Leveraging spectroscopic sensor measurements for development of models for reactions involving complex feedstocks

by

Karthik Srinivasan

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

Abstract

Chemically heterogeneous feedstocks are being increasingly used in process industries due to depletion of conventional feedstocks, to meet environmental demands and to recover value added products from wastes. Chemical modeling of reactive transformations of such complex feedstocks involves tracking the trajectories of multiple reactive species, typically through spectroscopic sensor measurements, to obtain atom level knowledge of the reacting species and can be challenging, especially without any human insight. Interpretation of spectroscopic signatures is an art and traditionally demands a level of domain expertise. Reaction models developed using sensor measurements also require domain expertise and are typically generated by suggesting model compounds for groups of substrates. Including human insight, however, leads to bias in modeling and does not allow for efficient exploration of the chemical space for all possible reactions in the system. Furthermore, updating these expert-guided models based on new operational data is quite cumbersome. This thesis aims to explore the usage of spectroscopic sensor measurements for automation of reaction and kinetic modeling of complex reaction systems by employing machine learning and chemometric methods. The methodologies developed are presented on hydrothermal liquefaction (HTL) of biomass as a case study by utilizing experimental Fourier Transform Infrared (FTIR) and Proton Nuclear Magnetic Resonance (^{1}H-NMR) spectroscopy.

Different methodologies for the identification of reaction networks from spectroscopic data are presented in decreasing order of human intervention required. Spectroscopic curve resolution techniques have been employed at different degrees of sensor data fusion to obtain interpretable and structurally consistent latent features of the reaction system. Sig-

nal level data fusion has been performed through Self Modeling Curve Resolution, while a higher order Joint Non-Negative Tensorial Factorization scheme has been applied at a contextual level to jointly analyze FTIR and ^{1}H-NMR spectroscopic data. Expert knowledge has been used in determination of reactive compounds and the subsequent reaction networks. In a step towards automation, extraction of functional group signatures of the reactive species has been performed through application of convolutional operations on the resolved FTIR spectrum and partial molecular fingerprints for each reactive species have been identified. A reaction network identification methodology that maps spectroscopic signatures to candidate molecules is presented. The network generation is constrained based on the causal structure inferred using Bayesian structure learning and domain knowledge, and employs algorithmically extracted reaction rules obtained through Atom-Atom Mapping (AAM) of reactions from a database. A one-shot molecular generation methodology is presented as the next step in automation thus subverting the need for Bayesian structure learning and spectroscopic deconvolution. Employing a graph neural network based hetero-autoencoder and generative adversarial networks, the molecular generation routine generates molecules constrained by the FTIR spectrum. Localized reaction networks for the process are identified by recursive application of reaction templates. The reaction networks identified have been found to be concordant with reactive transformations recorded in the literature for HTL of biomass. Mathematical modeling of the kinetics of the system based on temporal projection of latent features of the spectroscopic deconvolution has been performed by employing chemical reaction neural networks constrained based on the adjacency information obtained via Bayesian structure learning of the resolved spectrum. Benchmarking studies comparing these neural Ordinary Differential Equations with constrained alternating least squares and basis reduction techniques such as SIND-y is also presented for a synthetic system.

Preface

The research presented in this thesis was conducted under the guidance of Dr Vinay Prasad. The contributions and details of the chapters in this paper-based thesis is as follows,

Parts of Chapter 1 have been published as : Puliyanda, A., Srinivasan, K., Sivaramakrishnan, K., & Prasad, V. (2022). A review of automated and data-driven approaches for pathway determination and reaction monitoring in complex chemical systems. Digital Chemical Engineering, 2, 100009. https://doi.org/10.1016/ J.DCHE.2021.100009. Karthik Srinivasan, Anjana Puliyanda, Kaushik Sivaramakrishnan and Vinay Prasad were involved in conceptualization and writing the original draft of the manuscript.

Chapter 2 has been published as Sattari, F., Srinivasan, K., Puliyanda, A., & Prasad, V. (2023). Data Fusion-Based Approach for the Investigation of Reaction Networks in Hydrous Pyrolysis of Biomass. Industrial and Engineering Chemistry Research, 62(10), 4422–4432. https://doi.org/10.1021/acs.iecr.2c04309. Karthik Srinivasan and Fereshteh Sattari were involved in methodology, formal analysis, implementation and manuscript composition. Anjana Puliyanada provided help with the original draft of the manuscript. Vinay Prasad was involved in conceptutalization, methodology and writing original draft of manuscript.

Chapter 3 has been submitted to Jounal of Chemical Information and Modeling as : Srinivasan, K., Puliyanda, A., & Prasad, V. Identification of reaction network hypotheses for complex feedstocks from spectroscopic measurements with minimal human intervention. Karthik Srinivasan was responsible for Conceptualization, Methodology, Data Curation, Validation, Software, Formal analysis, Writing - Original Draft, Writing - Review

& Editing and Visualization. Anjana Puliyanda was responsible for Formal analysis and Writing-Review & Edit. Vinay Prasad was involved in Conceptualization, Methodology, Supervision, Writing – Review & Editing.

Chapter 4 will be submitted as : Srinivasan, K. & Prasad, V. Spectrum-constrained deep generative model for monitoring of complex reaction systems. Karthik Srinivasan was responsible for Conceptualization, Methodology, Data Curation, Validation, Software, Formal analysis, Writing - Original Draft, Writing - Review & Editing and Visualization. Vinay Prasad was involved in Conceptualization, Methodology, Supervision, Writing – Review & Editing.

Chapter 5 has been published as : Puliyanda, A., Srinivasan, K., Li, Z., & Prasad, V. (2023). Benchmarking chemical neural ordinary differential equations to obtain reaction network-constrained kinetic models from spectroscopic data. Engineering Applications of Artificial Intelligence, 125, 106690. https://doi.org/10.1016/j.engappai.2023.106690. Karthik Srinivasan and Anjana Puliyanda were involved in Conceptualization, Methodology, Results, Original draft preparation. Zukui Li and Vinay Prasad were responsible for Conceptualization, Methodology, Supervision, Writing – review & editing. Parts of this chapter are also a part of Anjana Puliyanda's thesis 'Machine learning-based monitoring of complex reactive systems'.

எண்பொருள வாகச் செலச்சொல்லித் தான்பிறர்வாய் நுண்பொருள் காண்ப தறிவு.

-Kural 424, 1st century BCE

'To speak so as that the meaning may easily enter the mind of the listener, and to discern the subtlest thought which may lie hidden in the words of others, this is wisdom.'

To my loving parents, Srinivasan and Latha

Acknowledgements

I would like to express my immense gratitude to my supervisor, Dr Vinay Prasad, for his constant guidance and grounding support during the course of my doctoral studies. Dr Prasad, thank you for creating an open and safe environment for discussion and tolerating my digressions during our meetings. Your ability to view the larger picture before cracking down on the specifics is something I hope to apply to my future career. I greatly appreciate your gentle nudges and silent backing that have helped me achieve much more than I ever hoped to.

I would like to thank Dr Jinfeng Liu and Dr Zukui Li for their guidance and support during my teaching pursuits. Dr Liu's mentoring during the TW Fraser and Shirley Russell Fellowship has been vital in polishing my instructional skills. I have learnt a great deal as a TA with Dr Prasad, Dr Li and Dr Liu. Working on the control lab has been a rewarding experience.

Thanks to my former and current lab colleagues Anjana, Kaushik, Pedro, Prince, Devavrat, Sushmita.M, Kiran, Kuldeep, Kevin and Shahdab. Our post group meeting coffees will always be memorable. A special thanks to Anjana for all the discussions and helping me find my footing in the initial days. It was a pleasure working with and learning from you. Thank you Kaushik for your help during my first few weeks in Edmonton. Devavrat, Sushmita, Kuldeep, Kevin, Pedro and Shahdab, thanks for putting up with my incessant chatter and for the fun trips.

My clan in Edmonton have made life in the cold a lot warmer. Thanks to my roommates Ankit and Bhubesh for the delicious food, good company and dealing with my crazies.

Sripradha, Manasaa and Suhasini for our impromptu meetups, walks and general banter. Sanjula for the mental health walks, food recommendations, LinkedIn posts and detailed travel itineraries.

Ragu, Miruna, Saurabh and Atul have been my constants since undergrad.

I owe all my success and good things in life to my parents and grand parents. Thank you for the incomparable love, warmth, support and trust that has allowed me to venture out and explore all the possibilities the world has to offer.

Edmonton 2024

Karthik Srinivasan

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Modeling of chemically heterogeneous feed stocks is of concern in the course of optimization of their transformation processes to meet environmental targets, product specifications and market demands [1]. Distributed feed stocks such as biomass pose a problem of being both physically and chemically heterogeneous and hence generating inferential mechanistic models for such systems involves tracking of numerous reactive species across various process conditions and requires complete information regarding all the species participating in the transformation process [2–4]. Measurements of the process from suitably placed sensors provide valuable process data that can be utilized to develop models in such scenarios. Identifying physically realizable models for such systems, especially from sensor measurements, proves to be challenging due to the following reasons:

(i) **Multitude of variables under consideration**. Sensor measurements provide information pertaining to the reaction mixture as a whole and contain signals corresponding to all measurable species present at the sampling instance. While delineating signatures specific to each component is challenging, lumped signatures corresponding to closely related species can be obtained. Developing a reaction network model from the same heavily relies on human intervention and is typically performed by defining model compounds for all major reactive species in the reaction mixture and hence is restricted based on technical expertise of the modeler [5, 6] .

(ii) **Lack of complete information in sensor measurements**.Typically, sensors are placed

to measure a state of interest and cannot provide a comprehensive view of the process without a model correlating unmeasured states to measured ones. In the context of reaction systems, spectroscopic sensors provide molecular level information but cannot represent all aspects of the reaction mixture [7]. Integrating information in a meaningful manner from multiple sensors can provide a holistic view of the system. In purely mathematical models, sensor noise can lead to causal mappings that are not physically realizable.

In this thesis we tackle the aforementioned challenges by incorporating machine learning and chemoinformatic principles to spectroscopic sensor measurements with the goal of reducing (if not eliminating) human intervention in modeling reactive systems.

The rest of this chapter is intended to provide an overview and lay ground for some concepts that provide context for other chapters in the thesis. A more thorough review of these concepts and the motivations for using the same are presented in each chapter as required. The final part of this chapter motivates the rest of the thesis and details its objectives.

1.1 Background

1.1.1 Multivariate curve resolution

Multivariate sensor measurements have allowed for efficient monitoring, control and optimization of processes where limited knowledge is available regarding the constituents *a priori* [8]. From the perspective of process modeling of reactive systems, sensors capable of tracking species level (atomic or molecular) transformations is essential. Spectroscopic sensor measurements provide molecular level information and are useful in identifying atomic or bond level signatures of the constituents of a reaction mixture [9]. These sensors typically capture the effects of interactions between the reaction mixture with electromagnetic radiation as in the case of infrared, UV-visible or Raman spectroscopy, or electrical or magnetic fields such as Nuclear Magnetic Resonance (NMR) or mass spectroscopy. Spectroscopic

sensor measurements have been widely used in characterization of reaction mixtures in pharmaceutical, fertilizer and upgrading processes to name a few [7, 10, 11]. While spectroscopic sensor measurements can be informative, they come at the cost of being high-dimensional, noisy and non-full rank.

The usage of spectroscopic sensors with multi-component systems such as reactors introduces an added layer of complexity involving overlapping peaks and correlated signatures of components evolving over time in the process [12]. The simplest way of dealing with the multi-component issue has been to use linear combination techniques to resolve the spectroscopic data. Principal Component Analysis (PCA), for example, has been used in analyzing anaerobic fermentation reactions and waste water treatment [13–15]. These linear combination methods (specifically PCA or Partial Least Squares (PLS)) aim to describe two-way data (time-evolving spectroscopic profiles) through orthogonal latent variables built as linear combinations of the original feature bases. Imposing orthogonality constraints on the decomposition results in projections that lack any physicochemical relevance and do not allow for incorporation of any chemical or instrumental knowledge into the decomposition [16, 17]. Multivariate curve resolution based on the alternating least squares algorithm (MCR-ALS) was developed to counteract this issue by imposing constraints such as non-negativity, unimodality or even hard constraints such as specific kinetic models [18]. Invoking Beer-Lambert's law, MCR-ALS bi-linearly decomposes a two-way data array into concentration and spectral profiles while incorporating additional instrumentational or physicochemical constraints. The resolved profiles suffer issues of rotational and intensity ambiguity and therefore require careful initialization. Evolving Factor Analysis (EFA) is typically used as a precursor to MCR-ALS in identifying the concentration profiles of each species. EFA performs singular value decomposition (SVD) along the temporal mode to identify the rise and fall of eigenvalues (to be interpreted as unique species) in the data [19]. These trajectories are used as an initialization for the concentration profiles in the MCR-ALS algorithm. Global optimization techniques have also been used in generating feasible initializations

for concentration and spectral profiles.

Higher-order decompositions of the evolving spectral data have been used to overcome rotational ambiguities. Parallel Factor Analysis (PARAFAC) has been used to perform trilinear decomposition of three-way data into a restricted Tucker-3 model [20]. PARAFAC assumes a common latent hyper-cube onto which the resolved profiles (two-way arrays) are projected to build the three dimensional data and preserves the inter-modal dependencies. The quality of the decomposition in all these cases is dependent on the chemical rank (number of species) provided as a hyperparameter to the algorithm. The chemical rank is usually identified through numerically computed metrics such as the Ratio of Derivatives [21] or through identification of orthogonal basis in SVD-like decompositions. Core consistency diagnostics [22] provides a high-order technique for chemical rank identification by analyzing the structure of the hyper-cube obtained as a result of multiple Tucker decompositions.

Data fusion in the context of spectroscopic measurements provides multiple perspectives of the reactive transformation and can be effectively used in developing molecular-level models of the process [23]. Signal-level data fusion such as concatenation of multiple sensor measurements enhances the feature space of the data at the cost of making it highly correlated [24]. Joint analysis of spectral signatures from multiple sources such as hierarchical clustering of peaks allows for better interpretation of the individual components post resolution. Contextual fusion of spectroscopic measurements incorporates knowledge specific to each type of sensor measurements such as instrumental constraints. Joint factorization schemes such as Joint non-negative matrix factorization [25] and Joint non-negative tensorial factorization [26] have been developed taking into account independence of features across sensor measurements while also constraining the decomposition to share factors common to all sets of signals.

1.1.2 Automated reaction outcome prediction

Prediction of reaction outcomes is an integral part in identifying models for a reactive system. The task of exploring the chemical space for plausible reaction outcomes has been approached in a multitude of ways. A common thread among these approaches is the representation of a molecule in a machine-readable format that is informative enough for the prediction task [27]. Automated reaction prediction methodologies can be broadly classified into i) rule-based approaches ii) Potential Energy Surface (PES) exploration and iii) machine learning aided approaches.

Rule-based techniques prescribe a set of valid rules according to which reactions are allowed to occur in the system. The molecules are typically represented as a molecular graph with the rules describing the changes to the connectivity of the graph. In scenarios where knowledge of the feasible reactions of a system is available *a priori*, reaction rules or templates can be manually encoded to capture specific chemical transformations around reaction centres [28, 29], but this comes at the cost of a limited library of reaction rules and performing modifications to the library is cumbersome. Automated reaction templating is an approach that aims to identify reaction rules from a database of chemical reactions. Atom-Atom Mapping (AAM) of reactions allows for tracking of reaction centres and templating reactions [30]. Regardless of their source, templates are applied recursively on the molecules to generate a reaction network [31].

Exploring the PES of a reaction mixture in attempt to identify plausible reaction networks is computationally expensive and requires extensive knowledge regarding the substrates. Typically a minima hopping approach is followed to identify routes connecting potential energy wells in the PES with a search for transition states along the routes [32]. Ab-initio molecular dynamics have also been used for simpler systems in identifying reaction events but still require knowledge of the system to identify reaction co-ordinates to track [33]. Hybrid approaches combining quantum mechanical calculations with rule-based methodologies are typically used to efficiently sample the PES with the rule-based

predictions restricting the region of the PES to be evaluated for the quantum mechanical simulations [34].

Machine learning models typically use neural networks as non-linear universal function approximators to learn mappings between substrates and products. The architecture of the machine learning model chosen relies heavily on the molecular representation used. Methodologies developed for natural language processing have been adapted to string-based molecular representations such as Simplified Molecular Input-Line Entry System (SMILES) and SELFIES. Neural network architectures capable of handling sequential information like recurrent neural networks (RNN) and transformers have been used in one step ahead prediction of reaction outcome based on SMILES string and its variants [35, 36]. *Seq2seq* architectures developed for neural machine translation have also been employed in reaction outcome prediction. These models consist of an encoder and a decoder where the encoder generates a context vector for the substrate by recurrently parsing the SMILES string of the molecule. The context vector is then used as an initialization for an RNN that token-wise predicts the outcome of the reaction. Molecular fingerprints and other vectoral representations are used in combination with deep neural networks to learn a probability distribution over a curated number of reaction classes. Another subclass of techniques perform convolution operations across the molecular graph to identify potential reaction outcomes. For example, a class of models perform repeated convolution operations on the molecular graphs of the reactants with an added global attention mechanism to generate a latent embedding of substrates, which is then input to a series of neural network layers to learn a reaction class distribution [37–39]. Machine learning models behave in a template-free manner, thus allowing for flexibility in terms of evaluating an entire molecule as well as being able to adapt to newer training information.

1.1.3 Generative models for molecules

Molecular discovery has been at the forefront of materials design and drug design in recent times. Conventionally, the task falls at the hands of chemists whose expert knowledge and intuition guides targeted development of new molecules suited for a specific purpose [40]. Traditional discovery methods are only able to sample a small section of the chemical space of theoretically feasible compounds and have higher resource requirements. Machine learning based generative models prove to be a suitable proxy for human intuition by learning latent features over multiple molecular structures [41]. Target-specific drug discovery has been a topic of interest, especially with the ability to machine learning algorithms to model Quantitative Structure Activity/Property Relationships (QSAR/QSPR) [42].

Generative algorithms learn to mimic the distribution of molecules across the support of the representation of the molecules supplied during their training phase. Learning the distribution allows for sampling of newer molecules in the inference phase. The representation of the molecule chosen allows for application of specific types of machine algorithms for the generative algorithm. Sequential notation of molecules such as SMILES and SELFIES have led to the adaptation of text-based models developed for language processing for molecular generation [43–45]. RNNs update their internal state one token at a time to generate an encoding of the molecule in a continuous space which is transformed through non-linear activations to predict tokens in the output sequence. The encodings are stochastically sampled in the inference phase to generate new molecules. Large language models have also been employed to this effect. Point cloud representation of molecules denote atoms as points in a space based on their Cartesian co-ordinates [46]. This three dimensional representation aids in capturing the effects of the different conformer states but can prove to be computationally challenging to obtain sufficient training samples. Graphical representations open up the avenue for the use of Message Passing Neural Networks (MPNNs) to extract latent encodings of the molecule [47, 48]. The direct generation of molecular graphs can be challenging though and have been limited to smaller molecules or a subset of molecules

containing a pre-defined set of atoms or bonds.

Generative models, initially developed for image generation, have been readily adapted towards molecular generation. Variational Auto Encoders (VAEs) consist of a decoder that aims to reconstruct the input from a latent embedding of it generated by an encoder [45, 49, 50]. A regularization term based on the Kullback-Leibler divergence (KL divergence) is added to the loss function to shape the encodings into a standard normal distribution. A closely related class of models are Generative Adversarial Networks (GANs), which generate a surrogate distribution for the distribution of training samples by setting the optimization problem as a min-max game [47, 51, 52]. This adversarial training allows for a more nuanced learning of the probability distribution and has been found to be more capable to generating more realistic samples as compared to VAEs [53]. Normalizing flow models generate invertible mappings of one random variable to another and can be used as a means of stochastically sampling from a latent distribution [54, 55].

Unconstrained generative models are used for exploratory purposes to identify newer sub-regions of the molecular space as well as add to training databases [56]. Evolutionary algorithms, for example, have been used to randomly mutate parts of the molecular graph to generate new molecules [57]. Constrained models have been built to deal with generating molecules that satisfy a particular threshold of property such as partition coefficient or drug-likeness [45, 50, 58]. Optimization routines or reward maximizing approaches such as reinforcement learning have been used to identify molecules with specific properties. Scaffold or structural constraints can also be provided to generative algorithms to restrict the search space to a particular section [59, 60]. In practice these models are built to improve upon a particular scaffold provided as an initialization to a fine-tuning model.

1.1.4 Kinetic Modeling

Kinetic modeling of reaction systems is derived from the law of mass action and addressing closure of mass balance across the reactor system [61]. Pooled approaches such as

the S-system formulation consider all influxes under one power law term and the effluxes under another while generalized mass action models (GMA) consider each reaction term individually [62]. The task of kinetic modeling can be broken down into identification of the topology of the kinetic scheme,i.e, the reaction network structure and the estimation of kinetic parameters. These tasks are cast as individual or joint optimization problems with structure being inferred directly through power law coefficients or as separate variables using integer programming [63]. Under the S-system or GMA the kinetic equations can be derived by constraining the parameter space with the reaction network topology to only include terms with direct influence. A more simplified model can be obtained by moving the equations to a logarithmic space followed by linearization around a specific operating point [64]. Complexities arising from usage of non-linear dynamical models (S-system for example) have been dealt with by describing the model as a weighted sum of linear polynomial basis functions. The basis functions represent a set of elementary reactions identified by model reduction [65]. Noisy process data can lead to multiple network models that result in the same temporal trajectory which is known as the *fundamental dogma of kinetics* [66]. Uncertainty in parameter estimation has been dealt with through stochastic modeling or through pruning of spurious network connections.

Progress of reactions for optimization and control is characterized through the concept of extents. For homogeneous systems, the reactants entering are either converted to products, remain unconverted in the reactor or leave the reactor. A mass balance closure of the reactor provides the extents and allows for identification of the reaction rates [67]. In heterogeneous systems additional mass and heat transfer terms need to be accounted for to achieve mass balance closure. The reaction rates can be deduced independently from the extent of reaction, which is not just a pure function of species concentration or reaction variants (except 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 [68]. 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 [69]. Tendency models are a parsimonious approach to approximate the kinetics of complex reaction systems [70] without prior mechanistic knowledge of the system to predict the dynamic reaction tendencies in transient batch operations [71]. 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 reactor configurations, is already a reduced model as redundant states are eliminated prior to identification, and facilitates estimation of unmeasured species concentration by reconciling the measured concentrations and inlet flowrates with the variant states transformed as extents; finally, integration of extents 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 [72, 73].

Inclusion of physical constraints on the data-driven inferential model subverts the need for user-defined polynomial basis functions and narrows down the solution space. Use of neural networks to learn non-linear dependencies between differentials of concentrations and concentrations of multiple species has been well studied in the literature [74]. Networks have been modified to incorporate network topology constraints and other coupled parameter functions such as the Arrhenius law identify kinetic parameters as the weights and biases of their hidden layers [75]. Evolutionary algorithms have been employed to incrementally build kinetic relationships between measured concentrations of multiple species with no prior knowledge. The genetic algorithm identifies a causally interpretable model for the data through functional forms generated (either mutated or retained) based on a lack

of fit cost [76]. The actual parameters themselves are identified through a secondary optimization task set. Stochastic block models and Markov chain models have also been used in inferring kinetic models from process data [77, 78].

1.2 Motivation

The overarching theme of this thesis is to harness information available from spectroscopic sensor measurements to generate physically realizable reaction models for complex reaction systems. As mentioned in Sections 1.1.1 and 1.1.2, techniques involved in identifying reaction networks typically involve complete or almost-complete information pertaining to molecules participating in the reaction. Sensor measurements and human knowledge are then used as a means of validating the automated prediction. From the perspective of process monitoring, reverse engineering this workflow becomes pertinent and has not been thoroughly investigated yet. A challenge in reversing this process is the lack of complete information regarding molecular structures and the effect of noise in developing inferential models.

Identification of reaction models from spectroscopic sensor measurements requires interpretable (machine or human) mapping of signal peaks to reactive entities. The task of unraveling multiple overlapping peaks in spectroscopic measurements is a well-studied problem, as outlined in Section 1.1.1. The first work in this thesis stems from the question of applicability of MCR techniques to complex reaction systems and the HTL of biomass is chosen as a candidate system. Joint inference of multiple sensor measurements is explored by a signal-level data fusion scheme and the extent of human intervention required in identifying reaction networks for the system is explored.

The task of generating reaction network hypotheses for complex feedstocks has always involved a large amount of human expertise and hence is also subject to human knowledge bias [7]. In complex systems, multiple reactive entites undergo simultaneous reactions and a means of automatically identifying these transformations was found to be lacking. Futher-

more, methodologies presented in literature have been developed for systems where at least the initial substrates are known [28, 31, 37]. Though rich in information, spectroscopic techniques chosen in this work (FTIR and ^{1}H-NMR) do not provide information regarding the complete molecular structure. A methodology dealing with partial information provided by spectroscopic measurements while concurrently leveraging domain knowledge of the process was found to be lacking in the literature.

Process sensor measurements can be noisy which leads to multiplicity in deconvolution of spectral peaks. Process noise was found to affect multiple steps of the network identification process developed in Chapters 2 and 3 and hence an one-shot molecular prediction routine that bypasses multiple modules in these workflows was required. Constrained generative models for molecules have been built to optimize a specific scalar or vector target property [45, 57], but have not been developed for continuous and correlated conditions such as spectroscopic measurements.

Developing kinetic models for complex systems has always been riddled with solution multiplicity and inclusion of physical constraints have been found to restrict the solution space. While chemical reaction networks provide a means of chemically modeling the system, control and optimization of the process require a numerical (kinetic) model. Temporal projections of deconvolution provide an interpretation of concentration, but inference of kinetic parameters from the same can be onerous due to process noise and requires infusion of process constraints to obtain causally interpretable kinetic models.

1.3 Objectives

The primary objective of this thesis is to explore the automation of identifying models for a complex reaction system from sensor measurements. To this effect we employ techniques of Multivariate Curve Resolution (MCR) in tandem with chemoinformatic and machine learning on spectroscopic sensor measurements to identify potential reaction network hypotheses and generate a kinetic model for the system.Figure 1.1 depicts the different routes

in which spectroscopic sensor measurements have been employed in this thesis. The following thesis objectives have been realized:

- Signal-level data fusion followed by MCR has been deployed to identify individual reactive components in hydrothermal pyrolysis of biomass. Expert knowledge has been used to develop a reaction network for the system and is used as a baseline for comparison with more complicated methodologies that follow.

- Structure preserving tensorial factorization has been used to jointly analyze multiple sensor measurements. An improved convolutional neural network has been trained to identify functional groups present in FTIR spectroscopy data. A rule-based reaction network generation scheme with in-built causal constraints has been developed as a step towards automating the reaction network identification.

- In order to subvert the need for curve resolution and causal inference techniques, a one-shot machine learning based generative model has been developed to propose candidate molecules for a reaction mixture directly from spectroscopic sensor measurements. Localized reaction networks for each operating condition have been identified in an automated fashion using a rule-based reaction prediction method.

- Projections of evolving spectroscopic data along the time mode have been utilized in identifying a kinetic model for the system by harnessing backpropagation algorithms used in neural network training. A structurally constrained neural ODE has been trained and benchmarked against an ALS-type constrained optimization routine.

1.4 Datasets

All methodologies described in this thesis are presented using hydrothermal liquefaction of biomass as a model system. A brief description of the experimental data acquisition and the structure of the dataset used in each subsequent chapter are as follows.

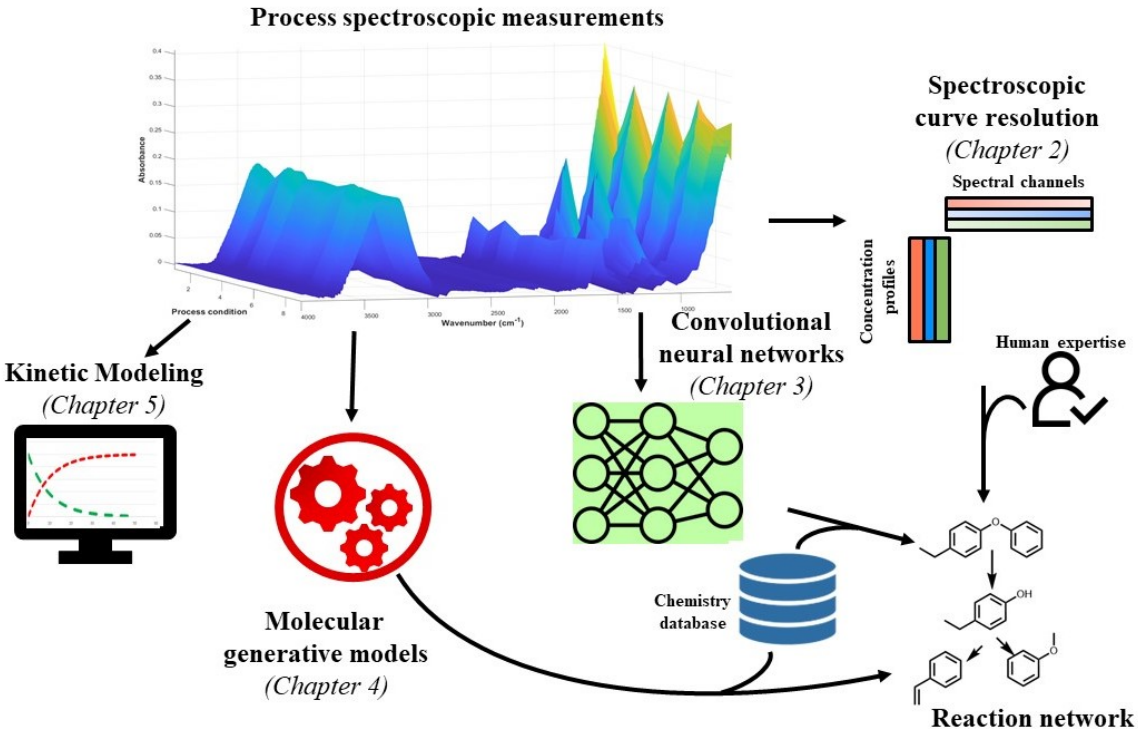


Figure 1.1: Usage of spectroscopic measurements of process in developing models for complex systems

Moneterey pine biomass was subjected to hydrothermal liquefaction in a batch reactor at temperatures of 150°C, 200°C and 250°C for batch times of 15, 20 and 25 minutes under acidic and alkaline conditions resulting in 27 process conditions. FTIR and ^{1}H-NMR measurements of the reactor effluent were obtained with water being the solvent medium.

In Chapter 2, data matrices of the shapes 27×1769 and 27×2084 were generated for the FTIR and ^{1}H-NMR measurements respectively by stacking spectra at each process condition together.

In Chapter 3, three dimensional tensors of shapes $3 \times 9 \times 1769$ and $3 \times 9 \times 2084$ for FTIR and ^{1}H-NMR were generated by combining the data for different catalytic conditions along the temperature mode. This 3-d tensors are a result of vertical stacking of 2-d matrices of data from Chapter 2.

In Chapter 4, each sensor measurement is treated individually, resulting in 27 1×1769 and 1×2084 data vectors.

Chapter 5 uses synthetically generated obtained by mixing spectra of compounds extracted from databases.

1.5 Thesis structure

The remainder of this thesis consists of five chapters arranged as follows:

Chapter 2 describes methodologies involved in identifying spectroscopic signatures of individual reacting components and using human intuition to develop plausible reaction networks for the HTL of biomass. The signal-level data fusion technique, Self-modeling MCR and Bayesian structure learning are elaborated upon and the validity of the generated reaction networks is described.

Chapter 3 details an automated workflow towards identifying reaction networks for a complex system. Two key results are presented in this chapter. An improved convolution-based functional group identifier is presented as means of automating reactive centre identification. This is followed by the description of a causally constrained rule-based reaction network generation routine. Potential limitations and areas of improvement are also discussed.

Chapter 4 describes an one-shot candidate molecule identification scheme and its subsequent use in reaction network generation. A hetero-autoencoder based on convolution operations on graph structures is used in tandem with a Generative Adversarial Network to generate molecules that conform to a given FTIR spectrum. Application of this generative model towards reaction network generation is also presented.

Chapter 5 focuses on usage of temporal concentration data obtained from MCR towards kinetic modeling of the system. A neural ODE solver constrained on the structure of the reaction network is presented and benchmarked against a structure constrained ALS routine towards kinetic parameter estimation. The effect of signal strength on deconvolution and Bayesian network structure learning are discussed.

Chapter 6 serves a conclusionary piece that summarizes the findings of this thesis along

with the limitations and includes potential avenues for exploration for further studies.

Chapter 2

Data Fusion-Based Approach for the Investigation of Reaction Networks in Hydrous Pyrolysis of Biomass

2.1 Abstract

In this work, we present and validate a methodology for generating reaction networks from spectroscopic data using data-driven methods by applying it to the hydrothermal liquefaction (HTL) of Monterrey pine biomass and its constituents, viz., cellulose and lignin. This work is presented as a step toward automated inference of chemistry of the hydrothermal liquefaction process, thus limiting the need for human expertise. Bayesian hierarchical clustering of spectra and self-modeling multivariate spectral curve resolution are used to generate groups of chemically similar species, the reaction networks among which have been developed using Bayesian networks. Fourier transform infrared spectroscopy and proton nuclear magnetic resonance spectroscopy-based measurements are used as input data. The data-driven reaction network includes pathways representing decomposition of the biomass components, large molecule hydrolysis, and reformation of produced molecules and is consistent with the literature. Furthermore, the comparison of the networks generated for biomass and its components (levoglucosan, representing cellulose, and 2-phenoxy-ethyl benzene, representing lignin) reveals the relationship between the biomass HTL reaction network and the reaction networks of the components. The data-driven approach provides

a diagnostic tool to identify the most probable reaction chemistry for complex biomass feedstocks and can be used for process understanding, design, and control.

2.2 Introduction

Biomass is a cheap, sustainable, CO2 – neutral renewable resource which can prove to be an effective substitute for conventional fuels [79]. The conversion of biomass to useful products occurs through biochemical or thermochemical pathways. Pyrolysis, a thermal conversion of biomass, generates bio-oil as a primary product with many uses, including serving as a precursor for chemicals in biorefineries [80, 81]. The gases produced during pyrolysis can be utilized in heat and power generation [82].

Hydrothermal liquefaction (HTL) is a conversion process that converts wet biomass to bio-fuels and other value-added chemicals [83]. Water at critical conditions is an essential reactant in the HTL process resulting in quick, homogenous and active reactions [84]. At temperatures below 400°C, where the HTL is usually performed, hot condensed water is used to produce biocrude with about 10-20% oxygen [85]. The product yield and physico-chemical properties of the HTL are mainly affected by variability in feedstock, processing conditions (temperature and reaction times) and the choice of catalyst[86].

Biomass can be characterized by the relative compositions of its three main constituents-cellulose, hemicellulose and lignin. The pyrolysis of biomass generates products that are equivalent to the total sum of the individual pyrolysis of the three constituents. Bio-oil is a combination of hundreds of compounds that are produced from the depolymerization of cellulose, hemicellulose and lignin. The oxygen and water composition of bio-oil ranges from 40-50% and 25-35%, respectively[87] and the chemical nature of bio-oil is firmly linked to the ratio of the components in the biomass [88]. Consequently, it is vital to comprehend the molecular composition of bio-oils to understand their properties and stability. Reaction networks provide a way to describe the synergy among the constituents of the system, providing a means for monitoring and control of the process. The development of reaction

networks for such a physically and chemically heterogeneous system requires much human expertise and is usually done through representative mixture models or through correlations [88, 89]. The workflow described herein infers chemistry of the process directly from spectral measurements of the feedstock, and unveils the causal information inherently present in the reactive transformations, thus reducing human bias.

In this work, we develop a diagnostic tool for the reactions of biomass and use it to describe the HTL process of Monterey pine whole biomass in the presence of water at different conditions. The diagnostic tool, which identifies the most probable chemistry, can then be used to advance systems engineering applications by way of monitoring, optimization and control of chemically reactive processes [90]. Importantly, the tool only needs spectroscopic (or similar) data that is easily available on-line or at-line in a process, i.e., the analytical or characterization requirements are not complex. It is also important to note that we do not attempt to develop an optimized HTL process, but instead use HTL at different operating conditions to elucidate the details of our approach and its diagnostic capabilities.

In this study, samples from the HTL process were studied and distinguished using Fourier Transform Infrared (FTIR) spectroscopy and proton nuclear magnetic resonance (^{1}H-NMR) spectroscopy. Spectroscopic measurements provide information at the molecular level along with physical process parameters like temperature, pressure, flow rate and liquid level being measured from other sensors [91]. A data fusion approach is subsequently used to combine information from the two types of spectroscopic measurements when inferring the chemistry of the process. The essence of data fusion is to link the data from several sensors to carry out deductions that cannot be acquired from a single sensor [92], as demonstrated by jointly analyzing spectroscopic data to capture complementary information in reactive systems [25, 26]. Input data from various sources might involve parametric data linked to the object identity, thereby providing a holistic view of the reaction scheme incorporating various distributed sources [93].

A major issue in computational fields like biology or chemical processes is the preva-

lence of high dimensional datasets to study the network architecture of the variables accurately [94, 95]. Functional connectivity is often illustrated in terms of statistical reliance, and it is also seen as a practical theory that controls the discovery of a functional connection without interpretability on how that connection was made. It could also be illustrated as a dependency test between two or more time series used to decline the null hypothesis of statistical independence. This is similar to evaluating the collective information and experimenting for critical departures from the null hypothesis [96]. The Granger causality technique and the Bayesian network inference technique are two procedures frequently used to infer interactions among a set of elements [97, 98]. Several studies have been conducted on the systematic and computationally intensive comparisons between the two techniques on the synthesized and experimental data, and it was inferred that the Bayesian network (BN) inference is more preferable to the Granger causality approach for small datasets [99]. This research generated a high dimensional dataset for a limited number of experimental samples; thus, the BN approach was used. A BN is a probabilistic structure learning method that represents the joint probability distribution among a set of random variable nodes as a product of the distributions of the child nodes conditioned on its parents, such that the likelihood score of the network structure, i.e. the Bayesian Information Criteria (BIC) is maximized. The Tabu [100] and Hill-climbing [101] heuristic structure learning techniques and the hybrid method (Max-min Hill-climbing) [95] were used to learn the optimal network structure that maximizes the BIC. Comparison of the reaction network generated by our method with reactions described in literature about the HTL of biomass indicates the high fidelity of our methodology. The transformation of cellulosic and hemi-cellulosic structures to furfural and its derivatives, as well as carbonyl structures has been captured by our network. Similarly, the decomposition of lignin type compounds to phenolic structures is represented in the network generated.

2.3 Materials and Methods

2.3.1 Materials and HTL

An overview of the entire workflow is described in Figure 1.

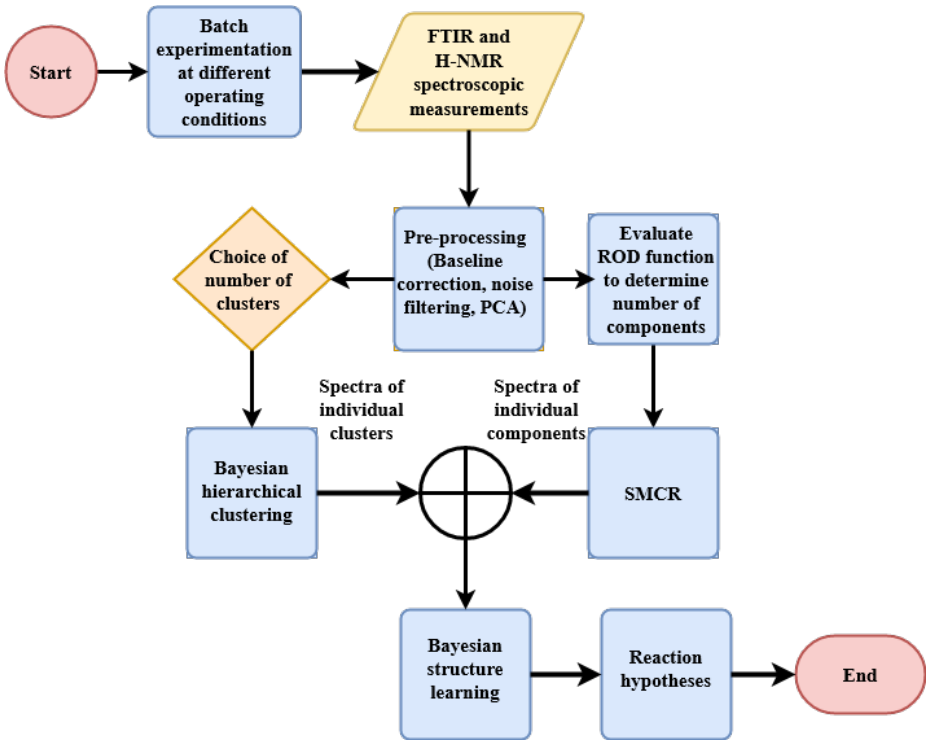


Figure 2.1: Flowchart of methods

The data used in this work was acquired from the experimental survey of hydrothermal decomposition of Monterey pine whole biomass obtained from Sigma Aldrich Canada. The refined biomass specimen was manufactured by thermal decomposition. A stainless-steel micro batch reactor of 24 cm length and 2.1cm width was used and the solvent used was subcritical water. The procedure for this experiment has been described in our previous work [24]. Twenty-seven liquid samples were examined in this research. This study was performed at different at temperatures of 150, 200 and 250°C, with reaction times of 15, 25 and 35 minutes for each temperature. The initial pressure was fixed at 0.1 MPa by shutting off the pressure relief valve. Catalysts are essential for HTL because they influence the rate of reaction and the structure of HTL products. Many homogeneous and heterogeneous

catalysts have been analyzed by other researchers for the catalysis of biomass HTL, even though the larger part of the work was centered around homogeneous catalysts (acid, alkali, and metal salts) because they are quite affordable [102]. A typical feature of homogeneous catalysts, also, is that they produce liquid products that are not affected by coking [103]. Due to this, 0.05M of sulfuric acid and 1M of sodium hydroxide were used as catalysts for every temperature-residence time combination in this work resulting in 27 process conditions. The volume ratio of biomass to the medium was 1:10, and the end products were stored in a glass beaker prior to analysis.

2.3.2 Spectroscopic analysis and data fusion

The use of FTIR and ^{1}H-NMR spectroscopic measurements in this study is motivated by the fact that inline spectroscopic measurements provide molecular level descriptions of the system while also being fast, reliable and low cost [104]. FTIR spectrometers also allow for the characterization of bio-oil specimen despite its high volatile component [105]. The IR spectra of the fluid specimens were obtained using an ABB MB 3000 FTIR spectrometer. The spectra were obtained at a resolution of 8 cm^{-1} in the normal spectra range of 4000-600 cm^{-1}. All the spectra had a numerical mean of 120 scans. The measurement was conducted on liquid samples with the aid of a pike miracle attenuated-total-reflectance attachment. The spectroscopic results obtained are illustrated in Figure 2 and Figure 3. The handbook of spectroscopic data was used to label the functional groups [106].

The hydrogen and carbon atoms in different functional groups in the reaction mixture were identified using 1H-NMR spectroscopy. This was performed using a NMReady instrument at a frequency of 60MHz and a full width half maximum (FWHM) resolution <1.0 HZ (20 ppb). Multi-sensor measurements of the same process produce data from different domains and capture multiple facets of the process that may not be uncovered with a single measurement [107]. Data fusion is a productive means of combining such independent measurements to reveal shared as well as distinct information. In this work, a signal-level

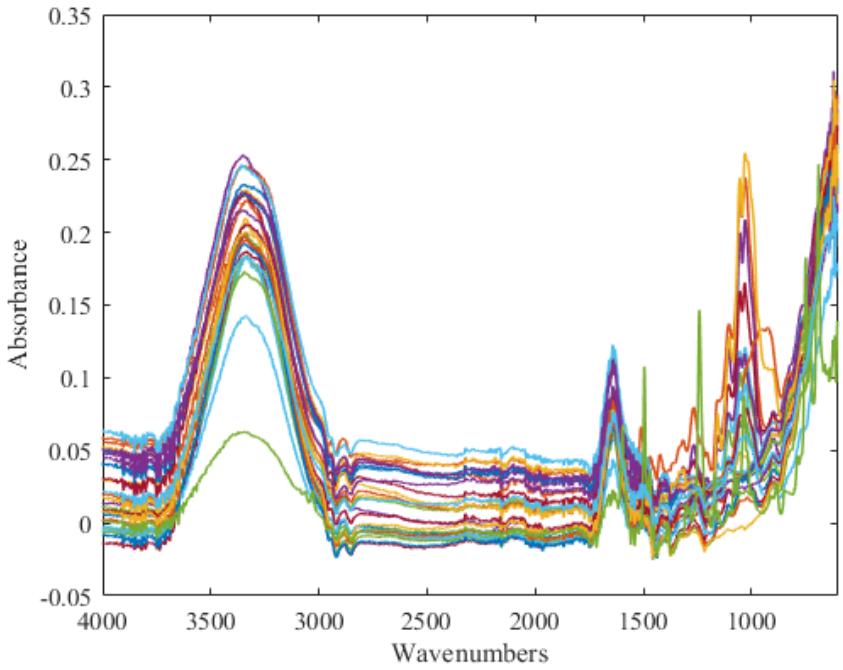


Figure 2.2: FTIR spectra of HTL of biomass under 27 different conditions.

data fusion approach was followed [108] to concatenate measurements from the two spectroscopic devices [24, 109]. The data was scaled and normalized with respect to minimum peak intensity prior to fusion, along with dimensional reduction (removal of uninformative variables/wave numbers) using Principal Component Analysis. The resultant signal had a better signal to noise ratio compared to the raw data from the individual signals. A key assumption considered in subsequent steps is that the FITR and ^{1}H-NMR spectra are informative enough and capture the dynamics of the transition between species. The informativeness of the data also depends on sufficient excitation of states (different reactions in the process) by the manipulated variables (temperature and batch time).

