in each dimension over the N_T time slices to produce q feature maps in the output $\phi \in \Re^{n_{\phi x} \times n_{\phi y} \times n_{\phi z} \times q}$, as given by the following equation, where $x_1 \in \{1, 2, \cdots, n_{\phi x}\}$, $x_2 \in \{1, 2, \cdots, n_{\phi y}\}$, $x_3 \in \{1, 2, \cdots, n_{\phi z}\}$, $j \in \{1, 2 \cdots, q\}$ and $b_j \in \Re$ is the bias term

$$\phi_{(j)}\left[x_1, x_2, x_3\right] = b_j + \sum_{i=1}^{N_T} \sum_{x_1'=1}^{n_{cx}} \sum_{x_2'=1}^{n_{cy}} \sum_{x_3'=1}^{n_{cz}} \mathcal{C}_{(i,j)}\left[x_1', x_2', x_3'\right] \mathcal{X}^{(i)}\left[x_1 + x_1' + s - 1, x_2 + x_2' + s - 1, x_3 + x_3' + s - 1\right] \tag{7.2}$$

$$n_\phi = \frac{N - n_u + 2P}{s} + 1 \tag{7.3}$$

The dimensions of the output (n_ϕ) across any specific axis is impacted by the respective kernel dimension (n_u), stride (s) and padding (P), if any, given an input of size N, as indicated by Equation 7.3. The purpose of padding is to preserve the input dimensions in the convolved output [395]. However, since the convolution operation is used to downsample the inputs for feature extraction, zero padding has been used in this work. The convolved features are then passed through the a nonlinear activation function that does not modify the dimensions.

$$f(\phi) = \max(0, \phi) \tag{7.4}$$

$$v' = f(W_{fc} v + b_{fc}) \tag{7.5}$$

As compared to activation functions like the *tanh, sigmoid*, the rectified linear unit, given by Equation 7.4, is preferred as it does not suffer from gradient saturation in the event of large magnitude inputs, thereby increasing the sensitivity of the model to input representations[396]. Following this, the pooling operation ($\mathcal{P} \in \Re^{n_{px} \times n_{py} \times n_{pz} \times q}$) is used to downsample the activated output, to make the encoded representations invariant to minor translations in the input [397], resulting in a pooled output $\phi_p \in \Re^{n_{\phi x'} \times n_{\phi y'} \times n_{\phi z'} \times q}$. This follows on the same lines as Equation 7.2 and Equation 7.3, except that there is no bias translation and the 3D max pooling kernel of dimension ($n_{px} \times n_{py} \times n_{pz}$) merely outputs a maximum valued scalar as it strides over s voxels at a time along the axes, for all the input feature maps. Several units comprising the

aforementioned convolution, activation and pooling operations can be hierarchically stacked to transform the input sample $\mathcal{X}^{(i)}$ to a tensor $\phi' \in \Re^{N'_x \times N'_y \times N'_z \times p}$, of p feature maps, before finally flattening it to result in a vector $v \in \Re^{N'_x.N'_y.N'_z.p \times 1}$ that is fed into a fully connected layer to result in an output feature vector $v' \in \Re^{f \times 1}$, given in Equation 7.5, where $W_{fc} \in \Re^{f \times N'_x.N'_y.N'_z.p}$ and $b_{fc} \in \Re^{f \times 1}$ are the weights and biases, parametrizing the fully connected layer, respectively. There can be many such fully connected layers as indicated by the schematic in Figure 7.2, to finally obtain f' latent features in the bottleneck layer of the encoder. The structure of the decoder is seen to mirror that of the encoder in reconstructing the input from the features of the bottleneck layer via a series of upsampling operations like deconvolution and unpooling. If the convolution operation is expressed as the multiplication of the *Toeplitz* block of strided kernel coefficients with the input, then the deconvolution can be expressed its inverse, where upsampling is achieved by mulipication with the transpose *Toeplitz* block [398]. Similarly, unpooling is performed by inserting the maximum values into their index positions, cached during the pooling operation.

Once trained, the bottleneck layer of the encoder is used to extract latent features from the AIMD trajectories of the reactant that are plugged into the quadratic distance-based classifier, to predict whether or not the solvent molecules reorganize significantly in the product configuration. This is based on the key assumption that samples with the same label should have similar latent features extracted by the autoencoder [399]. However, developing a classifier to discriminate between latent features is supervised, in that there is a requirement of ground truth labels, the generation of which is expensive and time consuming [400]. This is surmounted by calculating the root mean square deviation between features of the product and reactant trajectory for a system, followed by using KDE to probabilistically assess systems with a higher extent of reorganization using a threshold, based on which labels are assigned to the features extracted from the reactant trajectory samples to develop the Mahalanobis classifier, a choice deliberately made to also eliminate the cost of

training neural network classifiers [86].

$$\text{RMSD} = f_E(\mathcal{X}_{\text{product}}) - \frac{1}{T}\sum_{t=1}^{T} f_E(\mathcal{X}^{(t)}_{\text{reactant}}) \tag{7.6}$$

$$P(x) = \sum_{t=1}^{T} K\left(\frac{x - RMSD_t}{b}\right) \tag{7.7}$$

$$P(y=1|\text{RMSD}) = \frac{\sum_{t=1}^{T} P(\text{RMSD}_t|y=1)}{\sum_{t=1}^{T} P(\text{RMSD}_t|y=0) + \sum_{t=1}^{T} P(\text{RMSD}_t|y=1)} \tag{7.8}$$

For a particular system, the RMSD $\in \Re^{T\times 1}$, pointing to the deviation of features of the product trajectory samples from the average of features across the reactant trajectory samples, is given in Equation 7.4. The probability distribution of $x \in [\min(\text{RMSD}), \max(\text{RMSD})]$ for each of the systems is fit using kernel density estimation (Equation 7.7), where the choice of the kernel function (K) and bandwidth (b) are guided by grid search optimization [401]. The systems with the highest and least RMSD distributional modes are designated labels, $y = \{0, 1\}$ corresponding to a large and small extent of solvent reorganization, respectively. The posterior probability of the other systems being labelled 1, given their RMSDs and the assumption of equally likely priors is determined using Equation 7.8 [402]. If the posterior probability of the system exceeds a certain threshold, all the features corresponding to the samples in the reaction trajectory are designated a label 1, else 0, giving rise to labeled samples $\{f_E(\mathcal{X}^{(i)}_{\text{reactant}}), y^{(i)}\}$ for all $i \in \{1, 2, \cdots N\}$ supplied as training data to a quadratic classifier based on the Mahalanobis distance. The classifier is trained to detect solvent reorganization patterns from a reduced set of encoded features of the reactant trajectories in a kernel space ($f'_E(\mathcal{X}^{(i)}_{\text{reactant}})$), by assessing their closeness to positively labelled trajectories, using the mean and covariances of their kernelized features as

follows:

$$\left[f'_E(\mathcal{X}^{(i)}_{\text{reactant}}) - \overline{f_E}(\mathcal{X}^{(i)}_{\text{reactant}}|y^{(i)}=1)\right]^T Cov^{-1} \left[f'_E(\mathcal{X}^{(i)}_{\text{reactant}}) - \overline{f_E}(\mathcal{X}^{(i)}_{\text{reactant}}|y^{(i)}=1)\right] \tag{7.9}$$

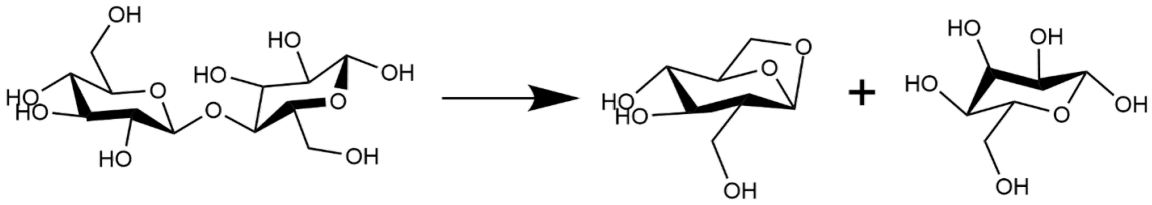
The loss function of the 3D CNN autoencoder is minimized by stochastic gradient descent, implemented using the *Adam* optimizer on PyTorch with a learning rate of 10^{-3}. The process of gradient descent involves computing the gradient of the loss function with respect to the weights of the layers and is efficiently performed via the backpropagation algorithm. The distance-based classifier is then implemented to assess the extent of solvent reorganization in newer systems by measuring the distance between their features and the distribution of features extracted from the labelled systems.

7.3 Results and Discussion

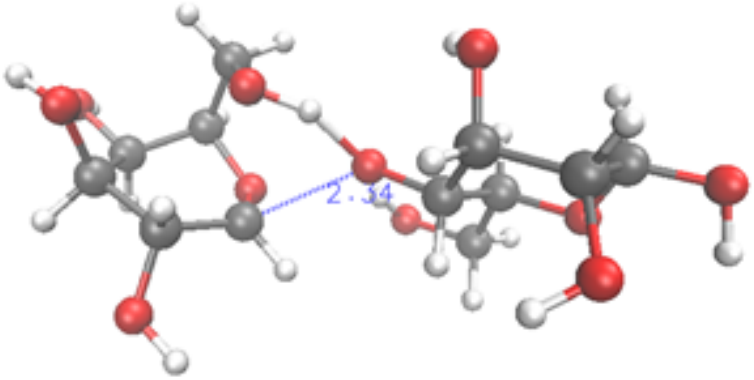
This section demonstrates the above developed framework with application to assessing the impact of solvent reorganization in two of the following reactive systems: (i) the condensed phase pyrolytic decomposition of cellobiose, and (ii) the aqueous phase acid catalyzed conversion of fructose to 5-hydroxyl methyl furfural (HMF). The probability distribution across the root mean square deviation (RMSD) between features of the reactant and product configurations of the reacting cellobiose system, extracted from the 3D CNN autoencoder have been used to assess the extent of solvent reorganization. Insights drawn from the map between features of the reactant cellobiose profiles and the extent of solvent reorganization, are shown to generalize well across the starkly different reacting fructose systems, when it comes to predicting the extent solvent reorganization from just its reactant profiles. This eliminates computational efforts in explicitly simulating the product HMF configurations to deduce the same. The results from the machine learning framework have been validated against a thermodynamic basis, for both the training process on the cellobiose systems as outlined

in Section 7.3.1, and, for the testing process on the fructose systems as outlined in Section 7.3.2.

7.3.1 Training the machine learning model on the cellobiose systems



(a) Glucose residues in cellobiose shift from ground state chair conformers to boat conformations prior to glycosidic bond cleavage [403]



(b) DFT calculated transition state for transglycosylation

Figure 7.3: Condensed phase pyrolytic decomposition of cellobiose

Pyrolytic cellobiose decomposition in the condensed phase is primarily initiated by the glycosidic C-O-C bond cleavage, as shown in Figure 7.3a. Gas phase DFT calculations carried out on an isolated molecule for the transglycosylation mechanism (Figure 7.3a) that exhibits one of the least enthalpic barriers for cellobiose decomposition is given in Figure 7.3b. In the condensed phase, the influence of the neighboring molecules could potentially alter the reaction chemistry and energetics. The reorganization of molecules around the reacting species long the reaction coordinates, sheds light on how the condensed phase affects reaction energetics. The solvent reorganization around the reactant cellobiose for the transglycosylation mechanism has been

simulated on GROMACS 2018.7 at four different temperature *viz.* 100 K, 500 K, 900 K and 1200 K. Configurations of the reacting species from the first-principles calculations has resulted in MD trajectories for the reactant and product profiles, recorded evert 500 simulation steps over a duration of 8ns.

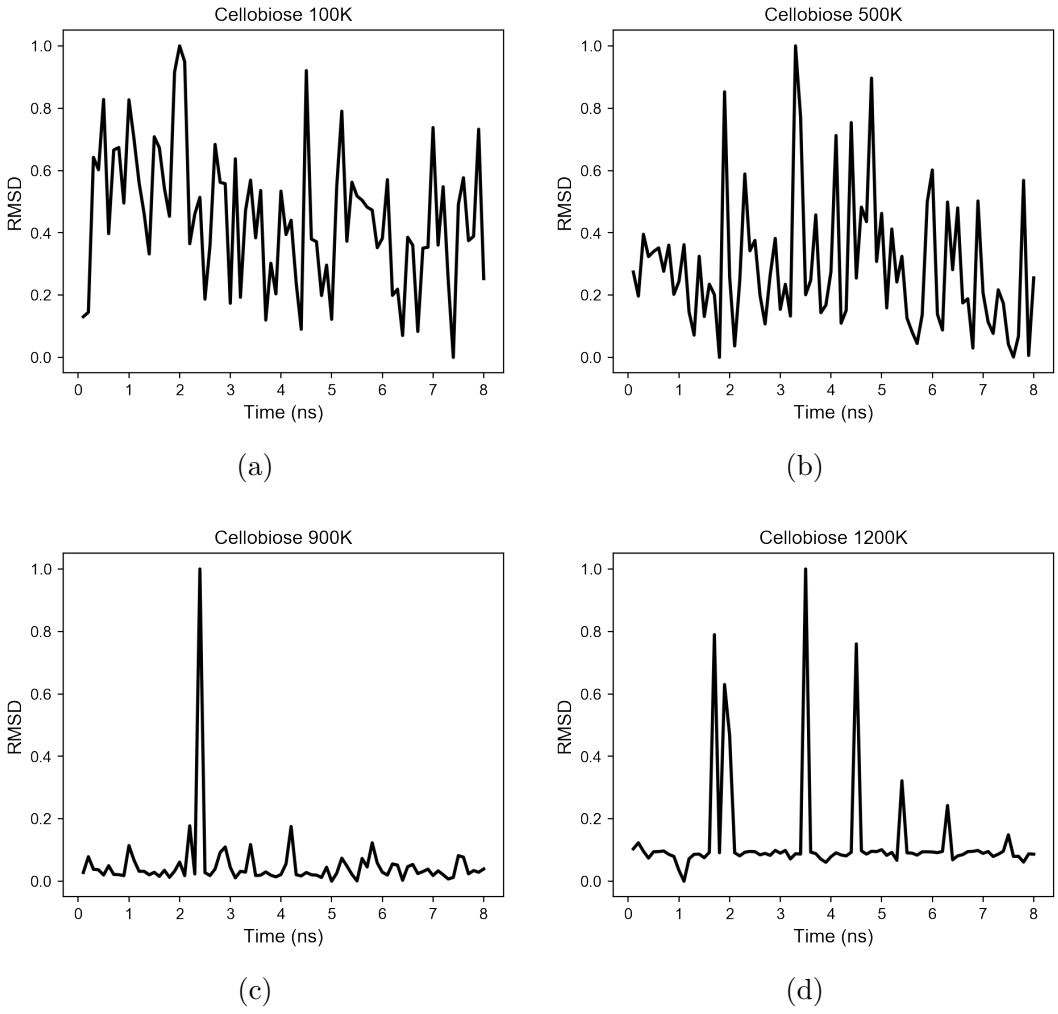
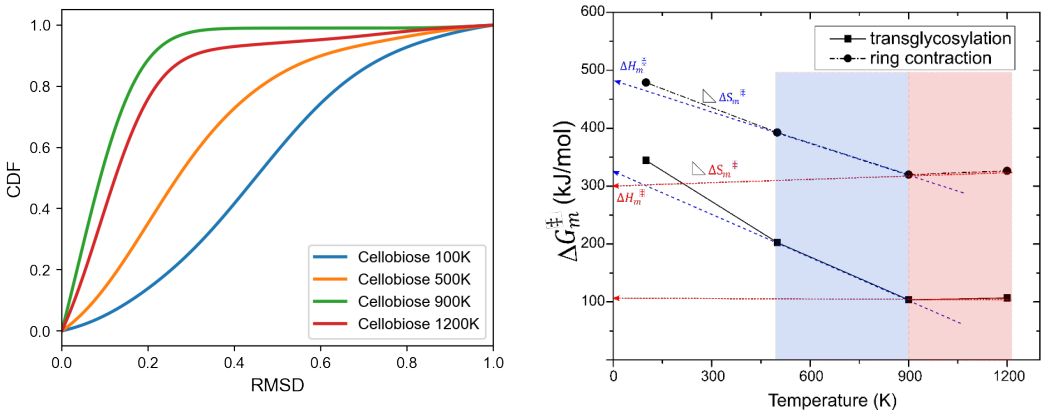


Figure 7.4: RMSD between encoded features of samples in the product trajectory and the mean encoded features across samples in the reactant profile

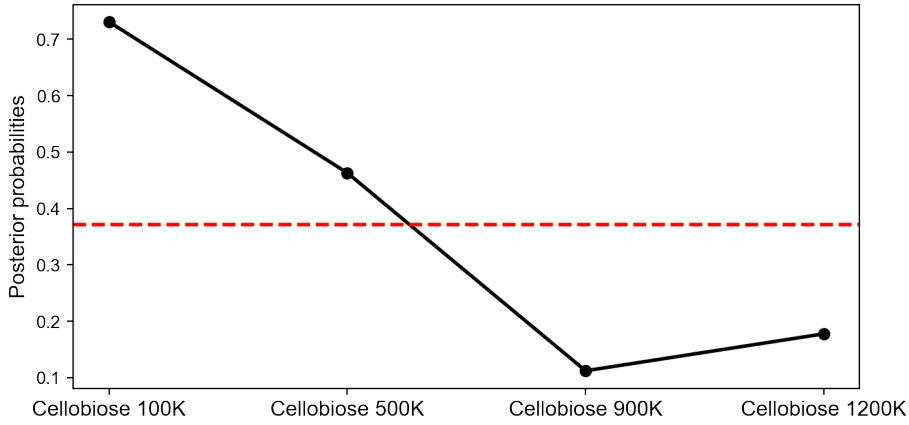
The voxels of atomic densities stacked over evert 100ps, results in 80 samples each, from the reactant and product profiles, at each finite temperature simulation of cellobiose. A 3D CNN autoencoder with a structure as given in Figure 7.2 is trained on these samples to extract encoded feature representations in the bottleneck layer. The root mean square deviation (RMSD) between these features of the samples in

the product profiles and the average of the encoded features across all samples in the reactant profiles have been indicated in Figure 7.4 across the four temperatures. Since the encoded features are extracted from equilibrium simulations of the cellobiose systems, it is permissible to average the encoded features across all samples in the reactant trajectory, as a reference against which the encoded features of the product trajectory are compared, when defining the RMSD. The use of RMSD as a descriptor of the extent of solvent reorganization circumvents the need of expensive sampling strategies [75] to compute metrics from the simulation data. It can be seen qualitatively from Figure 7.4c and Figure 7.4d that the effects of reorganization are less prominent at higher temperatures. A kernel density estimation is used to quantify the probability distributions of the RMSDs in accordance with Equation 7.7 using a Gaussian kernel (K). The bandwidth chosen by grid search cross validation is found to be optimal at 0.1, and results in cumulative density distributions given in Figure 7.5a, from which the CB 100K system and the CB 900K system, are seen to have the most (class 1) and least possible (class 0) solvent reorganizations, respectively. The density distributions of the CB 100K and CB 900K systems are then used to quantify the posterior probability of a system being recognized as significantly reorganized, conditioned on its RMSD, as outlined in Equation 7.8. The average of the posterior probabilities of across all the cellobiose systems is then used as threshold as shown in Figure 7.5c, to recognize if significant solvent reorganization has been observed in a system or not.

In Figure 7.5, the kernel density estimation of the RMSD to quantify the extent of solvent reorganization from the posterior computations (Figure 7.5c)is shown to be consistent with trends in the Gibbs free energy barrier for the transglycosylation mechanism of the reacting cellobiose molecules in the melt phase across the four different temperatures, as given by Figure 7.5b. The activation free energy barrier(FEB) is seen to decrease almost linearly with increasing temperatures and asymptotes above 900 K, at a constant value of $\sim$ 105 KJ/mol. The reduction in the FEB in going



(a) (b)



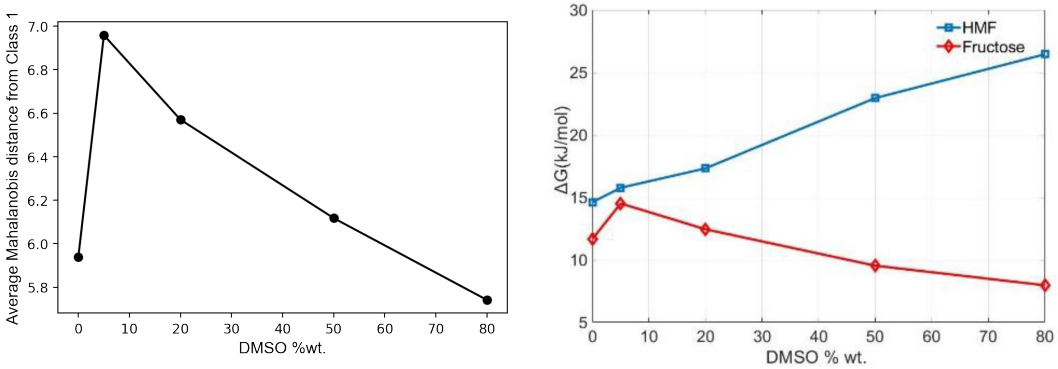
(c)

Figure 7.5: (a)Cumulative probability distributions of the RMSD, (b) Free energy barrier vs temperature profile for cellulose decomposition showing two reaction regimes transitioning at 900K. The slope and y-intercept gives the entropic and enthalpic contributions to the free energy barrier, respectively. The tangents are fitted between 500K-900K for the low temperature (blue dash lines) and 900K-1200K for high temperature regime (red dash lines), (c)Posterior probabilities

from 100 K to 900 K is 267.76 KJ/mol, and suggests a strong impact of the finite temperature environment on tranglycosylation. The slope and the y-intercept of the free energy vs. temperature plot gives the entropic and enthalpic contributions, respectively. The constant slope of the FEB curve at low temperature is indicative of the constant gain in entropy ($\Delta S_m^{\ddagger}$) of 334.69 J/mol-K for the decomposition of the cellobiose melt to LGA. At higher temperatures the FEB flattens indicating that the entropic contribution to the barrier is zero, making it an enthalpy-controlled regime.

Hence the linear slopes of the low temperature and the high temperature curves form distinct decomposition regimes [404]. The finite temperature affects the reorganization of the neighboring molecules in going from the crystal phase to the melt, and is seen to thus impact the cellulose chemistry. This validates the inferences drawn from the RMSD probability distributions in quantitatively deciphering the extents of reorganization in the decomposition of cellobiose across different temperatures.

7.3.2 Testing the machine learning model predictions on the fructose systems



(a) (b)

Figure 7.6: Predictions of the extent of solvent reorganization in fructose: (a) Average distance of the sample features of the fructose trajectory from the cellobiose 100 K and 500 K systems across different solvent composition, (b) Free energy difference between the FES minima corresponding to the migration of the hydronium ion from the bulk solvent to the first solvation shell of fructose at different DMSO concentrations.

This section focuses on using the above framework that has been trained on the cellobiose systems to decipher the extents of solvent reorganization in other reacting systems impacted by solvents. The aqueous phase acid-catalyzed reaction of the biomass-derived species, fructose, is a chosen as a system to test the model predictions. Polar aprotic solvents like dimethylsulfoxide (DMSO) are known to result in higher reactivities in the conversion of fructose to 5-hydroxyl methyl furfural (HMF). The stability of the catalyst (hydronium ion) in the first solvation shell of fructose as compared to the bulk of the solvent is seen to differ across the composition ratios

of the solvent:cosolvent, thereby impacting the reaction kinetics in the conversion of fructose to the final HMF product. It would be beneficial to assess from just the AIMD simulations of the fructose trajectory at different solvent compositions, the impact of solvent reorganization on the final product formation, without explicitly simulating the HMF trajectories. This is proposed to be achieved by obtaining the spatio-temporal encoded features of the reactant trajectory of the different fructose systems, from the 3D CNN autoencoder before classifying them using the Mahalanobis classifier in terms of the distance of the features of these systems from the distributions of features from the cellobiose systems, for which the ground truth of the extents of reorganization has already been established.

The AIMD simulations of the fructose systems on GROMACS 2018.7 at different solvent compositions have been recorded every 200 time steps for a duration of 10ns. Hence, a voxel sample over every 100ps, would lead to 100 samples from each of these trajectories. The spatio-temporal 3D CNN provides encoded latent features for the 100 samples of each fructose trajectory. Comparing the distances of each of these 100 fructose features from the distributions of features of the cellobiose systems at 100K and 500K (systems established to reorganize significantly), followed by a similar comparison against features from the lesser reorganizing cellobiose systems, before assigning a 0/1 class label to each of the samples in the fructose trajectory, does not have a sound basis. This is because we are interested in quantifying the overall extent of reorganization of the system as a whole using these equilibrium simulations, and even highly reorganizing systems may have points in their trajectories where reorganization is not significant. Therefore, the average of the Mahalanobis distance of features across all samples in the fructose trajectory, from the significantly reorganizing cellobiose systems, for different solvent compositions is shown in Figure 7.6a. When assessing the extent of solvent reorganization, it is better to compare the distance of the average sample features of the reactant trajectory from the highly reorganizing cellobiose systems, instead of how much further away they are from the weakly re-

organizing systems (*i.e* Class 0) because it is not that reorganization is completely absent in them. This would then bias the reference used to make the comparison.

The results in Figure 7.6a are seen to concur with the trends of the relative stability of hydronium ions at different DMSO concetrations given by the difference (ΔG) in the free energy surface (FES) minimum corresponding to the hydronium ion in the first solvation shell of fructose and that of the FES minimum corresponding to the hydronium ion in the bulk solvent [405], as shown in Figure 7.6b. The increase in ΔG when DMSO goes from 0 to 5 % wt can be attributed to the instability of DMSO molecules in the bulk solvent, generating a rich local domain of DMSO molecules near fructose while the water (solvent) molecules in the bulk stabilize the hydronium ion. However, as the DMSO concentrations increase from 5 to 80 % wt. a clear descending trend is observed in ΔG and the average distance from Class=1 cellobiose systems, suggesting that the relative stability of hydronium ions in the first solvation shell of fructose increases. This is largely owing to the limited availability of water molecules in the bulk to stabilize the hydronium ions, forcing them to interact with the reactant fructose, thereby making the effects of solvent reorganization more pronounced.

7.4 Conclusions

This study has demonstrated the effectiveness of a self-supervised framework of training predictive machine learning models that are not only computationally efficient to train but also propose to reduce the cost of AIMD simulations. A 3D CNN autoencoder for spatio-temporal feature extraction is trained on both the reactant and product configurations in the condensed phase reaction of cellobiose at different temperatures. The probability distributions across the RMSD of the features between the product and the reactant profiles are seen to show a higher probability of reorganization for the lower temperature finite simulations at 100 K and 500 K. These findings are consistent with the linear decrease in the free energy barrier with increasing temperatures. A quadratic classifier based on the Mahalonobis distance metric

is then used to calculate the average distance of the reactant features in the fructose trajectory from the distribution of the reactant features of the strongly reorganizing cellobiose systems, as a means of assessing the extent of solvent reorganization from just the reactant profiles of starkly different systems. The average Mahalanobis distance is seen to increase at first, and then decrease almost linearly with increasing concentrations of DMSO, consistent with the trends in the difference between the FES minima, thereby pointing to a larger impact of solvent reorganization with increasing DMSO concentrations when the fructose systems are seen to get *closer* to the cellobiose systems at 100 K and 500 K. The generalization of predictions to arrive at consistent results for systems other than what the ML models were trained on, can be used to limit computational efforts when simulating solvent effects in systems where its effects are found to be lower. This has the potential to accelerate the screening of systems for solvent impact when designing processes for chemical reactions.

Chapter 8

Concluding remarks and future scope

The use of machine learning to enhance the interpretability of data-driven models and also the automation capabilities of predictive models, is seen to make it an attractive tool for modeling complex reactive systems. This facilitates process monitoring strategies for the mitigation of the environmental impact in designing processes for the upgrading of complex feedstocks to produce chemicals of value and their compliance with pipeline transport. However, the setbacks of building inferential machine learning models for reactive systems where prior knowledge of species and reaction pathways underlying their conversion is obscure, concern the recovery of interpretable insights. Also, predictive ML models face deployment challenges when being used to limit the computational cost of AIMD simulations, when screening reactive systems on a mechanistic basis.

In this context, the following are the two overarching aims of this thesis: (i) the development of interpretable end to end machine learning models for species and reaction pathway identification, followed by estimating kinetics by using molecular-level information of reactive systems, from spectroscopic sensors (ii) the development of a computationally efficient self-supervised predictive ML model to increase the automation capabilities of AIMD simulations for reactive systems, where mechanistic knowledge is available. This chapter summarizes the key findings of this thesis and

highlights directions for future work.

8.1 Summary

Inferential machine learning models for black box systems.

Chapters 2, 3 and 4 demonstrate an increasing order of sophistication in spectroscopic decomposition to extract latent features to facilitate the generation of plausible reaction hypotheses. The first approach applies *i.e.* the self-modeling multivariate curve resolution (SMCR) of FTIR data, followed by the use of quantitative metrics based on the absorption intensity bands of the recovered spectral features to hypothesize reaction pathways. The next approach, developed the joint non-negative matrix factorization (JNMF) as a data fusion analogue of SMCR to extract complementary information from the FTIR and ^{1}H-NMR sensors, the latent features amongst which Bayesian structure learning is used to causally infer reaction pathways. Finally, joint non-negative tensor factorization (JNTF) was developed to be a structure-preserving data fusion analogue of JNMF. JNTF was seen to limit the loss of chemical information as opposed to JNMF where heuristically relaxing the latent factorization rank was seen to recover additional pathways that the JNTF was seen to recover without such relaxation heuristics to begin with. The latent factorization and causal inference were then used to interpret the dynamic modes identified from realtime spectroscopic data, via hidden semi-Markov models in Chapter 5. The time scales and cyclical dynamics of reaction mechanisms were inferred from the duration and transition dynamics of the modes, using spectroscopic data associated with realtime changes in operating temperatures. Reaction mechanisms at higher temperatures were seen to persist over longer durations than the ones at lower temperatures, and the cyclical reactions dynamics recovered points to the hydrolysis of esters to give alcohols that reversibly combine with acids to results in esters, while encountering thermal cracking and hydrogen transfer in the interim to form condensed products. Finally, in Chapter 6 the latent projections onto the temporal mode of data collection has been used to

develop kinetic models constrained by the adjacency matrix from the Bayesian network structure, causally inferred among the spectral features, by way of using the chemical reaction neural ODEs that are structured to incorporate the law of mass action and the Arrhenius law of temperature dependence. However, the kinetic model development was demonstrated using synthetic data from a reaaction template but not from spectroscopic data of complex feedstocks, owing to the lack of an exhaustive ground truth kinetic model for validation.

Predictive machine learning models for mechanistic systems.

Chapter 7 leverages self-supervised deep learning architectures to extract spatio-temporal features from the AIMD simulation trajectories of reacting systems to predict the extent of solvent reorganization from just the reactant profiles, so that in the event reorganization is found to be minimal, a decision can be made to eliminate the solvent molecules when simulating the final product profiles. This is proposed to save computational efforts of the AIMD simulations and increase its automation capacity when deployed to screen multiple reactant systems. In addition to using a self-supervised model that circumvents the need of expensive sampling calculations to compute target labels from the simulation data, the use of a simple distance-based quadratic classifier that obviates the need for training as with typical neural network classifiers that learn a decision boundary, the cost of developing predictive machine learning framework is considerably minimized. The probability distribution across the differences in the encoded features between the reactant and product simulations of the condensed phase pyrolytic decomposition of cellobiose, fit using kernel density estimation was seen to point to higher extents of solvent reorganization in systems at lower temperatures (100 K, 500K). This was then used as a basis to predict the extent of solvent reorganization in the acid catalyzed conversion of fructose to HMF, by computing the distance of the encoded features of the fructose trajectories from the features corresponding to the reactant trajectories of cellobiose at 100 K and 500 K, at different concentrations of DMSO. It was seen that the distance slightly increased

before decreasing almost linearly with increasing DMSO concentrations, pointing to more significant solvent reorganizations. The findings for the cellobiose and the fructose systems have been validated using a thermodynamic basis of these conversion processes.

8.2 Future research directions

The development of high fidelity models for reactive systems where the prior knowledge of the exhaustive enumeration of species and the underlying conversion pathways is obscure, could be used to further investigations in the following research directions:

- Control and optimization strategies for processes concerning complex reactive systems based on the system inferential models that have been developed, has the potential for realtime applications.

- Automated mapping of the reaction network hypotheses generated from the Bayesian networks to real chemistry in the databases, thereby surmounting the heuristics in validating the inferences drawn from the modelling efforts demonstrated in this thesis. The use of molecular fingerprint representations of the latent spectral features to query candidate molecules from a reaction template in the database, based on a similarity index could pave way for the automation.

- Attention-based feature selection using 3D-Resnet architectures to focus on only specific absorption intensity spectral channels by eliminating redundant information in spectroscopic data has the potential to accelerate spectroscopic deconvolutions, and when incorporated in the data fusion architectures for latent factorization that have been demonstrated in this thesis, could reduce the computational burden of these models for realtime applications.

The development of computationally efficient predictive machine learning models to increase the automation capabilities of AIMD simulations for mechanistic reactive

systems opens up the following research avenues:

- Machine learning surrogates for the long term predictions of the solvent coordinates in the final product profiles, given the reactant trajectories in systems where solvent reorganization is identified to be significant, using the framework demonstrated in Chapter 7.

- Data-driven approach of designing an interpretable low-dimensional manifold of collective variables onto which the evolution of the mechanistic simulations in the high dimensional space of the atomic coordinates can be effectively projected. Advances in this direction have the potential to use the low-dimensional manifold in simulating reaction events in reduced time.

- The reaction hypotheses generated from the inferential machine learning models that have been developed for reactive systems where the prior knowledge of the species and reactions is obscure, could be used to inform the search space of running mechanistic simulations, the insights from which could foster a synergy between data-driven models and first-principles simulations in modeling complex feedstocks across different length and time scales.

Bibliography

[1] B. M. Wise and N. B. Gallagher, "The process chemometrics approach to process monitoring and fault detection," *Journal of Process Control*, vol. 6, no. 6, pp. 329–348, 1996, ISSN: 0959-1524. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0959152496000091.

[2] D. C. Boffito and D. Fernandez Rivas, "Process intensification connects scales and disciplines towards sustainability," *The Canadian Journal of Chemical Engineering*, vol. 98, no. 12, pp. 2489–2506, 2020, ISSN: 0008-4034. DOI: 10.1002/cjce.23871. [Online]. Available: https://doi.org/10.1002/cjce.23871.

[3] M. A. Nemeth, "Multi- and Megavariate Data Analysis," *Technometrics*, vol. 45, no. 4, pp. 362–362, 2003, ISSN: 0040-1706. DOI: 10.1198/tech.2003.s162. arXiv: arXiv:1011.1669v3. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1198/tech.2003.s162.

[4] T. Kourti, "Process analytical technology beyond real-time analyzers: The role of multivariate analysis," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 3-4, pp. 257–278, 2006, ISSN: 10408347. DOI: 10.1080/10408340600969957.

[5] S. C. Rutan, A. de Juan, and R. Tauler, "2.06 - introduction to multivariate curve resolution," in *Comprehensive Chemometrics (Second Edition)*. Oxford: Elsevier, 2020, pp. 85–94, ISBN: 978-0-444-64166-3. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124095472148905.

[6] E. Skibsted and S. Engelsen, "Spectroscopy for process analytical technology (pat)," in Dec. 2010, 2651–2661, ISBN: 978-0-12-374417-3. DOI: 10.1016/B978-0-12-374413-5.00026-9.

[7] J. A. Keith *et al.*, "Combining machine learning and computational chemistry for predictive insights into chemical systems," *Chemical Reviews*, vol. 121, no. 16, pp. 9816–9872, 2021, ISSN: 0009-2665. DOI: 10.1021/acs.chemrev.1c00107. [Online]. Available: https://doi.org/10.1021/acs.chemrev.1c00107.

[8] A. S. Alshehri, R. Gani, and F. You, "Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: State-of-the-art and future directions," *Computers & Chemical Engineering*, vol. 141, p. 107 005, 2020, ISSN: 0098-1354. DOI: 10.1016/j.compchemeng.2020.107005. [Online]. Available: http://dx.doi.org/10.1016/j.compchemeng.2020.107005.

[9] S. Jamal, A. Grover, and S. Grover, "Machine learning from molecular dynamics trajectories to predict caspase-8 inhibitors against alzheimer's disease," *Frontiers in Pharmacology*, vol. 10, p. 780, 2019, ISSN: 1663-9812. DOI: 10.3389/fphar.2019.00780. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphar.2019.00780.

[10] L. M. Yañez Jaramillo and A. De Klerk, "Partial Upgrading of Bitumen by Thermal Conversion at 150-300 °c," *Energy and Fuels*, vol. 32, no. 3, pp. 3299–3311, 2018, ISSN: 15205029. DOI: 10.1021/acs.energyfuels.7b04145.

[11] J. Yue, J. C. Schouten, and T. A. Nijhuis, "Integration of Microreactors with Spectroscopic Detection for Online Reaction Monitoring and Catalyst Characterization," 2012. DOI: 10.1021/ie301258j.

[12] H. Fleischer, V. Q. Do, and K. Thurow, "Online Measurement System in Reaction Monitoring for Determination of Structural and Elemental Composition Using Mass Spectrometry," 2019. DOI: 10.1177/2472630318813838.

[13] P. W. Fedick, R. L. Schrader, S. T. Ayrton, C. J. Pulliam, and R. G. Cooks, "Process Analytical Technology for Online Monitoring of Organic Reactions by Mass Spectrometry and UV–Vis Spectroscopy," *Journal of Chemical Education*, vol. 96, pp. 124–131, 2018, ISSN: 0021-9584. DOI: 10.1021/acs.jchemed.8b00725.

[14] A. Puliyanda, K. Srinivasan, K. Sivaramakrishnan, and V. Prasad, "A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems," *Digital Chemical Engineering*, vol. 2, p. 100 009, 2022, ISSN: 2772-5081. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772508121000090.

[15] A. L. Pomerantsev and O. Y. Rodionova, "Process analytical technology: A critical view of the chemometricians," *Journal of Chemometrics*, vol. 26, no. 6, pp. 299–310, 2012, ISSN: 08869383. DOI: 10.1002/cem.2445.

[16] S. D. Brown, T. B. Blank, S. T. Sum, and L. G. Weyer, "Chemometrics," *Analytical Chemistry*, vol. 66, no. 12, pp. 315–359, 1994, ISSN: 15206882. DOI: 10.1021/ac00084a014.

[17] J. Felten, H. Hall, J. Jaumot, R. Tauler, A. De Juan, and A. Gorzsás, "Vibrational spectroscopic image analysis of biological material using multivariate curve resolution-alternating least squares (MCR-ALS)," *Nature Protocols*, vol. 10, no. 2, pp. 217–240, 2015, ISSN: 17502799. DOI: 10.1038/nprot.2015.008.

[18] B. Akbar, V. P. Gopi, and V. S. Babu, "Colon cancer detection based on structural and statistical pattern recognition," *Electronics and Communication Systems (ICECS), 2015 2nd International Conference on*, no. Icecs, pp. 1735–1739, 2015.

[19] L. Fortuna, S. Graziani, A. Rizzo, and M. G. Xibilia, *Soft Sensors for Monitoring and Control of Industrial Processes*. 2007, ISBN: 1846284805. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

[20] B. Lavine and J. Workman, "Chemometrics," *Analytical Chemistry*, vol. 78, no. 12, pp. 4137–4145, 2006, ISSN: 0003-2700. DOI: 10.1021/ac060717q. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/ac060717q.

[21] J. Laxalde, N. Caillol, F. Wahl, C. Ruckebusch, and L. Duponchel, "Combining near and mid infrared spectroscopy for heavy oil characterisation," *Fuel*, vol. 133, pp. 310–316, 2014, ISSN: 00162361. DOI: 10.1016/j.fuel.2014.05.041.

[22] K. Varmuza, "Clustering methods," in *Pattern Recognition in Chemistry*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1980, pp. 92–96, ISBN: 978-3-642-93155-0. DOI: 10.1007/978-3-642-93155-0_7. [Online]. Available: https://doi.org/10.1007/978-3-642-93155-0_7.

[23] R. M. Belchamber, D. Betteridge, Y. T. Chow, T. J. Sly, and A. P. Wade, "The application of computers in chemometrics and analytical chemistry," *Analytica Chimica Acta*, vol. 150, pp. 115–128, 1983, ISSN: 0003-2670. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267000854641.

[24] A. de Juan and R. Tauler, "Multivariate Curve Resolution (MCR) from 2000: Progress in concepts and applications," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 3-4, pp. 163–176, 2006, ISSN: 10408347. DOI: 10.1080/10408340600970005. arXiv: arXiv:1011.1669v3.

[25] D. T. Tefera, A Agrawal, L. M. Yanez Jaramillo, A De Klerk, and V Prasad, "Self-modeling multivariate curve resolution model for online monitoring of bitumen conversion using infrared spectroscopy," *Industrial and Engineering Chemistry Research*, vol. 56, no. 38, pp. 10 756–10 769, 2017, ISSN: 0888-5885. DOI: 10.1021/acs.iecr.7b01849. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85030255178{\&}doi=10.1021{\%}2Facs.iecr.7b01849{\&}partnerID=40{\&}md5=9a6ae31f7ad8d61841140cff98b126f4.

[26] R. A. Harshman and M. E. Lundy, "Img_3956 (1)," vol. 18, pp. 39–72, 1994, ISSN: 978-3-642-30122-3. arXiv: arXiv:1011.1669v3.

[27] B. S. Everitt and P. M. Kroonenberg, *Three-Mode Principal Component Analysis*. 1. 1986, vol. 42, p. 224, ISBN: 9066950021. DOI: 10.2307/2531268. [Online]. Available: https://www.jstor.org/stable/2531268?origin=crossref.

[28] H. R. Keller and D. L. Massart, "Evolving Factor Analysis References:," vol. 32, no. 1985, pp. 209–224, 1992.

[29] H. Shinzawa, J.-h. Jiang, and M. Iwahashi, "Self-modeling curve resolution (SMCR) by particle swarm optimization (PSO)," no. October 2017, 2007. DOI: 10.1016/j.aca.2006.12.004.

[30] A. Malik, A. De Juan, and R. Tauler, "Multivariate curve resolution: A different way to examine chemical data," *ACS Symposium Series*, vol. 1199, pp. 95–128, 2015, ISSN: 19475918. DOI: 10.1021/bk-2015-1199.ch005.

[31] E. Acar and B. Yener, "Unsupervised Multiway Data Analysis: A Literature Survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 1, pp. 6–20, 2009, ISSN: 10414347. DOI: 10.1109/TKDE.2008.112.

[32] R. Bro, J. J. Workman JR, P. R. Mobley, and B. R. Kowalski, "Review of chemometrics applied to spectroscopy: 1985-95, part 3: Multi-way analysis," *Applied Spectroscopy Reviews*, vol. 32, no. 3, pp. 237–261, 1997.

[33] E. R. Malinowski, "Determination of the number of factors and the experimental error in a data matrix," *Analytical Chemistry*, vol. 49, no. 4, pp. 612–617, 1977, ISSN: 0003-2700. DOI: 10.1021/ac50012a027. [Online]. Available: https://doi.org/10.1021/ac50012a027.

[34] E. E. Papalexakis and C. Faloutsos, "Fast efficient and scalable core consistency diagnostic for the parafac decomposition for big sparse tensors," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5441–5445. DOI: 10.1109/ICASSP.2015.7179011.

[35] A. de Juan and R. Tauler, "Chemometrics applied to unravel multicomponent processes and mixtures: Revisiting latest trends in multivariate resolution," *Analytica Chimica Acta*, vol. 500, no. 1, pp. 195–210, 2003, ANALYTICAL HORIZONS - An International Symposium celebrating the publication of Volume 500 of Analytica Chimica Acta, ISSN: 0003-2670. DOI: https://doi.org/10.1016/S0003-2670(03)00724-4. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003267003007244.

[36] S. Vernuccio and L. J. Broadbelt, "Discerning complex reaction networks using automated generators," pp. 1–20, 2019. DOI: 10.1002/aic.16663.

[37] K. Molga, "Reaction Chemistry & Engineering machine-processable forms : a modern playground for physical-organic chemistry †," no. i, pp. 1506–1521, 2019. DOI: 10.1039/c9re00076c.

[38] S. Rangarajan, T. Kaminski, E. Van Wyk, A. Bhan, and P. Daoutidis, "Language-oriented rule-based reaction network generation and analysis: Algorithms of RING," *Computers and Chemical Engineering*, vol. 64, pp. 124–137, 2014, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2014.02.007.

[39] P. P. Plehiers, G. B. Marin, C. V. Stevens, and K. M. V. Geem, "Automated reaction database and reaction network analysis : extraction of reaction templates using cheminformatics," *Journal of Cheminformatics*, pp. 1–18, 2018, ISSN: 1758-2946. DOI: 10.1186/s13321-018-0269-8. [Online]. Available: https://doi.org/10.1186/s13321-018-0269-8.

[40] W. Jin, C. W. Coley, R. Barzilay, and T. Jaakkola, "Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network," no. Nips, pp. 1–10, 2017, ISSN: 0163-6804. DOI: 10.1093/mnras/stt2264. arXiv: 1709.04555. [Online]. Available: http://arxiv.org/abs/1709.04555.

[41] M. H. S. Segler and M. P. Waller, "Modelling Chemical Reasoning to Predict and Invent Reactions," pp. 6118–6128, 2017. DOI: 10.1002/chem.201604556.

[42] T. F. G. G. Cova and A. A. C. C. Pais, "Deep Learning for Deep Chemistry : Optimizing the Prediction of Chemical Patterns," vol. 7, no. November, pp. 1–22, 2019. DOI: 10.3389/fchem.2019.00809.

[43] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, and K. F. Jensen, "Prediction of Organic Reaction Outcomes Using Machine Learning," *ACS Central Science*, vol. 3, no. 5, pp. 434–443, 2017, ISSN: 23747951. DOI: 10.1021/acscentsci.7b00064.

[44] M. H. S. Segler and M. P. Waller, "Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction," pp. 5966–5971, 2017. DOI: 10.1002/chem.201605499.

[45] J. N. Wei and D. Duvenaud, "Neural Networks for the Prediction of Organic Chemistry Reactions," 2016. DOI: 10.1021/acscentsci.6b00219.

[46] C. W. Coley *et al.*, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical Science*, vol. 10, no. 2, pp. 370–377, 2019, ISSN: 20416539. DOI: 10.1039/c8sc04228d.

[47] C. A. Grambow, L. Pattanaik, and W. H. Green, "Deep Learning of Activation Energies," *Journal of Physical Chemistry Letters*, vol. 11, no. 8, pp. 2992–2997, 2020, ISSN: 19487185. DOI: 10.1021/acs.jpclett.0c00500.

[48] B. Liu *et al.*, "Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models," *ACS Central Science*, vol. 3, no. 10, pp. 1103–1113, 2017, ISSN: 23747951. DOI: 10.1021/acscentsci.7b00303. arXiv: 1706.01643.

[49] F. Feng, L. Lai, and J. Pei, "Computational chemical synthesis analysis and pathway design," *Frontiers in Chemistry*, vol. 6, no. JUN, 2018, ISSN: 22962646. DOI: 10.3389/fchem.2018.00199.

[50] D. Fooshee *et al.*, "Deep learning for chemical reaction prediction," *Molecular Systems Design and Engineering*, vol. 3, no. 3, pp. 442–452, 2018, ISSN: 20589689. DOI: 10.1039/c7me00107j.

[51] P. Loskot, K. Atitey, and L. Mihaylova, "Comprehensive review of models and methods for inferences in bio-chemical reaction networks," *Frontiers in Genetics*, vol. 10, no. JUN, 2019, ISSN: 16648021. DOI: 10.3389/fgene.2019.00549. arXiv: 1902.05828.

[52] K. Modeling, A. N. Network, and N.-f. Models, "Modeling of Thermal Cracking of Heavy Liquid Hydrocarbon : Application of," pp. 1536–1547, 2011. DOI: 10.1021/ie1015552.

[53] A. Papachristodoulou and B. Recht, "Determining interconnections in chemical reaction networks," *Proceedings of the American Control Conference*, pp. 4872–4877, 2007, ISSN: 07431619. DOI: 10.1109/ACC.2007.4283084.

[54] N. Bhatt, N. Kerimoglu, M. Amrhein, W. Marquardt, and D. Bonvin, "Incremental Model Identification for Reaction Systems - A Comparison of Rate-based and Extent-based Approaches," 2011.

[55] N. Bhatt, M. Amrhein, and D. Bonvin, "Incremental identification of reaction and mass-transfer kinetics using the concept of extents," *Industrial and Engineering Chemistry Research*, vol. 50, no. 23, pp. 12 960–12 974, 2011, ISSN: 08885885. DOI: 10.1021/ie2007196.

[56] D. P. Searson, M. J. Willis, S. J. Horne, and A. R. Wright, "Inference of chemical reaction networks using hybrid S-system models," *Chemical Product and Process Modeling*, vol. 2, no. 1, 2007, ISSN: 19342659. DOI: 10.2202/1934-2659.1029.

[57] D. P. Searson, M. J. Willis, and A. Wright, *Reverse Engineering Chemical Reaction Networks from Time Series Data*. 2012, vol. 2, pp. 327–348, ISBN: 9783527324347. DOI: 10.1002/9783527645121.ch12.

[58] S. C. Burnham, D. P. Searson, M. J. Willis, and A. R. Wright, "Inference of chemical reaction networks," vol. 63, pp. 862–873, 2008. DOI: 10.1016/j.ces.2007.10.010.

[59] J. Srividhya, E. J. Crampin, P. E. Mcsharry, and S. Schnell, "Reconstructing biochemical pathways from," pp. 828–838, 2007. DOI: 10.1002/pmic.200600428.

[60] G. Craciun and C. Pantea, "Identifiability of chemical reaction networks," *Journal of Mathematical Chemistry*, vol. 44, no. 1, pp. 244–259, 2008, ISSN: 02599791. DOI: 10.1007/s10910-007-9307-x.

[61] W. Ji and S. Deng, *Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network*, 2020. arXiv: 2002.09062 [q-bio.MN].

[62] A. Chakraborty, A. Sivaram, L. Samavedham, and V. Venkatasubramanian, "Mechanism discovery and model identification using genetic feature extraction and statistical testing," *Computers and Chemical Engineering*, vol. 140, 2020, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2020.106900.

[63] W. Zhang, S. Klus, T. Conrad, and C. Sch, "Learning Chemical Reaction Networks from Trajectory Data," vol. 18, no. 4, pp. 2000–2046, 2019.

[64] P.-M. Jacob and A. Lapkin, "Prediction of Chemical Reactions Using Statistical Models of Chemical Knowledge," Aug. 2018. DOI: 10.26434/chemrxiv.6954908.v1. [Online]. Available: https://chemrxiv.org/articles/preprint/Prediction_of_Chemical_Reactions_Using_Statistical_Models_of_Chemical_Knowledge/6954908.

[65] M. Amrhein, N. Bhatt, B. Srinivasan, and D. Bonvin, "Extents of reaction and flow for homogeneous reaction systems with inlet and outlet streams," *Aiche Journal*, vol. 56, pp. 2873–2886, 2010.

[66] N. Bhatt, M. Amrhein, and D. Bonvin, "Extents of reaction, mass transfer and flow for gas-liquid reaction systems," *Industrial and Engineering Chemistry Research*, vol. 49, no. 17, pp. 7704–7717, 2010, ISSN: 08885885. DOI: 10.1021/ie902015t.

[67] C. Filippi-Bossy, J. Bordet, J. Villermaux, S. Marchal-Brassely, and C. Georgakis, "Batch reactor optimization by use of tendency models," *Computers and Chemical Engineering*, vol. 13, no. 1-2, pp. 35–47, 1989, ISSN: 00981354. DOI: 10.1016/0098-1354(89)89005-2.