2.3.3 Self-modeling multivariate curve resolution

SMCR is a method to bilinearly decompose a spectral data matrix $D \in \mathbb{R}^{m \times n}$ comprising m samples recorded across n spectral channels (wavenumbers for FTIR and chemical shifts for 1H-NMR, into the product of the concentration $C \in \mathbb{R}^{m \times r}$ and pseudo-spectra $S \in \mathbb{R}^{r \times n}$

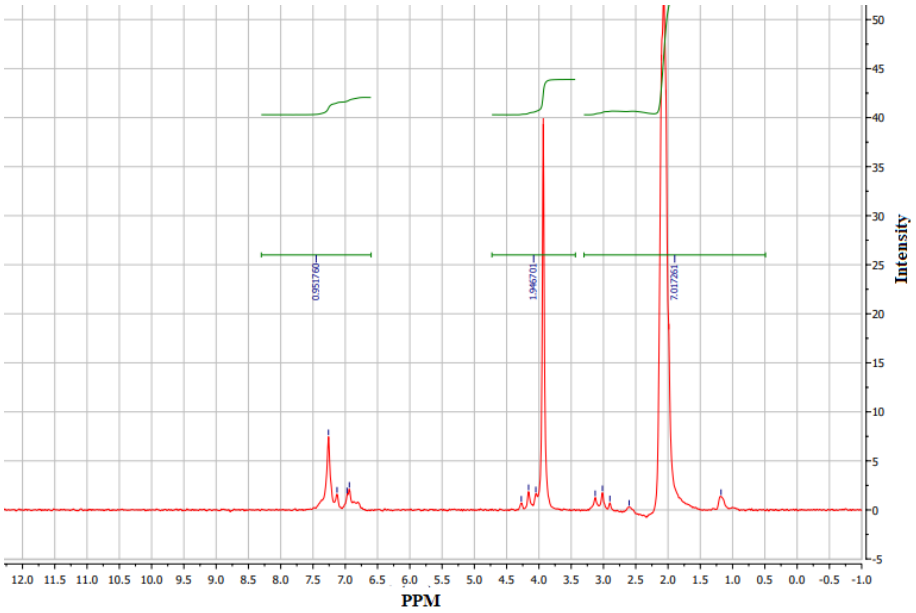


Figure 2.3: ^{1}H-NMR results of HTL of biomass.

of r components, assessed as the rank of the data matrix. The bilinear decomposition is achieved by minimizing the losses in Eqns 2.1a and 2.1b in an alternating least squares routine with non-negativity constraints on the factor matrices, to facilitate their interpretability according to the Beer Lambert's law.

$$\min_{S \geq 0} ||D - CS^T||_F^2 \tag{2.1a}$$

$$\min_{C \geq 0} ||D - CS^T||_F^2 \tag{2.1b}$$

The rank r, of the decomposition is calculated using the ratio of derivatives (ROD) of the empirical Malinowski's indicator function that is a measure of the discarded variance in the residual components after performing a PCA routine to stratify the systematic variations in the data from noise [21]. For the optimal rank r, eigenvalues arranged in decreasing order such that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_r$ explain the primary variance but the (n-r) smallest eigenvalues account for noise. In essence, the rank is calculated such that r is an integer taking values

$\in$ {1,2.. $min(n, m)$} that minimizes the residual variance given in Eq (2.2)

$$\min_r \sqrt{\frac{\sum_{k=r+1}^{n} \lambda_k}{m(n-r)}} \tag{2.2}$$

However, when dealing with noisy experimental data, the ratio of derivatives of the above empirical metric was found to be more sensitive in gleaning the optimal rank, as explained in our previous works [110, 111].

2.3.4 Bayesian hierarchical clustering (BHC)

Clustering is a non-supervised machine learning technique that combines or clusters data points that are similar into a group based on a given similarity metric [112]. The Bayesian technique offers a fundamental approach to data analysis and is fast gaining grounds in other disciplines like economics, signal processing, computational biology and genetics [113–115].

The BHC algorithm is a one-pass, bottom-up procedure that evaluates all the data points in its cluster and consequently joins pairs of clusters. The algorithm utilizes the concept of maximizing posterior probability to combine clusters [116]. The BHC algorithm enumerates the probability of combining clusters using the Bayes rule while the priors are modeled as a Dirichlet mixture model [117]. All wavenumbers are denoted as data points and are grouped into K cluster nodes of sizes T_1 ,T_2…. T_K such that D_i points belong to node T_i wherein the total data points $\sum_{k=1}^{K} D_K$ [118]. The initial point of clustering begins with as many nodes as the number of data points. The pairwise merge of nodes $T_i + T_j \rightarrow Tk$ is based on whether data points in the nodes D_i and D_j giving D_k maximize the posterior probability (γ_k) as given below: Null Hypothesis (H_0): Data in nodes is generated from same mixture component Alternate Hypothesis (H_a): Data in nodes are generated from distinct mixture components The likelihood of each of the hypotheses are evaluated as given below

$$\text{Likelihood}H_0 : P(D_k|H_0) = \sum_{\theta} P(D_k|\theta)Dir(\theta) \tag{2.3a}$$

$$\text{Likelihood}H_a : P(D_k|H_a) = P(D_i|T_i)P(D_j|T_j) \tag{2.3b}$$

$$\gamma_k = \frac{P(D_k|H_0)}{P(D_k|H0) + P(D_k|H_a)} \tag{2.3c}$$

Assuming that the data comprises random variables at the N nodes, each which has the following multinomial distribution $P(X_i = x_i|\theta) = \theta_i$; where i = 1, 2 ….N , enables the likelihood $P(D_k|\theta)$ to be expressed as a product of the probabilities of the mutually independent random variables. Here the parameters are $\theta = \theta_2, \theta_3 \ldots . \theta_N$ and $\theta_1 = 1 - \sum \theta_i$ where θ_i are the parameters that have a Dirichlet prior.

2.3.5 Bayesian structure learning

Relational information between the different reactive species in the system is inferred by learning directed acyclic graphical structures that best represent the dataset. The search space for such a problem is huge and becomes computationally infeasible for datasets with a large number of attributes. Therefore, the Bayesian formalism is used to determine the graph structure [119]. One class of algorithms aims to capture the dependencies in the data by applying a statistical hypothesis test, as seen with Bayesian clustering. In another approach, the structure learning is cast as a score-based optimization problem that aims to find a structure that maximizes the likelihood of conditional dependencies in the data [120]. Almost all such learning techniques use standard heuristic search techniques, such as greedy hill-climbing and simulated annealing, to locate high-scoring structures. The aforementioned ”generic” search procedure does not use information concerning the anticipated structure of the network to be studied. For instance, greedy hill-climbing and Tabu search techniques analyze all the possible local changes in all the steps and utilize the one that generates the largest improvement in the score. Max-min hill climbing is a hybrid approach that uses concepts from both algorithms. The Bayesian Information Criterion is

used as a scoring metric and the graph that maximizes this score is chosen. The details about the implementation of Bayesian learning for HTL data have been described in detail in our previous work [24, 109]. It is important to note here that the structure learning task identifies a non-loopy graph structure which is directed therefore rejecting the reversibility of transition (reaction) between species. Nonetheless, the approach is implemented as the graph structure learnt depicts the more dominant transition between the components, i.e., the direction of the reaction with a larger rate constant.

2.4 Results and Discussions

2.4.1 Spectroscopic analysis and data fusion

Lignin, cellulose and hemicellulose – the major constituents of biomass – contain carbon, hydrogen and oxygen. This is substantiated by the peaks observed in the FTIR spectra of the reaction mixture in Figure 2. The absorbance peaks from 3000 to 3500 cm^{-1} and from 1000 to 1750 cm^{-1} indicate the presence of C=O, C=C, C-O, C-H and O-H bonds pertaining to aldehyde, ketones, aromatics, acids, alcohols, ethers and aliphatic compounds, which is consistent with results obtained from GC-MS in other works [121].

The ^{1}H-NMR spectra of the bio-oil derived in the presence of NaOH and are shown in Figure 3. Hydrogens attached to aliphatic carbons showed peaks between 0.5-1.5 ppm. The presence of aromatic and olefinic groups was confirmed by the presence of peaks at 1.5 -3.0 ppm. The peaks in this region had the highest strength. The next portion of the spectrum showed a peak at 4.0 ppm, corresponding to protons of alcohols and carbon atoms next to aliphatic alcohols. The final peak in the spectrum was situated around 6.0-9.5 ppm representing the carbonyl hydrogens. The peaks in the 1H-NMR spectrum were consistent with the functional groups inferred from the FTIR spectra.

As mentioned earlier, a signal-level data fusion was performed with the datasets to infer the reaction network for the HTL process. At this level of data fusion, it is important to ensure similarity in the magnitude of the two signals, which was achieved by normalization

of the FTIR and the 1H-NMR signals. The signals were normalized by normalizing the unit length and highest peak intensity. The fused dataset was obtained by concatenating the two spectra.

The final signal consists of two regions. The first part consists of the IR spectrum, while the second part is the 1H-NMR spectrum, resulting in a total of 3665 variables. Principal Component Analysis (PCA) [122] was performed to remove noise in the data through reconstruction using MATLAB version 2018b and R version 3.5.1. PCA projects the dataset into a lower dimensional hyperplane obtained as a linear combination of original variables while retaining the maximum variance in the data. The first two principal components captured 95% of the total variance, and the data projected onto these components was reconstructed to the original dimensions. The final dataset used for clustering is depicted in Figure 4.

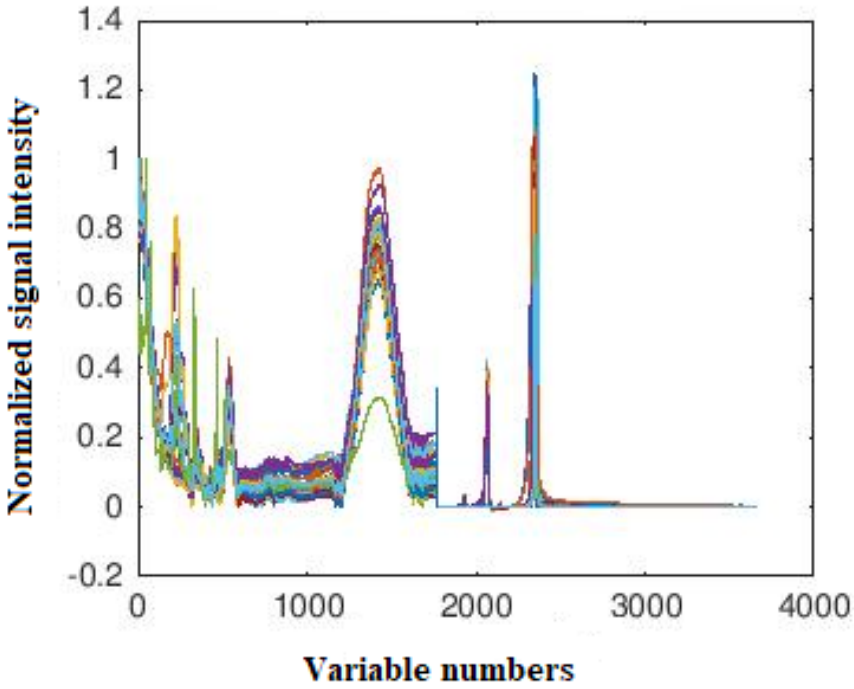


Figure 2.4: Final fused spectra for bio-oil.

2.4.2 Clustering and network generation

As mentioned earlier, the main constituents in the pyrolysis of biomass are alcohols, phenols, aromatics, carbonyls, aliphatics and gases [123]. Hence, clustering was performed with 3 to 6 clusters followed by Bayesian structure learning on the FTIR spectra of each cluster to obtain the structure of the reaction network. A minimum of 3 clusters are essential to generate a reaction network that does not lump multiple products into one node. The upper bound of 6 was chosen to curb the dimensionality of the reaction network graph. The Bayesian structure learning used Hill climbing, Tabu search and Max-Min hill climbing algorithms with the Bayesian Information Criterion as the scoring function. Structure learning revealed that the six-cluster network generated the reaction graph that was consistent across the three algorithms used, and hence, further analysis was performed based on this network. The wavenumbers placed in each cluster and the network structure are shown in Figure 5. The arc strength for each edge in the reaction graph depicts the dependency of one cluster on the other. Large negative values of arc strength represent a more favourable transition from the parent to the child node. It can be seen that the edge from cluster 3 to cluster 4 has the highest arc strength, followed by the edge from cluster 4 to cluster 5. Owing to the complex nature of biomass feedstock, numerous reaction mechanisms have been reported in the literature to describe the pyrolysis process [124, 125].

In general, the various reaction pathways can be summarised as: a) depolymerization of biomass into cellulose, lignin and hemicellulose, b) decarboxylation, decarbonylation, dehydration and decomposition of biomass monomers by cleavage, and c) recombination of reactive remains [126]. From analysis of the pseudo-component spectra and the Bayesian networks, this work proposes that the reaction mechanism for HTL of biomass will contain the reactions depicted in Figure 6. The explanation provided below highlights the types of conversions occurring between the components of the various clusters shown in Figure 4 and corroborates them with findings from the literature and from chemical reasoning. The first phase of HTL is the disintegration of the feedstock into the major constituent, viz, cel-

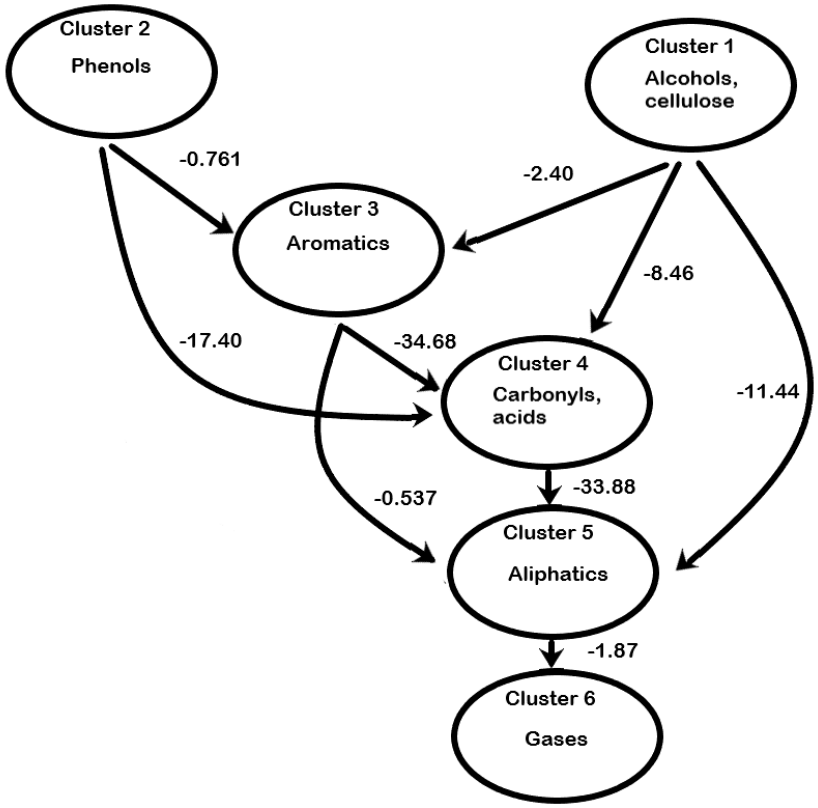


Figure 2.5: Six cluster Bayesian Network for HTL of biomass.

lulose, lignin and hemicellulose. This phase does not contribute to the actual pyrolysis but is essential for reaction modelling. Cellulose and hemicelluloses are the most abundant carbohydrates in lignocellulosic biomass. Various carbohydrates have different rates of hydrolysis. Cellulose hydrolyzes slower than hemicellulose because of the crystalline structure of cellulose. The different hydrolyzed products exist in the aqueous fraction derived after hydrothermal liquefaction of biomass. Once carbohydrates are placed under hydrothermal conditions, they undergo quick hydrolysis to form glucose and other saccharides. Alcohols are rarely recorded in HTL studies because single alcohols are found in small amounts, with vapor pressures corresponding with a meaningful fraction of the biocrude leading to coelution [127]. Nevertheless, some alcohols and saccharides were discovered in the first cluster in this study (1010, 1012, and 1019 cm^{-1}). The most abundant alcohols were the long straight chain and branched long chain alcohols produced as a result of the hydrolysis reaction.

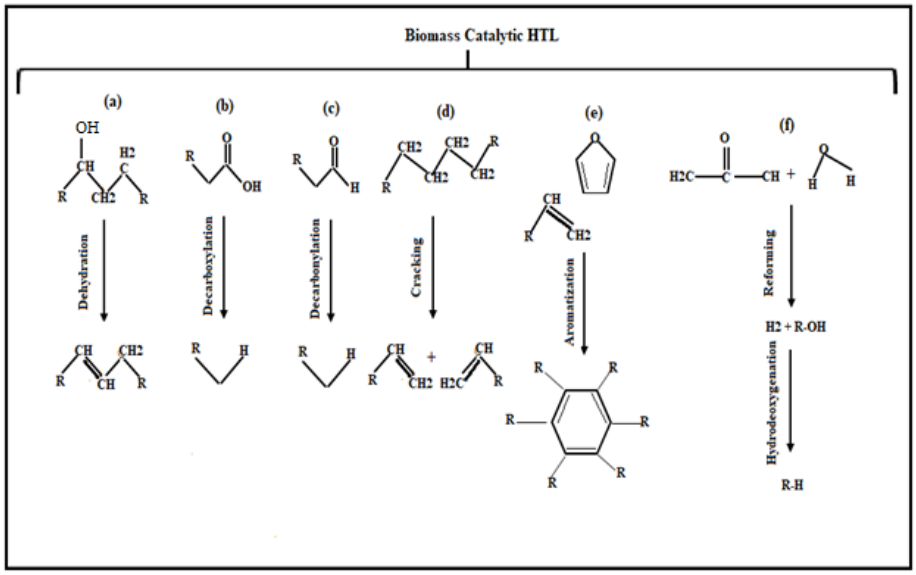


Figure 2.6: Proposed reactions for HTL of biomass.

Hydrothermal liquefaction of lignin results in the hydrolysis and cleavage of the ether and C-C bonds along with demethoxylation, alkylation and condensation along with counteraction between the reactions [128]. Cleavage of the β-O-4 ether bond results in the breakdown of lignin and its prototypical compounds, along with breaking the bond between Cα-Cβ [5]. However, the aromatic rings remain unchanged, resulting in biphenyl-type compounds indicating higher stability under hydrothermal reactions. Less severe conditions, such as low temperature and lesser reaction time, are required for the generation of phenolic monomers and dimers from lignin. These occur through the preliminary cleavage of ether bond and aliphatic C-C bond during hydrothermal liquefaction. An increase in temperature might result in demethoxylation and alkylation of lignin derived phenolic compounds. Alkyl phenols can also be obtained at high temperature [129]. Lin et al. discovered that during lignin liquefaction, the reactions involving intermediates with aliphatic side chains exhibited a huge reactivity and further combined with phenol or with themselves to change to the multi-condensed product [130]. Phenols are highly abundant in HTL product of carbohydrates and lignin-rich feedstocks and are a potential source of oxygen in the final biofuels. They are identified in cluster 2 (1099, 1100, and 1119 cm^{-1}) [131].

Wavenumbers in cluster 3 are linked to the oscillation of the benzene and the aromatic skeleton (1223, 1254, 1279, and 1500 cm^{-1}). The production of oxygenated aromatics from lignocellulose refining is common. The three monolignols of lignin are the prototypes of many aromatic compounds, together with the dehydration reactions of carbohydrates [132]. In contrast to the monofunctional ketones, the oxygenated aromatics naturally exhibit diversified functionalities resulting in complex compounds due to lignin's complex and heterogeneous arrangement. The monomers result from the thermal breakdown and hydrolysis of ether bonds [133]. At higher temperatures, the decomposition and dehydrogenation reaction of cyclic compounds from alkenes results in the production of aromatic hydrocarbons [134]. Aromatic structures were identified from the absorption bands around 1600 cm^{-1} and absorption between 3000 and 3050 cm^{-1} in FTIR spectra.

Under subcritical conditions, alkaline water and carbohydrates are known to form carboxylic acids like acetic, propionic, formic, and lactic acid via retro-aldol reactions. They can also be subjected to homogeneous and heterogeneous ketonic decarboxylation, generating a series of various ketones [133]. Figure 2 depicts carbonyl absorption at 1715 cm^{-1} and 1745 cm^{-1}, showing six-membered and five-membered cyclic ketones, respectively. Furthermore, compounds categorized in the first cluster can then further degrade to produce several oxygenated hydrocarbons like formic acid, lactic acid, hydroxymethyl furfural (HMF), and levullinic acid [135]. Carbonyls were found to be highly abundant in most HTL biocrudes, with the most abundant carbonyls being indenones, acetophenones, and a wide range of alkylated chromenones [136]. A recent study states that FTIR spectra of bio-oil in the region of 1490–1850 cm^{-1} could provide comprehensive information on several carbonyl groups in the bio-oil [137]. This work recorded the presence of ketones, aldehydes, and carboxylic acids in cluster 4 (1695, 1710, 1723, 1745, and 1749 cm^{-1}).

Short chained aliphatic hydrocarbons were identified in cluster 5 (1332, 1420, 1573, and 1665 cm^{-1}), and indicate the occurrence of C-C bond cleavage reactions. Glycerol conversion under near- and subcritical water conditions has been outlined to undergo C-C splitting

via an ionic and a radical pathway [133]. These compounds result in decarboxylation and decomposition reactions.

From a thermodynamic point of view, the thermochemical conversion of biomass, glucose and other organic components will result in light constituents. CH4 and CO2 are thermodynamically preferred products, with the CO and H2 yields remaining low [138]. The existence of these molecules can be traced to the last cluster (cluster 6) in the generated BN.

Based on the reactions described above and in Figure 6, along with the reaction network of Figure 5, a proposed mechanism of the HTL of biomass is illustrated in Figure 7 As an

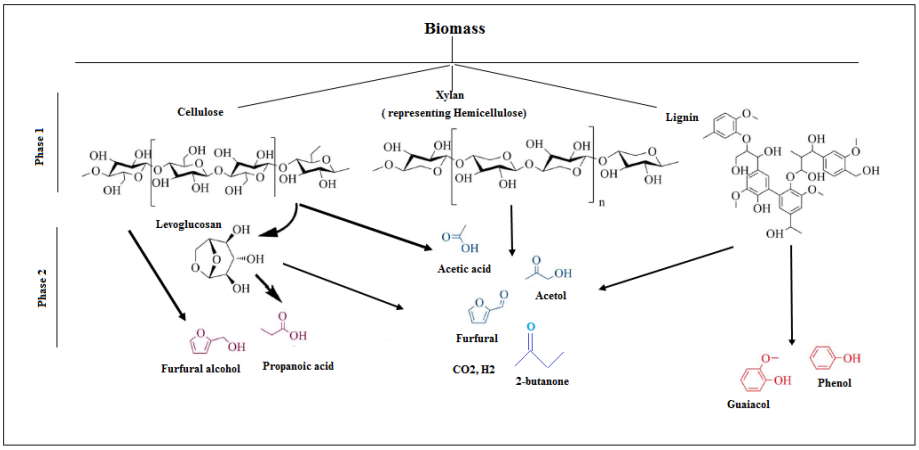


Figure 2.7: Two-phase reaction mechanism for HTL of biomass.

alternative approach, additional analysis was performed on the dataset using self modelling multivariate curve resolution (SMCR) [18]. The choice of the number of components, 3 in this case, was determined using the Ratio of Derivatives (ROD) function [21], as mentioned earlier.

The factorized spectral signatures for each component are shown in Figure 8. The concentration profile of each component obtained from SMCR was used in the determination of a Bayesian network using algorithms described earlier. The resultant network structure is depicted in Figure 9.

Component A1 contains the signatures of primary and secondary alcohols (1075-1010

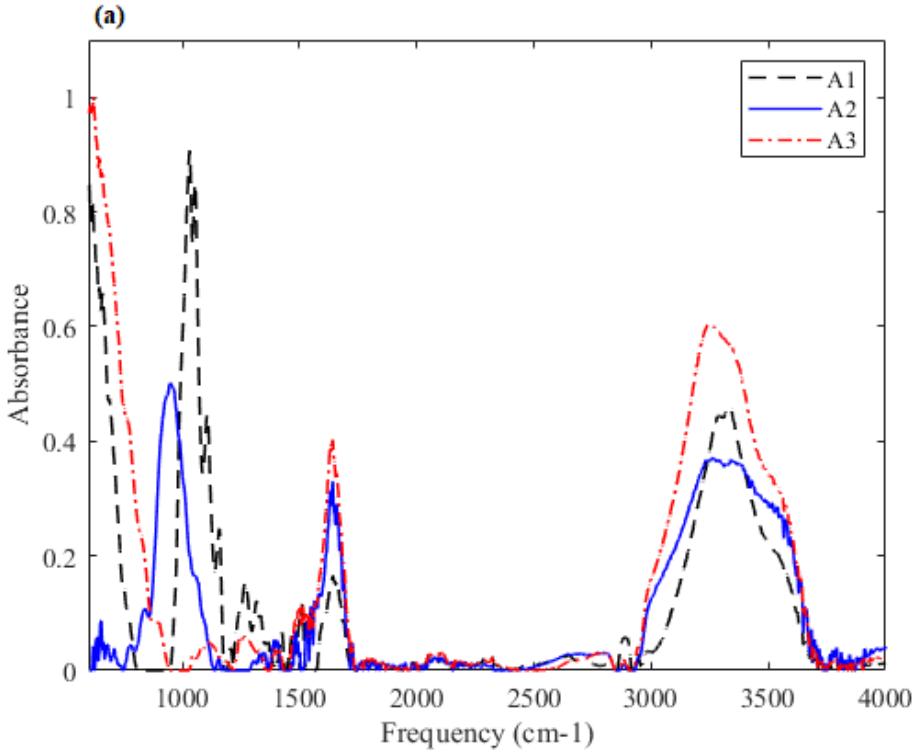


Figure 2.8: Resolved spectra for the pseudo components over the whole region (a) and the resolved spectra for the pseudo components focusing on the major peaks (b–d).

cm^{-1}) and ethers (1150-1070 cm^{-1}). The spectrum for A2 illustrates the presence of phenolic groups as indicated by broad absorbances between 3550-3200 cm^{-1}. A3 indicates the presence of carboxylic acids (3550-3500 cm^{-1}), ketones (3550-3205 cm^{-1}), and aryl aldehydes (1715-1695 cm^{-1}), as well as aromatic compounds due to the presence of C=C vibrations with absorbance peaks from 1625-1575 cm^{-1} or C-C in-ring stretching at 1500-1400 cm^{-1}.

2.4.3 Discussion

Since all lignocellulosic biomass is largely composed of three basic independent structural components (cellulose, hemicellulose, and lignin), any aggregative behavior of these components during pyrolysis describes the behavior of any lignocellulosic feed [139]. Furthermore, since biomass pyrolysis follows a complex network of reaction mechanisms, its chemistry can be simplified by studying the independent pyrolysis reactions of each indi-

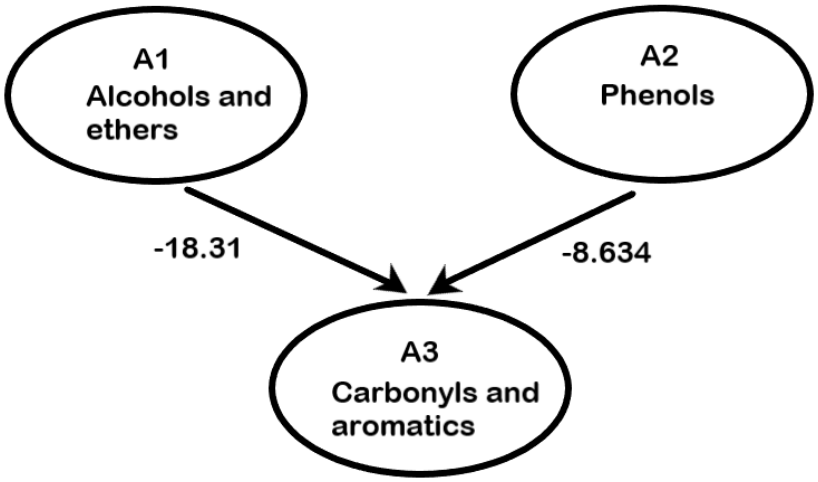


Figure 2.9: BN obtained through SMCR-ALS for the pseudo-components

vidual component. If synergistic effects occur, predicting a biomass feedstock's behavior would be considerably more complex. In our previous work [24, 109], cellulose and lignin independently underwent HTL conversion using their model components: levoglucosan and 2-Phenoxyethyl benzene, respectively, and their most probable BNs were developed. Figure 10 shows those networks along with the BN for the HTL of biomass developed in this work.

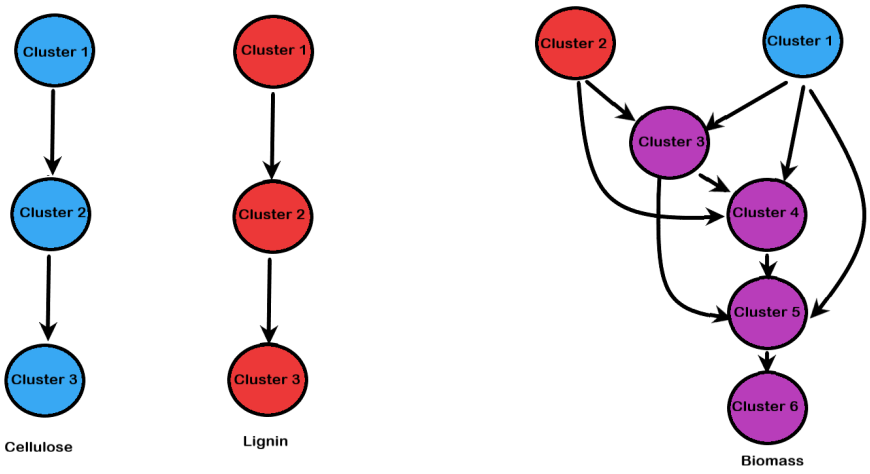


Figure 2.10: BNs for (a) cellulose, (b) lignin, and (c) biomass (data provided by data fusion).

After reviewing and comparing wave numbers in each cluster in each BN, it was inferred that the right side of biomass conversion from cluster 1 mostly represents hydrocarbon (cellulose or hemicellulose) conversion while the left side represents lignin conversion, though both have the same final products: aromatics, carbonyl groups, aliphatic hydrocarbons, and smaller molecules. In previous work, one of the major products identified from the HTL of cellulose was formaldehyde, identified by FTIR, 1HNMR, and GC-MS. Interestingly, the presence of formaldehyde was confirmed in the HTL of biomass by bands at 3308, 2982 and 2914 cm^{-1} for –CH stretch, a clear peak at 1636 cm^{-1} for –C=O for aldehyde, and bands at 1429, 1271, 1103, 1019, and 989 cm^{-1}. Following glycosidic bond cleavage, hydrogen from the hydroxyl group of the carbon atom 6C is transferred to 5C. This is conveyed by cleavage of the 5C–6C bond, resulting in formaldehyde formation. These conclusions are consistent with the 2-phase reaction pathway proposed earlier. The SMCR framework reveals a similar reaction mechanism as represented by the network structure derived after decomposition as depicted in Figure 11. The decomposition of cellulosic structures from previous works [24, 109] is well described by the Bayesian network for biomass. The signatures of alcohols, ethers and alkenes collected in SMCR analysis of levoglucosan are found in A1 of the biomass network. Similarly, phenolic and aromatic components of lignin breakdown can be traced to A1 and A2 of the biomass network. Overall, similarities between the final products were found in the last nodes of cellulose, lignin and biomass decomposition networks. The SMCR analysis was performed as an independent study to validate the reaction networks obtained through the hierarchical clustering approach as it does not use any information from the BHC approach. The SMCR – BN analysis is consistent with the BHC-BN approach, but provides little additional insight on the pyrolytic process as both methods yielded the results with similar levels of inference.

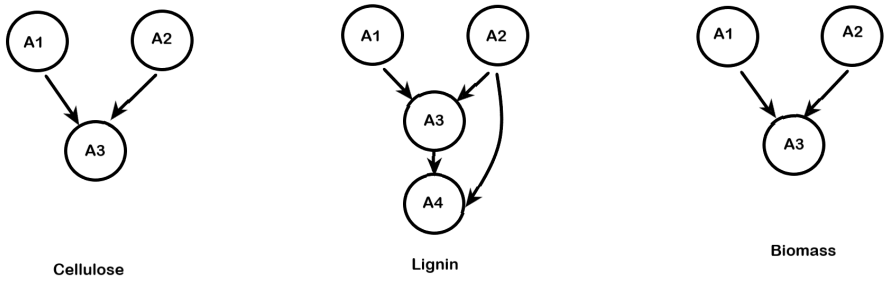


Figure 2.11: BNs from the SMCR-ALS method for (a) cellulose, (b) lignin, and (c) biomass conversion.

2.5 Conclusions

In this work, we present the validation of an approach for identifying reaction mechanisms for the conversion of complex biomass feedstocks. The approach uses spectroscopic data from multiple sources, data fusion, Bayesian clustering and Bayesian networks to identify reaction networks, and was applied to the hydrothermal liquefaction (HTL) of biomass in the temperature range from 150 -350°C. The pathway to biocrude identified from the reaction network represents decomposition of the biomass components, large molecule hydrolysis, and reformation of produced molecules. The reaction network hypothesized compares well with mechanisms reported in literature [89]. Importantly, comparison with analyses for the HTL of cellulose (represented by levoglucosan) and lignin (represented by 2 Phenoxyethyl benzene) revealed that the network hypothesized for biomass breakdown was a combination of the networks for individual decompositions of its components as A1 and A2 of the biomass network corresponded to elements in the cellulose and lignin networks. Thus, this work presents a data driven approach to infer reaction networks for complex reaction mixtures and relate them to the reaction networks for individual constituents of the feed, and can potentially be used to develop reaction hypotheses, process designs and process monitoring techniques [140] for biomass feedstocks of varying composition. The dominant reaction network that is well represented through the spectroscopic profiles is captured through the process. Intermediates with small lifetimes that do not show up in the measurements do not

surface in the reaction network. The kinetics of the decomposition process is not considered explicitly in this work. However, in the context of online reaction monitoring, sophisticated spectroscopic curve resolution algorithms [26] can be used to project realtime spectra onto the temporal data collection mode, and the spectroscopic channels which in accordance with Beer's law, gain interpretability as the pseudo-component concentrations and pseudo-component spectra, respectively. The kinetics of the underlying chemical transformations can then be assessed from the temporal concentration projections to further facilitate control and optimization [141]. The automated mapping of the hypothesized reaction networks to domain knowledge-based real chemistries [142], can aid the future development of an interpretable end-to-end pipeline to identify species, reactions and kinetics from spectroscopic data. Prior knowledge about the process in tandem with the reactions and species discovered through these methods can be applied towards experimental design in pinning down the specifics of a reaction. Information obtained through our methods on the conditions at which a particular transformation between species takes place can be looped back to experimental design to recover more data with sufficient excitation of variables in the region of interest to uncover more specific details of the chemistry.

Chapter 3

Identification of reaction network hypotheses for complex feedstocks from spectroscopic measurements with minimal human intervention

3.1 Abstract

In this work we detail an automated reaction network hypothesis generation protocol for processes involving complex feedstocks where information about the species and reactions involved is unknown. Our methodology is process agnostic and can be utilized in any reactive process with spectroscopic measurements that provide information on the evolution of the components in the mixture. We decompose the mixture spectra to obtain spectroscopic signatures of the individual components and use a 1-d convolutional neural network to automatically identify functional groups indicated by them. We employ atom-atom mapping to automatically recover reaction rules that are applied on candidate molecules identified from chemistry databases through fingerprint similarity. The method is tested on synthetic data and on spectroscopic measurements of lab-scale batch Hydrothermal Liquefaction (HTL) of biomass to determine the accuracy of prediction across datasets of varying complexities . Our methodology is able to identify reaction network hypotheses containing reaction networks close to the ground truth in the case of synthetic data and we are also able to recover candidate molecules and reaction networks close to the ones reported in previous literature

studies for biomass pyrolysis.

3.2 Introduction

With growing environmental concerns and lack of access to traditional precursors for chemical manufacturing, there has been a marked increase in attention towards alternative feedstocks for energy, fuel and chemical production [143–146]. The chemical complexity of both traditional and alternative feedstocks prove to be a challenge for reliable model development, monitoring and optimization of reactions underlying their thermal conversions. Their heterogeneous nature in both physical and chemical properties, does not allow for a straightforward analysis of the reactions these materials undergo in their upgrading or transformation process.

From the perspective of optimization and reliable control of the process, definition of a model that describes the various states of the system under different operating conditions is critical. However, development of explanatory models with just the knowledge available in literature, especially for complex systems, without concrete information on the constituents is impractical and can result in suboptimal models. Traditionally, models (reaction networks) for these complex systems are developed based on model compounds that describe the reactions occurring in the system [5, 6, 147]. The choice of model compounds used depends on the technical knowledge of the human expert describing the system and thus allows for a greater degree of variability. The integration of process measurements in the modeling process brings in more specific information indicative of the transformation of different chemical species in the system.

Spectroscopic measurements are popularly used to obtain molecular-level information in process industries owing to their relative swiftness, reliability and non-invasive nature [9]. Infrared spectroscopy, Nuclear Magnetic Resonance spectroscopy and Raman spectroscopy have found applications in various fields such as refining, pharmaceuticals and drug discovery, [7, 10, 148] though they come at the cost of being high dimensional, not of

full rank (i.e., the variables are dependent on each other) and noisy[104, 149]. While each type of spectroscopy provides a different perspective of the process, a combined analysis of distinctly different types of spectroscopic measurements can provide a holistic view of the system especially for chemically complex mixtures such as those encountered in biomass conversion.

Our previous works have demonstrated the use of Fourier Transform Infrared (FTIR) spectroscopy measurements and Proton Nuclear Magnetic Resonance (^{1}H- NMR) spectroscopy measurements in developing reaction models for these complex feedstocks [24–26, 109, 111, 150]. In one approach, information from each type of spectroscopic measurement was jointly decomposed in the temporal context of spectroscopic data collection, using a *Structure preserving Joint Non-Negative Tensor Factorization* scheme to extract signals of each component. This, in tandem with the usage of the Bayesian formalism to develop a graphical model, provided the directionality of reactions upon which chemistry inferred by human expertise was super-imposed to derive the reaction network for the system. In this work we present a reaction network hypothesis generation methodology as a means of identifying a set of reaction networks that correspond to the spectroscopic sensor measurements with minimal human intervention. This is a challenging task, and is accomplished by using an algorithmic approach to translate spectroscopic signatures of pseudo-components to representative (real) chemical species.

Automating reaction prediction is performed more commonly in the retrosynthetic problem where the precursors required to generate a specific compound are identified [151–155]. Broadly speaking, the automated reaction prediction task can be classified into i) Template/rule based approaches ii) Quantum mechanical calculations and iii) Machine learning approaches [156–158].

Rule-based network generation approaches use semantics that describe the transformation of the substrate into products encoded in a machine readable format. These semantics or reaction rules and are either manually encoded or algorithmically extracted as Simpli-

fied molecular-input line-entry system (SMILES) arbitrary target specification (SMARTS) strings or edits in the atom connectivity matrix [28, 29, 159, 160]. Manually encoded reaction rules are typically generated for a limited set of reactions pertinent to the system under consideration. The complexity of adding new rules to an existing databank increases based on the specificity of the rules. Algorithmically extracted rules typically employ the use of atom-atom mapping (which will be discussed in detail in Section 3.4.2). Quantum mechanical computation of the reaction network involves the exploration of the potential energy surface (PES) of the reaction mixture. The core idea in such exploration techniques involves a minima-hopping algorithm, where minima in the neighborhood of the initial equilibriated structure of the mixture are identified through different techniques, and a path connecting them that passes through a saddle point (transition state) is determined [32, 161–163]. Other *ab initio* techniques involve solving the equation of motion for the system and tracking the trajectory of the reactant molecule through the course of the simulation [29, 34, 164]. Due to the intractable nature of conducting quantum mechanical computations for reactions with a large number of species, they are often combined with rule-based techniques to identify reaction networks, [30, 165–167]. Machine learning (ML) algorithms aim to extract latent features of reaction types that are then used to predict reaction outcomes. A class of algorithms deal with ranking reactions or reaction rules under a specific context to identify the most probable product [37, 38, 168]. Due to the lack of a motif-based framework as in rule-based approaches, ML models present a more generalized approach towards solving the reaction network generation problem. Development of graph convolutional neural networks to encode information over graph structures have spawned a class of ML methods to retrosynthetically predict reaction networks [39, 169]. *seq2seq*, protocols typically used in natural language processing, have also been employed for network prediction [35, 153]. A more thorough review of the various techniques in automating reaction networks can be found in the literature [27, 141, 157, 170].

All the methods discussed above pertain to systems where the actual precursors/molecules

are known, which is hardly the case in case of complex reaction mixtures. Characterization of these mixtures in terms of relative compositions of various components does not provide a specific molecule from which to interpret the reaction network. In this work, we aim to develop the reaction network for these mixture systems with the partial information available from spectroscopic measurements. This fragmented information is used in generating plausible candidate molecules and reaction network hypotheses that correlate with the observed transformation of different components. We employ convolutional neural networks for automated functional group detection followed by molecular fingerprint matching to obtain candidate molecules. Reaction templates extracted through atom-atom mapping are applied on these candidate molecules to generate the reaction network hypotheses.

The goal of this work is to explore the possibility of automating the reaction network generation task for complex systems. It is vital to note that the network generation scheme presented in this work is built to identify descriptive and lumped reaction pathways rather than detailed elementary reaction mechanisms. This work is intended to be used as a hypothesis generation scheme where multiple hypotheses for the reaction network based on the spectroscopic sensor measurements are generated and is to be used a screening tool aimed at identifying the chemical space encompassing the true reaction network of the system. Further ranking or validation of these hypotheses through experimental or quantum mechanical based simulations are necessary to converge on a single best network.

The rest of the chapter is arranged as follows. In Section 3.3 we provide a brief description of the datasets used in this study. Section 3.4 describes in detail the methodolgies used in developing the reaction network followed by the results of this study and discussion in Section 3.5. Practical and theoretical limitations of this study have been discussed in Section 3.7

3.3 Datasets

We demonstrate the reaction network generation methodology on three datasets of varying availability of prior knowledge and complexity in this work. The first is synthetic data generated by mixing the FTIR spectra of four components (obtained from National Institute of Standards and Technology) based on a known reaction network architecture with a kinetic scheme imposed on it (B.1). The network generation methodology is agnostic to the dataset and generation pattern behind it; hence, using the synthetic data provides a means of validating the methodology.

The second dataset consists of FTIR and ^{1}H-NMR measurements for the hydothermal liquefaction of phenoxyethylbenzene (PEB), a model compound for lignin conversion, in a batch reactor.The data is obtained for batch time conditions of 15, 25 and 30 minutes.The usage of the PEB dataset provides a known starting point for the network generation algorithm, thereby providing another framework for validation. Additional details on the experimental procedure of the hydrothermal liquefaction can be found in our previous work. [24, 109]

The final dataset consists of FTIR and ^{1}H-NMR measurements for the hydothermal liquefaction of Monterey-pine biomass in a batch reactor. The three datasets encompass different degrees of prior knowledge about the system. The synthetic and biomass-HTL datasets assume minimal prior knowledge about the reaction system. The synthetic data is relatively less noisy and the ground truth is available for validation. The PEB dataset includes prior knowledge of the feedstock species and hence is used as a validation of the reaction template application scheme.

3.4 Methods

Figure 3.1 provides an overview of the different methods employed in this work. The spectroscopic sensor data contains multiple overlapping peaks corresponding to different

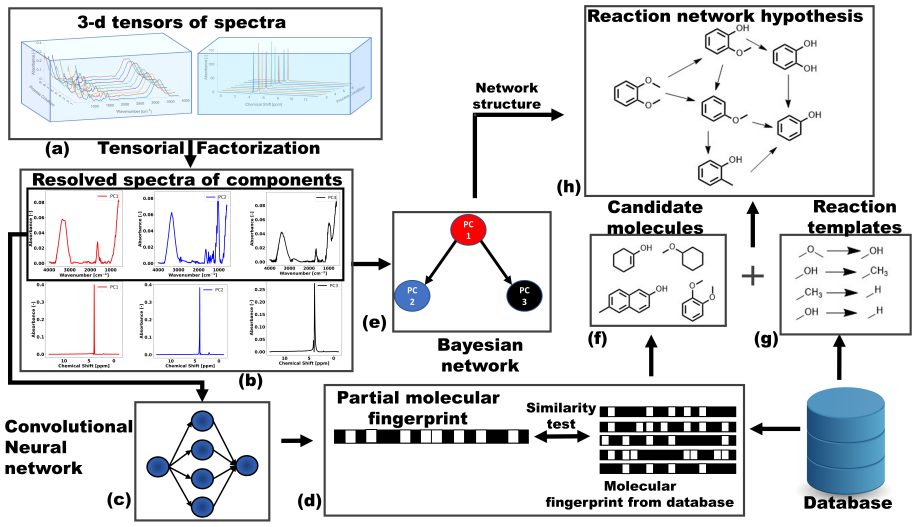


Figure 3.1: A representative overview of our methods

pseudo-components (PCs). These peaks are deconvoluted using a tensorial decomposition to obtain spectroscopic signatures of individual pseudo-components. A 1-d convolutional neural network identifies the different functional groups in the PC spectra and a partial molecular fingerprint is generated. Fingerprint similarity tests are conducted by matching the partial fingerprint to the fingerprints of molecules in chemistry databases. Candidate molecules are generated based on degree of similarity with some scaffolds of molecules being rejected based on prior knowledge of the process. Reaction templates are automatically extracted from a large number of reactions reported in literature through atom-atom mapping. These extracted templates are recursively applied to the candidate molecules to obtain reaction network hypotheses.