[68] C. Filippi *et al.*, "Tendency modeling of semibatch reactors for optimization and control," *Chemical Engineering Science*, vol. 41, no. 4, pp. 913–920, 1986, ISSN: 00092509. DOI: 10.1016/0009-2509(86)87175-5.

[69] D. Visser, R. Van Der Heijden, K. Mauch, M. Reuss, and S. Heijnen, "Tendency modeling: A new approach to obtain simplified kinetic models of metabolism applied to Saccharomyces cerevisiae," *Metabolic Engineering*, vol. 2, no. 3, pp. 252–275, 2000, ISSN: 10967176. DOI: 10.1006/mben.2000.0150.

[70] K. Yang *et al.*, "Predicting the young's modulus of silicate glasses using high-throughput molecular dynamics simulations and machine learning," *Scientific Reports*, vol. 9, no. 1, p. 8739, 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-45344-3. [Online]. Available: https://doi.org/10.1038/s41598-019-45344-3.

[71] F. Noé, A. Tkatchenko, K.-R. Müller, and C. Clementi, "Machine learning for molecular simulation," *Annual Review of Physical Chemistry*, vol. 71, no. 1, pp. 361–390, 2020, ISSN: 0066-426X. DOI: 10.1146/annurev-physchem-042018-052331. [Online]. Available: https://doi.org/10.1146/annurev-physchem-042018-052331.

[72] G. H. Gu, J. Noh, I. Kim, and Y. Jung, "Machine learning for renewable energy materials," *Journal of Materials Chemistry A*, vol. 7, no. 29, pp. 17 096–17 117, 2019, ISSN: 2050-7488. DOI: 10.1039/C9TA02356A. [Online]. Available: https://doi.org/10.1039/C9TA02356A.

[73] S. Xiao, R. Hu, Z. Li, S. Attarian, K.-M. Björk, and A. Lendasse, "A machine-learning-enhanced hierarchical multiscale method for bridging from molecular dynamics to continua," *Neural Computing and Applications*, vol. 32, no. 18, pp. 14 359–14 373, 2020, ISSN: 1433-3058. DOI: 10.1007/s00521-019-04480-7. [Online]. Available: https://doi.org/10.1007/s00521-019-04480-7.

[74] J. Gebhardt, M. Kiesel, S. Riniker, and N. Hansen, "Combining molecular dynamics and machine learning to predict self-solvation free energies and limiting activity coefficients," *Journal of Chemical Information and Modeling*, vol. 60, no. 11, pp. 5319–5330, 2020, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00479. [Online]. Available: https://doi.org/10.1021/acs.jcim.0c00479.

[75] A. S. Kelkar, B. C. Dallin, and R. C. Van Lehn, "Predicting hydrophobicity by learning spatiotemporal features of interfacial water structure: Combining molecular dynamics simulations with convolutional neural networks," *The Journal of Physical Chemistry B*, vol. 124, no. 41, pp. 9103–9114, 2020, ISSN: 1520-6106. DOI: 10.1021/acs.jpcb.0c05977. [Online]. Available: https://doi.org/10.1021/acs.jpcb.0c05977.

[76] W. F. D. Bennett *et al.*, "Predicting small molecule transfer free energies by combining molecular dynamics simulations and deep learning," *Journal of Chemical Information and Modeling*, vol. 60, no. 11, pp. 5375–5381, 2020, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.0c00318. [Online]. Available: https://doi.org/10.1021/acs.jcim.0c00318.

[77] S. Han, X. Li, L. Guo, H. Sun, M. Zheng, and W. Ge, "Refining fuel composition of rp-3 chemical surrogate models by reactive molecular dynamics and machine learning," *Energy & Fuels*, vol. 34, no. 9, pp. 11 381–11 394, 2020, ISSN: 0887-0624. DOI: 10.1021/acs.energyfuels.0c01491. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.0c01491.

[78] M. Gastegger, J. Behler, and P. Marquetand, "Machine learning molecular dynamics for the simulation of infrared spectra," *Chemical Science*, vol. 8, no. 10, pp. 6924–6935, 2017, ISSN: 2041-6520. DOI: 10.1039/C7SC02267K. [Online]. Available: https://doi.org/10.1039/C7SC02267K.

[79] F. Häse, I. F. Galván, A. Aspuru-Guzik, R. Lindh, and M. Vacher, "Machine learning for analysing ab initio molecular dynamics simulations," *Journal of Physics: Conference Series*, vol. 1412, p. 042 003, 2020, ISSN: 1742-6588. DOI: 10.1088/1742-6596/1412/4/042003. [Online]. Available: https://doi.org/10.1088/1742-6596/1412/4/042003.

[80] A. K. Chew, S. Jiang, W. Zhang, V. M. Zavala, and R. C. Van Lehn, "Fast predictions of liquid-phase acid-catalyzed reaction rates using molecular dynamics simulations and convolutional neural networks," *Chemical Science*, vol. 11, no. 46, pp. 12 464–12 476, 2020, ISSN: 2041-6520. DOI: 10.1039/D0SC03261A. [Online]. Available: https://doi.org/10.1039/D0SC03261A.

[81] D. Langary and Z. Nikoloski, "Inference of chemical reaction networks based on concentration profiles using an optimization framework," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 11, p. 113 121, 2019, ISSN: 1054-1500. DOI: 10.1063/1.5120598. [Online]. Available: https://doi.org/10.1063/1.5120598.

[82] T. I. Dearing, W. J. Thompson, C. E. Rechsteiner, and B. J. Marquardt, "Characterization of crude oil products using data fusion of process Raman, infrared, and nuclear magnetic resonance (NMR) spectra," *Applied Spectroscopy*, vol. 65, no. 2, pp. 181–186, 2011, ISSN: 00037028. DOI: 10.1366/10-05974. [Online]. Available: https://doi.org/10.1366/10-05974.

[83] L. Zhang and S. Zhang, "A unified joint matrix factorization framework for data integration," *arXiv:1707.08183*, 2017. eprint: arXiv:1707.08183.

[84] A. K. Smilde *et al.*, "Common and distinct components in data fusion," *Journal of Chemometrics*, vol. 31, no. 7, pp. 1–20, 2017, ISSN: 1099128X. DOI: 10.1002/cem.2900. arXiv: 1607.02328.

[85] A. Aspuru-Guzik, R. Lindh, and M. Reiher, "The matter simulation (r)evolution," *ACS Central Science*, vol. 4, no. 2, pp. 144–152, 2018, ISSN: 2374-7943. DOI: 10.1021/acscentsci.7b00550. [Online]. Available: https://doi.org/10.1021/acscentsci.7b00550.

[86] F. Babiloni *et al.*, "Mahalanobis distance-based classifiers are able to recognize eeg patterns by using few eeg electrodes," in *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, ser. 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol. 1, 2001, 651–654 vol.1. DOI: 10.1109/IEMBS.2001.1019019. [Online]. Available: https://doi.org/10.1109/IEMBS.2001.1019019.

[87] M. R. Gray, *Upgrading oilsands bitumen and heavy oil*. Edmonton, Canada: The University of Alberta Press, 2015.

[88] J. Belgrave, R. Moore, M. Ursenbach, and D. Bennion, "A Comprehensive Approach to In-Situ Combustion Modeling," *SPE Advanced Technology Series*, vol. 1, pp. 98–107, 1993.

[89] W. Shu and V. Venkatesan, "Kintetics Of Thermal Visbreaking Of A Cold Lake Bitumen," *Journal of Canadian Petroleum Technology*, vol. 23, no. 02, Mar. 1984, ISSN: 0021-9487. DOI: 10.2118/84-02-03. eprint: https://onepetro.org/JCPT/article-pdf/doi/10.2118/84-02-03/2168116/petsoc-84-02-03.pdf. [Online]. Available: https://doi.org/10.2118/84-02-03.

[90] M. L. Chacón-Patiño, S. M. Rowland, and R. P. Rodgers, "Advances in Asphaltene Petroleomics. Part 1: Asphaltenes Are Composed of Abundant Island and Archipelago Structural Motifs," *Energy and Fuels*, vol. 31, pp. 13 509–13 518, 2017.

[91] L. Fortuna, S. Graziani, A. Rizzo, and M. G. Xibilia, *Soft Sensors for Monitoring and Control of Industrial Processes*. 2007, ISBN: 1846284805. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

[92] H. R. Fischer and A. Cernescu, "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*, vol. 153, pp. 628–633, 2015.

[93] Y. Hou, L. Wang, D. Wang, M. Guo, P. Liu, and J. Yu, "Characterization of bitumen micro-mechanical behaviors using AFM, phase dynamics theory and MD simulation," *Materials*, vol. 10, pp. 1–16, 2017.

[94] B. Schuler, G. Meyer, D. Peña, O. C. Mullins, and L. Gross, "Unraveling the Molecular Structures of Asphaltenes by Atomic Force Microscopy," *Journal of the American Chemical Society*, vol. 137, pp. 9870–9876, 2015.

[95] J. Long, Z. Xu, and J. H. Masliyah, "Single Molecule Force Spectroscopy of Asphaltene Aggregates," *Langmuir*, vol. 23, pp. 6182–6190, 2007.

[96] A. J. Medford, M. R. Kunz, S. M. Ewing, T. Borders, and R. Fushimi, "Extracting Knowledge from Data through Catalysis Informatics," *ACS Catalysis*, vol. 8, pp. 7403–7429, 2018.

[97] N. Kumar, A. Bansal, G. S. Sarma, and R. K. Rawal, "Chemometrics tools used in analytical chemistry: An overview," *Talanta*, vol. 123, pp. 186–199, 2014.

[98] S. Yoon *et al.*, "Separation and characterization of bitumen from athabasca oil sand," *Korean Journal of Chemical Engineering*, vol. 26, no. 1, pp. 64–71, 2009, ISSN: 1975-7220. DOI: 10.1007/s11814-009-0011-3. [Online]. Available: https://doi.org/10.1007/s11814-009-0011-3.

[99] F. M. Adebiyi and V. Thoss, "Spectroscopic characterization of asphaltene fraction of Nigerian bitumen," *Petroleum Science and Technology*, vol. 33, pp. 245–255, 2015.

[100] J. C. Petersen, "An infrared study of hydrogen bonding in asphalt," *Fuel*, vol. 46, pp. 295–305, 1967.

[101] C. Varanda, I. Portugal, J. Ribeiro, C. M. Silva, and A. M. S. Silva, "NMR Spectroscopy in Bitumen Characterization," in *Analytical Characterization Methods for Crude Oiland Related Products*, Chichester, UK: John Wiley Sons Ltd, 2017, pp. 141–161.

[102] S. Niizuma, C. T. Steele, H. E. Gunning, and O. P. Strausz, "Electron spin resonance study of free radicals in Athabasca asphaltene," *Fuel*, vol. 56, pp. 249–256, 1977.

[103] T. F. Yen, J. G. Erdman, and A. J. Saraceno, "Investigation of the Nature of Free Radicals in Petroleum Asphaltenes and Related Substances by Electron Spin Resonance," *Analytical Chemistry*, vol. 34, pp. 694–700, 1962.

[104] J. H. Tannous and A. de Klerk, "Quantification of the Free Radical Content of Oilsands Bitumen Fractions," *Energy Fuels*, vol. 33, no. 8, pp. 7083–7093, 2019.

[105] G. Ding, Y. Hou, J. Peng, Y. Shen, M. Jiang, and G. Bai, "On-line near-infrared spectroscopy optimizing and monitoring biotransformation process of γ-aminobutyric acid," *Journal of Pharmaceutical Analysis*, vol. 6, pp. 171–178, 2016.

[106] M. Blanco and I. Villarroya, "NIR spectroscopy: A rapid-response analytical tool," *TrAC - Trends in Analytical Chemistry*, vol. 21, pp. 240–250, 2002.

[107] J. C. L. Alves, C. B. Henriques, and R. J. Poppi, "Determination of diesel quality parameters using support vector regression and near infrared spectroscopy for an in-line blending optimizer system," *Fuel*, vol. 97, pp. 710–717, 2012.

[108] E. Skibsted and S. B. Engelsen, "Spectroscopy for Process Analytical Technology (PAT)," in *Encyclopedia of Spectroscopy and Spectrometry*, Elsevier, 2010, pp. 2651–2661.

[109] M. Garrido, F. X. Rius, and M. S. Larrechi, "Multivariate curve resolution-alternating least squares (MCR-ALS) applied to spectroscopic data from monitoring chemical reactions processes," *Analytical and Bioanalytical Chemistry*, vol. 390, pp. 2059–2066, 2008.

[110] M. L. Selucky, Y. Chu, T. C. S. Ruo, and O. P. Strausz, "Chemical composition of Cold Lake bitumen," *Fuel*, vol. 57, pp. 9–16, 1978.

[111] K. Sivaramakrishnan, A. De Klerk, and V. Prasad, "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.

[112] P. Willet, *Similarity and Clustering in Chemical Information*. New York: Wiley, 1987.

[113] Y.-P. Wang, Y.-R. Zou, J.-T. Shi, and J. Shi, "Review of the chemometrics application in oil-oil and oil-source rock correlations," *J. Nat. Gas Geosci.*, vol. 3, pp. 217–232, 2018.

[114] R. Van de Vijver, B. R. Devocht, K. M. Van Geem, J. W. Thybaut, and G. B. Marin, "Challenges and opportunities for molecule-based management of chemical processes," *Current Opinion in Chemical Engineering*, vol. 13, pp. 142–149, 2016.

[115] J. C. Dellamorte, M. A. Barteau, and J. Lauterbach, "Opportunities for catalyst discovery and development: Integrating surface science and theory with high throughput methods," *Surface Science*, vol. 603, pp. 1770–1775, 2009.

[116] K. Sivaramakrishnan, A. Puliyanda, D. T. Tefera, A. Ganesh, S. Thirumalaivasan, and V. Prasad, "A Perspective on the Impact of Process Systems Engineering on Reaction Engineering," *Ind. Eng. Chem. Res.*, vol. 58, pp. 11 149–11 163, 2019.

[117] L. Wang, A. Zachariah, S. Yang, V. Prasad, and A. de Klerk, "Visbreaking Oilsands-Derived Bitumen in the Temperature Range of 340–400 °C," *Energy Fuels*, vol. 28, pp. 5014–5022, 2014.

[118] J. G. Speight, *The Chemistry and Technology of Petroleum*. New York: Marcel Dekker, 1991.

[119] F. E. Biasca, R. L. Dickenson, E. Chang, H. E. Johnson, R. T. Bailey, and D. R. Simbeck, "Future Technology In Heavy Oil Processing," in *Upgrading heavy crude oils and residue to transportation fuel: Technology, Economics, and Outlook*, Palo Alto, CA: SFA Pacific Inc., 2009.

[120] A. Tokarska, "Investigations on the processing of oil vacuum residue and its mixtures with coal and coal tars: Part 1. Primary conversion of crude materials," *Fuel*, vol. 75, pp. 1094–1100, 1996.

[121] M. L. Selucky, Y. Chu, T. Ruo, and O. P. Strausz, "Chemical composition of Athabasca bitumen," *Fuel*, vol. 56, pp. 369–381, 1977.

[122] A. M. McKenna, A. G. Marshall, and R. P. Rodgers, "Heavy Petroleum Composition. 4. Asphaltene Compositional Space," *Energy Fuels*, vol. 27, pp. 1257–1267, 2013.

[123] W. H. Press and S. A. Teukolsky, "Savitzky-Golay smoothing filters," *Comput. Phys.*, vol. 4, pp. 669–672, 1990.

[124] R. A. van den Berg, H. C. Hoefsloot, J. A. Westerhuis, A. K. Smilde, and M. J. van der Werf, "Centering, scaling, and transformations: improving the biological information content of metabolomics data," *BMC genomics*, vol. 7, p. 142, 2006.

[125] L. Eriksson, *Introduction to multi- and megavariate data analysis using projection methods (PCA PLS)*. Sweden: Umetrics, 1999.

[126] A. K. Smilde, M. J. van der Werf, S. Bijlsma, B. J. C. van der Werff-van der Vat, and R. H. Jellema, "Fusion of Mass Spectrometry-Based Metabolomics Data," *Analytical Chemistry*, vol. 77, pp. 6729–6736, 2005.

[127] H. C. Keun *et al.*, "Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling," *Analytica Chimica Acta*, vol. 490, pp. 265–276, 2003.

[128] R. P. Bro, "Tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, pp. 149–171, 1997.

[129] L. Duponchel, W. Elmi-Rayaleh, C. Ruckebusch, and J. P. Huvenne, "Multivariate Curve Resolution Methods in Imaging Spectroscopy: Influence of Extraction Methods and Instrumental Perturbations," *Journal of Chemical Information and Computer Sciences*, vol. 43, pp. 2057–2067, 2003.

[130] B. S. Everitt and P. M. Kroonenberg, *Three-Mode Principal Component Analysis: Theory and Applications*. Leiden, The Netherlands: DSWO Press, 1983.

[131] A. de Juan, J. Jaumot, and R. Tauler, "Multivariate Curve Resolution (MCR). Solving the mixture analysis problem," *Anal. Methods*, vol. 6, pp. 4964–4976, 2014.

[132] I. T. Jolliffe, *Principal Component Analysis*. Berlin, Heidelberg: Springer, 2011, pp. 1094–1096, ISBN: 0387954422. DOI: 10.2307/1270093. arXiv: arXiv:1011.1669v3.

[133] A. Elbergali, J. Nygren, and M. Kubista, "Automated procedure to predict the number of components in spectroscopic data," *Doktorsavhandlingar vid Chalmers Tekniska Hogskola*, vol. 379, pp. 143–158, 1999.

[134] E. R. Malinowski, "Window factor analysis: Theoretical derivation and application to flow injection analysis data," *Journal of Chemometrics*, vol. 6, pp. 29–40, 1992.

[135] R. Manne, H. Shen, and Y. Liang, "Subwindow factor analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 45, pp. 171–176, 1999.

[136] R. Y. Rubinstein and D. P. Kroese, *Simulation and the Monte Carlo method*. John Wiley Sons, 2016.

[137] Eberhart and Y. Shi, "Particle swarm optimization: Developments, applications and resources," in *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*, ser. Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), vol. 1, 2001, 81–86 vol. 1. DOI: 10.1109/CEC.2001.934374. [Online]. Available: https://doi.org/10.1109/CEC.2001.934374.

[138] H. Shinzawa, J.-H. Jiang, M. Iwahashi, and Y. Ozaki, "Robust Curve Fitting Method for Optical Spectra by Least Median Squares (LMedS) Estimator with Particle Swarm Optimization (PSO)," *Analytical Sciences*, vol. 23, pp. 781–785, 2007.

[139] A. de Juan, S. C. Rutan, and R. Tauler, "Two-Way Data Analysis: Multivariate Curve Resolution-Iterative Resolution Methods," in *Comprehensive Chemometrics*, Oxford: Elsevier, 2009, pp. 325–344.

[140] F. C. Sánchez, J. Toft, B. van den Bogaert, and D. L. Massart, "Orthogonal Projection Approach Applied to Peak Purity Assessment," *Analytical Chemistry*, vol. 68, pp. 79–85, 1996.

[141] F. Cuesta Sánchez, B. van den Bogaert, S. C. Rutan, and D. L. Massart, "Multivariate peak purity approaches," *Chemometrics and Intelligent Laboratory Systems*, vol. 34, no. 2, pp. 139–171, 1996, ISSN: 0169-7439. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0169743996000202.

[142] J. H. Jiang, S. Sasic, R.-Q. Yu, and Y. Ozaki, "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," *Journal of Chemometrics*, vol. 17, pp. 186–197, 2003.

[143] B. G. Vandeginste, W. Derks, and G. Kateman, "Multicomponent self-modelling curve resolution in high-performance liquid chromatography by iterative target transformation analysis," *Analytica Chimica Acta*, vol. 173, pp. 253–264, 1985.

[144] S. D. Brown, R. Tauler, and B. Walczak, *Comprehensive Chemometrics*. Oxford: Elsevier, 2009.

[145] J. De Leeuw, "Block-relaxation Algorithms in Statistics," in *Information Systems and Data Analysis*, Berlin: Springer, 1994, pp. 308–324.

[146] J. H. Jiang, Y. Liang, and Y. Ozaki, "Principles and methodologies in self-modeling curve resolution," *Chemometrics and Intelligent Laboratory Systems*, vol. 71, pp. 1–12, 2004.

[147] H. Shen, Y. Liang, O. M. Kvalheim, and R. Manne, "Determination of chemical rank of two-way data from mixtures using subspace comparisons," *Chemometrics and Intelligent Laboratory Systems*, vol. 51, pp. 49–59, 2000.

[148] S. L. Hao and L. M. Shao, "Determining the number of principal factors by eigenvector comparison of the original bi-linear data matrix and the one reconstructed from key variables," *Chemometrics and Intelligent Laboratory Systems*, vol. 149, pp. 17–23, 2015.

[149] E. R. Malinowski, *Factor Analysis in Chemistry*. Hoboken, NJ: Wiley, 2002.

[150] J. Mandel, "Use of the singular value decomposition in regression analysis," *The American Statistician*, vol. 36, pp. 15–24, 1982.

[151] M. Wasim and R. G. Brereton, "Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 72, pp. 133–151, 2004.

[152] W. Windig and J. Guilment, "Interactive self-modeling mixture analysis," *Analytical Chemistry*, vol. 63, pp. 1425–1432, 1991.

[153] E. R. Malinowski, "Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra," *Analytica Chimica Acta*, vol. 134, pp. 129–137, 1982.

[154] H. Shinzawa, J.-H. Jiang, M. Iwahashi, I. Noda, and Y. Ozaki, "Self-modeling curve resolution (SMCR) by particle swarm optimization (PSO)," *Analytica Chimica Acta*, vol. 595, pp. 275–281, 2007.

[155] J. J. Workman, P. R. Mobley, B. R. Kowalski, and R. Bro, "Review of Chemometrics Applied to Spectroscopy: 1985-95, Part I," *Applied Spectroscopy Reviews*, vol. 31, pp. 73–124, 1996.

[156] R. Tauler and M. Maeder, "Two-Way Data Analysis: Multivariate Curve Resolution – Error in Curve Resolution," in *Comprehensive Chemometrics*, Oxford: Elsevier, 2009, pp. 345–363.

[157] R. Tauler, A. Smilde, and B. Kowalski, "Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution," *Journal of Chemometrics*, vol. 9, pp. 31–58, 1995.

[158] E. Sanchez and B. R. Kowalski, "Generalized rank annihilation factor analysis," *Analytical Chemistry*, vol. 58, pp. 496–499, 1986.

[159] M. C. Antunes, J. E. J. Simão, A. C. Duarte, and R. Tauler, "Multivariate curve resolution of overlapping voltammetric peaks: quantitative analysis of binary and quaternary metal mixtures," *The Analyst*, vol. 127, pp. 809–817, 2002.

[160] R. H. Byrd, J. C. Gilbert, and J. Nocedal, "A trust region method based on interior point techniques for nonlinear programming," *Mathematical Programming*, vol. 89, pp. 149–185, 2000.

[161] J. Yen, J. Liao, Bogju Lee, and D. Randolph, "A hybrid approach to modeling metabolic systems using a genetic algorithm and simplex method," *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, vol. 28, pp. 173–191, 1998.

[162] M. Gen and R. Cheng, *Genetic Algorithms and Engineering Design*. New York: John Wiley Sons, 1997.

[163] J. C. Bansal, P. K. Singh, M. Saraswat, A. Verma, S. S. Jadon, and A. Abraham, "Inertia Weight strategies in Particle Swarm Optimization," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, Salamanca, Spain: IEEE, 2011, pp. 633–640.

[164] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *IEEE International of first Conference on Neural Networks*, Perth: IEEE, 1995, pp. 1942–1948.

[165] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, ser. 1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360), 1998, pp. 69–73. DOI: 10.1109/ICEC.1998.699146. [Online]. Available: https://doi.org/10.1109/ICEC.1998.699146.

[166] V. N. Vapnik, *The Nature of Statistical Learning Theory*. New York: Springer, 2000.

[167] A. B. Singer and P. I. Barton, "Global optimization with nonlinear ordinary differential equations," *J. Glob. Optim.*, vol. 34, pp. 159–190, 2006.

[168] K. Kumar, "Application of Akaike information criterion assisted probabilistic latent semantic analysis on non-trilinear total synchronous fluorescence spectroscopic data sets: automatizing fluorescence based multicomponent mixture analysis," *Anal. Chim. Acta*, vol. 1062, pp. 60–67, 2019.

[169] A. A. Neath and J. E. Cavanaugh, "The Bayesian information criterion: background, derivation, and applications," *WIREs Computational Statistics*, vol. 4, pp. 199–203, 2012, ISSN: 1939-5108. DOI: 10.1002/wics.199.

[170] O. P. Strausz and E. M. Lown, *The Chemistry of Alberta Oil Sands, Bitumens and Heavy Oils*. Calgary, AB: Alberta Energy Research Institute, 2003.

[171] P. R. Craddock, T. V. Le Doan, K. Bake, M. Polyakov, A. M. Charsky, and A. E. Pomerantz, "Evolution of Kerogen and Bitumen during Thermal Maturation via Semi-Open Pyrolysis Investigated by Infrared Spectroscopy," *Energy Fuels*, vol. 29, pp. 2197–2210, 2015, ISSN: 0887-0624. DOI: 10.1021/ef5027532.

[172] R. H. Potts, "Carboxylic acids (manufacture)," in *Kirk-Othmer Encyclopedia of Chemical Technology*, 3rd, New York: Wiley, 1978, pp. 835–845.

[173] D. C. Cronauer, D. M. Jewell, Y. T. Shah, and R. J. Modi, "Mechanism and Kinetics of Selected Hydrogen Transfer Reactions Typical of Coal Liquefaction," *Industrial Engineering Chemistry Fundamentals*, vol. 18, pp. 153–162, 1979, ISSN: 0196-4313. DOI: 10.1021/i160070a011.

[174] R. Billmers, L. L. Griffith, and S. E. Stein, "Hydrogen transfer between anthracene structures," *The Journal of Physical Chemistry*, vol. 90, pp. 517–523, 1986.

[175] I. W. C. E. Arends and P. Mulder, "Study of Hydrogen Shuttling Reactions in Anthracene/9,10-DihydroanthraceneNaphthyl-X Mixtures," *Energy Fuels*, vol. 10, pp. 235–242, 1996.

[176] L. M. Yañez Jaramillo and A. De Klerk, "Partial Upgrading of Bitumen by Thermal Conversion at 150-300 °C," *Energy Fuels*, vol. 32, pp. 3299–3311, 2018.

[177] D. L. Silverstein, R. M.; Webster, F. X.; Kiemle, D. J.; Bryce, *Spectrometric Identification of Organic Compounds.* New York: John Wiley Sons, 2014.

[178] S. J. Blanksby and G. B. Ellison, "Bond dissociation energies of organic molecules," *Accounts of Chemical Research*, vol. 36, no. 4, pp. 255–263, 2003, ISSN: 0001-4842. DOI: 10.1021/ar020230d. [Online]. Available: https://doi.org/10.1021/ar020230d.

[179] J. H. Tannous and A. de Klerk, "Methyl and Hydrogen Transfer in Free Radical Reactions," *Energy Fuels*, vol. 34, no. 2, pp. 1698–1709, 2020.

[180] K. N. Jha, D. S. Montgomery, and O. P. Strausz, "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*, O. P. Strausz and E. M. Lown, Eds., New York: Verlag Chemie, 1978, pp. 33–54.

[181] K. Sivaramakrishnan, "Application of chemometric and experimental tools for monitoring processes of industrial importance," PhD, University of Alberta, 2019.

[182] F. Khorasheh and M. R. Gray, "High-pressure thermal cracking of n-hexadecane in aromatic solvents," *Industrial Engineering Chemistry Research*, vol. 32, pp. 1864–1876, 1993.

[183] I. N. Evdokimov, "The Importance of Asphaltene Content in Petroleum II—Multipeak Viscosity Correlations," *Petroleum Science and Technology*, vol. 28, pp. 920–924, 2010.

[184] I. A. Wiehe, "A phase-separation kinetic model for coke formation," *Industrial Engineering Chemistry Research*, vol. 32, pp. 2447–2454, 1993.

[185] C. M. Blanchard and M. R. Gray, "Free radical chain reactions of bitumen residue," *ACS Division of Fuel Chemistry, Preprints*, vol. 42, pp. 137–141, 1997.

[186] A. Zachariah, L. Wang, S. Yang, V. Prasad, and A. De Klerk, "Suppression of coke formation during bitumen pyrolysis," *Energy and Fuels*, vol. 27, pp. 3061–3070, 2013.

[187] I. Noda, "Determination of Two-Dimensional Correlation Spectra Using the Hilbert Transform," *Applied Spectroscopy*, vol. 54, pp. 994–999, 2000.

[188] E. Skibsted and S. Engelsen, "Spectroscopy for process analytical technology (pat)," English, in *Encyclopedia of spectroscopy and spectrometry*, J. Lindon, G. Tranter, and D. Koppenaal, Eds., 2. United States: Academic Press, 2010, vol. 3, pp. 2651–2661, ISBN: 978-0-12-226680-5.

[189] M. A. Nemeth, "Multi- and megavariate data analysis," *Technometrics*, vol. 45, no. 4, pp. 362–362, 2003, ISSN: 0040-1706. DOI: 10.1198/tech.2003.s162. arXiv: arXiv:1011.1669v3. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1198/tech.2003.s162.

[190] T. Kourti, "Process analytical technology beyond real-time analyzers: The role of multivariate analysis," *Critical Reviews in Analytical Chemistry*, vol. 36, no. 3-4, pp. 257–278, 2006, ISSN: 10408347. DOI: 10.1080/10408340600969957.

[191] X. Fu, K. Huang, N. D. Sidiropoulos, and W. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 36, no. 2, pp. 59–80, 2019, ISSN: 1558-0792. DOI: 10.1109/MSP.2018.2877582.

[192] R. Tauler, B. Kowalski, and S. Fleming, "Multivariate curve resolution applied to spectral data from multiple runs of an industrial process," *Analytical Chemistry*, vol. 65, no. 15, pp. 2040–2047, 1993. DOI: 10.1021/ac00063a019. eprint: https://doi.org/10.1021/ac00063a019. [Online]. Available: https://doi.org/10.1021/ac00063a019.

[193] D. T. Tefera, L. M. Yañez Jaramillo, R. Ranjan, C. Li, A. De Klerk, and V. Prasad, "A Bayesian learning approach to modeling pseudoreaction networks for complex reacting systems: Application to the mild visbreaking of bitumen," *Industrial and Engineering Chemistry Research*, vol. 56, no. 8, pp. 1961–1970, 2017, ISSN: 15205045. DOI: 10.1021/acs.iecr.6b04437.

[194] W. Chen, L. T. Biegler, and S. G. Munoz, "An approach for simultaneous estimation of reaction kinetics and curve resolution from process and spectral data.," *JOURNAL OF CHEMOMETRICS*, vol. 30, no. 9, pp. 506 –522, n.d. ISSN: 08869383. [Online]. Available: https://login.ezproxy.library.ualberta.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=edswsc&AN=000383703000002&site=eds-live&scope=site.

[195] H. Abdollahi and R. Tauler, "Uniqueness and rotation ambiguities in multivariate curve resolution methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 108, no. 2, pp. 100–111, 2011, ISSN: 0169-7439. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169743911001110.

[196] L. Wang, A. Zachariah, S. Yang, V. Prasad, and A. de Klerk, "Visbreaking oilsands-derived bitumen in the temperature range of 340–400 °c," *Energy & Fuels*, vol. 28, no. 8, pp. 5014–5022, 2014. DOI: 10.1021/ef501128p. eprint: https://doi.org/10.1021/ef501128p. [Online]. Available: https://doi.org/10.1021/ef501128p.

[197] L. M. Yañez Jaramillo and A. de Klerk, "Partial upgrading of bitumen by thermal conversion at 150–300 °c," *Energy & Fuels*, vol. 32, no. 3, pp. 3299–3311, 2018. DOI: 10.1021/acs.energyfuels.7b04145. eprint: https://doi.org/10.1021/acs.energyfuels.7b04145. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.7b04145.

[198] R. Kannan *et al.*, "Deep data analysis via physically constrained linear unmixing: universal framework, domain examples, and a community-wide platform," *Advanced Structural and Chemical Imaging*, vol. 4, no. 1, 2018, ISSN: 21980926. DOI: 10.1186/s40679-018-0055-8. [Online]. Available: https://doi.org/10.1186/s40679-018-0055-8.

[199] J. Wang, F. Tian, W. Liu, X. Wang, W. Zhang, and K. Yamanishi, "Ranking preserving nonnegative matrix factorization," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, International Joint Conferences on Artificial Intelligence Organization, Jul. 2018, pp. 2776–2782. DOI: 10.24963/ijcai.2018/385. [Online]. Available: https://doi.org/10.24963/ijcai.2018/385.

[200] R. Du, D. Kuang, B. Drake, and H. Park, "Dc-nmf: Nonnegative matrix factorization based on divide-and-conquer for fast clustering and topic modeling," *Journal of Global Optimization*, vol. 68, no. 4, 777–798, Aug. 2017, ISSN: 0925-5001. DOI: 10.1007/s10898-017-0515-z. [Online]. Available: https://doi.org/10.1007/s10898-017-0515-z.

[201] C. Gobinet, E. Perrin, and R. Huez, "Application of non-negative matrix factorization to fluorescence spectroscopy," in *2004 12th European Signal Processing Conference*, 2004, pp. 1095–1098.

[202] H.-T. Gao, T.-H. Li, K. Chen, W.-G. Li, and X. Bi, "Overlapping spectra resolution using non-negative matrix factorization," *Talanta*, vol. 66, no. 1, pp. 65 –73, 2005, ISSN: 0039-9140. DOI: https://doi.org/10.1016/j.talanta.2004.09.017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0039914004005715.

[203] R. Luce, P. Hildebrandt, U. Kuhlmann, and J. Liesen, "Using separable nonnegative matrix factorization techniques for the analysis of time-resolved raman spectra," *Applied Spectroscopy*, vol. 70, no. 9, pp. 1464–1475, 2016. DOI: 10.1177/0003702816662600. eprint: https://doi.org/10.1177/0003702816662600. [Online]. Available: https://doi.org/10.1177/0003702816662600.

[204] S. Jia and Y. Qian, "A complexity constrained nonnegative matrix factorization for hyperspectral unmixing," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4666 LNCS, no. 1, pp. 268–276, 2007, ISSN: 03029743. DOI: 10.1007/978-3-540-74494-8_34.

[205] Y. Meng, R. Shang, L. Jiao, W. Zhang, and S. Yang, "Dual-graph regularized non-negative matrix factorization with sparse and orthogonal constraints," *Engineering Applications of Artificial Intelligence*, vol. 69, no. May 2017, pp. 24–35, 2018, ISSN: 09521976. DOI: 10.1016/j.engappai.2017.11.008. [Online]. Available: https://doi.org/10.1016/j.engappai.2017.11.008.

[206] X. Wang, T. Zhang, and X. Gao, "Multiview clustering based on non-negative matrix factorization and pairwise measurements," *IEEE Transactions on Cybernetics*, vol. 49, no. 9, pp. 3333–3346, 2019, ISSN: 21682267. DOI: 10.1109/TCYB.2018.2842052.

[207] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011, ISSN: 01628828. DOI: 10.1109/TPAMI.2010.231.

[208] M. Zitnik and B. Zupan, "Penalized matrix tri-factorization (DFMF)," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 1, pp. 41–53, 2015, ISSN: 0162-8828. DOI: 10.1109/TPAMI.2014.2343973. arXiv: 1307.0803. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6867358.

[209] B. Ray, W. Liu, and D. Fenyo, "Adaptive multiview nonnegative matrix factorization algorithm for integration of multimodal biomedical data," *Cancer Informatics*, vol. 16, 2017, ISSN: 11769351. DOI: 10.1177/1176935117725727.

[210] Y. Zheng, "Methodologies for cross-domain data fusion: An overview," *IEEE Transactions on Big Data*, vol. 1, no. 1, pp. 16–34, 2015, ISSN: 2332-7790. DOI: 10.1109/TBDATA.2015.2465959. [Online]. Available: http://ieeexplore.ieee.org/document/7230259/.

[211] J. Wang, F. Tian, H. Yu, C. H. Liu, K. Zhan, and X. Wang, "Diverse non-negative matrix factorization for multiview data representation," *IEEE Transactions on Cybernetics*, vol. 48, no. 9, pp. 2620–2632, 2018, ISSN: 21682267. DOI: 10.1109/TCYB.2017.2747400.

[212] Y. Kim and S. Choi, "Weighted nonnegative matrix factorization," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 1541–1544. DOI: 10.1109/ICASSP.2009.4959890.

[213] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using L_{21}-norm," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11, Glasgow, Scotland, UK: Association for Computing Machinery, 2011, 673–682, ISBN: 9781450307178. DOI: 10.1145/2063576.2063676. [Online]. Available: https://doi.org/10.1145/2063576.2063676.

[214] C. J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007, ISSN: 08997667. DOI: 10.1162/neco.2007.19.10.2756.

[215] C. J. Lin, "On the convergence of multiplicative update algorithms for nonnegative matrix factorization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1589–1596, 2007, ISSN: 10459227. DOI: 10.1109/TNN.2007.895831.

[216] S. Kritchman and B. Nadler, "Determining the number of components in a factor model from limited noisy data," *Chemometrics and Intelligent Laboratory Systems*, vol. 94, no. 1, pp. 19–32, 2008, ISSN: 01697439. DOI: 10.1016/j.chemolab.2008.06.002.

[217] A. Elbergali, J. Nygren, and M. Kubista, "Automated procedure to predict the number of components in spectroscopic data," *Doktorsavhandlingar vid Chalmers Tekniska Hogskola*, vol. 379, no. 1492, pp. 143–158, 1999, ISSN: 0346718X. DOI: 10.1016/S0003-2670(98)00640-0.

[218] M. Wasim and R. G. Brereton, "Determination of the number of significant components in liquid chromatography nuclear magnetic resonance spectroscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 72, no. 2, pp. 133–151, 2004, ISSN: 01697439. DOI: 10.1016/j.chemolab.2004.01.008.

[219] M. Meloun, J. Capek, P. Miksik, and R. Brereton, "Critical comparison of methods predicting the number of components in spectroscopic data," *Analytica Chimica Acta*, vol. 423, pp. 51–68, 2000.

[220] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, ser. NIPS'00, Denver, CO: MIT Press, 2000, 535–541.

[221] C. Boutsidis and E. Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008, ISSN: 00313203. DOI: 10.1016/j.patcog.2007.09.010.

[222] N. Sauwen *et al.*, "The successive projection algorithm as an initialization method for brain tumor segmentation using non-negative matrix factorization," *PLoS ONE*, vol. 12, no. 8, pp. 1–17, 2017, ISSN: 19326203. DOI: 10.1371/journal.pone.0180268.

[223] M Chu, F Diele, R Plemmons, and S Ragni, "Optimality, computation, and interpretations of nonnegative matrix factorizations," *Siam Journal on Matrix Analysis*, pp. 4–8030, 2004. [Online]. Available: http://scholar.google.com/scholar?q=intitle:OPTIMALITY,+COMPUTATION,+AND+INTERPRETATION+OF+NONNEGATIVE+MATRIX+FACTORIZATIONS+(VERSION:+October+18,+2004){\#}0.

[224] E. Y. Kang, I. Shpitser, C. Ye, and E. Eskin, "Detecting the presence and absence of causal relationships between expression of yeast genes with very few samples," in *Research in Computational Molecular Biology*, S. Batzoglou, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 466–481, ISBN: 978-3-642-02008-7.

[225] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000. DOI: 10.1089/106652700750050961. eprint: https://doi.org/10.1089/106652700750050961. [Online]. Available: https://doi.org/10.1089/106652700750050961.

[226] S. Triantafillou, V. Lagani, C. Heinze-Deml, A. Schmidt, J. Tegner, and I. Tsamardinos, "Predicting causal relationships from biological data: Applying automated causal discovery on mass cytometry data of human immune cells," *Scientific Reports*, vol. 7, no. 1, p. 12 724, 2017, ISSN: 2045-2322. DOI: 10.1038/s41598-017-08582-x. [Online]. Available: https://doi.org/10.1038/s41598-017-08582-x.

[227] D. K. Agrafiotis, D. Bandyopadhyay, J. K. Wegner, and H. Van Vlijmen, "Recent advances in chemoinformatics," *Journal of Chemical Information and Modeling*, vol. 47, no. 4, pp. 1279–1293, 2007, ISSN: 15499596. DOI: 10.1021/ci700059g.

[228] M. Młyńczak, *Data-driven causal path discovery without prior knowledge - a benchmark study*, 2018. arXiv: 1807.02348 [stat.AP].

[229] L. Uusitalo, "Advantages and challenges of bayesian networks in environmental modelling," *Ecological Modelling*, vol. 203, no. 3, pp. 312 –318, 2007, ISSN: 0304-3800. DOI: https://doi.org/10.1016/j.ecolmodel.2006.11.033. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0304380006006089.

[230] D. Heckerman, "A bayesian approach to learning causal networks," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'95, Montréal, Qué, Canada: Morgan Kaufmann Publishers Inc., 1995, 285–295, ISBN: 1558603859.

[231] D. Freedman and P. Humphreys, "Are there algorithms that discover causal structure?" *Synthese*, vol. 121, no. 1, pp. 29–54, 1999, ISSN: 1573-0964. DOI: 10.1023/A:1005277613752. [Online]. Available: https://doi.org/10.1023/A:1005277613752.

[232] S. L. Lauritzen, "The em algorithm for graphical association models with missing data," *Computational Statistics & Data Analysis*, vol. 19, no. 2, pp. 191–201, 1995, ISSN: 0167-9473. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0167947393E0056A.

[233] D. Heckerman, "Bayesian networks for data mining," *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 79–119, 1997, ISSN: 1573-756X. DOI: 10.1023/A:1009730122752. [Online]. Available: https://doi.org/10.1023/A:1009730122752.

[234] C. Berzan, *An exploration of structure learning in bayesian networks*, Tufts University Senior Honors Thesis, 2012.

[235] Z. Xu and S. N. Srihari, "Bayesian network structure learning using causality," in *2014 22nd International Conference on Pattern Recognition*, 2014, pp. 3546–3551. DOI: 10.1109/ICPR.2014.610.

[236] P. Leray and O. Francois, "Bnt structure learning package: Documentation and experiments," PSI, LITIS Laboratory, INSA de Rouen, Avenue de lUniversit-BP 8, 76801 Saint-tienne-du-Rouvray Cedex, Tech. Rep. 2004/PhLOF, 2004.

[237] B. Selman and C. P. Gomes, "Hill-climbing search," Nature Encyclopedia of Cognition, Nature Publ., 2002, pp. 333–336.

[238] X. Bai and R. Padman, "Tabu search enhanced markov blanket classifier for high dimensional data sets," in *The Next Wave in Computing, Optimization, and Decision Technologies*, B. Golden, S. Raghavan, and E. Wasil, Eds., Boston, MA: Springer US, 2005, pp. 337–354, ISBN: 978-0-387-23529-5.

[239] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Machine Learning*, vol. 65, no. 1, pp. 31–78, 2006, ISSN: 1573-0565. DOI: 10.1007/s10994-006-6889-7. [Online]. Available: https://doi.org/10.1007/s10994-006-6889-7.

[240] S. Mani, C. F. Aliferis, and A. Statnikov, "Bayesian algorithms for causal data mining," in *Proceedings of Workshop on Causality: Objectives and Assessment at NIPS 2008*, I. Guyon, D. Janzing, and B. Schölkopf, Eds., ser. Proceedings of Machine Learning Research, vol. 6, Whistler, Canada: PMLR, 2010, pp. 121–136. [Online]. Available: http://proceedings.mlr.press/v6/mani10a.html.

[241] J. H. Tannous and A. de Klerk, "Quantification of the free radical content of oilsands bitumen fractions," *Energy & Fuels*, vol. 33, no. 8, pp. 7083–7093, 2019. DOI: 10.1021/acs.energyfuels.9b01115. eprint: https://doi.org/10.1021/acs.energyfuels.9b01115. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.9b01115.

[242] N. Naghizada, G. H. C. Prado, and A. de Klerk, "Uncatalyzed hydrogen transfer during 100–250 $^{\circ}C$ conversion of asphaltenes," *Energy & Fuels*, vol. 31, no. 7, pp. 6800–6811, 2017. DOI: 10.1021/acs.energyfuels.7b00661. eprint: https://doi.org/10.1021/acs.energyfuels.7b00661. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.7b00661.

[243] E. M. Lown and O. P. Strausz, *The chemistry of Alberta oil sands, bitumens and heavy oils.* Alberta Energy Research Institute : Calgary, AB, Canada, 2003, ISBN: 978-0-7785-3096-1. [Online]. Available: https://login.ezproxy.library.ualberta.ca/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=geh&AN=2008-022984&site=ehost-live&scope=site.

[244] F. Cong, Q. H. Lin, L. D. Kuang, X. F. Gong, P. Astikainen, and T. Ristaniemi, "Tensor Decomposition of EEG Signals: A Brief Review," *Journal of Neuroscience Methods*, vol. 248, pp. 59–69, 2015, ISSN: 1872678X. DOI: 10.1016/j.jneumeth.2015.03.018. [Online]. Available: http://dx.doi.org/10.1016/j.jneumeth.2015.03.018.

[245] A. Cichocki *et al.*, "Tensor Decompositions for Signal Processing Applications: From Two-Way to Multiway Component Analysis," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 145–163, 2015, ISSN: 10535888. DOI: 10.1109/MSP.2013.2297439. arXiv: 1403.4462.

[246] J. Sun, S. Papadimitriou, C. Y. Lin, N. Cao, S. Liu, and W. Qian, "MultiVis: Content-Based Social Network Exploration Through Multi-Way Visual Analysis," *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, vol. 2, pp. 1057–1068, 2009. DOI: 10.1137/1.9781611972795.91.

[247] A. Zare, A. Ozdemir, M. A. Iwen, and S. Aviyente, "Extension of PCA to Higher Order Data Structures: An Introduction to Tensors, Tensor Decompositions, and Tensor PCA," *Proceedings of the IEEE*, vol. 106, no. 8, pp. 1341–1358, 2018, ISSN: 00189219. DOI: 10.1109/JPROC.2018.2848209. arXiv: 1803.00704.

[248] E. Acar, R. Bro, and A. K. Smilde, "Data fusion in metabolomics using coupled matrix and tensor factorizations," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1602–1620, 2015, ISSN: 00189219. DOI: 10.1109/JPROC.2015.2438719.

[249] L. De Lathauwer and E. Kofidis, "Coupled matrix-tensor factorizations — the case of partially shared factors," *2017 51st Asilomar Conference on Signals, Systems, and Computers*, pp. 711–715, 2017.

[250] A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, and V. Prasad, "Data fusion by joint non-negative matrix factorization for hypothesizing pseudochemistry using bayesian networks," *React. Chem. Eng.*, vol. 5, pp. 1719–1737, 9 2020. DOI: 10.1039/D0RE00147C. [Online]. Available: http://dx.doi.org/10.1039/D0RE00147C.

[251] J. D. Carroll and J. J. Chang, "Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of "Eckart-Young" Decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970, ISSN: 00333123. DOI: 10.1007/BF02310791.

[252] L. R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, vol. 31, pp. 279–311, 1966c.

[253] Q. Zhao, C. F. Caiafa, A. Cichocki, L. Zhang, and A. H. Phan, "Slice oriented tensor decomposition of eeg data for feature extraction in space, frequency and time domains," in *Neural Information Processing*, C. S. Leung, M. Lee, and J. H. Chan, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 221–228, ISBN: 978-3-642-10677-4.

[254] S. Rabanser, O. Shchur, and S. Günnemann, "Introduction to tensor decompositions and their applications in machine learning," *arXiv preprint arXiv:1711.10781*, 2017.

[255] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[256] M. Genicot, P. A. Absil, R. Lambiotte, and S. Sami, "Coupled Tensor Decomposition: A Step Towards Robust Components," *European Signal Processing Conference*, vol. 2016-Novem, pp. 1308–1312, 2016, ISSN: 22195491. DOI: 10.1109/EUSIPCO.2016.7760460.

[257] K. Takeuchi, R. Tomioka, K. Ishiguro, A. Kimura, and H. Sawada, "Non-Negative Multiple Tensor Factorization," *Proceedings - IEEE International Conference on Data Mining, ICDM*, pp. 1199–1204, 2013, ISSN: 15504786. DOI: 10.1109/ICDM.2013.83.

[258] E. Acar, D. M. Dunlavy, and T. G. Kolda, "Scalable Optimization Approach for Fitting Canonical Tensor Decompositions ,," vol. 25, no. February, pp. 67–86, 2011.

[259] R. Bro and H. Kiers, "A new efficient method for determining the number of components in parafac models," *Journal of Chemometrics*, vol. 17, pp. 274 –286, Jun. 2003. DOI: 10.1002/cem.801.

[260] K. Liu, H. So, J. P. C. L. da Costa, and L. Huang, "Core consistency diagnostic aided by reconstruction error for accurate enumeration of the number of components in parafac models," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6635–6639. DOI: 10.1109/ICASSP.2013.6638945.

[261] K. R. Murphy, C. A. Stedmon, D. Graeber, and R. Bro, "Fluorescence Spectroscopy and Multi-Way Techniques. PARAFAC," *Analytical Methods*, vol. 5, no. 23, pp. 6557–6566, 2013, ISSN: 17599660. DOI: 10.1039/c3ay41160e.

[262] W. S. DeSarbo, "An application of parafac to a small sample problem, demonstrating preprocessing, orthogonality constraints, and split-half diagnostic techniques," *Research Methods for Multimode Data Analysis p*, pp. 602–642, 1984.

[263] Q. Zhao, G. Zhou, T. Adali, L. Zhang, and A. Cichocki, "Kernelization of Tensor-Based Models for Multiway Data Analysis: Processing of Multidimensional Structured Data," *IEEE Signal Processing Magazine*, vol. 30, no. 4, pp. 137–148, 2013, ISSN: 10535888. DOI: 10.1109/MSP.2013.2255334.

[264] S. Zafeiriou and M. Petrou, "Nonnegative Tensor Factorization as an Alternative Csiszar-Tusnady Procedure: Algorithms, Convergence, Probabilistic Interpretations and Novel Probabilistic Tensor Latent Variable Analysis Algorithms," *Data Mining and Knowledge Discovery*, vol. 22, no. 3, pp. 419–466, 2011, ISSN: 13845810. DOI: 10.1007/s10618-010-0196-4.

[265] A. Cichocki, "Alternating least squares and related algorithms for nmf and sca problems," in *Nonnegative Matrix and Tensor Factorizations*, John Wiley Sons, Ltd, 2009, ch. 4, pp. 203–266, ISBN: 9780470747278. DOI: https://doi.org/10.1002/9780470747278.ch4. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470747278.ch4. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470747278.ch4.

[266] J. P. Royer, N. Thirion-Moreau, and P. Comon, "Computing the Polyadic Decomposition of Nonnegative Third Order Tensors," *Signal Processing*, vol. 91, no. 9, pp. 2159–2171, 2011, ISSN: 01651684. DOI: 10.1016/j.sigpro.2011.03.006.

[267] D. M. Dunlavy, T. G. Kolda, and E. Acar, "Poblano v1.0: A matlab toolbox for gradient-based optimization," Sandia National Laboratories, Tech. Rep. SAND2010-1422, 2010. DOI: 10.2172/989350. [Online]. Available: http://www.osti.gov/scitech/biblio/989350 (visited on 04/17/2014).

[268] E. C. Chi and T. G. Kolda, *Making tensor factorizations robust to non-gaussian noise*, 2010. arXiv: 1010.3043 [math.NA].

[269] H. Huang and C. Ding, "Robust Tensor Factorization Using R1 Norm," *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 1, pp. 1–8, 2008. DOI: 10.1109/CVPR.2008.4587392.

[270] S. A. Vorobyov, Y. Rong, N. D. Sidiropoulos, and A. B. Gershman, "Robust iterative fitting of multilinear models," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2678–2689, 2005, ISSN: 19410476. DOI: 10.1109/TSP.2005.850343.

[271] A. H. Phan and A. Cichocki, "Fast Nonnegative Tensor Factorization for Very Large-Scale Problems Using Two-Stage Procedure," *CAMSAP 2009 - 2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pp. 297–300, 2009. DOI: 10.1109/CAMSAP.2009.5413274.

[272] A. H. Phan and A. Cichocki, "Block Decomposition for Very Large-Scale Nonnegative Tensor Factorization," *CAMSAP 2009 - 2009 3rd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, pp. 316–319, 2009. DOI: 10.1109/CAMSAP.2009.5413268.

[273] K. Sivaramakrishnan, A. Puliyanda, A. de Klerk, and V. Prasad, "A data-driven approach to generate pseudo-reaction sequences for the thermal conversion of athabasca bitumen," *Reaction Chemistry & Engineering*, vol. 6, no. 3, pp. 505–537, 2021. DOI: 10.1039/D0RE00321B. [Online]. Available: https://doi.org/10.1039/D0RE00321B.

[274] E. Fumoto, S. Sato, and T. Takanohashi, "Determination of carbonyl functional groups in heavy oil using infrared spectroscopy," *Energy & Fuels*, vol. 34, no. 5, pp. 5231–5235, 2020, ISSN: 0887-0624. DOI: 10.1021/acs.energyfuels.9b02703. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.9b02703.

[275] Y. Rao and A. de Klerk, "Characterization of heteroatom-containing compounds in thermally cracked naphtha from oilsands bitumen," *Energy & Fuels*, vol. 31, no. 9, pp. 9247–9254, 2017, ISSN: 0887-0624. DOI: 10.1021/acs.energyfuels.7b01646. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.7b01646.

[276] D. G. Lee, Y. Yan, and W. D. Chandler, "Measurement of equilibrium constants for the formation of esters from aliphatic carboxylic acids and alcohols," *Analytical Chemistry*, vol. 66, no. 1, pp. 32–34, 1994, ISSN: 0003-2700. DOI: 10.1021/ac00073a007. [Online]. Available: https://doi.org/10.1021/ac00073a007.

[277] C. Rüchardt, M. Gerst, and M. Nölke, "The uncatalyzed transfer hydrogenation of α-methylstyrene by dihydroanthracene or xanthene—a radical reaction," *Angewandte Chemie International Edition in English*, vol. 31, no. 11, pp. 1523–1525, 1992, ISSN: 0570-0833. DOI: 10.1002/anie.199215231. [Online]. Available: https://doi.org/10.1002/anie.199215231.

[278] N. Naghizada, G. H. C. Prado, and A. de Klerk, "Uncatalyzed hydrogen transfer during 100–250 °c conversion of asphaltenes," *Energy & Fuels*, vol. 31, no. 7, pp. 6800–6811, 2017, ISSN: 0887-0624. DOI: 10.1021/acs.energyfuels.7b00661. [Online]. Available: https://doi.org/10.1021/acs.energyfuels.7b00661.

[279] M. R. Gray and W. C. McCaffrey, "Role of chain reactions and olefin formation in cracking, hydroconversion, and coking of petroleum and bitumen fractions," *Energy & Fuels*, vol. 16, no. 3, pp. 756–766, 2002, ISSN: 0887-0624. DOI: 10.1021/ef010243s. [Online]. Available: https://doi.org/10.1021/ef010243s.

[280] A. F. Parsons, *An Introduction to Free Radical Chemistry*. Blackwell Science, 2000.

[281] S. Patai, *The Chemistry of Acid Derivatives, Supplement B, Part 2*, ser. Patai's Chemistry of Functional Groups. Wiley, 1979, pp. 859–914, ISBN: 9780471996118. [Online]. Available: https://books.google.ca/books?id=cFuTzgEACAAJ.

[282] A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, and V. Prasad, "Structure-preserving joint non-negative tensor factorization to identify reaction pathways using bayesian networks," *Journal of Chemical Information and Modeling*, vol. 61, no. 12, pp. 5747–5762, 2021, ISSN: 1549-9596. DOI: 10.1021/acs.jcim.1c00789. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c00789.

[283] V. Venkatasubramanian, "Process Fault Detection and Diagnosis: Past, Present and Future," *IFAC Proceedings Volumes*, vol. 34, no. 27, pp. 1–13, 2001, ISSN: 14746670. DOI: 10.1016/s1474-6670(17)33563-2. [Online]. Available: http://dx.doi.org/10.1016/S1474-6670(17)33563-2.

[284] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis part I: Quantitative model-based methods," *Computers and Chemical Engineering*, vol. 27, no. 3, pp. 293–311, 2003, ISSN: 00981354. DOI: 10.1016/S0098-1354(02)00160-6.

[285] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 313–326, 2003, ISSN: 00981354. DOI: 10.1016/s0098-1354(02)00161-8.

[286] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, "A review of process fault detection and diagnosis: Part iii: Process history based methods," *Computers & Chemical Engineering*, vol. 27, no. 3, pp. 327 –346, 2003, ISSN: 0098-1354. DOI: https://doi.org/10.1016/S0098-1354(02)00162-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S009813540200162X.

[287] T. M. Floyd, M. A. Schmidt, and K. F. Jensen, "Silicon micromixers with infrared detection for studies of liquid-phase reactions," *Industrial & Engineering Chemistry Research*, vol. 44, no. 8, pp. 2351–2358, 2005, ISSN: 0888-5885. DOI: 10.1021/ie049348j. [Online]. Available: https://doi.org/10.1021/ie049348j.

[288] J. S. Moore and K. F. Jensen, "Automated multitrajectory method for reaction optimization in a microfluidic system using online ir analysis," *Organic Process Research & Development*, vol. 16, no. 8, pp. 1409–1415, 2012, ISSN: 1083-6160. DOI: 10.1021/op300099x. [Online]. Available: https://doi.org/10.1021/op300099x.

[289] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, and A. A. Lapkin, "Machine learning meets continuous flow chemistry: Automated optimization towards the pareto front of multiple objectives," *Chemical Engineering Journal*, vol. 352, pp. 277–282, 2018, ISSN: 1385-8947. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1385894718312634.

[290] K.-U. Klatt and W. Marquardt, "Perspectives for process systems engineering—personal views from academia and industry," *Computers & Chemical Engineering*, vol. 33, no. 3, pp. 536–550, 2009, ISSN: 0098-1354. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0098135408001737.

[291] R. Zimmerleiter *et al.*, "Probeless non-invasive near-infrared spectroscopic bioprocess monitoring using microspectrometer technology," *Analytical and Bioanalytical Chemistry*, vol. 412, no. 9, pp. 2103–2109, 2020, ISSN: 1618-2650. DOI: 10.1007/s00216-019-02227-w. [Online]. Available: https://doi.org/10.1007/s00216-019-02227-w.

[292] M. Rößler, P. U. Huth, and M. A. Liauw, "Process analytical technology (pat) as a versatile tool for real-time monitoring and kinetic evaluation of photocatalytic reactions," *Reaction Chemistry & Engineering*, vol. 5, no. 10, pp. 1992–2002, 2020. DOI: 10.1039/D0RE00256A. [Online]. Available: https://doi.org/10.1039/D0RE00256A.

[293] A. Golabgir and C. Herwig, "Combining mechanistic modeling and raman spectroscopy for real-time monitoring of fed-batch penicillin production," *Chemie Ingenieur Technik*, vol. 88, no. 6, pp. 764–776, 2016, ISSN: 0009-286X. DOI: 10.1002/cite.201500101. [Online]. Available: https://doi.org/10.1002/cite.201500101.

[294] J. N. Pauk *et al.*, "Advances in monitoring and control of refolding kinetics combining pat and modeling," eng, *Applied microbiology and biotechnology*, vol. 105, no. 6, pp. 2243–2260, 2021, 33598720[pmid], ISSN: 1432-0614. DOI: 10.1007/s00253-021-11151-y. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/33598720.

[295] J. Meyer-Kirschner *et al.*, "In-line monitoring of monomer and polymer content during microgel synthesis using precipitation polymerization via raman spectroscopy and indirect hard modeling," *Applied Spectroscopy*, vol. 70, no. 3,

pp. 416–426, 2016, ISSN: 0003-7028. DOI: 10.1177/0003702815626663. [Online]. Available: https://doi.org/10.1177/0003702815626663.

[296] N. Kail, W. Marquardt, and H. Briesen, "Process analysis by means of focused beam reflectance measurements," *Industrial & Engineering Chemistry Research*, vol. 48, no. 6, pp. 2936–2946, 2009, ISSN: 0888-5885. DOI: 10.1021/ie800839s. [Online]. Available: https://doi.org/10.1021/ie800839s.

[297] C. Houben, G. Nurumbetov, D. Haddleton, and A. A. Lapkin, "Feasibility of the simultaneous determination of monomer concentrations and particle size in emulsion polymerization using in situ raman spectroscopy," *Industrial & Engineering Chemistry Research*, vol. 54, no. 51, pp. 12 867–12 876, 2015, ISSN: 0888-5885. DOI: 10.1021/acs.iecr.5b02759. [Online]. Available: https://doi.org/10.1021/acs.iecr.5b02759.

[298] S. J. Qin, "Statistical process monitoring: Basics and beyond," *Journal of Chemometrics*, vol. 17, no. 8-9, pp. 480–502, 2003, ISSN: 08869383. DOI: 10.1002/cem.800.

[299] S. J. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annual Reviews in Control*, vol. 36, no. 2, pp. 220–234, 2012, ISSN: 13675788. DOI: 10.1016/j.arcontrol.2012.09.004. [Online]. Available: http://dx.doi.org/10.1016/j.arcontrol.2012.09.004.

[300] J. MacGregor and A. Cinar, "Monitoring, fault diagnosis, fault-tolerant control and optimization: Data driven methods," *Computers and Chemical Engineering*, vol. 47, pp. 111–120, 2012, ISSN: 00981354. DOI: 10.1016/j.compchemeng.2012.06.017. [Online]. Available: http://dx.doi.org/10.1016/j.compchemeng.2012.06.017.

[301] Z. Ge, Z. Song, and F. Gao, "Review of recent research on data-based process monitoring," *Industrial and Engineering Chemistry Research*, vol. 52, no. 10, pp. 3543–3562, 2013, ISSN: 08885885. DOI: 10.1021/ie302069q.

[302] B. Pretzner, C. Taylor, F. Dorozinski, M. Dekner, A. Liebminger, and C. Herwig, *Multivariate monitoring workflow for formulation, fill and finish processes*, 2020. DOI: 10.3390/bioengineering7020050. [Online]. Available: https://doi.org/10.3390/bioengineering7020050.

[303] M. S. Afzal and A. W. Al-Dabbagh, "Forecasting in Industrial Process Control: A Hidden Markov Model Approach," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 14 770–14 775, 2017, ISSN: 24058963. DOI: 10.1016/j.ifacol.2017.08.2591. [Online]. Available: https://doi.org/10.1016/j.ifacol.2017.08.2591.

[304] H. Alshraideh and G. Runger, "Process monitoring using hidden markov models," *Quality and Reliability Engineering International*, vol. 30, no. 8, pp. 1379–1387, 2014, ISSN: 10991638. DOI: 10.1002/qre.1560.

[305] M. Quiñones-Grueiro, A. Prieto-Moreno, C. Verde, and O. Llanes-Santiago, "Data-driven monitoring of multimode continuous processes: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 189, no. April, pp. 56–71, 2019, ISSN: 18733239. DOI: 10.1016/j.chemolab.2019.03.012.

[306] X. Wang, X. Wang, Z. Wang, and F. Qian, "A novel method for detecting processes with multi-state modes," *Control Engineering Practice*, vol. 21, no. 12, pp. 1788–1794, 2013, ISSN: 09670661. DOI: 10.1016/j.conengprac.2013.08.016. [Online]. Available: http://dx.doi.org/10.1016/j.conengprac.2013.08.016.

[307] S. Chen, Q. Jiang, and X. Yan, "Multimodal process monitoring based on transition-constrained Gaussian mixture model," *Chinese Journal of Chemical Engineering*, no. xxxx, 2020, ISSN: 10049541. DOI: 10.1016/j.cjche.2020.08.021. [Online]. Available: https://doi.org/10.1016/j.cjche.2020.08.021.

[308] C. Zhao, J. Chen, and H. Jing, "Condition-driven data analytics and monitoring for wide-range nonstationary and transient continuous processes," *IEEE Transactions on Automation Science and Engineering*, pp. 1–12, 2020. DOI: 10.1109/TASE.2020.3010536.

[309] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989. DOI: 10.1109/5.18626.

[310] P. Giudici and I. Abu Hashish, "A hidden Markov model to detect regime changes in cryptoasset markets," *Quality and Reliability Engineering International*, vol. 36, no. 6, pp. 2057–2065, 2020, ISSN: 10991638. DOI: 10.1002/qre.2673.

[311] M. E. Cholette and D. Djurdjanovic, "Degradation modeling and monitoring of machines using operation-specific hidden Markov models," *IIE Transactions (Institute of Industrial Engineers)*, vol. 46, no. 10, pp. 1107–1123, 2014, ISSN: 15458830. DOI: 10.1080/0740817X.2014.905734.

[312] D. Psimadas, P. Georgoulias, V. Valotassiou, and G. Loudos, "Molecular Nanomedicine Towards Cancer :" *Journal of pharmaceutical sciences*, vol. 101, no. 7, pp. 2271–2280, 2012. DOI: 10.1002/jps.

[313] F. Eskandanian and B. Mobasher, "Detecting changes in user preferences using hidden markov models for sequential recommendation tasks," *arXiv*, no. October, 2018. arXiv: 1810.00272.

[314] M. S. Afzal, W. Tan, and T. Chen, "Process monitoring for multimodal processes with mode-reachability constraints," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 5, pp. 4325–4335, 2017, ISSN: 02780046. DOI: 10.1109/TIE.2017.2677351.

[315] D. Wu, M. Chen, and D. Zhou, "Multimode process monitoring with mode transition constraints," *Proceedings of 2019 11th CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes, SAFEPROCESS 2019*, pp. 513–518, 2019. DOI: 10.1109/SAFEPROCESS45799.2019.9213368.

[316] L. Wang, C. Yang, and Y. Sun, "Multimode Process Monitoring Approach Based on Moving Window Hidden Markov Model," *Industrial and Engineering Chemistry Research*, vol. 57, no. 1, pp. 292–301, 2018, ISSN: 15205045. DOI: 10.1021/acs.iecr.7b03600.

[317] C. Ning, M. Chen, and D. Zhou, "Hidden Markov model-based statistics pattern analysis for multimode process monitoring: An index-switching scheme," *Industrial and Engineering Chemistry Research*, vol. 53, no. 27, pp. 11 084–11 095, 2014, ISSN: 15205045. DOI: 10.1021/ie5002394.

[318] P. Peng, J. Zhao, Y. Zhang, and H. Zhang, "Hidden markov model combined with kernel principal component analysis for nonlinear multimode process fault detection," *IEEE International Conference on Automation Science and Engineering*, vol. 2019-Augus, pp. 1586–1591, 2019, ISSN: 21618089. DOI: 10.1109/COASE.2019.8843205.

[319] M. M. Rashid and J. Yu, "Hidden markov model based adaptive independent component analysis approach for complex chemical process monitoring and fault detection," *Industrial and Engineering Chemistry Research*, vol. 51, no. 15, pp. 5506–5514, 2012, ISSN: 08885885. DOI: 10.1021/ie300203u.

[320] M. Galagedarage Don and F. Khan, "Dynamic process fault detection and diagnosis based on a combined approach of hidden Markov and Bayesian network model," *Chemical Engineering Science*, vol. 201, pp. 82–96, 2019, ISSN: 00092509. DOI: 10.1016/j.ces.2019.01.060. [Online]. Available: https://doi.org/10.1016/j.ces.2019.01.060.

[321] L. M. Owsley, L. E. Atlas, and G. D. Bernard, "Self-organizing feature maps and hidden Markov models for machine-tool monitoring," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2787–2798, 1997, ISSN: 1053587X. DOI: 10.1109/78.650105.

[322] L. Wang, C. Yang, Y. Sun, H. Zhang, and M. Li, "Effective variable selection and moving window HMM-based approach for iron-making process monitoring," *Journal of Process Control*, vol. 68, pp. 86–95, 2018, ISSN: 09591524. DOI: 10.1016/j.jprocont.2018.04.008. [Online]. Available: https://doi.org/10.1016/j.jprocont.2018.04.008.

[323] F. Wang, S. Tan, and H. Shi, "Hidden Markov model-based approach for multimode process monitoring," *Chemometrics and Intelligent Laboratory Systems*, vol. 148, pp. 51–59, 2015, ISSN: 18733239. DOI: 10.1016/j.chemolab.2015.08.025. [Online]. Available: http://dx.doi.org/10.1016/j.chemolab.2015.08.025.

[324] J. Yu, "Hidden Markov models combining local and global information for nonlinear and multimodal process monitoring," *Journal of Process Control*, vol. 20, no. 3, pp. 344–359, 2010, ISSN: 09591524. DOI: 10.1016/j.jprocont.2009.12.002. [Online]. Available: http://dx.doi.org/10.1016/j.jprocont.2009.12.002.

[325] F. Cartella, J. Lemeire, L. Dimiccoli, and H. Sahli, "Hidden semi-markov models for predictive maintenance," *Mathematical Problems in Engineering*, vol. 2015, 2015, ISSN: 15635147. DOI: 10.1155/2015/278120.

[326] A. Sabanovic, "SMC Framework in Motion Montrol Systems," *International Journal of Adaptive Control and Signal Processing*, vol. 21, no. February, pp. 731–744, 2007. DOI: 10.1002/acs.

[327] N. Sammaknejad, B. Huang, W. Xiong, A. Fatehi, F. Xu, and A. Espejo, "Operating condition diagnosis based on hmm with adaptive transition probabilities in presence of missing observations," *AIChE Journal*, vol. 61, no. 2, pp. 477–493, 2015. DOI: https://doi.org/10.1002/aic.14661. eprint: https://aiche.onlinelibrary.wiley.com/doi/pdf/10.1002/aic.14661. [Online]. Available: https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.14661.

[328] S. Z. Yu, "Hidden semi-Markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215–243, 2010, ISSN: 00043702. DOI: 10.1016/j.artint.2009.11.011. [Online]. Available: http://dx.doi.org/10.1016/j.artint.2009.11.011.

[329] N. Sammaknejad, Y. Zhao, and B. Huang, "A review of the Expectation Maximization algorithm in data-driven process identification," *Journal of Process Control*, vol. 73, pp. 123–136, 2019, ISSN: 09591524. DOI: 10.1016/j.jprocont.2018.12.010. [Online]. Available: https://doi.org/10.1016/j.jprocont.2018.12.010.

[330] Z. Lou and Y. Wang, "Multimode Continuous Processes Monitoring Based on Hidden Semi-Markov Model and Principal Component Analysis," *Industrial and Engineering Chemistry Research*, vol. 56, no. 46, pp. 13 800–13 811, 2017, ISSN: 15205045. DOI: 10.1021/acs.iecr.7b01721.

[331] B. G. Keller, A. Kobitski, A. Jäschke, G. U. Nienhaus, and F. Noé, "Complex RNA folding kinetics revealed by single-molecule FRET and hidden Markov models," *Journal of the American Chemical Society*, vol. 136, no. 12, pp. 4534–4543, 2014, ISSN: 15205126. DOI: 10.1021/ja4098719.

[332] Z. Li, J. Guo, W. Jiao, P. Xu, B. Liu, and X. Zhao, "Random linear interpolation data augmentation for person re-identification," *Multimedia Tools and Applications*, vol. 79, no. 7, pp. 4931–4947, 2020, ISSN: 1573-7721. DOI: 10.1007/s11042-018-7071-5. [Online]. Available: https://doi.org/10.1007/s11042-018-7071-5.

[333] S. Vaseghi, "State duration modelling in hidden markov models," *Signal Process.*, vol. 41, pp. 31–41, 1995.

[334] M. Russell and R. Moore, "Explicit modelling of state occupancy in hidden markov models for automatic speech recognition," in *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 10, 1985, pp. 5–8. DOI: 10.1109/ICASSP.1985.1168477.

[335] T. Liu, J. Chen, and G. Dong, "Singular spectrum analysis and continuous hidden markov model for rolling element bearing fault diagnosis," *Journal of Vibration and Control*, vol. 21, no. 8, pp. 1506–1521, 2015. DOI: 10.1177/1077546313496833. eprint: https://doi.org/10.1177/1077546313496833. [Online]. Available: https://doi.org/10.1177/1077546313496833.

[336] F. Tian, Q. Zhou, and C. Yang, "Gaussian mixture model-hidden markov model based nonlinear equalizer for optical fiber transmission," *Opt. Express*, vol. 28, no. 7, pp. 9728–9737, 2020. DOI: 10.1364/OE.386476. [Online]. Available: http://www.opticsexpress.org/abstract.cfm?URI=oe-28-7-9728.

[337] I. Visser, "Seven things to remember about hidden markov models: A tutorial on markovian models for time series," *Journal of Mathematical Psychology*, vol. 55, no. 6, pp. 403 –415, 2011, ISSN: 0022-2496. DOI: https://doi.org/10.1016/j.jmp.2011.08.002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022249611000691.

[338] Shun-Zheng Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden markov model," *IEEE Transactions on Signal Processing*, vol. 54, no. 5, pp. 1947–1951, 2006. DOI: 10.1109/TSP.2006.872540.

[339] Guorong Xuan, Wei Zhang, and Peiqi Chai, "Em algorithms of gaussian mixture model and hidden markov model," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, vol. 1, 2001, 145–148 vol.1. DOI: 10.1109/ICIP.2001.958974.

[340] A. Benouareth, A. Ennaji, and M. Sellami, "Semi-continuous hmms with explicit state duration for unconstrained arabic word modeling and recognition," *Pattern Recognition Letters*, vol. 29, no. 12, pp. 1742 –1752, 2008, ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2008.05.008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167865508001670.

[341] N. Nguyen, "Hidden Markov Model for Stock Trading," *International Journal of Financial Studies*, vol. 6, no. 2, p. 36, 2018, ISSN: 2227-7072. DOI: 10.3390/ijfs6020036.

[342] J. Pohle, R. Langrock, F. M. van Beest, and N. M. Schmidt, "Selecting the number of states in hidden markov models: Pragmatic solutions illustrated using animal movement," *Journal of Agricultural, Biological and Environmental Statistics*, vol. 22, no. 3, pp. 270–293, 2017, ISSN: 1537-2693. DOI: 10.1007/s13253-017-0283-8. [Online]. Available: https://doi.org/10.1007/s13253-017-0283-8.

[343] J. Bloit and X. Rodet, "Short-time viterbi for online hmm decoding: Evaluation on a real-time phone recognition task," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2121–2124. DOI: 10.1109/ICASSP.2008.4518061.

[344] J. Lember and A. A. Koloydenko, "Bridging viterbi and posterior decoding: A generalized risk approach to hidden path inference based on hidden markov models," *J. Mach. Learn. Res.*, vol. 15, no. 1, 1–58, Jan. 2014, ISSN: 1532-4435.

[345] K. Wang *et al.*, "Kinetic and data-driven reaction analysis for pharmaceutical process development," *Industrial & Engineering Chemistry Research*, vol. 59, no. 6, pp. 2409–2421, 2020, ISSN: 0888-5885. DOI: 10.1021/acs.iecr.9b03578. [Online]. Available: https://doi.org/10.1021/acs.iecr.9b03578.

[346] S. Stocker, G. Csányi, K. Reuter, and J. T. Margraf, "Machine learning in chemical reaction space," *Nature Communications*, vol. 11, no. 1, p. 5505, 2020, ISSN: 2041-1723. DOI: 10.1038/s41467-020-19267-x. [Online]. Available: https://doi.org/10.1038/s41467-020-19267-x.

[347] W. Ji and S. Deng, "Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network," *The Journal of Physical Chemistry A*, vol. 125, no. 4, pp. 1082–1092, 2021, ISSN: 1089-5639. DOI: 10.1021/acs.jpca.0c09316. [Online]. Available: https://doi.org/10.1021/acs.jpca.0c09316.

[348] M. Sedighi, K. Keyvanloo, and J. Towfighi, "Modeling of thermal cracking of heavy liquid hydrocarbon: Application of kinetic modeling, artificial neural network, and neuro-fuzzy models," *Industrial & Engineering Chemistry Research*, vol. 50, no. 3, pp. 1536–1547, 2011, ISSN: 0888-5885. DOI: 10.1021/ie1015552. [Online]. Available: https://doi.org/10.1021/ie1015552.

[349] G. Craciun and C. Pantea, "Identifiability of chemical reaction networks," *Journal of Mathematical Chemistry*, vol. 44, pp. 244–259, 2008.

[350] F. Santosa and B. Weitz, "An inverse problem in reaction kinetics," *Journal of Mathematical Chemistry*, vol. 49, no. 8, pp. 1507–1520, 2011, ISSN: 1572-8897. DOI: 10.1007/s10910-011-9835-2. [Online]. Available: https://doi.org/10.1007/s10910-011-9835-2.

[351] M. Hoffmann, C. Fröhner, and F. Noé, "Reactive sindy: Discovering governing reactions from concentration data," *The Journal of Chemical Physics*, vol. 150, no. 2, p. 025 101, 2019, ISSN: 0021-9606. DOI: 10.1063/1.5066099. [Online]. Available: https://doi.org/10.1063/1.5066099.

[352] D. F. Anderson, B. Joshi, and A. Deshpande, "On reaction network implementations of neural networks," *Journal of The Royal Society Interface*, vol. 18, no. 177, p. 20 210 031, XXXX. DOI: 10.1098/rsif.2021.0031. [Online]. Available: https://doi.org/10.1098/rsif.2021.0031.

[353] B. Sen and S. Menon, "Representation of chemical kinetics by artificial neural networks for large eddy simulations," in ser. Joint Propulsion Conferences. American Institute of Aeronautics and Astronautics, 2007, 0. DOI: 10.2514/6.2007-5635. [Online]. Available: https://doi.org/10.2514/6.2007-5635.

[354] N. V. Muravyev, G. Luciano, H. L. Ornaghi, R. Svoboda, and S. Vyazovkin, *Artificial neural networks for pyrolysis, thermal analysis, and thermokinetic studies: The status quo*, 2021. DOI: 10.3390/molecules26123727. [Online]. Available: https://doi.org/10.3390/molecules26123727.

[355] N. Shenvi, J. M. Geremia, and H. Rabitz, "Efficient chemical kinetic modeling through neural network maps," *The Journal of Chemical Physics*, vol. 120, no. 21, pp. 9942–9951, 2004, ISSN: 0021-9606. DOI: 10.1063/1.1718305. [Online]. Available: https://doi.org/10.1063/1.1718305.

[356] H.-J. Zander, R. Dittmeyer, and J. Wagenhuber, "Dynamic modeling of chemical reaction systems with neural networks and hybrid models," *Chemical Engineering & Technology*, vol. 22, no. 7, pp. 571–574, 1999, ISSN: 0930-7516. DOI: 10.1002/(SICI)1521-4125(199907)22:7⟨571::AID-CEAT571⟩3.0.CO;2-5. [Online]. Available: https://doi.org/10.1002/(SICI)1521-4125(199907)22:7⟨571::AID-CEAT571⟩3.0.CO;2-5.

[357] G. S. Gusmão, A. P. Retnanto, S. C. da Cunha, and A. J. Medford, *Kinetics-informed neural networks*, 2020. arXiv: 2011.14473 [cs.LG].

[358] W. Ji, W. Qiu, Z. Shi, S. Pan, and S. Deng, "Stiff-pinn: Physics-informed neural network for stiff chemical kinetics," *The Journal of Physical Chemistry A*, vol. 125, no. 36, pp. 8098–8106, 2021, ISSN: 1089-5639. DOI: 10.1021/acs.jpca.1c05102. [Online]. Available: https://doi.org/10.1021/acs.jpca.1c05102.

[359] I. M. Galván, J. M. Zaldívar, H. Hernández, and E. Molga, "The use of neural networks for fitting complex kinetic data," *Computers & Chemical Engineering*, vol. 20, no. 12, pp. 1451–1465, 1996, ISSN: 0098-1354. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0098135495002316.

[360] E. J. Molga, B. A. A. van Woezik, and K. R. Westerterp, "Neural networks for modelling of chemical reaction systems with complex kinetics: Oxidation of 2-octanol with nitric acid," *Chemical Engineering and Processing: Process Intensification*, vol. 39, no. 4, pp. 323–334, 2000, ISSN: 0255-2701. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0255270199000938.

[361] O. Owoyele and P. Pal, "Chemnode: A neural ordinary differential equations framework for efficient chemical kinetic solvers," *Energy and AI*, p. 100 118, 2021, ISSN: 2666-5468. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666546821000677.

[362] S. Kim, W. Ji, S. Deng, Y. Ma, and C. Rackauckas, "Stiff neural ordinary differential equations," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 9, p. 093 122, 2021, ISSN: 1054-1500. DOI: 10.1063/5.0060697. [Online]. Available: https://doi.org/10.1063/5.0060697.

[363] P. Kollenz, D.-P. Herten, and T. Buckup, "Unravelling the kinetic model of photochemical reactions via deep learning," *The Journal of Physical Chemistry B*, vol. 124, no. 29, pp. 6358–6368, 2020, ISSN: 1520-6106. DOI: 10.1021/acs.jpcb.0c04299. [Online]. Available: https://doi.org/10.1021/acs.jpcb.0c04299.

[364] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," *Advances in Neural Information Processing Systems*, 2018.

[365] R. T. Q. Chen, B. Amos, and M. Nickel, "Learning neural event functions for ordinary differential equations," *International Conference on Learning Representations*, 2021.

[366] A.Kramida, Yu.Ralchenko, J.Reader, and and NIST ASD Team, NIST Atomic Spectra Database (ver. 5.9), [Online]. Available: https://physics.nist.gov/asd [2017, April 9]. National Institute of Standards and Technology, Gaithersburg, MD. 2021.

[367] F. Bahrpeyma, M. Roantree, P. Cappellari, M. Scriney, and A. McCarren, "A methodology for validating diversity in synthetic time series generation," *MethodsX*, vol. 8, p. 101 459, 2021, ISSN: 2215-0161. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2215016121002521.

[368] C. J. Giunta, "What's in a name? amount of substance, chemical amount, and stoichiometric amount," *Journal of Chemical Education*, vol. 93, no. 4, pp. 583–586, 2016, ISSN: 0021-9584. DOI: 10.1021/acs.jchemed.5b00690. [Online]. Available: https://doi.org/10.1021/acs.jchemed.5b00690.

[369] R. E. Ferner and J. K. Aronson, "Cato guldberg and peter waage, the history of the law of mass action, and its relevance to clinical pharmacology," *British Journal of Clinical Pharmacology*, vol. 81, no. 1, pp. 52–55, 2016, ISSN: 0306-5251. DOI: 10.1111/bcp.12721. [Online]. Available: https://doi.org/10.1111/bcp.12721.

[370] N. Kriegeskorte and T. Golan, "Neural network models and deep learning," *Current Biology*, vol. 29, no. 7, R231–R236, 2019, ISSN: 0960-9822. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982219302040.

[371] G. S. Mittal, "Chapter 18 - artificial neural network (ann) based process modeling," in *Handbook of Farm, Dairy and Food Machinery Engineering (Second Edition)*. San Diego: Academic Press, 2013, pp. 467–473, ISBN: 978-0-12-385881-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123858818000185.

[372] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6391–6438, 2021, ISSN: 1573-7462. DOI: 10.1007/s10462-021-09975-1. [Online]. Available: https://doi.org/10.1007/s10462-021-09975-1.

[373] D. C. Psichogios and L. H. Ungar, "A hybrid neural network-first principles approach to process modeling," *AIChE Journal*, vol. 38, no. 10, pp. 1499–1511, 1992, ISSN: 0001-1541. DOI: 10.1002/aic.690381003. [Online]. Available: https://doi.org/10.1002/aic.690381003.

[374] F. Steyer and K. Sundmacher, "Cyclohexanol production via esterification of cyclohexene with formic acid and subsequent hydration of the esterreaction kinetics," *Industrial & Engineering Chemistry Research*, vol. 46, no. 4, pp. 1099–1104, 2007, ISSN: 0888-5885. DOI: 10.1021/ie060781y. [Online]. Available: https://doi.org/10.1021/ie060781y.

[375] J. Dalgaard, T. Kocka, and J. Pena, *On local optima in learning bayesian networks*, ISSN ; -, 2003.

[376] N. V. Orupattur, S. H. Mushrif, and V. Prasad, "Catalytic materials and chemistry development using a synergistic combination of machine learning and ab initio methods," *Computational Materials Science*, vol. 174, p. 109 474, 2020, ISSN: 0927-0256. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025619307736.

[377] M. Rupp *et al.*, "Machine learning estimates of natural product conformational energies," *PLOS Computational Biology*, vol. 10, no. 1, pp. 1–8, Jan. 2014. DOI: 10.1371/journal.pcbi.1003400. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003400.

[378] T. Morawietz and N. Artrith, "Machine learning-accelerated quantum mechanics-based atomistic simulations for industrial applications," *Journal of Computer-Aided Molecular Design*, vol. 35, no. 4, pp. 557–586, 2021, ISSN: 1573-4951. DOI: 10.1007/s10822-020-00346-6. [Online]. Available: https://doi.org/10.1007/s10822-020-00346-6.

[379] J. Xu, X.-M. Cao, and P. Hu, "Perspective on computational reaction prediction using machine learning methods in heterogeneous catalysis," *Physical Chemistry Chemical Physics*, vol. 23, no. 19, pp. 11 155–11 179, 2021, ISSN: 1463-9076. DOI: 10.1039/D1CP01349A. [Online]. Available: https://doi.org/10.1039/D1CP01349A.

[380] C. Schran, F. L. Thiemann, P. Rowe, E. A. Müller, O. Marsalek, and A. Michaelides, "Machine learning potentials for complex aqueous systems made simple," *Proceedings of the National Academy of Sciences*, vol. 118, no. 38, e2110077118, 2021. DOI: 10.1073/pnas.2110077118. [Online]. Available: http://www.pnas.org/content/118/38/e2110077118.abstract.

[381] W. Jia *et al.*, *Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning*, 2020. arXiv: 2005.00223 [physics.comp-ph].

[382] P. Pattnaik, S. Raghunathan, T. Kalluri, P. Bhimalapuram, C. V. Jawahar, and U. D. Priyakumar, "Machine learning for accurate force calculations in molecular dynamics simulations," *The Journal of Physical Chemistry A*, vol. 124, no. 34, pp. 6954–6967, 2020, ISSN: 1089-5639. DOI: 10.1021/acs.jpca.0c03926. [Online]. Available: https://doi.org/10.1021/acs.jpca.0c03926.

[383] W. Plazinski, A. Plazinska, and A. Brzyska, "Efficient sampling of high-energy states by machine learning force fields," *Phys. Chem. Chem. Phys.*, vol. 22, pp. 14 364–14 374, 25 2020. DOI: 10.1039/D0CP01399D. [Online]. Available: http://dx.doi.org/10.1039/D0CP01399D.

[384] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, "Towards exact molecular dynamics simulations with machine-learned force fields," *Nature Communications*, vol. 9, no. 1, p. 3887, 2018, ISSN: 2041-1723. DOI: 10.1038/s41467-018-06169-2. [Online]. Available: https://doi.org/10.1038/s41467-018-06169-2.

[385] L. Böselt, M. Thürlemann, and S. Riniker, "Machine learning in qm/mm molecular dynamics simulations of condensed-phase systems," *Journal of Chemical Theory and Computation*, vol. 17, no. 5, pp. 2641–2658, 2021, ISSN: 1549-9618. DOI: 10.1021/acs.jctc.0c01112. [Online]. Available: https://doi.org/10.1021/acs.jctc.0c01112.

[386] V. Botu and R. Ramprasad, "Adaptive machine learning framework to accelerate ab initio molecular dynamics," *International Journal of Quantum Chemistry*, vol. 115, no. 16, pp. 1074–1083, 2015, ISSN: 0020-7608. DOI: 10.1002/qua.24836. [Online]. Available: https://doi.org/10.1002/qua.24836.

[387] F. Häse, I. Fdez. Galván, A. Aspuru-Guzik, R. Lindh, and M. Vacher, "How machine learning can assist the interpretation of ab initio molecular dynamics simulations and conceptual understanding of chemistry," *Chemical Science*, vol. 10, no. 8, pp. 2298–2307, 2019, ISSN: 2041-6520. DOI: 10.1039/C8SC04516J. [Online]. Available: https://doi.org/10.1039/C8SC04516J.

[388] S. Jiang and V. M. Zavala, "Convolutional neural nets in chemical engineering: Foundations, computations, and applications," *AIChE Journal*, vol. 67, no. 9, e17282, 2021, ISSN: 0001-1541. DOI: 10.1002/aic.17282. [Online]. Available: https://doi.org/10.1002/aic.17282.

[389] T. W. Walker, A. K. Chew, R. C. Van Lehn, J. A. Dumesic, and G. W. Huber, "Rational design of mixed solvent systems for acid-catalyzed biomass conversion processes using a combined experimental, molecular dynamics and machine learning approach," *Topics in Catalysis*, vol. 63, no. 7, pp. 649–663, 2020, ISSN: 1572-9028. DOI: 10.1007/s11244-020-01260-9. [Online]. Available: https://doi.org/10.1007/s11244-020-01260-9.

[390] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015, ISSN: 1476-4687. DOI: 10.1038/nature14539. [Online]. Available: https://doi.org/10.1038/nature14539.

[391] V. K. Ramaswamy, S. C. Musson, C. G. Willcocks, and M. T. Degiacomi, "Deep learning protein conformational space with convolutions and latent interpolations," *Physical Review X*, vol. 11, no. 1, p. 011 052, 2021. DOI: 10.1103/PhysRevX.11.011052. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevX.11.011052.

[392] K. Ryczko, K. Mills, I. Luchak, C. Homenick, and I. Tamblyn, "Convolutional neural networks for atomistic systems," *Computational Materials Science*, vol. 149, pp. 134–142, 2018, ISSN: 0927-0256. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025618301526.

[393] R. Singh *et al.*, *3d deep learning with voxelized atomic configurations for modeling atomistic potentials in complex solid-solution alloys*, 2018. arXiv: 1811.09724 [cond-mat.mtrl-sci].

[394] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, "Gromacs: Fast, flexible, and free," *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, 2005, ISSN: 0192-8651. DOI: 10.1002/jcc.20291. [Online]. Available: https://doi.org/10.1002/jcc.20291.

[395] K. O'Shea and R. Nash, *An introduction to convolutional neural networks*, 2015. arXiv: 1511.08458 [cs.NE].

[396] T. Szandała, "Review and comparison of commonly used activation functions for deep neural networks," in *Bio-inspired Neurocomputing*, A. K. Bhoi, P. K. Mallick, C.-M. Liu, and V. E. Balas, Eds. Singapore: Springer Singapore, 2021, pp. 203–224, ISBN: 978-981-15-5495-7. DOI: 10.1007/978-981-15-5495-7_11. [Online]. Available: https://doi.org/10.1007/978-981-15-5495-7_11.

[397] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[398] P. C. Hansen, "Deconvolution and regularization with toeplitz matrices," *Numerical Algorithms*, vol. 29, no. 4, pp. 323–378, 2002, ISSN: 1572-9265. DOI: 10.1023/A:1015222829062. [Online]. Available: https://doi.org/10.1023/A:1015222829062.

[399] D. Bank, N. Koenigstein, and R. Giryes, *Autoencoders*, 2021. arXiv: 2003.05991 [cs.LG].

[400] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: An overview and application in radiology," *Insights into Imaging*, vol. 9, no. 4, pp. 611–629, 2018, ISSN: 1869-4101. DOI: 10.1007/s13244-018-0639-9. [Online]. Available: https://doi.org/10.1007/s13244-018-0639-9.

[401] S. Weglarczyk, "Kernel density estimation and its application," *ITM Web Conf.*, vol. 23, 2018. DOI: 10.1051/itmconf/20182300037. [Online]. Available: https://doi.org/10.1051/itmconf/20182300037.

[402] A. K. Ghosh, P. Chaudhuri, and D. Sengupta, "Classification using kernel density estimates," *Technometrics*, vol. 48, no. 1, pp. 120–132, 2006, ISSN: 0040-1706. DOI: 10.1198/004017005000000391. [Online]. Available: https://doi.org/10.1198/004017005000000391.

[403] V. Seshadri and P. R. Westmoreland, "Concerted reactions and mechanism of glucose pyrolysis and implications for cellulose kinetics," *The Journal of Physical Chemistry A*, vol. 116, no. 49, pp. 11 997–12 013, 2012, ISSN: 1089-5639. DOI: 10.1021/jp3085099. [Online]. Available: https://doi.org/10.1021/jp3085099.

[404] C. Krumm, J. Pfaendtner, and P. J. Dauenhauer, "Millisecond pulsed films unify the mechanisms of cellulose fragmentation," *Chemistry of Materials*, vol. 28, no. 9, pp. 3108–3114, 2016, ISSN: 0897-4756. DOI: 10.1021/acs.chemmater.6b00580. [Online]. Available: https://doi.org/10.1021/acs.chemmater.6b00580.

[405] J. C. Velasco Calderón, S. Jiang, and S. H. Mushrif, "Understanding the effect of solvent environment on the interaction of hydronium ion with biomass derived species: A molecular dynamics and metadynamics investigation," *ChemPhysChem*, vol. 22, no. 21, pp. 2222–2230, 2021, ISSN: 1439-4235. DOI: 10.1002/cphc.202100485. [Online]. Available: https://doi.org/10.1002/cphc.202100485.

[406] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *1998 IEEE International Conference on Evolutionary Computation Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98TH8360)*, Anchorage, AK: IEEE, 2002, pp. 69–73.

[407] R. C. Eberhart and Y. Shi, "Tracking and optimizing dynamic systems with particle swarms," in *Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546)*, ser. Proceedings of the 2001 Congress on Evolutionary Computation (IEEE Cat. No.01TH8546), vol. 1, 2001, 94–100 vol. 1. DOI: 10.1109/CEC.2001.934376. [Online]. Available: https://doi.org/10.1109/CEC.2001.934376.

[408] J. Xin, G. Chen, and Y. Hai, "A particle swarm optimizer with multi-stage linearly-decreasing inertia weight," in *2009 International Joint Conference on Computational Sciences and Optimization*, ser. 2009 International Joint Conference on Computational Sciences and Optimization, vol. 1, 2009, pp. 505–508. DOI: 10.1109/CSO.2009.420. [Online]. Available: https://doi.org/10.1109/CSO.2009.420.

[409] M. S. Arumugam and M. V. C. Rao, "On the performance of the particle swarm optimization algorithm with various inertia weight variants for computing optimal control of a class of hybrid systems," *Discrete Dynamics in Nature and Society*, vol. 2006, pp. 1–17, 2006.

[410] J. J. Moré and D. C. Sorensen, "Computing a trust region step," *Siam Journal on Scientific and Statistical Computing*, vol. 4, pp. 553–572, 1983.

[411] T. Steihaug, "The conjugate gradient method and trust regions in large scale optimization," *SIAM Journal on Numerical Analysis*, vol. 20, pp. 626–637, 1983.

[412] R. H. Byrd, R. B. Schnabel, and G. A. Shultz, "Approximate solution of the trust region problem by minimization over two-dimensional subspaces," *Mathematical Programming*, vol. 40, no. 1, pp. 247–263, 1988, ISSN: 1436-4646. DOI: 10.1007/BF01580735. [Online]. Available: https://doi.org/10.1007/BF01580735.

[413] T. F. Coleman and A. Verma, "A preconditioned conjugate gradient approach to linear equality constrained minimization," *Computational Optimization and Applications*, vol. 20, no. 1, pp. 61–72, 2001, ISSN: 1573-2894. DOI: 10.1023/A:1011271406353. [Online]. Available: https://doi.org/10.1023/A:1011271406353.

[414] H.-C. Wu, "The karush–kuhn–tucker optimality conditions in multiobjective programming problems with interval-valued objective functions," *European Journal of Operational Research*, vol. 196, no. 1, pp. 49–60, 2009, ISSN: 0377-2217. DOI: https://doi.org/10.1016/j.ejor.2008.03.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0377221708002877.

[415] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*. WORLD SCIENTIFIC, 2002, p. 308, ISBN: 978-981-238-151-4. DOI: 10.1142/5089. [Online]. Available: https://doi.org/10.1142/5089.

[416] R. Fletcher, *Practical Methods of Optimization*, Second. New York, NY, USA: John Wiley & Sons, 1987.

[417] K. Schittkowski, "Nlpql: A fortran subroutine solving constrained nonlinear programming problems," *Annals of Operations Research*, vol. 5, no. 2, pp. 485–500, 1986, ISSN: 1572-9338. DOI: 10.1007/BF02022087. [Online]. Available: https://doi.org/10.1007/BF02022087.

[418] S. P. Han, "A globally convergent method for nonlinear programming," *Journal of Optimization Theory and Applications*, vol. 22, no. 3, pp. 297–309, 1977, ISSN: 1573-2878. DOI: 10.1007/BF00932858. [Online]. Available: https://doi.org/10.1007/BF00932858.

[419] M. J. D. Powell, "A fast algorithm for nonlinearly constrained optimization calculations," in *Numerical Analysis*, G. A. Watson, Ed., Berlin, Heidelberg: Springer Berlin Heidelberg, 1978, pp. 144–157, ISBN: 978-3-540-35972-2.

[420] R. A. Waltz, J. L. Morales, J. Nocedal, and D. Orban, "An interior algorithm for nonlinear optimization that combines line search and trust region steps," *Mathematical Programming*, vol. 107, no. 3, pp. 391–408, 2006, ISSN: 1436-4646. DOI: 10.1007/s10107-004-0560-5. [Online]. Available: https://doi.org/10.1007/s10107-004-0560-5.

[421] V.-D. Nguyen, "Contributions to fast matrix and tensor decompositions," Ph.D. dissertation, Université d'Orléans, 2016.

[422] S. J. Blanksby and G. B. Ellison, "Bond dissociation energies of organic molecules," *Accounts of Chemical Research*, vol. 36, no. 4, pp. 255–263, 2003, ISSN: 0001-4842. DOI: 10.1021/ar020230d. [Online]. Available: https://doi.org/10.1021/ar020230d.

[423] A. M. McKenna, A. G. Marshall, and R. P. Rodgers, "Heavy petroleum composition. 4. asphaltene compositional space," *Energy & Fuels*, vol. 27, no. 3, pp. 1257–1267, 2013, ISSN: 0887-0624. DOI: 10.1021/ef301747d. [Online]. Available: https://doi.org/10.1021/ef301747d.

Appendix A: Chapter 2

A.1 Introduction

Detailed information on extraction of concentration and spectral profiles of the liquid products obtained at different experimental conditions of thermal conversion by self-modeling curve resolution (SMCR) and a parallel method to detect the underlying network structure from Bayesian cluster groups is provided in the manuscript. However, some sections do not require that all figures, plots and tables be supplied in the manuscript itself, at the same time not causing difficulty for the readers in relating to the global aim of the study. These additional details are given in this Supporting Information document.

A.2 Experimental

All experimental details are provided in the main manuscript.

A.3 Methods and parameters used

A.3.1 FTIR data available

All the data regarding the FTIR spectra of the liquid products from thermal conversion at different temperatures and residence times is provided in the manuscript itself.

A.3.2 Pre-processed and residual data for temperatures of 420°C, 400°C, 380°C, 300°C

The FTIR spectra of liquid samples obtained after thermal conversion at 350°C after baseline correction and SG filtering are provided in the manuscript. The respective plots along with the residual obtained from smoothing and the raw data for the other 4 temperatures are given in Figure A.1, Figure A.2, Figure A.3 and Figure A.4.

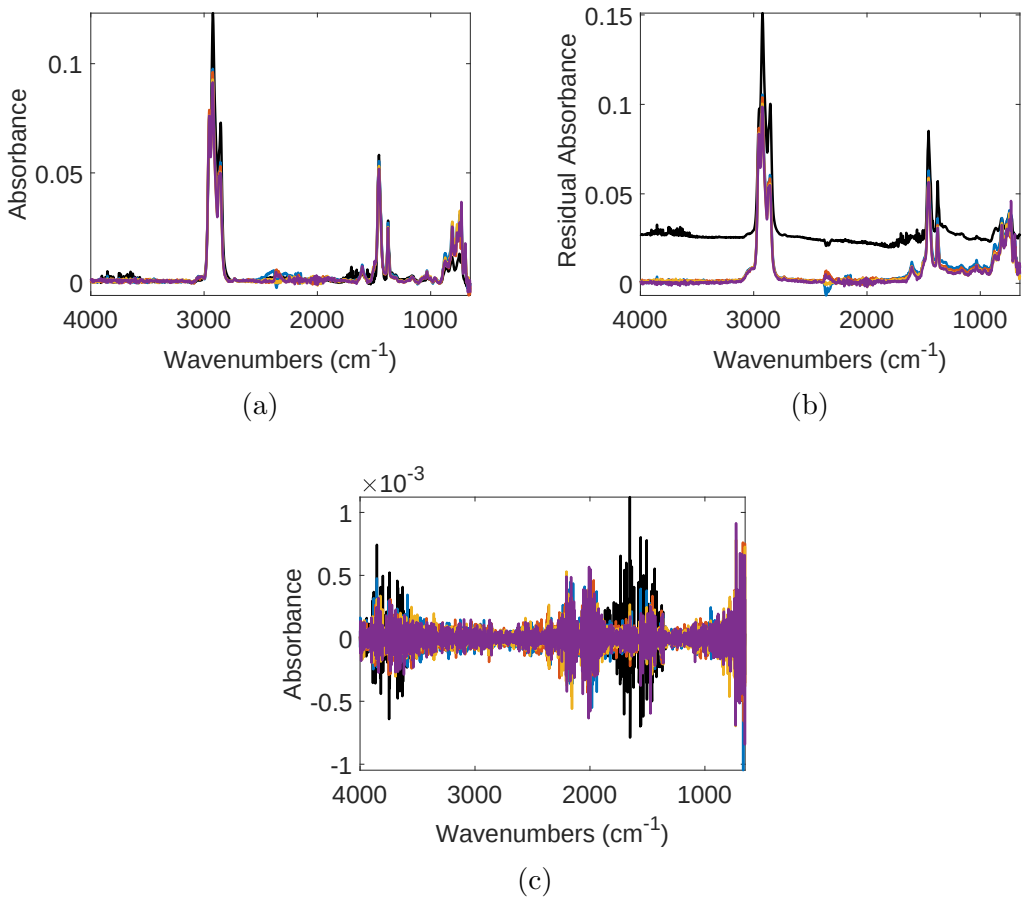
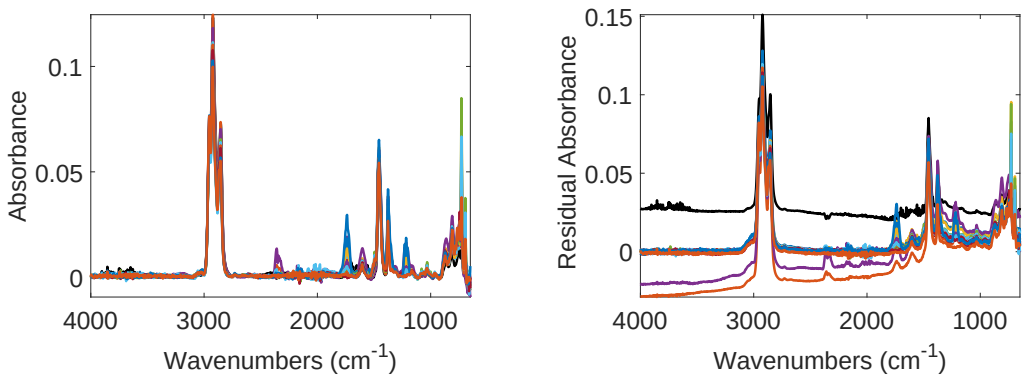


Figure A.1: Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 420°C; (c) residual after smoothing.