The tensorial decomposition and the formulation of the Bayesian network have been discussed in detail in our previous work[26, 111, 150]. In brief, the joint non-negative tensorial factorization (JNTF) algorithm jointly deconvolves two three-dimensional tensors consisting of FTIR and ^{1}H-NMR at different process conditions (temperature and residence time), respectively. The resultant decomposition produces individual spectroscopic signatures (FTIR and ^{1}H-NMR) of the k unique pseudo-components (PCs) along with the concentration profiles of each component along the residence time and temperature modes.

The choice of the number of components in the decomposition was generated through core consistency diagnostics [22]. The joint decomposition allows for constraints to be placed on the concentration profiles in such a way that they are shared amongst the two tensors. The variability in information between the two decompositions is restricted to occur only in the spectral modes while also maintaining a non-negativity constraint on these modes.

A Bayesian network formalism was used to identify the structure for the reaction scheme and provide a skeletal structure onto which the chemistry of the process would be superimposed. The network is a directed acyclic graph (DAG) indicating the root node and its subsequent children. The structure learning task was performed using 'greedy' search algorithms such as Hill-climbing and Tabu search and the hybrid Max-Min Hill climbing algorithm with the Bayesian Information Criterion as the scoring metric. The graph generated has nodes representing the PCs with arc strengths indicating the strength of connections between them.

The tensorial factorization [26] and Bayesian network generation [111] aspects of the pipeline have been studied in our previous works. This work aims to introduce the functional group classification and reaction network generation routines and discuss the results of the same.

3.4.1 Functional group identification

In the context of the proposed framework, evaluation of the deconvolved spectra to identify the different functional groups present in each component is essential for identifying the chemistry of the process. Typically, analysis of FTIR or ^{1}H-NMR spectra is done through use of expert knowledge along with lookup tables indicating the peak positions for different functional groups. In-built programming in FTIR measurement devices also allow for automation in detection through comparison with a database of spectra for known molecules. A similarity metric is used in determining the closeness of the spectrum of the sample to the spectra in the database [171, 172]. Library search methods require extensive effort in

compilation and pre-processing of candidate spectra and cannot extrapolate to spectra of new components.

Another approach involves the use of machine learning models such as support vector machines [173–175], k-nearest neighbors [176, 177] and Principal Component Analysis (PCA) [177] to interpret the spectra. While these models perform well in identification, they present a narrow scope with models being trained on a particular set of examples (eg., plastics) to solve a specific task. Applying these models in our case severely restricts the search space, i.e, models trained on a specific set of molecules and their functional groups will only be able to identify the spectroscopic signatures of functional groups in that context, thus leading to ineffective network generation.

It is thus imperative to generate a classification scheme where the spectroscopic signatures of a wide range of molecules are learnt in the context of the characteristic peaks of functional groups present. Artificial neural networks (ANNs) are a class of ML models that learn complex nonlinear information from the data by representing them as nested functions of simpler functions known as activations.

In this work, we present a classification scheme using a Convolutional Neural Network (CNN) [178]. Originally developed for image classification, the CNN extracts information (called features) from a image by moving a filter of a pre-determined size across the data and performing a convolution operation between the filter and the image. The convolution operation identifies regions that match the filter. Through the course of training, the CNN learns these features that best help the model in distinguishing between classes. The dimensionality of the input data determines the number of directions in which the filters move and consequently the dimension of the CNN. Previous studies have incorporated the use of 2-D CNNs and feed-forward neural networks in individual classification of functional groups, which allows for the extraction of the peak height characteristic along with the wavenumber of the peak as features [179–181]. In this work, we employ a 1-D CNN that moves only across the wavenumber axis to classify the functional groups. The advantage of using such

a methodology is two-fold: i) it allows for a lower complexity in training the model, and ii) it disregards the peak height information. The need for the classifier to be indifferent to the peak height is due to the intensity ambiguity possible in the tensorial decomposition [26].

In the implementation, gas phase FTIR spectra of 11062 molecules were automatically scraped from the NIST database. The International Chemical Identifier (InChI) for each of molecule was used in generating the labels for classification. The label vector is a bit vector of size 13 with each bit representing the presence or absence of a particular functional group. This leads to a multi-label classification task that forces the CNN to jointly learn features for the different functional groups in the presence of other groups. The number of labels to be classified into was chosen based on the availability of the functional groups in training dataset as well as domain expertise on plausible functional groups present in the feedstock.

The spectra were filtered to include only absorbance values with wavenumbers in cm^{-1} within a range of 400 - 4000 cm^{-1}. The resolution of the extracted spectra were down-sampled to 4 cm^{-1} to match with the minimum resolution of the scraped spectra. This was performed by grouping wavenumbers into sections of 4 cm^{-1} with the absorbance taking the average values across the grouped wavenumbers. Finally, spectra were rescaled between 0 and 1. Other pre-processing steps included baseline correction and the use of a Savitzky-Golay filter of window length 19 and polynomial order 2 to remove noise and smoothen the peaks. No derivative information was used in the filter. The training data consisted of 7743 such spectrum (input)-functional group (output) pairs and the testing dataset contained 3319 such samples.

The convolution layers consisted of 10 filters of size 10 along with rectified linear activation. In order to avoid overfitting, the random node dropout with a dropout fraction of 0.2 was used at the end of the convolutional layers. Fully connected layers were used on the representations learnt by the CNN to classify the spectrum. The binary cross entropy (BCE) loss was applied on each label to train the model. The output of the CNN includes a

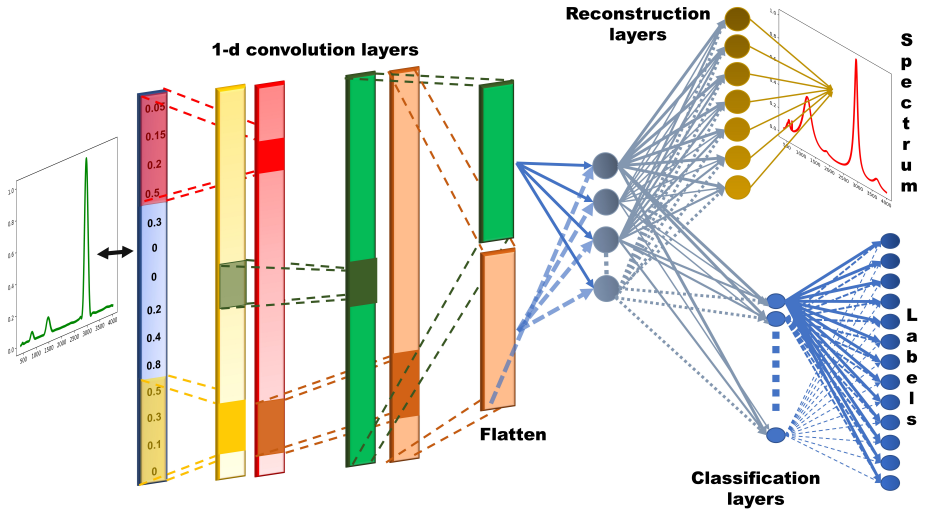


Figure 3.2: Architecture of the neural network used in classification of functional groups

reconstruction branch in addition to the classification architecture. The features extracted by the convolutions were used in reconstructing the input spectrum along with the classification of the functional groups. The spectrum was reconstructed as the sum of Cauchy distributions. The choice of reconstruction using Cauchy distributions was motivated by other works involving curve fitting of FTIR spectra, where parameters of the distribution used to fit a spectrum were used in analysis of the spectrum.[182–184]. As is evident from Figure 3.2, classification and reconstruction are performed based on the shared feature set from the convolutional layers. The improvement in reconstruction allows for more explanatory and discriminatory features to be extracted by the CNN, thus enhancing classification. Additionally, the weighting of the BCE loss was included based on the proportion of positive and negative samples available in the training dataset, which resulted in a weighted BCE (WBCE) loss given as

$$WBCE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} w_j \cdot [y_{i,j} \cdot log(p(y_{i,j})) + (1 - y_{i,j} \cdot log(1 - p(y_{i,j})))] \quad (3.1)$$

where N represents the total number of samples, C is the total number of classes (13 in this case), w_j is the weight associated with each class, $y_{i,j}$ is the binary label for the j^{th} class in the i^{th} sample and $p(y_{i,j})$ is the neural network prediction of the binary label for j^{th} class in the i^{th} sample

The Cauchy or Lorentzian distribution can be given as

$$f(x; x_0, \gamma) = \frac{1}{\pi\gamma \left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]} \tag{3.2}$$

where x_0 represents the position of the peak of the distribution and γ is the half-width at half maximum. The modified architecture thus uses a custom built layer that reconstructs the input spectrum as the weighted sum of 15 Lorentzians. If the output of the previous layer of the neural network was $y \in \mathbb{R}^{m \times 1}$ and weights and bias of the Lorentzian layer are $W \in \mathbb{R}^{15 \times m}$ and $b \in \mathbb{R}^{15 \times 1}$, respectively, the Lorentzian layer computes the distribution as

$$f(x; W, y, b) = \sum_{i=1}^{15} \beta_i \cdot \frac{1}{\pi b \left[1 + \left(\frac{x - W_i * y}{b_i}\right)^2\right]} \tag{3.3}$$

where W_i and b_i are the i^{th} row and the i^{th} element of the W and b matrices respectively. β_i's are the weights of the summation of the 15 Lorentzians and are learnt as parameters of the neural network. The loss in reconstruction was computed as the Kullback-Leibler divergence between the softmax of the input and the softmax of reconstructed spectra.

3.4.2 Reaction network generation

As mentioned earlier, automating reaction network generation consists of determining candidate molecules for each PC and the network propagation through application of reaction rules. Identification of the functional groups in the spectra provides information on the possible reaction centres in the PCs.

Determination of candidate molecules

The selection of candidate molecules for the network was done by comparing substructures identified by the CNN with molecules present in chemistry databases. The database of molecules was obtained from US patent literature from 1976-2013 [185][37]. The initial pre-processing step to ensure that all the atoms had the correct valency was performed through sanitization checks in *RDKit* [186]. The molecules were represented using their SMILES strings, which indicate the connectivity of the atoms.

The search for the most structurally similar candidate molecule was done by comparison of molecular fingerprints [187–191]. Molecular fingerprints are a many-to-one mapping of the connectivity and properties of atoms in a molecule, usually represented in a vectorized form, with each position in the vector being indicative of a specific property. Though initially used in identification of isomeric structures, molecular fingerprints proved to be effective in comparison of molecules and sub-structure detection. The notion of similarity requires a distance metric that compares the 'closeness' of two molecules. Identifiers of a molecule such as its name, formula and Chemical abstracts service (CAS) ID, for example, do not allow for direct computation of the distance metric. Fingerprints, on the other hand, include bit-vector representation, thus allowing for mathematical computations of distance.

Structural fingerprints are bit-vectors with each bit corresponding to a particular substructure. Turning 'on' a bit corresponds to the presence of that particular substructure within the molecule. The size of the vector allows for extensiveness in capturing all the sub-structures in the molecule, though it does not allow for a one-to-one mapping between molecules and fingerprints.

Another class of fingerprints takes the connectivity information of atoms into account. The structural fingerprints lump groups of molecules into substructures while connectivity fingerprints consider each atom in the molecule individually and represent the neighborhood information of each atom in its bits. Morgan fingerprints iteratively update the identifiers of each atom in a molecule based on the neighbors at different bond radii and hash them into a fingerprint [192]. The information gleaned by explicit consideration of the neighbors is important, especially when matching candidate molecules to then be used in reactions. The reactivity of a molecule (or the reaction centres within a molecule) differs based on its substituents. This modification in the reactivity is to be accounted for in the reaction network generation to ensure its extrapolation for monitoring and control purposes.

In this work, the Molecular ACCess System (MACCS) keys and Morgan fingerprints were used in determining the most structurally similar candidate molecule to be applied

in the reaction network formulation. For each PC, the functional groups identified in the previous step are converted into SMARTS strings that indicate information about fragments of the molecule [181]. Determining the candidate molecule reduces to identification of the scaffold structure and possible placement of the functional groups that best match the partial information available from the spectra. The SMILES string is then converted into both MACCS and Morgan fingerprints, which are then compared with the fingerprints of the molecules from the database using the Tanimoto similarity coefficient. The molecules with high degree of similarity were further filtered based on domain knowledge on the possible scaffolds present in the system. The Tanimoto coefficient between fingerprint vectors A and B is

$$\text{Tanimoto similarity} = \frac{A \cap B}{A \cup B} \tag{3.4}$$

The coefficient is the ratio of number of common bits between the two fingerprints to the total number of bits in both fingerprints. The molecules that show a high degree of similarity to the identified partial fingerprint are used in building the network.

The task of similarity detection between fingerprints of PCs and actual molecules is done at multiple stages. The notion of similarity is invoked in identifying the starting point of network generation (candidate molecule detection). The network generation then continues and the pruning of the network is done again through a similarity match between the children nodes in the network and the partial fingerprint of the corresponding PC in the Bayesian network. This allows for intelligent pruning of the reaction network graph to only include pathways that are indicated by the spectra of the process. In a scenario with multiple candidate molecules being chosen for a particular species, the network generation is performed with all candidates.

Reaction templating and network generation

The reaction network generation is performed by consecutive application of reaction templates to molecules in each generation of a graph. Reaction templates or reaction rules encode the transformation of substrate(s) into product(s). This usage of reaction templates is widespread in retrosynthetic analysis where the requisite substrates for a target molecule are identified through backwards propagation of the reaction template. The transmutative information encoded in a template is restricted to the reacting atoms in a molecule. While obvious for a human expert, computer-aided reaction network generation requires a mathematical formalism that helps in identification of these reaction centres. Previous works have used template-free approaches where ANNs were used in classification of molecules as electron donors or acceptors, thereby generating the product as combination of the substrates at the electron transfer locations [193]. However, the more prevalent formalism involves the description of molecules as molecular graphs [159]. In a molecular graph, the nodes represent the atoms and the edges define the bonds between the atoms. Information pertinent to the different atoms in the molecule, such as its type, charge, valency, etc. are contained as features of the particular node. The information about the type of bond (single, double, triple, ionic) are represented as weights of the graph's edges. More extensive information such as bond angle, bond length, shape of molecule, hybridization, etc, are not incorporated in the rudimentary molecular graph, though such attributes can be encoded in more complex models.

This representation allows for a straightforward definition of a reaction. In the context of the molecular graph, a reaction is a change in the edge properties (eg, double bond to single bond in saturation), or the adjacency matrix (breaking or formation of bonds), or a change in the node attributes (change in charge, etc). This definition allows for a straightforward identification of the reacting atoms. The graph-based formalism does not incorporate ordinality; hence, finding the difference in connectivity between the reactant molecule and product molecule involves computing the difference between the isomorphs of their re-

spective molecular graphs. Therefore, atom-atom mapping is used to place 'name tags' for each atom in both the reactant and product molecules [30, 151, 153, 194]. The atom mapping number of a particular atom is the same in the both reactant and products. This mapping constraints the graph to a particular sequence, thus allowing for the identification of the reaction centre in a single iteration. For each mapped atom in the reactant, the atom (node) with the same mapping number in the product is identified. If the attributes of that node or the edges of the node change, then the atom is identified as a reaction centre.

Automation of the template generation protocol begins with a curated SMILES strings of reactions in the US patent database [185]. In earlier work by Lowe [185], the US patent office data was text-mined and data regarding reactions, solvents, yields, etc were curated into a database. Atom-atom mapping of the reactions were performed in the previous work using the Indigo Toolkit [195].This curated dataset was used in this work. The database contained reactions commonly hypothesized to occur in HTL biomass such as decarboxylation, dehydration, cracking, etc. While the reaction strings used in this study were atom mapped already, automated atom mapping techniques exist that can be used to place tags on atoms. The reaction centres for a particular reaction were identified using the aforementioned criteria. The templates were then generated by considering the reacting atoms and their one bond neighbours to incorporate the connectivity information of reaction centre and also to incorporate specificity in the templates. Based on the connectivity information, the templates were encoded as SMIRKS strings [196]. SMIRKS string uses a SMARTS representation for querying atoms that match a particular substructure, which is essential for the network generation scheme [197]. The SMARTS string incorporates the connectivity between the atoms and also places additional constraints on the type of atom (alipahtic/aromatic carbon, primary/secondary alcohol, etc,). Furthermore, sanity checks based on charge, valency and hybridization are placed on the extracted template to ensure feasibility of the reaction.

A reaction is 'performed' by the application of a template to the substrate. The algorithm first checks for a substructure match between the reactant part of the template and the sub-

strate. This is done by solving the Ullman sub-graph isomorphism problem [198]. Once a substructure match is found, the edits in the molecular graph of the reactant are performed based on the template and products are generated. The network generation begins with the root node on the DAG obtained as the Bayesian network. The candidate molecule identified through fingerprint similarity is used as the substrate. All the templates generated are applied on the substrate to check for substructure matches. No filtering of template based on conditions, kinetics or yields were performed. If a match occurs, the reaction is 'carried out' and products are generated. In certain scenarios, the candidate molecule can be one of the substrates in a reaction. The reactant side of the template in such a case consists of multiple units, but, as the starting substrate is just a single molecule, a search of all templates is carried out to identify reactions where the molecule under consideration matches one of the reactant template units. In case of a match, the reaction is carried out by artificially introducing the remaining substrates. The products generated then are used as substrates for the next phase of the network generation with the templates being applied on them. The network generation procedure is continued in a breadth-first fashion for a pre-determined number of generations. At the end of the generation step, similarity tests are performed between the functional groups of the other PCs and the elements of the network. This allows for pruning of the network to discard pathways that do not match with the information present in the spectroscopic sensor measurements. The graph structure of the Bayesian network is enforced in such a way that the functional groups identified for the other PCs match with the functional groups of the molecules obtained as products. It is to be noted that in the case of the product being formed as a result of multi-step reaction, the intermediate molecule is also included in the reaction network. Chemically valid pathways between the different molecules that have not been identified by the Bayesian graph structure learning algorithm are still included in the reaction network for completeness.

Other runs of the network generation steps are performed with the candidate molecules for the other PCs as well. For a component that is only a child node in the Bayesian network,

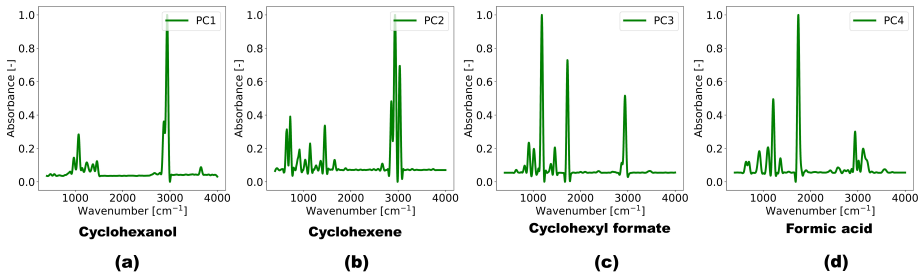


Figure 3.3: (a) FTIR spectrum of pseudo-component 1 (PC1), (b) FTIR spectrum of PC2, (c) FTIR spectrum of PC3, (d) FTIR spectrum of PC4

i.e., only the product of a reaction, the same network generation protocol is followed but with an inversion of the templates (similar to a retrosynthetic approach). For an intermediate node, both the forward and backward generation steps are followed. This generates multiple viable hypotheses for the reaction network. Performing multiple trials of the network generation with different starting points allows for a more exhaustive search while also refining the network. For example, in a scenario where the network generated from an intermediate molecule does not produce any products that match the fingerprints of its children, a modification is performed to the candidate molecule based on the reaction networks obtained from the other PCs. Reversal of templates also results in detection of other substrates necessary for the reaction that are not captured by the spectroscopic data.

3.5 Results and Discussion

3.5.1 Results

As mentioned in Section 3.3, synthetic data generated from known molecules and reactions were used as one of the validation datasets. The deconvolved spectra for the synthetic data are presented in Figure 3.3.

Evaluation of classifier performance

The ^{1}H-NMR spectra was used in the study as means of constraining the tensorial factorization.Though the factorization scheme generates individual 1H-NMR profiles for each

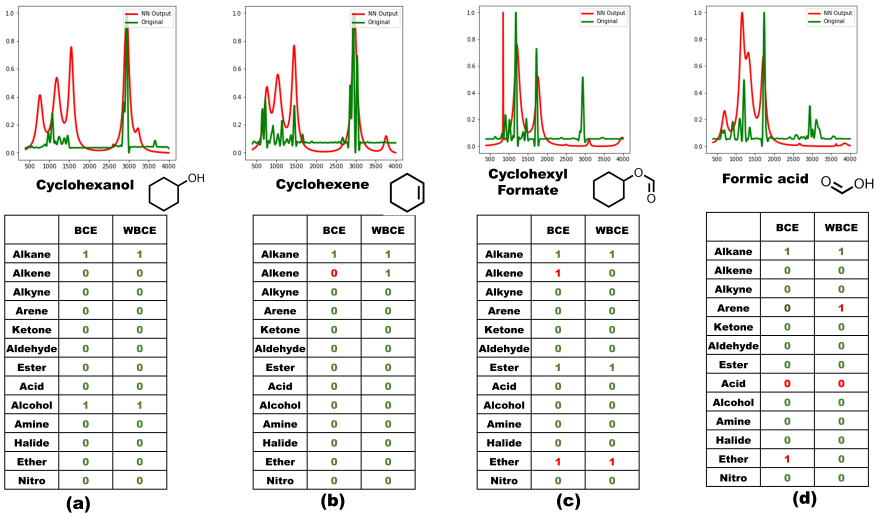
component, the resolution of the signatures was not high enough to be informative. Hence only FTIR measurements were used for classification by the CNN. The presence of functional groups in each PC identified by the deconvolution was determined using a 1-d CNN. The spectrum of each PC was rescaled to have absorbance between 0.0 and 1.0 along with down-sampling to match the resolution of 4 cm^{-1}. The predictive power of the CNN trained with BCE loss across different functional groups is described in Table B.1 and Figure B.2. On comparison, the accuracy, precision, recall, F1 score and specificity of the 1-d CNN was on par or better than architecture described in literature [199, 200]. The classification of alcohols showed the worst performance (F1 score of 0.88) in comparison to Wang et.al[199]. Weighing of the loss term for the less represented classes resulted in an improvement in the F1 score to 0.90 with the recall improving to 0.915 in the case of alcohols though this decreased the precision to 0.89. An overall increase in recall and decrease in precision was noted with the use of WBCE loss. In order to account for the overall drop in precision, the CNN trained on BCE loss was used for further inference. In this case of a prediction of 0 for the alkene, alkyne, alcohol, acid and aldehyde, the same spectrum was classified using the CNN trained on WBCE loss to ensure that the functional group was actually absent and in the case of a mismatch in classification by the two networks, the WBCE classification was chosen for these cases. While this methodology introduces more false positives into the classification, the detection of the presence of a functional group was deemed more vital for reaction network generation.

Comparison of the molecular F1 score and molecular perfection rate (MPR) as introduced by Fine et.al[181] in Table B.2 reveals that our model performs better in the recognizing all the functional groups indicated by the spectrum with the addition of reconstruction drastically improving the MPR. Weighing of the classes improves the recall with a little drop to the MPR (owing to misclassification of highly represented groups).

Network generation for synthetic data

The classification of the decomposed spectra from the synthetic data is shown in Figure 3.4. The green entries in tables indicate the correct identification (presence or absence) of a functional group in the PC while the red entries indicate a wrong classification (presence or absence) of a functional group. In the case of cyclohexanol (PC1), it can be seen that both the networks trained on BCE and weighted BCE (WBCE) losses were able to recover the correct functional groups indicated by the spectrum. In the case of cyclohexene (PC2), the WBCE-network is able to detect the C=C, but not the BCE-network. The BCE-network was able to identify the ester part of cyclohexyl formate (PC3), but it also identified a C=C bond, while the WBCE-network did not identify the C=C bond and was also able to identify the ester group. However, both networks identified an ether group, which was not present in the molecule. Unfortunately, neither of the networks were able to correctly identify formic acid. Analysis of the reason behind this failure revealed that the deconvolution of the spectroscopic mixture resulted in the peak corresponding to the OH group of the acid being assigned to PC1 (alcohol), while peaks of the ester corresponding to the C-O bond of the ester were assigned to PC4, explaining the ether classification by the BCE-network. Futhermore, since C=O and aromatic C=C stretches occur close together, the WBCE-network assigns an aromatic functional group to the molecule. A comparison of the original spectrum each component with the deconvoluted spectrum is presented in Figure B.3.

The reaction network identified by our method for the synthetic data is presented in Figure 3.5. The reaction networks shown have been filtered from all generated network hypotheses based on domain knowledge (i.e., scaffold information of molecules) of the process. As mentioned in Section 3.4, the partial-MACCS keys of each of the PCs were generated based on the detected functional groups. A fingerprint similarity match with molecules present in the database was performed to generate an initial candidate for each of the PCs. As an initial pass to match multiple candidates, when aromatic rings have not been



	BCE	WBCE
Alkane	1	1
Alkene	0	0
Alkyne	0	0
Arene	0	0
Ketone	0	0
Aldehyde	0	0
Ester	0	0
Acid	0	0
Alcohol	1	1
Amine	0	0
Halide	0	0
Ether	0	0
Nitro	0	0

(a)

	BCE	WBCE
Alkane	1	1
Alkene	0	1
Alkyne	0	0
Arene	0	0
Ketone	0	0
Aldehyde	0	0
Ester	0	0
Acid	0	0
Alcohol	0	0
Amine	0	0
Halide	0	0
Ether	0	0
Nitro	0	0

(b)

	BCE	WBCE
Alkane	1	1
Alkene	1	0
Alkyne	0	0
Arene	0	0
Ketone	0	0
Aldehyde	0	0
Ester	1	1
Acid	0	0
Alcohol	0	0
Amine	0	0
Halide	0	0
Ether	1	1
Nitro	0	0

(c)

	BCE	WBCE
Alkane	1	1
Alkene	0	0
Alkyne	0	0
Arene	0	1
Ketone	0	0
Aldehyde	0	0
Ester	0	0
Acid	0	0
Alcohol	0	0
Amine	0	0
Halide	0	0
Ether	1	0
Nitro	0	0

(d)

Figure 3.4: (a) Classification for cyclohexanol, (b) Classification for cyclohexene, (c) Classification for cyclohexyl formate, (d) Classification for formic acid

detected by the classifier, the functional groups detected were attached to both straight chain and cyclic scaffolds with the cyclic scaffold being chosen based on prior knowledge of the system. The degree of filtration brought about by including scaffold information is provided in Table 3.1 The introduction scaffold information brings about a 4 to 10 fold reduction in the number of molecules considered. But most of the filtering of the molecule occurs at the functional group matching stage and during the enforcement of the graph constraint. There is no distinction between label-based filtration and scaffold-based filtration for PC4 as the aromatic scaffold has been incorporated into the functional group labels. It is also important to note that the Bayesian network identified from the spectra forms a subset of the overall chemical reaction network. Compounds in the chemical reaction network whose signatures have not been indicated by the pseudocomponent spectra are still included if one of their products (or subsequent products) are indicated by the spectroscopic data.

The templates were applied to the initial candidate molecule for the root node of the Bayesian network and the resultant network generated is shown in Figure 3.5(c). The reaction network developed from candidates for PC2 is shown in Figure 3.5(d). In this run of network generation, the reaction templates were also applied in reverse to the molecule to check for a molecule similar to PC1 being a precursor for the candidate molecule in

Table 3.1: Filtering of molecules based on scaffold information for synthetic data
Total number of molecules : 1808254

Molecule	Filtered based on labels	Filtered by addition of scaffold	Filtered by graph-constraint only using labels	Filtered by graph-constraint using labels and scaffold
Cyclohexanol	614	137	26	1
Cyclohexene	2096	203	34	1
Cyclohexyl formate	6420	995	107	1
Formic acid	79264	79264	0	0

question. It is interesting to note that even though formic acid was not identified correctly by the CNN, the reversal of the template allowed for the discovery of formic acid as the other node in the network. Similarly, reversed templates were applied to the candidate molecule for PC3 and the network was propagated in the reverse direction upto PC1 (Figure 3.5(e)). The network generation algorithm in each case was run for 4 generations and similarity tests were performed between fingerprints of the other PCs and products of the network. The graph constraint was enforced in such a way that multiple step reactions were also considered. Additional arcs identified by the network generation algorithm have also been included in the reaction network showcase the ability of the algorithm in discovering interactions that are not captured by the sensor measurements. Based on the information obtained from classification of PC4, none of the networks generated with any initial molecule for PC4 showed nodes similar to candidates for the other PCs and hence no viable network was generated as seen in Figure 3.5(f)

Reaction network for a known starting molecule

The fidelity of the network generation algorithm in generating networks that are close to the ones indicated in literature was tested out by applying the algorithm to a known starting molecule and examining the resulting network structure. 2-Phenoxyethylbenzene (PEB), a model compound often used to describe reactions of lignin-type compounds was used as

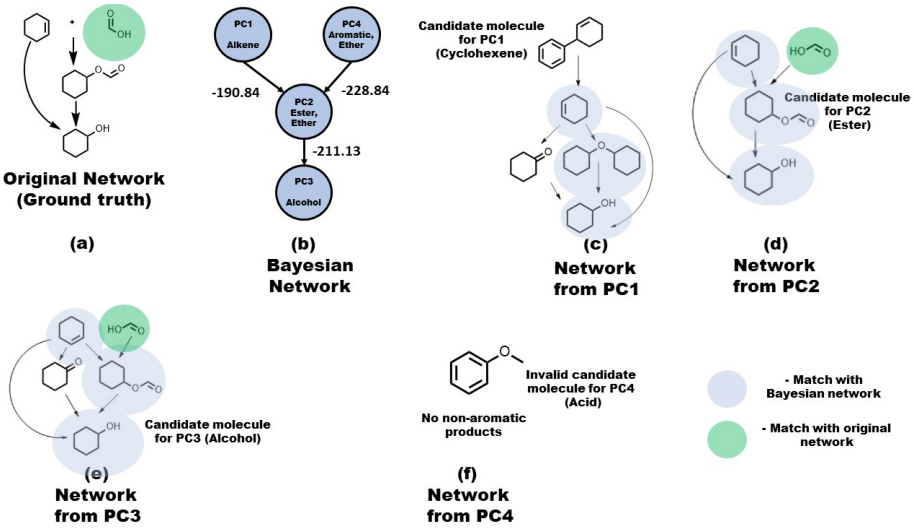


Figure 3.5: (a) Ground truth network used in data generation, (b) Bayesian Network identified from spectra (c) Reaction network starting from candidate molecule for PC2, (d) Reaction network starting from candidate molecule for PC3, (e) Reaction network starting from candidate molecule for PC1, (f) Reaction network starting from candidate molecule for PC4

the starting molecule.

The network structure identified using Bayesian structure learning on the factorized spectroscopic data obtained from experiments indicated a 3 node graph with polycyclic aromatic compounds converting into aromatic alcohols, carbonyls and alkenes (Figure 3.6(a)). The reaction network obtained by using PEB as the starting point is depicted in Figure 3.6(b). The network showcased the cleavage of the ether bond to produce alcohols (phenol). Oxidation of the phenolic compounds generated carbonyls and acids. The cleavage of the ether followed by dehydration of the alcohol also produced styrene-like molecules. Similar reactions have been described in the literature: the end products of HTL of lignin components of biomass consist of phenolic and carbonyl compounds [5, 24].

Validation of workflow for Biomass HTL

Having validated our methods with the ground truth for the synthetic data, the methods were validated using the spectroscopic data for HTL of biomass. The results of the tensorial factorization are presented in Figure 3.7. 4 components (excluding water) were identified by core consistency diagnostics and the resultant projections of the data (except for water)

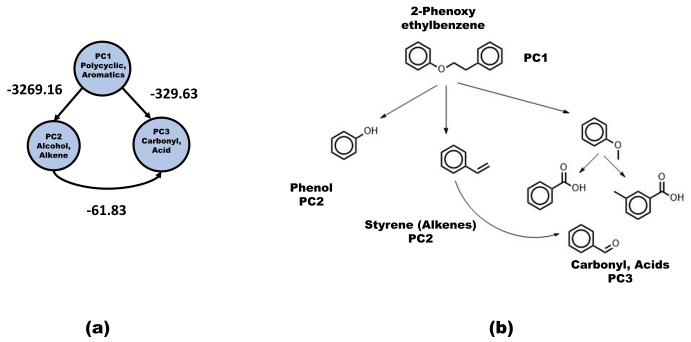


Figure 3.6: (a) Bayesian network obtained for PEB hydrolysis, (b) Candidate reaction network starting from PEB

along the residence time, process conditions and wavenumber/chemical shift modes are shown in Figures 3.7(b), 3.7(c), 3.7(d), 3.7(e), respectively. Since the tensorial decomposition assumes the data to be a linear combination of the component spectra, it is assumed that PC1 to PC4 contain no water as water's spectrum has been extracted separately. Only spectra of PC1 to PC4 were used for further studies. Exclusion of water as a bystander/solvent molecule is done based on knowledge of the process. The results of the decomposition along including water is presented in Figure B.4. The factorized spectra were fed as inputs to the CNN classifier to identify the functional groups present in each PC. From Figure 3.8, it can be seen that PC1 consists of aromatics with ester groups. Visual inspection of the spectrum also indicated peaks at 1040-1250 cm^{-1} indicating the presence of ethers as well as peaks corresponding to aromatic C-H and C=C bends and stretches (3000-3200 cm^{-1} and 1420-1650 cm^{-1}). PC2 consisted of aromatic alcohols, and PC 3 contained carbonyls (aldehydes). PC4 consisted of simple aromatic compounds with alkane side chains. The predictions of the classifier were also verified by visual inspection.

The Bayesian network structure along with the identified reaction networks are presented in Figure 3.9. These networks are results of the iterative application of the templates on the candidate molecules followed by similarity tests. This routine of template application and

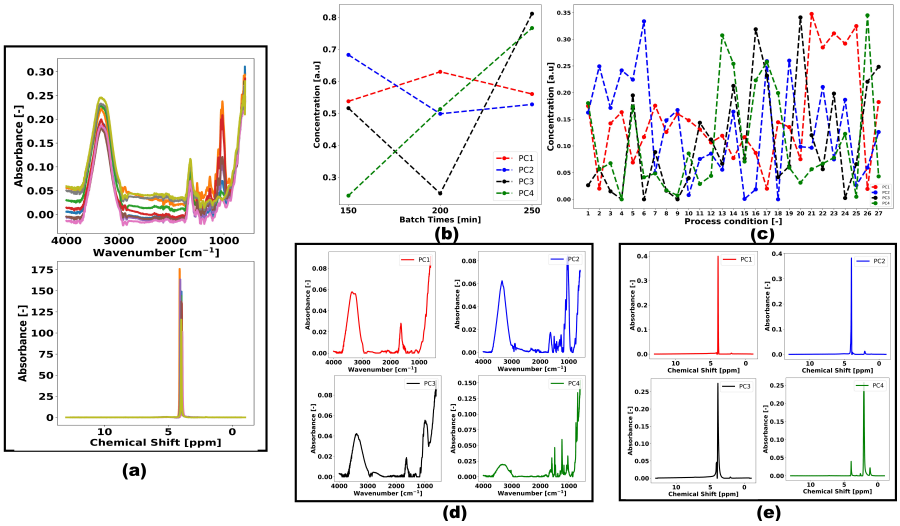
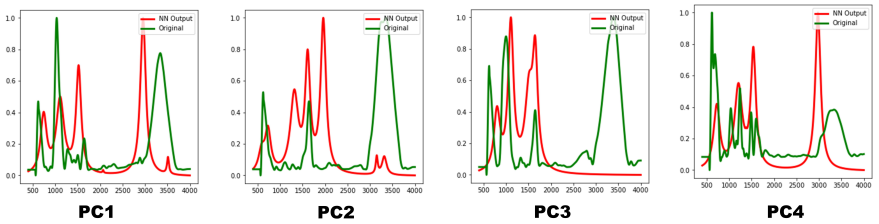


Figure 3.7: (a) FTIR and ^{1}H-NMR spectra for HTL of Biomass, (b) Projection along residence time mode, (c) Projection along process condition mode, (d) Resolved FTIR spectra of each PC, (e) Resolved ^{1}H-NMR spectra of each PC



	PC1	PC2	PC3	PC4
Alkane	0	0	1	1
Alkene	0	0	1	0
Alkyne	1	0	0	0
Arene	1	1	0	1
Ketone	0	0	0	0
Aldehyde	0	0	1	0
Ester	0	0	0	0
Acid	0	0	0	0
Alcohol	0	1	0	0
Amine	0	0	0	0
Halide	0	0	0	0
Ether	1	0	0	0
Nitro	0	0	0	0

Figure 3.8: Classification of biomass PCs

similarity test was performed for multiple epochs until structures were found that matched with the connectivity information of the Bayesian network as well as the functional groups detected. The network structure revealed a transformation of the ether group into alcohols and carbonyls, finally cleaving to form substituted aromatics. The reaction network identified from the candidate for PC1 is depicted in Figure 3.9(b). The course of reactions as seen in panel b followed the trends seen in pyrolysis of lignin components of biomass, while panel c described the breakdown of the cellulosic structures. Scaffold information used for the synthetic data only included information on polycyclic compounds for PC1 and mono-cyclic compounds for PC2, PC3 and PC4. No explicit information on actual scaffold structure of molecule was provided.

Panels d, e and f depict the networks generated by applying templates in the forward and reverse directions on the candidate molecules for PC2, PC3 and PC4, respectively. As mentioned earlier, different runs of the network generation algorithm were conducted with PC2, PC3 and PC4 as starting points. Based on their positions on the Bayesian network, the reaction templates were applied only in the reverse direction (for PC2 and PC4) or in both forward and reverse directions (PC3) to generate additional reaction network hypotheses. It is important to note here that the networks presented in this work have been filtered from all generated hypotheses based on domain knowledge of the process regarding molecular scaffold information. These networks also seemed to match with the general trend noticed in hydrolysis of biomass [24, 89, 109]. It also interesting to note that the candidate molecules identified in each step closely resemble model compounds presented in literature for the thermal pyrolysis of biomass [6, 201].

3.5.2 Discussion

We have discussed the generation of a reaction network hypothesis for the HTL process of biomass by incorporating human expertise in transforming the PC network to real-life chemistries in our previous work [24, 109, 201]. In this work, we present methods to auto-

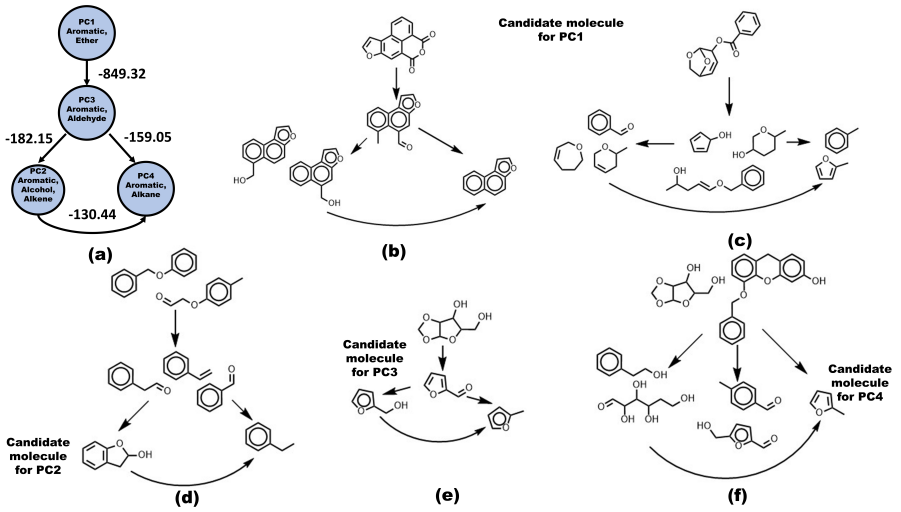


Figure 3.9: (a) Bayesian network for biomass, (b) Reaction network starting from lignin-like candidate molecules for PC1, (c) Reaction network starting from cellulose-like candidate molecule for PC2, (d) Reaction network starting from candidate molecule for PC3, (e) Reaction network starting from candidate molecule for PC4

mate these routines with some added human intuition by way of using 1-d CNNs as well as reaction templating and fingerprinting schemes.

Human expertise in deciphering a FTIR spectrum consists of identification of the position and shape of peaks in the spectrum and assigning functional groups based on these features. This contextual mapping of spectral peaks was what we sought to achieve in an automated fashion through the use of the CNN, and the analysis of the feature space learnt by the neural network validated the same. As seen in Figure B.4, the CNN extracted information about specific wavenumber regions showcased by the functional group. To understand the peak shape information extracted by the CNN, the filters learnt during training were inspected. From the visualization of the filters, it was evident that the different filters learnt corresponded to different peak shapes (see Figure B.5). Each panel in Figure B.6 depicts the filter learnt by the CNN (top and middle) and the activation map (output of the first convolutional layer) with respect to that filter (bottom) for methanol. This is important, since different functional groups show up in the FTIR spectrum not just at different wavenumbers but also as different shapes. Alcohols, for example, show a wider peak for their O-H bond

than the sp^3 C-H bond.

The reaction network identified for the synthetic data also included certain compounds that were not included in data generation, such as cyclohexanone, which was identified as a potential candidate for PC3. This was due to the fact that cyclohexanone also undergoes reactions similar to what has been described about PC3 based on the data shown to the network generation algorithm. The ability of our methodology to identify other potential candidates for a certain PC its ability to effectively explore the chemical space. While the networks presented in this work have been filtered to include reasonable scaffolds for candidate molecules, the reaction networks generated for different scaffolds only differ in the base scaffold structure and not in the transformation between the species. A representative image showcasing the inclusion of scaffold information in reaction network generation has been presented in Figure 3.10 It is to note here that the synthetic was generated to ensure kinetics of all the reactions considered were activated by the temperature profiles leading easier deconvolution. While the focus of this work is not test the effectiveness of the deconvolution technique, we do acknowledge that it plays an important role in determining the reaction network. It is also important to note that each component in the synthetic data had only one functional group. The molecular F1 score and MPR are known to decrease when number of functional groups in a molecule increase[181] potentially leading to misidentification of functional groups indicated by the spectrum. With respect to HTL of biomass, literature reports the breakdown of lignin-like compounds with the cleavage of the ether and C-C bonds [202]. Phenols are found in abundance as products of HTL and are formed through ether cleavage [202, 203]. Dehydration reactions have also been reported in literature resulting in alkenes [204]. Alkylated aromatic structures have also been proposed as products of the thermal pyrolysis process [24]. The hydrous pyrolysis of carbohydrate-like molecules typically results in various saccharides and glucose. Under the conditions of this study (temperature, acidic/alkaline nature), carbohydrates has been known to undergo oxidation to carboxylic acids via retro-aldol condensations which undergo decarboxylation

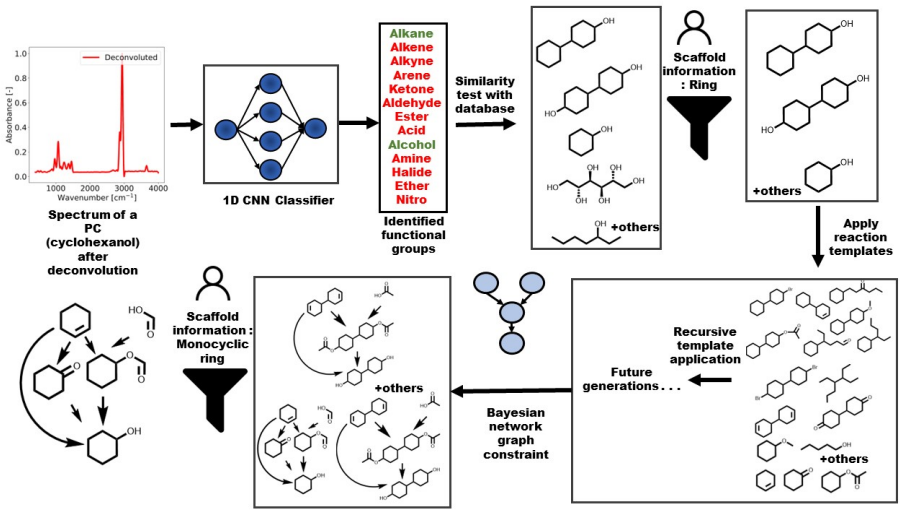


Figure 3.10: A representative workflow of the reaction network generation algorithm showcasing the incorporation of scaffold information.

reactions to form various carbonyls [133]. Furthermore, carbohydrates can degrade to form oxygenated compounds such as hydroxymethyl furfural (HMF) and lactic acid [205]. These concurrent reactions were captured by the reaction networks generated by our methods. As seen in Figure 3.9, panels b and d described the breakdown of lignin-type compounds. It is important to highlight here that phenols and alkylated aromatics, along with styrene-like structures, has been correctly identified in the reaction network. Similarly, the breakdown of carbohydrates have also been recovered in panels c,e and f of Figure 3.9. HMF is typically used as a model compound in describing reactions of biomass feedstocks [135] and the presence of substituted furfural-type molecules in the automatically generated networks points at its fidelity towards identifying reactions and compounds describing biomass conversion.