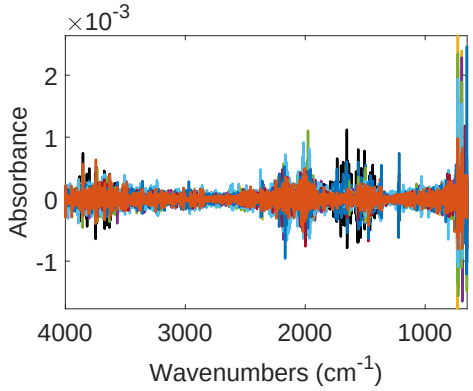
A.3.3 SMCR-ALS and SMCR-ALS-PSO methods

To deal with some of the limitations of MCR like rotational and intensity ambiguities, datasets from different runs and techniques are combined together into a single data



(a)

(b)



(c)

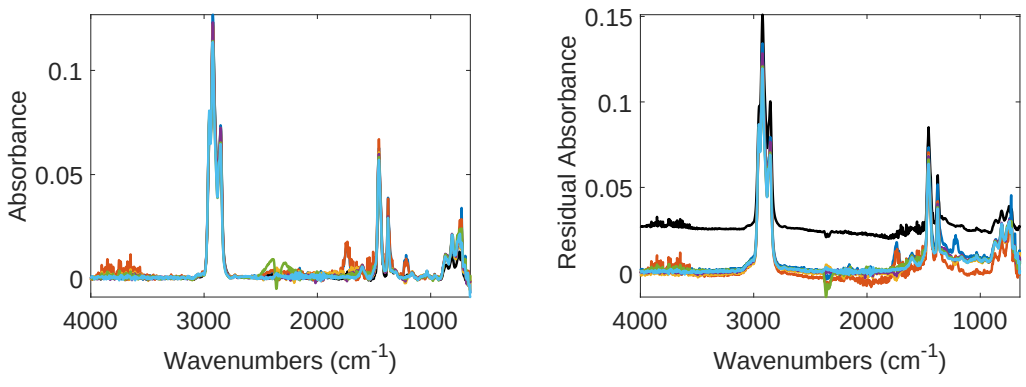
Figure A.2: Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 400°C; (c) residual after smoothing.

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 A.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 A.2.

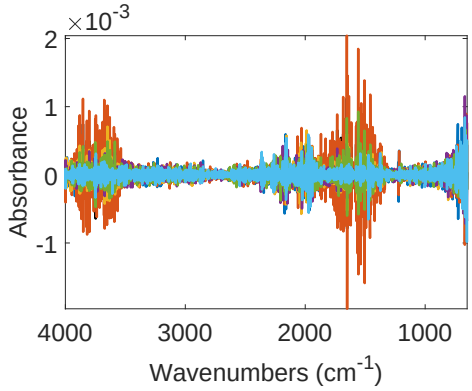
$$[D_1 D_2 D_3 \cdots D_n] = C\,[S_1 S_2 S_3 \cdots S_n]^T + [E_1 E_2 E_3 \cdots E_n] \tag{A.1}$$

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix} S^T \tag{A.2}$$

Intensity ambiguity is represented by:



(a) (b)



(c)

Figure A.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°C; (c) residual after smoothing.

$$D = (C^k)(S^{\frac{1}{k}})^T \tag{A.3}$$

where k is a scalar.

Rotational ambiguity is given in equation A.4 by:

$$D = (CT)(T^{-1}S^T) + E \tag{A.4}$$

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

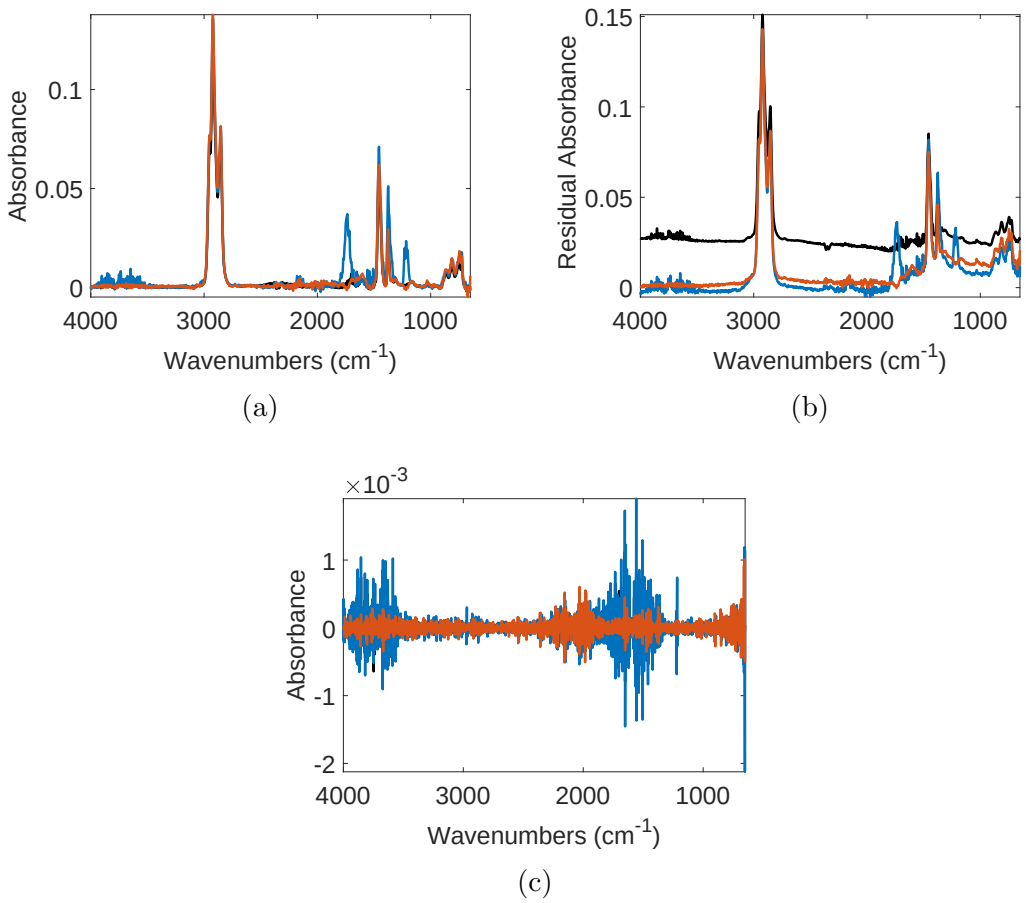


Figure A.4: Plots of: (a) Baseline corrected and smoothed data; (b) the raw FTIR spectra of the liquid products from thermal conversion of Athabasca bitumen at 300°C; (c) residual after smoothing.

in the manuscript. The respective equations of the alternative minimization of the Frobenius norm of the residual are given below:

$$\min_{S\geq 0}(||D - CS^T||^2) \tag{A.5}$$

$$\min_{C\geq 0}(||D^T - SC^T||^2) \tag{A.6}$$

Table A.1 gives some of the common strategies of choosing the inertia weight parameter for velocity updating in PSO.

As mentioned in the manuscript, 'fmincon' was used to further carry out a local search for the PSO-optimized concentration profiles inside the ALS loop. The next

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

Type of strategy	**Remarks**
Constant [406]	A value between 0.7 – 1 shows lower error but larger number of iterations for convergence
Random [407]	Increases convergence in early stages of PSO; Gives faster overall convergence
Linearly decreasing [408]	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 [409]	Falls in between constant and random inertia weight strategies; takes global and local best particle positions into consideration but gives large error

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.

a 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. [410] The trust region sub-problem is expressed in equation A.7 as:

$$\min_s q(s), s \in N \tag{A.7}$$

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:

$$\min \frac{1}{2} s^T H s + s^T g \qquad \text{for} ||Ds|| \leq \Delta \tag{A.8}$$

In equation A.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.[411],[412] 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

$$H.s_2 = -g \tag{A.9}$$

The solution to equation A.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.

b 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. [413] 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. [414] The solution also comprises of a reflection step that delineates the step size.

c 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:

$$\nabla f(x_s) + \sum_{i=1}^{m} \lambda_i . \nabla G_i(x_s) = 0$$
$$\lambda_i . G_i(x_s) = 0 \quad \text{and} \quad \lambda_i \geq 0 \tag{A.10}$$

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. [415] 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 A.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. [416],[417].

The solution of the quadratic sub-problem is used to form a search direction for the variable x as:

$$x_{k+1} = x_k + \alpha_k d_k \tag{A.11}$$

Here, d_k is the search direction and

$$\alpha_k$$

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 [417] 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:

i Updating the Hessian (H_k) of the Lagrangian formulation

The Lagrangian formulation of the quadratic problem is given by the following equation:

$$L(\lambda, x) = f(x) + \sum \lambda_i . g_i(x) \tag{A.12}$$

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.

ii 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 A.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 A.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.

iii 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 A.9.

iv Merit function and step length

A merit function proposed by Han [418] is used and a penalty parameter was introduced by Powell. [419] 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. [420] 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 A.13:

$$\min_{x,s} f(x,s) = \min f(x,s) - \mu \sum \ln(s_i) \tag{A.13}$$

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 A.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.

A.3.4 Bayesian networks

Figure A.5 shows the importance index for the first 1550 wavenumbers. The procedure behind the choice of these wavenumbers is given in the main manuscript.

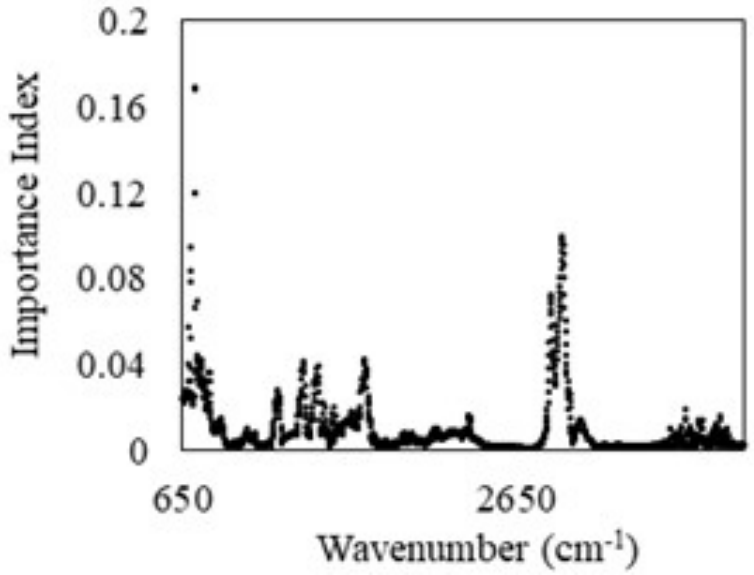
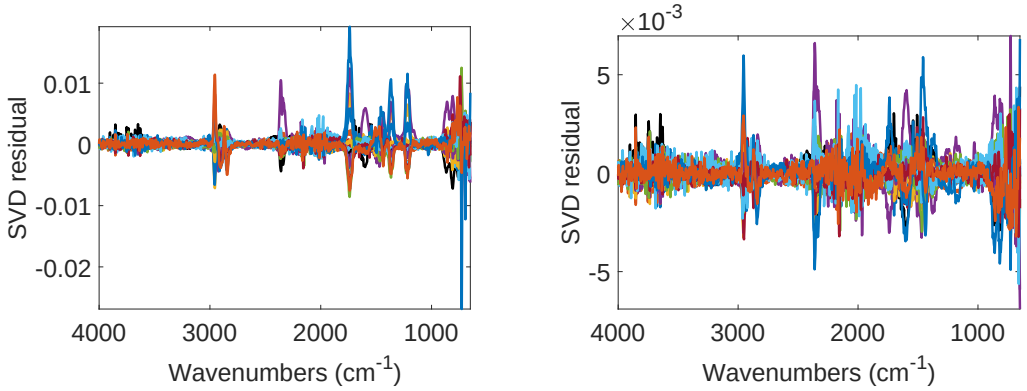


Figure A.5: Plot of importance index of the selected 1550 wavenumbers.

A.4 Results and Discussion

A.4.1 Rank determination of each sub-matrix

Figure A.6 gives the plots of residuals obtained after performing SVD on the 400°C 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°C, 350°C, 380°C, 420°C) are given in Figure A.7, Figure A.8, Figure A.9 and Figure A.10 respectively.



(a) (b)

Figure A.6: Residuals obtained after performing SVD on the 400°Cdata set considering: (a) 2 components and (b) 4 components.

The values of performance indicators (LOF and R^2) for SVD with 2, 3 and 4 pseudo-components are given in Table A.2.

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

	300°C		350°C			380°C			420°C		
# Components	2	3	2	3	4	2	3	4	2	3	4
LOF	2.38	8.27E-14	3.09	2.17	1.54	7.2	4.99	3.71	4.72	2.93	1.83
R^2	99.94	100	99.9	99.95	99.97	99.48	99.75	99.86	99.78	99.91	99.96

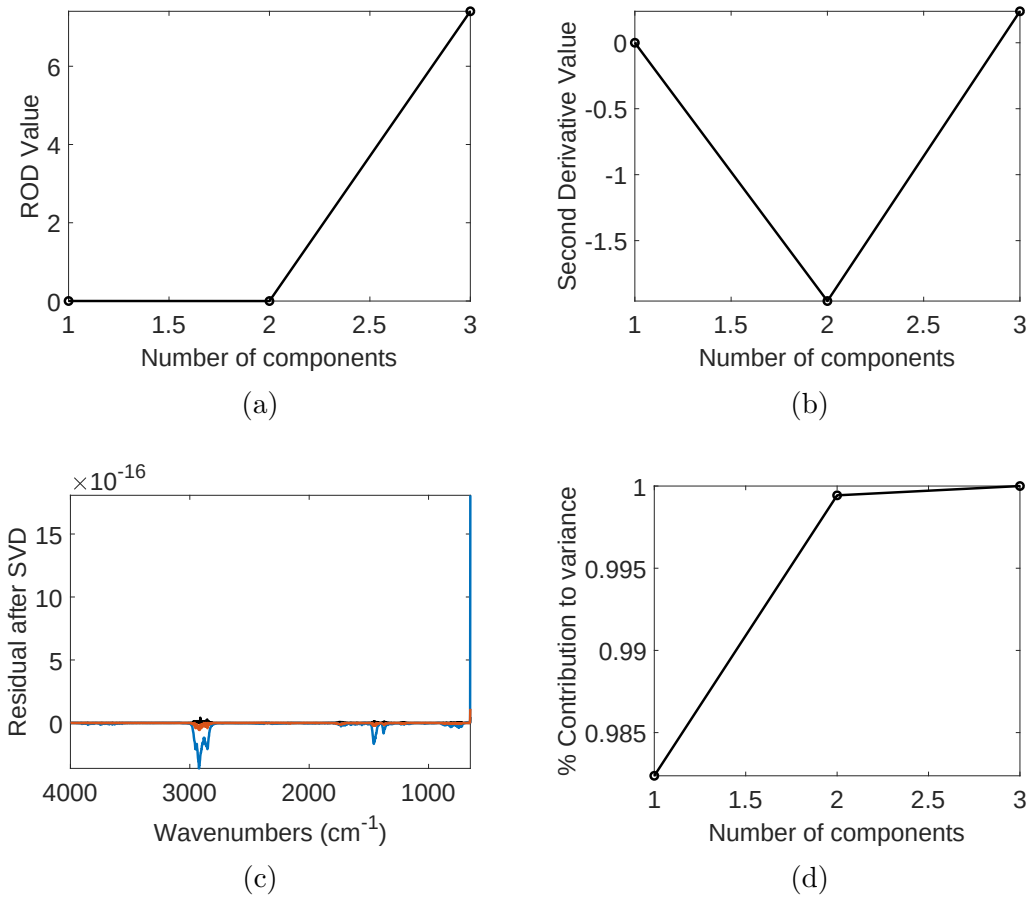


Figure A.7: Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each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

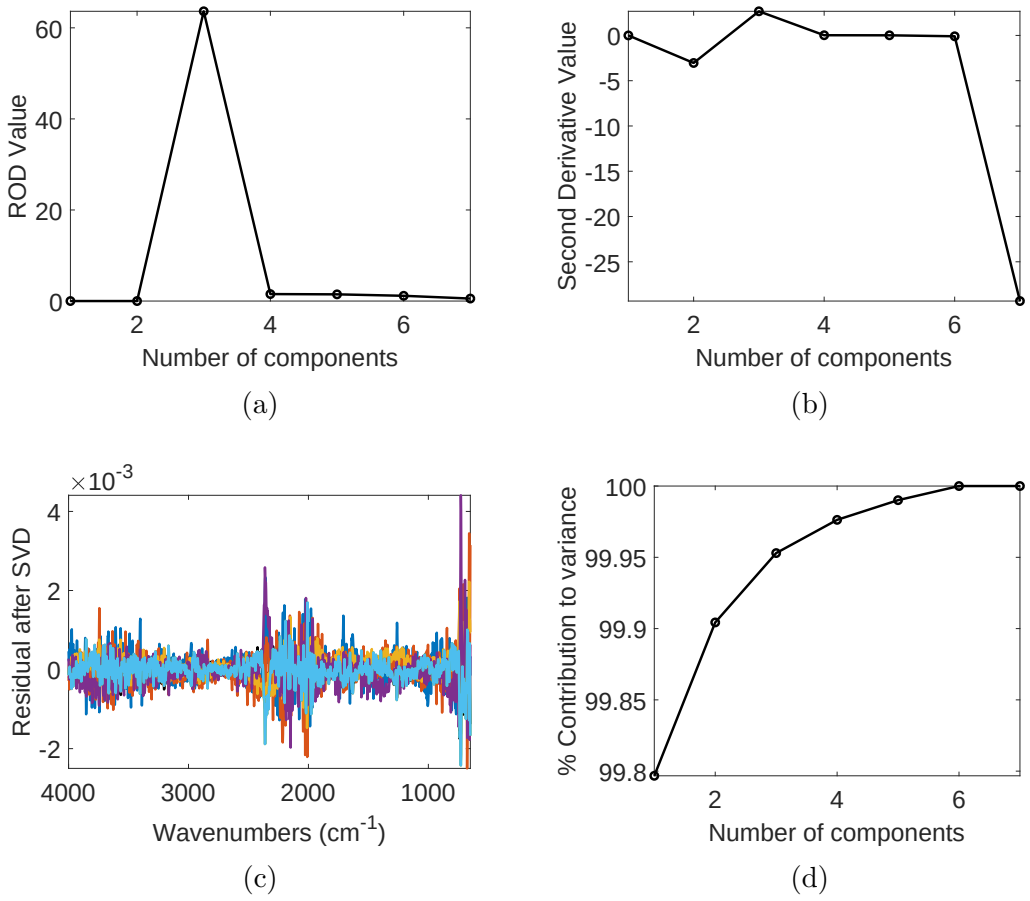


Figure A.8: Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each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

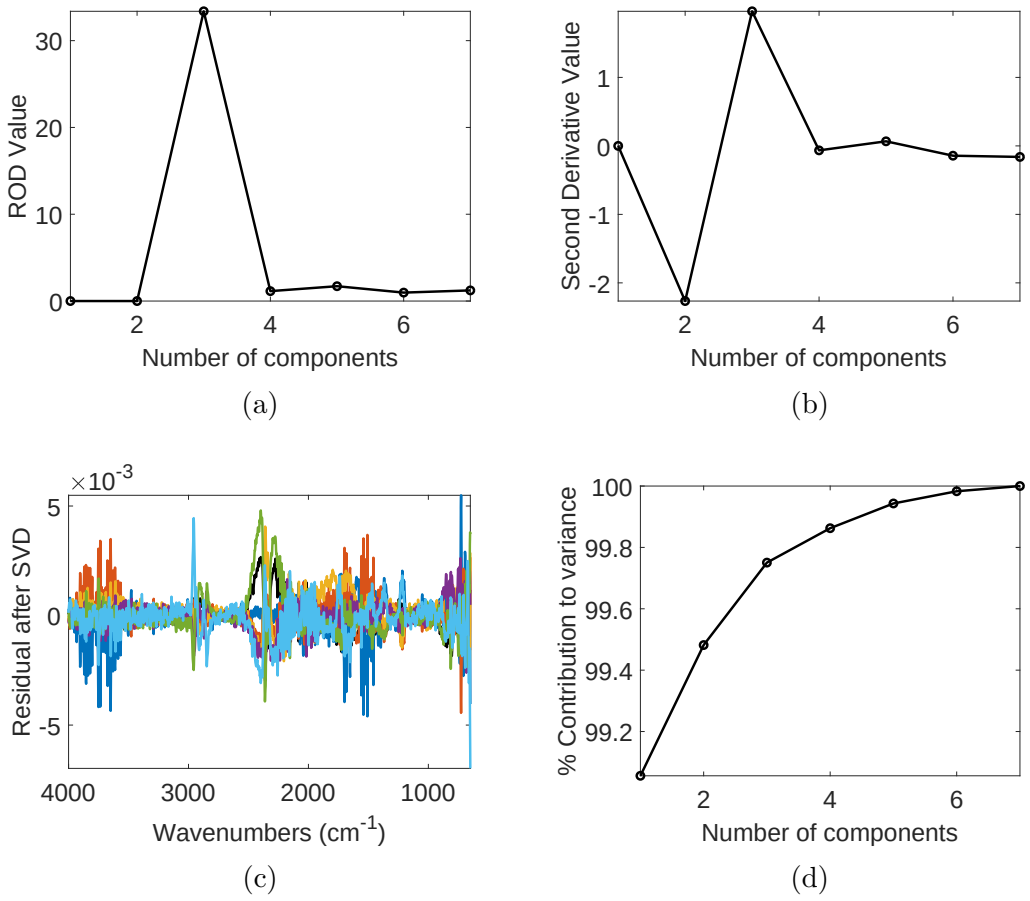


Figure A.9: Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each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

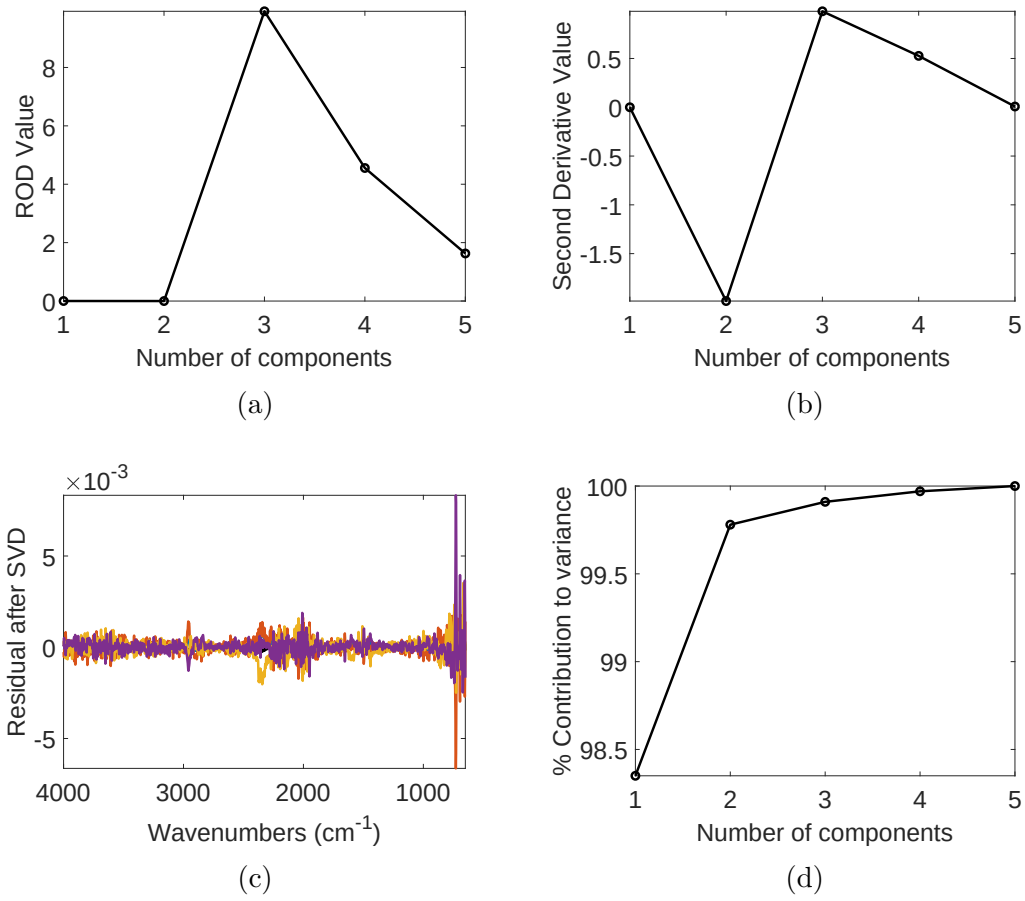


Figure A.10: Plots for (a) ROD with respect to each component; (b) SD with respect to each component; (c) Residual after performing SVD considering 3 components on the FTIR data set for all 1738 wavenumbers; (d) Percentage contribution to the variance explained by the eigenvalues corresponding to each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

A.4.2 Initial concentration estimates

The initial estimates of concentration profiles at 300°C are given in Figure A.11.

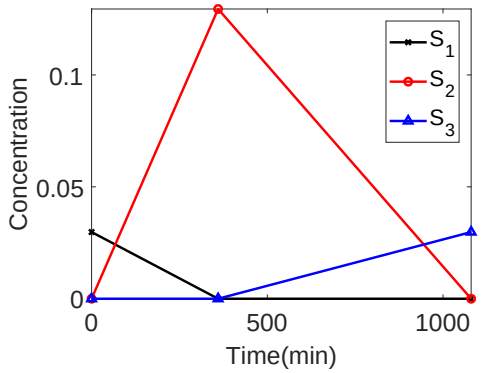


Figure A.11: Initial concentration estimates for S1, S2 and S3 at 300°C.

A.4.3 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 A.12.

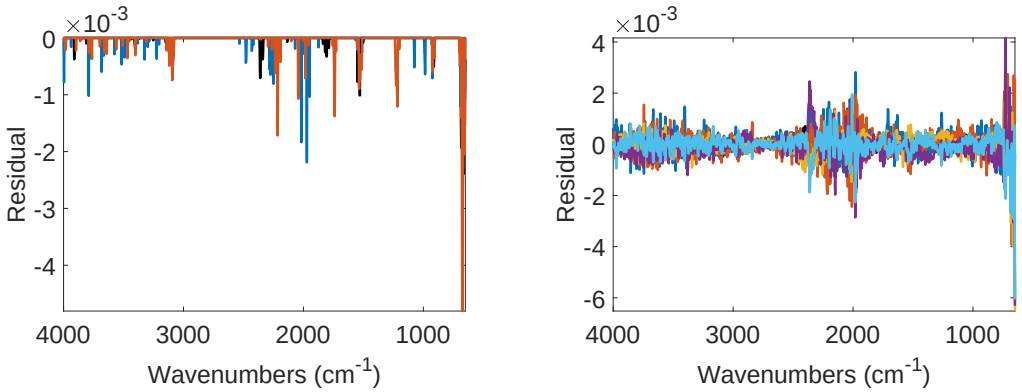
A.4.4 PSO-optimized concentration and spectral profiles

Results at 300°C

The concentration and spectral profiles when the ALS-PSO algorithm was used to resolve the FTIR spectra obtained at 300°C for Athabasca bitumen is given in Figure A.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 A.17b). Discussion on the differences of these profiles with respect to ALS-optimized results in terms of resolution quality and convergence speed is provided in the manuscript.

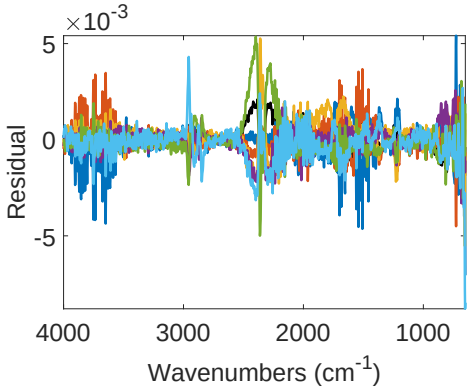
Results at 350°C

The concentration and spectral profiles when the ALS-PSO algorithm was used to resolve the FTIR spectra obtained at 350°C for Athabasca bitumen are given in Figure

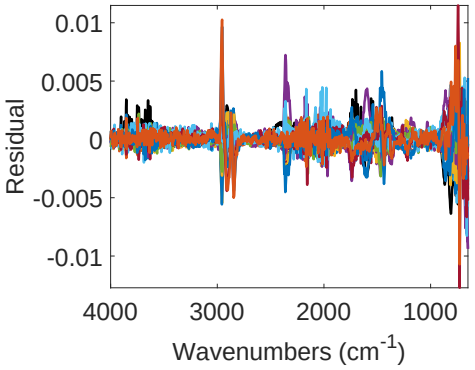


(a)

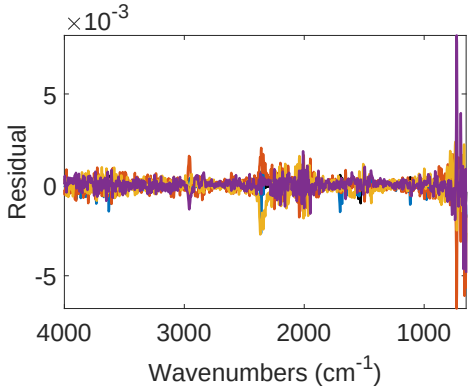
(b)



(c)



(d)



(e)

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

A.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

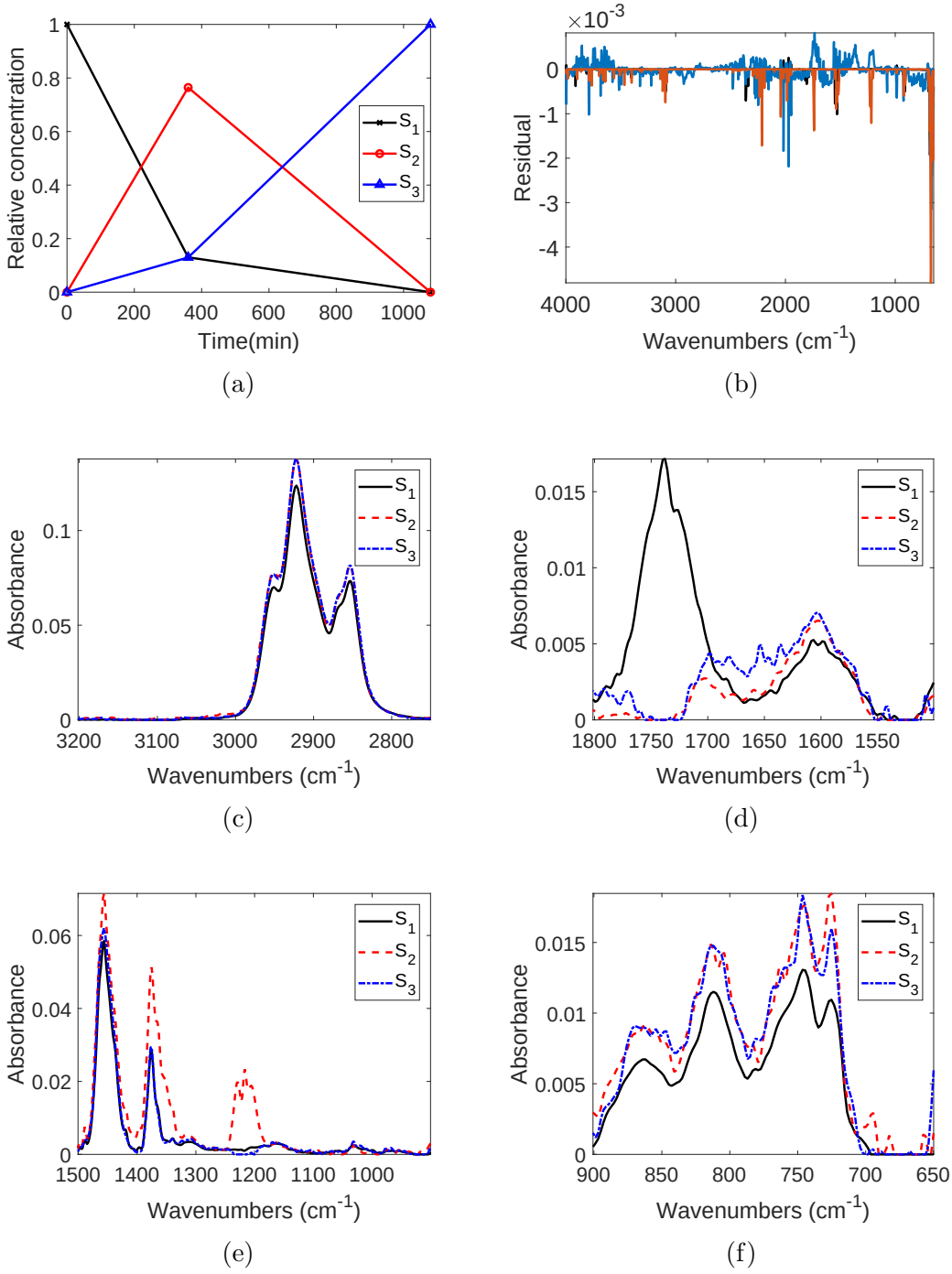


Figure A.13: Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 300°C. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown 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}.

quality and convergence speed is given in the manuscript.

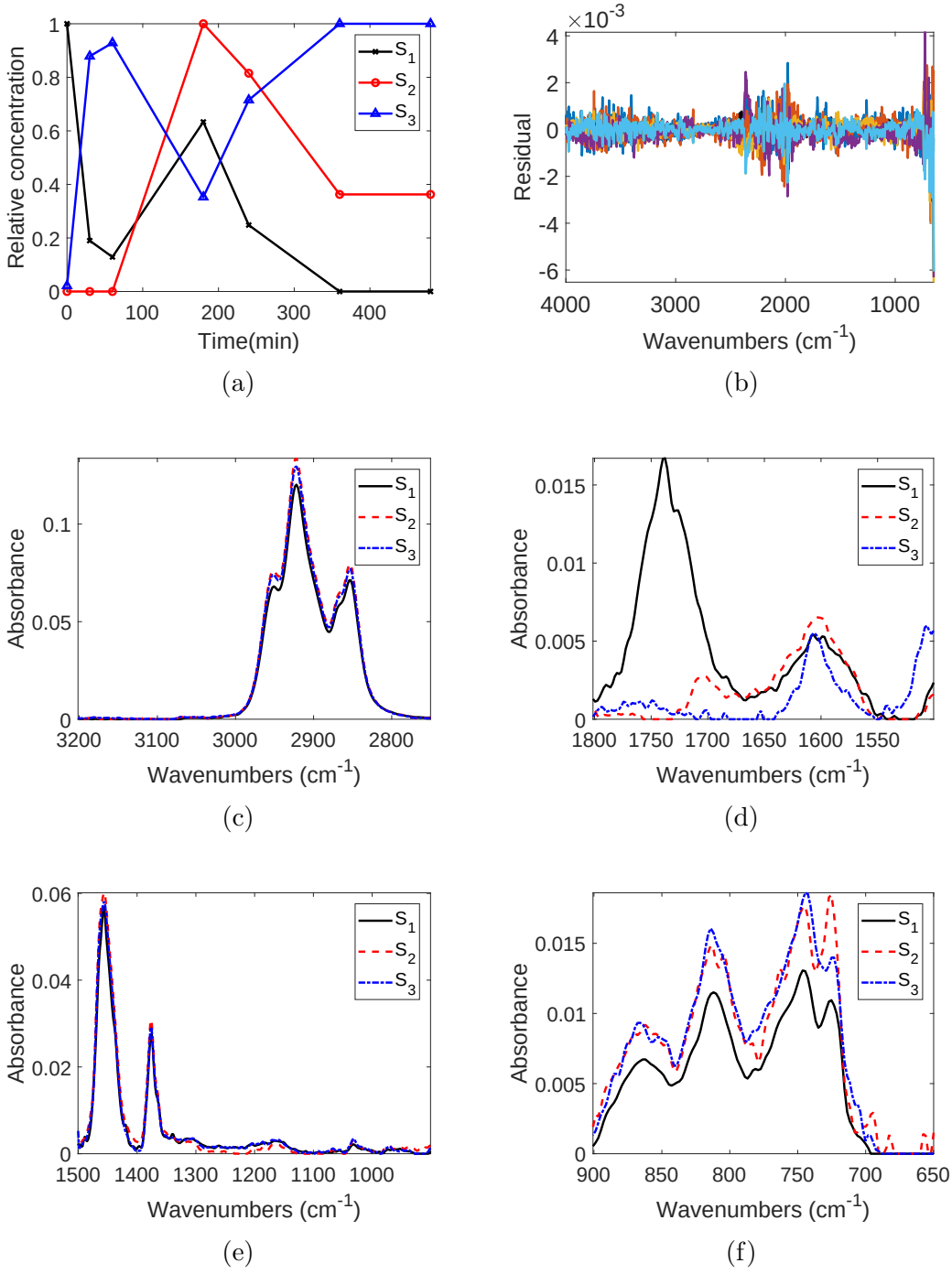


Figure A.14: Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 350°C. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown 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}.

Results at 380°C

Figure A.15 provides the ALS-PSO-resolved concentration and spectral profiles for the 380 °C dataset. The residual plot when the reproduced matrix is subtracted from the original data matrix is also provided in the figure.

Results at 400°C

Figure A.16 gives the ALS-PSO resolved final profiles for the dataset obtained at 400°C. The residual plot when the reproduced matrix is subtracted from the original data matrix is also provided in the figure.

Results at 420°C

Figure A.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.

Comparison of ALS and ALS-PSO methods

The results and corresponding discussion of this section are provided in the manuscript itself.

A.4.5 BHC and associated chemical signatures relative to the clusters

The variation of effective intensity for each wavenumber in this cluster is shown in Figure A.18. Other necessary information regarding this section is provided in the manuscript.

A.4.6 Deriving chemical reaction pathway through Bayesian networks applied on the BHC clusters

All the required details of this section are provided in the manuscript.

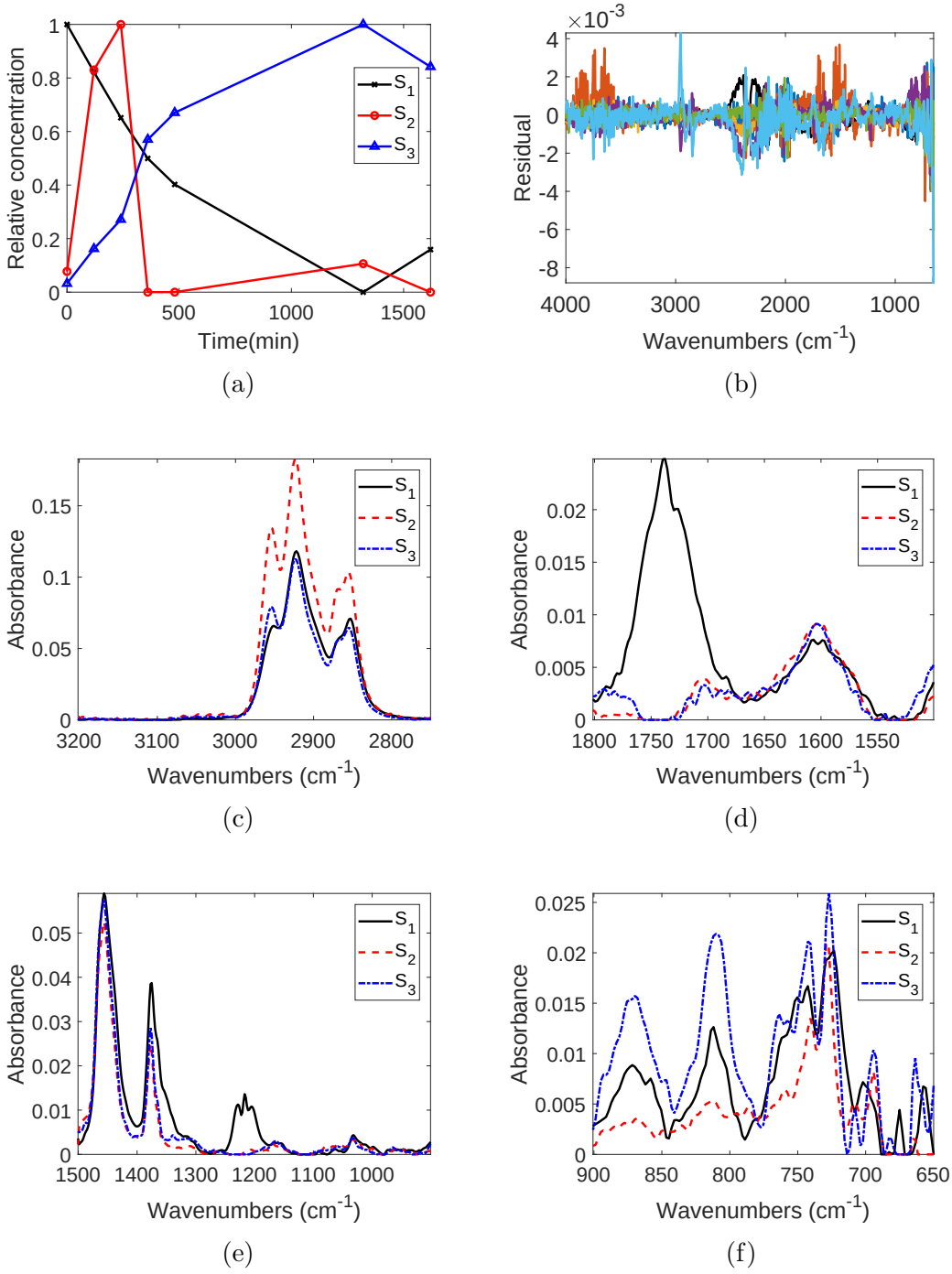


Figure A.15: Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 380°C. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown 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}.

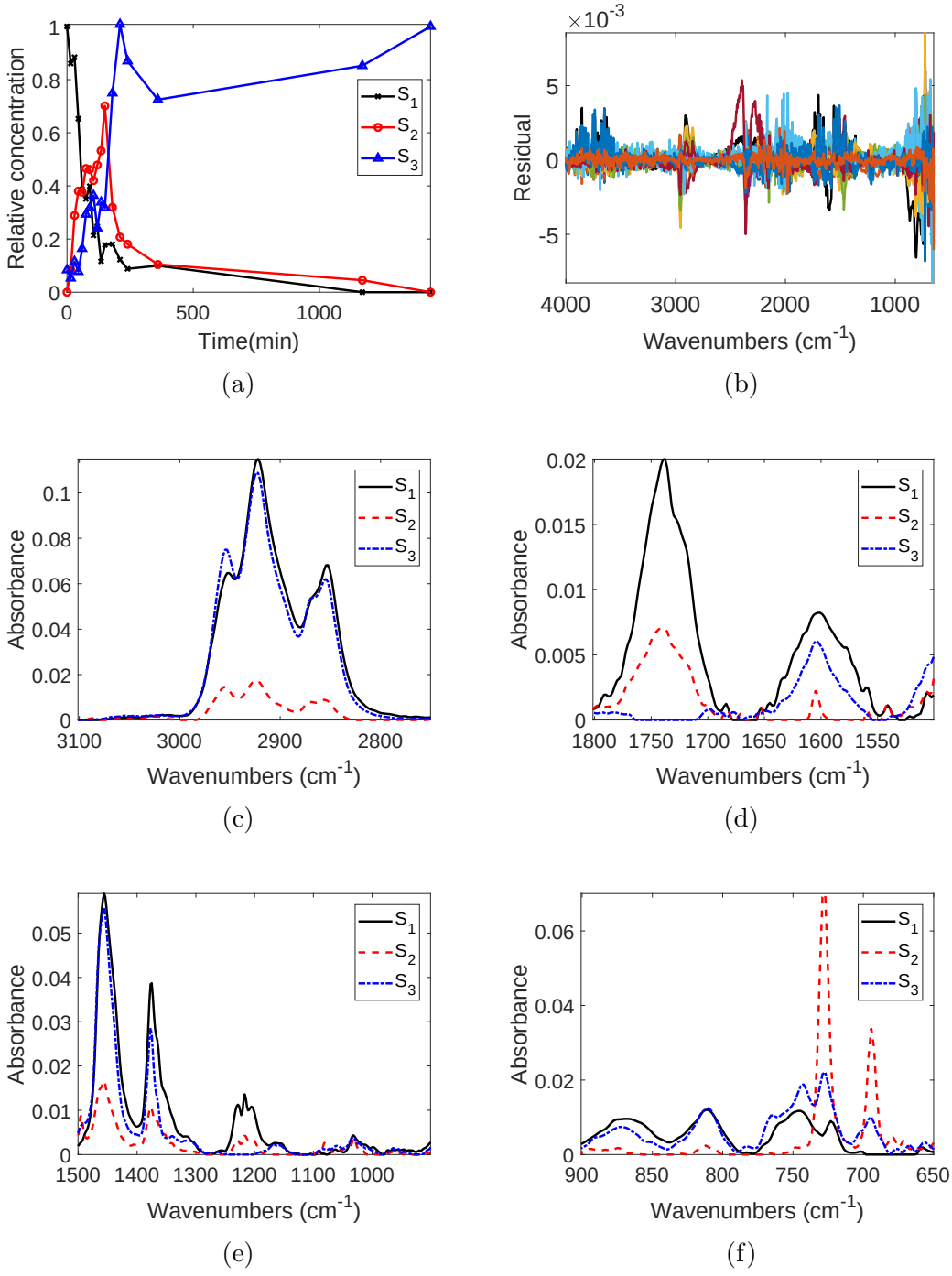


Figure A.16: Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 400°C. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown 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}.

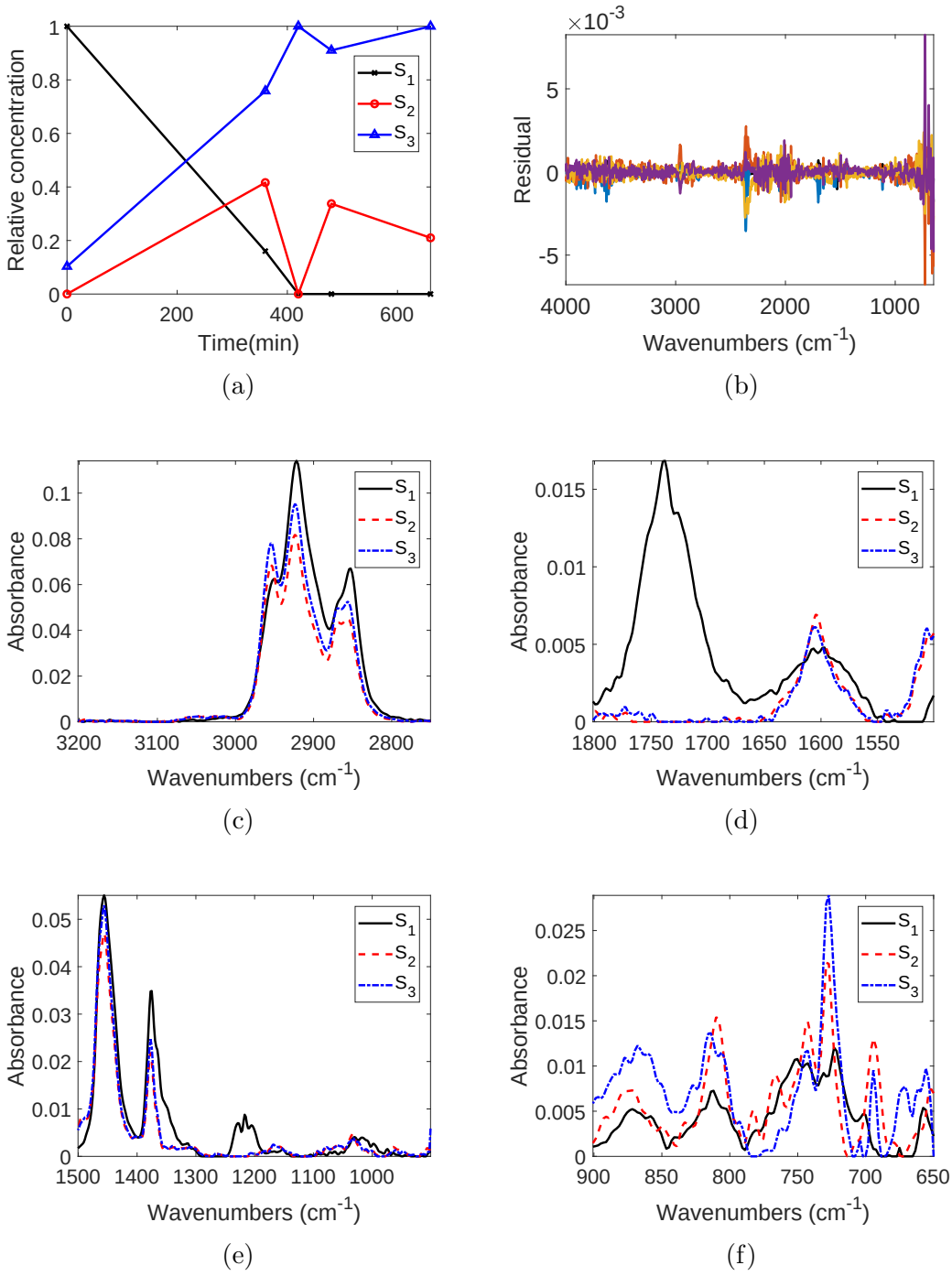


Figure A.17: Results of SMCR-ALS-PSO applied to FTIR spectra of liquid products from thermal conversion of Athabasca bitumen at 420°C. The profiles are arranged as: (a) concentration vs. reaction time for the three pseudo-components; (b) residual plot; and resolved spectra for each pseudo-component shown 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}.

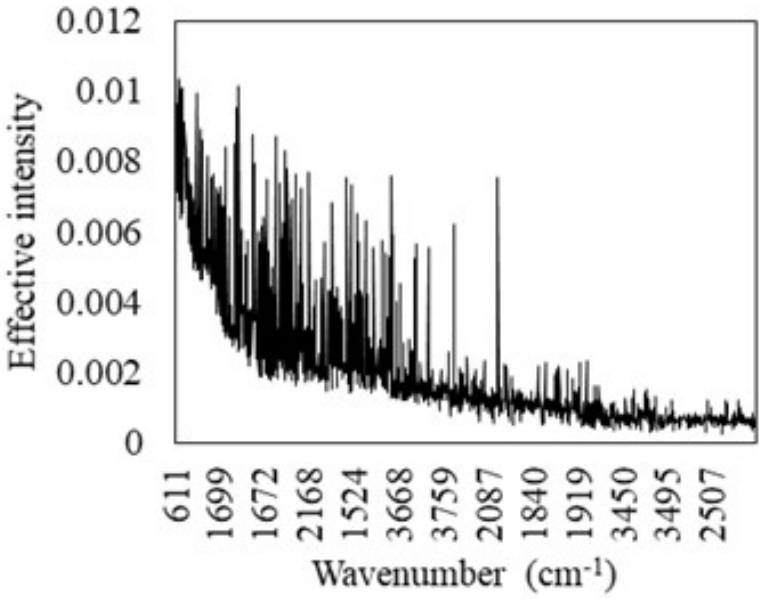
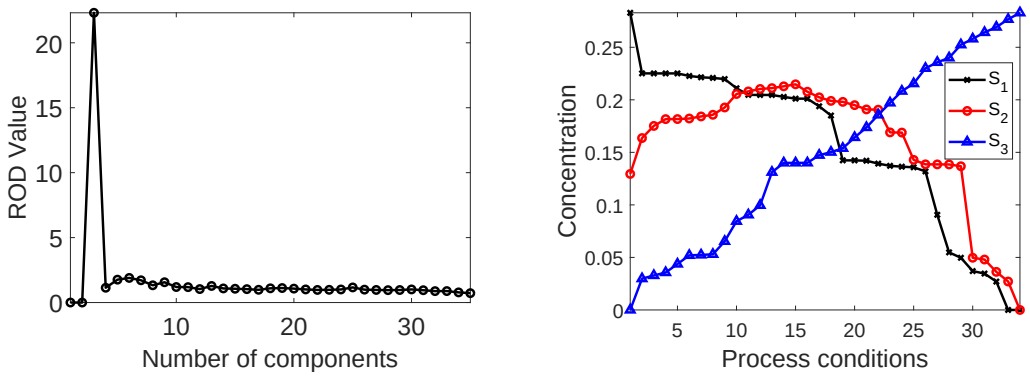


Figure A.18: Effective intensity for each wavenumber in the fifth cluster (Table 14 in the manuscript). Some of the important peaks are indicated.



(a) (b)

Figure A.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.

A.4.7 ALS-optimized profiles for the global model

Figure A.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.

Appendix B: Chapter 3

B.1 Process Conditions

Table B.1: Process conditions for spectral data collection

Spectral sensor	Process conditions	
	Temperature($°C$)	Residence time (min)
FTIR	150	66, 126, 186, 246, 306, 366, 426, 486
	200	66, 126, 186, 246, 306, 486
	250	246
	300	126, 186, 246, 306, 366, 426, 486
	340	6, 66, 126, 246, 486
	360	6, 16.02, 25.98, 36, 66, 246, 583.02
	400	6, 16.02, 25.98, 36, 66, 96, 126
^{1}H-NMR	150	60, 120, 180, 240, 300, 360, 420, 480
	200	60, 120, 180, 240, 300, 360, 420, 480
	250	60, 120, 180, 240, 300, 360, 420, 480
	300	60, 120, 180, 240, 300, 360, 420, 480

B.2 Highly correlated spectral channels

Table B.2: Wavenumbers of FTIR spectra with correlation > 0.9

609.46	619.11	1323.07	1911.32	1448.43	2634.57	3116.74	3193.89	3242.10	3627.84	1920.96	3946.07
763.76	802.00	1810.00	1911.32	734.83	2644.21	2297.05	3203.53	1390.57	3637.48	2750.29	3946.07
783.04	841.00	1294.14	1920.96	1631.66	2644.21	2287.41	3213.17	1911.32	3637.48	3830.35	3946.07
599.82	870.00	1487.01	1930.60	2605.64	2644.21	2239.19	3222.82	2711.72	3637.48	1303.78	3955.71
648.04	937.00	1890.00	1930.60	1467.72	2653.86	1053.06	3232.46	1361.64	3647.12	1853.46	3955.71
667.32	966.00	1420.00	1940.25	1853.46	2653.86	1564.16	3232.46	3242.10	3647.12	2682.79	3955.71
898.76	1010.00	1600.00	1949.89	1949.89	2663.50	1940.25	3232.46	1409.86	3656.77	3685.70	3955.71
1014.48	1050.00	1197.71	1969.18	1062.70	2682.79	2991.37	3232.46	2557.42	3656.77	1101.27	3965.35
1014.48	1080.00	1245.92	1978.82	1650.95	2682.79	2412.77	3242.10	3637.48	3656.77	1776.31	3965.35
1081.99	1100.00	1640.00	1978.82	2615.28	2682.79	3087.81	3251.75	1487.01	3666.41	2557.42	3965.35
811.97	1140.00	1770.00	1988.46	1323.07	2692.43	1458.08	3261.39	2499.56	3666.41	3531.40	3965.35
763.76	1160.00	1700.00	2007.75	1814.88	2692.43	2769.58	3271.03	3531.40	3666.41	1004.84	3975.00
1139.85	1170.00	1020.00	2036.68	2644.21	2692.43	821.62	3280.68	3569.98	3676.05	1660.59	3975.00
1188.06	1220.00	1700.00	2036.68	1390.57	2702.07	3261.39	3280.68	1487.01	3685.70	2162.04	3975.00
898.76	1260.00	1550.00	2046.32	1882.39	2702.07	2846.72	3290.32	2499.56	3685.70	3483.19	3975.00
1101.27	1270.00	860.00	2055.97	744.47	2711.72	831.26	3309.61	3531.40	3685.70	3946.07	3975.00

Table B.2 continued from previous page

1014.48	1280.00	1062.70	2075.25	1487.01	2711.72	667.32	3319.25	1294.14	3695.34	1400.22	3984.64
1101.27	1290.00	744.47	2084.90	1882.39	2711.72	2866.01	3328.89	1853.46	3695.34	1920.96	3984.64
706.00	1310.00	1487.01	2084.90	734.83	2721.36	667.32	3348.18	2624.93	3695.34	2731.00	3984.64
1024.13	1320.00	1872.74	2084.90	1525.58	2721.36	2663.50	3357.82	3656.77	3695.34	3782.13	3984.64
1101.27	1330.00	1294.14	2094.54	2499.56	2721.36	2914.23	3367.47	1400.22	3704.98	1110.92	3994.29
1207.35	1370.00	1043.41	2104.18	1110.92	2731.00	2952.80	3377.11	1901.67	3704.98	1805.24	3994.29
1280.00	1390.00	1630.00	2104.18	1795.60	2731.00	3377.11	3386.75	2721.36	3704.98	2567.07	3994.29
1332.71	1400.00	706.00	2113.83	2567.07	2731.00	3386.75	3396.40	705.90	3714.63	3550.69	3994.29
638.39	1420.00	1487.01	2113.83	1255.57	2740.65	2875.65	3406.04	2200.62	3714.63		
1313.43	1430.00	1024.13	2123.47	1824.53	2740.65	1612.37	3415.68	1380.93	3724.27		
831.26	1460.00	1795.60	2123.47	2586.35	2740.65	3299.96	3415.68	1506.30	3733.91		
1091.63	1480.00	1720.00	2142.76	1294.14	2750.29	2644.21	3425.33	1506.30	3743.56		
1101.27	1490.00	1710.00	2162.04	2094.54	2750.29	1544.87	3434.97	1352.00	3753.20		
1080.00	1500.00	1390.57	2171.69	946.98	2759.93	2084.90	3434.97	3222.82	3753.20		
1014.48	1520.00	1100.00	2181.33	1564.16	2759.93	811.97	3444.61	1323.07	3762.84		
1024.13	1530.00	908.00	2190.97	1978.82	2759.93	3261.39	3444.61	2509.21	3762.84		
734.83	1540.00	2162.04	2200.62	918.05	2769.58	1390.57	3454.26	3560.33	3762.84		
1525.58	1540.00	1371.29	2229.55	2316.34	2779.22	1911.32	3454.26	1294.14	3772.49		
879.48	1554.51	1730.00	2248.83	821.62	2798.51	2711.72	3454.26	2036.68	3772.49		
1438.79	1564.16	1371.29	2268.12	937.34	2808.15	1583.44	3463.90	2972.09	3772.49		
744.47	1583.44	2258.48	2277.76	1168.78	2817.79	2538.14	3463.90	1043.41	3782.13		
1290.00	1590.00	1371.29	2297.05	2769.58	2827.44	1043.41	3473.54	1920.96	3782.13		
1515.94	1602.73	725.00	2316.34	2779.22	2837.08	1631.66	3473.54	2972.09	3782.13		
1564.16	1612.37	975.91	2325.98	2325.98	2846.72	2596.00	3473.54	1014.48	3791.77		
1602.73	1620.00	976.00	2345.27	2335.63	2856.37	1033.77	3483.19	1535.23	3791.77		
1602.73	1631.66	957.00	2364.56	1959.53	2875.65	1564.16	3483.19	2084.90	3791.77		
1544.87	1641.30	2350.00	2374.20	1901.67	2885.30	2104.18	3483.19	3001.02	3791.77		
1303.78	1650.95	1350.00	2393.49	1949.89	2894.94	3454.26	3483.19	1361.64	3801.42		
1024.13	1660.59	1747.38	2403.13	667.32	2914.23	1535.23	3492.83	3733.91	3801.42		
1583.44	1660.59	2133.11	2412.77	667.32	2933.51	1930.60	3492.83	1400.22	3811.06		
1294.14	1670.23	2268.12	2422.42	2904.58	2943.16	3463.90	3492.83	1901.67	3811.06		
734.83	1679.88	2297.05	2432.06	1689.52	2952.80	2190.97	3502.47	2721.36	3811.06		
1477.37	1679.88	1370.00	2451.35	705.90	2962.44	1091.63	3512.12	3772.49	3811.06		
1053.06	1689.52	2130.00	2460.99	1438.79	2962.44	1689.52	3512.12	3222.82	3820.70		
1564.16	1689.52	2210.00	2470.63	1853.46	2962.44	2547.78	3512.12	1419.50	3830.35		
1400.22	1699.16	638.00	2489.92	2731.00	2962.44	744.47	3521.76	2721.36	3830.35		
1361.64	1728.09	1080.00	2499.56	1400.22	2972.09	1525.58	3521.76	3811.06	3830.35		
686.61	1760.00	2480.00	2499.56	2489.92	2972.09	1901.67	3521.76	3618.19	3839.99		
1072.34	1776.31	1430.00	2509.21	1053.06	2981.73	3434.97	3521.76	3724.27	3849.63		
1062.70	1785.95	2490.00	2509.21	1679.88	2981.73	1313.43	3531.40	1795.60	3859.28		
1689.52	1785.95	1070.00	2528.49	2181.33	2981.73	1872.74	3531.40	2981.73	3859.28		
1400.22	1795.60	2080.00	2528.49	1043.41	2991.37	2682.79	3531.40	1226.64	3868.92		
1081.99	1805.24	1930.00	2538.14	1795.60	2991.37	1043.41	3541.05	3020.30	3868.92		
1593.09	1805.24	1795.60	2547.78	2605.64	2991.37	1554.51	3541.05	1255.57	3878.56		
1245.92	1814.88	1110.92	2557.42	1265.21	3001.02	1949.89	3541.05	2499.56	3878.56		
1641.30	1814.88	2499.56	2557.42	2181.33	3001.02	2991.37	3541.05	3695.34	3878.56		
1284.50	1824.53	1390.57	2567.07	1419.50	3010.66	1274.85	3550.69	1795.60	3888.21		
1679.88	1824.53	1901.67	2567.07	3001.02	3010.66	1843.81	3550.69	3001.02	3888.21		
1323.07	1834.17	1062.70	2576.71	628.75	3029.95	2605.64	3550.69	1419.50	3897.85		
1814.88	1834.17	1602.73	2576.71	2788.86	3039.59	1014.48	3560.33	676.97	3907.49		
1400.22	1843.81	2080.00	2576.71	783.04	3058.88	1689.52	3560.33	1699.16	3907.49		
1824.53	1843.81	1290.00	2586.35	2142.76	3068.52	2528.49	3560.33	2692.43	3907.49		
1490.00	1850.00	1680.00	2586.35	1197.71	3087.81	3550.69	3560.33	3762.84	3907.49		
744.00	1860.00	744.00	2596.00	3058.88	3097.45	3001.02	3569.98	1776.31	3917.14		
1535.23	1863.10	1480.00	2596.00	773.40	3116.74	1409.86	3579.62	2981.73	3917.14		
1004.84	1870.00	1860.00	2596.00	715.54	3126.38	2123.47	3579.62	705.90	3926.78		
1544.87	1870.00	1070.00	2605.64	1352.00	3136.03	3454.26	3579.62	1776.31	3926.78		
1004.84	1880.00	1580.00	2605.64	773.40	3145.67	2403.13	3589.26	2702.07	3926.78		
1520.00	1882.39	2080.00	2605.64	3116.74	3145.67	1265.21	3598.91	3772.49	3926.78		
734.83	1890.00	1270.00	2615.28	3097.45	3155.31	1978.82	3598.91	1284.50	3936.42		

Table B.2 continued from previous page

1470.00	1892.03	1670.00	2615.28	2837.08	3164.96	2962.44	3598.91	1834.17	3936.42		
1853.46	1892.03	706.00	2624.93	2769.58	3174.60	2248.83	3608.55	2615.28	3936.42		
1467.72	1901.67	1440.00	2624.93	850.55	3184.24	2229.55	3618.19	3637.48	3936.42		
1853.46	1901.67	1840.00	2624.93	773.40	3193.89	1371.29	3627.84	1110.92	3946.07		

Table B.3: Wavenumbers of FTIR spectra with correlation < -0.9

609.46	638.39	918.05	2094.54	2065.61	2798.51	1014.48	3107.09	2277.76	3367.47	3425.33	3753.2
705.9	783.04	773.4	2113.83	898.76	2808.15	1535.23	3107.09	2287.41	3377.11	3039.59	3762.84
821.62	898.76	1149.49	2123.47	1785.95	2808.15	1911.32	3107.09	2239.19	3386.75	1149.49	3772.49
686.61	966.27	1477.37	2142.76	2711.72	2808.15	2962.44	3107.09	1747.38	3396.4	754.11	3782.13
792.69	1024.13	1882.39	2142.76	1699.16	2817.79	1390.57	3116.74	1747.38	3406.04	2846.72	3782.13
783.04	1053.06	802.33	2162.04	2731	2817.79	1940.25	3116.74	1747.38	3415.68	2306.7	3791.77
754.11	1081.99	927.69	2171.69	1776.31	2827.44	2740.65	3116.74	1737.74	3425.33	1448.43	3801.42
811.97	1101.27	648.04	2190.97	2750.29	2827.44	1303.78	3126.38	1737.74	3434.97	754.11	3811.06
1043.41	1130.2	956.62	2200.62	2171.69	2837.08	1785.95	3126.38	2065.61	3444.61	3097.45	3811.06
1062.7	1149.49	889.12	2229.55	2075.25	2846.72	2644.21	3126.38	1149.49	3454.26	2046.32	3820.7
1072.34	1178.42	869.83	2268.12	2200.62	2856.37	2663.5	3136.03	3280.68	3454.26	831.26	3830.35
879.48	1226.64	1448.43	2297.05	2470.63	2866.01	1920.96	3145.67	628.75	3473.54	3280.68	3830.35
763.76	1255.57	1487.01	2306.7	2432.06	2875.65	3010.66	3145.67	715.54	3483.19	2894.94	3839.99
773.4	1274.85	1882.39	2306.7	1188.06	2894.94	1409.86	3155.31	3107.09	3483.19	2335.63	3849.63
918.05	1294.14	2181.33	2316.34	1236.28	2904.58	1853.46	3155.31	3087.81	3492.83	1178.42	3859.28
1149.49	1313.43	2065.61	2335.63	1352	2914.23	2702.07	3155.31	2316.34	3502.47	1959.53	3868.92
1130.2	1332.71	1988.46	2354.91	1342.36	2923.87	1284.5	3164.96	918.05	3512.12	811.97	3878.56
995.2	1371.29	1419.5	2374.2	2210.26	2933.51	2162.04	3164.96	3251.75	3512.12	3261.39	3878.56
1159.13	1390.57	1930.6	2383.84	1737.74	2943.16	898.76	3174.6	3271.03	3521.76	2798.51	3888.21
918.05	1409.86	879.48	2412.77	1728.09	2952.8	1496.65	3174.6	3078.16	3531.4	956.62	3897.85
773.4	1429.15	956.62	2441.7	821.62	2972.09	1978.82	3174.6	3029.95	3541.05	792.69	3907.49
1371.29	1448.43	648.04	2470.63	1130.2	2981.73	2972.09	3174.6	1168.78	3550.69	3145.67	3907.49
715.54	1467.72	975.91	2480.28	1178.42	2991.37	1400.22	3184.24	628.75	3560.33	1458.08	3917.14
1178.42	1496.65	1149.49	2489.92	1178.42	3001.02	2557.42	3184.24	3078.16	3560.33	811.97	3926.78
1207.35	1515.94	628.75	2509.21	1159.13	3010.66	1062.7	3193.89	937.34	3569.98	3164.96	3926.78
1216.99	1544.87	1159.13	2518.85	2933.51	3020.3	1689.52	3193.89	773.4	3579.62	2808.15	3936.42
1506.3	1573.8	1458.08	2528.49	1641.3	3029.95	2499.56	3193.89	3126.38	3579.62	840.9	3946.07
1130.2	1612.37	783.04	2547.78	2624.93	3029.95	1448.43	3203.53	3319.25	3589.26	3261.39	3946.07
1197.71	1641.3	850.55	2557.42	1255.57	3039.59	2904.58	3213.17	2808.15	3598.91	3058.88	3955.71
1207.35	1670.23	715.54	2576.71	1650.95	3039.59	763.76	3232.46	1959.53	3608.55	1458.08	3965.35
802.33	1699.16	2142.76	2586.35	2576.71	3039.59	3126.38	3232.46	2634.57	3618.19	783.04	3975
937.34	1708.81	715.54	2615.28	1033.77	3049.23	898.76	3251.75	2538.14	3627.84	3145.67	3975
1573.8	1728.09	1226.64	2634.57	1544.87	3049.23	1496.65	3251.75	811.97	3637.48	2798.51	3984.64
1438.79	1747.38	2142.76	2644.21	1920.96	3049.23	1882.39	3251.75	3164.96	3637.48	811.97	3994.29
937.34	1766.67	1969.18	2663.5	2972.09	3049.23	2759.93	3251.75	3299.96	3647.12	3164.96	3994.29
783.04	1785.95	1757.02	2673.14	1332.71	3058.88	2123.47	3261.39	2769.58	3656.77		
1458.08	1795.6	2142.76	2682.79	1805.24	3058.88	1043.41	3271.03	792.69	3666.41		
1178.42	1824.53	918.05	2702.07	2653.86	3058.88	1554.51	3271.03	3155.31	3666.41		
715.54	1853.46	773.4	2721.36	1101.27	3068.52	1978.82	3271.03	2846.72	3676.05		
715.54	1882.39	1178.42	2731	1612.37	3068.52	3001.02	3271.03	2837.08	3685.7		
1737.74	1901.67	831.26	2750.29	2113.83	3068.52	2094.54	3280.68	1130.2	3695.34		
1178.42	1920.96	1053.06	2769.58	946.98	3078.16	1419.5	3290.32	3251.75	3695.34		
1159.13	1940.25	1679.88	2769.58	1535.23	3078.16	1371.29	3299.96	3058.88	3704.98		
879.48	1969.18	2181.33	2769.58	1911.32	3078.16	696.25	3309.61	927.69	3714.63		
725.18	1988.46	1053.06	2779.22	2981.73	3078.16	1766.67	3319.25	619.11	3724.27		
1747.38	1998.11	1940.25	2779.22	1535.23	3087.81	1728.09	3328.89	879.48	3733.91		
599.82	2027.04	705.9	2788.86	1930.6	3087.81	1236.28	3338.54	1834.17	3733.91		
1197.71	2046.32	1487.01	2788.86	946.98	3097.45	3222.82	3338.54	3377.11	3733.91		
792.69	2065.61	2499.56	2788.86	1564.16	3097.45	2451.35	3348.18	1930.6	3743.56		
937.34	2075.25	1081.99	2798.51	1978.82	3097.45	2403.13	3357.82	879.48	3753.2		

Table B.4: Chemical shifts of ^{1}H-NMR spectra with correlation > 0.9

-0.997	-0.956	3.031	3.642	3.316	5.595	6.042	7.548	4.415	9.460	4.415	11.943
-0.916	-0.549	1.650	3.682	4.171	5.595	1.119	7.589	5.717	9.460	7.874	11.943
-0.468	-0.427	2.584	3.682	5.554	5.595	2.218	7.629	8.158	9.460	9.745	11.943
-0.712	-0.305	3.560	3.682	4.293	5.676	4.049	7.629	5.432	9.542	8.606	11.983
-0.427	-0.224	3.070	3.723	2.136	5.717	5.595	7.629	2.340	9.583	9.054	12.024
-0.875	-0.142	2.050	3.764	3.601	5.717	2.096	7.670	3.235	9.583	5.147	12.065
-0.183	-0.102	3.230	3.764	4.659	5.717	2.991	7.670	4.049	9.583	7.019	12.065
-0.956	0.020	2.014	3.805	2.096	5.798	3.805	7.670	4.985	9.583	8.036	12.105
-0.997	0.102	2.909	3.805	2.991	5.798	4.659	7.670	7.914	9.583	11.292	12.105
-0.102	0.142	3.720	3.805	3.845	5.798	6.205	7.670	8.077	9.623	5.961	12.146
0.142	0.183	2.950	3.845	4.700	5.798	1.648	7.711	9.135	9.664	10.925	12.146
0.061	0.264	3.800	3.845	4.496	5.839	1.973	7.751	3.764	9.745	9.013	12.187
-0.956	0.346	2.750	3.886	4.008	5.880	2.869	7.751	4.740	9.745	4.822	12.227
-0.102	0.387	3.560	3.886	5.758	5.880	3.682	7.751	6.246	9.745	9.827	12.227
0.102	0.427	3.360	3.927	4.618	5.920	4.862	7.751	8.565	9.745	-0.509	12.309
0.264	0.468	2.050	3.967	4.537	5.961	1.933	7.792	8.158	9.827	10.763	12.309
0.346	0.509	2.991	3.967	3.764	6.002	3.031	7.792	4.252	9.867	8.972	12.390
0.387	0.549	3.805	3.967	5.147	6.002	3.886	7.792	8.077	9.867	2.502	12.431
0.264	0.590	3.153	4.008	3.438	6.083	4.700	7.792	9.745	9.867	6.246	12.431
0.061	0.631	1.892	4.049	5.595	6.083	6.612	7.792	3.031	9.908	7.589	12.431
-0.142	0.671	3.072	4.049	2.747	6.124	3.194	7.833	3.886	9.908	11.861	12.472
-0.956	0.712	3.927	4.049	3.601	6.124	4.049	7.833	4.700	9.908	5.107	12.594
0.671	0.712	3.030	4.089	4.618	6.124	4.862	7.833	7.792	9.908	8.525	12.594
-0.549	0.793	3.890	4.089	1.811	6.165	7.670	7.833	5.229	9.949	0.061	12.675
0.793	0.875	2.665	4.130	3.113	6.165	3.601	7.874	3.357	10.071	7.385	12.716
0.793	0.956	3.560	4.130	3.967	6.165	4.659	7.874	4.333	10.071	-0.916	12.878
-0.549	1.040	3.276	4.171	2.502	6.205	6.002	7.874	8.077	10.071	12.349	12.878
0.916	1.080	4.130	4.171	4.130	6.205	3.194	7.914	9.867	10.071		
-0.916	1.160	2.460	4.211	5.920	6.205	4.049	7.914	8.565	10.111		
0.793	1.200	3.276	4.211	2.747	6.246	4.862	7.914	2.625	10.152		
0.916	1.280	4.090	4.211	3.560	6.246	7.833	7.914	6.205	10.152		
1.038	1.360	3.190	4.252	4.700	6.246	3.805	7.955	7.548	10.152		
1.404	1.530	4.049	4.252	2.258	6.287	4.740	7.955	3.560	10.274		
1.445	1.690	2.502	4.293	4.415	6.287	6.246	7.955	4.496	10.274		
1.648	1.770	3.360	4.293	6.246	6.287	1.933	7.996	7.467	10.274		
1.811	1.851	4.211	4.293	2.258	6.327	2.828	7.996	8.931	10.274		
1.770	1.973	3.438	4.333	3.194	6.327	3.642	7.996	5.920	10.356		
1.690	2.050	4.293	4.333	4.903	6.327	4.496	7.996	7.548	10.396		
2.055	2.096	3.110	4.374	6.002	6.368	5.920	7.996	2.828	10.518		
2.014	2.177	4.089	4.374	6.246	6.409	7.833	7.996	4.374	10.518		
1.811	2.260	3.600	4.415	6.287	6.449	4.130	8.036	5.961	10.518		
1.973	2.299	1.970	4.456	2.177	6.490	7.833	8.036	7.670	10.518		
2.055	2.340	2.950	4.456	3.113	6.490	4.293	8.077	4.985	10.559		
1.648	2.502	3.760	4.456	4.659	6.490	8.036	8.077	7.019	10.559		
1.485	2.543	2.299	4.496	2.462	6.531	6.490	8.118	8.077	10.600		
2.462	2.543	3.642	4.496	3.276	6.531	3.967	8.158	5.391	10.640		
2.177	2.584	2.096	4.537	4.415	6.531	6.083	8.158	7.711	10.681		
1.933	2.625	3.357	4.537	6.124	6.531	3.479	8.199	1.038	10.763		
1.648	2.665	4.250	4.537	1.973	6.571	4.415	8.199	5.920	10.763		
2.543	2.665	2.140	4.578	3.113	6.571	5.880	8.199	10.559	10.763		
2.136	2.706	3.030	4.578	1.933	6.612	5.758	8.240	2.909	10.925		
1.770	2.747	3.850	4.578	2.828	6.612	8.077	8.280	4.252	10.925		
2.665	2.747	3.190	4.618	3.642	6.612	2.991	8.321	5.798	10.925		
2.218	2.790	4.330	4.618	4.618	6.612	3.845	8.321	7.792	10.925		
1.811	2.828	2.180	4.659	6.124	6.612	4.659	8.321	-0.793	10.966		
2.706	2.828	3.070	4.659	4.903	6.653	6.205	8.321	0.956	10.966		
2.177	2.869	3.890	4.659	-0.916	6.694	4.333	8.362	8.728	11.007		
1.729	2.909	1.890	4.700	0.916	6.734	3.316	8.403	9.664	11.047		
2.625	2.909	3.030	4.700	5.798	6.734	4.252	8.403	3.153	11.088		

Table B.4 continued from previous page

2.014	2.950	3.886	4.700	4.944	6.775	5.717	8.403	4.008	11.088	
2.909	2.950	1.811	4.740	1.200	6.816	8.280	8.403	5.880	11.088	
2.462	2.991	2.869	4.740	2.828	6.816	2.869	8.443	8.769	11.088	
1.851	3.031	3.764	4.740	6.694	6.816	3.682	8.443	9.216	11.210	
2.747	3.031	4.700	4.740	6.409	6.856	4.578	8.443	9.257	11.292	
2.136	3.072	4.089	4.781	1.200	6.897	8.036	8.443	9.623	11.332	
3.031	3.072	2.014	4.822	2.543	6.938	2.787	8.525	0.509	11.373	
2.340	3.113	3.190	4.822	5.025	6.938	5.025	8.525	2.665	11.414	
1.729	3.153	4.010	4.822	6.856	6.938	6.734	8.525	3.601	11.414	
2.620	3.150	2.340	4.862	2.543	6.978	4.089	8.565	4.659	11.414	
1.970	3.190	3.280	4.862	1.200	7.019	5.798	8.565	9.583	11.414	
2.869	3.194	4.290	4.862	2.665	7.019	8.403	8.565	5.066	11.454	
2.218	3.230	1.810	4.903	5.920	7.019	7.792	8.606	6.897	11.454	
3.113	3.230	2.710	4.903	1.445	7.060	4.496	8.687	5.514	11.495	
2.340	3.280	3.520	4.903	2.869	7.060	8.036	8.687	8.809	11.495	
3.230	3.276	4.500	4.903	6.612	7.060	5.758	8.728	6.449	11.536	
2.828	3.320	2.790	4.944	1.322	7.100	1.973	8.769	11.332	11.576	
2.050	3.357	4.860	4.944	2.869	7.100	3.601	8.769	1.322	11.617	
2.991	3.357	3.110	4.985	-0.916	7.141	4.537	8.769	2.177	11.658	
2.258	3.398	4.500	4.985	6.694	7.141	8.321	8.769	3.276	11.658	
3.276	3.398	2.010	5.025	2.421	7.263	5.514	8.850	4.130	11.658	
2.787	3.438	2.909	5.025	1.729	7.426	5.391	8.891	4.985	11.658	
1.973	3.479	3.723	5.025	2.665	7.426	4.781	8.931	7.792	11.658	
2.909	3.479	4.659	5.025	4.211	7.426	8.565	8.931	10.274	11.658	
2.014	3.520	4.862	5.066	2.258	7.467	8.606	9.013	3.398	11.698	
2.950	3.520	3.031	5.107	3.276	7.467	8.728	9.135	4.252	11.698	
1.973	3.560	2.706	5.147	4.130	7.467	8.809	9.216	6.083	11.698	
2.869	3.560	4.618	5.147	4.985	7.467	4.496	9.338	8.565	11.698	
1.933	3.601	4.171	5.473	2.462	7.507	8.158	9.338	9.013	11.739	
3.031	3.601	3.886	5.554	4.578	7.507	-0.509	9.420	5.636	11.820	
2.014	3.642	4.985	5.554	6.531	7.507	3.520	9.460	11.292	11.820	

Table B.5: Chemical shifts of ^{1}H-NMR spectra with correlation < -0.9

1.322	2.380	2.787	7.222	0.387	9.298	0.183	11.129	9.501	12.756
1.445	2.421	5.107	7.222	-0.712	9.664	0.387	11.292	12.716	12.797
2.421	2.665	1.689	7.263	1.322	9.949	0.346	11.332		
-0.875	5.229	2.625	7.263	2.828	9.949	-0.509	11.576		
1.729	5.229	3.642	7.263	-0.671	10.030	0.509	11.576		
1.363	5.269	1.526	7.304	0.346	10.030	0.590	11.739		
1.322	5.310	2.625	7.304	9.786	10.437	0.387	11.820		
0.142	5.351	6.490	7.304	0.224	10.478	0.387	11.983		
0.509	5.391	7.345	7.426	0.305	10.600	0.509	12.105		
5.229	6.571	7.263	7.996	0.102	10.640	10.437	12.349		
5.269	7.019	0.305	8.362	5.269	10.681	10.234	12.634		
1.770	7.182	0.142	8.891	0.061	10.722	11.576	12.675		
2.747	7.182	0.305	9.013	7.222	10.763	3.072	12.756		
5.025	7.182	0.264	9.135	1.160	10.844	3.927	12.756		
1.770	7.222	0.346	9.257	0.305	11.007	5.880	12.756		

Table B.6: Wavenumbers of FTIR and chemical shifts of ^{1}H-NMR spectra with cross-correlation > 0.7

750.25	-1.00	3299.96	1.85	2960.52	3.44	3230.53	5.35	1409.86	8.24	1035.70	10.60
850.55	-0.96	1580.00	1.89	1487.01	3.48	1093.56	5.39	2096.47	8.24	1317.28	10.60
607.53	-0.92	2659.64	1.89	1679.88	3.48	2092.61	5.39	2725.22	8.24	1845.74	10.60
2864.08	-0.92	3508.26	1.89	2655.78	3.48	2752.22	5.39	3596.98	8.24	2582.49	10.60

Table B.6 continued from previous page

2335.62	-0.88	1620.00	1.93	892.98	3.52	3728.13	5.39	3932.57	8.24	3454.26	10.60
935.41	-0.83	2650.00	1.93	1645.16	3.52	892.98	5.59	1610.44	8.32	3789.85	10.60
619.10	-0.79	3470.00	1.93	2879.51	3.52	1633.59	5.59	3481.26	8.32	1240.14	10.64
2933.51	-0.79	1614.30	1.97	885.26	3.56	1984.61	5.59	1247.85	8.36	2559.35	10.64
2324.05	-0.71	2678.93	1.97	1637.44	3.56	3504.40	5.59	1780.17	8.36	2910.37	10.68
854.40	-0.63	3490.00	1.97	2648.07	3.56	1247.85	5.64	2513.06	8.36	2486.06	10.72
2856.37	-0.63	1900.00	2.01	3492.83	3.56	1687.59	5.64	2983.66	8.36	3890.14	10.72
2362.63	-0.59	3300.00	2.01	1579.58	3.60	2096.47	5.64	3728.13	8.36	3338.54	10.76
2366.48	-0.55	1550.00	2.05	1880.46	3.60	2717.50	5.64	997.12	8.44	2262.34	10.84
2042.47	-0.51	2660.00	2.05	2659.64	3.60	3593.12	5.64	1629.73	8.44	931.55	10.97
777.26	-0.39	689.00	2.10	881.40	3.64	3948.00	5.64	1934.46	8.44	900.69	11.01
2042.47	-0.39	1640.00	2.10	1618.16	3.64	1444.58	5.72	2894.94	8.44	1325.00	11.01
669.25	-0.35	2879.51	2.10	1957.60	3.64	1645.16	5.72	1081.99	8.48	2177.47	11.01
2374.20	-0.35	665.39	2.14	2960.52	3.64	1911.32	5.72	2115.76	8.48	2987.52	11.01
3280.00	-0.31	1610.44	2.14	1463.86	3.68	2655.78	5.72	3531.40	8.48	3789.85	11.01
2864.08	-0.26	2632.64	2.14	3388.68	3.68	707.82	5.76	3948.00	8.48	1000.98	11.09
973.98	-0.18	3442.68	2.14	1525.58	3.72	1309.57	5.76	1070.41	8.61	1587.30	11.09
3311.53	-0.18	1568.01	2.18	1814.88	3.72	2732.93	5.76	1429.15	8.61	1849.60	11.09
2350.00	-0.14	1950.00	2.18	2574.78	3.72	3955.71	5.76	1656.73	8.61	2613.35	11.09
862.12	-0.10	3380.00	2.18	3485.11	3.72	3539.12	5.88	1903.60	8.61	3500.54	11.09
3141.81	-0.10	1483.15	2.22	1564.15	3.76	1290.28	6.08	2605.64	8.61	1112.85	11.29
923.84	-0.06	1876.60	2.22	1865.03	3.76	1687.59	6.08	2987.52	8.61	1703.02	11.29
2852.51	-0.06	2945.09	2.22	2682.79	3.76	2084.90	6.08	3681.84	8.61	2524.63	11.29
854.40	-0.02	893.00	2.26	1452.29	3.80	3454.26	6.08	946.98	8.77	3002.95	11.29
3265.25	-0.02	1960.00	2.26	2594.07	3.80	730.97	6.16	1514.01	8.77	3774.42	11.29
1140.00	0.02	3407.97	2.26	3465.83	3.80	1606.59	6.16	1687.59	8.77	703.97	11.33
3284.53	0.02	1600.00	2.30	1610.44	3.85	1903.60	6.16	1984.61	8.77	1429.15	11.33
1143.70	0.06	2880.00	2.30	2628.78	3.85	2690.50	6.16	2713.64	8.77	2524.63	11.33
3261.39	0.06	889.12	2.34	1467.72	3.89	889.12	6.33	1012.55	8.85	3569.98	11.33
1135.99	0.10	2069.47	2.34	1024.13	3.93	2945.09	6.33	1402.15	8.85	3909.42	11.33
3292.25	0.10	3420.00	2.34	1514.01	3.93	1575.73	6.49	2729.07	8.85	1490.87	11.41
1124.42	0.14	1992.32	2.38	1687.59	3.93	3396.40	6.49	3704.98	8.85	1672.16	11.41
3122.52	0.14	2435.92	2.38	1980.75	3.93	2887.23	6.57	1024.13	8.89	2536.21	11.41
916.00	0.18	1236.28	2.42	2705.93	3.93	1444.58	6.61	1305.71	8.89	3434.97	11.41
2856.37	0.18	2300.00	2.42	3936.42	3.93	607.53	6.69	1880.46	8.89	2189.05	11.58
846.69	0.22	3855.42	2.42	1614.30	3.97	2925.80	6.69	2586.35	8.89	3924.85	11.58
2833.22	0.22	2910.00	2.50	2644.21	3.97	2937.37	6.82	3512.12	8.89	2933.51	11.62
835.12	0.26	1460.00	2.54	3504.40	3.97	3342.39	6.90	3793.70	8.89	1548.73	11.66
2813.94	0.26	3430.00	2.54	1529.44	4.01	603.68	7.02	1039.56	9.01	1872.74	11.66
808.11	0.31	2910.00	2.58	1695.30	4.01	3365.54	7.02	1413.72	9.01	2690.50	11.66
2783.08	0.31	1448.43	2.62	2077.18	4.01	2941.23	7.06	2729.07	9.01	1008.70	11.70
765.68	0.35	3299.96	2.62	3434.97	4.01	2675.07	7.10	3693.41	9.01	1602.73	11.70
1170.71	0.35	1957.60	2.67	1533.30	4.05	877.55	7.14	1390.57	9.13	1861.17	11.70
3265.25	0.35	3384.82	2.67	1930.60	4.05	1359.72	7.18	2972.09	9.13	2628.78	11.70
935.41	0.39	2880.00	2.71	3481.26	4.05	2293.19	7.18	703.97	9.26	900.69	11.74
3122.52	0.39	889.00	2.75	2536.21	4.09	3816.85	7.18	1398.29	9.26	2115.76	11.74
850.55	0.43	3350.00	2.75	1579.58	4.13	1969.18	7.22	2189.05	9.26	3577.69	11.74
2829.37	0.43	2670.00	2.79	2644.21	4.13	2462.92	7.22	2995.23	9.26	615.25	11.82
831.26	0.47	689.00	2.83	1305.71	4.17	1224.71	7.26	3774.42	9.26	2169.76	11.82
2817.79	0.47	2940.00	2.83	1641.30	4.17	2297.05	7.26	1776.31	9.30	3654.84	11.82
823.54	0.51	1470.00	2.87	1915.17	4.17	3836.13	7.26	1051.13	9.50	1008.70	11.98
2802.36	0.51	3360.00	2.87	3469.69	4.17	2281.62	7.30	1514.01	9.50	1290.28	11.98
811.97	0.55	2050.00	2.91	1587.30	4.21	746.40	7.43	1865.03	9.50	1811.03	11.98
2813.94	0.55	3430.00	2.91	2648.07	4.21	3234.39	7.43	2698.22	9.50	2559.35	11.98
854.40	0.59	2680.00	2.95	3465.83	4.21	1564.15	7.47	3604.69	9.50	3006.80	11.98
3292.25	0.59	881.40	2.99	2594.07	4.25	2590.21	7.47	1552.58	9.58	3778.27	11.98
1170.71	0.63	2034.75	2.99	1575.73	4.29	865.97	7.59	1980.75	9.58	1016.41	12.11
927.69	0.67	3407.97	2.99	1980.75	4.29	2046.32	7.67	3396.40	9.58	1298.00	12.11
2351.05	0.71	2536.21	3.03	3461.97	4.29	746.40	7.75	1043.41	9.62	2000.04	12.11
2362.63	0.96	3442.68	3.03	1568.01	4.33	2046.32	7.75	1325.00	9.62	2709.79	12.11

Table B.6 continued from previous page

970.12	1.00	1895.89	3.07	1853.46	4.33	3415.68	7.75	1911.32	9.62	3635.55	12.11
3346.25	1.00	3299.96	3.07	2648.07	4.33	1544.87	7.79	2624.93	9.62	3940.28	12.11
2366.48	1.04	1570.00	3.11	1444.58	4.46	1853.46	7.79	3539.12	9.62	1066.56	12.19
966.27	1.08	2880.00	3.11	1645.16	4.46	2648.07	7.79	3920.99	9.62	1402.15	12.19
3360.00	1.08	885.00	3.15	2084.90	4.46	889.12	8.00	1062.70	9.83	1849.60	12.19
2920.00	1.12	1930.00	3.15	3407.97	4.46	2894.94	8.00	1398.29	9.83	2555.49	12.19
2929.66	1.16	2960.00	3.15	1606.59	4.58	1004.84	8.04	2092.61	9.83	2983.66	12.19
2864.08	1.32	1440.00	3.19	3380.97	4.58	1301.86	8.04	2744.50	9.83	3704.98	12.19
2038.61	1.40	2880.00	3.19	1587.30	4.66	1606.59	8.04	3658.70	9.83	892.98	12.23
989.41	1.44	1470.00	3.23	2675.07	4.66	1857.31	8.04	3994.28	9.83	1267.14	12.23
3380.00	1.44	1900.00	3.23	3508.26	4.66	2601.78	8.04	1583.44	9.91	1548.73	12.23
2933.51	1.49	2950.00	3.23	1595.01	4.70	3473.54	8.04	1865.03	9.91	1814.88	12.23
1450.00	1.53	1450.00	3.28	1895.89	4.70	3971.14	8.04	2640.36	9.91	2100.33	12.23
3380.97	1.53	1980.00	3.28	2690.50	4.70	1089.70	8.08	711.68	9.95	2721.36	12.23
2914.23	1.57	3400.00	3.28	1440.72	4.82	1502.44	8.08	2266.19	9.95	3531.40	12.23
1463.86	1.61	1580.00	3.32	1641.30	4.82	1799.45	8.08	3589.26	9.95	3955.71	12.23
3388.68	1.61	2883.37	3.32	1949.89	4.82	2084.90	8.08	2181.33	10.03	2362.63	12.31
2891.08	1.65	1529.44	3.36	2952.80	4.82	2759.93	8.08	3890.14	10.03	804.26	12.67
889.12	1.69	1895.89	3.36	1463.86	4.90	3701.13	8.08	1108.99	10.07	2786.93	12.67
3377.11	1.69	2956.66	3.36	3388.68	4.90	1043.41	8.16	1533.30	10.07	1213.14	12.76
2883.37	1.73	1467.72	3.40	2879.51	5.03	1490.87	8.16	1795.60	10.07	3161.10	12.76
885.26	1.77	1656.73	3.40	1379.00	5.23	1676.02	8.16	2050.18	10.07		
2941.23	1.77	2636.50	3.40	2428.20	5.23	1953.75	8.16	2736.79	10.07		
1448.43	1.81	665.39	3.44	1236.28	5.27	2729.07	8.16	3662.55	10.07		
3361.68	1.81	1606.59	3.44	2420.49	5.27	3697.27	8.16	2007.75	10.48		
1602.73	1.85	1946.03	3.44	676.96	5.35	1074.27	8.24	3677.98	10.48		

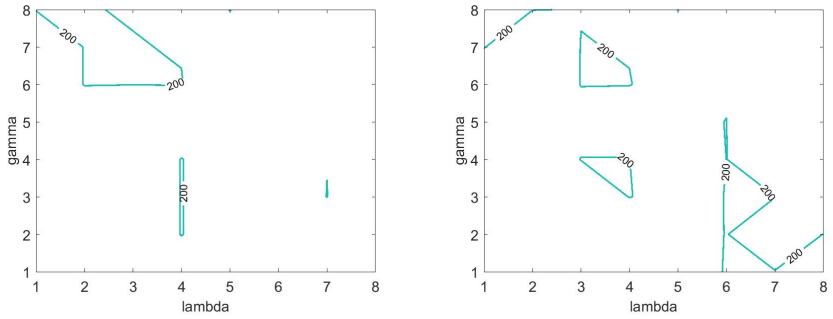
Table B.7: Wavenumbers of FTIR and chemical shifts of ^{1}H-NMR spectra with cross-correlation < -0.7

615.247	-0.997	3569.976	0.468	1517.867	2.218	2281.622	3.682	3870.849	7.100	1182.277	11.088
3874.707	-0.997	676.965	0.509	3211.243	2.218	3851.563	3.682	2455.202	7.141	3091.665	11.088
3010.661	-0.956	2111.898	0.509	1517.867	2.258	3072.378	3.723	2069.468	7.182	808.115	11.292
2023.179	-0.916	3002.946	0.509	3249.816	2.258	1213.136	3.764	3427.254	7.182	2806.221	11.292
622.962	-0.875	3920.995	0.509	1718.449	2.299	3735.842	3.764	2910.370	7.222	781.113	11.332
2470.632	-0.875	2092.612	0.549	3245.959	2.299	2003.893	3.805	1448.434	7.263	2019.322	11.332
1988.463	-0.834	3600.835	0.549	1436.862	2.340	715.538	3.845	3380.966	7.263	3307.677	11.332
3832.276	-0.834	1699.162	0.590	3161.097	2.340	3238.244	3.845	2887.226	7.304	3072.378	11.414
2459.060	-0.793	3670.267	0.590	1452.292	2.380	1189.992	3.927	1236.280	7.426	854.403	11.576
2007.750	-0.712	1776.309	0.631	3384.824	2.380	3110.952	3.927	2412.772	7.426	657.678	11.617
3758.986	-0.712	3631.694	0.631	2925.799	2.421	1521.724	3.967	1201.564	7.467	2451.345	11.617
2189.045	-0.631	2127.328	0.671	1236.280	2.502	3735.842	3.967	3751.272	7.467	1436.862	11.658
642.249	-0.590	1386.717	0.712	2389.628	2.502	3049.234	4.008	3546.832	7.589	3743.557	11.658
3882.421	-0.590	692.394	0.956	3901.708	2.502	1197.707	4.049	1969.177	7.670	3083.950	11.698
3708.841	-0.549	3172.669	0.956	2262.335	2.543	1189.992	4.089	1213.136	7.751	819.687	11.739
2470.632	-0.509	2150.472	0.997	3627.836	2.543	1216.994	4.130	2393.485	7.751	719.396	11.820
1240.138	-0.387	3805.275	0.997	1733.878	2.584	1189.992	4.171	673.107	7.792	1151.419	11.820
2486.061	-0.387	2439.773	1.038	3026.090	2.584	3272.960	4.171	3095.522	7.792	3265.246	11.820
3720.413	-0.387	1344.286	1.078	1213.136	2.625	1749.308	4.211	1714.592	7.996	1556.440	11.983
1768.594	-0.346	2466.775	1.078	2293.194	2.625	3743.557	4.211	3566.119	7.996	3176.527	11.983
3797.560	-0.346	1942.175	1.119	3855.420	2.625	1205.422	4.293	3049.234	8.036	919.978	12.105
3797.560	-0.305	3608.550	1.119	1992.321	2.665	3731.985	4.293	1521.724	8.077	3122.524	12.105
2462.917	-0.264	2462.917	1.160	3731.985	2.665	3107.094	4.333	3272.960	8.077	1139.847	12.187
1768.594	-0.183	2219.904	1.322	1737.736	2.706	1737.736	4.456	3103.237	8.158	3103.237	12.187
3758.986	-0.183	3612.407	1.322	3195.813	2.706	3585.406	4.456	823.544	8.240	1556.440	12.227
2192.903	-0.142	2262.335	1.404	1363.573	2.747	1749.308	4.578	3049.234	8.240	3180.384	12.227
615.247	-0.102	3816.847	1.404	2405.057	2.747	3839.991	4.578	2304.766	8.321	2239.191	12.309
2497.633	-0.102	2277.764	1.445	3870.849	2.747	1749.308	4.659	1178.420	8.362	3905.566	12.309
3924.852	-0.102	3816.847	1.445	2254.620	2.787	3743.557	4.659	3191.956	8.362	2123.470	12.675

Table B.7 continued from previous page

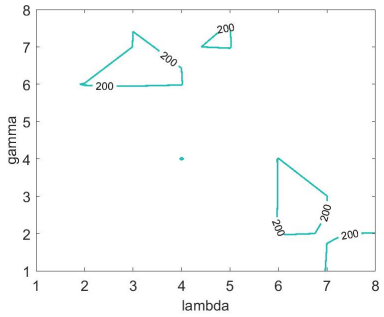
2162.044	-0.061	2243.048	1.485	3855.420	2.787	3026.090	4.700	3029.947	8.443	3577.691	12.675
3596.978	-0.061	3211.243	1.485	2003.893	2.828	1193.849	4.822	754.112	8.484	734.825	12.756
904.548	-0.020	1730.021	1.526	3820.704	2.828	3103.237	4.822	2802.364	8.484	1494.723	12.756
2547.779	-0.020	2416.629	1.526	1753.165	2.869	1232.423	4.903	1560.298	8.606	1676.018	12.756
676.965	0.020	680.822	1.567	3616.264	2.869	3022.233	4.903	3191.956	8.606	1980.749	12.756
2185.188	0.020	2258.478	1.567	1714.592	2.909	1359.715	5.025	3049.234	8.769	2952.800	12.756
3782.130	0.020	3245.959	1.567	3218.958	2.909	3735.842	5.025	916.120	8.850		
1706.877	0.061	1726.164	1.607	1359.715	2.950	2933.514	5.229	796.543	8.891		
2748.361	0.061	2412.772	1.607	3095.522	2.950	2667.357	5.269	2817.793	8.891		
3886.279	0.061	3901.708	1.607	1209.279	2.991	862.117	5.351	1132.132	9.013		
2165.901	0.102	2246.906	1.648	2397.342	2.991	3265.246	5.351	808.115	9.135		
3720.413	0.102	3222.815	1.648	1120.560	3.031	1132.132	5.391	2833.223	9.135		
1394.432	0.142	1375.145	1.689	2385.770	3.031	3164.955	5.391	850.545	9.257		
2505.348	0.142	2393.485	1.689	680.822	3.072	3161.097	5.595	2833.223	9.257		
3693.411	0.142	3901.708	1.689	2142.757	3.072	1178.420	5.636	2829.365	9.298		
1240.138	0.183	2235.334	1.729	3836.133	3.072	3126.381	5.636	3276.818	9.501		
2505.348	0.183	3242.102	1.729	1749.308	3.113	2304.766	5.717	2003.893	9.583		
3878.564	0.183	1359.715	1.770	3585.406	3.113	3747.414	5.717	3839.991	9.583		
1922.888	0.224	2412.772	1.770	1517.867	3.153	3107.094	5.880	2790.792	9.623		
2968.230	0.224	1201.564	1.811	3203.528	3.153	3199.671	6.083	719.396	9.827		
3874.707	0.224	2385.770	1.811	1510.152	3.194	3037.662	6.165	3041.519	9.827		
1429.148	0.264	3870.849	1.811	3731.985	3.194	1213.136	6.327	1216.994	9.908		
2543.922	0.264	2003.893	1.851	1741.593	3.235	2408.914	6.327	3188.099	9.908		
3720.413	0.264	3801.417	1.851	3801.417	3.235	1197.707	6.490	2879.511	9.949		
1078.129	0.305	1737.736	1.892	1745.450	3.276	2393.485	6.490	811.972	10.030		
2000.035	0.305	3249.816	1.892	3743.557	3.276	626.819	6.571	2856.367	10.030		
2725.217	0.305	1510.152	1.933	3195.813	3.316	2250.763	6.571	3064.664	10.071		
3685.697	0.305	3195.813	1.933	1737.736	3.357	3242.102	6.571	827.401	10.478		
3986.570	0.305	1363.573	1.973	3751.272	3.357	1733.878	6.612	2825.508	10.478		
1703.019	0.346	3095.522	1.973	3033.805	3.398	1379.002	6.694	819.687	10.600		
2532.349	0.346	1213.136	2.014	1220.851	3.438	3527.545	6.694	3056.949	10.600		
3639.409	0.346	2389.628	2.014	3735.842	3.438	2258.478	6.816	846.688	10.640		
3951.854	0.346	3870.849	2.014	3033.805	3.479	3863.135	6.816	2852.509	10.640		
1919.031	0.387	1969.177	2.055	1216.994	3.520	2451.345	6.897	757.969	10.722		
2597.924	0.387	3735.842	2.055	3195.813	3.520	2266.192	6.978	2316.338	10.722		
3782.130	0.387	1722.306	2.096	1517.867	3.560	1359.715	7.019	2216.047	10.763		
1332.714	0.427	3245.959	2.096	3211.243	3.560	2389.628	7.019	3546.832	10.763		
2185.188	0.427	1541.011	2.136	1737.736	3.601	3901.708	7.019	2061.753	10.844		
3635.551	0.427	3238.244	2.136	3751.272	3.601	2285.479	7.060	700.109	10.966		
3959.568	0.427	1521.724	2.177	2300.909	3.642	3870.849	7.060	3832.276	10.966		
2119.613	0.468	3211.243	2.177	680.822	3.682	2293.194	7.100	1170.705	11.007		