3.6 Conclusions

In this work we present a methodology to automate the mapping of data-driven hypotheses to candidate chemistry for reaction transformations in systems where accurate information about the species and reactions are absent, which is a formidable task. The reaction network

identified for biomass seems to fit well with the kinds of reactions and products described for the HTL process in the literature. The CNN-classifier identified ethers, alcohols, aromatics and carbonyls in the reaction mixture. Our method identified both the ligninic and cellulosic aspects of the hydrothermal pyrolysis of biomass. In our previous , we showed that these two pathways were the most significant reaction schemes occurring during the HTL process and this work was able to extract networks that indicate these transformations.

Though we present the application of our methods on the HTL process as a validation, the deconvolution and functional group detection are process-agnostic and can be used in developing reaction networks for any process with spectroscopic measurements of the process available. The tensorial factorization scheme could be applied to different types of spectroscopy measurements. The CNN is trained on FTIR spectroscopy measurements that are informative about the functional groups. Other types of spectroscopic measurements that provide a similar information could be used in training such that the molecular fingerprint obtained is informative enough. The lack of scaffold information in FTIR spectroscopy measurements forces a degree of human intervention as described in this work. With the addition of scaffold information through other sensor measurements the routine can be truly automated. The other steps of the methodology do not depend on any type of data and can be readily adapted to other systems. Furthermore, the reaction networks can be modified based on the operating conditions by introducing data points at the new operating range into the network generation scheme.

Pruning of the reaction network is done at present based on similarity matches with functional groups and with some human expertise on the system under consideration. Information about recurrent sub-structures present in the reaction network provides a very informative criterion on the branches to be included in the pruned network. Furthermore, kinetic information about the different reactions can also provide a means of pruning the network. Our other studies involve the estimation of kinetic parameters for the reactions and extending it towards control and monitoring of the process [206].

3.7 Limitations

1. Analyses performed in our other works[206] indicate that the results of the deconvolution are influenced by the amount of noise present in the spectroscopic measurement. Though not a theoretical limitation, this does a pose a practical limitation on the validity of the deconvolution for samples with low signal-to-noise ratios.

2. The type of spectroscopic measurement provided to the system affects the expanse of the networks generated. FTIR spectroscopy as discussed in this work does not provide information about the scaffold of the molecule and is limited to the functional groups to be detected. This results in multiple candidate molecules (and hence reaction networks) with similar reactions but differing in just the scaffold structure and requires an additional layer of human input in filtering out these reaction networks based on domain knowledge. It is important to note that this a limitation due to the nature of information present in the sample itself and not necessarily of the workflow. Other types of spectroscopic measurements can be readily incorporated into the deconvolution information and classifiers built to jointly identify functional groups and scaffolds can be incorporated. More specific candidate molecules can then be deducted thus reducing the number of hypotheses generated. This forms a part of our future work.

3. The number of functional group classes chosen for the study were restricted to 13 to ensure sufficient training samples for each class. While this places a limit on the extensiveness of the chemical space search, the functional groups chosen do encompass a majority the most frequently occurring functional groups in the database. Similarly, extension of the classifier to aqueous phase is restricted by the lack of training data and can lead to incorrect assignment of functional groups.

Chapter 4

Spectrum-constrained deep generative model for monitoring of complex reaction systems

4.1 Abstract

Identifying molecular structures of components of a reaction mixture from spectroscopic measurements is not a trivial task. Development of chemical models for complex reaction mixtures is not straightforward owing to difficulties faced while interpreting sensor measurements and the expanse of the chemical space. In this work we detail a molecular generative model that can be used for identification of the molecular structure of components in the reaction mixture given the infrared spectrum of the reaction mixture. A generative adversarial network conditioned on the spectrum of molecules is detailed in this work wherein the generator generates latent representations of molecules pertaining to the given spectrum condition. The latent representation is decoded to generate a string representation of the molecule using a pre-trained Long Short Term Memory(LSTM) decoder. The decoder is trained to generate SMILES and SELFIES strings of molecules given their molecular structure by encoding the structure using Message Passing Neural Networks (MPNNs). We show that our *Graph2SMILES* encoder is able to populate the encoding space based on molecular sub-structures and translates the molecular structure with an average BLEU score of 0.91 with 82.7% validity of translation. The methodology is tested on experimen-

tal data obtained from batch hydrothermal conversion of biomass and reaction networks for the system under multiple operating condition are developed depicting the breakdown of cellulosic structures.

4.2 Introduction

Optimizing the valorization of distributed feeds forms a key aspect in establishing an effective circular economy of chemical and biological products and wastes [1, 2]. Recovery and upgrading techniques associated with such physically and chemically heterogeneous feedstocks involve numerous reactions of multiple reactive species and developing a first-principles based model is intractable due to the sheer number of variables in the system. Traditional approaches to identifying the reactive species and the reactions in systems involving complex feedstocks rely heavily on domain expertise and are based on handcrafted model compounds for the system under consideration [5, 6]. The choice of model compounds is subject to human bias and the modeling methodology does generalize well across different classes of reaction systems.

An alternative is to utilize sensor measurements of the reaction to track the reaction progress [9, 13]. In this regard, spectroscopic sensor measurements provide a quick and reliable means of extracting chemical information in a process. For a well-defined system (i.e., systems where both species and reactions are known) spectroscopic measurements can track the relative change in chemical composition of the species across the operating conditions. For systems where species and reactions are unknown *a priori*, spectroscopic measurements can provide a means of identifying the different reactive species in the feedstock. Identification of the individual species in evolving spectroscopic data forms the key aspect of multivariate curve resolution [18].

Our previous works have successfully employed spectroscopic measurements to identify reaction networks in complex systems such as bitumen and biomass [26, 142]. Curve resolution in these previous works was performed using non-negative tensorial factorization to

identify spectroscopic signatures of individual reactive components. These signatures were used as an input to a Bayesian structure learning algorithm to identify the structure of the reaction network followed by identification of reactions and species through human expertise or automated chemical database searches. An issue encountered with curve resolution, especially for measurements with low signal-to-noise ratio, is that the resultant deconvolutions of spectra do not necessarily correspond to the signatures of the actual molecules in the reaction mixture. Similarly, the Bayesian structure identified as the reaction network tends to be fully connected with spurious links between some species [206]. This requires additional human expertise in removal of such spurious arcs or in weighting of regions of interest in the deconvoluted spectra. In this work we aim to develop an one-shot generation of molecules that represent the reaction mixture at a given process condition based on the spectroscopic measurement of the process at that condition.

Machine learning-based molecule generation schemes employ Recurrent Neural Networks (RNNs) [207, 208], Variational Auto-Encoders (VAEs) [49, 50], Generative Adversarial Networks (GANs) [47, 51, 52, 209], Transformers [210–213] and Reinforcement learning [44, 214, 215]. The generative approach used varies based on the representation of molecule being generated. RNNs are typically employed in sequential generation of molecules represented as Simplified molecular-input line-entry system (SMILES) strings or a sequence of graph edits to molecules represented as graphs [48, 216, 217]. VAEs and GANs have been employed to molecules represented as strings, graphs or as molecular fingerprints [47, 49, 51]. Typically, one-shot generation of the molecular graph is a hard problem due to the permutation invariance of the adjacency matrix, but methodologies have been developed for small molecule generation in the literature [47]. An easier approach is to generate a latent vector representation of the molecule, which is then decoded by a hetero-decoder to generate a molecule.

Conditional molecular generation has been a topic of active research in drug discovery and researchers have employed VAEs and GANs to generate drug-like molecules that

match a prescribed condition (scaffold structure [59, 60, 218–220], ligand-binding properties [221–223], target genome expression [42, 224], etc). But from a process engineering perspective, the task of identifying molecular entities from sensor measurements of the reaction has not been explored in the literature. It is important to note that several studies have been performed in the literature towards automated elucidation of molecular structure from spectroscopic measurements of pure compounds [225, 226]. Most of these studies require some knowledge such about the molecule whose structure is to be identified. It is not uncommon to use the molecular formula to limit the structure elucidation to a space of possible isomers to reduce the computational complexity of the problem. In the case of structure identification for mixtures, existing methodologies require a list of possible components in the mixture.

The focus of this work is to infer the reaction network of a complex reaction system at different operating conditions solely from spectroscopic measurements of the reaction process and without *a priori* knowledge of the species. To this effect, we employ a conditional GAN-based molecular generation routine that incorporates the spectroscopic sensor measurements to generate latent representations of molecules that are decoded using a RNN-based decoder trained on encodings obtained from Graph Convolutional neural networks. We identify the reactions occurring at a particular condition by applying algorithmically extracted reaction templates.

4.3 Methods

The molecular generation routine is carried out using a GAN trained to generate latent vector representation of molecules conditioned on the input spectrum. The latent vector is decoded into a SMILES string using a pre-trained RNN-based decoder.

4.3.1 Datasets and data preparation

Reaction smarts [197] from the US Patent Office database [185] were parsed to identify individual molecules. The Atom Atom Mapping ids of the atoms were stripped and sanity checks were performed using RDKit [186] to generate a database of SMILES. Infrared spectra of molecules were obtained through web scraping from the National Institute of Standards and Technology Chemistry webbook. The spectra were pre-processed as mentioned in Chapter 3 to fall between 400 cm^{-1} and 4000 cm^{-1} with absorbances scaled between 0 and 1. SMILES string of the molecules whose spectra were collected were concatenated to database. Unique smiles were identified and any salts were removed. SMILES and Self-Referencing Embedded Strings (SELFIES) [227] were generated for molecules in the list and were tokenized using a regular expression [228] and in-built tokenizing functions, respectively. The dataset generated 1.24 million unique samples of SMILES/SELFIES strings and 11062 IR spectra. Only tokens that had a frequency greater than 2000 were retained with other tokens being represented by an identifier for unknown atom or unknown number in the dataset. Experimental FTIR spectroscopic measurements from our previous studies [201] were used as a test dataset to present our methods. More details regarding the dataset are made available in Appendix C.1

4.3.2 Graph2SMILES translator

A graph-based encoder-decoder architecture is employed in this work. The encoder consists of multiple graph-convolutional layers with the molecular graph as the input. In the most general form the convolutions update the state vector h_v of each node at a step t based on the message function M_t and an update function U as given by

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^{(t)}, h_w^{(t)}, e_{vw}) \tag{4.1a}$$

$$h_v^{(t+1)} = U(h_v^{(t)}, m_v^{(t+1)}) \tag{4.1b}$$

As can be seen from Equation 4.1, the message m_v^{t+1} of node v depends on the states of the node and its neighbors ($N(v)$) along with the with the edge-feature (e_{vw}) for the edge connecting the nodes and the neighbors. In this work, we use a version of the graph convolution as defined by Gilmer et al. [229], which allows for a vector-valued edge-feature. The states of all nodes in the graph are pooled to generate a context vector z which encodes the characteristics of each node and edge of the molecular graph. More details regarding the features used for nodes and edges can be found in Appendix C.2.

The context vector is used as an initialization of the states of a RNN-based decoder. This formulation is similar to that of a *seq2seq* neural machine translation. The Long Short-Term Memory (LSTM) formulation of the RNN is used as the decoder to predict the SMILES string of the given molecular graph one character at a time. The decoder is trained without any teacher forcing. The token-level prediction of SMILES is tasked as a multi-class classification problem with the decoder predicting the conditional probability of the next token in the SMILES sequence given the previous tokens of the sequence. The frequency of occurrence of each token varies drastically across the training data and hence Focal loss [230] is chosen as the loss function to be minimized as opposed to the traditional cross-entropy loss. The Focal loss includes a modulating term that further penalizes incorrect predictions across hard-to-classify samples and is given by

$$\text{Focal loss } = -\alpha_c(1-p_c)^{\gamma} log(p_c) \tag{4.2}$$

where p_c denotes the predicted probability of the sample belonging to class c, γ is the focusing parameter and α_c is the weight associated with class c.

4.3.3 Spectrum constrained GAN

The spectrum-constrained GAN consists of two networks known as the Generator and Critic, which are trained in an adversarial fashion. The Critic (D) assigns scores to its inputs with a higher score for 'real' samples or samples belonging to the training data distribution. The Generator (G) generates samples intended at fooling the Critic, i.e, the Generator tries

to produce 'fake' samples that the Critic identifies as real (assigns a higher score). The conditional GAN is trained to minimize the 1-Wasserstein distance between the distribution of the training data and the distribution of the samples generated by the generator [231]. The 1-Lipschitz continuity of the function formulated as the Critic network is ensured by using a gradient penalty scheme [232]. The Generator (L_G) and Critic (L_D) losses are given as

$$L_D = \mathbb{E}_{z\sim P_{real}, c\in C}[-D(z,c)] + \mathbb{E}_{x\sim \mathcal{N}(0,1), C\in C}\left[D(G(x,c),c)\right] + \lambda[\|\Delta_{\hat{x}} D(\hat{x}) - 1\|_2^2] \quad (4.3a)$$

$$L_G = \mathbb{E}_{x\sim \mathcal{N}(0,1), C\in C}\left[-D(G(x,c),c)\right] + \beta f(z,c,l) \quad (4.3b)$$

where c represents the conditioning spectrum and λ is the regularization weight of the gradient penalty. The f term corresponds to the loss associated with the generated molecule containing the functional groups indicated by the spectrum and is formulated as the cross-entropy loss of a classifier trained on predicting functional groups based on the spectrum and molecular latent vector z. This term of the loss is high when the Generator generates a molecule that does not contain the functional groups indicated by the conditioning spectrum c and β is the regularization term associated with this loss. Details on the structure of the Generator, Critic and the functional group penalty classifier can be found in Appendix C.4.

The GAN generates latent representations of the molecule that matches the spectrum condition given, which is then decoded using the pre-trained decoder mentioned in Section 4.3.2. Figure 4.1(a) and Figure 4.1(b) depict schematics of the *Graph2SMILES* and spectrum-constrained GAN, respectively.

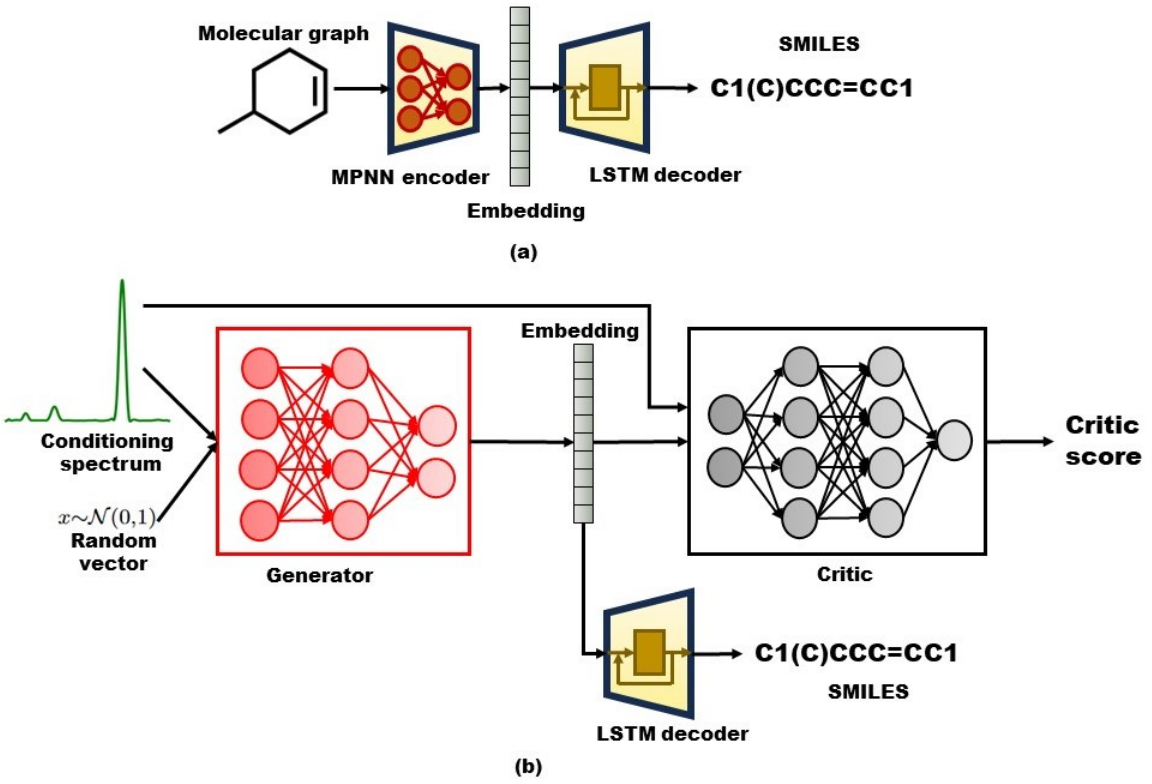


Figure 4.1: (a) Schematic of *Graph2SMILES* translator. (b) Schematic of spectrum constrained GAN

Identification of the reaction network at the operating condition is performed by application of reaction templates to the molecule generated by the GAN. The reaction template encodes reaction rules indicating the conversion substrate(s) into product(s). In this work we employ algorithmically extracted reaction templates. Since the spectrum of the mixture is provided as a conditioning input, the generated molecules contain motifs of all components in the reaction mixture at the sampling instant. Substructure matching is performed between the generated molecule and templates of reactants in the bank of reaction templates. If a substructure match occurs at any part of the molecule, a reaction is performed at the location and products are generated. Further generations of the reaction network are generated by recursive application of the templates on the products of the previous generation. Th reaction network developed at a process condition comprises of all possible reactions occurring at the condition. Further details on the reaction template generation and

the reaction network generation can be found in our previous work [142].

4.4 Results and discussions

4.4.1 Graph2SMILES translator

The *Graph2SMILES* translator was trained using teacher forcing to enhance speed of convergence. A variant of the *Graph2SMILES* architecture was also trained to predict SELFIES strings, which we refer to as *Graph2SELFIES* henceforth. BLEU score [233], used to compute the effectiveness of text translation, was employed to check the accuracy of translation between molecular structure and SMILES [234]. The SMILES-to-SMILES translator developed by Winter et al. (henceforth called the RNN-translator) [235] was used a baseline. The average BLEU score of translation between the two architectures is given in Table 4.1. Appendix C.2 shows the BLEU scores for different testing samples for both *Graph2SMILES* and *Graph2SELFIES* translators.

Model	BLEU Score	No.of molecules with BLEU Score <1	No.of molecules with BLEU Score <1 but Tanimoto similarity =1	Mean Tanimoto similarity of molecule with BLEU score < 1
RNN translator	1	0	-	-
Graph2SMILES	0.91	259	135	0.944
Graph2SELFIES	0.859	351	131	0.821

Table 4.1: Comparison prediction capabilities of different translator based on BLEU score

BLEU score performs a character-level match between the reference (ground truth SMILES) and the hypothesis (predicted SMILES). In the case of SMILES, a strict adherence to a particular ordering rule is not necessary at all times. A molecule can have multiple equivalent SMILES strings and hence a mismatch of tokens at a particular position does not always imply an incorrect translation. Therefore, a test for the similarity between structures of the ground truth string and the predicted (translated) string was performed by computing the

Tanimoto coefficient of similarity between the MACCS fingerprints of the two strings. Additionally, a test for the validity of the translated SMILES was performed. Validity in this context is defined as the ability of the RDKit SMILES parser to generate a valid MolFile of the input SMILES. The SMILES translation was found to have a validity of 82.7 % out of 124,600 samples tested. The mean Tanimoto coefficient between reference and predicted SMILES across all valid predicted SMILES was found to be 0.95. The *Graph2SELFIES* translation provided valid molecules 100% of the time. This is due to the inherent nature of the string as SELFIES was developed with purpose of being an injective mapping, i.e, every SELFIES string always produces a valid molecule. The mean Tanimoto similarity was found to be 0.82.

In an attempt to understand some properties of the encoding dimension of the translators, studies were performed on encodings obtained from a subset of the testing dataset containing 500 samples of varied lengths of SMILES strings. Nonlinear Principal Component Analysis (PCA) with a radial basis function kernel was applied to the encodings from RNN-translator, *Graph2SMILES* and *Graph2SELFIES* translators. The projections showed two distinct cluster in the case of RNN and *Graph2SELFIES* translator and three clusters for the *Graph2SMILES* translator. K-means clustering was performed on the lower dimensional data followed by a test of inter-cluster and intra-cluster structural similarity of molecules. The clustering is depicted in Figure 4.2(a), (b) and (c). The RNN-translator showed lower intra-cluster similarity as compared to the *Graph2SMILES* and *Graph2SELFIES* translators, as seen in Figure 4.2 (d), (e) and (f). An analysis of the length of SMILES/SELFIES sequences in each cluster was then performed. The distribution of sequence length across the clusters is depicted in Figure 4.3. The sequence lengths were found to be fairly evenly distributed across all clusters in the case of *Graph2SMILES* and *Graph2SELFIES* translators while RNN-translator showed a significant demarcation in the sequence length between clusters. This lead to the conclusion that molecules whose SMILES strings had similar lengths were placed together in the encoding space of the RNN-translator.

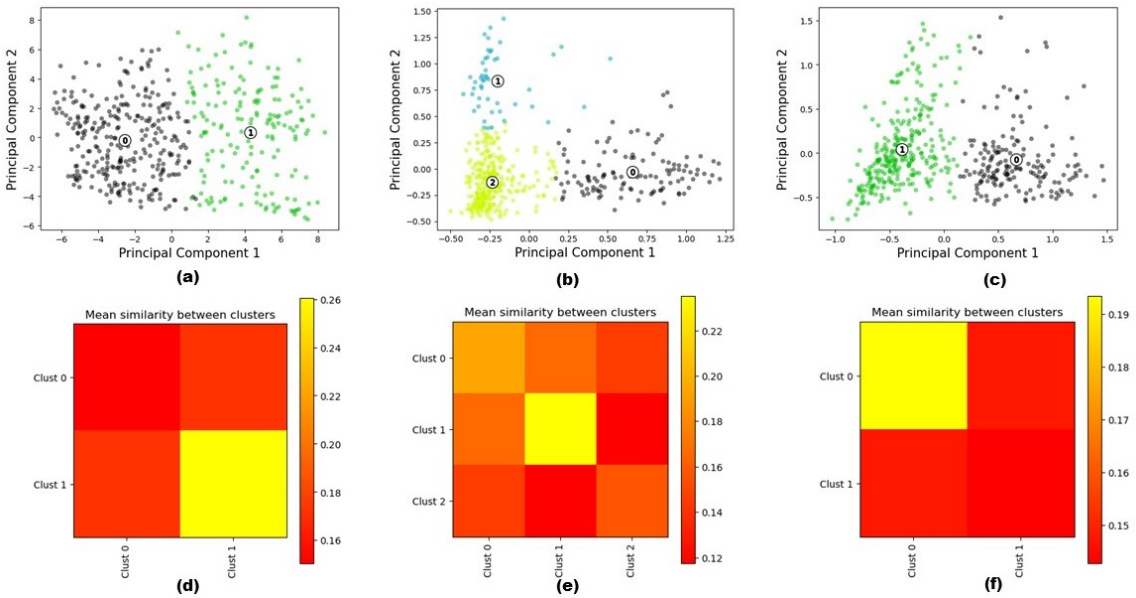


Figure 4.2: (a) K-means clustering of RNN translator encodings (b) K-means clustering of *Graph2SMILES* translator encodings (c) K-means clustering of *Graph2SELFIES* translator encodings (d) Mean intra-cluster Tanimoto similarity for RNN-translator encodings (e) Mean intra-cluster Tanimoto similarity for *Graph2SMILES* translator encodings (f) Mean intra-cluster Tanimoto similarity for *Graph2SELFIES* translator encodings

To further test this hypothesis, the t- Stochastic Neighbor Embedding(t-SNE) [236] plot of the encodings from all translators were obtained as shown in Figure 4.4. The distribution of sequence length across the lower dimension manifold correlated well with the hypothesis. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [237] was performed as an alternative to K-means on the t-SNE manifold to identify density-based clusters. Analysis of the intra-cluster similarity showcased that structurally similar molecules were placed close to each other in the encoding space of all translators but sequence length distributions of each cluster again showed a clear demarcation in the case of RNN translator as seen in Appendix C.3. The RNN-translator encodes molecules based on the length of the SMILES string (Dimension 2 of t-SNE plot), with structurally similar molecules being placed together in the region corresponding to a sequence length range. The *Graph2SMILES* and *Graph2SELFIES* translators, on the other hand, encode molecules only based on molecular structure with similar structures occupying a region of the encoder

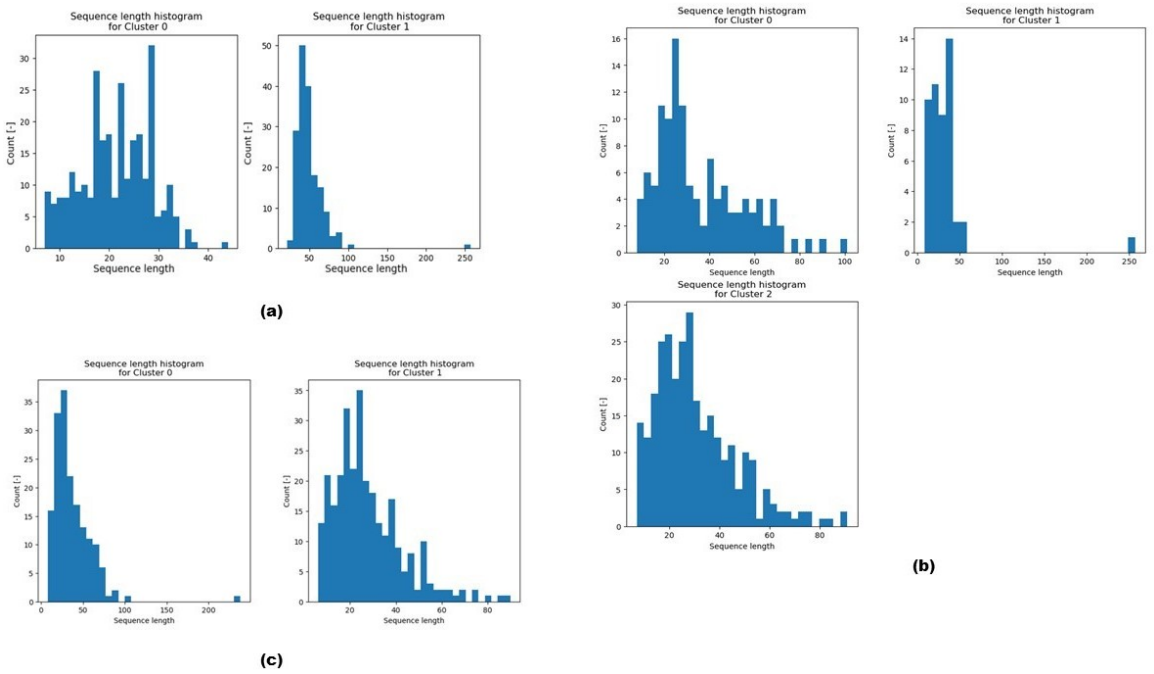


Figure 4.3: (a) Distribution of sequence length in each K means cluster of RNN-translator encodings, (b) Distribution of sequence length in each K means cluster of *Graph2SMILES* translator encodings, (c) Distribution of sequence length in each K means cluster of *Graph2SELFIES* translator encodings

space.

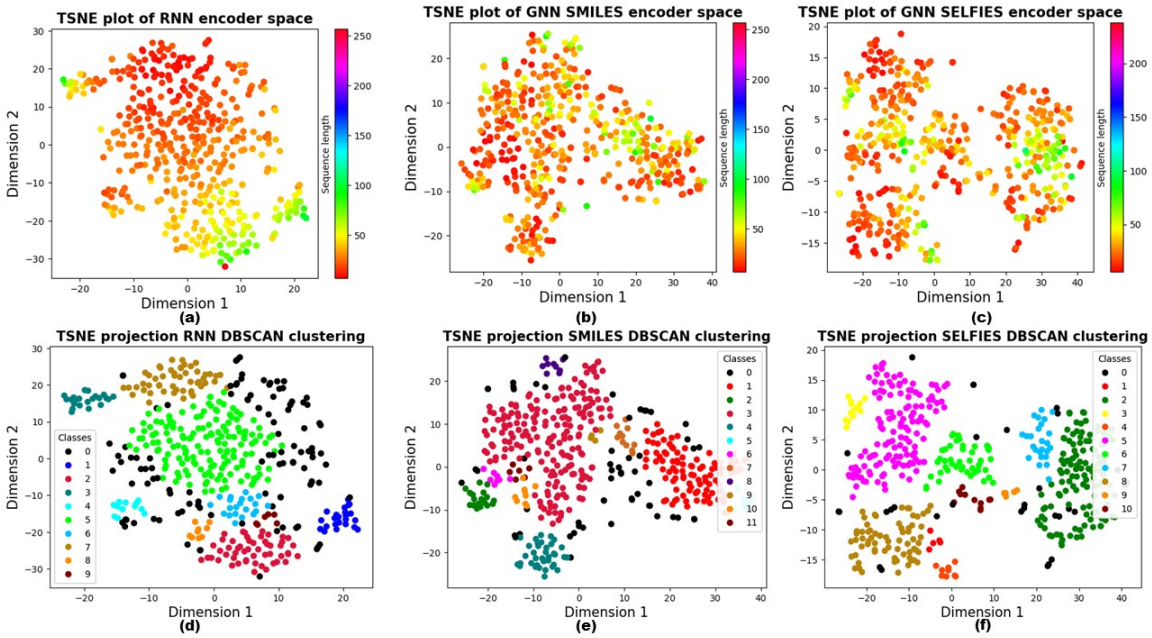


Figure 4.4: (a) DBSCAN clustering of RNN translator t-SNE encodings (b) DBSCAN clustering of *Graph2SMILES* translator t-SNE encodings (c) DBSCAN clustering of *Graph2SELFIES* translator t-SNE encodings (d) Mean intra-cluster Tanimoto similarity for RNN translator t-SNE encodings (e) Mean intra-cluster Tanimoto similarity for *Graph2SMILES* translator t-SNE encodings (f) Mean intra-cluster Tanimoto similarity for *Graph2SELFIES* translator t-SNE encodings

4.4.2 Spectrum constrained GAN

The GAN was trained until saturation of the Generator and Critic. Initial training of the GAN was performed based on the formulation provided in 4.3b with the β value set to zero, i.e., no penalty was added for the functional group identification. In the inference phase, ten thousand random vectors sampled from a normal distribution were provided as the input to the generator along with one sample spectrum to generate hundred molecules for a given spectrum. To understand the range of molecules generated for a given spectrum condition, the Tanimoto similarity between the MACCS fingerprints of the generated molecules was computed. The distribution of similarity with a set of generated molecules is shown in Figure 4.5. The histogram reveals a good distribution across molecules with an average similarity of 0.37 (Figure 4.5(a)). Analysis of the similarity between the molecule whose IR spectrum (termed as ground truth) was provided as input and the generated molecules

showcased poor match between the molecular structures as seen in Figure 4.5(b). A few examples of generated molecules with the highest similarity to the ground truth are shown in Figure 4.5(c). A mean similarity of 0.361 was found across the hundred molecules generated for each of the 256 spectra.

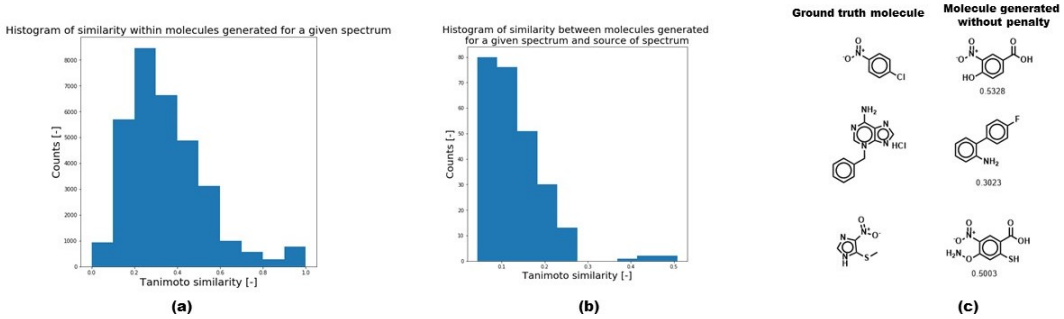


Figure 4.5: (a) Distribution of Tanimoto similarity between molecules generated for a given condition, (b) Distribution of Tanimoto similarity between generated and molecule whose spectrum was given as condition, (c) Few examples of molecules generated along with ground truth. Tanimoto similarity is given below each generated molecule.

This mismatch in the structural similarity between the ground truth and the generated molecules prompted the inclusion of the penalty term for a functional group classifier during training. The hyperparameter β was maintained at a value of 10 across all epochs. The histogram of the Tanimoto similarity for the same testing set is shown in Figure 4.6(a). Similar to the previous case a hundred molecules were generated for each spectrum condition. The histogram is skewed more towards the right, indicating a larger fraction of similar looking molecules generated for a given spectral condition with a mean Tanimoto coefficient of 0.965. Though this indicates a lack of variability in the generation scheme, molecules corresponding to given IR spectrum tend to be similar in structure thereby explaining the skew in the distribution. A test of similarity between the ground truth molecule and the generated molecules also reveals a greater degree of similarity as see in Figure 4.6(b). The mean Tanimoto coefficient across all testing samples was found to be 0.724. A few examples of molecules generated for different spectra conditions with the highest degree of similarity to ground truth along with molecules generated without functional group penalty are show in

Figure 4.6(c).

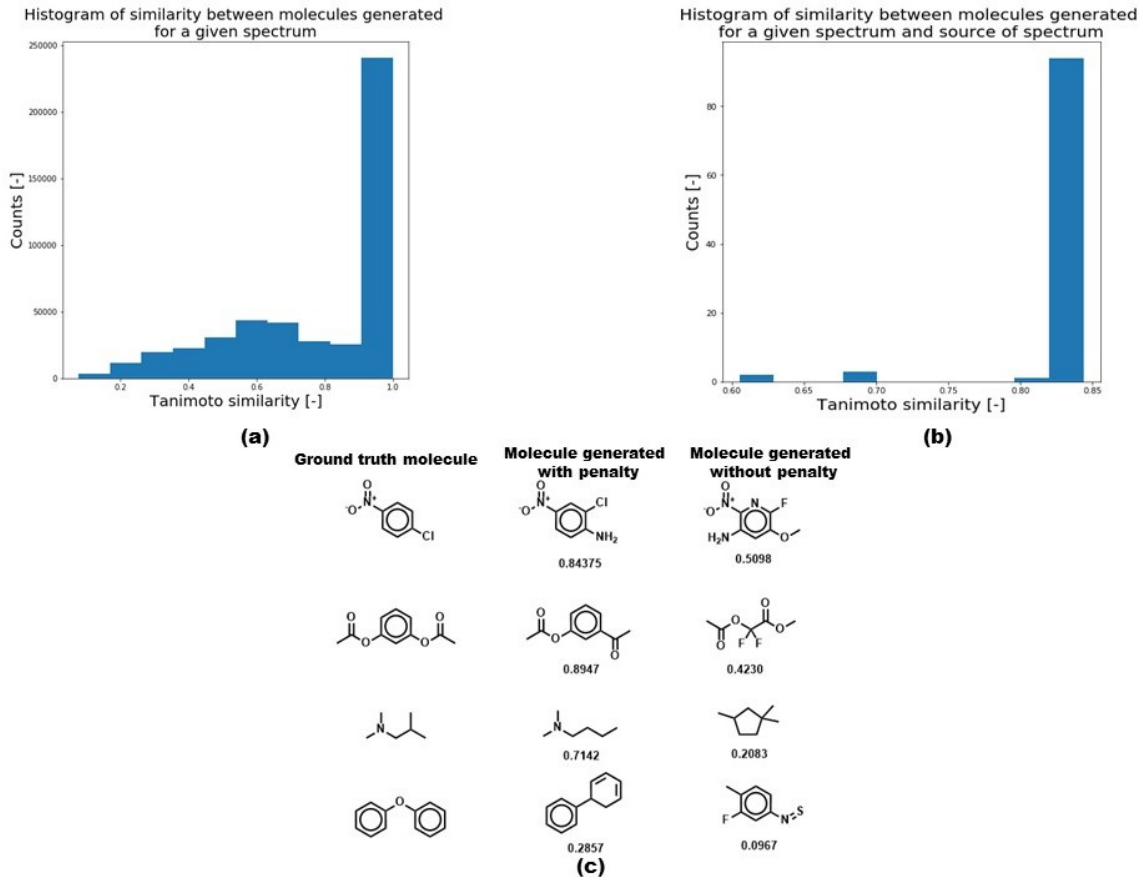


Figure 4.6: (a) Distribution of Tanimoto similarity between molecules generated for a given condition with functional group penalty, (b) Distribution of Tanimoto similarity between generated and molecule whose spectrum was given as condition with functional group penalty, (c) Few examples of molecules generated with functional group penalty, without functional group penalty and the ground truth. Tanimoto similarity is given below each generated molecule.

4.4.3 Biomass molecular generation

FTIR spectra at different operating conditions were used as inputs to the GAN and 10 molecules per spectrum were generated. Unique molecules generated for each batch operating condition are depicted in Figure 4.7. Hydroxyl groups are predominantly noted in all structures, with some generated molecules also indicating amines. The amine groups were found to be an artifact of larger weights provided to the amine classification by the func-

tional group penalizer. As amine groups are not frequently observed as products of HTL of pine biomass, molecules containing only amine groups were excluded from further study. At higher temperatures (250°C), molecular scaffolds similar to those of furan rings were generated. Furfural is commonly obtained as a product of biomass pyrolysis and is formed through dehydration of pentose sugars. To investigate the reaction routes undertaken at each

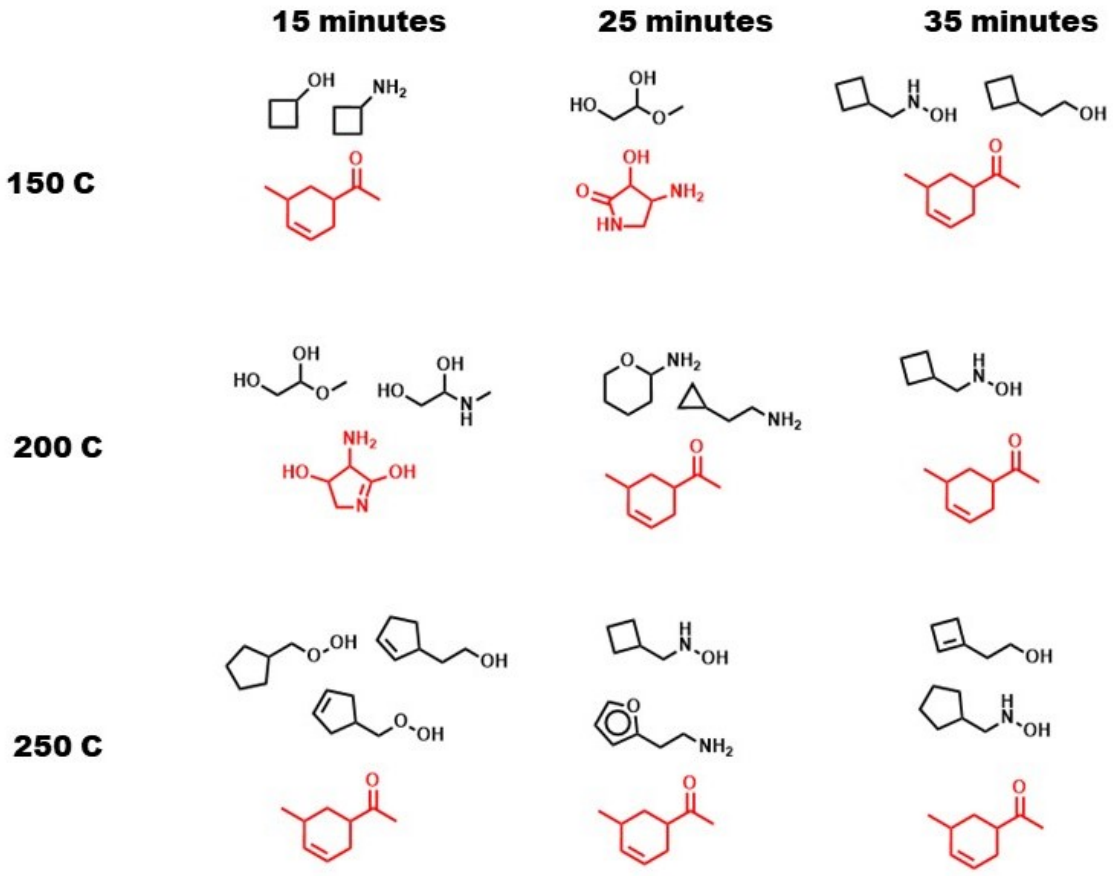


Figure 4.7: Molecules generated at different batch process conditions for HTL of biomass. Molecules shown in red have been generated without functional group penalty.

operating condition, reaction networks were generated by recursive application of reaction templates extracted automatically from literature. The reaction templates were applied in both forward and reverse directions to each generated molecule to construct a local reaction network at that process condition. Molecular structural similarity tests were performed between other generated molecules and the nodes of the reaction network to identify in-

tersections between the networks. The networks generated from each unique molecule is presented in Figure 4.8. The dashed arrows represent a retrosynthetic route obtained by application of inverted templates. Ellipses are included to indicate products obtained as result of multi-step reactions.

The networks generated from each molecule contain molecules whose structures match with the other generated molecules. The networks generated tend to focus on the dehydration of alcohol groups along with ring opening mechanisms. No obvious routes depicting the lignin pyrolysis were identified by the netowrk generation algorithm. In lignin pyrolysis, polyaromatic compounds cleave at β O-4 linkages to form phenolic compounds, side chains oxidise to form acidic and carbonyl compounds [238, 239]. Aromatic scaffolds were not generated by the generative model, hence the ligninic-phase of the pyrolysis was not captured. Carbonyl groups were identified as a part of multiple networks. Furfural-like molecules were generated by the GAN and also were identified as products in multiple reaction networks capturing the cellulosic reactions of biomass [240, 241].

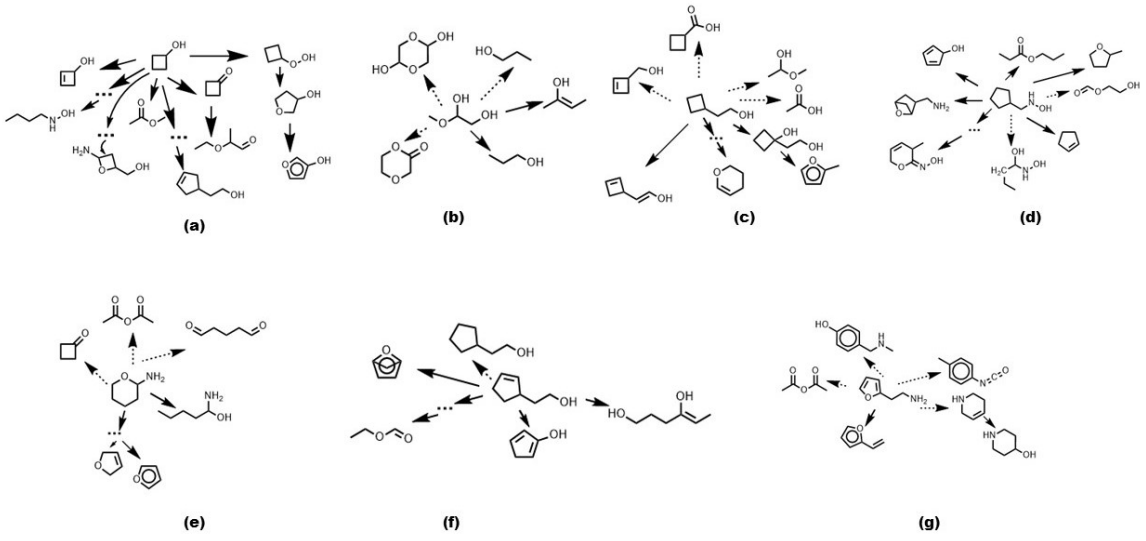


Figure 4.8: Reaction networks generated from each unique molecule generated at each process condition. (a) 150 degC and 15 minutes, (b) 150 degC and 25 minutes, (c) 150 degC and 35 minutes,(d) 200 degC and 25 minutes, (e) 250 degC and 15 minutes, (f) 250 degC and 25 minutes, (g) 250 degC and 35 minutes

4.5 Conclusion

This works presents a spectrum-constrained molecular generation methodology by training a conditional GAN to predict latent representations of molecules. The latent representations are obtained as context vectors from a molecular graph to SMILES/SELFIES generation routine. Employing graph convolutions, the molecular graph is converted to latent representation which is initialized as the hidden state for a LSTM-based decoder. The GAN loss function is modified to incorporate a functional group penalty term which penalizes generation of molecules with different functional groups than the ones indicated by the FTIR spectrum provided as a constraint. The methodology is applied on experimental measurements of HTL of biomass to infer reaction networks of the process. The graph-based translator is able to achieve a performance close to benchmark translators in the literature and is seen to populate the latent space based on structural similarity rather string length.

The dominant reactions captured by the spectroscopic information have been captured by the network generation algorithm. Breakdown of the cellulosic structures is well represented in the reaction networks. The methodology bypasses the need for resolution techniques and Bayesian structure learning algorithms, which are affected by noise, and can prove to be a substitute for techniques mentioned in Chapters 2 and 3. The method is affected by lack of scaffold information in the conditioning spectrum and thus generates networks that representative of the reactions rather than the actual molecules themselves. Nonetheless, the methodology is able to provide lumped characteristic reaction networks that describe the progression of the HTL process and can be used as a initial hypothesis generation mechanism for further detailed studies.

Chapter 5

Benchmarking chemical neural ordinary differential equations to obtain reaction network-constrained kinetic models from spectroscopic data

5.1 Abstract

Kinetic model identification relies on accurate concentration measurements and physical constraints to limit solution multiplicity. Not having these measurements and prior knowledge of species and reactions creates considerable challenges that are currently unresolved. We address these by developing a data-driven framework using realtime spectroscopic data, comprising: (i) multivariate curve resolution to deconvolve the spectra of the reacting mixture into those of its pseudocomponents and their corresponding concentrations, which enables species identification without prior information, (ii) Bayesian structure learning among the pseudocomponent spectra enables hypothesizing reaction pathways, and (iii) neural ordinary differential equations (ODE) that are physically constrained by the hypothesized reaction network and the laws of mass action and temperature dependence are trained to learn kinetic models from the temporal concentration projections of the realtime spectra. The predictive performance of the constrained neural ODEs is limited by the accuracy of spectral deconvolution in the presence of noise, and has been benchmarked against a constrained regression approach by varying signal/noise ratios in synthetic spectroscopic data

of reacting mixtures. Although the hypothesized reaction network differs from the actual reaction template, owing to noise, the network-constrained neural ODEs are seen to result in a 75.2% and 68.15% decrease in the root mean squared error (RMSE) of the concentration profile predictions as compared to the constrained regression method, when trained on time projected concentration data of the synthetic spectra generated at a signal to noise ratio of 35 and 100, respectively

5.2 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 [242]. 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 [243]. 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 use of reactors with spectroscopic sensors that provide molecular-level information of the reactive mixtures [244, 245], which is then used as a basis for developing data-driven models for species identification and the generation of plausible reaction hypotheses [110], thereby marking the contributions of systems engineering tools towards modeling reactive chemical processes [90] . Upon species identification, reaction pathways can be deduced by perceiving chemistry as a series of graph transformations in the space of all possible reactions [194], 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[246]. 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 [25, 26]. 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. Once spectroscopic data of reacting systems has been used to identify species and hypothesize pathways, the next step is to develop a kinetic model.

The kinetic model function described by ODEs, Markov processes and state space representations using the law of mass action kinetics, S-system or polynomial models [247] is characterized by a structure that is derived from the reaction pathways among the species and a set of parameters (rate constants, stoichiometric coefficients, orders). Estimating the parameters by fitting the model to experimental concentration data [248] is known as the inverse problem in chemical kinetics and could lead to multiple solutions resulting from the same reaction dynamics [66]. 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 [249]. However, in the absence of prior knowledge of the network topology, the structure is learned by virtue of kinetic parameter estimation [250] resulting in larger degrees of freedom that challenge a unique solution owing to the *fundamental dogma of chemical kinetics* [66]. Additionally, when it comes to the inverse problem, obtaining measurements of non-equilibrium temporal concentrations of the species is often challenging [251]. Also, the knowledge of physical laws such as mass action and Arrhenius temperature dependence encapsulated in a system of coupled ODEs represented by the kinetic model function 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 [252], and its sparse variant with a curated library of vector-valued ansatz functions called 'reactive sparse identification of non-linear dynamics (SINDy)' [253], 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 [254]. 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 [255]. Instead of using simplifying assumptions like quasi-steady state, neural networks have been used for thermokinetic modeling [256] by learning a functional mapping between the true kinetic model (encompassing all mechanisms and transport limitations) and the time evolution of species concentration [257]. 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 [258]. A physics-informed neural network used to model chemical kinetics [259, 260], 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 [261], and also from reaction colorimeter data for a heterogeneous oxidation process [262]. 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 [263]. 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 [264], and differ from physics-informed neural networks in that they can model irregular and incompletely 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 [265]. 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 [266]. 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.

Research initiatives in reaction monitoring of chemical systems is lacking in a data-driven framework to estimate reaction network-constrained kinetics when prior knowledge of the underlying species and reactions is unknown [141]. Even in reacting systems where the species and reactions are identified *a priori*, the time scales of reactions make it difficult for a measurement probe to obtain temporal concentrations of each reacting species, to later fit a kinetic model. There is a significant knowledge gap due to the absence of an end to end framework for identifying species, hypothesizing their reaction pathways, and developing kinetics using readily available measurement data of the reacting mixture via molecular-level spectral probes, even when directly measuring concentrations is challenging. The present work seeks to bridge this gap by addressing the following objectives in this chapter:

1. Species are identified by way of their pseudocomponent spectra and corresponding

concentrations that are obtained by deconvolving realtime Fourier transform infrared (FTIR) spectra using multivariate curve resolution algorithms, following which Bayesian structure learning among the pseudocomponent spectra is used to hypothesize reaction pathways [25, 26].

2. Kinetic models are developed using the temporal concentrations of the pseudocomponents obtained from spectral deconvolution by training chemical reaction neural ODEs that are rightly called so because they are physically constrained by the reaction network inferred from the structure of the Bayesian network, the laws of mass action and temperature dependence.

The performance of the neural ODEs in modeling the kinetics is benchmarked against :(i) a baseline model, by way of a simple feed forward neural network that is not physically constrained, and (ii) a physically constrained regression model solved via the alternating least squares (ALS) routine. The neural ODEs are assessed against their benchmarks for their ability to predict concentration profiles when trained on time resolved projections of noisy spectroscopic data, and when constrained by a hypothesized reaction network, whose structure could differ from the true reaction template.

The synthetic data generation procedure using a reaction template from database, is outlined in Section 5.3.1. The details for constrained kinetic parameter estimation are described in Section 5.3, comprising Section 5.3.2 and Section 5.3.3, that present methods used to approximate the kinetic model functions via the chemical reaction neural ODEs, and the constrained regression approach, respectively. Section 5.4 presents the results of recovering pathway hypotheses, and kinetics from the spectroscopic data. A comparison of the prediction results from both the methods, has been presented in the context of being constrained by the reaction network hypotheses deciphered from latent features of noisy spectroscopic data. Finally, Section 5.5 summarizes the findings of this chapter, and highlights its contribution to advancing data-driven online monitoring of chemical process systems.

5.3 Methods

The present work motivates a methodological framework that uses realtime spectroscopic data to decipher categories of underlying species, the plausible reaction pathways among them, followed by developing pseudokinetic models constrained by the hypothesized reaction network structure, as a way to advance the online monitoring of chemical feedstocks in the absence of prior knowledge of its underlying species, reactions or kinetics. The framework has been demonstrated on temporally generated synthetic spectra by convolving the pure component spectra of species from a known reaction template with a power law kinetic model, to generate spectra for the reacting mixture, comprising 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 previous works [25, 26, 110]. 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. Bayesian network learning is then used to identify possible reactions (which are causal relationships) between the pseudo-components. The pseudo-component spectra are 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) [25, 26]. Multiple score-search methods (Hill climbing, Tabu search and maximum-minimum hill climbing) are used to ensure that the most probable causal relationships are identified in the directed acyclic graph that encodes the proposed reaction network. 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. The adjacency matrix of a finite graph indicates the connectivity of the graph(i.e., whether pairs of vertices are adjacent or not in the graph), and in our case, identifies the reactions between various pseudo-components. A detailed description of the synthetic data generation process is outlined in Section 5.3.1.

5.3.1 Description of datasets

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 biomass where the ground truth concerning species enumeration, their reactions pathways and kinetics may not yet have 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 National Institute of Standards and Technology (NIST) database [267]. 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 for concentration of the n^{th} species (C_n), where $n \in \{1, 2, \cdots N_S\}$ and $m \in \{1, 2, \cdots N_R\}$ indicate a specific species and reaction, respectively.

$$\frac{dC_n}{dt} = \sum_{m=1}^{N_R} \mathbb{1}(Adj_{mn} = 1) K_m \prod_{n=1}^{N_S} C_n^{\mathrm{O}_n} - \sum_{m=1}^{N_R} \mathbb{1}(Adj_{mn} = -1) K_m \prod_{n=1}^{N_S} C_n^{\mathrm{O}_n} \tag{5.1}$$

The ODEs in Equation 5.1 are parametrized by the kinetic parameters *viz.* the order of the n^{th} species (O_n), and the rate constant of the m^{th} reaction pathway (K_m) that are modeled to account for their temperature dependence in accordance with the Arrhenius law:

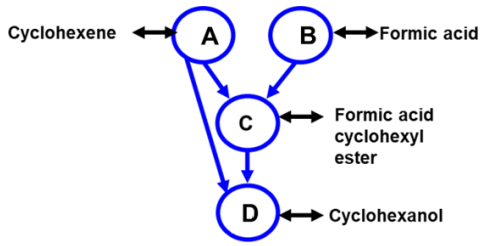
$$K_m = K_{m0} e^{\frac{-E_a}{RT}} \tag{5.2}$$

The ODEs are also constrained by the adjacency matrix ($Adj \in \text{Re}^{N_R \times N_S}$) using an indicator function ($\mathbb{1}$) as given in Equation 5.1, where N_R and N_S refer to the total number of reaction pathways and total number of species, respectively. 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.

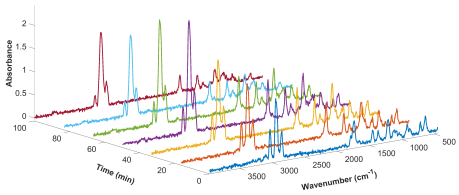
The reaction template that has been chosen for this study is shown in Figure 5.1a, and is seen to have a total of $N_S = 4$ species that are undergoing $N_R = 2$ reactions, with rate constants K_1 and K_2 as shown below

$$\begin{aligned} A + B &\xrightarrow{K_1} C \\ A + C &\xrightarrow{K_2} D \end{aligned} \tag{5.3}$$

For the above reaction template, the ODEs in Equation 5.1 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 $T \in [200°C, 400°C]$ is used to perturb the system dynamics via the rate constants, as modeled in Equation 5.2. 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* [268]. Illustrative samples of the synthetic spectra at a select few points in the time interval are shown in Figure 5.1b.



(a) Reaction network template from database, with annotations for the pure species at the nodes and directed edges indicating reactive transformation to product nodes



(b) Synthetic FTIR spectra

Figure 5.1: Synthetic FTIR spectroscopic data generated from the reaction network template for cyclohexane esterification with formic acid.

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 concentration vector $C(t) = [C_1(t), C_2(t), \cdots C_n(t)]^{\mathrm{T}}$

$$\frac{dC}{dt} = f_\theta(t, T, C_1(t), C_2(t), \cdots, C_n(t)) \tag{5.4}$$

In this chapter, the kinetic model functions of Equation 5.4 are approximated using the reaction network constrained power law model of Equation 5.1 that is solved via two methods that involve training: (i) chemical neural ODEs, and (ii) constrained regression models for constrained kinetic parameter estimation, the underpinnings of which are described in Section 5.3.2 and Section 5.3.3, going forward.

5.3.2 Chemical reaction neural ODEs

Let us consider the following reaction involving 4 species, typically represented as a chemical reaction (Equation 5.5), with reactants on the left and products on the right, prefixed by their respective stoichiometric coefficients [269]:

$$\nu_A A + \nu_B B \xrightarrow{k} \nu_C C + \nu_D D \tag{5.5}$$

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 [270].

$$r = \frac{-1}{\nu_A}\frac{dC_A}{dt} = \frac{-1}{\nu_B}\frac{dC_B}{dt} = \frac{1}{\nu_C}\frac{dC_C}{dt} = \frac{1}{\nu_D}\frac{dC_D}{dt} \tag{5.6}$$

The kinetic rate expression based on the law of mass action [271] is as follows

$$r = kC_A^a C_B^b \tag{5.7}$$

In Equation 5.7, 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 5.2, the rate expression in Equation 5.7 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.

$$r = \exp\left[\ln k_0 - \frac{E_a}{RT} + a \ln C_A + b \ln C_B\right] \tag{5.8}$$

Building on this principle, the time rate of change of concentrations for species that partake in multiple reactions indicated by the reaction template of Equation 5.3 that has been used in this work, are given as follows:

$$\begin{aligned}
\frac{dC_A}{dt} &= -\nu_A^{(1)} r_1 - \nu_A^{(2)} r_2 \\
\frac{dC_B}{dt} &= -\nu_B^{(1)} r_1 \\
\frac{dC_C}{dt} &= \nu_C^{(1)} r_1 - \nu_C^{(2)} r_2 \\
\frac{dC_D}{dt} &= \nu_D^{(2)} r_2
\end{aligned} \tag{5.9}$$

In Equation 5.9, $\nu_{\mathbb{S}}^{(\mathbb{R})}$ is the stoichiometric coefficient of the species $\mathbb{S} \in \{A, B, C, D\}$ participating in a reaction $\mathbb{R} \in \{1, 2\}$, the rates $r_{\mathbb{R}}$ of which are as follows:

$$\begin{aligned}
r_1 &= \exp\left[\ln k_0^{(1)} - \frac{E_a^{(1)}}{RT} + a^{(1)} \ln C_A + b^{(1)} \ln C_B\right] \\
r_2 &= \exp\left[\ln k_0^{(2)} - \frac{E_a^{(2)}}{RT} + a^{(2)} \ln C_A + c^{(2)} \ln C_C\right]
\end{aligned} \tag{5.10}$$

The activation energies and the pre-exponential rate constants of the reactions, are given by $E_a^{(\mathbb{R})}$ and $\ln k_0^{(\mathbb{R})}$, respectively, while the order of the species, $\mathbb{O} \in \{a, b, c, d\}$ with respect to the reactions is indicated as $\mathbb{O}^{(\mathbb{R})}$, in Equation 5.10.

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 [272]. 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) [273]. Deep neural networks comprise more hyperparameters than shallow neural networks, and thereby suffer from overfitting due to the model complexity (and the associated extra degrees of freedom); this is sought to be handled by effective regularization of the parameters, [274] and the reconciliation of domain knowledge into the network structure [253, 275]. 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 5.2. 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

$$X_t = [\ln C_1(t), \ln C_2(t), \cdots \ln C_{N_S}(t), -1/T(t)]^T \tag{5.11}$$

such that $X_t \in \mathbb{R}^{(N_S+1)\times 1}$ 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 $H_t \in \mathbb{R}^{N_R\times 1}$ 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 $\widehat{\dot{C}}_t \in \mathbb{R}^{N_S\times 1}$. 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 $\widehat{C}_t \in \mathbb{R}^{N_S\times 1}$ given by

$$\widehat{C}_t = \left[\widehat{C}_1(t), \widehat{C}_2(t), \cdots \widehat{C}_{N_S}(t)\right]^T \tag{5.12}$$

The parameters of the network denoted by θ comprise the weights of the first two layers, denoted by $W^{(1)} \in \mathrm{Re}^{N_R \times (N_S+1)}$ and $W^{(2)} \in \mathrm{Re}^{N_S \times N_R}$, and the bias associated with the first layer, denoted by $b^{(1)} \in \mathrm{Re}^{N_R \times 1}$. 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 5.7 and Equation 5.8. The weights of the network are regularized by the adjacency matrix $Adj \in \mathrm{Re}^{N_R \times N_S}$ as illustrated in the following set of equations in the forward pass of the neural ODE, where $\mathbb{1}$ is the indicator function, while **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.

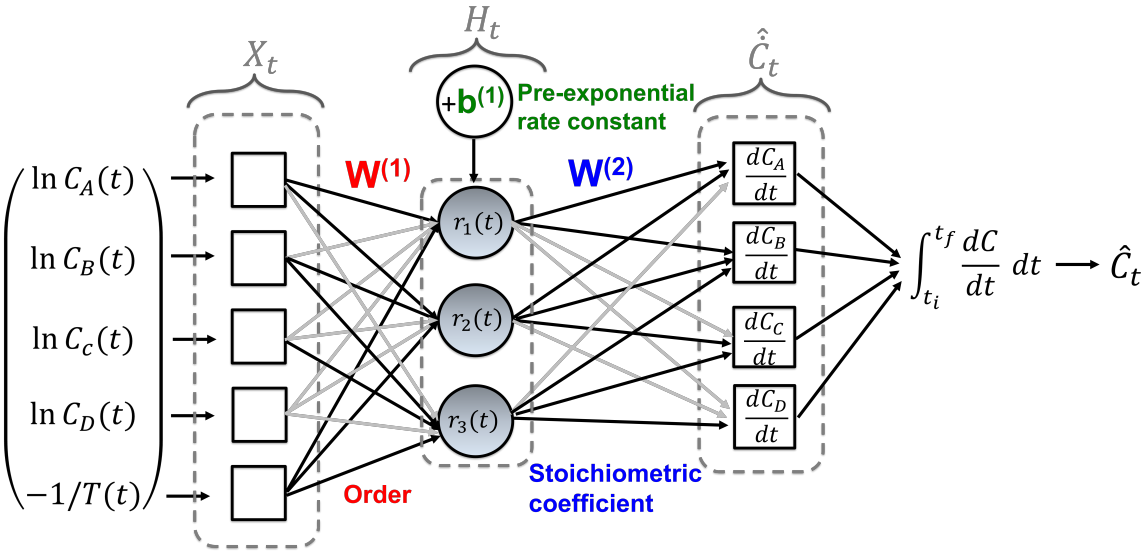


Figure 5.2: Schematic representation of the chemical reaction neural ODE

$$H_t = \exp\left[\left(W^{(1)} * \left[\mathbb{1}(Adj = -1) | \mathbf{1}^{N_R \times 1}\right]\right) X_t + b^{(1)}\right] \tag{5.13}$$

$$\dot{\widehat{C}}_t = \left(W^{(2)} * \mathbb{1}(Adj \neq 0)^T\right) H_t \tag{5.14}$$

$$\begin{aligned} \widehat{C}_t &= \widehat{C}_{t-1} + \int_{t-1}^{t} \dot{\widehat{C}}_{t-1}\, dt \\ &= \text{ODESolve}(\widehat{C}_{t-1}, \dot{\widehat{C}}_{t-1}, [t-1, t], \theta) \end{aligned} \tag{5.15}$$

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 in Equation 5.16. 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 5.13-5.15 as constrained by the adjacency matrix, are forced towards sparsity.

At this point, it is worthwhile to make a distinction between the following two kinds of modeling frameworks that arise from coupling the universal function approximation power of neural networks with differential equation modeling to account for the physics that governs the dynamics of the system being modeled: (i) neural ODEs use a neural network to parametrize the continuous dynamics of states (species concentration in this chapter) as a system of ODEs that are solved at the output by using a standalone differential equation solver to minimize the solution difference from the (in)directly measured model states (deduced via latent projections from experimentally measured spectra in this chapter), so that the neural network acts as a surrogate of the system dynamics [276], (ii)physics informed neural networks on the other hand builds neural network surrogates for ODE/PDE solutions, whereby automatic differentiation is used to compute derivatives of the neural network outputs with respect to its inputs and model parameters, the residual difference of which from the ODE/PDE solutions at fixed collocation points, is minimized during training [277].

$$L(\theta) = \sum_{t} \left(C_t - \widehat{C}_t\right)^2 + \sum_{t} \left(\dot{C}_t - \widehat{\dot{C}}_t\right)^2 + \alpha \left(W^{(1)} * \left[\mathbb{1}(Adj \neq -1) | \mathbf{0}^{N_R \times 1}\right]\right) + W^{(2)} * \mathbb{1}(Adj = 0)^T \tag{5.16}$$

This work proposes a neural ODE framework to learn a kinetic model from the concen-

tration projections of species that have been deciphered from constrained latent factorization of spectroscopic data of the reacting mixture by minimizing the loss function in Equation 5.16 that involves solving the ODE in the forward pass, and the continuous backpropagation of the gradient that requires solving the augmented ODE backwards in time [276], as it involves computing the derivatives of the ODE solution with respect to the network parameters. This has been implemented using adjoint sensitivity analysis by framing a set of auxiliary ODEs, the solution of which is evaluated to provide the aforementioned derivatives, while training the neural ODE. The PyTorch library, *torchdiffeq* [276, 278], encapsulates code for the same, and has been used to train the neural ODE presented in this work. Its performance has been compared against a baseline model comprising a simple feedforward neural network (FFN) with the species concentrations and temperature supplied as inputs, followed by a linear activation for the hidden layer with as many nodes as the number of reactions, and finally an output layer that predicts the concentration rate change of all the species ($\widehat{\dot{C}}$). The FFN does not account for i) the physical constraints when it comes to either the transformations applied to the data that is input, or the activation functions applied to the hidden layer, and the ii) temporal nature of the data samples. The FFN is trained on independent samples of species concentrations, and its predictions are temporally integrated post facto, to obtain concentration predictions $\widehat{C}$ of the species. The predictive power and the interpretability of the chemical reaction neural ODE that results from accounting for the complexities neglected herein by the FFN baseline, has been assessed to strengthen the merit of the framework.

5.3.3 Constrained regression

Additionally, we present another method to infer the kinetic parameters while honouring the topology of the reaction network. Under this approach, the parameter inference is formulated as a matrix factorization problem, by representing the rate laws as a linear combination of concentrations of the reacting species and temperature. Similar to the previous approach,

a logarithmic transform is applied to the concentration profiles to represent the rate law in a linear form as presented in Equation 5.8. The factorization of the rate law can be written as:

$$\ln\left(\frac{\mathrm{dC_i}}{\mathrm{d}t}\right) = \left[\ln C_1, \ln C_2, \cdots, \ln C_n, \frac{1}{T}\right]\left[n_1, n_2, \cdots, n_n, \frac{-E_a}{R}\right]^T + \ln k_0 \qquad (5.17)$$

This factorization is a natural product of the linearization of the rate law. The rate of transformation of an i^{th} species is written as the product of all n species in the reaction mixture along with the temperature conditions, weighted by their reaction order and the exponential terms in an Arrhenius type rate law. The natural logarithm of the pre-exponential factor is represented as an intercept in the concentration space of the n components.The determination of kinetic parameters based on this factorization amounts to determination of the second matrix in the decomposition and the intercept that best fits the experimental rate. This is a classical linear regression and it is straightforward to obtain its solution.

As discussed earlier, this decomposition does not necessarily lead to an unique solution. It therefore becomes essential to incorporate any *a priori* information available to ensure an interpretable decomposition. The network topology derived from causal inference of the component profiles provides additional constraints that can restrict the solution space. The adjacency matrix of the reaction network can be used to incorporate this relational information. Thus, computing the decomposition was set as an optimization problem aimed at minimizing the error of reconstruction of the rates with additional penalty terms added to incorporate the adjacency information. The objective function for one component is given by

$$\begin{aligned} &\min_{\theta_1,\theta_2} ||r - exp(X\theta_1 + \frac{1}{T}\theta_2 + ln(k_0)||_2^2 + \lambda_1 h(C_{recon}, C) + \lambda_2 g(Adj, \theta_1) \\ &\text{S.T.} \qquad lb \leq \theta_1, \theta_2 \leq ub \end{aligned} \qquad (5.18)$$

$X \in \mathbb{R}^{m\times n}$ represents a matrix consisting of the logarithm of the concentrations of each species along its columns for m time points. $\theta_1 \in \mathbb{R}^{n\times 1}$ represents the vector of orders

associated with the reaction. θ_2 represents the $-\frac{E_a}{R}$ parameter. The rate r of each species is obtained through numerical differentiation of the concentration of each species. f and g correspond to the penalty terms included to incorporate the error in reconstruction of the concentration and the adjacency matrix(*Adj*), respectively, which are weighted by λ_1 and λ_2.

Function g compares the values of θ_1 with the row of the adjacency matrix corresponding to the species under consideration, thus steering the decomposition towards a structure that conforms to the network architecture. Function h computes the norm of the difference between the concentrations reconstructed by solving the rate law ODEs based on the calculated parameters and the actual concentration of the species. The ODEs in this case were solved using the *odeint* function in the *SciPy* package of Python. The factorizations are obtained for each component using an Alternating Least Squares (ALS)- type approach. In the ALS approach a multi- objective optimization problem is solved by individually solving each objective function in turn and updating the inital guess for each problem based on the solution of the preceding problems. The algorithm begins at the root node of the graph and computes the kinetic parameters associated with it. It then moves in a breadth-first approach, calculating the parameters for all species at the same level in the graph before moving to the subsequent level. Parent nodes of a node N in a graph are all the nodes in the immediate higher level that connected directly to N. Similarly, children of a node N are all the nodes in the immediate lower level that are directly connected to N. It can be seen that for nodes in intermediate levels in the reaction graph, the rate law is of the form,

$$r_i = f(\text{concentration of parent of } i^{th} \text{ species}) - k_i * C_i^{n_i} \prod_j C_j^{n_j} \tag{5.19}$$

where j corresponds to the other participants in the reactions in which the i^{th} species participates. $f(C_{parents})$ represents the rate of formation of the i^{th} species dependent only on the parents of the species. Since this form of the rate law is not linearly separable in its logarithmic form, the logarithmic transform is applied to $r_i - f(C_{parents})$, therefore

estimating the parameters for the reaction in which i is the reactant. Hence, the optimization problem is transformed into:

$$\begin{aligned} &\min_{\theta_1,\theta_2} ||y - exp(X\theta_1 + \frac{1}{T}\theta_2 + ln(k_0)||_2^2 + \lambda_1 h(C_{recon}, C) + \lambda_2 g(Adj, \theta_1) \\ &\text{S.T.} \qquad lb \leq \theta_1, \theta_2 \leq ub \end{aligned} \tag{5.20}$$

where $y = r_i - f(C_{parents})$. This protocol allows for lower transmission of error in the optimization routine as the parameters in $f(C_{parents})$ have already been computed in previous steps of the ALS routine. When a species is the substrate in multiple reactions, a similar approach is followed where the parameters for each of those reaction is computed iteratively by moving the other terms in the rate law to the LHS. The workflow of the algorithm is depicted in Figure 5.3. From the flowchart choice of objective function to minimize is based on whether it is strictly one of the root nodes,i.e., has no parents or an intermediate node.

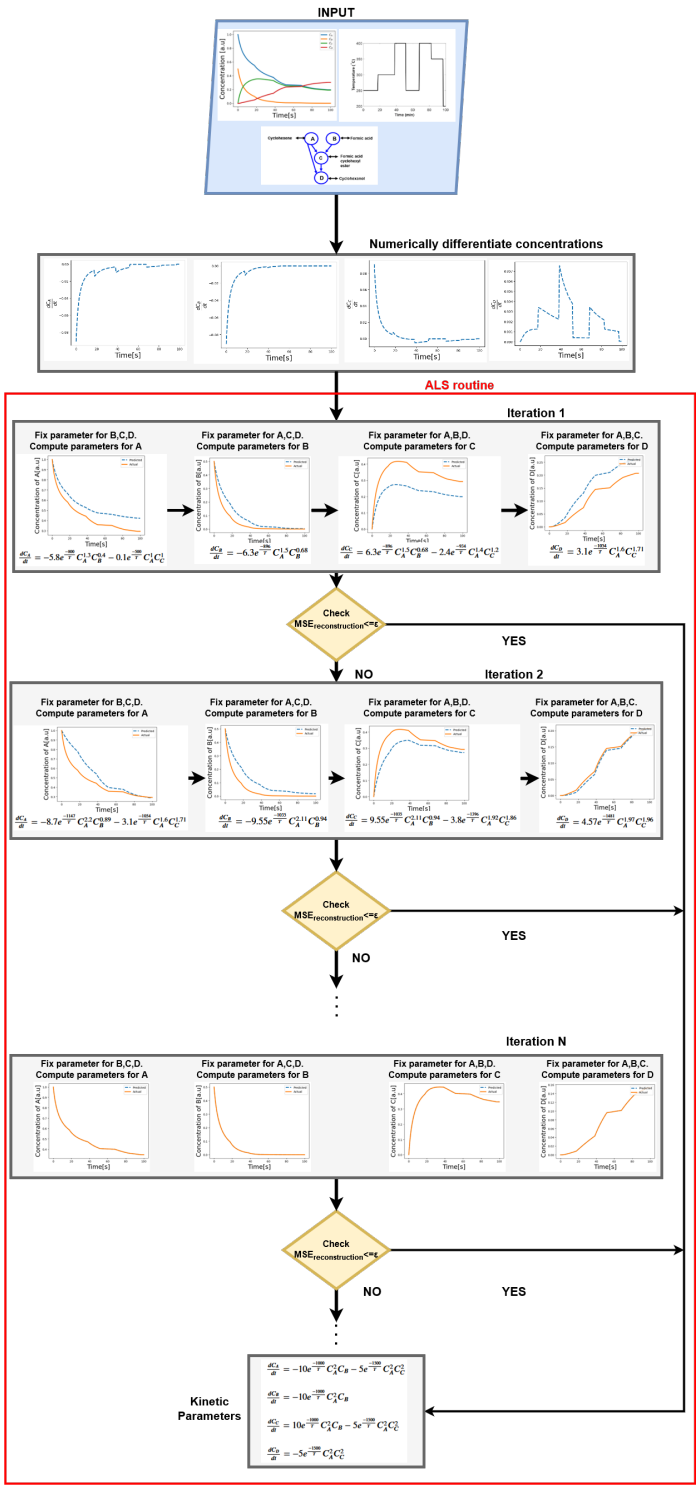


Figure 5.3: Workflow of parameter estimation using the ALS approach.

As a baseline test for the regression method, a SINDy regression is performed using the *PySINDy* [279, 280] package on Python. This technique evaluates the rate of change of a

species as a minimal linear combination of the various function types present in the feature library. This allows for one to uncover a functional form for the ODE that describes the transformation of the species without any prior knowledge on the reaction scheme. As reaction kinetics are generally polynomial ODEs, the feature library was set to be all polynomial functions up to an order of 4 ,i.e., the function library consists of terms such as C_A,C_AC_B,$C_A^2C_B$,C_CC_A,$C_BC_D^2$,etc along with a exponential functions to incorporate the temperature dependency terms. The sparsity hyper-parameter to limit model complexity was set between 0.01 and 0.005 to ensure that all ODEs have at least one functional form associated with them.

5.4 Results and Discussion

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 [281]. Temporal concentration profiles are obtained by solving the system of ODEs for the reaction template, as outlined in Section 5.3.1. A multi-level pseudorandom temperature signal as shown in Figure 5.4a was used to perturb the the kinetic model of Equation 5.1. In this study, temporally generated synthetic spectra for the reacting mixture is posed to be a signal of the pure component spectra of the species (Figure 5.4b) combined in proportion to the species concentrations over time, and is representative of spectra recorded for a reacting system in realtime, based on Beer's law. The ability to perfectly deconvolve the mixture spectra into those of its pure components through curve resolution techniques (Section 5.3), is limited by the noise in the mixture spectra. At the outset, neural ODE predictions and that from the constrained regression are tested on the temporal concentration data of the species obtained from perfect deconvolution of the mixture spectra, which in the noise-free case corresponds to the data obtained from solving the ODEs of Section 5.3.1. The kinetic mechanism is seen to comprise 4 species undergoing 2 reactions, as indicated by Equation 5.3. Random initial concentration values

were used for the reacting species (A and B) to obtain concentration profiles using Equation 5.1 that 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 both the methods are compared against the profiles of the temporal concentrations recovered from solving the ODEs, as shown in Figure 5.4c. The root mean squared error (RMSE) between the data and predicted concentrations is used a metric for comparison, and is calculated as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{T}(C_t - \widehat{C}_t)^2}{T}} \tag{5.21}$$

It can clearly be seen that constraints on the neural networks' structure and parameters prevent it from overfitting the data, while the regression method is able to more closely predict the concentration profiles, registering a 52.17% decrease in the RMSE over the chemical neural ODE (Table 5.1). Hence, in the future when the model is trained on synthetically generated noisy data, the chemical reaction neural ODE is expected to run a low risk of fitting the noise. Both the models are shown to perform better than their respective baselines as shown in Figure D.4a of Appendix D.3. The chemical neural ODE and the constrained regression ALS routine are seen to register an 84.11% and 94.23% decrease in the RMSE, respectively (Table D.1).

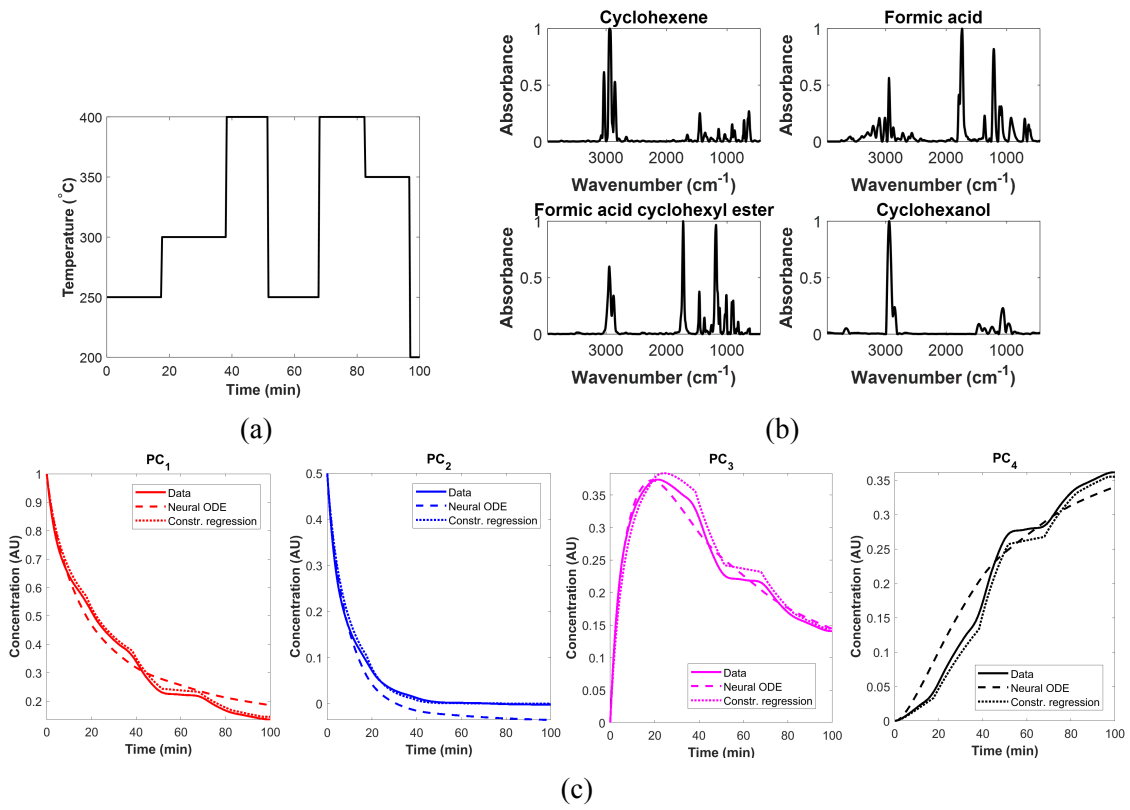


Figure 5.4: (a) Multi-level pseudo random temperature signal, (b) Pure component spectra from the database, (c) Predictions of the chemical reaction neural ODE and the constrained regression model compared against the temporal concentration data obtained by solving a known ODE system for kinetics.

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, is investigated.

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 5.3.1, before it undergoes 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 D.1

of the D.1. The time resolved species concentrations in Figure D.1a are used to train kinetic models that are constrained by the hypothesized reaction network among the species, presented in Figure D.1b, and has been inferred by Bayesian structure learning among the pseudocomponent spectra in Figure D.1c obtained by deconvolving the reacting mixture spectra. The similarity of the recovered pseudo-component spectra (Figure D.1c) with the pure component spectra (Figure 5.4b) 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 5.1a), respectively. Consequently, the causally inferred reaction network among the pseudo-components spectra (Figure D.1b), when compared with the reaction template structure (Figure 5.1a), 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 D.1b with the highest arc strength points to the conversion of compound A to compound C. The directed arc strength between two nodes in a Bayesian network is the mutual information between the nodes conditioned on the joint distribution of all the other parent nodes [25]. In the event that peaks in PC_2, corresponding to compound C are confounded with PC_4, which corresponds 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:

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

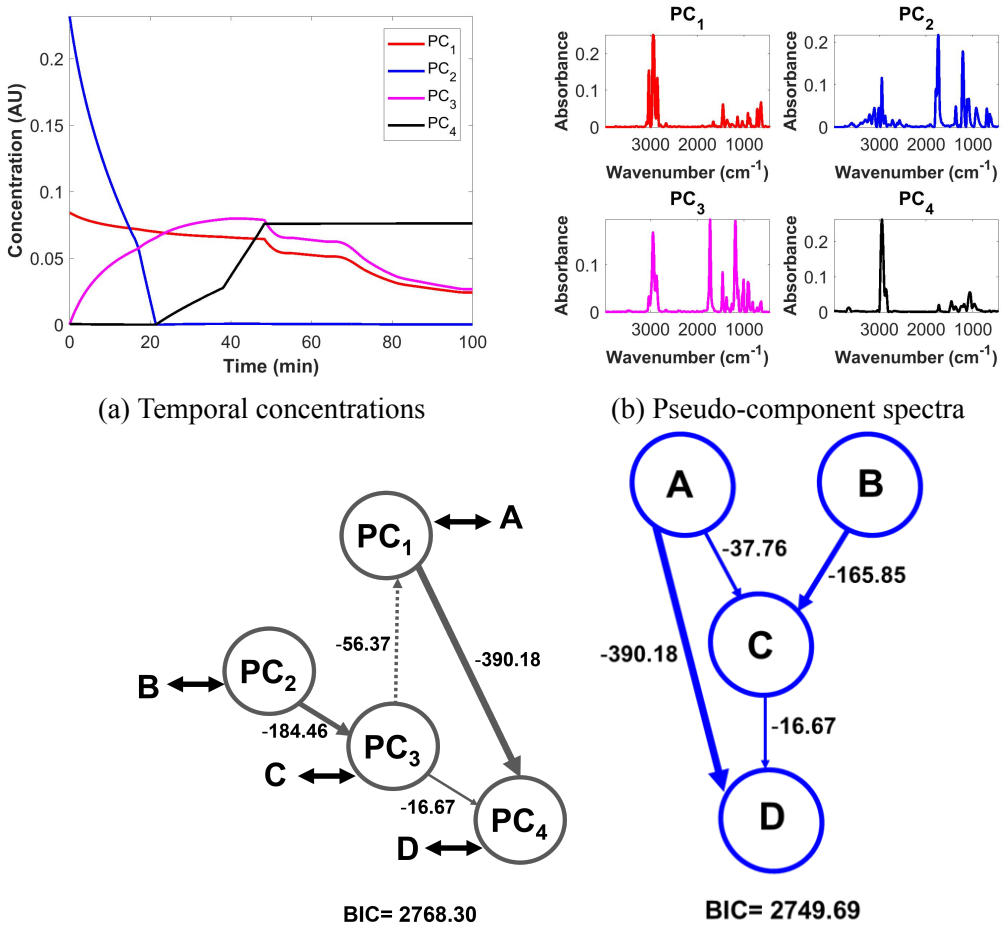
The predictions of the chemical neural ODE with the above adjacency constraints are compared against the reconstructed data from integration of the smoothed time derivative of the noisy concentration profiles from spectral deconvolution as given in Figure D.2. The neural

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.

However, the matrix factorization technique fails to recover meaningful parameters in this case as seen in Figure D.2. It can be seen that the concentration profiles from the spectral deconvolution in Figure D.1a is not congruent with the profiles expected from the Bayesian network in Figure D.1b. The network indicates that PC_3 is only a substrate in the reaction scheme and is never generated, which implies that its concentration keeps decreasing through the course of time. However, the concentration profile for PC_3 shows an increase and a decrease. This mismatch between the reaction network and the concentration profiles is not handled well by the constrained regression method. The ALS routine in tandem with the penalty parameters imposes a stricter constraint on the optimization to follow the adjacency matrix and hence fails to recover the kinetic parameters or even the trends for PC_3 and PC_4, owing to which the neural ODE method is found to reduce the RMSE by 75.2% as compared to the ALS regression (Table 5.1). Although the neural ODE out performs its baseline, the ALS routine is seen to fall short of its baseline SINDy model that better captures the noisy oscillations in the concentrations due to the inclusion of sinusoidal terms in the basis function library, which are not physically relevant to the present reaction scheme. The results are included in Figure D.4b and Figure D.1 of D.3.

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 5.5a) and pseudo-component spectra (Figure 5.5b) 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 5.4b). 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 5.5c. The skeleton of the inferred network structure is exactly the same as the reaction template (Figure 5.1a), 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 [282]. In Bayesian structure learning, the BIC is the log likelihood, computed as a difference between the mutual information and entropy for a given graph structure among the nodes, penalized by the number of directed edges constructed [25]. The solution multiplicity in the space of plausible graph structures is verified by computing the arc strengths and the BIC score shown in Figure 5.5d, given the directed edges among the compounds nodes from the reaction template network structure of Figure 5.1a. The arc strengths and the BIC score, given the network structure from the template in Figure 5.5d, are found to be comparable to those when the unknown network structure is inferred by heuristic score-search algorithms as shown in Figure 5.5c. 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.



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

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 D.3 of the D.2. 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 5.5b, 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 fac-

tor 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 those of cyclohexene and its derivatives (compounds A and D), as seen in Figure D.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 D.3b concur with those, given the reaction template structure, outlined in Figure D.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 5.5c), 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 are compared against the reconstructed data from integration of the smoothed time derivative of the temporal concentration projections from spectral resolution (Figure 5.5a), as shown in Figure 5.6. 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 differs slightly from the original reaction template.

Table 5.1: Comparison of chemical neural ODEs to the constrained regression framework.

	Root mean squared error (RMSE)		
Type of data	Neural ODE	Constr. regression	% Improvement over constr. regression
Without noise	0.0276	0.0132	-52.17%
SNR= 35	0.0278	0.1121	75.20%
SNR= 100	0.01	0.0314	68.15%

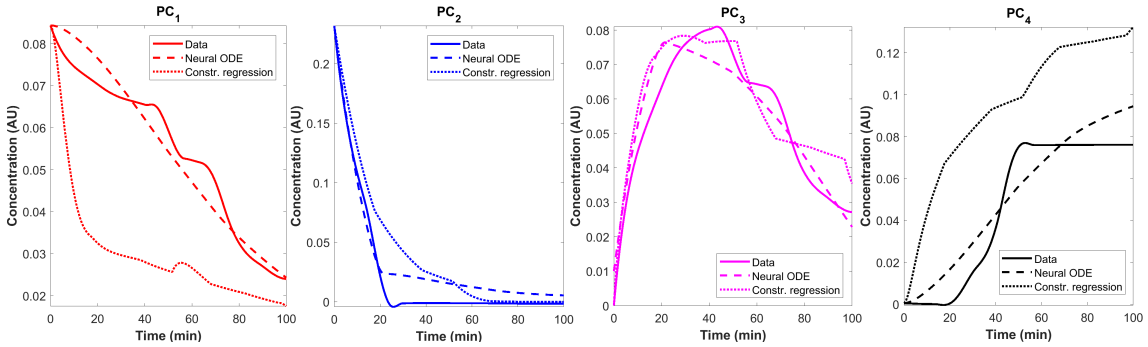


Figure 5.6: Comparison of the predictions from the chemical neural ODE and constrained regression against the reconstructed data from integration of the smoothed time derivative of temporal concentration obtained by the deconvolution of synthetic spectroscopic data, at a signal to noise ratio of 100.

Since the higher SNR results in a network structure that conforms better to the reaction profiles, the constrained regression approach is able to capture the trends in concentration profile as seen in Figure 5.6. The most deviation in prediction is seen for PC_1 and PC_4. PC_1 is the point of mismatch between the actual and reconstructed reaction networks as shown in Figure 5.5c. Thus, a larger difference is noticed in the case of PC_1. The mismatch in PC_1 is compensated in PC_4, which is a direct descendent of PC_1. Even in such a case, the method still recovers the general trend in the profile. The variations of the predictions between the constrained regression and the neural ODE, can therefore be attributed to the lack of robustness of the former in reconciling the resolved temporal concentration profiles with the reaction network structure inferred from noisy spectra of the reacting mixture, which is bound to deviate from the actual structure of the ground truth reaction template.

The neural ODE is seen to achieve a 68.15% decrease in the RMSE as compared to

the constrained regression method (Table 5.1), when it comes to quantifying the prediction performances in Figure 5.6. Both the models are seen to outperform their baselines as shown in Figure D.4c and Figure D.1 of D.3, but it is interesting to note that in the absence of noise and low noise (synthetic data with a SNR of 100), the RMSE decrease by the neural ODE is $> 80\%$ as compared to its baseline concentration predictions. However, in the presence of higher noise (synthetic data with SNR of 35), the neural ODE outperforms its baseline only by $\sim 35\%$, alluding to its ability to not fit noise owing to the incorporation of physical constraints.