B.3 For $\alpha = 10^{-3}$

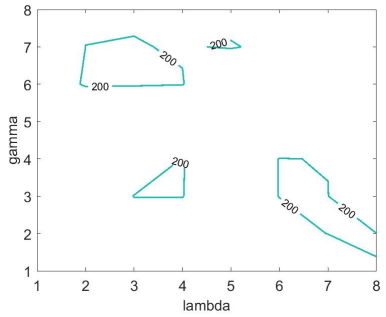


(a) $\beta = 10^{-3}$

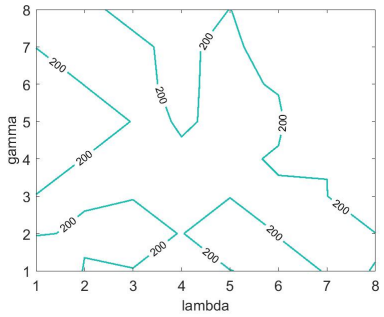
(b) $\beta = 10^{-2}$



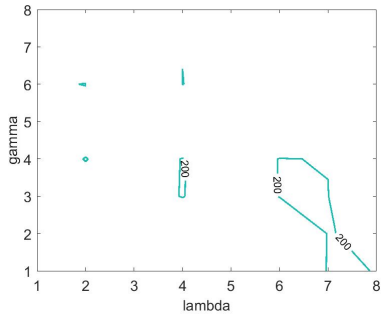
(c) $\beta = 10^{-1}$



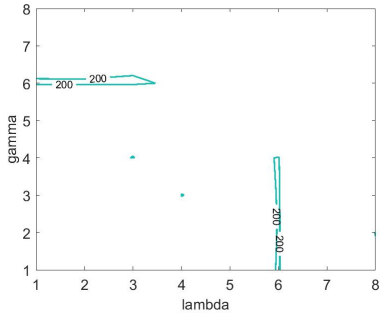
(d) $\beta = 10^{0}$



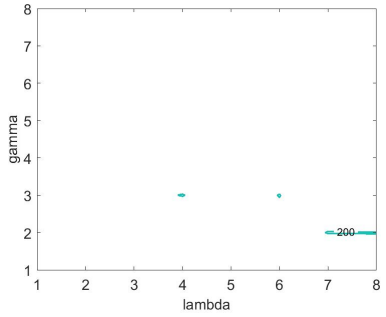
(e) $\beta = 0$



(f) $\beta = 10^{1}$

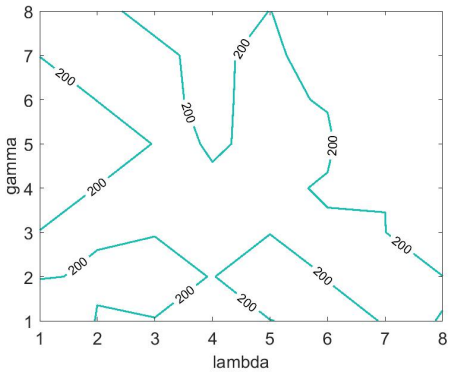


(g) $\beta = 10^{2}$

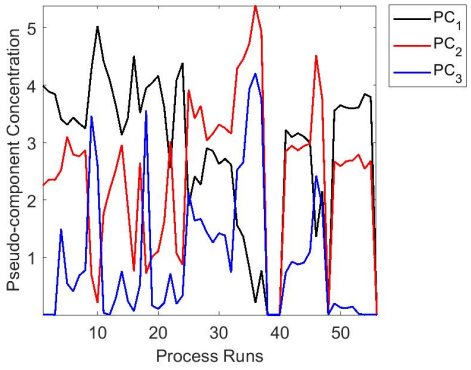


(h) $\beta = 10^{3}$

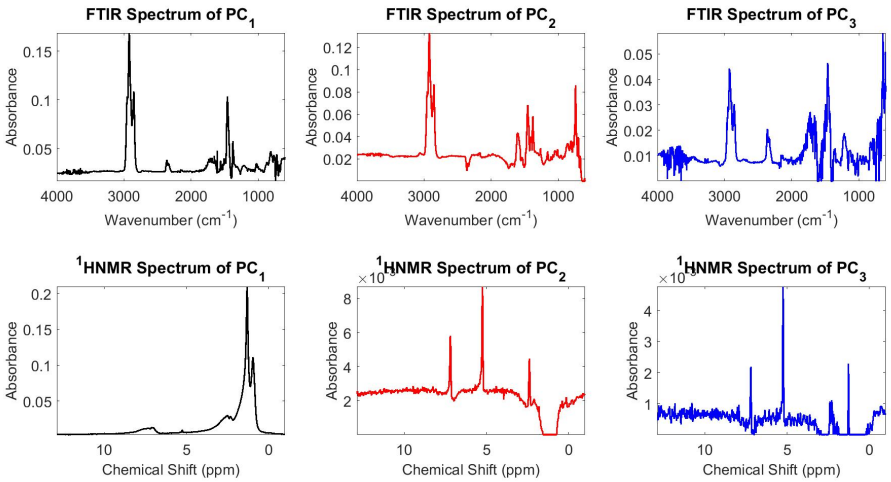
Figure B.1: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 0$



(b) Concentration



(c) Pseudo-component spectra

Figure B.2: JNMF profiles for $\alpha = 10^{-3}, \beta = 0, \gamma = 0, \lambda = 10^{-1}$

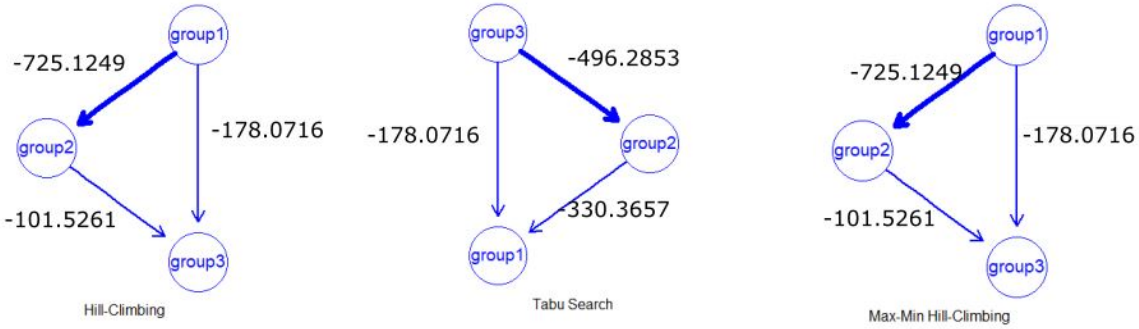
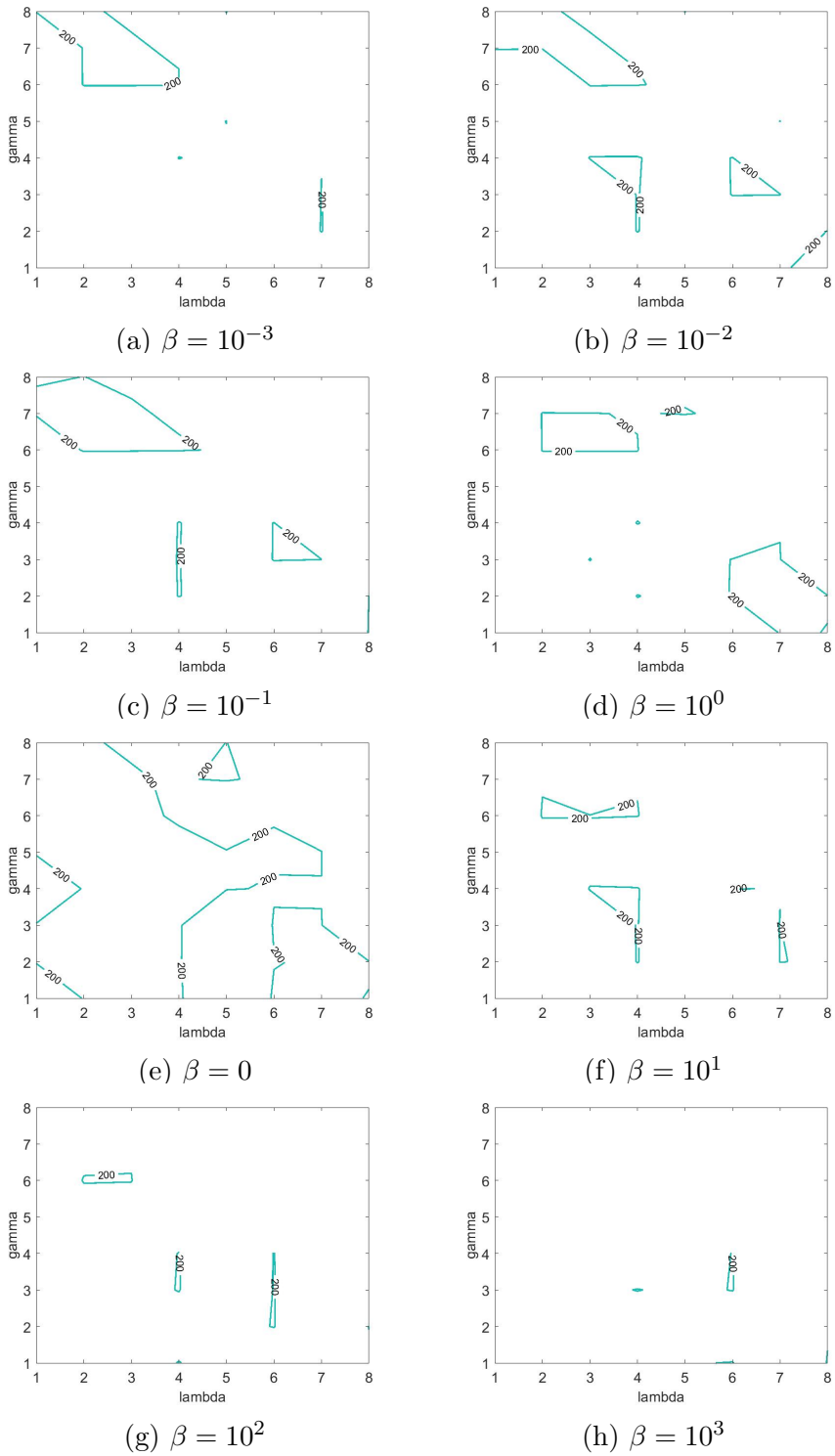


Figure B.3: Bayesian networks constructed from the PC spectra

B.4 For $\alpha = 10^{-2}$



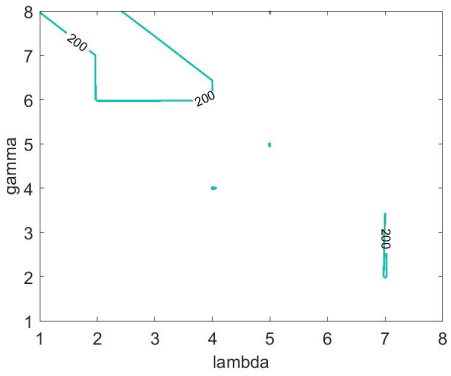
(a) $\beta = 10^{-3}$ (b) $\beta = 10^{-2}$

(c) $\beta = 10^{-1}$ (d) $\beta = 10^{0}$

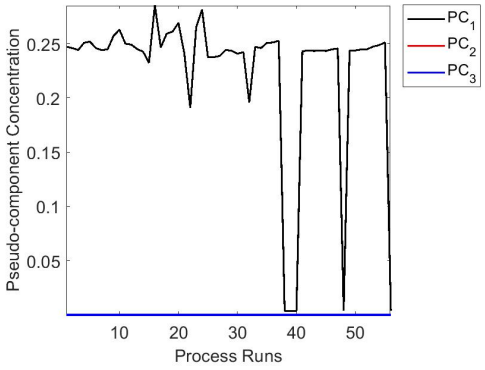
(e) $\beta = 0$ (f) $\beta = 10^{1}$

(g) $\beta = 10^{2}$ (h) $\beta = 10^{3}$

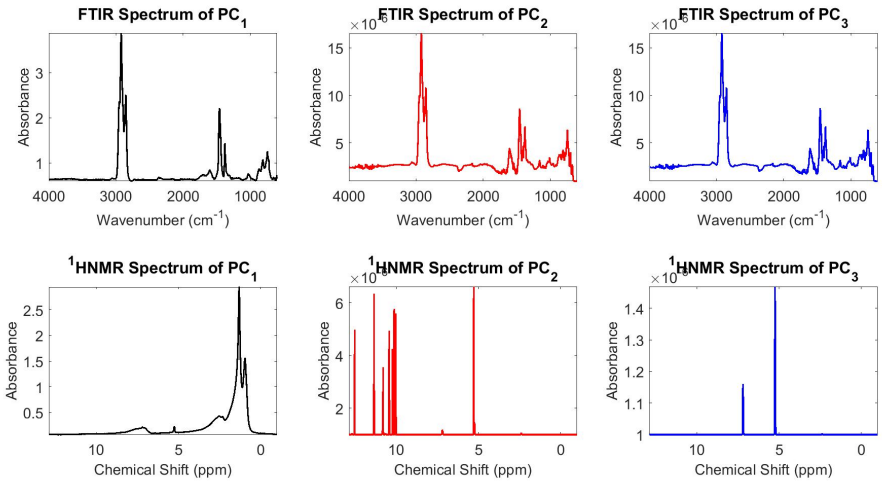
Figure B.4: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 10^{-3}$



(b) Concentration



(c) Pseudo-component spectra

Figure B.5: JNMF profiles for $\alpha = 10^{-2}, \beta = 10^{-3}, \gamma = 10^{1}, \lambda = 10^{-2}$

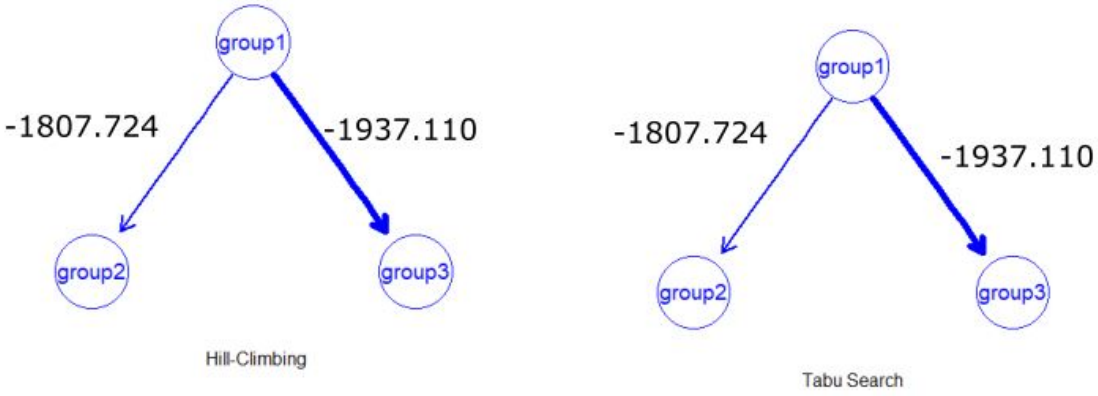
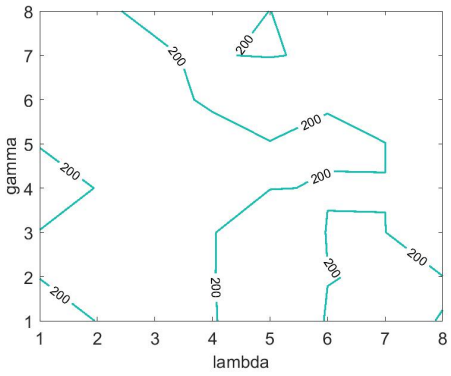
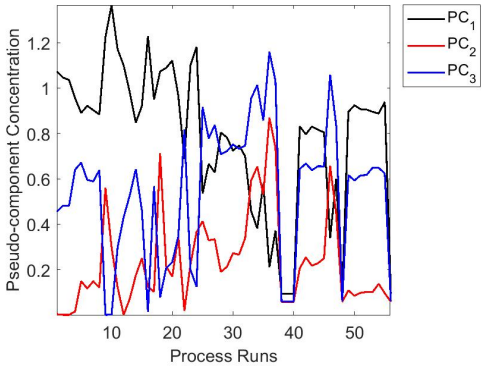


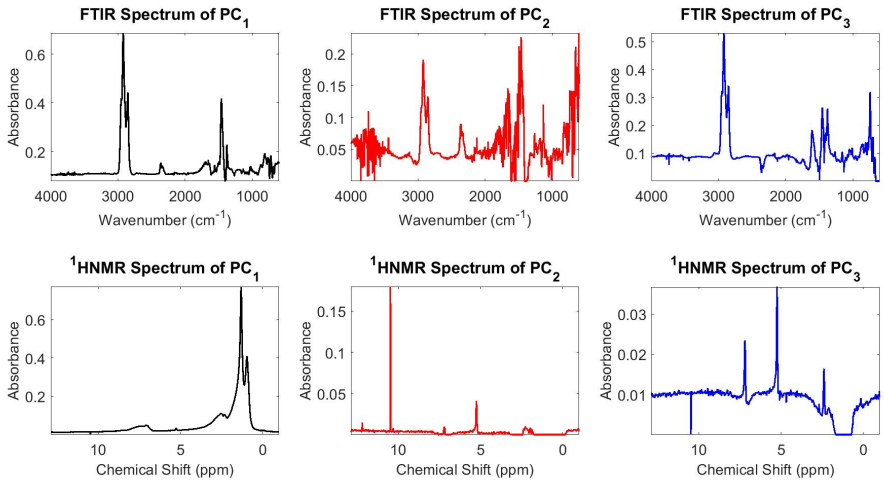
Figure B.6: Bayesian networks constructed from the PC spectra



(a) $\beta = 0$



(b) Concentration



(c) Pseudo-component spectra

Figure B.7: JNMF profiles for $\alpha = 10^{-2}, \beta = 0, \gamma = 10^{-1}, \lambda = 10^{-2}$

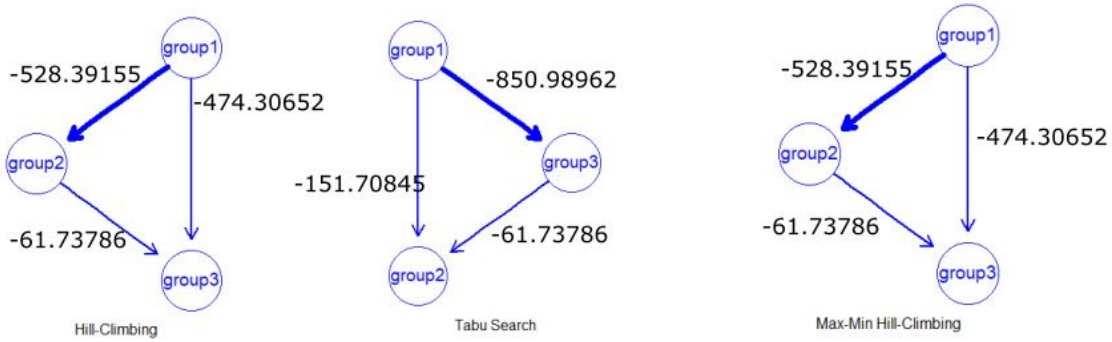
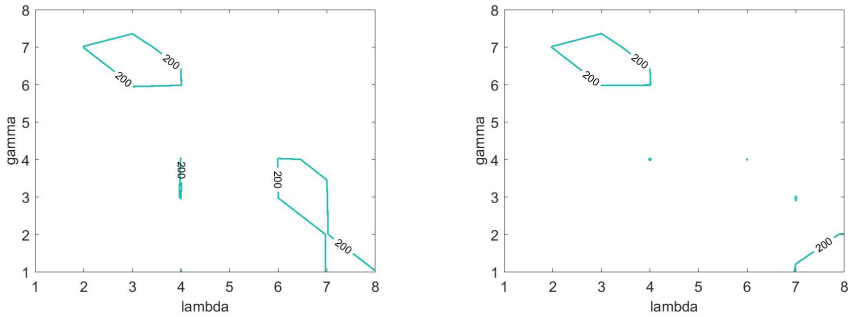


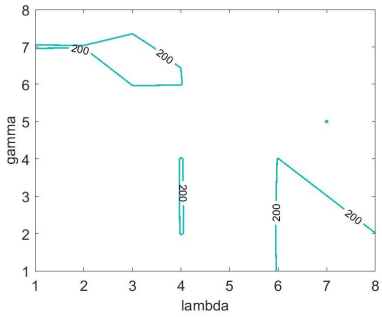
Figure B.8: Bayesian networks constructed from the PC spectra

B.5 For $\alpha = 10^0$

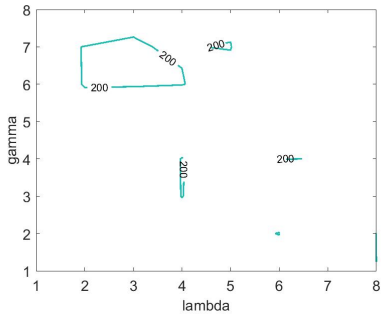


(a) $\beta = 10^{-3}$

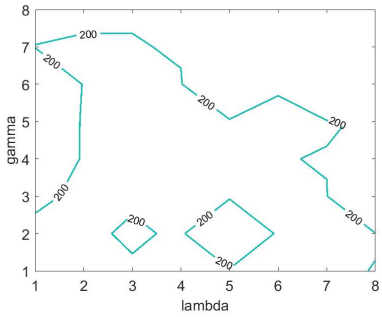
(b) $\beta = 10^{-2}$



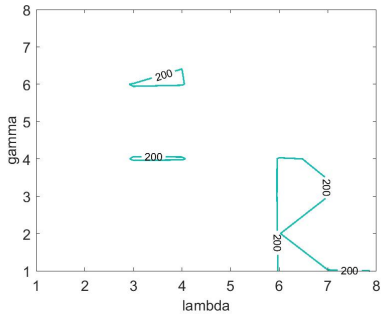
(c) $\beta = 10^{-1}$



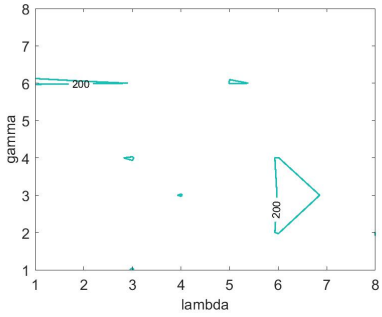
(d) $\beta = 10^0$



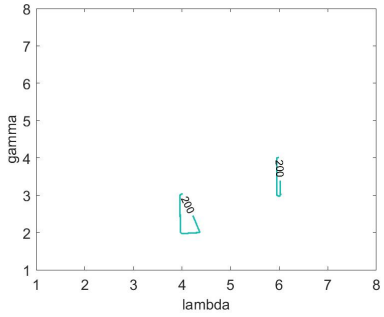
(e) $\beta = 0$



(f) $\beta = 10^1$

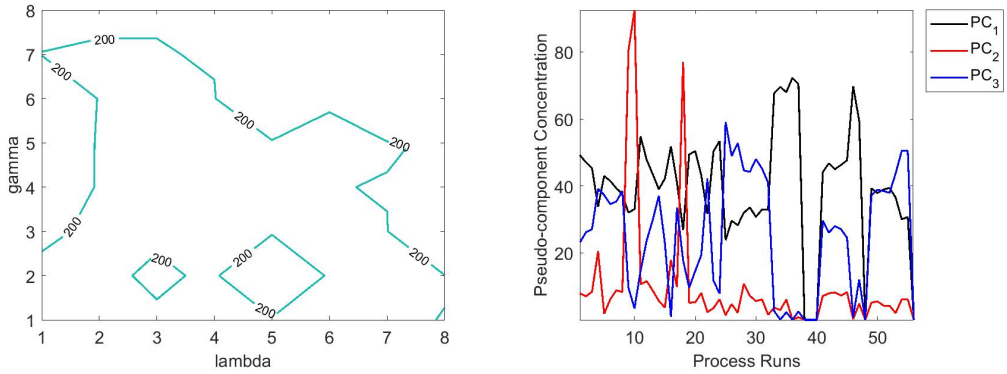


(g) $\beta = 10^2$

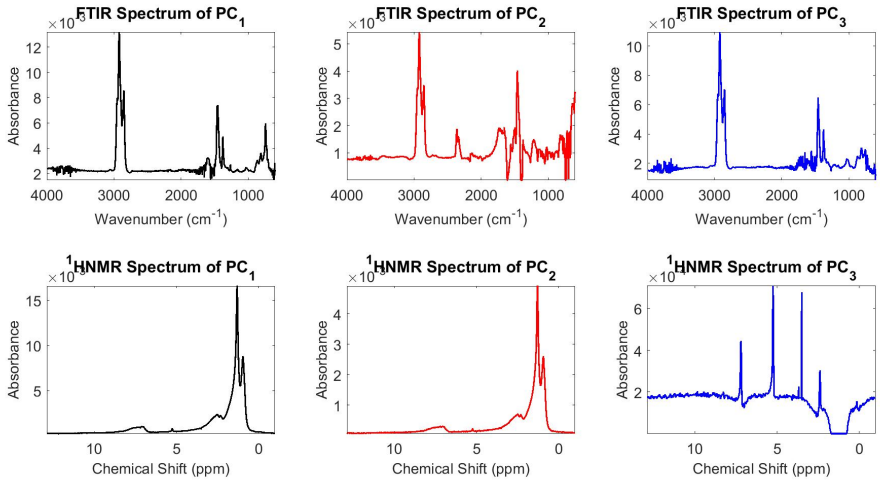


(h) $\beta = 10^3$

Figure B.9: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 0$ (b) Concentration



(c) Pseudo-component spectra

Figure B.10: JNMF profiles for $\alpha = 10^0, \beta = 0, \gamma = 0, \lambda = 10^0$

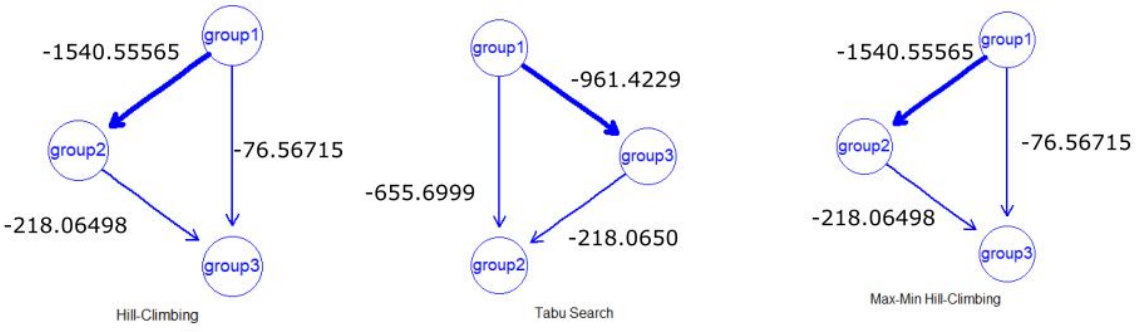
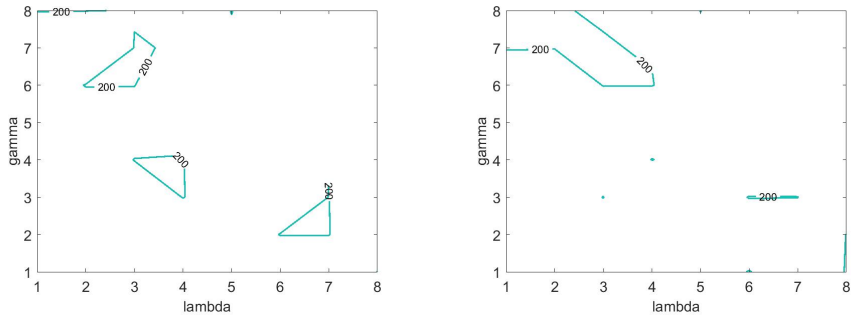


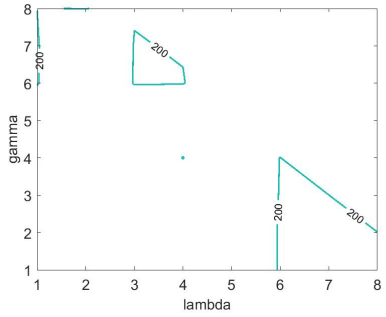
Figure B.11: Bayesian networks constructed from the PC spectra

B.6 For $\alpha = 0$

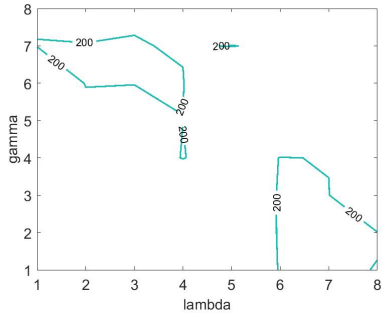


(a) $\beta = 10^{-3}$

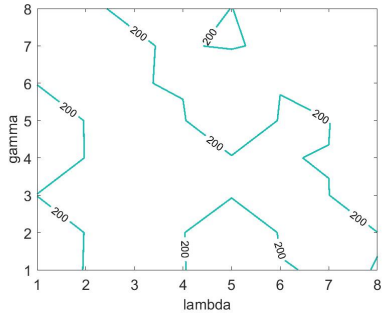
(b) $\beta = 10^{-2}$



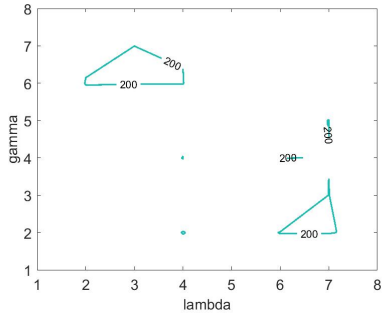
(c) $\beta = 10^{-1}$



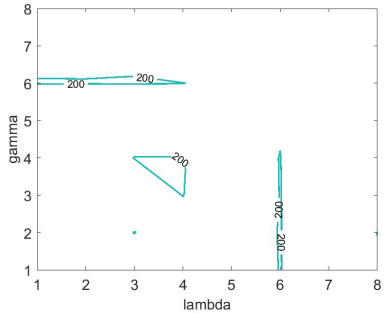
(d) $\beta = 10^{0}$



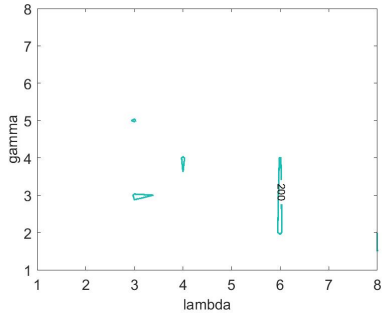
(e) $\beta = 0$



(f) $\beta = 10^{1}$

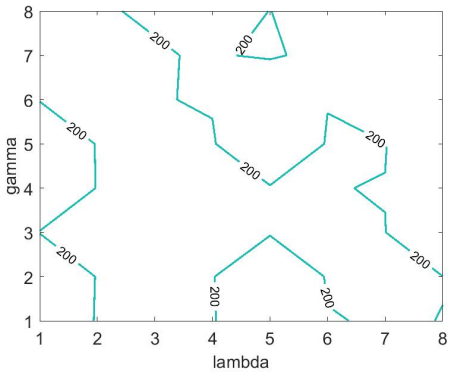


(g) $\beta = 10^{2}$

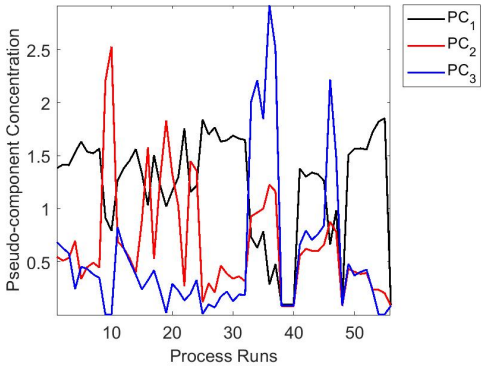


(h) $\beta = 10^{3}$

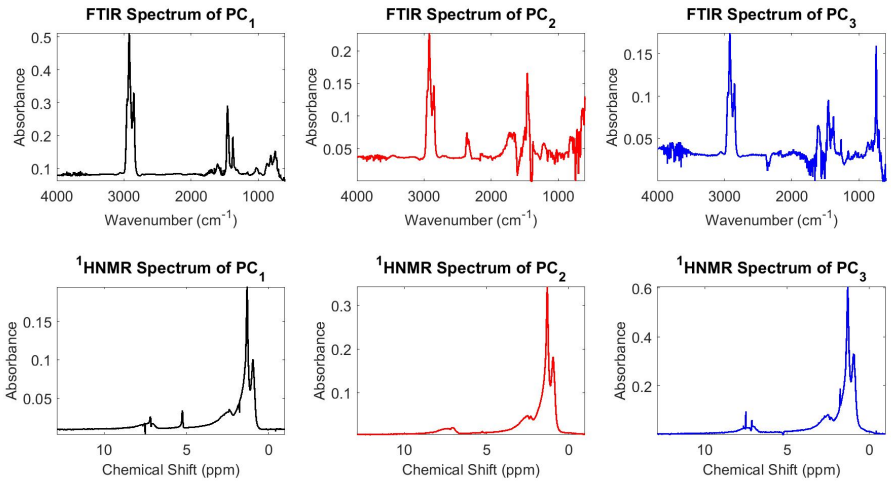
Figure B.12: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 0$



(b) Concentration



(c) Pseudo-component spectra

Figure B.13: JNMF profiles for $\alpha = 0, \beta = 0, \gamma = 10^{-2}, \lambda = 10^{-2}$

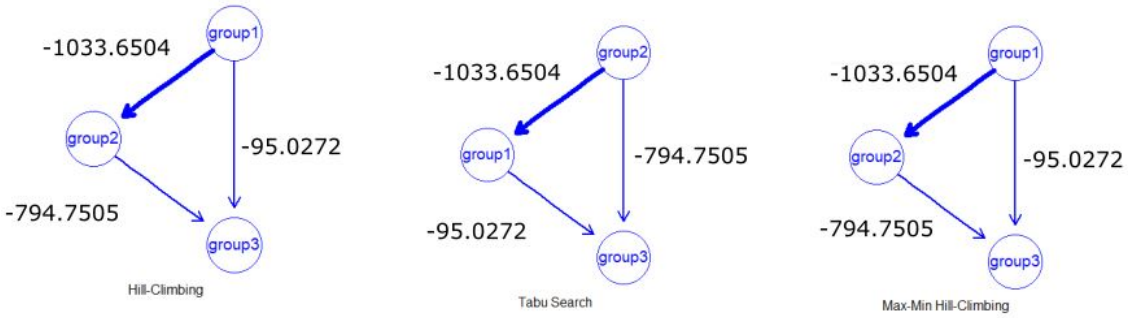
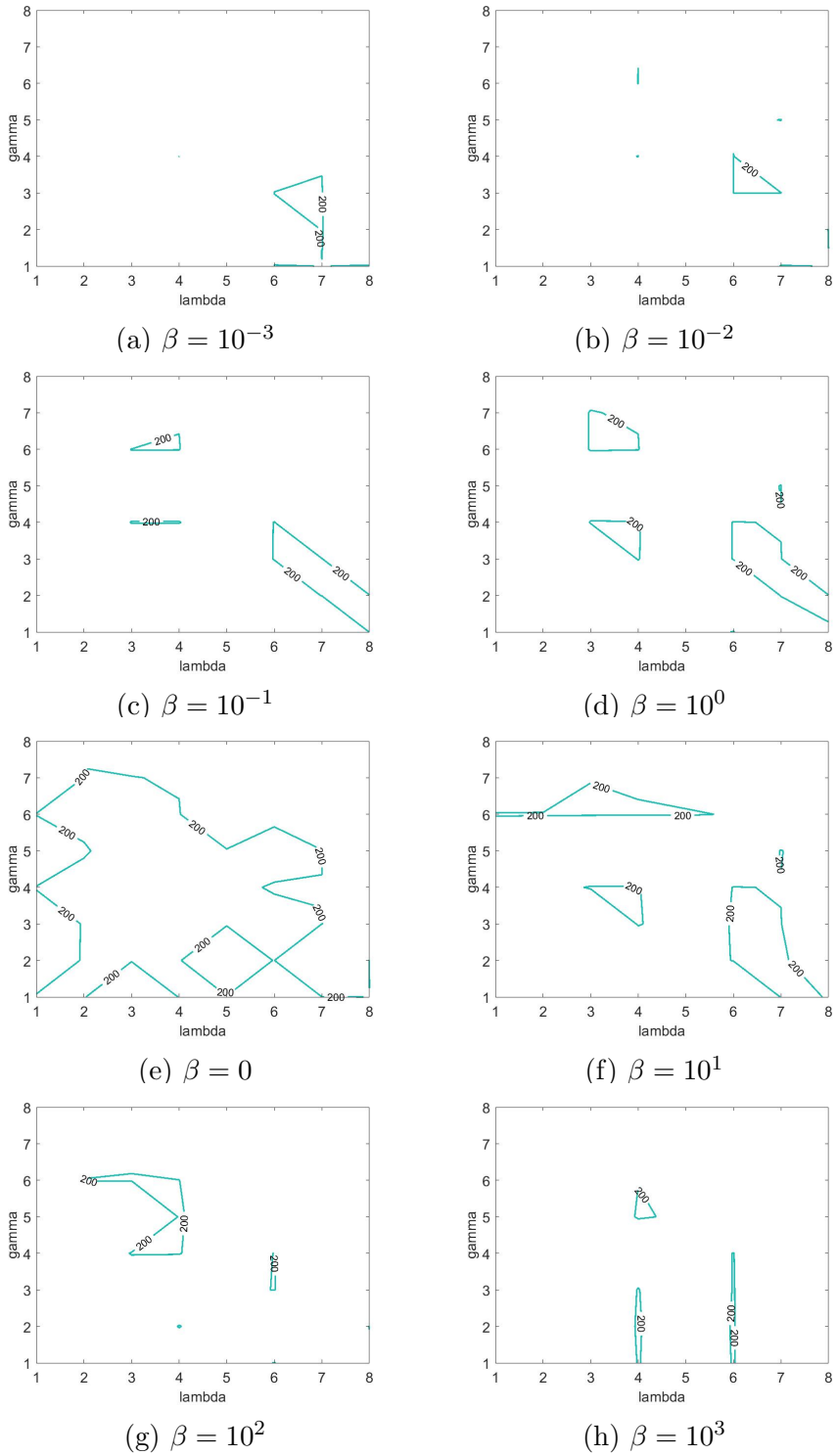


Figure B.14: Bayesian networks constructed from the PC spectra

B.7 For $\alpha = 10$



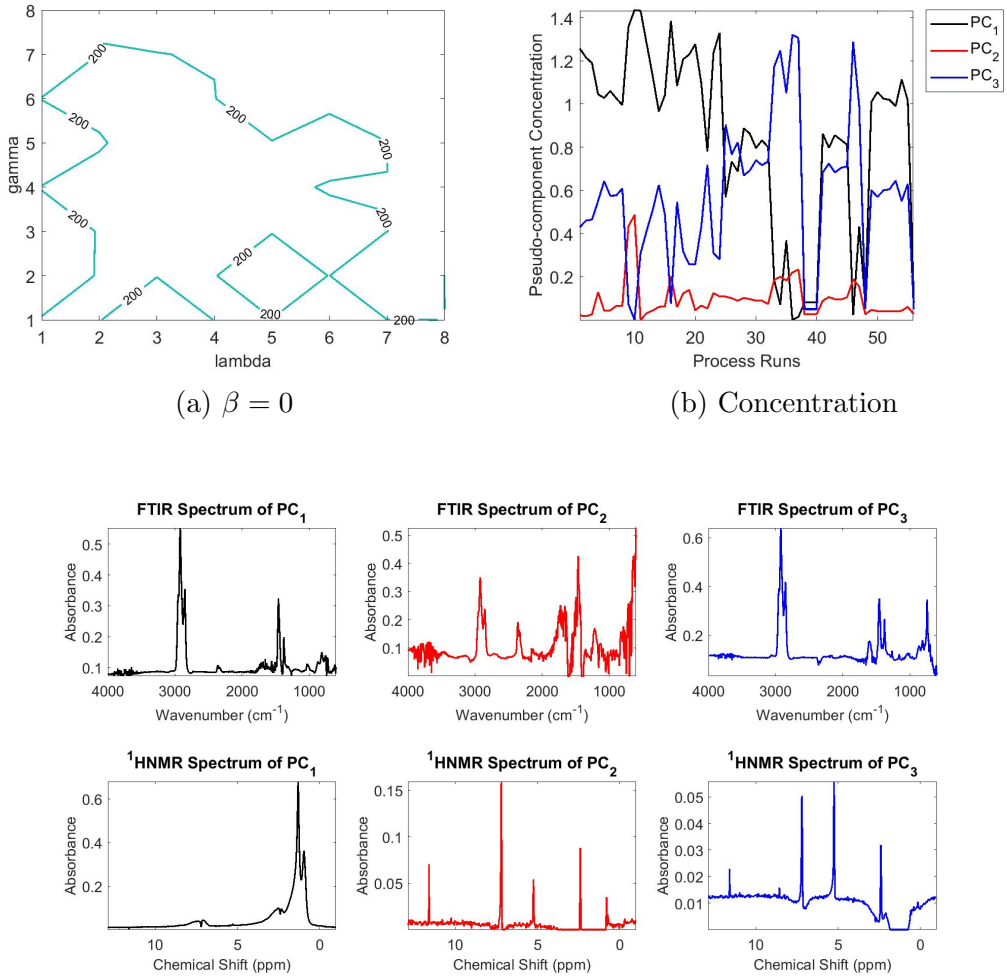
(a) $\beta = 10^{-3}$ (b) $\beta = 10^{-2}$

(c) $\beta = 10^{-1}$ (d) $\beta = 10^{0}$

(e) $\beta = 0$ (f) $\beta = 10^{1}$

(g) $\beta = 10^{2}$ (h) $\beta = 10^{3}$

Figure B.15: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 0$ (b) Concentration

(c) Pseudo-component spectra

Figure B.16: JNMF profiles for $\alpha = 10^1, \beta = 0, \gamma = 10^{-1}, \lambda = 0$

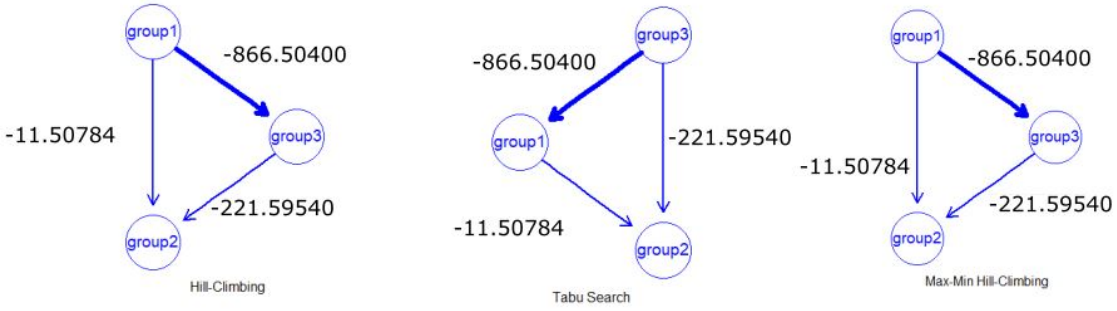
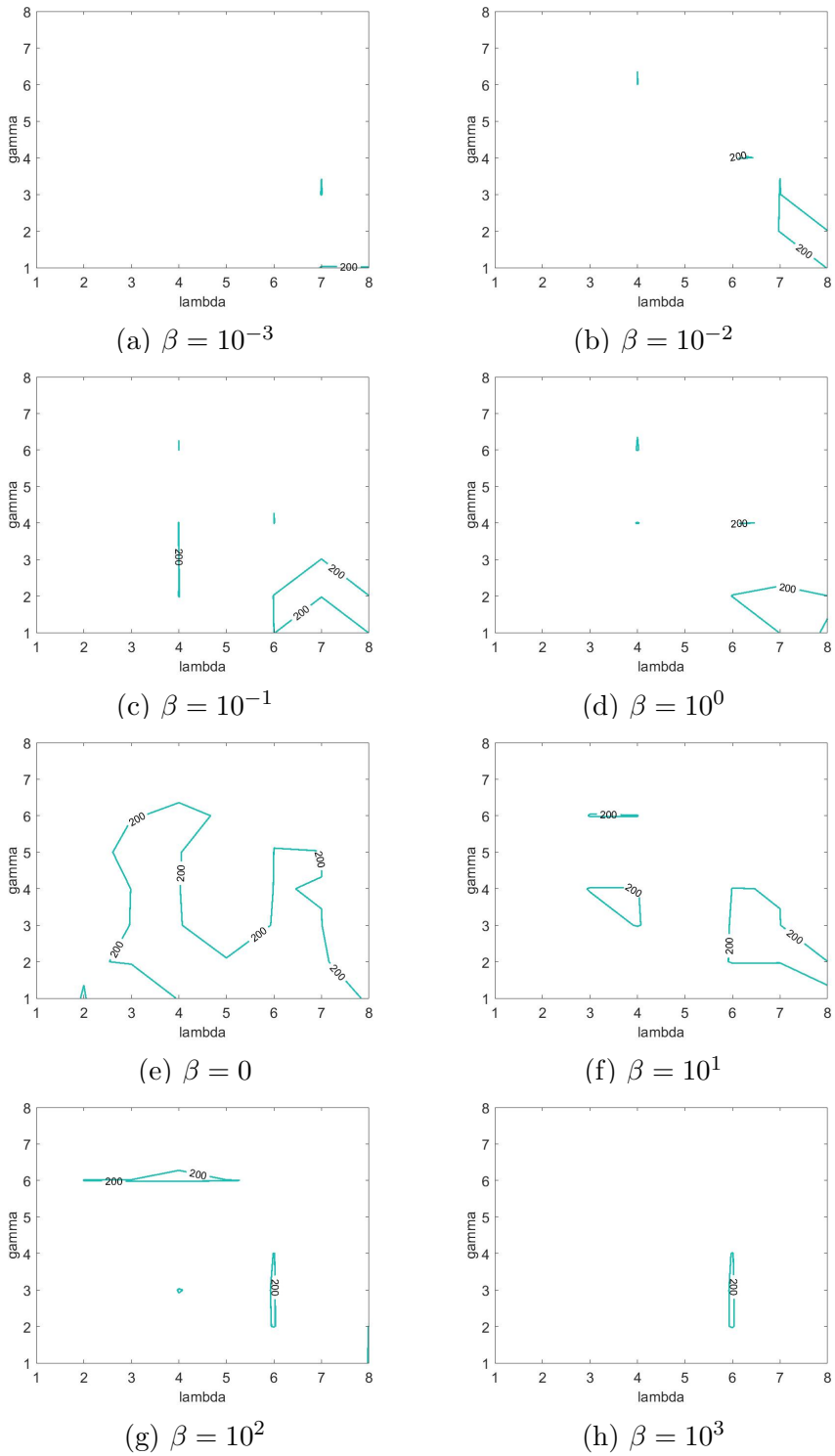


Figure B.17: Bayesian networks constructed from the PC spectra

B.8 For $\alpha = 10^2$



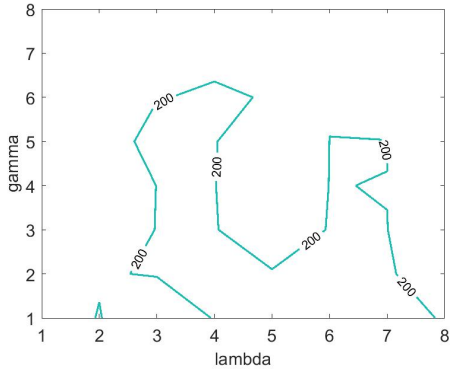
(a) $\beta = 10^{-3}$ (b) $\beta = 10^{-2}$

(c) $\beta = 10^{-1}$ (d) $\beta = 10^{0}$

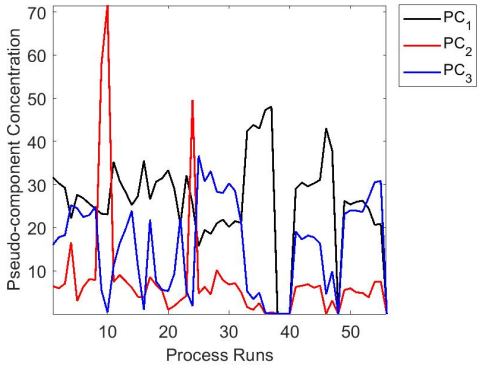
(e) $\beta = 0$ (f) $\beta = 10^{1}$

(g) $\beta = 10^{2}$ (h) $\beta = 10^{3}$

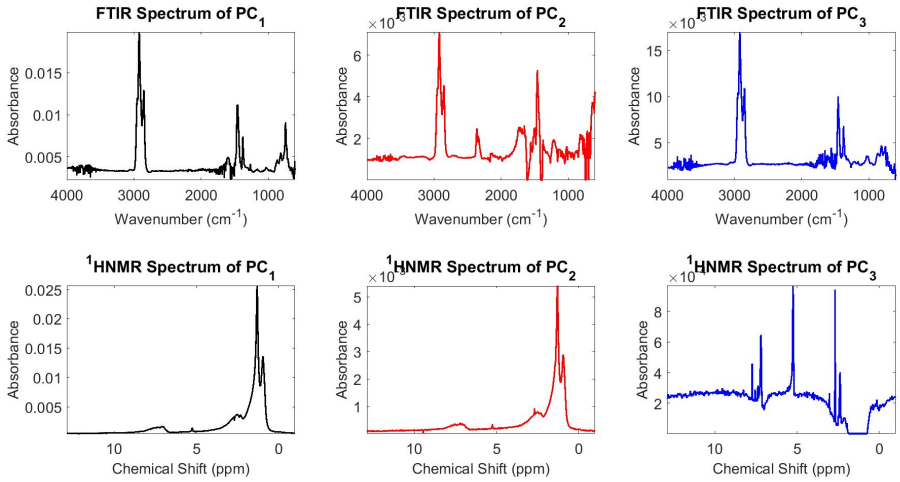
Figure B.18: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 0$



(b) Concentration



(c) Pseudo-component spectra

Figure B.19: JNMF profiles for $\alpha = 10^2, \beta = 0, \gamma = 10^{-2}, \lambda = 10^0$

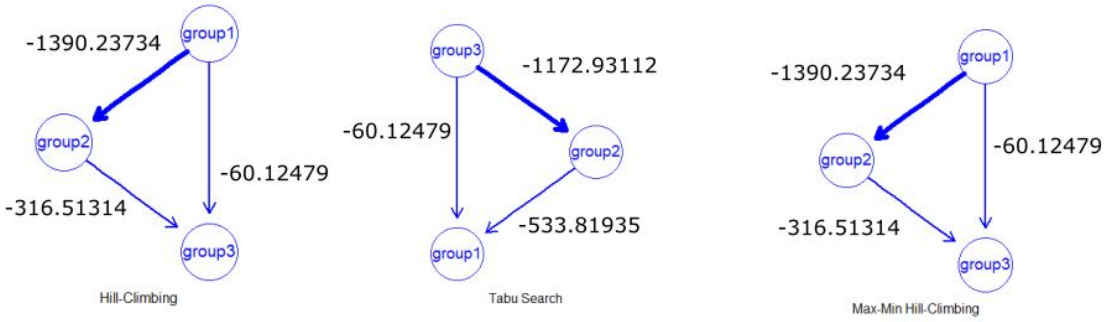
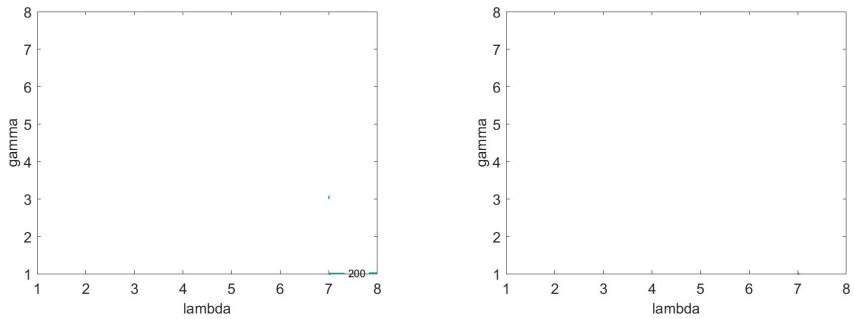