An important point to note is that the results provided do not depend significantly on hyperparameter tuning, which means that the algorithm is robust. In general, the number of neurons per layer and the number of layers in a neural ODE constitute its hyperparameters, while the values assumed by the weights and biases of such a network constitute its parameters. The loss regularizer weights are not parameters of the neural ODE, but are used to tune the target for the neural ODEs to predict. The hyperparameters of the neural ODE are guided by physics, in terms of the number of neurons being equal to the number of species in the input and output layers, while the number of neurons in the single hidden layer correspond to the number of reaction paths, meaning that they do not need to be tuned. The initial values of the weights and biases come from Xavier initialization, and during the process of training, the layer weights are regulated by the reaction adjacency matrix. The choice of λ_1 and λ_2, i.e., the loss regularizer weights, are fixed a priori, and are not treated as parameters of the neural ODE. Therefore, the exercise of parameter selection distils down to trying out different initialization schemes other than Xavier to train the neural ODEs, and comparing the results of the same to choose an optimal parameter combination. Even then, it is not so much a parameter combination exercise, as it is an exercise in trying to check whether the solution converges to the same predictions independent of the choice of parameter initializations, which has been done.

5.5 Conclusions

We have presented a chemically constrained neural ODE and constrained regression method 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 and also regularize the regression method. Both the methods have been 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 models that learns the system kinetics, (ii) limit the tendency of the models 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. The constrained regression method seems to be heavily dependent on the veracity of the computed network architecture. The method is challenged to recover useful kinetic information from the data, when presented with a highly incongruous adjacency matrix and concentration profiles, resulting from noisy reacting mixture spectra. Nevertheless, the method is successful in recovering concentration profiles in the case of minor mismatch between the trends in the species concentration profiles and the reaction network by way of the adjacency matrix. Despite the above limitations, when chemical reaction neural ODEs are used for kinetic modeling, the framework presented in this work is shown to have the potential to reliably develop an end-to-end modeling framework for

species, reaction pathway and kinetic model identification of reactive systems without reliance on prior knowledge, purely by using spectroscopic data, even in the presence of noise. The predictions of the neural ODE are seen to achieve a 75.2% and 68.15% decrease in the RMSE as compared to the constrained ALS regression, when the reacting mixture spectra is contaminated with noise at SNR thresholds of 35 and 100, respectively. Also, when benchmarked against the feedforward neural network as its baseline, the neural ODE is seen to outperform by $> 80\%$ and $\sim 35\%$ for cases of low/no noise, and high noise, respectively. This indicates its ability to refrain from fitting noise and can be attributed to the incorporation of physical constraints by way of the hypothesized reaction network structure, and the laws of mass action and temperature dependence.

The framework demonstrated in this chapter has the potential to advance the online monitoring of reacting mixtures when prior knowledge of the underlying species, reaction pathways and kinetics is lacking. The use of realtime spectroscopic measurements that hold molecular-level insights of an obscure reacting system, supplemented by physical constraints for spectral deconvolution, reaction pathway inference and kinetic modeling, has been shown to provide a strong basis for data-driven species identification, generation of reaction hypotheses and pseudokinetics, respectively. Future work seeks to extend this framework to complex hydrocarbon systems such as biomass.

Chapter 6

Conclusions & Future Work

Automated reaction network discovery has been a topic of interest in many fields such as pharmaceuticals, oil and gas and waste recovery and the networks have been built based on prior knowledge of the process and its components in most cases. Automatically identifying models for reaction systems that reconcile with spectroscopic sensor measurements is a formidable task riddled with challenges arising at algorithmic and instrumental levels. With heavy reliance on human domain expertise, reaction networks developed based on model compounds tend to be suited for a specific purpose and cannot be easily modified without human intervention. Machine learning-based approaches, on the other hand, enable extraction of non-linear features from process data thereby providing significant insight from spectroscopic measurements but come at the cost of being non-interpretable. Furthermore, partial information provided by spectroscopic measurements cannot be directly adapted to methods available in literature. The aim of thesis was to explore the automation of chemical reaction network discovery and the effects of incorporating network constraints in to kinetic modeling of complex reaction systems directly from spectroscopic sensor measurements.

6.1 Summary

Chapter 2 aims to explore the use of spectroscopic sensor measurements of a complex transformation process such as HTL of biomass to build a reaction network with heavy reliance on human expertise in inferring information from spectra and mapping network topology to

real-life chemistry. Self-Modeling Multivariate Curve Resolution was used to obtain FTIR spectra of pseudo-components. The structure of the reaction network was learnt through Bayesian structure learning algorithms applied to the deconvolved spectra. Expert knowledge was applied to infer functional groups from spectral peaks and map network structure to real-life chemistry. In Chapter 3, higher-order constrained tensorial factorization was employed to jointly deconvolute FTIR and ^{1}H-NMR data. Latent factorization projected along wavenumber mode was inferred as pseudo-component spectra. Convolution operations on FTIR data was used to automatically infer functional groups and molecular fingerprinting techniques were used to identify candidate molecules from the literature. Network structure constraints from Bayesian structure learning and reaction templates derived from literature were jointly employed in generating reaction networks for both synthetic and real-life process data. Minimal domain knowledge was provided to make the solution space more compact and the methodology was able to recover lumped reaction networks for the HTL of biomass. The effect of process noise in both network structure learning and spectra deconvolution was found to be detrimental and hence a one-shot molecular generation methodology to bypass them was developed in Chapter 4. Spectrum-constrained molecular generation was introduced by employing adversarial training of neural networks to model the distribution of molecular inputs. Low dimensional representation of molecular structures were obtained using a hetero-autoencoder built using MPNNs. Evidence indicated that graph convolutional encoders were capable to generating latent representations focused on molecular structure in contrast to string-based encodings which were found to be reliant on length of the string representation. Candidate molecules representing a mixture spectra at an operating condition were generated and reaction templates were applied to develop chemical reaction networks. Projections of the tensorial data along the temporal mode were used to identify kinetic parameters for a synthetic system in Chapter 5. The adjacency matrix from the reaction network identified was utilized as a constraint on a neural ODE network structured based on the law of mass action and Arrhenius law and trained using

backpropagation. The chemical neural ODE was compared against an ALS-type optimization scheme constrained on the adjacency information at the effectiveness each protocol in inferring kinetic parameters was studied.

6.2 Future work

The work presented in this thesis aims to bridge the gap between expert developed and purely data driven models for reaction systems identified from sensor measurements. Some avenues for further research in this area are presented below.

(i) FT-IR spectroscopy used in this thesis provides rich information on functional groups present in the reaction mixture and can identify reaction centres, but does not enable identification of scaffold structures of molecules especially when peaks overlap. Inclusion of other types of spectroscopic techniques such as a mass spectrometry or ^{13}C-NMR, can provide information regarding scaffold structures of the molecules to further remove human intervention required in reaction network prediction.

(ii) The reaction network generated in this thesis are hypotheses or the true reaction network for the system. While attempts have been made to validate them to literature, discovery of the true reaction network for the system requires fine tuning of the hypotheses through first principles based quantum mechanical calculations. Molecular dynamics simulations of the hypothesised reactions can prove to be a means of validation and can incorporate solvent effects and thermodynamic constraints on feasibility.

(iii) Models developed in this thesis provide both chemical and numerical perspective of the reaction system and are well suited for product optimization and reactor control. Reaction networks can be used in tandem with kinetic models to obtain estimates for properties of interest such as boiling point or viscosity through machine learning predictions which allows for quantitative optimization of process parameters and operating conditions.

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, pp. 329–348, 1996, issn: 0959-1524. doi: 10.1016/0959-1524(96)00009-1. [Online]. Available: http://dx.doi.org/10.1016/0959-1524(96)00009-1.

[2] P. Roy, A. K. Mohanty, P. Dick, and M. Misra, "A review on the challenges and choices for food waste valorization: Environmental and economic impacts," *ACS Environmental Au*, vol. 3, pp. 58–75, 2023, issn: 2694-2518. doi: 10.1021/ACSENVIRONAU.2C00050/. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acsenvironau.2c00050http : / / dx . doi . org / 10 . 1021 / ACSENVIRONAU . 2C00050 / ASSET / IMAGES/LARGE/VG2C00050_0005.JPEG.

[3] N. Tsolakis and J. S. Srai, "Mapping supply dynamics in renewable feedstock enabled industries: A systems theory perspective on 'green' pharmaceuticals," *Operations Management Research*, vol. 11, pp. 83–104, 2018, issn: 1936-9743. doi: 10 . 1007 / S12063 - 018 - 0134 - Y / TABLES / 2. [Online]. Available: https : / / link . springer . com / article / 10 . 1007 / s12063 - 018 - 0134 - yhttp : / / dx . doi . org / 10 . 1007 / S12063-018-0134-Y/TABLES/2.

[4] A. S. Y. Wong and W. T. S. Huck, "Grip on complexity in chemical reaction networks," *Beilstein Journal of Organic Chemistry*, vol. 13, p. 1486, 2017, issn: 1860-5397. doi: 10.3762/BJOC.13.147. [Online]. Available: http://dx.doi.org/10.3762/BJOC.13.147https://www.ncbi.nlm.nih.gov/pubmed/28845192.

[5] Y. Z. H. W. C. W. Y. F. Q. Guo, "Theoretical study on the pyrolysis process of lignin dimer model compounds," *Acta. Chim. Sinica.*, vol. 9, pp. 893–900, 2009.

[6] C. Yang *et al.*, "Hydrothermal liquefaction and gasification of biomass and model compounds: A review," *Green Chemistry*, vol. 22, pp. 8210–8232, 23 2020, issn: 1463-9270. doi: 10.1039/D0GC02802A. [Online]. Available: http://dx.doi.org/10.1039/D0GC02802A.

[7] X. Yang *et al.*, "Determination of 10-hda in royal jelly by atr-ftmir and nir spectral combining with data fusion strategy," *Optik*, vol. 203, 2020, issn: 0030-4026. doi: 10.1016/j.ijleo.2019.164052. [Online]. Available: http://dx.doi.org/10.1016/j.ijleo.2019.164052.

[8] A. L. Pomerantsev and O. Y. Rodionova, "Process analytical technology: A critical view of the chemometricians," *Journal of Chemometrics*, vol. 26, pp. 299–310, 2012, issn: 1099-128X. doi: 10.1002/CEM.2445. [Online]. Available: http://dx.doi.org/10.1002/CEM.2445.

[9] E. Skibsted and S. B. Engelsen, "Spectroscopy for process analytical technology (pat)," in J. Lindon, G. Tranter, and D. Koppenaal, Eds., 2. Academic Press, 2010, vol. 3, pp. 2651–2661, isbn: 9780122266805.

[10] H. Liu, Y. Chen, C. Shi, X. Yang, and D. Han, "Ft-ir and raman spectroscopy data fusion with chemometrics for simultaneous determination of chemical quality indices of edible oils during thermal oxidation," *LWT*, vol. 119, 2020, issn: 0023-6438. doi: 10.1016/j.lwt.2019.108906. [Online]. Available: http://dx.doi.org/10.1016/j.lwt.2019.108906.

[11] Y. Li, Y. Huang, J. Xia, Y. Xiong, and S. Min, "Quantitative analysis of honey adulteration by spectrum analysis combined with several high-level data fusion strategies," *Vibrational Spectroscopy*, vol. 108, p. 103 060, 2020, issn: 0924-2031. doi: 10.1016/j.vibspec.2020.103060. [Online]. Available: http://dx.doi.org/10.1016/j.vibspec.2020.103060.

[12] S. J. Mazivila and J. L. Santos, "A review on multivariate curve resolution applied to spectroscopic and chromatographic data acquired during the real-time monitoring of evolving multi-component processes: From process analytical chemistry (pac) to process analytical technology (pat)," *TrAC Trends in Analytical Chemistry*, vol. 157, p. 116 698, Dec. 2022, issn: 0165-9936. doi: 10.1016/J.TRAC.2022.116698.

[13] N. D. Lourenço, J. A. Lopes, C. F. Almeida, M. C. Sarraguça, and H. M. Pinheiro, "Bioreactor monitoring with spectroscopy and chemometrics: A review," *Analytical and Bioanalytical Chemistry*, vol. 404, pp. 1211–1237, 2012, issn: 1618-2642. doi: 10.1007/S00216-012-6073-9/TABLES/6. [Online]. Available: https://link.springer.com/article/10.1007/s00216-012-6073-9http://dx.doi.org/10.1007/S00216-012-6073-9/TABLES/6https://www.ncbi.nlm.nih.gov/pubmed/22644146.

[14] N. Yang, C. Guerin, N. Kokanyan, and P. Perré, "Raman spectroscopy applied to online monitoring of a bioreactor: Tackling the limit of detection," *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 304, p. 123 343, 2024, issn: 1386-1425. doi: 10.1016/J.SAA.2023.123343. [Online]. Available: http://dx.doi.org/10.1016/J.SAA.2023.123343https://www.ncbi.nlm.nih.gov/pubmed/37690399.

[15] A. Brächer, L. M. Kreußer, S. Qamar, A. Seidel-Morgenstern, and E. von Harbou, "Application of quantitative inline nmr spectroscopy for investigation of a fixed-bed chromatographic reactor process," *Chemical Engineering Journal*, vol. 336, pp. 518–530, 2018, issn: 1385-8947. doi: 10.1016/j.cej.2017.12.004. [Online]. Available: http://dx.doi.org/10.1016/j.cej.2017.12.004.

[16] A. de Juan, S. Mas, M. Maeder, and R. Tauler, "A perspective on modeling evolution," *Journal of Chemometrics*, vol. 34, 2020, issn: 1099-128X. doi: 10.1002/CEM.3205. [Online]. Available: http://dx.doi.org/10.1002/CEM.3205.

[17] R Tauler, "Application of non-linear optimization methods to the estimation of multivariate curve resolution solutions and of their feasible band boundaries in the investigation of two chemical and environmental simulated data sets," *Analytica Chimica Acta*, vol. 595, pp. 289–298, 2007, issn: 0003-2670. doi: 10.1016/J.ACA.2006.12.043. [Online]. Available: http://dx.doi.org/10.1016/J.ACA.2006.12.043.

[18] R. Tauler, "Multivariate curve resolution applied to second order data," *Chemometrics and Intelligent Laboratory Systems*, vol. 30, pp. 133–146, 1995, issn: 0169-7439. doi: 10.1016/0169-7439(95)00047-X. [Online]. Available: http://dx.doi.org/10.1016/0169-7439(95)00047-X.

[19] M. Maeder, "Evolving factor analysis for the resolution of overlapping chromatographic peaks," *Anal. Chem*, vol. 59, pp. 527–530, 1987.

[20] R. Bro, "Parafac. tutorial and applications," *Chemometrics and Intelligent Laboratory Systems*, vol. 38, pp. 149–171, 1997, issn: 0169-7439. doi: 10.1016/S0169-7439(97)00032-4. [Online]. Available: http://dx.doi.org/10.1016/S0169-7439(97)00032-4.

[21] E. R. Malinowski, "Determination of the number of factors and the experimental error in a data matrix," *Analytical Chemistry*, vol. 49, pp. 612–617, 1977, issn: 1520-6882. doi: 10.1021/ac50012a027. [Online]. Available: http://dx.doi.org/10.1021/ac50012a027.

[22] R. Bro and H. A. L. Kiers, "A new efficient method for determining the number of components in parafac models," *Journal of Chemometrics*, vol. 17, pp. 274–286, 2003, issn: 1099-128X. doi: 10.1002/CEM.801. [Online]. Available: http://dx.doi.org/10.1002/CEM.801.

[23] Z. J. Lu, Q. Xiang, and L. Xu, "An application case study on multi-sensor data fusion system for intelligent process monitoring," *Procedia CIRP*, vol. 17, pp. 721–725, 2014, issn: 2212-8271. doi: 10.1016/j.procir.2014.01.122. [Online]. Available: http://dx.doi.org/10.1016/j.procir.2014.01.122.

[24] F. Sattari, D. Tefera, K. Sivaramakrishnan, S. H. Mushrif, and V. Prasad, "Chemoinformatic investigation of the chemistry of cellulose and lignin derivatives in hydrous pyrolysis," *Industrial and Engineering Chemistry Research*, vol. 59, pp. 11 582–11 595, 25 2020, issn: 1520-5045. doi: 10.1021/acs.iecr.0c01592. [Online]. Available: http://dx.doi.org/10.1021/acs.iecr.0c01592.

[25] A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. de Klerk, and V. Prasad, "Data fusion by joint non-negative matrix factorization for hypothesizing pseudo-chemistry using bayesian networks," *Reaction Chemistry and Engineering*, vol. 5, pp. 1719–1737, 9 2020, issn: 2058-9883. doi: 10.1039/d0re00147c. [Online]. Available: http://dx.doi.org/10.1039/d0re00147c.

[26] A. Puliyanda, K. Sivaramakrishnan, Z. Li, A. D. 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, pp. 5747–5762, 12 2021, issn: 1549-960X. doi: 10.1021/ACS.JCIM.1C00789/. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c00789http://dx.doi.org/10.1021/ACS.JCIM.1C00789/ASSET/IMAGES/LARGE/CI1C00789_0017.JPEGhttps://www.ncbi.nlm.nih.gov/pubmed/34813321.

[27] C. W. Coley, *Defining and exploring chemical spaces*, Feb. 2021. doi: 10.1016/j.trechm.2020.11.004.

[28] S. Rangarajan, A. Bhan, and P. Daoutidis, "Language-oriented rule-based reaction network generation and analysis: Description of ring," *Computers and Chemical Engineering*, vol. 45, pp. 114–123, 2012, issn: 0098-1354. doi: 10.1016/j.compchemeng.2012.06.008. [Online]. Available: http://dx.doi.org/10.1016/j.compchemeng.2012.06.008.

[29] C. W. Gao, J. W. Allen, W. H. Green, and R. H. West, "Reaction mechanism generator: Automatic construction of chemical kinetic mechanisms," *Computer Physics Communications*, vol. 203, pp. 212–225, 2016, issn: 0010-4655. doi: 10.1016/j.cpc.2016.02.013. [Online]. Available: http://dx.doi.org/10.1016/j.cpc.2016.02.013.

[30] 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*, vol. 10, pp. 1–18, 2018, issn: 1758-2946. doi: 10.1186/s13321-018-0269-8. [Online]. Available: https://doi.org/10.1186/s13321-018-0269-8http://dx.doi.org/10.1186/s13321-018-0269-8.

[31] U. Gupta, T. Le, W. S. Hu, A. Bhan, and P. Daoutidis, "Automated network generation and analysis of biochemical reaction pathways using ring," *Metabolic Engineering*, vol. 49, pp. 84–93, 2018, issn: 1096-7184. doi: 10.1016/j.ymben.2018.07.009. [Online]. Available: https://doi.org/10.1016/j.ymben.2018.07.009.

[32] S. Maeda, K. Ohno, and K. Morokuma, "Systematic exploration of the mechanism of chemical reactions: The global reaction route mapping (grrm) strategy using the addf and afir methods," *Physical Chemistry Chemical Physics*, vol. 15, pp. 3683–3701, 2013, issn: 1463-9076. doi: 10.1039/c3cp44063j. [Online]. Available: http://dx.doi.org/10.1039/c3cp44063j.

[33] L.-P. Wang, A. Titov, R. Mcgibbon, F. Liu, V. S. Pande, and T. J. Martínez, "Discovering chemistry with an ab initio nanoreactor," *NATURE CHEMISTRY* |, vol. 6, 2014. doi: 10.1038/NCHEM.2099. [Online]. Available: www.nature.com/naturechemistryhttp://dx.doi.org/10.1038/NCHEM.2099.

[34] D. Rappoport, C. J. Galvin, D. Y. Zubarev, and A. Aspuru-Guzik, "Complex chemical reaction networks from heuristics-aided quantum chemistry," *Journal of Chemical Theory and Computation*, vol. 10, pp. 897–907, 2014, issn: 1549-9626. doi: 10.1021/ct401004r. [Online]. Available: http://dx.doi.org/10.1021/ct401004r.

[35] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, and T. Laino, ""found in translation": Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models," *Chemical Science*, vol. 9, pp. 6091–6098, 2018, issn: 2041-6539. doi: 10.1039/c8sc02339e. [Online]. Available: http://dx.doi.org/10.1039/c8sc02339e.

[36] J. Nam and J. Kim, "Linking the neural machine translation and the prediction of organic chemistry reactions," 2016. [Online]. Available: http://arxiv.org/abs/1612.09529.

[37] 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*, 2017. doi: 10.1021/acscentsci.7b00064. [Online]. Available: http://dx.doi.org/10.1021/acscentsci.7b00064.

[38] W. Jin, C. W. Coley, R. Barzilay, and T. Jaakkola, *Predicting organic reaction outcomes with weisfeiler-lehman network*, 2017. [Online]. Available: http://arxiv.org/abs/1709.04555v3.

[39] C. W. Coley *et al.*, "A graph-convolutional neural network model for the prediction of chemical reactivity," *Chemical Science*, vol. 10, pp. 370–377, 2 2019, issn: 2041-6539. doi: 10.1039/c8sc04228d. [Online]. Available: http://dx.doi.org/10.1039/c8sc04228d.

[40] C. Bilodeau, W. Jin, T. Jaakkola, R. Barzilay, K. F. Jensen, and C. F. K. Jensen, "Generative models for molecular discovery: Recent advances and challengesmachine learning," 2022. doi: 10.1002/wcms.1608. [Online]. Available: https://wires.onlinelibrary.wiley.com/doi/10.1002/wcms.1608http://dx.doi.org/10.1002/wcms.1608.

[41] D. M. Anstine and O. Isayev, "Generative models as an emerging paradigm in the chemical sciences," *Journal of the American Chemical Society*, vol. 145, pp. 8736–8750, 2023, issn: 1520-5126. doi: 10.1021/JACS.2C13467/ASSET/IMAGES/LARGE/JA2C13467_0005.JPEG. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/jacs.2c13467http://dx.doi.org/10.1021/JACS.2C13467/ASSET/IMAGES/LARGE/JA2C13467_0005.JPEGhttps://www.ncbi.nlm.nih.gov/pubmed/37052978.

[42] X. Zeng *et al.*, "Deep generative molecular design reshapes drug discovery," *Cell Reports Medicine*, vol. 3, p. 100 794, 12 2022. doi: 10.1016/j.xcrm.2022.100794. [Online]. Available: https://doi.org/10.1016/j.xcrm.2022.100794.

[43] M. H. S. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," 2017. doi: 10.1021/acscentsci.7b00512. [Online]. Available: https://pubs.acs.org/sharingguidelineshttp://dx.doi.org/10.1021/acscentsci.7b00512.

[44] M. Olivecrona, T. Blaschke, O. Engkvist, and H. Chen, "Molecular de-novo design through deep reinforcement learning," *Journal of Cheminformatics*, vol. 9, pp. 1–14, 2017, issn: 1758-2946. doi: 10.1186/S13321-017-0235-X. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0235-x.

[45] R. Gómezgómez-Bombarelli *et al.*, "Automatic chemical design using a data-driven continuous representation of molecules," vol. 17, p. 17, 2024. doi: 10.1021/acscentsci.7b00572. [Online]. Available: https://pubs.acs.org/sharingguidelineshttp://dx.doi.org/10.1021/acscentsci.7b00572.

[46] G. N. C. Simm, R. Pinsler, and J. M. Hernández-Lobato, *Reinforcement learning for molecular design guided by quantum mechanics*, Nov. 2020. [Online]. Available: https://proceedings.mlr.press/v119/simm20b.html.

[47] N. D. Cao and T. Kipf, "Molgan: An implicit generative model for small molecular graphs," 2018. [Online]. Available: https://arxiv.org/abs/1805.11973v2.

[48] Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia, "Learning deep generative models of graphs," 2018.

[49] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular graph generation," 2018.

[50] M. J. Kusner, B. Paige, and J. M. Hernández-Lobato, "Grammar variational autoencoder," 2017. [Online]. Available: http://opensmiles.org/spec/open-smiles-2-grammar.html.

[51] G. Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. Luis, C. Farias, and A. Aspuru-Guzik, "Objective-reinforced generative adversarial networks (organ) for sequence generation models," 2017. [Online]. Available: https://arxiv.org/abs/1705.10843v3.

[52] O. Prykhodko *et al.*, "A de novo molecular generation method using latent vector based generative adversarial network," *Journal of Cheminformatics*, vol. 11, 2019, issn: 1758-2946. doi: 10.1186/s13321-019-0397-9. [Online]. Available: http://dx.doi.org/10.1186/s13321-019-0397-9.

[53] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," [Online]. Available: https://tfhub.dev/s?q=biggan.

[54] C. Zang and F. Wang, "Moflow: An invertible flow model for generating molecular graphs," vol. 10, doi: 10.1145/3394486.3403104. [Online]. Available: https://doi.org/10.1145/3394486.3403104.

[55] K. Madhawa, K. Ishiguro, K. Nakago, and M. Abe, "Graphnvp: An invertible flow model for generating molecular graphs,"

[56] M. H. Segler, T. Kogej, C. Tyrchan, and M. P. Waller, "Generating focused molecule libraries for drug discovery with recurrent neural networks," *ACS central science*, vol. 4, no. 1, pp. 120–131, 2018.

[57] J. H. Jensen, "A graph-based genetic algorithm and generative model/monte carlo tree search for the exploration of chemical space," *Chemical Science*, vol. 10, pp. 3567–3572, 2019, issn: 2041-6539. doi: 10.1039/C8SC05372C. [Online]. Available: http://dx.doi.org/10.1039/C8SC05372C.

[58] G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, and A. L. Hopkins, "Quantifying the chemical beauty of drugs," *Nature Chemistry 2011 4:2*, vol. 4, pp. 90–98, 2012, issn: 1755-4349. doi: 10.1038/nchem.1243. [Online]. Available: https://www.nature.com/articles/nchem.1243http://dx.doi.org/10.1038/nchem.1243https://www.ncbi.nlm.nih.gov/pubmed/22270643.

[59] W. Jin, K. Yang, R. Barzilay, and T. Jaakkola, "Learning multimodal graph-to-graph translation for molecular optimization," *7th International Conference on Learning Representations, ICLR 2019*, 2018. [Online]. Available: https://arxiv.org/abs/1812.01070v3.

[60] W. Jin, R. Barzilay, and T. Jaakkola, "Hierarchical generation of molecular graphs using structural motifs," 2020. [Online]. Available: https://github.com/wengong-jin/hgraph2graph.

[61] H. Resat, L. Petzold, and M. F. Pettigrew, "Kinetic modeling of biological systems," *Methods in molecular biology (Clifton, N. J.)*, vol. 541, p. 311, 2009, issn: 1064-3745. doi: 10.1007/978-1-59745-243-4_14. [Online]. Available: http://dx.doi.org/10.1007/978-1-59745-243-4_14https://www.ncbi.nlm.nih.gov/pubmed/19381542.

[62] F. S. Wang, C. L. Ko, and E. O. Voit, "Kinetic modeling using s-systems and linlog approaches," *Biochemical Engineering Journal*, vol. 33, pp. 238–247, 2007, issn: 1369-703X. doi: 10.1016/J.BEJ.2006.11.002. [Online]. Available: http://dx.doi.org/10.1016/J.BEJ.2006.11.002.

[63] L. P. D. Oliveira, D. Hudebine, D. Guillaume, and J. J. Verstraete, "A review of kinetic modeling methodologies for complex processes," *Oil
Gas Science and Technology-Rev. IFP Energies nouvelles*, vol. 71, 2016. doi: 10.2516/ogst/2016011. [Online]. Available: http://dx.doi.org/10.2516/ogst/2016011.

[64] E. O. Voit, *Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists*. Cambridge University Press, 2000.

[65] J. Srividhya, E. J. Crampin, P. E. McSharry, and S. Schnell, "Reconstructing biochemical pathways from time course data," *PROTEOMICS*, vol. 7, pp. 828–838, 2007, issn: 1615-9861. doi: 10.1002/PMIC.200600428. [Online]. Available: http://dx.doi.org/10.1002/PMIC.200600428https://www.ncbi.nlm.nih.gov/pubmed/17370261.

[66] G. Craciun and C. Pantea, "Identifiability of chemical reaction networks," *Journal of Mathematical Chemistry*, vol. 44, pp. 244–259, 2008, issn: 0259-9791. doi: 10.1007/S10910-007-9307-X/METRICS. [Online]. Available: https://link.springer.com/article/10.1007/s10910-007-9307-xhttp://dx.doi.org/10.1007/S10910-007-9307-X/METRICS.

[67] 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, issn: 1547-5905. doi: 10.1002/AIC.12125. [Online]. Available: http://dx.doi.org/10.1002/AIC.12125.

[68] 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, pp. 7704–7717, 2010, issn: 0888-5885. doi: 10.1021/IE902015T. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ie902015thttp://dx.doi.org/10.1021/IE902015T.

[69] C Filippi-Bossy, J Bordet, J Villermaux, S Marchal-Brassely, and C Georgakis, "Batch reactor optimization by use of tendency models," *Computers & Chemical Engineering*, vol. 13, pp. 35–47, 1989, issn: 0098-1354. doi: 10.1016/0098-1354(89)89005-2. [Online]. Available: http://dx.doi.org/10.1016/0098-1354(89)89005-2.

[70] C Filippi *et al.*, "Tendency modeling of semibatch reactors for optimization and control," *Chemical Engineering Science*, vol. 41, pp. 913–920, 1986, issn: 0009-2509. doi: 10.1016/0009-2509(86)87175-5. [Online]. Available: http://dx.doi.org/10.1016/0009-2509(86)87175-5.

[71] D. Visser, R. V. D. 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, pp. 252–275, 2000, issn: 1096-7176. doi: 10.1006/MBEN.2000.0150. [Online]. Available: http://dx.doi.org/10.1006/MBEN.2000.0150https://www.ncbi.nlm.nih.gov/pubmed/11056067.

[72] N Bhatt, N Karimoglu, M Amrhein, W Marquardt, and D Bonvin, "Incremental model identification for reaction systems-a comparison of rate-based and extent-based approaches," *Chem. Eng. Sci*, 2011.

[73] 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, pp. 12 960–12 974, 2011, issn: 0888-5885. doi: 10.1021/IE2007196. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ie2007196http://dx.doi.org/10.1021/IE2007196.

[74] C. J. Taylor *et al.*, "An automated computational approach to kinetic model discrimination and parameter estimation," *Reaction Chemistry Engineering*, vol. 6, pp. 1404–1411, 2021, issn: 2058-9883. doi: 10.1039/D1RE00098E. [Online]. Available: http://dx.doi.org/10.1039/D1RE00098E.

[75] W. Ji and S. Deng, *Autonomous discovery of unknown reaction pathways from data by chemical reaction neural network*.

[76] A. Chakraborty, A. Sivaram, L. Samavedham, and V. Venkatasubramanian, "Mechanism discovery and model identification using genetic feature extraction and statistical testing," *Computers & Chemical Engineering*, vol. 140, p. 106 900, 2020, issn: 0098-1354. doi: 10.1016/J.COMPCHEMENG.2020.106900. [Online]. Available: http://dx.doi.org/10.1016/J.COMPCHEMENG.2020.106900.

[77] W. Zhang, S. Klus, T. Conrad, and C. Schütte, "Learning chemical reaction networks from trajectory data *," *J. Applied Dynamical Systems*, vol. 18, pp. 2000–2046, 2019. doi: 10.1137/19M1265880. [Online]. Available: https://doi.org/10.1137/19M1265880http://dx.doi.org/10.1137/19M1265880.

[78] P.-M. Jacob and A. A. Lapkin, "Prediction of chemical reactions using statistical models of chemical knowledge,"

[79] A. Kruse, A. Krupka, V. Schwarzkopf, C. Gamard, and T. Henningsen, "Influence of proteins on the hydrothermal gasification and liquefaction of biomass. 1. comparison of different feedstocks," *Industrial and Engineering Chemistry Research*, vol. 44, pp. 3013–3020, 9 2005, issn: 08885885. doi: 10.1021/ie049129y.

[80] K. E. Balat H. Balat M Balat M, "Main routes for the thermo-conversion of biomass into fuels and chemicals. part 1: Pyrolysis systems.," *Energy Convers Manag*, vol. 50, pp. 3147–57, 12 2009.

[81] F. L. Tiilikkala J. Tiilikkala K, "History and use of wood pyrolysis liquids as biocide and plant protection product history and use of wood pyrolysis liquids as biocide and plant protection product.," *The Open Agriculture Journal*, vol. 4, pp. 111–18, 2010.

[82] M. G. de Haan AB Vitasari C, "Conceptual process design of an integrated bio-based acetic acid, glycolaldehyde, and acetol production in a pyrolysis oil-based biorefinery," *Chem Eng Res*, vol. 95, pp. 133–143, 2015.

[83] R. F. B. W. L. B. D. P. W, "Hydrothermalliquefaction (htl) of microalgae for biofuel production: State of the artreview and future prospeects," *Biomass Bioenergy*, vol. 53, pp. 113–127, 2013.

[84] S. KB, N. S, M. J, F. M, F. VV, and K. HC, "On water: Unique reactivity of organic compounds in aqueous suspension.," *Angew Chem Int Ed*, vol. 44, pp. 3275–9, 2005.

[85] Jr. and R. P. L. M. F. M. J. A. J. W. T. A. A. P. F. Vogel, "Thermochemical biofuel production in hydrothermal media: A review of sub- and supercritical water technologies," *Energy and Enviromental Sicence*, pp. 32–65, 1 2008.

[86] A. Uihlein and L. Schebek, "Environmental impacts of a lignocellulose feedstock biorefinery system: An assessment," *Biomass and Bioenergy*, vol. 33, pp. 793–802, 5 2009, issn: 09619534. doi: 10.1016/j.biombioe.2008.12.001.

[87] T. J. E. E. J. Soltes, "Pyrolysis, in: I.s. goldstein (ed.), organic chemicals from biomass," *CRC Press, Boca Raton*, pp. 63–99, 1981.

[88] S. M. Subramanya and P. E. Savage, "Identifying and modeling interactions between biomass components during hydrothermal liquefaction in sub-, near-, and supercritical water," *ACS Sustainable Chemistry and Engineering*, vol. 9, pp. 13 874–13 882, 41 Feb. 2021, issn: 21680485. doi: https://doi.org/10.1021/acssuschemeng.1c04810. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acssuschemeng.1c04810.

[89] B. Hao, D. Xu, G. Jiang, T. A. Sabri, Z. Jing, and Y. Guo, "Chemical reactions in the hydrothermal liquefaction of biomass and in the catalytic hydrogenation upgrading of biocrude," *Green Chemistry*, vol. 23, pp. 1562–1583, 4 2021, issn: 1463-9270. doi: 10.1039/D0GC02893B. [Online]. Available: http://dx.doi.org/10.1039/D0GC02893B.

[90] 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," *Industrial and Engineering Chemistry Research*, vol. 58, pp. 11 149–11 163, 2019, issn: 1520-5045. doi: 10.1021/ACS.IECR.9B00280/. [Online]. Available: https://pubs-acs-org.login.ezproxy.library.ualberta.ca/doi/full/10.1021/acs.iecr.9b00280http://dx.doi.org/10.1021/ACS.IECR.9B00280/ASSET/IMAGES/LARGE/IE-2019-00280G_0001.JPEG.

[91] J. M. Chalmers, "Spectroscopy in process analysis," *Sheffield Academic Press*, 2000.

[92] S. A. H. M. D. L. Hall, *Mathemathical Techniques in Multisensor Data Fusion*. ARTEC House, INC, 2004.

[93] F. Castanedo, "A review of data fusion techniques," *The Scientific World Journal*, vol. 2013, 2013, issn: 1537744X. doi: 10.1155/2013/704504.

[94] D. Friedman, N. Nachman, and I. Peer, "Learning bayesian network structure from massive datasets: The "sparse candidate" algorithm," 1999, pp. 206–215.

[95] I. Tsamardinos, L. E. Brown, and C. F. Aliferis, "The max-min hill-climbing bayesian network structure learning algorithm," *Springer Science*, pp. 31–78, 2006. doi: 10.1007/s10994-006-6889-7.

[96] K. Friston, R. Moran, and A. K. Seth, "Analysing connectivity with granger causality and dynamic causal modelling," *Current Opinion in Neurobiology*, vol. 23, pp. 172–178, 2 2013, issn: 0959-4388. doi: 10.1016/j.conb.2012.11.010.

[97] D. M. F. J. G. S. W. J, "Uncovering interactions in the frequence domain.," *PLoS Computational Biology*, vol. 4, 5 2008.

[98] G. D. K. Y. K. N. J. C. S. E. A. S. M. G. J. F. G. M. A. J. R. Y. H, "Bayesian networks approach for predicting protein-protein interactions from genomic data.," *Science*, 302: 449, 2003.

[99] C. Zou, K. J. Denby, and J. Feng, "Granger causality vs. dynamic bayesian network inference: A comparative study," *BMC Bioinformatics*, vol. 10, 2009, issn: 14712105. doi: 10.1186/1471-2105-10-401.

[100] H. Jia, Y. Li, B. Dong, and H. Ya, "An improved tabu search approach to vehicle routing problem," *Procedia - Social and Behavioral Sciences*, vol. 96, pp. 1208–1217, Cictp 2013, issn: 1877-0428. doi: 10.1016/j.sbspro.2013.08.138.

[101] T. Jiang, G. Ren, and X. Zhao, "Evacuation route optimization based on tabu search algorithm and hill-climbing algorithm," *Procedia - Social and Behavioral Sciences*, vol. 96, pp. 865–872, Cictp 2013, issn: 1877-0428. doi: 10.1016/j.sbspro.2013.08.098.

[102] D. J. G. et al Yeh TM, "Hydrothermal catalytic production of fuels and chemicals from aquatic biomass.," *J Chem Technol Biotechnol*, vol. 88, pp. 13–24, 1 2013.

[103] S. PE, "A perspective on catalysis in sub- and supercritical water.," *J Supercrit Fluids*, vol. 47, pp. 407–414, 3 2009.

[104] T. Kourti, "Process analytical technology beyond real-time analyzers: The role of multivariate analysis," *Critical Reviews in Analytical Chemistry*, vol. 36, pp. 257–278, 3-4 2006. doi: 10.1080/10408340600969957. [Online]. Available: https://doi.org/10.1080/10408340600969957http://dx.doi.org/10.1080/10408340600969957.

[105] M. Staš, D. Kubička, J. Chudoba, and M. Pospíšil, "Overview of analytical methods used for chemical characterization of pyrolysis bio-oil," *Energy and Fuels*, vol. 28, pp. 385–402, 1 2014, issn: 08870624. doi: 10.1021/ef402047y.

[106] R Bassilakis, R. M. Carangelo, and M. A. Wojtowicz, "Tg-ftir analysis of biomass pyrolysis," *Fuel*, vol. 80, pp. 1765–1786, 12 2001, issn: 0016-2361.

[107] D. L. Hall and J Llinas, "An introduction to multisensor data fusion," *Proceedings of the IEEE*, vol. 85, pp. 6–23, 1 1997.

[108] J. Dong, D. Zhuang, Y. Huang, and J. Fu, "Advances in multi-sensor data fusion: Algorithms and applications," pp. 7771–7784, 1 2009. doi: 10.3390/s91007771.

[109] F. Sattari and V. Prasad, "Data-driven hypotheses of reaction networks for thermochemical conversion of a physical mixture of levoglucosan and 2-phenoxyethyl benzene," *Canadian Journal of Chemical Engineering*, vol. 100, S154–S159, S1 2022, issn: 1939-019X. doi: 10.1002/cjce.24071. [Online]. Available: http://dx.doi.org/10.1002/cjce.24071.

[110] K. Sivaramakrishnan, A. Puliyanda, A. D. Klerk, and V. Prasad, "A data-driven approach to generate pseudo-reaction sequences for the thermal conversion of athabasca bitumen," *Reaction Chemistry and Engineering*, vol. 6, pp. 505–537, 3 2021, issn: 2058-9883. doi: 10.1039/D0RE00321B. [Online]. Available: http://dx.doi.org/10.1039/D0RE00321B.

[111] D. T. Tefera, A. Agrawal, L. M. Y. Jaramillo, A. D. 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, pp. 10 756–10 769, 38 2017, issn: 1520-5045. doi: 10.1021/acs.iecr.7b01849. [Online]. Available: http://dx.doi.org/10.1021/acs.iecr.7b01849.

[112] H. L. Hammer, A. Yazidi, and B. J. Oommen, ""anti-bayesian" flat and hierarchical clustering using symmetric quantiloids," *Information Sciences*, vol. 418-419, pp. 495–512, August 2016 2017, issn: 00200255. doi: 10.1016/j.ins.2017.08.017.

[113] K. S. F.-S. S, "Model-based clustering of multiple time series.," *Journal of Business and Economic Statistics*, vol. 26, pp. 78–89, 2008.

[114] D. A. F. W. J. E. D. M, "Bayesian unsupervised signal classification by dirichlet process mixtures of gaussian processes.," 2007.

[115] R. B. B. M, "The bayesian revolution in genetics," *Nat Rev Genet*, pp. 251–261, 4 2004.

[116] R. S. Savage *et al.*, "R / bhc : Fast bayesian hierarchical clustering for microarray data," vol. 9, pp. 1–9, 2009. doi: 10.1186/1471-2105-10-242.

[117] M. G. Tadesse, N. Sha, and M. Vannucci, "Bayesian variable selection in clustering high-dimensional data," *Journal of the American Statistical Association*, vol. 100, pp. 602–617, 470 2005, issn: 01621459. doi: 10.1198/016214504000001565.

[118] K. A. Heller and Z. Ghahramani, "Bayesian hierarchical clustering," 2005.

[119] S. M. Lee and P. A. Abbott, "Bayesian networks for knowledge discovery in large datasets: Basics for nurse researchers," *Journal of Biomedical Informatics*, vol. 36, pp. 389–399, 4-5 2003, issn: 1532-0464. doi: 10.1016/J.JBI.2003.09.022. [Online]. Available: http://dx.doi.org/10.1016/J.JBI.2003.09.022https://www.ncbi.nlm.nih.gov/pubmed/14643735.

[120] D. Koller and A. Pfeffer, "Probabilistic frame-based systems," 1998. [Online]. Available: www.aaai.org.

[121] L. Zhang, C. Shen, and R. Liu, "Gc – ms and ft-ir analysis of the bio-oil with addition of ethyl acetate during storage," vol. 2, pp. 1–6, January 2014. doi: 10.3389/fenrg.2014.00003.

[122] B. G. H. Dunteman, *Principal Components Analysis*. SAGD publications, Inc, 1976.

[123] D. C. Elliott, P. Biller, A. B. Ross, A. J. Schmidt, and S. B. Jones, "Hydrothermal liquefaction of biomass: Developments from batch to continuous process," *Bioresource Technology*, vol. 178, pp. 147–156, 2015, issn: 18732976. doi: 10.1016/j.biortech.2014.09.132.

[124] L. Abella, S. Nanbu, and K. Fukuda, "A theoretical study on levoglucosan pyrolysis reactions yielding aldehydes and a ketone in biomass," 2007.

[125] S Li, J Lyons-Hart, J Banyasz, and K Shafer, "Real-time evolved gas analysis by ftir method: An experimental study of cellulose pyrolysis," *Fuel*, vol. 80, pp. 1809–1817, 12 Feb. 2001, issn: 00162361. doi: 10.1016/S0016-2361(01)00064-3.

[126] Z. Y., "Hydrothermal liquefaction to convert biomass into crude oil. in: Biofuels from agricultural wastes and byproducts.," *Hoboken, NJ: Wiley-Blackwell*, pp. 201–232, 2010.

[127] C. W.-T. Z. P. D. Y. G. C. Z. Y, "An investigation of reaction pathways of hydrothermal liquefaction using chlorella pyrenoidosa and spirulina platensis.," *Energy Convers Manag*, vol. 96, pp. 330–339, 2015.

[128] J. F. J. C. S. K. X. Li, "Classified separation of lignin hydrothermal liquefied products," *Ind. Eng. Chem. Res.*, vol. 50, pp. 11 288–11 296, 2011.

[129] S. S. H. K. K. Ehara, "Characterization of the lignin-derived products from wood as treated in supercritical water, j. wood sci. 48 (2002) 320–325.," *J. Wood Sci*, vol. 48, pp. 320–325, 2002.

[130] Y. Y. N. S. L. L. M. Yoshioka, "Liquefaction mechanism of lignin in the presence of phenol at elevated temperature without catalysts. studies on β-o-4 lignin model compound iii. multi-condensation," *Holzforschung*, vol. 51, pp. 333–337, 1997.

[131] R. B. Madsen, H. Zhang, P. Biller, A. H. Goldstein, and M. Glasius, "Characterizing semivolatile organic compounds of biocrude from hydrothermal liquefaction of biomass," *Energy and Fuels*, vol. 31, pp. 4122–4134, 4 2017, issn: 15205029. doi: 10.1021/acs.energyfuels.7b00160.

[132] L. N. H. G. T. GA, "Catalytic conversion of sugars to fuels.," *John Wiley & Sons*, pp. 273–279, 2011.

[133] T. H. Pedersen, C. U. Jensen, L Sandström, and L. A. Rosendahl, "Full characterization of compounds obtained from fractional distillation and upgrading of a htl biocrude," *Applied Energy*, vol. 202, pp. 408–419, 2017, issn: 0306-2619. doi: 10.1016/j.apenergy.2017.05.167. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030626191730733Xhttp://dx.doi.org/10.1016/j.apenergy.2017.05.167.

[134] X. Zhang, T. Wang, L. Ma, and J. Chang, "Vacuum pyrolysis of waste tires with basic additives," *Waste Management*, vol. 28, pp. 2301–2310, 11 2008, issn: 0956053X. doi: 10.1016/j.wasman.2007.10.009.

[135] M. S. M. G. Wahyudiono, "Conversion of biomass model compound under hydrothermal conditions using batch reactor," *Fuel*, pp. 1656–1664, 2009.

[136] X. Huang, D. guo Cheng, F. Chen, and X. Zhan, "Reaction pathways of hemicellulose and mechanism of biomass pyrolysis in hydrogen plasma: A density functional theory study," *Renewable Energy*, vol. 96, pp. 490–497, 2016.

[137] M. D. H. M. G. R. L. X. L. C. and Li, "An ft-ir spectroscopic study of carbonyl functionalities in bio-oils. fuel 90," *Fuel*, vol. 90, 3417 2011.

[138] A Kruse and A Gawlik, "Biomass conversion in water at 330-410 °c and 30-50 mpa. identification of key compounds for indicating different chemical reaction pathways," *Industrial and Engineering Chemistry Research*, vol. 42, pp. 267–279, 2 2003, issn: 08885885. doi: 10.1021/ie0202773.

[139] P. D. Muley, C. Henkel, K. K. Abdollahi, C. Marculescu, and D. Boldor, "A critical comparison of pyrolysis of cellulose, lignin, and pine sawdust using an induction heating reactor," *Energy Conversion and Management*, vol. 117, pp. 273–280, 2016.

[140] 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. doi: 10.1016/J.DCHE.2021.100009. [Online]. Available: http://dx.doi.org/10.1016/J.DCHE.2021.100009.

[141] A. Puliyanda, Z. Li, and V. Prasad, "Real-time monitoring of reaction mechanisms from spectroscopic data using hidden semi-markov models for mode identification," *Journal of Process Control*, vol. 117, pp. 188–205, 2022, issn: 0959-1524. doi: 10.1016/J.JPROCONT.2022.07.011. [Online]. Available: http://dx.doi.org/10.1016/J.JPROCONT.2022.07.011.

[142] K. Srinivasan, A. Puliyanda, and V. Prasad, "Automated generation of reaction network hypotheses for complex feedstocks," *2022 IEEE International Symposium on Advanced Control of Industrial Processes, AdCONIP 2022*, vol. 125, pp. 234–239, 2022, issn: 0952-1976. doi: 10.1109/ADCONIP55568.2022.9894209. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623008746http://dx.doi.org/10.1109/ADCONIP55568.2022.9894209.

[143] *Feedstock Technologies | Department of Energy*, n.d. [Online]. Available: https://www.energy.gov/eere/bioenergy/feedstock-technologies (visited on 03/01/2023).

[144] S. I. Mussatto *et al.*, "Sustainable aviation fuels: Production, use and impact on decarbonization," *Comprehensive Renewable Energy*, pp. 348–371, 2022. doi: 10.1016/B978-0-12-819727-1.00057-1. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-819727-1.00057-1.

[145] S. V. Mohan, "Reorienting waste remediation towards harnessing bioenergy: A paradigm shift," *Industrial Wastewater Treatment, Recycling and Reuse*, pp. 235–281, 2014. doi: 10.1016/B978-0-08-099968-5.00006-4. [Online]. Available: http://dx.doi.org/10.1016/B978-0-08-099968-5.00006-4.

[146] C. Ling, J. Hamilton, and B. Khandelwal, "Feedstock and pathways for alternative aviation fuels," *Aviation Fuels*, pp. 1–22, 2021. doi: 10.1016/B978-0-12-818314-4.00004-2. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-818314-4.00004-2.

[147] R. Obeid, D. M. Lewis, N. Smith, T. Hall, and P. V. Eyk, "Reaction kinetics and characterization of species in renewable crude from hydrothermal liquefaction of mixtures of polymer compounds to represent organic fractions of biomass feedstocks," *Energy and Fuels*, vol. 34, pp. 419–429, 1 2020, issn: 1520-5029. doi: 10.1021/ACS.ENERGYFUELS.9B02936/ASSET/IMAGES/LARGE/EF9B02936_0005.JPEG. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.energyfuels.9b02936http://dx.doi.org/10.1021/ACS.ENERGYFUELS.9B02936/ASSET/IMAGES/LARGE/EF9B02936_0005.JPEG.

[148] C. Fernández, M. P. Callao, and M. S. Larrechi, "Uv-visible-dad and 1h-nmr spectroscopy data fusion for studying the photodegradation process of azo-dyes using mcr-als," *Talanta*, vol. 117, pp. 75–80, 2013, issn: 0039-9140. doi: 10.1016/j.talanta.2013.08.004. [Online]. Available: http://dx.doi.org/10.1016/j.talanta.2013.08.004.

[149] M. A. Nemeth, "Multi- and megavariate data analysis," *Technometrics*, vol. 45, pp. 362–362, 4 2003. doi: 10.1198/tech.2003.s162. [Online]. Available: http://dx.doi.org/10.1198/tech.2003.s162.

[150] D. T. Tefera, L. M. Y. Jaramillo, R. Ranjan, C. Li, A. D. 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, pp. 1961–1970, 8 2017, issn: 1520-5045. doi: 10.1021/acs.iecr.6b04437. [Online]. Available: http://dx.doi.org/10.1021/acs.iecr.6b04437.

[151] A. A. Lee *et al.*, "Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space," *Chemical Communications*, vol. 55, pp. 12 152–12 155, 81 2019, issn: 1364-548X. doi: 10.1039/c9cc05122h. [Online]. Available: http://dx.doi.org/10.1039/c9cc05122hhttps://www.ncbi.nlm.nih.gov/pubmed/31497831.

[152] J. J. Naveja, B. A. Pilón-Jiménez, J. Bajorath, and J. L. Medina-Franco, "A general approach for retrosynthetic molecular core analysis," *Journal of Cheminformatics*, vol. 11, pp. 1–9, 1 2019, issn: 1758-2946. doi: 10.1186/s13321-019-0380-5. [Online]. Available: https://doi.org/10.1186/s13321-019-0380-5http://dx.doi.org/10.1186/s13321-019-0380-5.

[153] K. Lin, Y. Xu, J. Pei, and L. Lai, "Automatic retrosynthetic route planning using template-free models," *Chemical Science*, vol. 11, pp. 3355–3364, 12 2020, issn: 2041-6539. doi: 10.1039/c9sc03666k. [Online]. Available: http://dx.doi.org/10.1039/c9sc03666k.

[154] I. A. Watson, J. Wang, and C. A. Nicolaou, "A retrosynthetic analysis algorithm implementation," *Journal of Cheminformatics*, vol. 11, pp. 1–12, 1 2019, issn: 1758-2946. doi: 10.1186/s13321-018-0323-6. [Online]. Available: https://doi.org/10.1186/s13321-018-0323-6http://dx.doi.org/10.1186/s13321-018-0323-6.

[155] A. A. Lapkin *et al.*, "Automation of route identification and optimisation based on data-mining and chemical intuition," *Faraday Discussions*, vol. 202, pp. 483–496, 2017, issn: 1364-5498. doi: 10.1039/c7fd00073a. [Online]. Available: http://dx.doi.org/10.1039/C7FD00073Ahttp://dx.doi.org/10.1039/c7fd00073a.

[156] J. P. Unsleber and M. Reiher, "The exploration of chemical reaction networks," *Annual Review of Physical Chemistry*, vol. 71, pp. 121–142, 1 2020, issn: 0066-426X. doi: 10.1146/annurev-physchem-071119-040123. [Online]. Available: http://dx.doi.org/10.1146/annurev-physchem-071119-040123http://arxiv.org/abs/1906.10223.

[157] G. N. Simm, A. C. Vaucher, and M. Reiher, "Exploration of reaction pathways and chemical transformation networks," *Journal of Physical Chemistry A*, vol. 123, pp. 385–399, 2 2019, issn: 1520-5215. doi: 10.1021/acs.jpca.8b10007. [Online]. Available: http://dx.doi.org/10.1021/acs.jpca.8b10007http://arxiv.org/abs/1810.07490.

[158] M. H. S. Segler and M. P. Waller, "Neural-symbolic machine learning for retrosynthesis and reaction prediction," *Chemistry - A European Journal*, vol. 23, pp. 5966–5971, 25 2017, issn: 1521-3765. doi: 10.1002/chem.201605499. [Online]. Available: http://dx.doi.org/10.1002/chem.201605499.

[159] S. Rangarajan, A. Bhan, and P. Daoutidis, "Rule-based generation of thermochemical routes to biomass conversion," *Industrial and Engineering Chemistry Research*, vol. 49, pp. 10 459–10 470, 21 2010, issn: 0888-5885. doi: 10.1021/ie100546t. [Online]. Available: http://dx.doi.org/10.1021/ie100546t.

[160] I. Ismail, H. B. V. A. Stuttaford-Fowler, C. O. Ashok, C. Robertson, and S. Habershon, "Automatic proposal of multistep reaction mechanisms using a graph-driven search," *Physical Chemistry*, 2019. doi: 10.1021/acs.jpca.9b01014. [Online]. Available: https://pubs.acs.org/sharingguidelineshttp://dx.doi.org/10.1021/acs.jpca.9b01014.

[161] J. A. Varela, S. A. VÃ¡zquez, and E MartÃnez-NÃºÃ±ez, "An automated method to find reaction mechanisms and solve the kinetics in organometallic catalysis," *Chem. Sci.*, vol. 8, pp. 3843–3851, 2017. doi: 10.1039/C7SC00549K. [Online]. Available: http://dx.doi.org/10.1039/C7SC00549K.

[162] A. L. Dewyer, A. J. Argüelles, and P. M. Zimmerman, "Methods for exploring reaction space in molecular systems," *Wiley Interdisciplinary Reviews: Computational Molecular Science*, vol. 8, pp. 1–20, 2 2018, issn: 1759-0884. doi: 10.1002/wcms.1354. [Online]. Available: http://dx.doi.org/10.1002/wcms.1354.

[163] G. N. Simm and M. Reiher, "Context-driven exploration of complex chemical reaction networks," *Journal of Chemical Theory and Computation*, vol. 13, pp. 6108–6119, 12 2017, issn: 1549-9626. doi: 10.1021/acs.jctc.7b00945. [Online]. Available: http://dx.doi.org/10.1021/acs.jctc.7b00945http://arxiv.org/abs/1709.02479.

[164] J. W. Kim, Y. Kim, K. Y. Baek, K. Lee, and W. Y. Kim, "Performance of ace-reaction on 26 organic reactions for fully automated reaction network construction and microkinetic analysis," *Journal of Physical Chemistry A*, vol. 123, pp. 4796–4805, 22 2019, issn: 1520-5215. doi: 10.1021/acs.jpca.9b02161. [Online]. Available: http://dx.doi.org/10.1021/acs.jpca.9b02161https://www.ncbi.nlm.nih.gov/pubmed/31074624.

[165] S. J. Klippenstein and C. Cavallotti, *Ab initio kinetics for pyrolysis and combustion systems*, 1st ed. Elsevier B.V., 2019, vol. 45, isbn: 9780444640871. doi: 10.1016/B978-0-444-64087-1.00002-4. [Online]. Available: http://dx.doi.org/10.1016/B978-0-444-64087-1.00002-4.

[166] C Cavallotti, M Pelucchi, Y Georgievskii, and S. J. Klippenstein, "Estoktp: Electronic structure to temperature-and pressure-dependent rate constants-a code for automatically predicting the thermal kinetics of reactions," *Chemical Theory and Computation*, 2018. doi: 10.1021/acs.jctc.8b00701. [Online]. Available: https://pubs.acs.org/sharingguidelineshttp://dx.doi.org/10.1021/acs.jctc.8b00701.

[167] A. Ahmed *et al.*, "Small ester combustion chemistry: Computational kinetics and experimental study of methyl acetate and ethyl acetate," *Proceedings of the Combustion Institute*, vol. 37, pp. 419–428, 1 2019, issn: 1540-7489. doi: 10.1016/j.proci.2018.06.178. [Online]. Available: http://dx.doi.org/10.1016/j.proci.2018.06.178.

[168] C. W. Coley, W. H. Green, and K. F. Jensen, "Machine learning in computer-aided synthesis planning," *Accounts of Chemical Research*, 2018. doi: 10.1021/acs.accounts.8b00087. [Online]. Available: https://pubs.acs.org/sharingguidelineshttp://dx.doi.org/10.1021/acs.accounts.8b00087.

[169] S. Ishida, K. Terayama, R. Kojima, K. Takasu, and Y. Okuno, "Prediction and interpretable visualization of retrosynthetic reactions using graph convolutional networks," *Journal of Chemical Information and Modeling*, vol. 59, pp. 5026–5033, 12 2019, issn: 1520-5142. doi: 10.1021/acs.jcim.9b00538. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.9b00538https://www.ncbi.nlm.nih.gov/pubmed/31769668.

[170] V. Venkatasubramanian and V. Mann, "Artificial intelligence in reaction prediction and chemical synthesis," *Current Opinion in Chemical Engineering*, vol. 36, p. 100 749, 2022, issn: 2211-3398. doi: 10.1016/J.COCHE.2021.100749. [Online]. Available: http://dx.doi.org/10.1016/J.COCHE.2021.100749.

[171] *How to use software and spectral libraries for IR and Raman compound identification - 2020 - Wiley Analytical Science*, n.d. [Online]. Available: https://analyticalscience.wiley.com/do/10.1002/was.00080147 (visited on 03/01/2023).

[172] *Automatic and Interactive IR Spectrum Interpretation - 2010 - Wiley Analytical Science*, n.d. [Online]. Available: https://analyticalscience.wiley.com/do/10.1002/gitlab.555 (visited on 03/01/2023).

[173] R. Nalla, R. Pinge, M. Narwaria, and B. Chaudhury, "Priority based functional group identification of organic molecules using machine learning," *ACM International Conference Proceeding Series*, vol. 18, pp. 201–209, 2018. doi: 10.1145/3152494.3152522. [Online]. Available: https://dl.acm.org/doi/10.1145/3152494.3152522http://dx.doi.org/10.1145/3152494.3152522.

[174] J. Yang, J. Xu, X. Zhang, C. Wu, T. Lin, and Y. Ying, "Deep learning for vibrational spectral analysis: Recent progress and a practical guide," *Analytica chimica acta*, vol. 1081, pp. 6–17, 2019, issn: 1873-4324. doi: 10.1016/J.ACA.2019.06.012. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31446965/http://dx.doi.org/10.1016/J.ACA.2019.06.012https://www.ncbi.nlm.nih.gov/pubmed/31446965.

[175] Z. Wang, X. Feng, J. Liu, M. Lu, and M. Li, "Functional groups prediction from infrared spectra based on computer-assist approaches," *Microchemical Journal*, vol. 159, p. 105 395, 2020, issn: 0026-265X. doi: 10.1016/J.MICROC.2020.105395. [Online]. Available: http://dx.doi.org/10.1016/J.MICROC.2020.105395.

[176] G. Renner, T. C. Schmidt, and J. Schram, "A new chemometric approach for automatic identification of microplastics from environmental compartments based on ft-ir spectroscopy," *Analytical Chemistry*, 2017. doi: 10.1021/ACS.ANALCHEM.7B02472. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.analchem.7b02472http://dx.doi.org/10.1021/ACS.ANALCHEM.7B02472.

[177] K. Judge, C. W. Brown, and L. Hamel, "Sensitivity of infrared spectra to chemical functional groups," *Analytical Chemistry*, 2008. doi: 10.1021/AC8000429. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/ac8000429http://dx.doi.org/10.1021/AC8000429.

[178] Y Lecun, L Bottou, Y Bengio, and P Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, pp. 2278–2324, 11 1998. doi: 10.1109/5.726791. [Online]. Available: http://dx.doi.org/10.1109/5.726791.

[179] A. A. Enders, N. M. North, C. M. Fensore, J. Velez-Alvarez, and H. C. Allen, "Functional group identification for ftir spectra using image-based machine learning models," *Analytical Chemistry*, vol. 93, pp. 9711–9718, 28 2021, issn: 1520-6882. doi: 10.1021/ACS.ANALCHEM.1C00867/. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.analchem.1c00867http://dx.doi.org/10.1021/ACS.ANALCHEM.1C00867/ASSET/IMAGES/LARGE/AC1C00867_0005.JPEG.

[180] S. Jiang *et al.*, "Using atr-ftir spectra and convolutional neural networks for characterizing mixed plastic waste," *Computers and Chemical Engineering*, vol. 155, p. 107 547, 2021, issn: 0098-1354. doi: 10.1016/J.COMPCHEMENG.2021.107547. [Online]. Available: http://dx.doi.org/10.1016/J.COMPCHEMENG.2021.107547.

[181] J. A. Fine, A. A. Rajasekar, K. P. Jethava, and G. Chopra, "Spectral deep learning for prediction and prospective validation of functional groups," *Chemical Science*, vol. 11, pp. 4618–4630, 18 2020, issn: 2041-6539. doi: 10.1039/C9SC06240H. [Online]. Available: http://dx.doi.org/10.1039/C9SC06240H.

[182] R. Fan, X. Yang, C. F. Drury, and Z. Zhang, "Curve-fitting techniques improve the mid-infrared analysis of soil organic carbon: A case study for brookston clay loam particle-size fractions," *Scientific Reports 2018 8:1*, vol. 8, pp. 1–10, 1 2018, issn: 2045-2322. doi: 10.1038/s41598-018-30704-2. [Online]. Available: https://www.nature.com/articles/s41598-018-30704-2http://dx.doi.org/10.1038/s41598-018-30704-2https://www.ncbi.nlm.nih.gov/pubmed/30111781.

[183] A. Sadat and I. J. Joye, "Peak fitting applied to fourier transform infrared and raman spectroscopic analysis of proteins," *Applied Sciences 2020, Vol. 10, Page 5918*, vol. 10, p. 5918, 17 2020, issn: 2076-3417. doi: 10.3390/APP10175918. [Online]. Available: http://dx.doi.org/10.3390/APP10175918.

[184] M. Bradley, *Curve Fitting in Raman and IR Spectroscopy: Basic Theory of Line Shapes and Applications*, n.d. [Online]. Available: https://assets.thermofisher.com/TFS-Assets/CAD/Application-Notes/AN50733_E.pdf.

[185] D. M. Lowe, "Extraction of chemical structures and reactions from the literature," 2012.

[186] *RDKit: Open-source cheminformatics.* n.d. [Online]. Available: https://www.rdkit.org.

[187] L. Xue and J. Bajorath, "Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening," *Combinatorial Chemistry and High Throughput Screening*, vol. 3, pp. 363–372, 5 2012, issn: 1386-2073. doi: 10.2174/1386207003331454. [Online]. Available: http://dx.doi.org/10.2174/1386207003331454https://www.ncbi.nlm.nih.gov/pubmed/11032954.

[188] A. Cereto-Massagué, M. J. Ojeda, C. Valls, M. Mulero, S. Garcia-Vallvé, and G. Pujadas, "Molecular fingerprint similarity search in virtual screening," *Methods*, vol. 71, pp. 58–63, C 2015, issn: 1095-9130. doi: 10.1016/j.ymeth.2014.08.005. [Online]. Available: http://dx.doi.org/10.1016/j.ymeth.2014.08.005https://www.ncbi.nlm.nih.gov/pubmed/25132639.

[189] N. Schneider, D. M. Lowe, R. A. Sayle, and G. A. Landrum, "Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity," *Journal of Chemical Information and Modeling*, vol. 55, pp. 39–53, 1 2015, issn: 1549-960X. doi: 10.1021/CI5006614/SUPPL_FILE/CI5006614_SI_004.PDF. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/ci5006614http://dx.doi.org/10.1021/CI5006614/SUPPL_FILE/CI5006614_SI_004.PDFhttps://www.ncbi.nlm.nih.gov/pubmed/25541888.

[190] E. F.-D. Gortari, C. R. García-Jacas, K. Martinez-Mayorga, and J. L. Medina-Franco, "Database fingerprint (dfp): An approach to represent molecular databases," *Journal of Cheminformatics*, vol. 9, pp. 1–9, 1 2017, issn: 1758-2946. doi: 10.1186/S13321-017-0195-1/FIGURES/6. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-017-0195-1http://dx.doi.org/10.1186/S13321-017-0195-1/FIGURES/6.

[191] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of mdl keys for use in drug discovery," *Journal of Chemical Information and Computer Sciences*, 2002. doi: 10.1021/ci010132r. [Online]. Available: https://pubs.acs.org/sharingguidelineshttp://dx.doi.org/10.1021/ci010132r.

[192] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, pp. 742–754, 2010.

[193] D. Fooshee *et al.*, "Deep learning for chemical reaction prediction," *Molecular Systems Design and Engineering*, vol. 3, pp. 442–452, 3 2018, issn: 2058-9689. doi: 10.1039/c7me00107j. [Online]. Available: http://dx.doi.org/10.1039/c7me00107j.

[194] S. Stocker, G. Csányi, K. Reuter, and J. T. Margraf, "Machine learning in chemical reaction space," *Nature Communications*, vol. 11, 1 2020, issn: 2041-1723. doi: 10.1038/s41467-020-19267-x. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/33127879.

[195] *Indigo Toolkit*. [Online]. Available: https://lifescience.opensource.epam.com/indigo/index.html (visited on 11/17/2023).

[196] *Daylight Theory: SMIRKS - A Reaction Transform Language*, n.d. [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (visited on 03/02/2023).

[197] *Daylight Theory: SMARTS - A Language for Describing Molecular Patterns*, n.d. [Online]. Available: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html (visited on 07/15/2020).

[198] J. R. Ullmann, "An algorithm for subgraph isomorphism," *J. ACM*, vol. 23, pp. 31–42, 1 1976, issn: 0004-5411. doi: 10.1145/321921.321925. [Online]. Available: https://doi.org/10.1145/321921.321925http://dx.doi.org/10.1145/321921.321925.

[199] T. Wang, Y. Tan, Y. Z. Chen, and C. Tan, "Infrared spectral analysis for prediction of functional groups based on feature-aggregated deep learning," *Journal of Chemical Information and Modeling*, vol. 63, pp. 4615–4622, 15 2023, issn: 1549-960X. doi: 10.1021/ACS.JCIM.3C00749/. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.3c00749http://dx.doi.org/10.1021/ACS.JCIM.3C00749/ASSET/IMAGES/LARGE/CI3C00749_0003.JPEGhttps://www.ncbi.nlm.nih.gov/pubmed/37531205.

[200] G. Jung, S. G. Jung, and J. M. Cole, "Automatic materials characterization from infrared spectra using convolutional neural networks," *Chemical Science*, vol. 14, pp. 3600–3609, 13 2023, issn: 2041-6539. doi: 10.1039/D2SC05892H. [Online]. Available: http://dx.doi.org/10.1039/D2SC05892H.

[201] F. Sattari, K. Srinivasan, A. Puliyanda, and V. Prasad, "Data fusion-based approach for the investigation of reaction networks in hydrous pyrolysis of biomass," *Industrial and Engineering Chemistry Research*, vol. 62, pp. 4422–4432, 2023, issn: 1520-5045. doi: 10.1021/ACS.IECR.2C04309/. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.iecr.2c04309http://dx.doi.org/10.1021/ACS.IECR.2C04309/ASSET/IMAGES/LARGE/IE2C04309_0012.JPEG.

[202] Y. S. Choi *et al.*, "Pyrolysis reaction networks for lignin model compounds: Unraveling thermal deconstruction of β-o-4 and α-o-4 compounds," *Green Chem.*, vol. 18, pp. 1762–1773, 2016. doi: 10.1039/C5GC02268A. [Online]. Available: http://dx.doi.org/10.1039/C5GC02268A.

[203] R. B. Madsen, H. Zhang, P. Biller, A. H. Goldstein, and M. Glasius, "Characterizing semivolatile organic compounds of biocrude from hydrothermal liquefaction of biomass," *Energy Fuels*, vol. 31, no. 4, pp. 4122–4134, Apr. 2017.

[204] S. S. Toor, L. Rosendahl, and A. Rudolf, "Hydrothermal liquefaction of biomass: A review of subcritical water technologies," *Energy*, vol. 36, pp. 2328–2342, 5 2011, issn: 0360-5442. doi: 10.1016/j.energy.2011.03.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360544211001691http://dx.doi.org/10.1016/j.energy.2011.03.013.

[205] S. Yin and Z. Tan, "Hydrothermal liquefaction of cellulose to bio-oil under acidic, neutral and alkaline conditions," *Applied Energy*, vol. 92, pp. 234–239, 2012, issn: 0306-2619. doi: 10.1016/j.apenergy.2011.10.041. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261911006957http://dx.doi.org/10.1016/j.apenergy.2011.10.041.

[206] A. Puliyanda, K. Srinivasan, Z. Li, and V. Prasad, "Benchmarking chemical neural ordinary differential equations to obtain reaction network-constrained kinetic models from spectroscopic data," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106 690, 2023, issn: 0952-1976. doi: 10.1016/j.engappai.2023.106690. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0952197623008746http://dx.doi.org/10.1016/j.engappai.2023.106690.

[207] J. Zou, L. Zhao, and S. Shi, "Generation of focused drug molecule library using recurrent neural network," *Journal of molecular modeling*, vol. 29, 12 Dec. 2023, issn: 0948-5023. doi: 10.1007/S00894-023-05772-5. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/37932607/.

[208] F. Grisoni, M. Moret, R. Lingwood, and G. Schneider, "Bidirectional molecule generation with recurrent neural networks," *J. Chem. Inf. Model*, vol. 60, p. 8, 2020. doi: 10.1021/acs.jcim.9b00943. [Online]. Available: https://dx.doi.org/10.1021/acs.jcim.9b00943http://dx.doi.org/10.1021/acs.jcim.9b00943.

[209] H. Wang *et al.*, *Graphgan: Graph representation learning with generative adversarial nets*. [Online]. Available: https://cdn.aaai.org/ojs/11872/11872-13-15400-1-2-20201228.pdf.

[210] V. Bagal, R. Aggarwal, P. K. Vinod, and U. D. Priyakumar, "Molgpt: Molecular generation using a transformer-decoder model," *Journal of Chemical Information and Modeling*, vol. 2022, 2021, issn: 1520-5142. doi: 10.1021/acs.jcim.1c00600. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c00600http://dx.doi.org/10.1021/acs.jcim.1c00600https://www.ncbi.nlm.nih.gov/pubmed/34694798.

[211] S. Ekins *et al.*, "Exploiting machine learning for end-to-end drug discovery and development," *Nat. Mater.*, vol. 18, pp. 435–441, 2019, issn: 1476-4660. doi: 10.1038/s41563-019-0338-z. [Online]. Available: http://dx.doi.org/10.1038/s41563-019-0338-zhttps://www.ncbi.nlm.nih.gov/pubmed/31000803.

[212] E. Mazuz, G. Shtar, B. Shapira, and L. Rokach, "Molecule generation using transformers and policy gradient reinforcement learning," *Scientific Reports* |, vol. 13, p. 8799, 2023. doi: 10.1038/s41598-023-35648-w. [Online]. Available: http://dx.doi.org/10.1038/s41598-023-35648-w.

[213] J. Wang *et al.*, "Multi-constraint molecular generation based on conditional transformer, knowledge distillation and reinforcement learning," *Nature Machine Intelligence 2021 3:10*, vol. 3, pp. 914–922, 2021, issn: 2522-5839. doi: 10.1038/s42256-021-00403-1. [Online]. Available: https://www.nature.com/articles/s42256-021-00403-1http://dx.doi.org/10.1038/s42256-021-00403-1.

[214] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, "Optimization of molecules via deep reinforcement learning," *Scientific Reports 2019 9:1*, vol. 9, pp. 1–10, 2019, issn: 2045-2322. doi: 10.1038/s41598-019-47148-x. [Online]. Available: https://www.nature.com/articles/s41598-019-47148-xhttp://dx.doi.org/10.1038/s41598-019-47148-xhttps://www.ncbi.nlm.nih.gov/pubmed/31341196.

[215] M. Goel, S. Raghunathan, S. Laghuvarapu, and U. D. Priyakumar, "Molegular: Molecule generation using reinforcement learning with alternating rewards," *Journal of Chemical Information and Modeling*, vol. 61, pp. 5815–5826, 2021, issn: 1549-960X. doi: 10.1021/acs.jcim.1c01341. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.1c01341http://dx.doi.org/10.1021/acs.jcim.1c01341https://www.ncbi.nlm.nih.gov/pubmed/34866384.

[216] J. You, B. Liu, R. Ying, V. Pande, and J. Leskovec, "Graph convolutional policy network for goal-directed molecular graph generation,"

[217] B. Samanta *et al.*, "Nevae: A deep generative model for molecular graphs *," *Journal of Machine Learning Research*, vol. 21, pp. 1–33, 2020.

[218] X. Liu, K. Ye, H. W. T. van Vlijmen, A. P. IJzerman, and G. J. P. van Westen, "Drugex v3: Scaffold-constrained drug design with graph transformer-based reinforcement learning," *Journal of Cheminformatics*, vol. 15, pp. 1–14, 2023, issn: 1758-2946. doi: 10.1186/S13321-023-00694-Z. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-023-00694-zhttp://dx.doi.org/10.1186/S13321-023-00694-Z.

[219] J. Choi, S. Seo, and S. Park, "Coma: Efficient structure-constrained molecular generation using contractive and margin losses," *Journal of Cheminformatics*, vol. 15, pp. 1–13, 2023, issn: 1758-2946. doi: 10.1186/S13321-023-00679-Y/. [Online]. Available: http://dx.doi.org/10.1186/S13321-023-00679-Y/FIGURES/4.

[220] M. Langevin, H. Minoux, M. Levesque, and M. Bianciotto, "Scaffold-constrained molecular generation," *Journal of Chemical Information and Modeling*, vol. 60, pp. 5637–5646, 2020, issn: 1549-960X. doi: 10.1021/ACS.JCIM.0C01015/. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.0c01015http://dx.doi.org/10.1021/ACS.JCIM.0C01015/https://www.ncbi.nlm.nih.gov/pubmed/33301333.

[221] H. Qian, C. Lin, D. Zhao, S. Tu, and L. Xu, "Alphadrug: Protein target specific de novo molecular generation," *PNAS Nexus*, vol. 1, pp. 1–12, 2022, issn: 2752-6542. doi: 10.1093/PNASNEXUS/PGAC227. [Online]. Available: https://dx.doi.org/10.1093/pnasnexus/pgac227http://dx.doi.org/10.1093/PNASNEXUS/PGAC227.

[222] W. Zhung, H. Kim, and W. Y. Kim, "A protein-ligand interaction-focused 3d molecular generative framework for generalizable structure-based drug design," *ChemRxiv*, 2023. doi: 10.26434/CHEMRXIV-2023-JSJWX. [Online]. Available: http://dx.doi.org/10.26434/CHEMRXIV-2023-JSJWX.

[223] S. Li *et al.*, "Ls-molgen: Ligand-and-structure dual-driven deep reinforcement learning for target-specific molecular generation improves binding affinity and novelty," *Journal of Chemical Information and Modeling*, vol. 63, pp. 4207–4215, 2023, issn: 1549-960X. doi: 10.1021/ACS.JCIM.3C00587. [Online]. Available: https://pubs.acs.org/doi/full/10.1021/acs.jcim.3c00587http://dx.doi.org/10.1021/ACS.JCIM.3C00587https://www.ncbi.nlm.nih.gov/pubmed/37341350.

[224] O. Méndez-Lucio, B. Baillif, D. A. Clevert, D. Rouquié, and J. Wichard, "De novo generation of hit-like molecules from gene expression signatures using artificial intelligence," *Nature Communications 2020 11:1*, vol. 11, pp. 1–10, 1 Jan. 2020, issn: 2041-1723. doi: 10.1038/s41467-019-13807-w. [Online]. Available: https://www.nature.com/articles/s41467-019-13807-w.

[225] Z. Huang, M. S. Chen, C. P. Woroch, T. E. Markland, and M. W. Kanan, "A framework for automated structure elucidation from routine nmr spectra," *Chemical Science*, vol. 12, pp. 15 329–15 338, 2021, issn: 2041-6539. doi: 10.1039/D1SC04105C. [Online]. Available: http://dx.doi.org/10.1039/D1SC04105C.

[226] M. Alberts, T. Laino, and A. C. Vaucher, "Leveraging infrared spectroscopy for automated structure elucidation," 2023. doi: 10.26434/CHEMRXIV-2023-5V27F. [Online]. Available: http://dx.doi.org/10.26434/CHEMRXIV-2023-5V27F.

[227] A. Lo, R. Pollice, A. Nigam, A. D. White, M. Krenn, and A. Aspuru-Guzik, "Recent advances in the self-referencing embedded strings (selfies) library," *Digital Discovery*, vol. 2, pp. 897–908, 4 2023. doi: 10.1039/D3DD00044C. [Online]. Available: http://dx.doi.org/10.1039/D3DD00044C.

[228] P. Schwaller *et al.*, "Mapping the space of chemical reactions using attention-based neural networks," *Nature Machine Intelligence*, vol. 3, pp. 144–152, 2 2021, issn: 2522-5839. doi: 10.1038/s42256-020-00284-w. [Online]. Available: https://doi.org/10.1038/s42256-020-00284-w.

[229] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *Proceedings of the 34th International Conference on Machine Learning*, D. Precup and Y. W. Teh, Eds., ser. Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 1263–1272. [Online]. Available: https://proceedings.mlr.press/v70/gilmer17a.html.

[230] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 318–327, 2017, issn: 1939-3539. doi: 10.1109/TPAMI.2018.2858826. [Online]. Available: https://arxiv.org/abs/1708.02002v2http://dx.doi.org/10.1109/TPAMI.2018.2858826https://www.ncbi.nlm.nih.gov/pubmed/30040631.

[231] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017. [Online]. Available: https://arxiv.org/abs/1701.07875v3.

[232] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans montreal institute for learning algorithms," 2017. [Online]. Available: https://github.com/igul222/improved_wgan_training..

[233] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 311–318. doi: 10.3115/1073083.1073135. [Online]. Available: https://aclanthology.org/P02-1040.

[234] K. Rajan, A. Zielesny, and C. Steinbeck, "Stout: Smiles to iupac names using neural machine translation," *Journal of Cheminformatics 2021 13:1*, vol. 13, pp. 1–14, 2021, issn: 1758-2946. doi: 10.1186/S13321-021-00512-4. [Online]. Available: https://jcheminf.biomedcentral.com/articles/10.1186/s13321-021-00512-4http://dx.doi.org/10.1186/S13321-021-00512-4.

[235] R. Winter, F. Montanari, F. Noé, and D. A. Clevert, "Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations," *Chemical Science*, vol. 10, pp. 1692–1701, 2019, issn: 2041-6539. doi: 10.1039/C8SC04175J. [Online]. Available: http://dx.doi.org/10.1039/C8SC04175J.

[236] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008. [Online]. Available: http://jmlr.org/papers/v9/vandermaaten08a.html.

[237] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96, Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[238] J. Schuler *et al.*, "Hydrothermal liquefaction of lignin," *Journal of Biomaterials and Nanobiotechnology*, vol. 8, pp. 96–108, 1 Dec. 2016, issn: 2158-7027. doi: 10.4236/JBNB.2017.81007. [Online]. Available: http://www.scirp.org/journal/PaperInformation.aspx?PaperID=73655http://www.scirp.org/Journal/Paperabs.aspx?paperid=73655https://www.scirp.org/journal/paperinformation.aspx?paperid=73655.

[239] B. Ciuffi, M. Loppi, A. M. Rizzo, D. Chiaramonti, and L. Rosi, "Towards a better understanding of the htl process of lignin-rich feedstock," *Scientific Reports*, vol. 11, p. 15 504, 1 2021, issn: 2045-2322. doi: 10.1038/s41598-021-94977-w. [Online]. Available: https://doi.org/10.1038/s41598-021-94977-w.

[240] B. Qiu, J. Shi, W. Hu, J. Gao, S. Li, and H. Chu, "Construction of hydrothermal liquefaction system for efficient production of biomass-derived furfural: Solvents, catalysts and mechanisms," *Fuel*, vol. 354, p. 129 278, Dec. 2023, issn: 0016-2361. doi: 10.1016/J.FUEL.2023.129278.

[241] S. Yin, Y. Pan, and Z. Tan, "Hydrothermal conversion of cellulose to 5-hydroxymethyl furfural," *International Journal of Green Energy*, vol. 8, pp. 234–247, 2 Feb. 2011, issn: 15435075. doi: 10.1080/15435075.2010.548888. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/15435075.2010.548888.

[242] D. C. Boffito and D. F. Rivas, "Process intensification connects scales and disciplines towards sustainability," *The Canadian Journal of Chemical Engineering*, vol. 98, pp. 2489–2506, 12 2020, issn: 0008-4034. doi: 10.1002/cjce.23871. [Online]. Available: https://doi.org/10.1002/cjce.23871http://dx.doi.org/10.1002/cjce.23871.

[243] K. Wang *et al.*, "Kinetic and data-driven reaction analysis for pharmaceutical process development," *Industrial & Engineering Chemistry Research*, vol. 59, pp. 2409–2421, 6 2020, issn: 0888-5885. doi: 10.1021/acs.iecr.9b03578. [Online]. Available: https://doi.org/10.1021/acs.iecr.9b03578http://dx.doi.org/10.1021/acs.iecr.9b03578.

[244] J. Yue, J. C. Schouten, and T. A. Nijhuis, "Integration of microreactors with spectroscopic detection for online reaction monitoring and catalyst characterization," *Industrial & Engineering Chemistry Research*, vol. 51, pp. 14 583–14 609, 45 2012, issn: 0888-5885. doi: 10.1021/ie301258j. [Online]. Available: https://doi.org/10.1021/ie301258jhttp://dx.doi.org/10.1021/ie301258j.

[245] H. Fleischer, V. Q. Do, and K. Thurow, "Online measurement system in reaction monitoring for determination of structural and elemental composition using mass spectrometry," *SLAS TECHNOLOGY: Translating Life Sciences Innovation*, vol. 24, pp. 330–341, 3 2019, issn: 2472-6303. doi: 10.1177/2472630318813838. [Online]. Available: https://doi.org/10.1177/2472630318813838http://dx.doi.org/10.1177/2472630318813838.

[246] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, "Neural networks for the prediction of organic chemistry reactions," *ACS Central Science*, vol. 2, pp. 725–732, 10 2016, issn: 2374-7943. doi: 10.1021/acscentsci.6b00219. [Online]. Available: http://dx.doi.org/10.1021/acscentsci.6b00219.

[247] M. Pellegrini, J. Vohradsky, P. Loskot, K. Atitey, and L. Mihaylova, "Comprehensive review of models and methods for inferences in bio-chemical reaction networks," *Frontiers in Genetics | www. frontiersin. org*, vol. 1, p. 549, 2019. doi: 10.3389/fgene.2019.00549. [Online]. Available: www.frontiersin.orghttp://dx.doi.org/10.3389/fgene.2019.00549.

[248] 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, pp. 1536–1547, 3 2011, issn: 0888-5885. doi: 10.1021/ie1015552. [Online]. Available: https://doi.org/10.1021/ie1015552http://dx.doi.org/10.1021/ie1015552.

[249] F. Santosa and B. Weitz, "An inverse problem in reaction kinetics," *Journal of Mathematical Chemistry*, vol. 49, pp. 1507–1520, 8 2011, issn: 1572-8897. doi: 10.1007/s10910-011-9835-2. [Online]. Available: https://doi.org/10.1007/s10910-011-9835-2http://dx.doi.org/10.1007/s10910-011-9835-2.

[250] A. Papachristodoulou and B. Recht, "Determining interconnections in chemical reaction networks," in *2007 American Control Conference*, ser. 2007 American Control Conference, 2007, pp. 4872–4877. doi: 10.1109/ACC.2007.4283084. [Online]. Available: https://doi.org/10.1109/ACC.2007.4283084.

[251] 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, p. 113 121, 11 2019, issn: 1054-1500. doi: 10.1063/1.5120598. [Online]. Available: https://doi.org/10.1063/1.5120598http://dx.doi.org/10.1063/1.5120598.

[252] S. C. Burnham, D. P. Searson, M. J. Willis, and A. R. Wright, "Inference of chemical reaction networks," *Chemical Engineering Science*, vol. 63, pp. 862–873, 4 2008, issn: 0009-2509. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0009250907007786.

[253] M. Hoffmann, C. Fröhner, and F. Noé, "Reactive sindy: Discovering governing reactions from concentration data," *The Journal of Chemical Physics*, vol. 150, p. 025 101, 2 2019, issn: 0021-9606. doi: 10.1063/1.5066099. [Online]. Available: https://doi.org/10.1063/1.5066099http://dx.doi.org/10.1063/1.5066099.

[254] D. F. Anderson, B. Joshi, and A. Deshpande, "On reaction network implementations of neural networks," *Journal of The Royal Society Interface*, vol. 18, p. 20 210 031, 177 2021. doi: 10.1098/rsif.2021.0031. [Online]. Available: https://doi.org/10.1098/rsif.2021.0031http://dx.doi.org/10.1098/rsif.2021.0031.

[255] B. Sen and S. Menon, "Representation of chemical kinetics by artificial neural networks for large eddy simulations," in American Institute of Aeronautics and Astronautics, 2007, 0. doi: 10.2514/6.2007-5635. [Online]. Available: https://doi.org/10.2514/6.2007-5635http://dx.doi.org/10.2514/6.2007-5635.

[256] 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.

[257] N. Shenvi, J. M. Geremia, and H. Rabitz, "Efficient chemical kinetic modeling through neural network maps," *The Journal of Chemical Physics*, vol. 120, pp. 9942–9951, 21 2004, issn: 0021-9606. doi: 10.1063/1.1718305. [Online]. Available: https://doi.org/10.1063/1.1718305http://dx.doi.org/10.1063/1.1718305.

[258] 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, pp. 571–574, 7 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.CO2-5http://dx.doi.org/10.1002/(SICI)1521-4125(199907)22:7<571::AID-CEAT571>3.0.CO2-5.

[259] G. S. Gusmão, A. P. Retnanto, S. C. da Cunha, and A. J. Medford, *Kinetics-informed neural networks*, 2020. arXiv: 2011.14473 [cs.LG].

[260] 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, pp. 8098–8106, 36 2021, issn: 1089-5639. doi: 10.1021/acs.jpca.1c05102. [Online]. Available: https://doi.org/10.1021/acs.jpca.1c05102http://dx.doi.org/10.1021/acs.jpca.1c05102.

[261] 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, Jan. 1996, issn: 0098-1354. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0098135495002316.

[262] 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, pp. 323–334, 4 2000, issn: 0255-2701. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0255270199000938.

[263] 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.

[264] 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, p. 093 122, 9 2021, issn: 1054-1500. doi: 10.1063/5.0060697. [Online]. Available: https://doi.org/10.1063/5.0060697http://dx.doi.org/10.1063/5.0060697.

[265] 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, pp. 1082–1092, 4 2021, issn: 1089-5639. doi: 10.1021/acs.jpca.0c09316. [Online]. Available: https://doi.org/10.1021/acs.jpca.0c09316http://dx.doi.org/10.1021/acs.jpca.0c09316.

[266] 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, pp. 6358–6368, 29 2020, issn: 1520-6106. doi: 10.1021/acs.jpcb.0c04299. [Online]. Available: https://doi.org/10.1021/acs.jpcb.0c04299http://dx.doi.org/10.1021/acs.jpcb.0c04299.

[267] 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.

[268] 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.

[269] O. Levenspiel, *Chemical reaction engineering*. John Wiley, 1972, isbn: 9780852265130. [Online]. Available: https://books.google.ca/books?id=1eBXYgEACAAJ.

[270] C. J. Giunta, "What's in a name? amount of substance, chemical amount, and stoichiometric amount," *Journal of Chemical Education*, vol. 93, pp. 583–586, 4 2016, issn: 0021-9584. doi: 10.1021/acs.jchemed.5b00690. [Online]. Available: https://doi.org/10.1021/acs.jchemed.5b00690http://dx.doi.org/10.1021/acs.jchemed.5b00690.

[271] 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, pp. 52–55, 1 2016, issn: 0306-5251. doi: 10.1111/bcp.12721. [Online]. Available: https://doi.org/10.1111/bcp.12721http://dx.doi.org/10.1111/bcp.12721.

[272] N. Kriegeskorte and T. Golan, "Neural network models and deep learning," *Current Biology*, vol. 29, R231–R236, 7 2019, issn: 0960-9822. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982219302040.

[273] G. S. Mittal, "Chapter 18 - artificial neural network (ann) based process modeling," in Academic Press, 2013, pp. 467–473, isbn: 9780123858818. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123858818000185.

[274] M. M. Bejani and M. Ghatee, "A systematic review on overfitting control in shallow and deep neural networks," *Artificial Intelligence Review*, vol. 54, pp. 6391–6438, 8 2021, issn: 1573-7462. doi: 10.1007/s10462-021-09975-1. [Online]. Available: https://doi.org/10.1007/s10462-021-09975-1http://dx.doi.org/10.1007/s10462-021-09975-1.

[275] D. C. Psichogios and L. H. Ungar, "A hybrid neural network-first principles approach to process modeling," *AIChE Journal*, vol. 38, pp. 1499–1511, 10 1992, issn: 0001-1541. doi: 10.1002/aic.690381003. [Online]. Available: https://doi.org/10.1002/aic.690381003http://dx.doi.org/10.1002/aic.690381003.

[276] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," *Advances in Neural Information Processing Systems*, 2018.

[277] M Raissi, P Perdikaris, and G. E. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019, issn: 0021-9991. doi: 10.1016/j.jcp.2018.10.045. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021999118307125http://dx.doi.org/10.1016/j.jcp.2018.10.045.

[278] R. T. Q. Chen, B. Amos, and M. Nickel, "Learning neural event functions for ordinary differential equations," *International Conference on Learning Representations*, 2021.