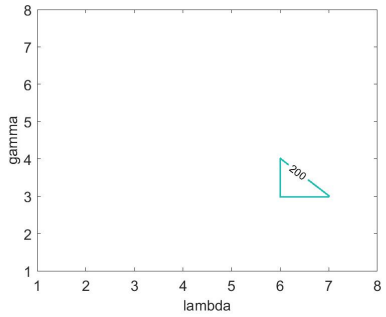
Figure B.20: Bayesian networks constructed from the PC spectra

B.9 For $\alpha = 10^3$

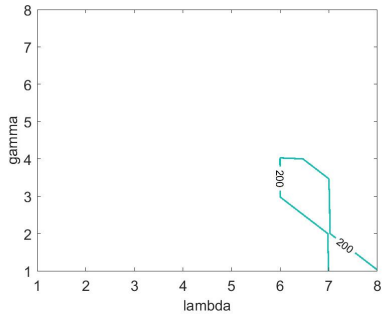


(a) $\beta = 10^{-3}$

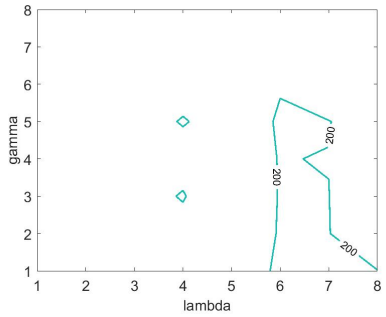
(b) $\beta = 10^{-2}$



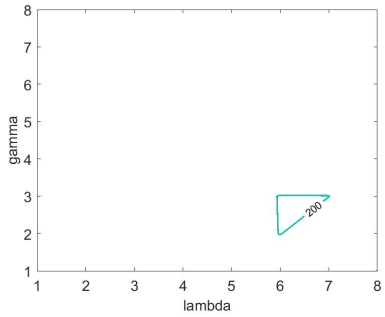
(c) $\beta = 10^{-1}$



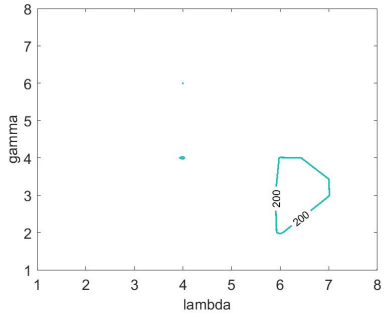
(d) $\beta = 10^{0}$



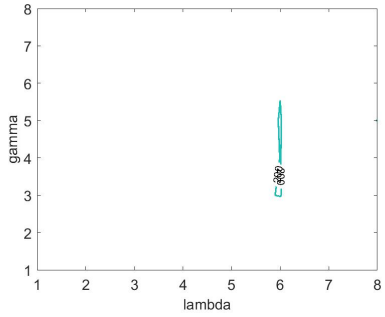
(e) $\beta = 0$



(f) $\beta = 10^{1}$

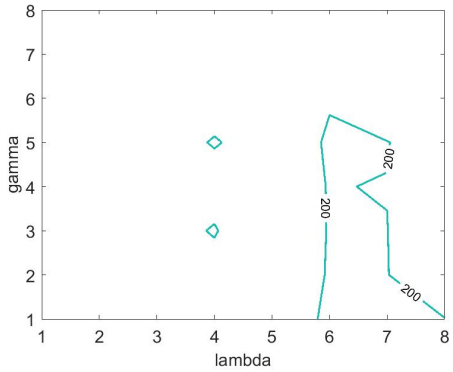


(g) $\beta = 10^{2}$

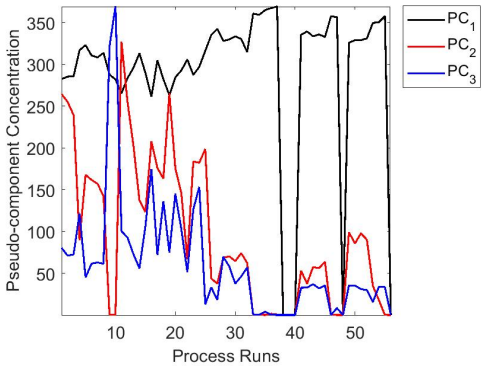


(h) $\beta = 10^{3}$

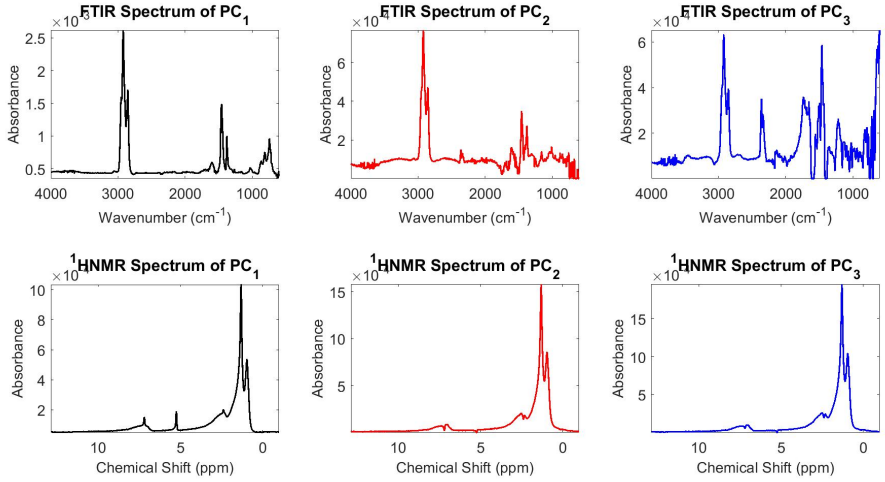
Figure B.21: Isocontours for reconstruction error $E \leq 200$



(a) $\beta = 0$



(b) Concentration



(c) Pseudo-component spectra

Figure B.22: JNMF profiles for $\alpha = 10^3, \beta = 0, \gamma = 10^{-3}, \lambda = 10^1$

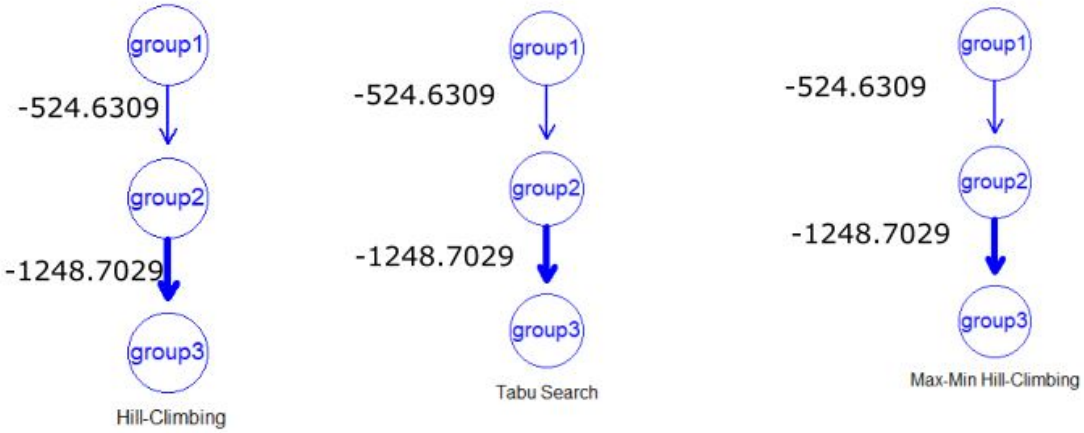


Figure B.23: Bayesian networks constructed from the PC spectra

B.10 ROD=4 for orthogonal case

Relaxing the rank in the factorization of the objective beyond 4 factors in noisy pseudocomponents which is undesirable. The reported results of the factorization for a rank of 4 pseudocomponents given below did not satisfy the convergence criteria within 5000 iterations and had a relatively higher reconstruction error.

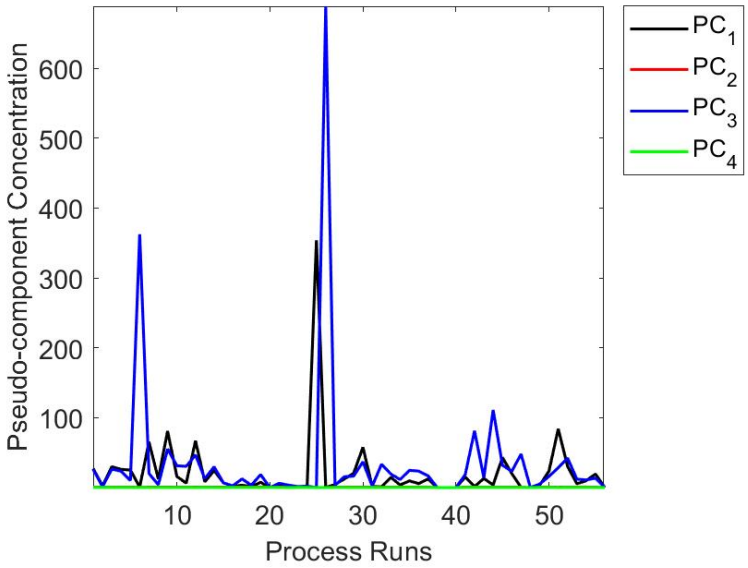


Figure B.24: Concentration profiles

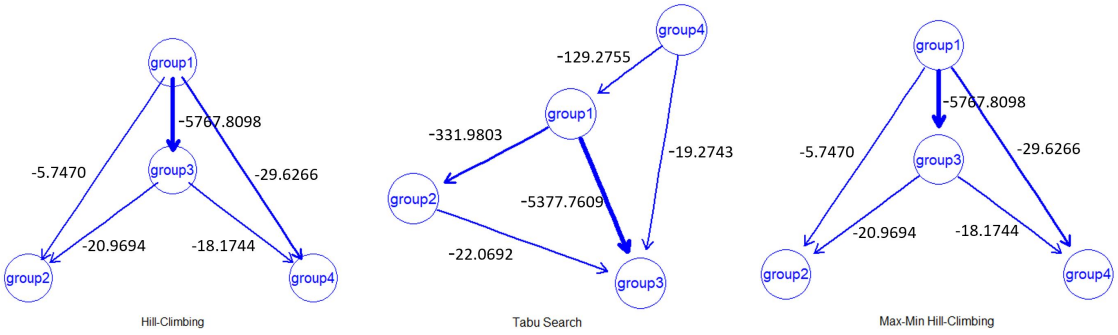
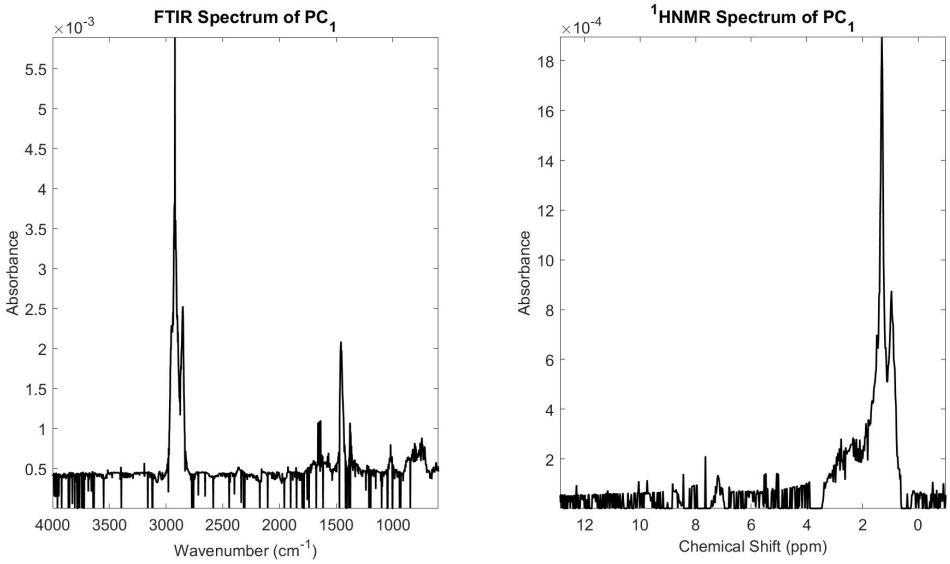
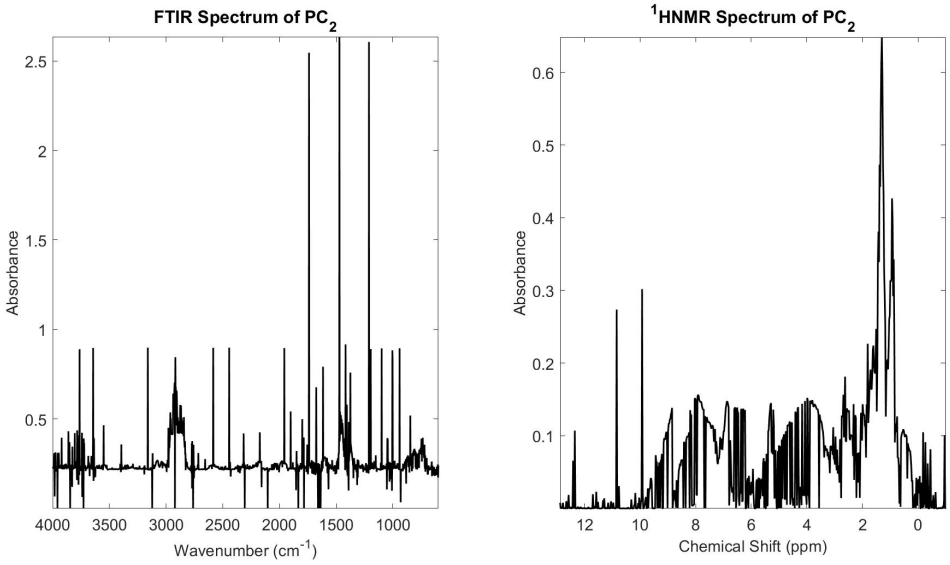


Figure B.25: Bayesian networks constructed from the PC spectra

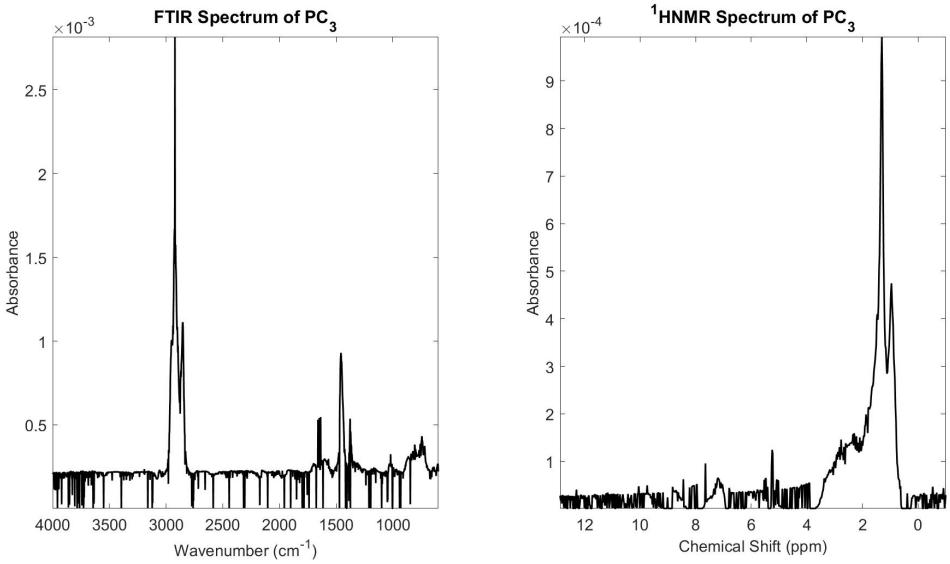


(a) PC_1

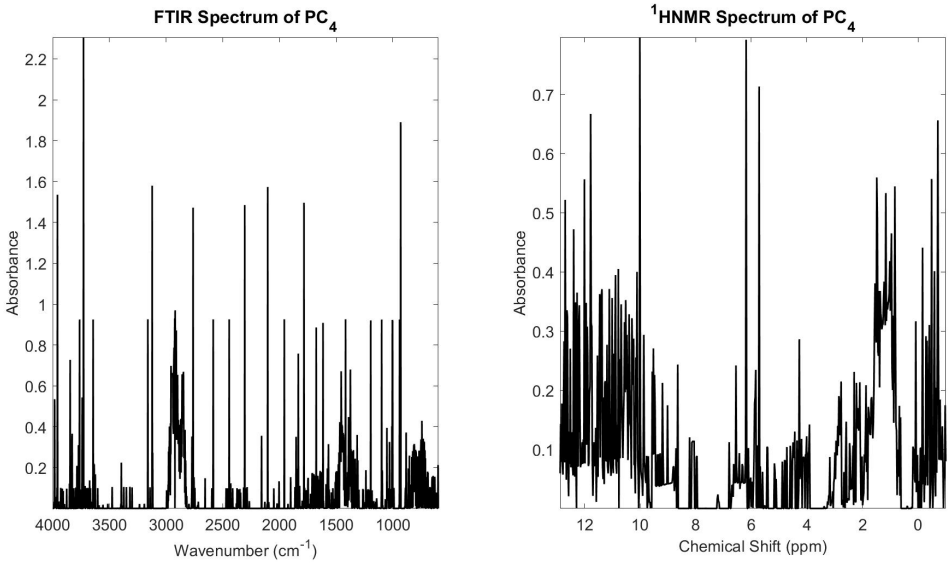


(b) PC_2

Figure B.26: Pseudo-component spectra for rank= 4



(c) PC_3



(d) PC_4

Figure B.26: Pseudo-component spectra for rank= 4

Appendix C: Chapter 4

C.1 Process Conditions

Table C.1: Process conditions for spectral data collection

Spectral sensor	Process conditions	
	Temperature($°C$)	Residence time (min)
FTIR	150	66, 126, 186, 246, 306, 366, 426, 486
	200	66, 126, 186, 246, 306, 486
	250	246
	300	126, 186, 246, 306, 366, 426, 486
	340	6, 66, 126, 246, 486
	360	6, 16.02, 25.98, 36, 66, 246, 583.02
	400	6, 16.02, 25.98, 36, 66, 96, 126
^{1}H-NMR	150	60, 120, 180, 240, 300, 360, 420, 480
	200	60, 120, 180, 240, 300, 360, 420, 480
	250	60, 120, 180, 240, 300, 360, 420, 480
	300	60, 120, 180, 240, 300, 360, 420, 480

C.2 Robust formulation of JNTF using subtensors

This section outlines the approach to gradient-based optimization of simultaneously solving for mode matrices. Individual sub-problems in eqn C.1-eqn C.3 are simple rank R approximations of the mode-n matricized tensors, solved in an ALS-based round robbin scheme.

$$\min_{\mathbf{A}} \sum_{i=1}^{I_1} \sum_{j=1}^{I_2 I_3} \sqrt{(\mathcal{Z}_{(1)} - (\mathbf{A}(\mathbf{C} \odot \mathbf{B})^T)_{ij})^2} \tag{C.1}$$

$$\min_{\mathbf{B}} \sum_{i=1}^{I_2} \sum_{j=1}^{I_3 I_1} \sqrt{(\mathcal{Z}_{(2)} - (\mathbf{B}(\mathbf{C} \odot \mathbf{A})^T)_{ij})^2} \tag{C.2}$$

$$\min_{\mathbf{C}} \sum_{i=1}^{I_3} \sum_{j=1}^{I_1 I_2} \sqrt{(\mathcal{Z}_{(3)} - (\mathbf{C}(\mathbf{B} \odot \mathbf{A})^T)_{ij})^2} \tag{C.3}$$

It is desired to combine these into a single objective function designed to minimize the L_{21} norm of the n^{th} mode matricized tensor. The L_{21} norm of a certain matrix $\mathbf{X}_{m \times n}$ is as given below:

$$||\mathbf{X}||_{21} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} x_{ji}^2} \tag{C.4}$$

From matrix algebra it is known that $||\mathbf{X}||_F^2 = Tr[\mathbf{X}\mathbf{X}^T]$. Following on these lines for an L_{21} norm has a similar expression in terms of the trace $||\mathbf{X}||_{21} = Tr[\mathbf{X}\ \mathbf{D}\ \mathbf{X}^T]$, with an additional diagonal scaling matrix $\mathbf{D}$ defined as follows:

$$\mathbf{D}(\mathbf{X}) = \frac{I_{n \times n}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \text{for any } \mathbf{X}_{m \times n} \tag{C.5}$$

Using eqn C.4 and eqn C.5 in eqn 15 we have the following formulation of the objective function:

$$\begin{aligned} \min_{\mathbf{A},\mathbf{B},\mathbf{C} \geq 0} F(\mathbf{A},\mathbf{B},\mathbf{C}) = Tr\Big(& \{\mathcal{W}_{(1)} * [\mathcal{Z}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T]\}\mathbf{D}_1\{\mathcal{W}_{(1)} * [\mathcal{Z}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T]\}^T \\ & + \{\mathcal{W}_{(2)} * [\mathcal{Z}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^T]\}\mathbf{D}_2\{\mathcal{W}_{(2)} * [\mathcal{Z}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^T]\}^T \\ & + \{\mathcal{W}_{(3)} * [\mathcal{Z}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^T]\}\mathbf{D}_3\{\mathcal{W}_{(3)} * [\mathcal{Z}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^T]\}^T \Big) \end{aligned} \tag{C.6}$$

where $\mathbf{D}_1 = \mathbf{D}(\mathcal{W}_{(1)} * [\mathcal{Z}_{(1)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T])$, $\mathbf{D}_2 = \mathbf{D}(\mathcal{W}_{(2)} * [\mathcal{Z}_{(2)} - \mathbf{B}(\mathbf{C} \odot \mathbf{A})^T])$, $\mathbf{D}_3 = \mathbf{D}(\mathcal{W}_{(3)} * [\mathcal{Z}_{(3)} - \mathbf{C}(\mathbf{B} \odot \mathbf{A})^T])$ are the diagonal scaling matrices for the n^{th} mode matricized tensor.

The gradients of the objective function in eqn C.6 with respect to each of the factor matrices is given below:

$$\begin{aligned}\nabla F_{\mathbf{A}} &= \mathcal{W}_{(1)} * \Big(\mathbf{A}(\mathbf{C}\odot\mathbf{B})^T - \mathcal{Z}_{(1)}\Big)\mathbf{D}_1(\mathbf{C}\odot\mathbf{B}) \\ &+ \frac{\partial(\mathbf{C}\odot\mathbf{A})}{\partial\mathbf{A}}\mathbf{B}^T\mathcal{W}_{(2)} * \Big(\mathbf{B}(\mathbf{C}\odot\mathbf{A})^T - \mathcal{Z}_{(2)}\Big)\mathbf{D}_2(\mathbf{C}\odot\mathbf{A}) \\ &+ \frac{\partial(\mathbf{B}\odot\mathbf{A})}{\partial\mathbf{A}}\mathbf{C}^T\mathcal{W}_{(3)} * \Big(\mathbf{C}(\mathbf{B}\odot\mathbf{A})^T - \mathcal{Z}_{(3)}\Big)\mathbf{D}_3(\mathbf{B}\odot\mathbf{A})\end{aligned} \tag{C.7}$$

$$\begin{aligned}\nabla F_{\mathbf{B}} &= \mathcal{W}_{(2)} * \Big(\mathbf{B}(\mathbf{C}\odot\mathbf{A})^T - \mathcal{Z}_{(2)}\Big)\mathbf{D}_2(\mathbf{C}\odot\mathbf{A}) \\ &+ \frac{\partial(\mathbf{C}\odot\mathbf{B})}{\partial\mathbf{B}}\mathbf{A}^T\mathcal{W}_{(1)} * \Big(\mathbf{A}(\mathbf{C}\odot\mathbf{B})^T - \mathcal{Z}_{(1)}\Big)\mathbf{D}_1(\mathbf{C}\odot\mathbf{B}) \\ &+ \frac{\partial(\mathbf{B}\odot\mathbf{A})}{\partial\mathbf{B}}\mathbf{C}^T\mathcal{W}_{(3)} * \Big(\mathbf{C}(\mathbf{B}\odot\mathbf{A})^T - \mathcal{Z}_{(3)}\Big)\mathbf{D}_3(\mathbf{B}\odot\mathbf{A})\end{aligned} \tag{C.8}$$

$$\begin{aligned}\nabla F_{\mathbf{C}} &= \mathcal{W}_{(3)} * \Big(\mathbf{C}(\mathbf{B}\odot\mathbf{A})^T - \mathcal{Z}_{(3)}\Big)\mathbf{D}_3(\mathbf{B}\odot\mathbf{A}) \\ &+ \frac{\partial(\mathbf{C}\odot\mathbf{B})}{\partial\mathbf{C}}\mathbf{A}^T\mathcal{W}_{(1)} * \Big(\mathbf{A}(\mathbf{C}\odot\mathbf{B})^T - \mathcal{Z}_{(1)}\Big)\mathbf{D}_1(\mathbf{C}\odot\mathbf{B}) \\ &+ \frac{\partial(\mathbf{C}\odot\mathbf{A})}{\partial\mathbf{C}}\mathbf{B}^T\mathcal{W}_{(2)} * \Big(\mathbf{B}(\mathbf{C}\odot\mathbf{A})^T - \mathcal{Z}_{(2)}\Big)\mathbf{D}_2(\mathbf{C}\odot\mathbf{A})\end{aligned} \tag{C.9}$$

To tackle the derivatives of the Khatri-rao ($\odot$) aka columnwise Kronecker product ($|\otimes|$) in the above expression for gradients we resort to the use of vectorizing the product expressions using principles of tensor algebra, computing the gradients of the vectors and then re-shaping them to matrices.

For example let us say we have two matrices $\mathbf{X}_1$ and $\mathbf{X}_2$ of dimensions $m \times n$ and $p \times n$ respectively, then the derivative of their column-wise Kronecker product is given by:

$$\frac{\partial(\mathbf{X}_1\odot\mathbf{X}_2)}{\partial\mathbf{X}_i} = \text{Reshape}\left(\frac{\partial\, vec\{\mathbf{X}_1\odot\mathbf{X}_2\}}{\partial x_i}\right) = \text{Reshape}\Big(K_i^T K_i\, vec\{\mathbf{X}_i\}\Big) \tag{C.10}$$

Expressions for K_i come from the following two equations from tensor algebra:

$$vec\{\mathbf{X}_1\odot\mathbf{X}_2\} = \quad ([I_N\odot\mathbf{X}_1]\otimes I_P)\, vec\{\mathbf{X}_2\} = \quad K_2 vec\{\mathbf{X}_2\} \tag{C.11}$$

$$vec\{\mathbf{X}_1\odot\mathbf{X}_2\} = \quad [I_{MN}\odot(\mathbf{X}_2\,[I_N\otimes 1_{1\times M}])]\, vec\{\mathbf{X}_1\} = \quad K_1 vec\{\mathbf{X}_1\} \tag{C.12}$$

It can be seen that the gradient computation of mode matricized tensors involve the derivatives of the Khatri-Rao products of the matrix modes, the computation of which is memory intensive for large-scale tensors making it challenge in the implementation of JNTF [271]. Hence a large tensor is typically divided into subtensors, parallelzing the JNTF over the small-sized subtensors using the divide and conquer technique [421].

The concepts discussed in this section are now put together as we extend it to the framework of *joint* weighted robust non-negative tensor factorization with respect to our case of factorizing tensor blocks of FTIR and ^{1}H-NMR data. Since the dimension of the spectral channel modes are much higher than that of the process modes, it is proposed to divide the tensors into subtensors along the spectral channel modes. Hence, the grid tensor factorization (GTF) is also implemented in the high dimensional mode of wavenumbers/chemical shifts. Let N_1, N_2 be the number of FTIR and HNMR subtensors respectively. For FTIR i=1 for HNMR i=2 : $\mathcal{Z}^{[n_i]} \in \Re^{I_1 \times I_2 \times K_{n_i}}$ So from CPD $\mathcal{Z}^{[n_i]} \approx I \times_1 \mathbf{A}^{[n_i]} \times_2 \mathbf{B}^{[n_i]} \times_3 \mathbf{H}_i^{[n_i]}$ where $n_i = 1, 2 \cdots N_i$ and $\mathbf{A}^{[n_i]} \in \Re^{I_1 \times R}, \mathbf{B}^{[n_i]} \in \Re^{I_2 \times R}, \mathbf{H}_i^{[n_i]} \in \Re^{K_{n_i} \times R}$ such that $\sum_{n_i=1}^{N_i} K_{n_i} = I_3$ followed by $\mathbf{H}_i = [\mathbf{H}_i^{[1]T}, \mathbf{H}_i^{[2]T}, \cdots \mathbf{H}_i^{[N_i]T}]^T$

The objective function :

$$\min_{\mathbf{A}^{[n_i]}, \mathbf{B}^{[n_i]}, \mathbf{H}_i^{[n_i]} \geq 0} \sum_{i=1,2} \sum_{n_i=1}^{N_i} ||\mathcal{W}^{[n_i]} * (\mathcal{Z}^{[n_i]} - [[\mathbf{A}^{[n_i]}, \mathbf{B}^{[n_i]}, \mathbf{H}_i^{[n_i]}]])||_{21} \tag{C.13}$$

Writing out eqn C.13 out explicitly in terms of the matricized n-mode tensor:

$$\begin{aligned} \min_{\mathbf{A}^{[n_i]}, \mathbf{B}^{[n_i]}, \mathbf{H}_i^{[n_i]} \geq 0} F(\mathbf{A}, \mathbf{B}, \mathbf{H}_i) = \sum_{i=1,2} \sum_{n_i=1}^{N_i} & \left|\left| \mathcal{W}_{(1)}^{[n_i]} * [\mathcal{Z}_{(1)}^{[n_i]} - \mathbf{A}^{[n_i]} (\mathbf{H}_i^{[n_i]} \odot \mathbf{B}^{[n_i]})^T] \right|\right|_{21} + \\ & \left|\left| \mathcal{W}_{(2)}^{[n_i]} * [\mathcal{Z}_{(2)}^{[n_i]} - \mathbf{B}^{[n_i]} (\mathbf{H}_i^{[n_i]} \odot \mathbf{A}^{[n_i]})^T] \right|\right|_{21} \\ & + \left|\left| \mathcal{W}_{(3)}^{[n_i]} * [\mathcal{Z}_{(3)}^{[n_i]} - \mathbf{H}_i^{[n_i]} (\mathbf{B}^{[n_i]} \odot \mathbf{A}^{[n_i]})^T] \right|\right|_{21} \end{aligned} \tag{C.14}$$

Using eqn C.4 and eqn C.5 in eqn C.14 we have the following formulation of the

objective function:

$$\min_{\mathbf{A}^{[n_i]},\mathbf{B}^{[n_i]},\mathbf{H}_i^{[n_i]}\geq 0} F(\mathbf{A},\mathbf{B},\mathbf{H}_i)=\sum_{i=1,2}\sum_{n_i=1}^{N_i} Tr\Big(\{\mathcal{W}_{(1)}^{[n_i]}*[\mathcal{Z}_{(1)}^{[n_i]}-\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T]\}\mathbf{D}_1^{[n_i]}\{\mathcal{W}_{(1)}^{[n_i]}*[\mathcal{Z}_{(1)}^{[n_i]}-\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T]\}^T$$
$$+\{\mathcal{W}_{(2)}^{[n_i]}*[\mathcal{Z}_{(2)}^{[n_i]}-\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}\mathbf{D}_2^{[n_i]}\{\mathcal{W}_{(2)}^{[n_i]}*[\mathcal{Z}_{(2)}^{[n_i]}-\mathcal{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}^T$$
$$+\{\mathcal{W}_{(3)}^{[n_i]}*[\mathcal{Z}_{(3)}^{[n_i]}-\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}\mathbf{D}_3^{[n_i]}\{\mathcal{W}_{(3)}^{[n_i]}*[\mathcal{Z}_{(3)}^{[n_i]}-\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T]\}^T\Big) \tag{C.15}$$

The gradients of the objective function wrt to the factor matrices are given below:

$$\begin{aligned}\nabla F_{\mathbf{A}} = &\sum_{i=1,2}\sum_{n_i=1}^{N_i}\mathcal{W}_{(1)}^{[n_i]}*\Big(\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T-\mathcal{Z}_{(1)}^{[n_i]}\Big)\mathbf{D}_1^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})\\ &+\frac{\partial(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{A}^{[n_i]}}\mathbf{B}^{[n_i]T}\mathcal{W}_{(2)}^{[n_i]}*\Big(\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(2)}^{[n_i]}\Big)\mathbf{D}_2^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})\\ &+\frac{\partial(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{A}^{[n_i]}}\mathbf{H}_i^{[n_i]T}\mathcal{W}_{(3)}^{[n_i]}*\Big(\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]}T)-\mathcal{Z}_{(3)}^{[n_i]}\Big)\mathbf{D}_3^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})\end{aligned} \tag{C.16}$$

$$\begin{aligned}\nabla F_{\mathbf{B}} = &\sum_{i=1,2}\sum_{n_i=1}^{N_i}\mathcal{W}_{(2)}^{[n_i]}*\Big(\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(2)}^{[n_i]}\Big)\mathbf{D}_2^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})\\ &+\frac{\partial(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})}{\partial\mathbf{B}^{[n_i]}}\mathbf{A}^{[n_i]T}\mathcal{W}_{(1)}^{[n_i]}*\Big(\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T-\mathcal{Z}_{(1)}^{[n_i]}\Big)\mathbf{D}_1^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})\\ &+\frac{\partial(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{B}^{[n_i]}}\mathbf{H}_i^{[n_i]T}\mathcal{W}_{(3)}^{[n_i]}*\Big(\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(3)}^{[n_i]}\Big)\mathbf{D}_3^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})\end{aligned} \tag{C.17}$$

$$\begin{aligned}\nabla F_{\mathbf{H}_i} = &\mathcal{W}_{(3)}^{[n_i]}*\Big(\mathbf{H}_i^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(3)}^{[n_i]}\Big)\mathbf{D}_3^{[n_i]}(\mathbf{B}^{[n_i]}\odot\mathbf{A}^{[n_i]})\\ &+\frac{\partial(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})}{\partial\mathbf{H}_i^{[n_i]}}\mathbf{A}^{[n_i]T}\mathcal{W}_{(1)}^{[n_i]}*\Big(\mathbf{A}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})^T-\mathcal{Z}_{(1)}^{[n_i]}\Big)\mathbf{D}_1^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{B}^{[n_i]})\\ &+\frac{\partial(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})}{\partial\mathbf{H}_i^{[n_i]}}\mathbf{B}^{[n_i]T}\mathcal{W}_{(2)}^{[n_i]}*\Big(\mathbf{B}^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})^T-\mathcal{Z}_{(2)}^{[n_i]}\Big)\mathbf{D}_2^{[n_i]}(\mathbf{H}_i^{[n_i]}\odot\mathbf{A}^{[n_i]})\end{aligned} \tag{C.18}$$

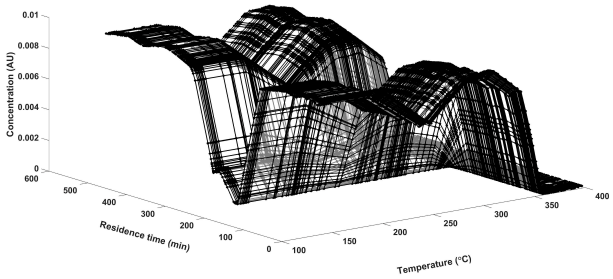
The above problem has been formulated as a gradient-based optimization and is solved using the LBFGSB solver of the Poblano optimization toolbox developed by Sandia Laboratories on Matlab [267].

C.3 NTF of synthetically generated FTIR spectra

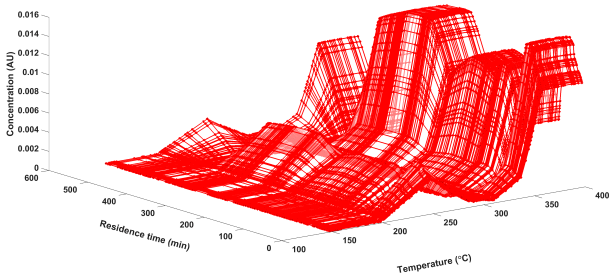
Section 4.1 describes the results of performing robust non-negative tensor factorization on the FTIR spectra for the 41 temperature and residence time conditions given in Table **??**, in addition to the baseline spectrum. The absence of spectral data across certain reaction times at each temperature are accorded as missing values, and are imputed in the process of factorization. In this section, we investigate the results of NTF in the event of being able to collect data extensively across all times at each temperature, at several intermediate temperature conditions. The spectral data at the intermediate temperature-time conditions have been generated synthetically by random interpolation of the existing spectral data in Table **??**, followed by baseline correction before being fed into the NTF objective.

Figure C.1 provides the concentration profiles across the reaction space of temperature and residence times, for the 4 pseudo-components, while Figure C.2 gives the extracted spectral profiles obtained by projection onto the FTIR spectral channels for the 4 pseudo-components. It can be seen that the concentration surface of PC_3 is more pronounced at intermediate residence times, whereas PC_1 is seen to have a sharp decreasing trend, while PC_2 and PC_4 have smaller increases in concentration, that later rise at higher temperatures. It can be inferred that PC_1 represents a class of starting reactants that finally give rise to a class of final products, represented by PC_3, while PC_2 and PC_4 could be treated as a class of reaction intermediates obtained by various mechanisms underlying the conversion of $PC_1 \rightarrow PC_3$.

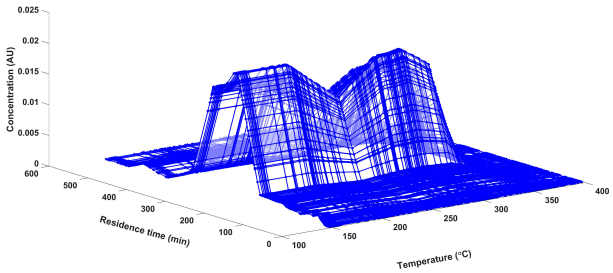
The reaction mechanisms inferred by using Bayesian structure learning among the pseudo-component spectra of Figure C.2, as given by Figure C.3 is found to corroborate with the qualitative inferences drawn from the concentration profiles. The details of the reaction mechanisms underlying the hypotheses generated from the Bayesian networks can be deciphered by chemically interpreting the functional groups in the spectra of the associated pseudo-components.



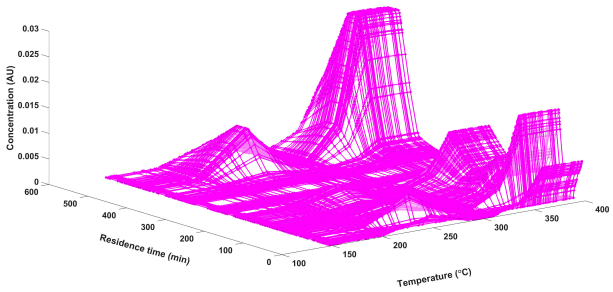
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure C.1: Concentrations of the pseudo-components across the reaction space of the synthetic FTIR dataset

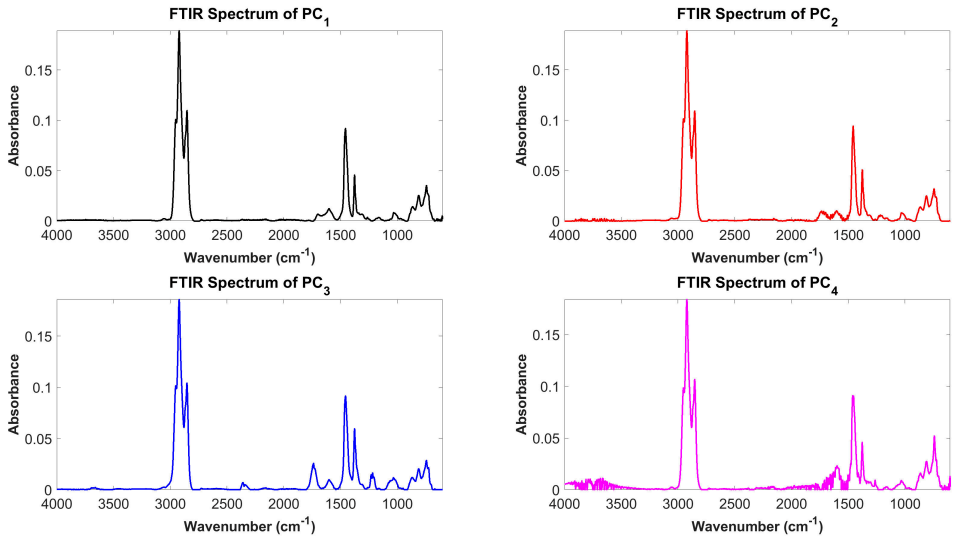


Figure C.2: Spectra of pseudo-components from the synthetic FTIR tensor decomposition

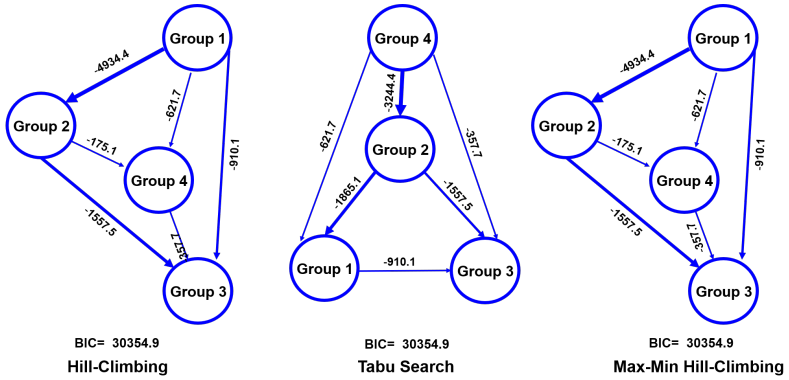
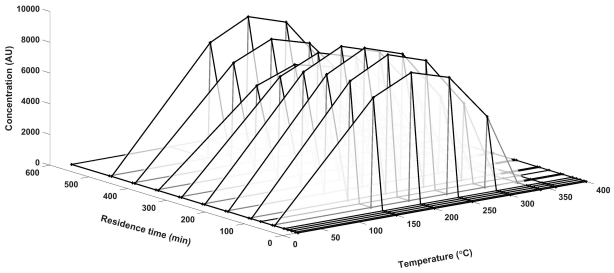


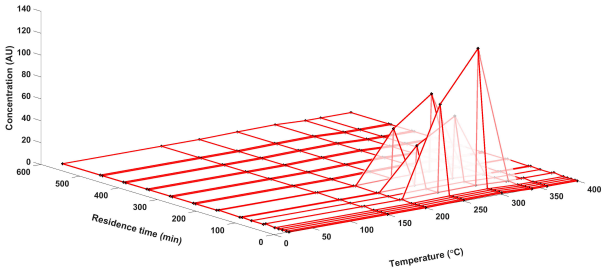
Figure C.3: Bayesian networks from the synthetic FTIR pseudo-component spectra

C.4 Individual analysis of ^{1}H-NMR data

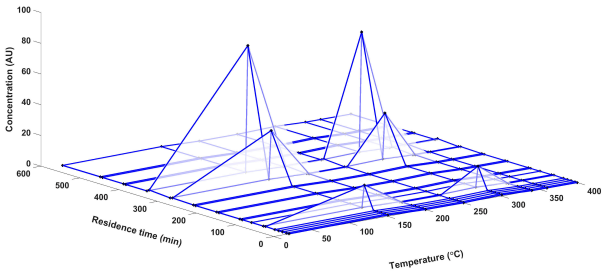
The extracted concentration profiles of the pseudo-components from tensor decomposition in the reaction space of the temperature and residence time modes are given in Figure C.4. The extracted ^{1}H-NMR profiles for the 4 pseudo-components are given in Figure C.5. The Bayesian networks depicting causal relationships among the 4 groups are given in Figure C.6. Hill climbing and the maximum minimum hill climbing score search methods result in similar network structures that indicate PC_1 as the reactant species. The concentration of PC_1 is seen to be much higher than the



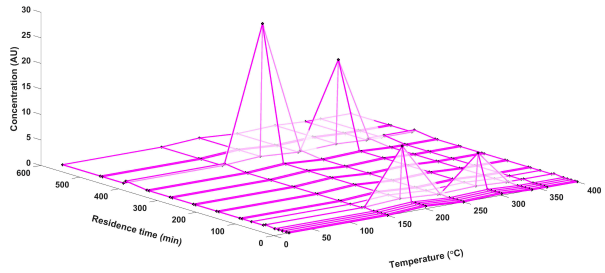
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure C.4: Concentrations of the pseudo-components across the reaction space of the ^{1}H-NMR spectra

other pseudo-components at all temperatures and residence times, corroborate with PC_1 being the starting reactant species. The concentrations of PC_2 are prominent at higher temperatures and lower residence times, while PC_3 and PC_4 appear at lower temperatures reacted over longer durations, towards the later part of the reaction residence time, indicating that they represent a class of the product species, as indicated by the Bayesian networks as well.

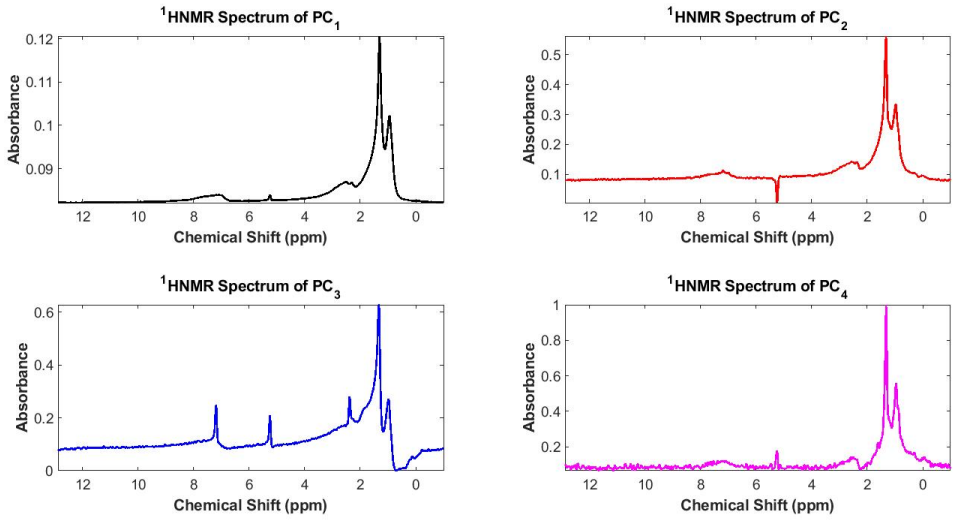


Figure C.5: Spectra of pseudo-components from ^{1}HNMR tensor decomposition

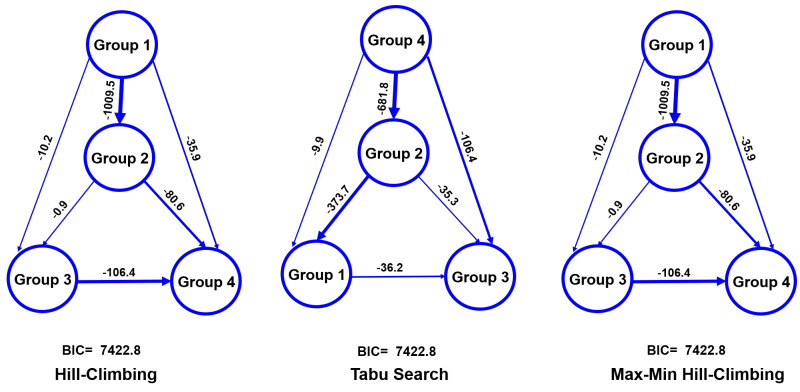


Figure C.6: Bayesian networks from the unique ^{1}H-NMR pseudo-component spectra

^{1}H-NMR spectra alone does not provide as much information as the FTIR spectra, especially in the aromatic region since it just shows a single overlapped lump from 7

– 9 ppm (Figure C.5), except for PC_3 that has a distinct aromatic hydrogen peak at ∼7.2 ppm. Peaks for aliphatic methylene and methyl protons are distinct and common to all pseudo-components with CH_2 showing higher intensity. All spectra also show the peak for benzylic proton at ∼2.5 ppm confirming the presence of aromatics, but this does not indicate the number of substitutions. Another distinct characteristic of ^{1}H-NMR profiles is the peak at ∼5.2 ppm that depicts hydrogen from methylene chloride which points to the solvent that remains in the converted samples. This is present in all pseudo-components, although an inverted peak in PC_2, and also falls in the olefinic range. Overall, not much conversion chemistry can be proposed from ^{1}H-NMR profiles alone so it is worthwhile to look at the joint decomposition in section 4.2.

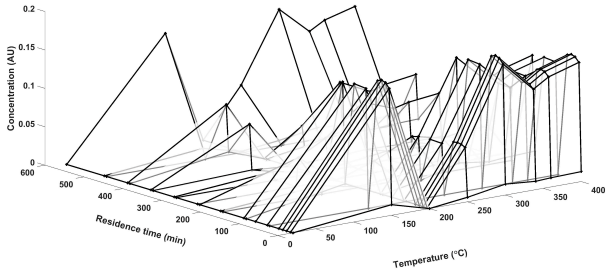
C.5 Gaussian tensor factorization

C.5.1 Individual tensor factorization of FTIR spectra

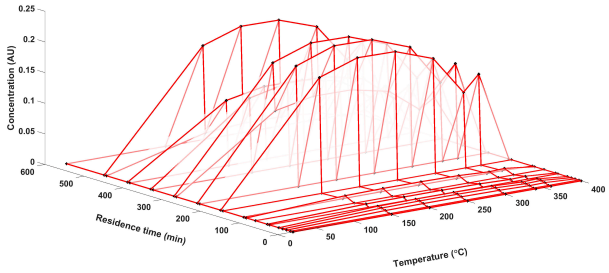
The major peaks in the FTIR spectra of the pseudo-components have been tabulated in Table C.2

Table C.2: Absorption regions for all groups in robust FTIR formulation.

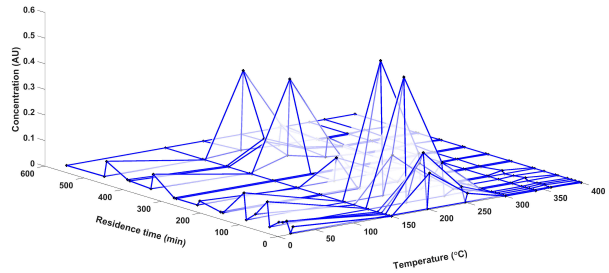
Wavenumber (cm^{-1})	Functional group	Vibration type	PCs/groups present
1597	C=C aromatic	Stretch	PC1, PC2, PC4
1701	C=O of carboxylic acid	Stretch	All 4 PCs
1172 – 1203	C-O of acyl group	Stretch	PC1, PC2, PC3
1018	C-O of aliphatics	Stretch	PC1, PC2, PC3
862	C-H in p-substituted aromatics	Bend	Least intensity but present in all 4 PCs
810	C-H in m-substituted aromatics	Bend	Clearly present in all 4 PCs
740	C-H in o-substituted aromatics	Bend	All 4 PCs but highest for PC4
723	C-H in mono-substituted aromatics	Bend	All 4 PCs – as a shoulder with 740 cm-1
1730	C=O in esters/anhydrides	Stretch	PC4
2360	S-H in thiols	Stretch	PC2, PC3
2150	Alkyne triple bond	Stretch	PC3, PC4



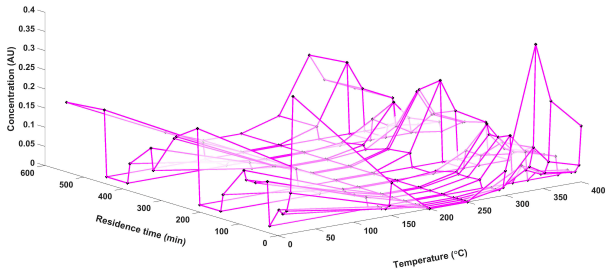
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure C.7: Concentrations of the pseudo-components across the reaction space of the FTIR spectra

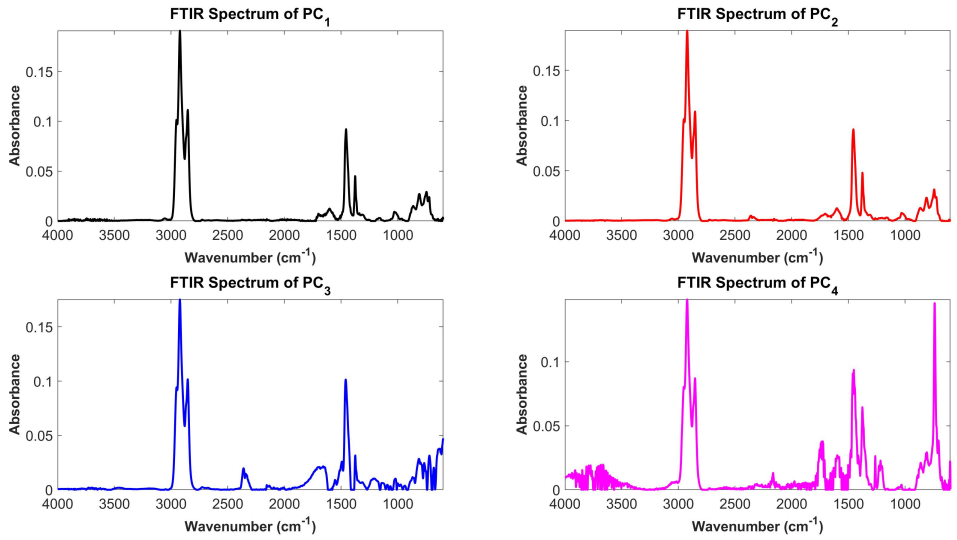


Figure C.8: Spectra of pseudo-components from FTIR tensor decomposition

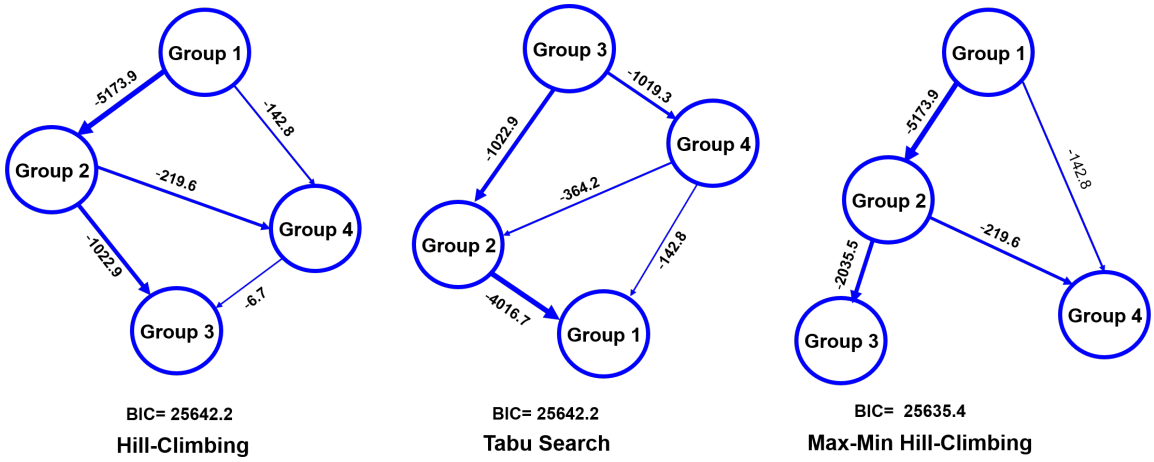


Figure C.9: Bayesian networks from the unique FTIR pseudo-component spectra

---

For PC_1, absorption at 1700 cm^{-1} indicated the presence of carboxylic acid and its co-existence with C-O acylic group at 1175 cm^{-1} confirmed this observation. Presence of aliphatic alcohol was also marked by absorption at 1018 cm^{-1}. All sp^2 C-H bends for aromatics in the 700 – 900 cm^{-1} region were of almost equal intensity (0.035 units) except the p-compounds as already mentioned. The representative compounds for each group are shown in Figure C.10, Figure C.11, Figure C.12 and Figure C.13

that depict the proposed reaction pathways based on the results of Bayesian networks from Gaussian tensor decomposition of FTIR data. Compound (1) is a representative molecule for G1 since it has a carboxylic acid, aliphatic alcohol in the naphthene ring, a side chain and an aromatic ring that is substituted in o-, m- and p- positions. The chemical composition of G2 species is not much different than G1 but it was speculated to be a condensed version of the tri-cyclic compound (1), where the middle or the third ring becomes aromatic in addition to the already existing aromatic first ring. When the middle ring turns aromatic, it leads to a phenolic entity (compound (2)) while if the end ring turns aromatic, it remains an aliphatic alcoholic species. Probability of the end ring turning into aromatic is lower than that of the middle ring due to the requirement of the loss of a lower number of hydrogens but since G2 has a higher intensity for alkoxy C-O absorption (Figure C.8 and Table 1), compound (3) could represent G2 species better. Nevertheless, both compound (2) and compound (3) in Figure C.10 are good representatives of G2/PC2.

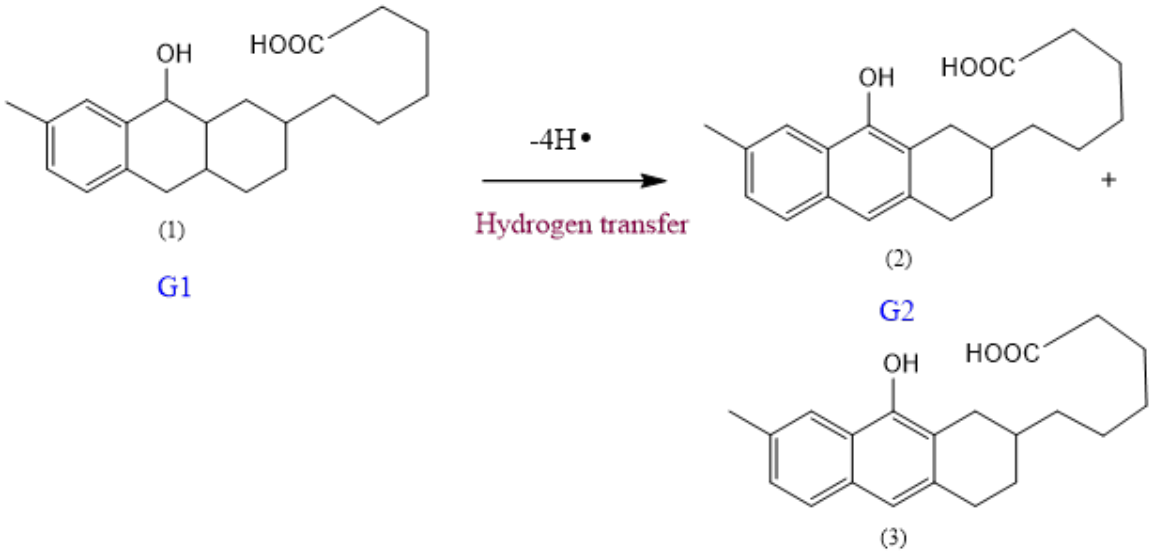


Figure C.10: Proposed reaction pathway of group 1 to group 2 conversion.

Moving on to PC_3, it was interesting to note that although it had aromatic C-H bends in the 700 – 900 cm^{-1} region, it had more olefinic characteristics due to the C=C stretch at 1650 cm^{-1} (Table 1).In order to realize PC_3, we need to look at Figure

C.11 that gives the conversion pathway of G2 to G3 species as proposed from the developed Bayesian networks.

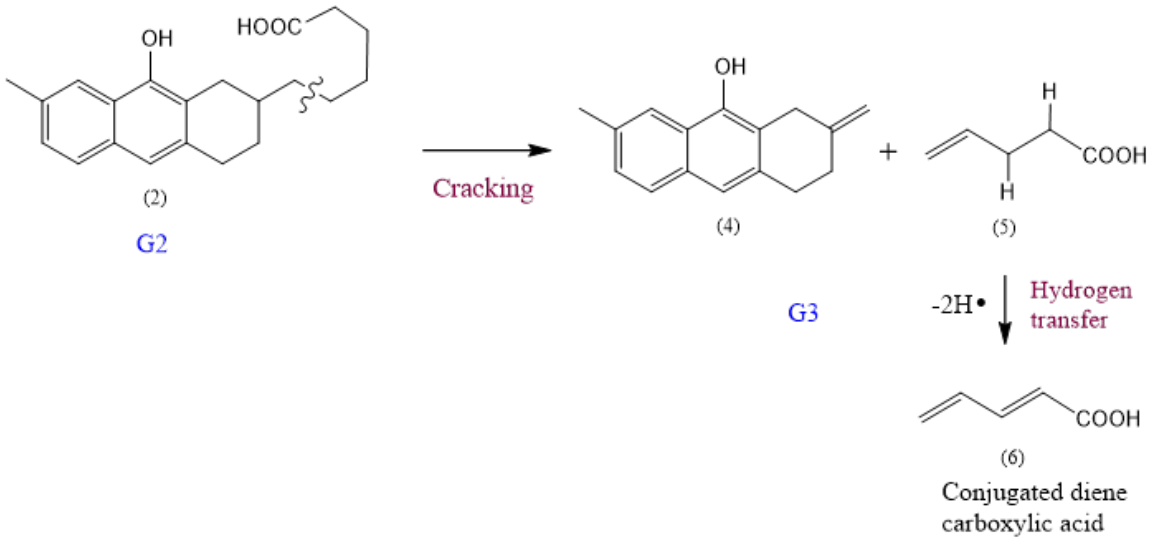


Figure C.11: Proposed reaction pathway of group 2 to group 3 conversion.

Compounds (4), (5) and (6) are all representatives of G3, where (4) has aromatic and olefinic C=C bonds, (5) is an olefinic carboxylic acid while (6) has conjugated C=C double bonds with the C=O of the carboxylic acid group. In a similar way, to realize the composition of PC4, we look at Figure C.12 and Figure C.13 that depict the conversion of G2 to G4 and G1 to G4 respectively. Stretching of C=O at 1730 cm^{-1} and the absorption of acylic C-O at 1202 cm^{-1} for PC_4 indicated the presence of ester/anhydride-type species. Furthermore, among the aromatic C-H bends, the intensity for the ortho-substituted aromatics was the highest.

Compound (9) in Figure C.12 and compound (13) in Figure C.13 are good representatives for G4 species. Compound (9) has 3 fused aromatic rings out of which the first and the 3rd ring excluding the middle one is ortho-substituted while compound (13) is entirely ortho-substituted. Although compound (13) is stabilized by tautomerism due to the olefinic conjugation with the C=O of the ester group, compound (9) is a better representation of G4 since the middle ring has para- and meta- substitutions as well.

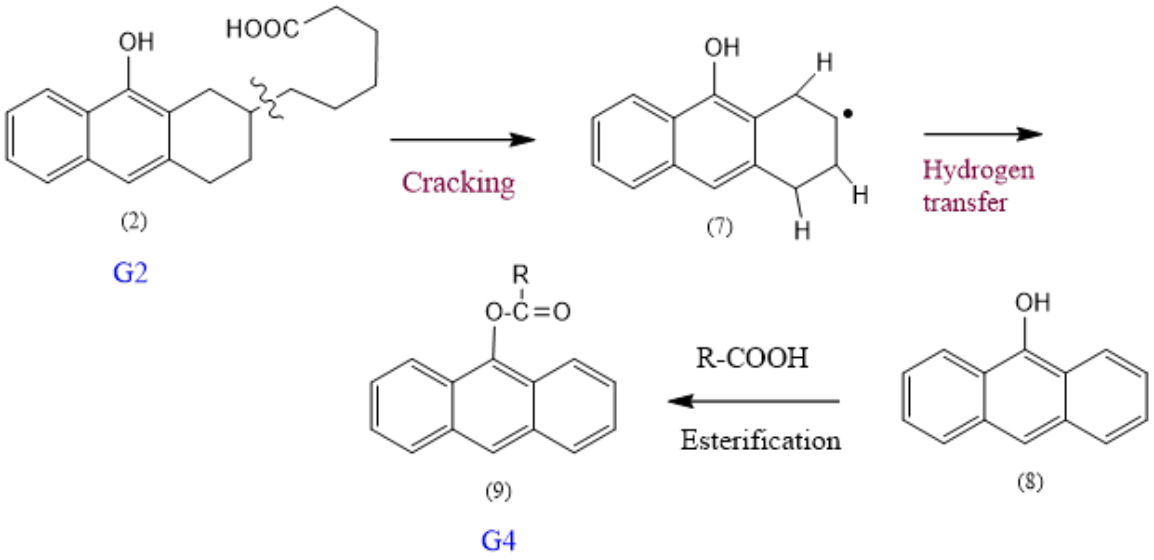


Figure C.12: Proposed reaction pathway of group 2 to group 4 conversion.

Once the representative molecules for each pseudo-component were identified, a reaction pathway was developed according the algorithms in Bayesian structure learning. Here, the hill-climbing and MMHC networks are chosen and the reason for this has been highlighted at the start of this section. From Figure C.9, it can be seen that G1 →G2 has maximum arc-strength indicating the most probable reaction, followed by G2 → G3, G2 → G4 and G1 → G4 in decreasing order. The proposed conversion chemistries are given in Figure C.10– C.13. Conversion of G1 → G2 is the easiest since it involves only hydrogen transfer from the middle or end rings to terminate other free radicals in the bitumen matrix or alternatively get transferred to other aromatics. Bond dissociation energy of benzylic C-H is 301 kJ/mol, which is 30 kJ/-mol lesser than C-H in aliphatics. [422] Compound (2) can undergo cracking in the aliphatic side chain to yield olefins (4) and (5), which can further lose 2 hydrogens to give a conjugated diene pentanoic acid. The conjugated dienoic acid is stabilized by double bond resonance. This chemistry provides a path from G2 → G3, that requires an additional step as compared to G1 →G2 and is depicted in Figure C.11. The same sequence of reactions is possible with compound (3) as the starting material for G2 but in that case, only the olefin will be present only in the carboxylic acid product

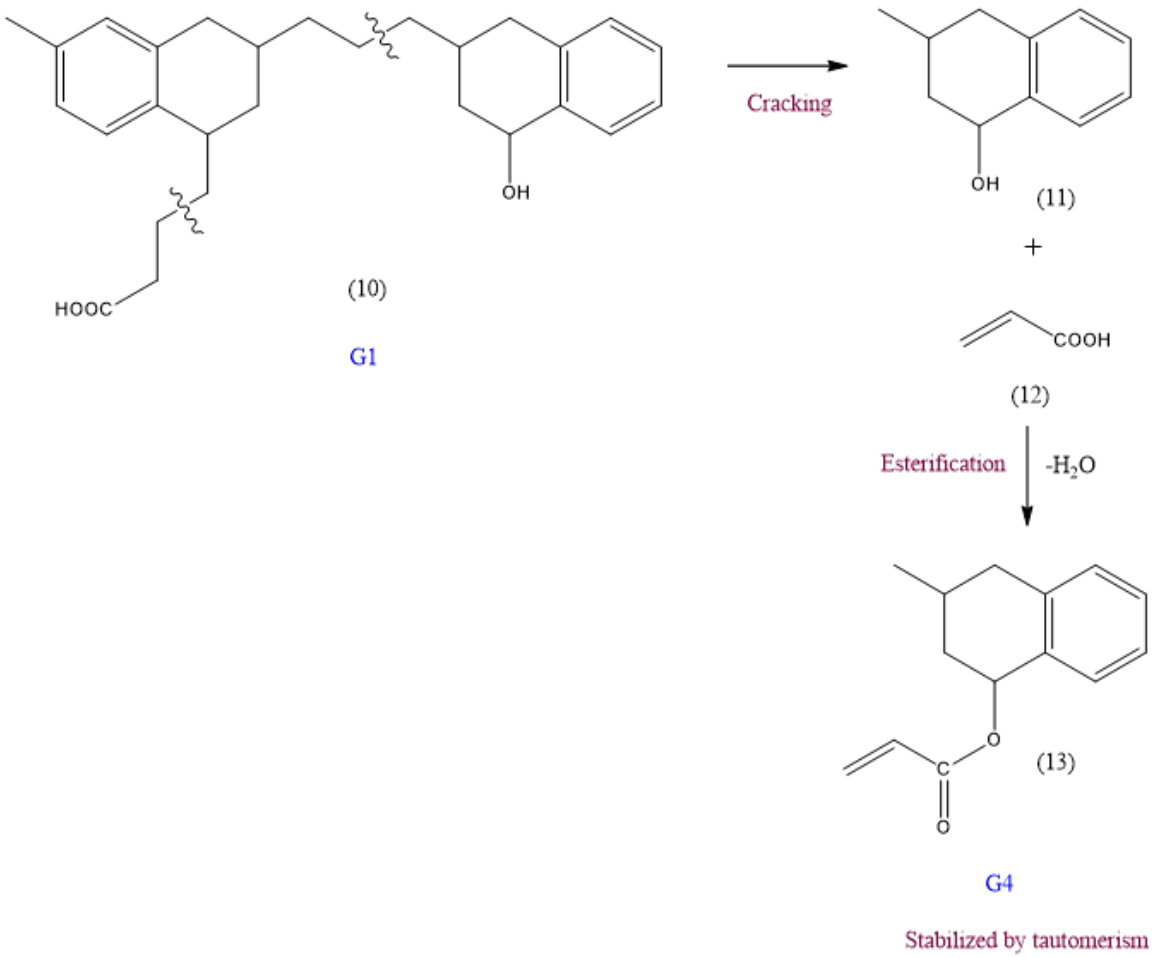


Figure C.13: Proposed reaction pathway for group 1 to group 4 conversion.

and the benzylic free radical would be stabilized by a hydrogen or an alkyl free radical. In this case, the alkoxy group in the middle ring would also exist, supporting the absorption at 1018 cm^{-1} for G2 species.

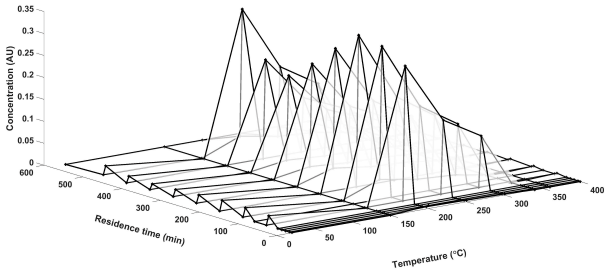
Next, in order to account for the formation of esters from G2-type species, cracking at the carbon attached to the naphthene ring in compound (2) needs to be considered (Figure C.12). This would not be possible in (3) since it is much more difficult to break an sp^2 C-sp^3 C bond rather than an sp^3 C-C bond at these milder reaction conditions of ¡ 400 °C. Once the sp^3 C-C bond breaks, the ring can lose 3 more H free radicals to produce a tri-cyclic condensed aromatic phenol (8) (Figure C.12). This can add to a carboxylic acid from the reaction medium to give an ester (9),

that has all the characteristics of a G4 entity. The path from G2 → G4 involved an additional esterification step apart from cracking and hydrogen transfer through hydrogen disproportionation and hence is concomitant with the Bayesian networks produced from HC and MMHC where this path is the third most probable. G4 → G3 would involve hydrolysis of an ester but that requires the presence of water which is unlikely at these temperatures of bitumen conversion. This could be an explanation for the absence of this path in the MMHC network and being the least probable pathway in the HC network.

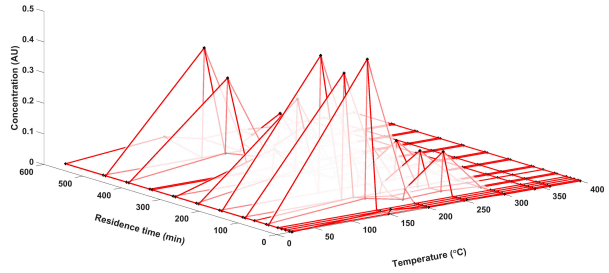
Lastly, to explain the conversion of G1 → G4 even if that was the least probable pathway in the MMHC-produced network, we consider a separate compound that satisfied the absorptions of G1 (compound (10) in Figure C.13). This has characteristics to the archipelago structure [423] of asphaltenes where 2 aromatic cores are bridged by aliphatic chains. Compound (10) can crack in the aliphatic bridge and yield an o-substituted alcohol (11) while the other part is m- and p-substituted as well and is not shown. The side chain possessing a COOH group in (10) can crack and add to (11) and compound (13), which is an ester and also stabilized by tautomerism between the C=C and C=O groups. Compound (13) is another representative of G4.

C.5.2 Individual tensor factorization of ^{1}H-NMR spectra

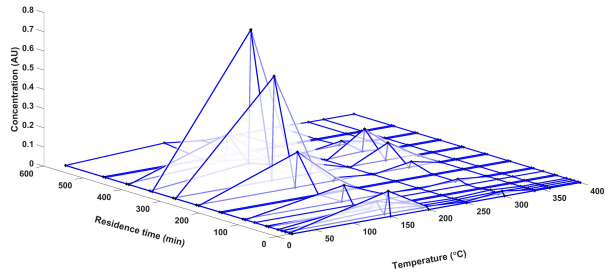
The drawback of this section is that NMR spectra alone does not provide as much information as the FTIR spectra, especially in the aromatic region since it just shows a single overlapped lump from 7 – 9 ppm (Figure C.15). Peaks for aliphatic methylene and methyl protons are distinct and common to all pseudo-components with CH_2 showing higher intensity. All spectra also show the peak for benzylic proton at ∼2.5 ppm confirming the presence of aromatics but does not indicate the number of substitutions. One interesting observation was that PC_3 and PC_4 showed a peak for hydrogen attached to an alkyne group at 3.1 ppm and this was also reflected in the FTIR spectra for the same pseudo-components (Figure C.8). Triple bonds are quite



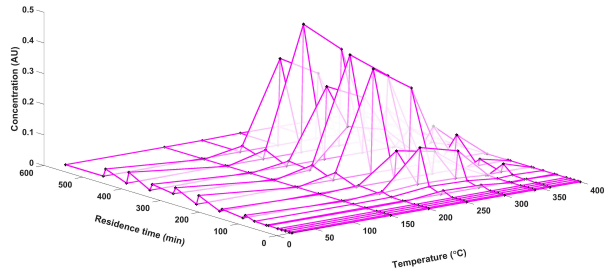
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure C.14: Concentrations of the pseudo-components across the reaction space of the ^{1}H-NMR spectra

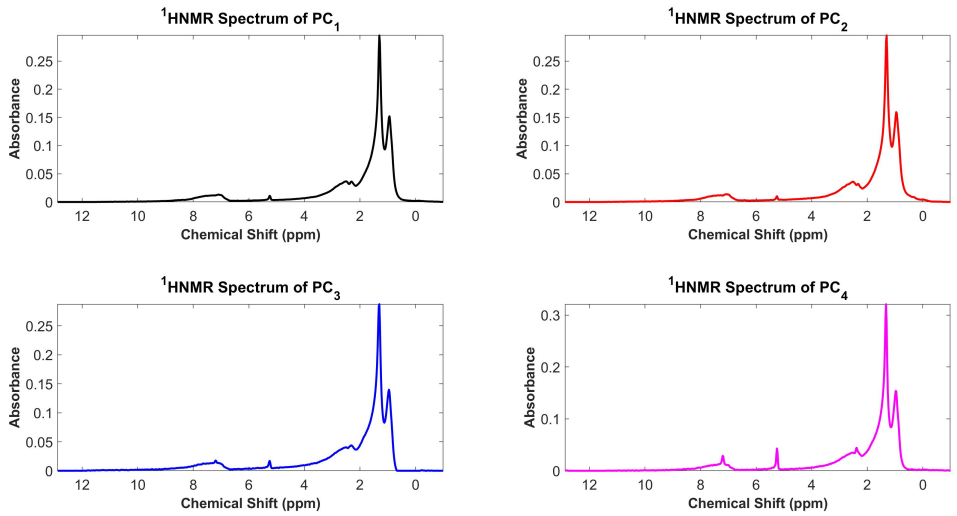


Figure C.15: Spectra of pseudo-components from ^{1}H-NMR tensor decomposition

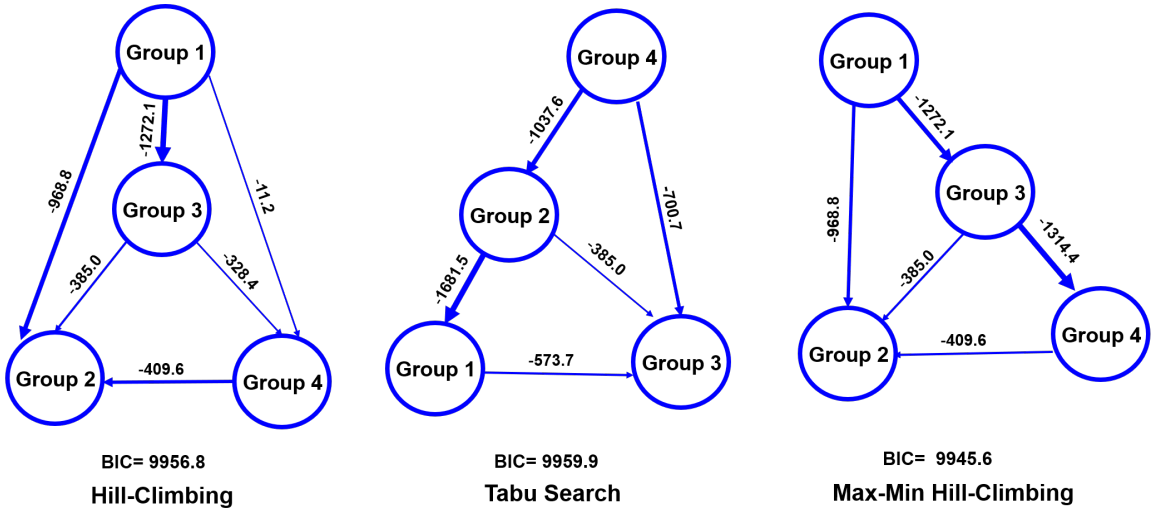


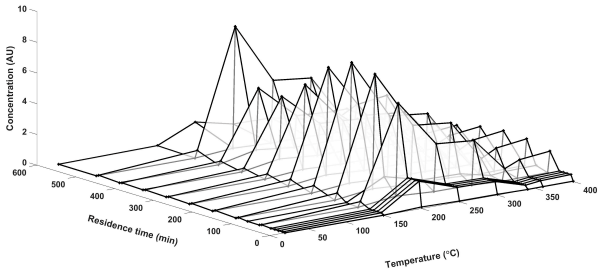
Figure C.16: Bayesian networks from the unique ^{1}H-NMR pseudo-component spectra

stable and their possible participation in the reaction could be such that hydrogens from disproportionation could add across the triple bond. Another distinct characteristic of NMR profiles is the peak at ∼5.2 ppm that depicts hydrogen from methylene chloride that might be remaining in the converted samples. This is present in all pseudo-components but of higher intensity in PC_3, PC_4 and also falls in the olefin range. Overall, not much conversion chemistry can be proposed from NMR profiles alone so it is worthwhile to look at the joint decomposition in Section C.5.3.

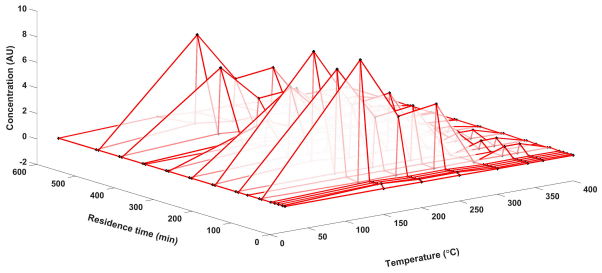
C.5.3 Joint Gaussian tensor factorization

Figure C.17, Figure C.18 and Figure C.19 give the concentration profiles across time and temperature modes, spectral profiles for all 4 pseudo-components and the Bayesian networks obtained from tensor decomposition of FTIR and ^{1}H-NMR fused data, respectively, for the non-robust formulation. The absorption peaks for all pseudo-components were similar to those reported in Table 1. The Bayesian network structures are as reported in Figure C.19. Here, G1 →G3 was the most probable pathway which meant cracking leading to olefin formation had a higher chance of occurring than hydrogen transfer. Alcohol groups in these olefins have more probability of finding carboxylic acids from the matrix to yield an ester (G4) and this pathway is the second most and third most probable in the MMHC and HC network, respectively. This is because carboxylic acids are more prominent and available to react in bitumen than alcohols. [278]

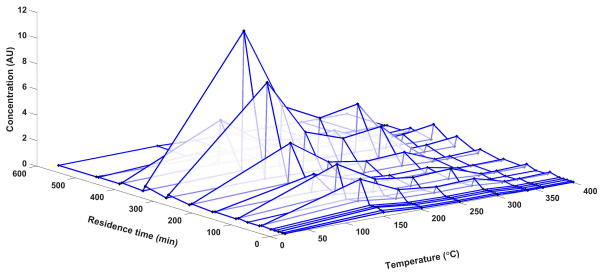
G4→ G2 had the lowest arc strength in the MMHC network and this meant that hydrolysis of esters was least probable which corroborated with the observations from the robust method. In conclusion, the robust method indicates a better flow in the reaction chemistry as hydrogen transfer occurs more easily than cracking. Also, a conjugated double bonded carboxylic acid like (6) that belongs to G3 would react slower than an unconjugated carboxylic acid (G2) to yield G4 esters, which is captured in robust formulation. Hence, overall, it is suggested that the robust formulation gives a better representation of bitumen conversion chemistry at these process conditions.



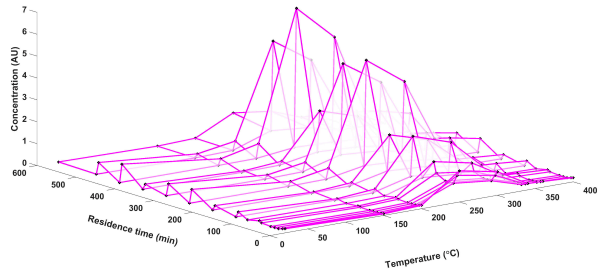
(a) Pseudo-component 1



(b) Pseudo-component 2



(c) Pseudo-component 3



(d) Pseudo-component 4

Figure C.17: Concentrations of the pseudo-components across the reaction space from the joint decomposition of FTIR and ^{1}H-NMR spectra

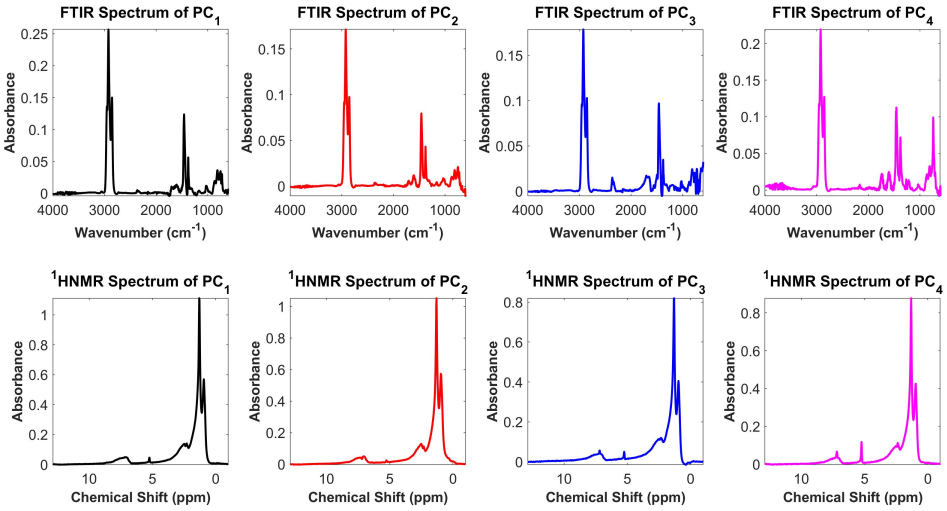


Figure C.18: Spectra of pseudo-components from joint tensor decomposition

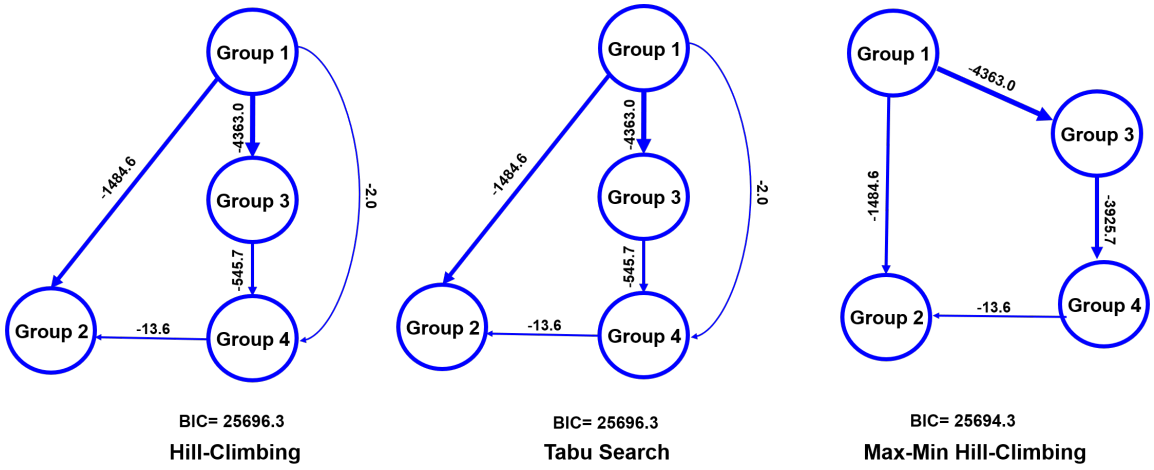


Figure C.19: Bayesian networks from the unique joint pseudo-component spectra

Appendix D: Chapter 5

D.1 Process Conditions

Table D.1: Process conditions for spectral data collection

Spectral sensor	Process conditions	
	Temperature($°C$)	Residence time (min)
FTIR	150	66, 126, 186, 246, 306, 366, 426, 486
	200	66, 126, 186, 246, 306, 486
	250	246
	300	126, 186, 246, 306, 366, 426, 486
	340	6, 66, 126, 246, 486
	360	6, 16.02, 25.98, 36, 66, 246, 583.02
	400	6, 16.02, 25.98, 36, 66, 96, 126
^{1}H-NMR	150	60, 120, 180, 240, 300, 360, 420, 480
	200	60, 120, 180, 240, 300, 360, 420, 480
	250	60, 120, 180, 240, 300, 360, 420, 480
	300	60, 120, 180, 240, 300, 360, 420, 480

D.2 Additional figures for the case studies investigated

D.2.1 Decreasing temperature sequence

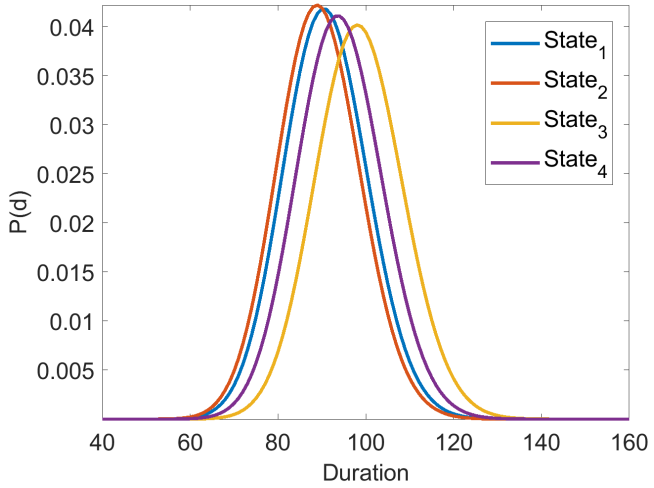


Figure D.1: Duration distribution of the identified modes

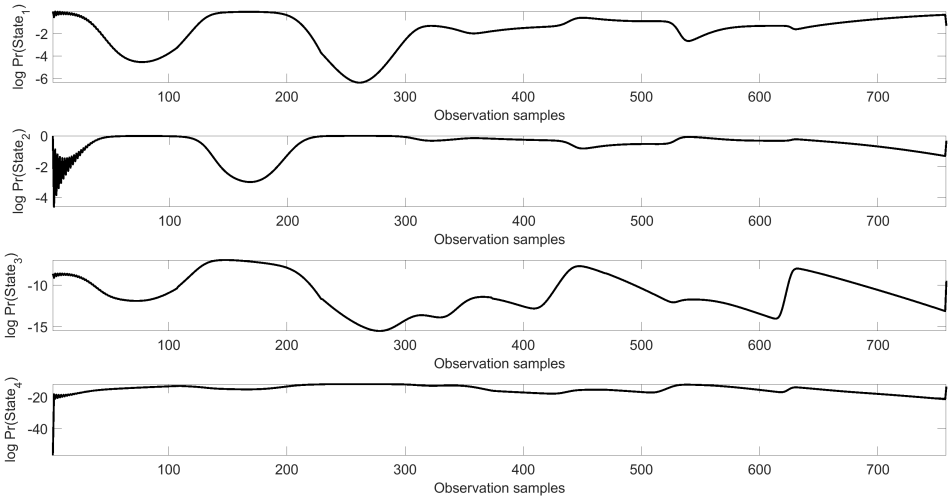


Figure D.2: Posterior probabilities of the states

D.2.2 Randomized temperature sequence using 4 states

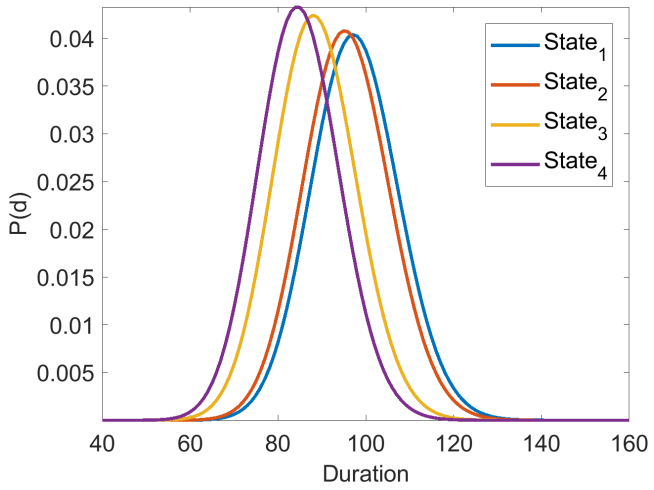


Figure D.3: Duration distribution of the identified modes

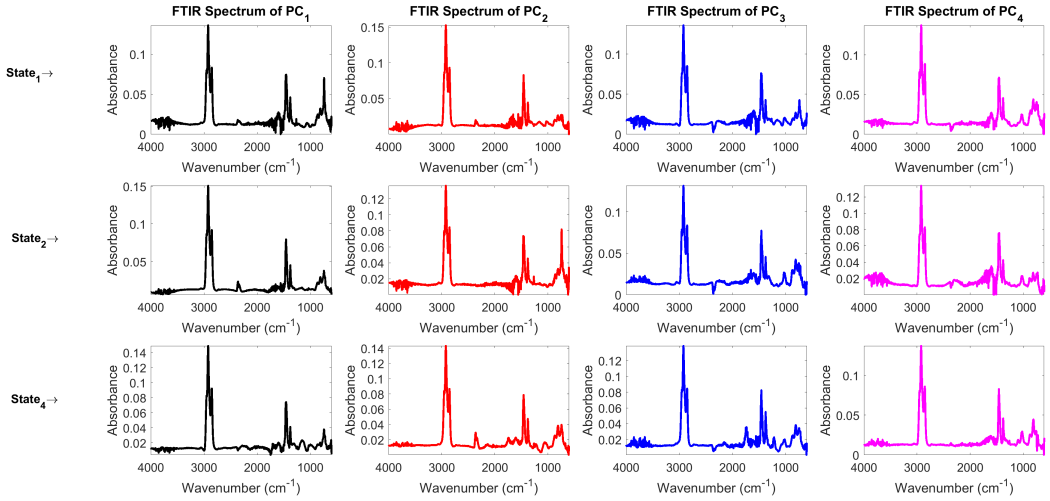


Figure D.4: Pseudocomponent spectra associated with the modes

D.2.3 Randomized temperature sequence using 6 states

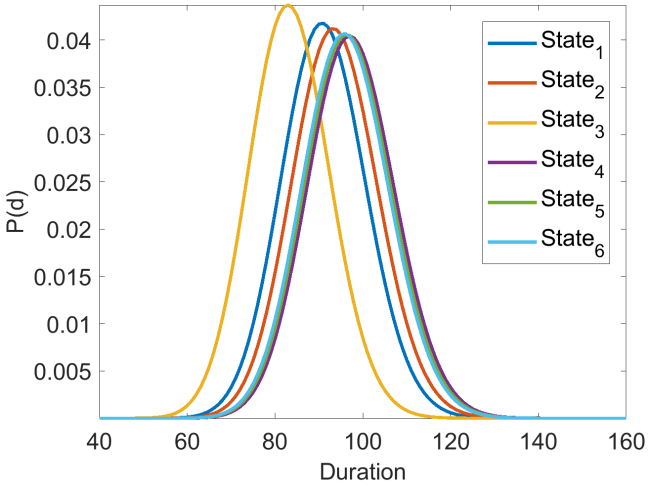


Figure D.5: Duration distribution of the identified modes

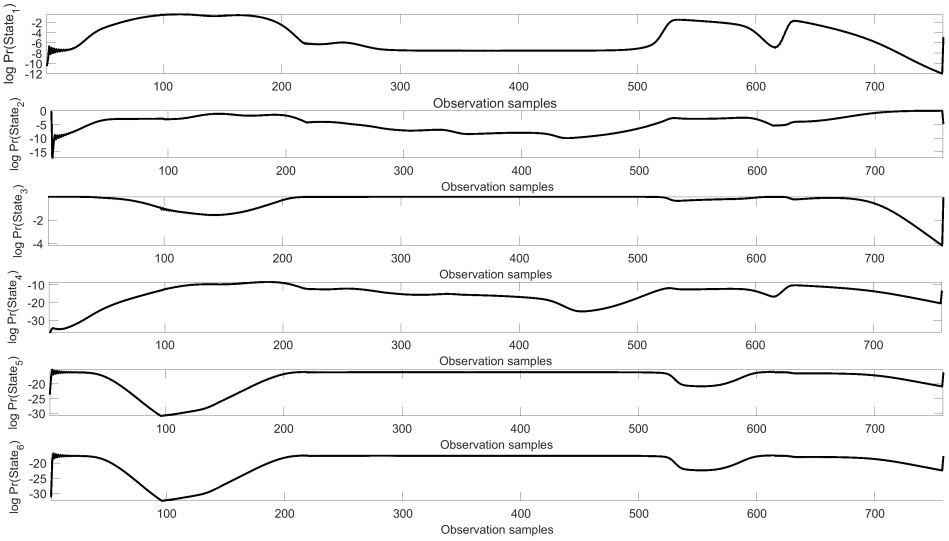
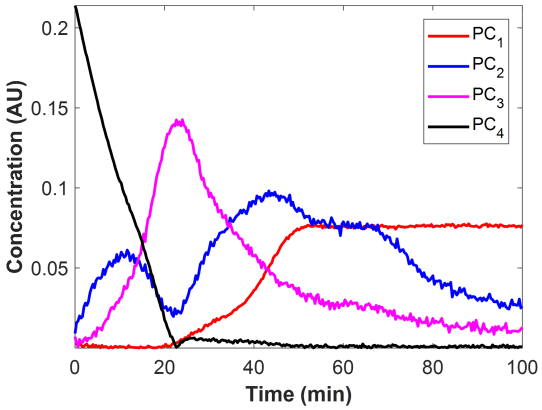
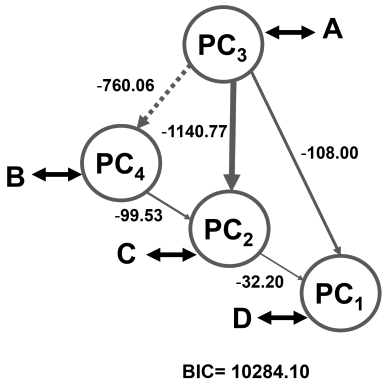


Figure D.6: Posterior probabilities of the states

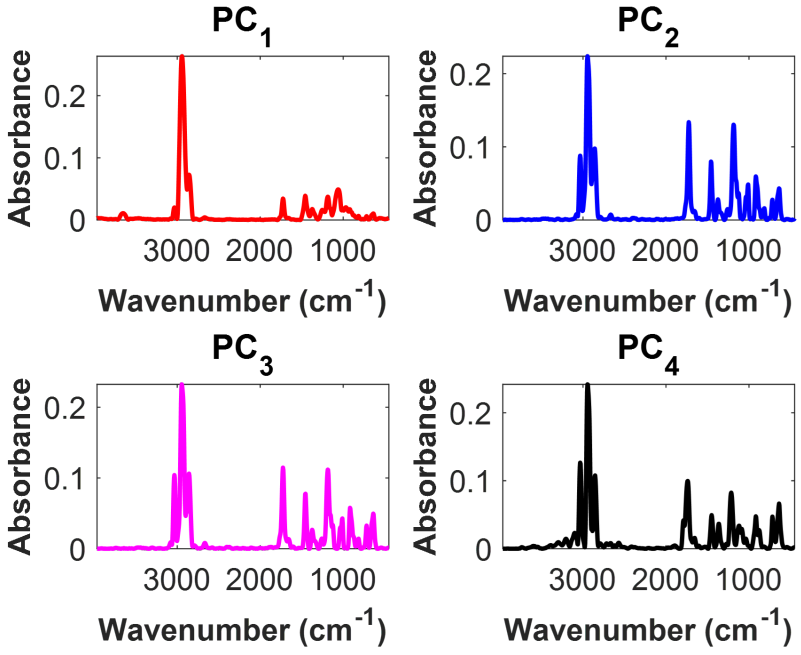
Appendix E: Chapter 6



(a) Temporal concentrations



(b) Reaction network structure, arc strengths and score inferred from the pseudo-component spectra



(c) Pseudo-component spectra

Figure E.1: Spectral deconvolution and causal inference using noisy synthetic data at a signal to noise ratio of 35.

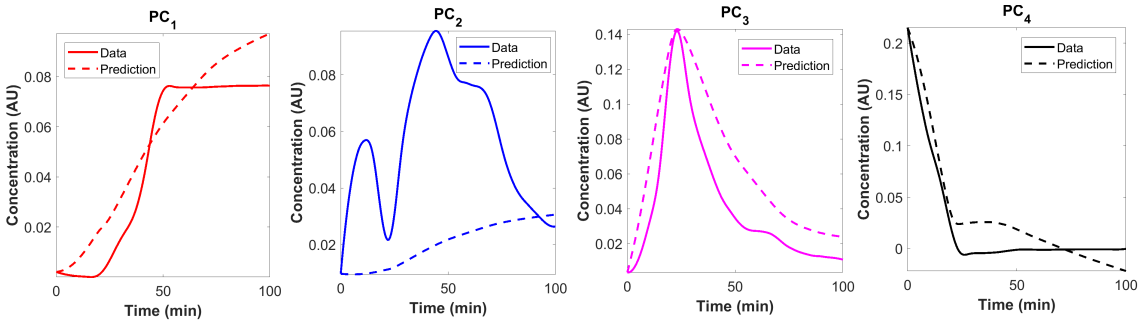
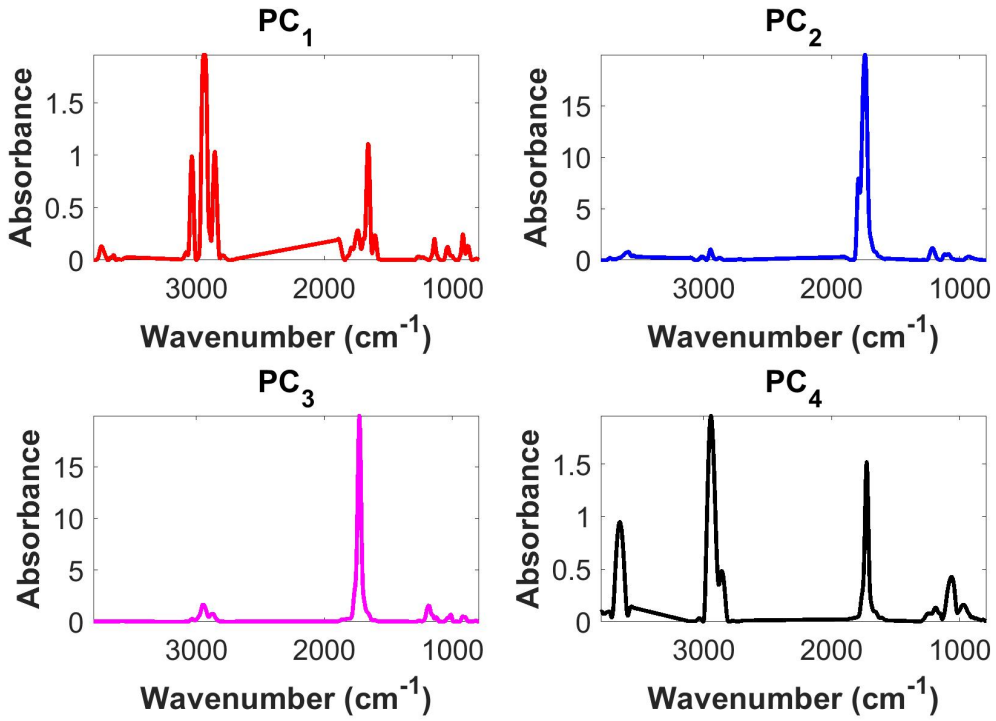
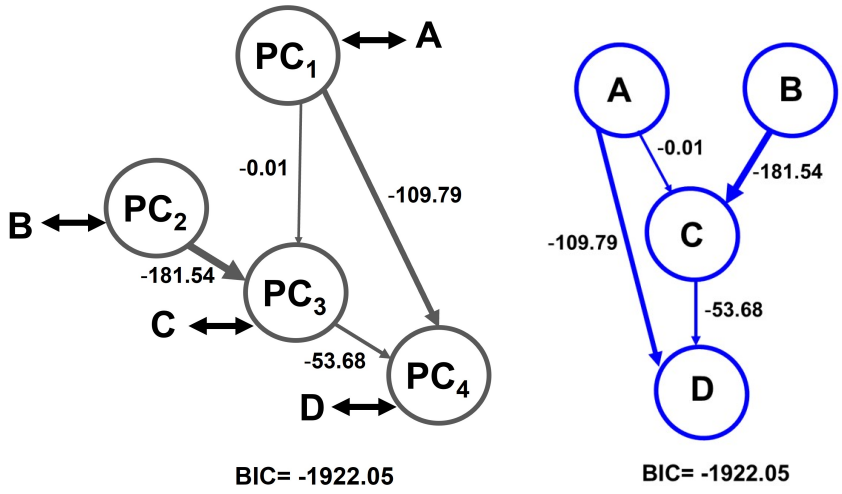


Figure E.2: Comparison of the predictions from the chemical neural ODE against the reconstructed data from integration of the smoothed time derivative of temporal concentration obtained by the deconvolution of synthetic spectroscopic data, at a signal to noise ratio of 35.



(a) Preferentially weighted pseudo-component spectra after deconvolution



(b) Reaction network structure, arc strengths and score inferred from the preferentially weighted pseudo-component spectra

(c) Arc strengths and score inferred from the preferentially weighted spectra, given the reaction template

Figure E.3: Preferential weighting of the wavenumber absorption bands of the deconvolved pseudo-component spectra followed by causal inference using noisy synthetic data at a signal to noise ratio of 100.