[279] B. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J Kutz, and S. Brunton, "Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data," *Journal of Open Source Software*, vol. 5, p. 2104, 49 2020. doi: 10.21105/joss.02104. [Online]. Available: https://doi.org/10.21105/joss.02104http://dx.doi.org/10.21105/joss.02104.

[280] A. A. Kaptanoglu *et al.*, "Pysindy: A comprehensive python package for robust sparse system identification," *Journal of Open Source Software*, vol. 7, p. 3994, 69 2022. doi: 10.21105/joss.03994. [Online]. Available: https://doi.org/10.21105/joss.03994http://dx.doi.org/10.21105/joss.03994.

[281] 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, pp. 1099–1104, 4 2007, issn: 0888-5885. doi: 10.1021/ie060781y. [Online]. Available: https://doi.org/10.1021/ie060781yhttp://dx.doi.org/10.1021/ie060781y.

[282] J. Dalgaard, T. Kocka, and J. Pena, *On local optima in learning bayesian networks*, ISSN ; -, May 2003.

[283] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 618–626, 2017, issn: 1550-5499. doi: 10.1109/ICCV.2017.74. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.74.

[284] B. D. Mistry, *A handbook of Spectroscopic Data Chemistry (UV, IR, PMR, 13CNMR and Mass Spectroscopy)*. Oxford Book Company, 2009, isbn: 9788189473860. doi: 10.1063/1.2719251. [Online]. Available: http://dx.doi.org/10.1063/1.2719251.

Appendix A: Chapter 2

A.1 Bayesian hierarchical clustering

Bayesian hierarchical clustering is applied as a pre-processing to the process spectroscopic measurements to reduce the dimensionality of the data. The Bayesian hierarchical clustering algorithm considers each wavenumber as an individual cluster with the probability of combining to form a cluster being set equal to the α hyper-parameter of the Dirichlet distribution. This prior of clustering is updated in a recursive fashion as the agglomerative clustering combines more sub-trees to form larger clusters as ,

$$d_k = \alpha\Gamma(n_k) + d_{left,k}d_{right,k} \tag{A.1a}$$

$$\pi_k = \frac{\alpha\Gamma(n_k)}{d_k} \tag{A.1b}$$

n_k indicates the number of data points in the k^{th} cluster, $d_{left,k}$ and $d_{right,k}$ represent the d values associated with the left and right sub-trees in the dendrogram of the clustering as depicted in Figure A.1.

The clustered wavenumber regions are assigned to functional groups through comparison with standard handbooks for IR spectroscopy detection. The choice of number of clusters is based on expert knowledge and is restricted to 6 to ensure enough resolution of the individual functional groups.

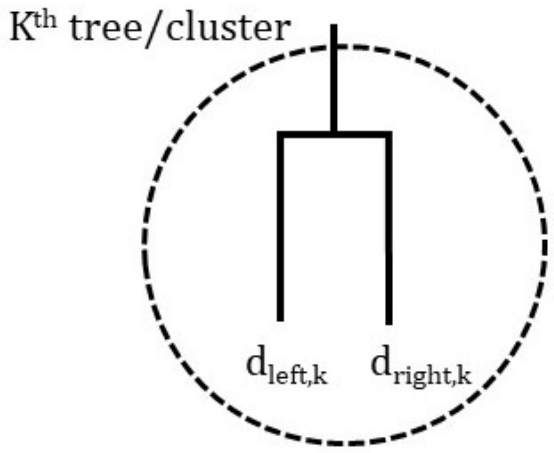


Figure A.1: Tree structure of Bayesian hierarchical clustering

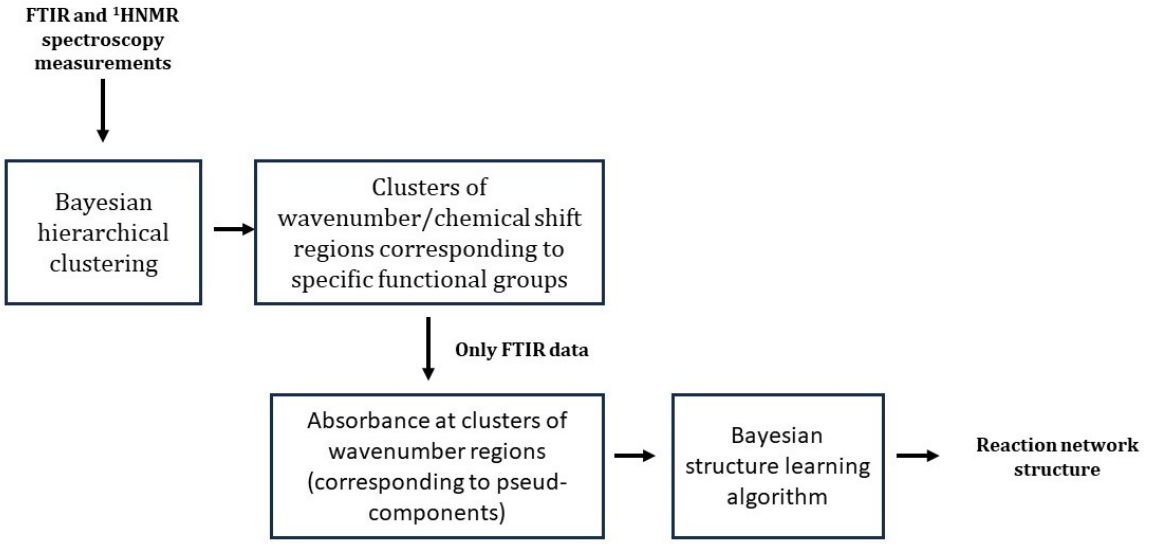
A.2 Bayesian structure learning

The absorbance values of each cluster is used as the nodes of the Bayesian structure learning algorithm. Only FTIR spectra were used for the structure learning task as the resolution of the ^{1}H-NMR data was insufficient for reliable determination of reaction network. The Bayesian Information Score used in the structure determination can be given as,

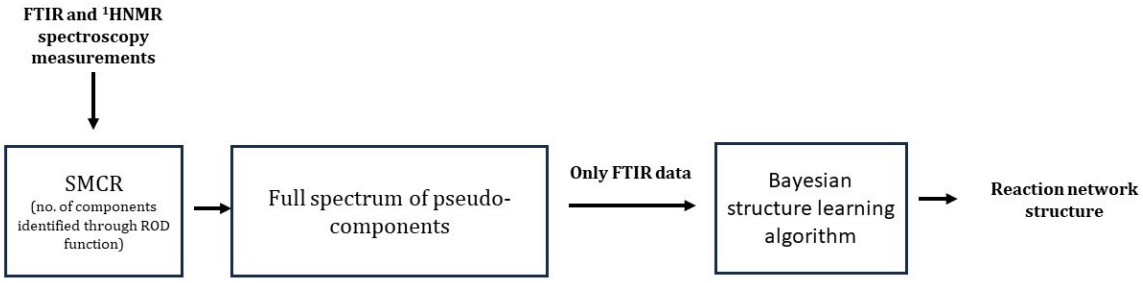
$$\text{score}_{\text{BIC}} = M \sum_{i=1}^{n} I_p(X_i; Pa_{X_i}) - M \sum_{i=1}^{n} H(X_i) - \log\left(\frac{M}{2}\text{DIM(DAG)}\right) \qquad \text{(A.2)}$$

where M is the number of samples, Pa_{X_i} represents the parents of X_i node, I_p is the mutual information function and H is the entropy function. The BIC score quantifies the Markov property of DAG generated with the first two terms and penalises for the graph complexity in the last term.

In the case of SMCR, the absorbance of the entire pseudo-component spectra is used for the structure determination. A flow sheet of the process for structure from Bayesian hierarchical clustering and SMCR are given in Figure A.2a and A.2b respectively.



(a)

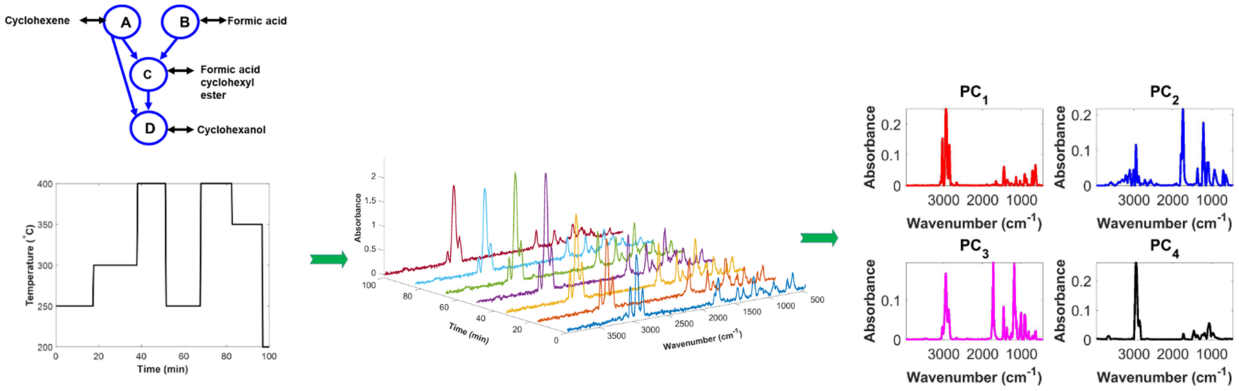


(b)

Figure A.2: Flowchart of Bayesian network construction for a) Bayesian hierarchical clustering b) SMCR

Appendix B: Chapter 3

B.1 Synthetic data generation



$$\dot{C}_A = -K_1 {C_A}^a {C_B}^b - K_2 {C_A}^a {C_C}^c$$
$$\dot{C}_B = -K_1 {C_A}^a {C_B}^b$$
$$\dot{C}_C = K_1 {C_A}^a {C_B}^b - K_2 {C_A}^a {C_C}^c$$
$$\dot{C}_D = K_2 {C_A}^a {C_C}^c$$

Where a=2,b=1,c=1, K_1=10, K_2=5

- Temperature dependence is accounted by the Arrhenius law

$$K_i = K_{i0} e^{-\frac{1}{T}}$$

Figure B.1: Scheme for synthetic data generation

B.2 Comparison of classifier with literature

Table B.1: Comparison of classification metrics with Wang et.al [199]

Functional Group	Accuracy		Precision		Recall		F1 Score		Specificity	
	1D-CNN	Wang et.al	1D-CNN	Wang et.al	1D-CNN	Wang et.al	1D-CNN	Wang et.al	1D-CNN	Wang et.al
Alkane	0.973	0.939	0.980	0.94	0.988	0.94	0.984	0.94	0.892	0.865
Alkene	0.970	0.954	0.904	0.854	0.864	0.791	0.884	0.822	0.986	0.989
Alkyne	0.996	0.993	0.945	0.934	0.877	0.875	0.909	0.904	0.999	0.999
Arene	0.968	0.974	0.975	0.978	0.971	0.976	0.973	0.977	0.964	0.97
Ketone	0.979	0.978	0.889	0.893	0.875	0.868	0.883	0.880	0.989	0.99
Aldehyde	0.996	0.997	0.943	0.973	0.879	0.900	0.910	0.935	0.999	0.999
Ester	0.984	0.983	0.932	0.933	0.931	0.913	0.932	0.923	0.991	0.992
Acid	0.990	0.989	0.953	0.923	0.903	0.928	0.927	0.925	0.997	0.994
Alcohol	0.989	0.975	0.919	0.956	0.844	0.951	0.881	0.953	0.996	0.984
Amine	0.973	0.966	0.897	0.91	0.900	0.871	0.898	0.890	0.984	0.982
Halide	0.972	0.912	0.901	0.853	0.850	0.828	0.875	0.840	0.988	0.945
Ether	0.966	0.963	0.932	0.879	0.932	0.862	0.932	0.870	0.978	0.98
Nitro	0.989	0.994	0.927	0.952	0.882	0.934	0.904	0.943	0.996	0.997

Table B.2: Comparison of overall classification metrics with Fine et.al [181]

Model	Molecular Precision	Molecular Recall	Molecular F1 score	Molecular Perfection Rate
MLP	0.9459	0.9392	0.943	0.815
CNN w/o recon	0.939	0.899	0.919	0.72
CNN w/ recon BCE	0.949	0.949	0.949	0.849
CNN w/ recon WBCE	0.9323	0.9533	0.943	0.83
Fine et.al	-	-	0.931	0.749

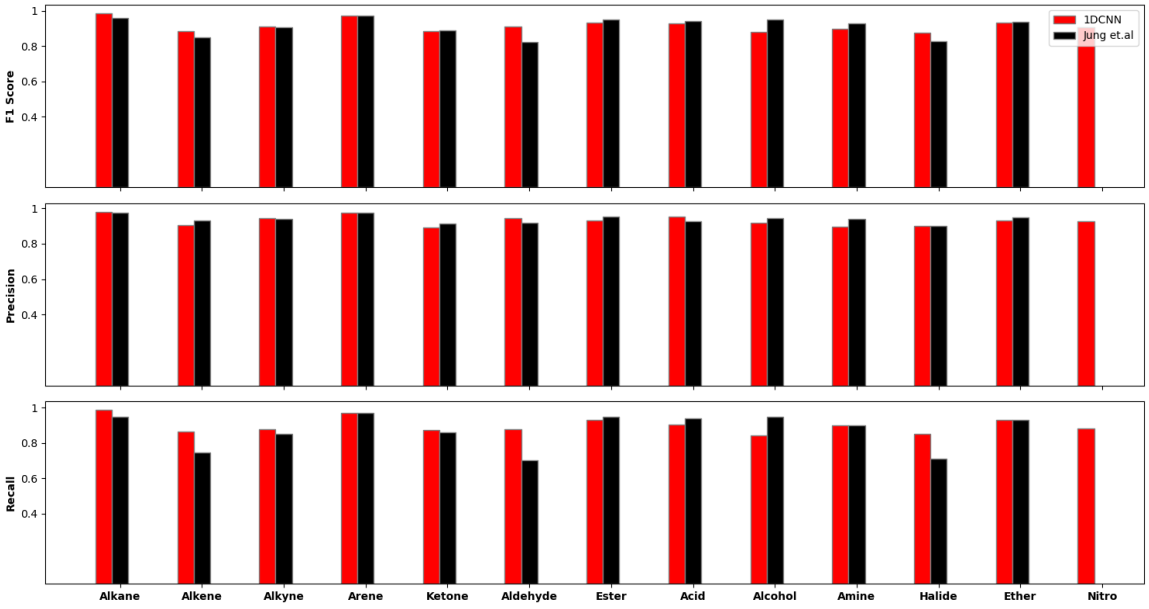


Figure B.2: Comparison of classification metrics with Jung et.al[200]

B.3 Deconvolution of Synthetic data

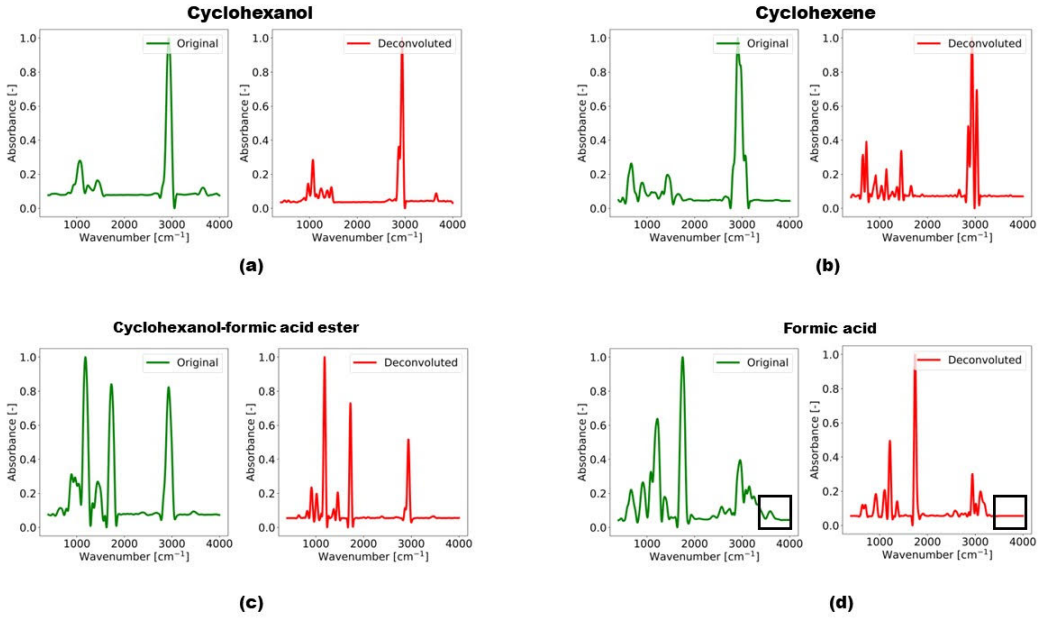


Figure B.3: Comparison of original and deconvoluted spectrum for synthetic data

B.4 Deconvolution of biomass data with water

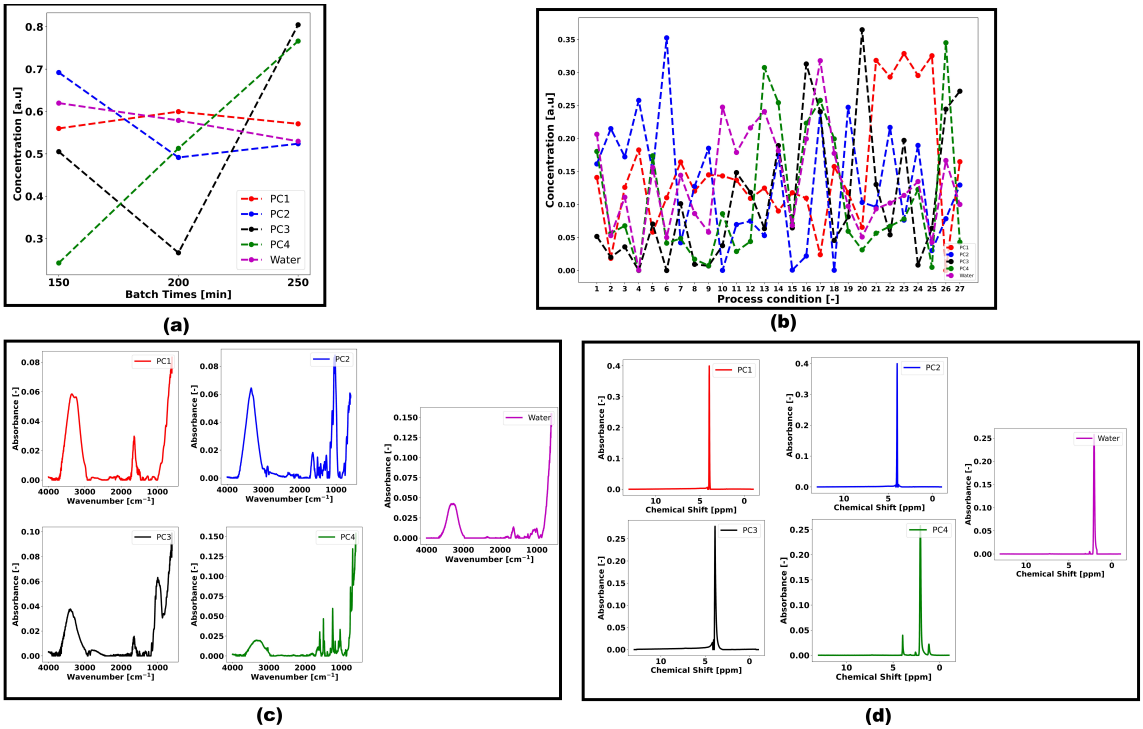


Figure B.4: Deconvolution of biomass HTL data: (a)Projection along residence time mode, (b) Projection along process condition mode, (c) Resolved FTIR spectra of each PC + Water, (d) Resolved ^{1}H-NMR spectra of each PC + Water

B.5 Grad-CAM

In order to infer the wavenumbers responsible for classification into different functional groups, saliency analysis, specifically, the computation of Grad-CAM (Class Activation Maps) was performed on the neural network [283]. Grad-CAM uses the gradient of the score of the output with respect to a certain class (or a combination of classes in this case) in determining a localization map of the features responsible for classification into those class(es). The global averaged value (for Z number of neurons (indexed by i) in the convolutional layer) of the gradient of the output (y_c) labelled as a particular set of classes (c) with respect to the k^{th} activation (A_k) of the convolution layer provides a weighting matrix α_k^c which, when multiplied by the activations, generates the Grad-CAM for the particular class.

$$\alpha_k^c = \frac{1}{Z} \sum_i \frac{\partial y_c}{\partial A_k^i} \tag{B.1a}$$

$$\text{Grad-CAM} = \text{ReLU}\left(\sum_k \alpha_k^c A_k \right) \tag{B.1b}$$

Typically used in image-classification problems, Grad-CAM can be applied to the 1-D CNN problem as well to understand the importance of certain wavenumber regions in identifying a particular functional group. The Grad-CAM values for a simple classification (methane) and for classification of multiple functional groups (methanol, acetic acid) are presented in Figure B.5. As seen from Panel (a), the wavenumber regions highlighted by Grad-CAM in the classification of methane fall in the wavenumbers corresponding to the sp^3 C-H stretch (2850-3000 cm^{-1}) and peaks corresponding to sp^3 C-H bend frequencies occur at 1380-1460 cm^{-1} . In the case of methanol, the O-H and C-O stretch frequencies fall in wavenumber regions of 3400-3600 cm^{-1} and 1000-1260 cm^{-1}, which are highlighted by the saliency analysis along with the sp^3 C-H bends and stretches. For the case of acetic acid, the saliency map indicates that the CNN gives more importance to features (peaks) corresponding to the C=O stretch (1700-1730 cm^{-1}), O-H stretch (2400-3400 cm^{-1}), and

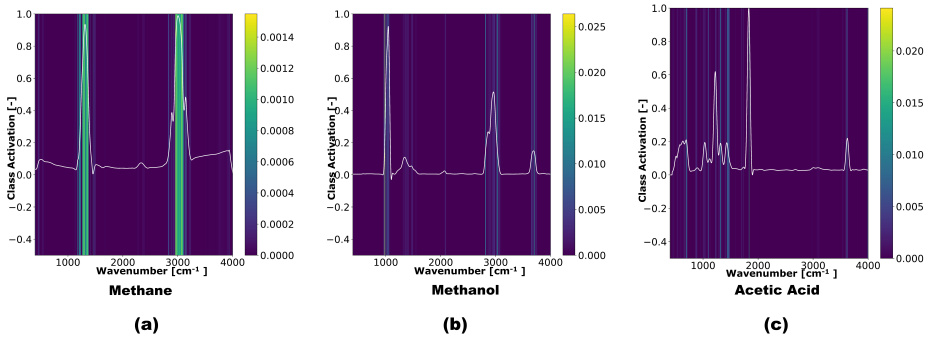


Figure B.5: Saliency Analysis: (a) Grad-CAM for methane, (b) Grad-CAM for methanol, (c) Grad-CAM for acetic acid

C-O stretch (1210-1320 $cm^{-}1$)[284].

B.6 CNN features

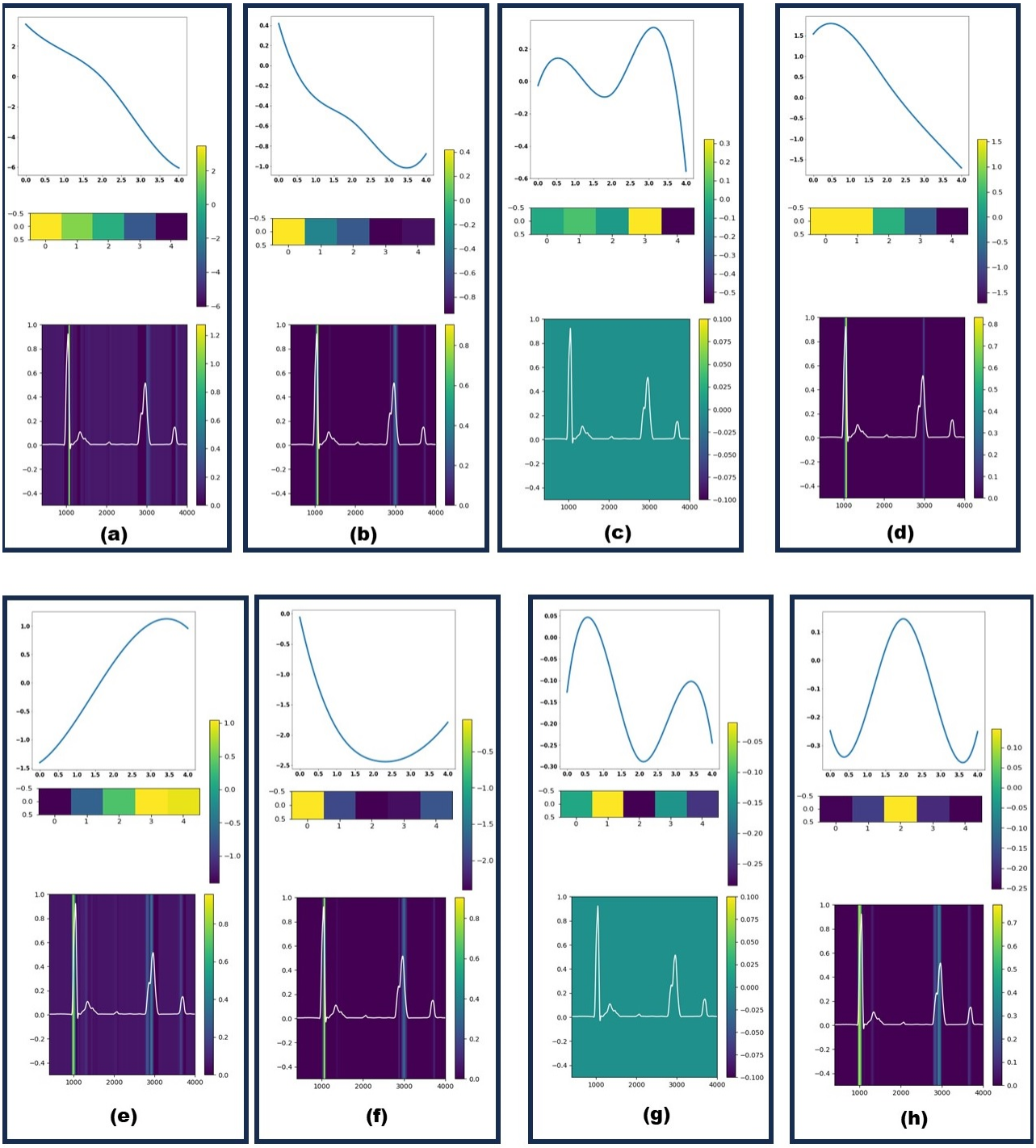


Figure B.6: Filters learnt by the CNN

B.7 Distribution of training data

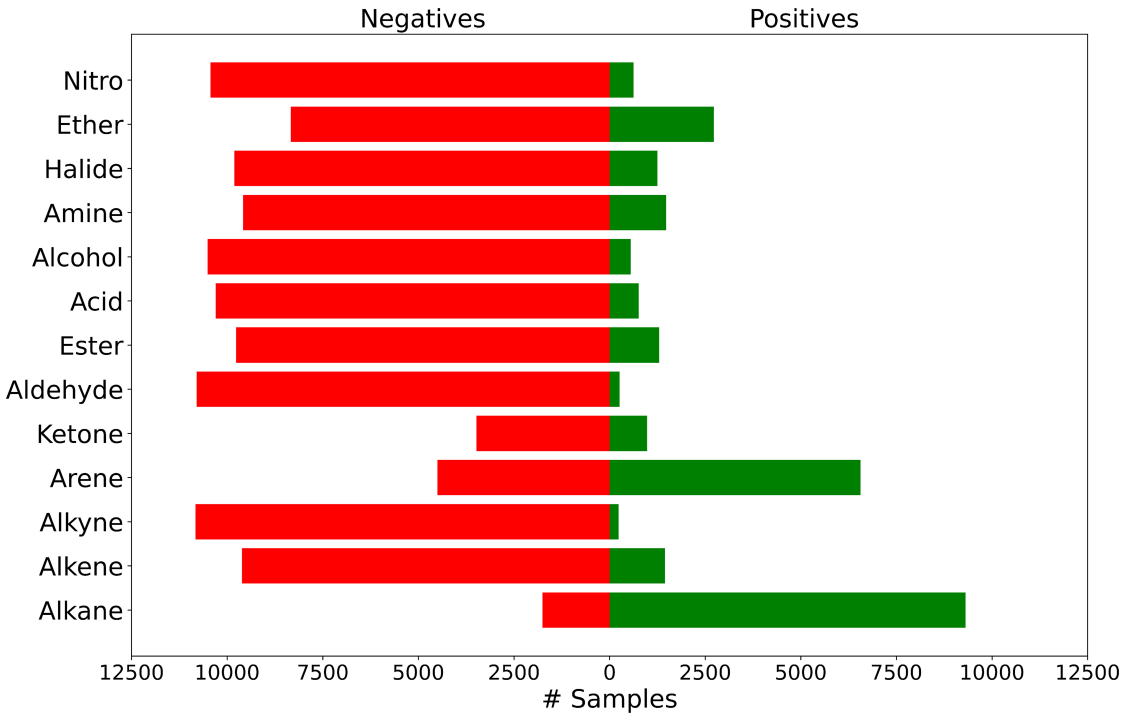


Figure B.7: Distribution of samples for CNN training

Appendix C: Chapter 4

C.1 Dataset characteristics

C.1.1 Distribution of sequence length of training samples

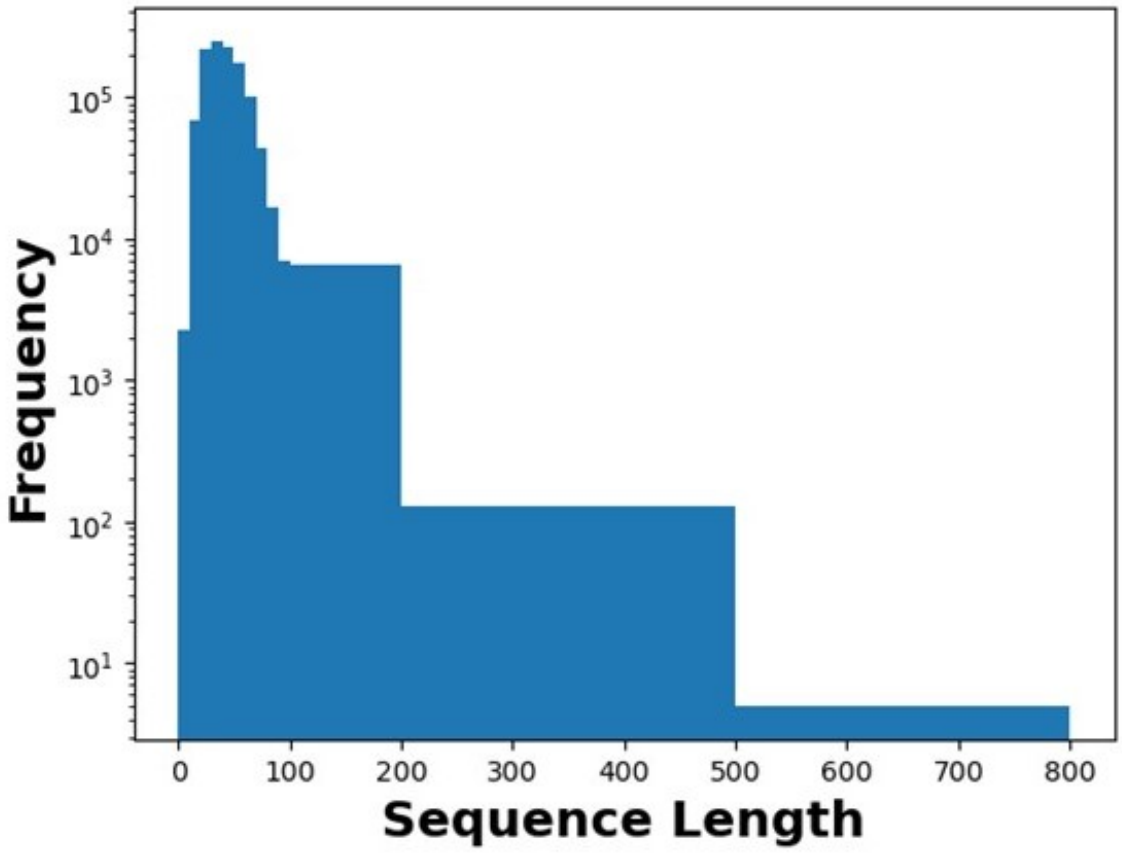


Figure C.1: Histogram of distribution of sequence length across entire dataset

C.1.2 Token frequency for strings used in prediction

Token	Frequency
[C]	19799996
[=C]	6714258
[Ring1]	4220684
[Branch1]	3874419
[=Branch1]	2528569
[N]	2363501
[O]	2136863
[=O]	1505421
[Ring2]	1236100
[Branch2]	1033005
[=N]	928818
[F]	685692
[#Branch1]	554077
[S]	532575
[=Branch2]	529263
[#Branch2]	379404
[Cl]	334948
[#C]	275349
[P]	197259
[NH1]	126625
[=Ring1]	111307
[Br]	104477
[O-1]	82977
[N+1]	65032
[#N]	59262
[Si]	34220
[=Ring2]	29116
[=S]	21773
[I]	21480
[=N+1]	18460
[B]	10569
[=N-1]	4147
[N-1]	2347

(a) Frequency of tokens in SMILES

Token	Frequency
c	12895831
C	9865284
(	5595427
)	5595427
O	3279613
1	3252674
2	2328722
=	1946997
N	1555030
n	1260670
3	1038570
F	685811
-	370405
4	335940
Cl	334964
S	239324
[nH]	126628
s	113713
#	111435
Br	104479
[O-]	82994
o	81561
[N+]	77774
5	74430
[Si]	34225
I	21526
P	13349
6	12782
B	10575
[N-]	6417
[n+]	6070
7	2160

(b) Frequency of tokens in SELFIES

Table C.1: Frequency of occurrence of tokens in SMILES and SELFIES strings

C.1.3 Features used for MPNN

Node features
Atomic number
Degree of connectivity
Formal charge
Hybridization
Aromaticity

(a) Node features

Edge features
Bond type
Ring membership
Aromaticity

(b) Edge features

Table C.2: Node and edge features for MPNNs

C.2 BLEU Scores for testing data

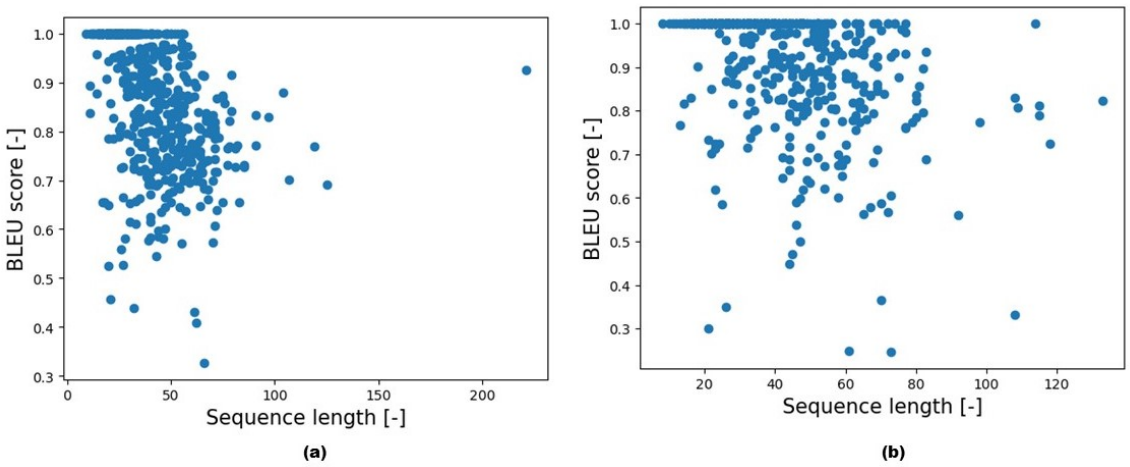


Figure C.2: BLEU scores for samples (a) *Graph2SMILES* (b) *Graph2SELFIES*

C.3 Characteristics of DBSCAN clusters

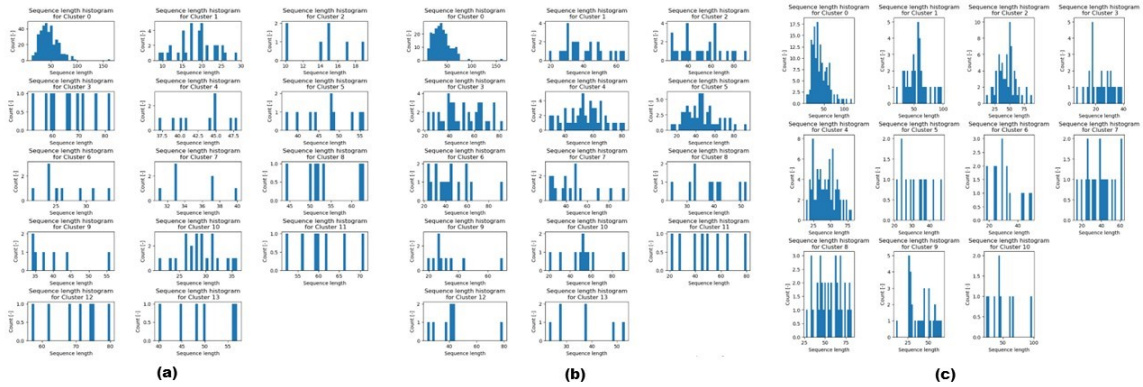


Figure C.3: Sequence length distribution across DBSCAN clusters for (a) RNN-translator t-SNE encodings, (b) *Graph2SMILES* translator t-SNE encodings and (c) *Graph2SELFIES* translator encodings.

C.4 Details on spectrum-constrained GAN

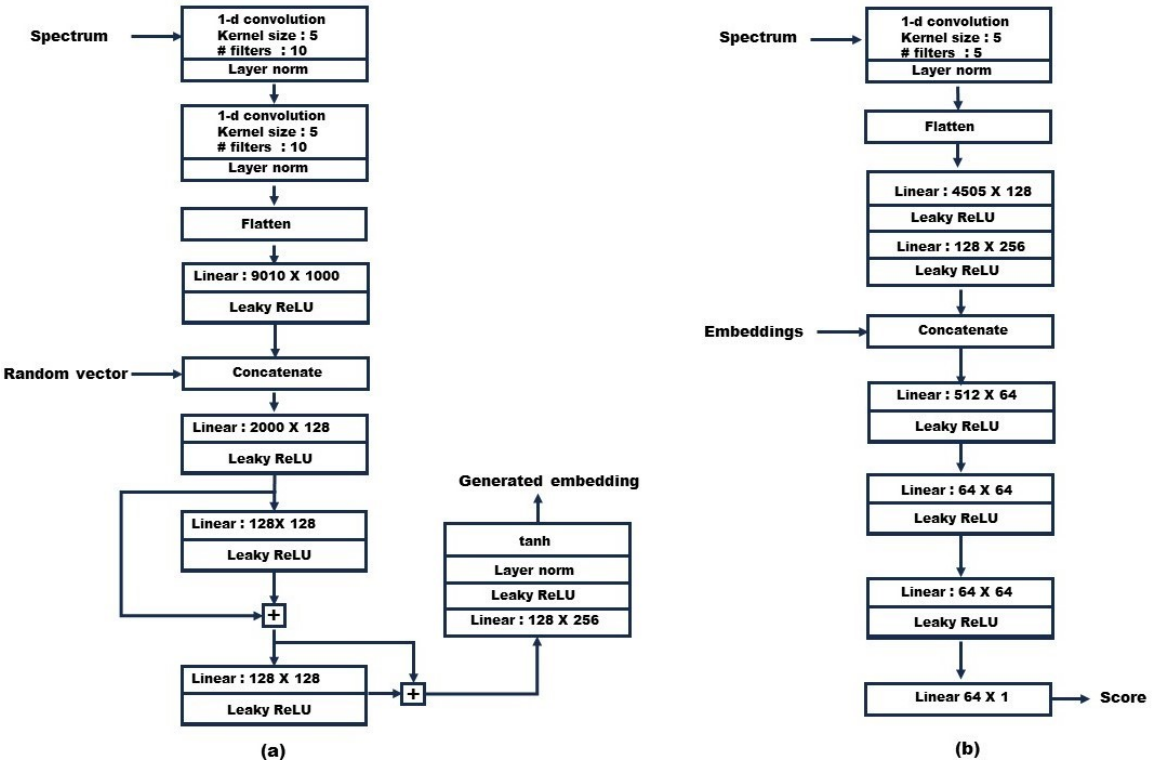


Figure C.4: Architecture of (a) Generator and (b) Discriminator

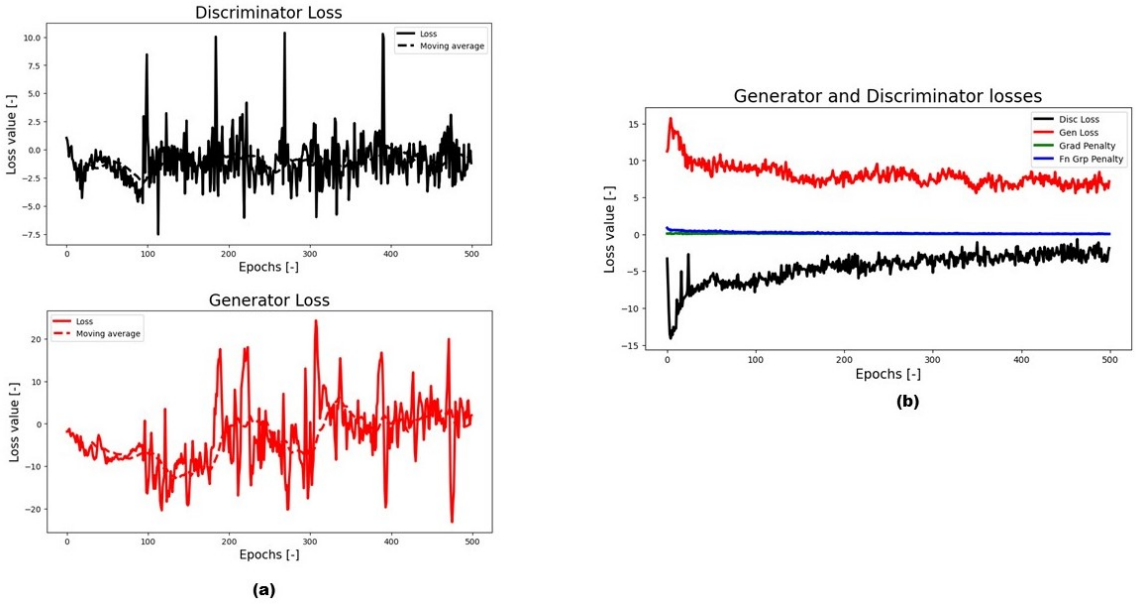
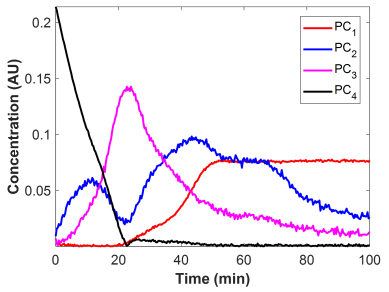


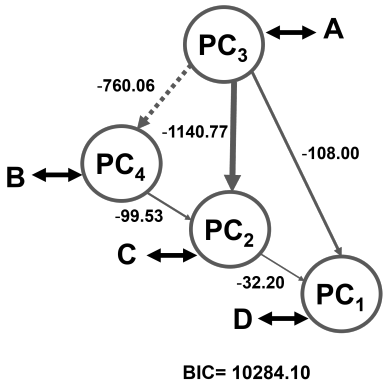
Figure C.5: Plots of losses during training of GANs (a) without functional group penalty and (b) with functional group penalty.

Appendix D: Chapter 5

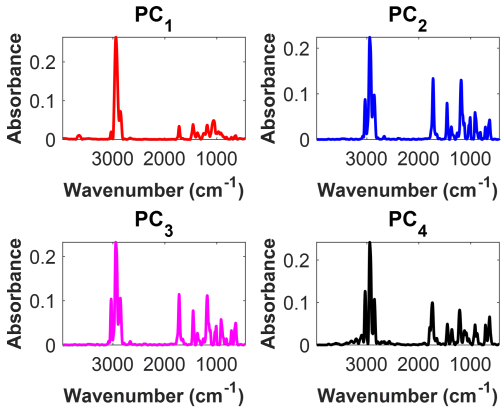
D.1 Model results for synthetic data with signal to noise ratio of 35



(a) Temporal concentrations



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



(c) Pseudo-component spectra

Figure D.1: Spectral deconvolution and causal inference using noisy synthetic data at a signal to noise ratio of 35.

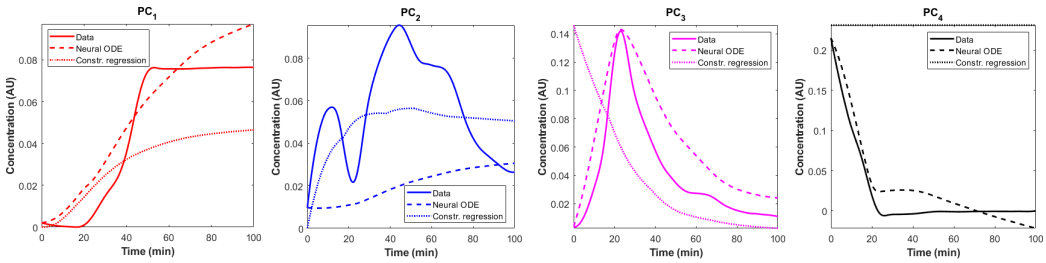
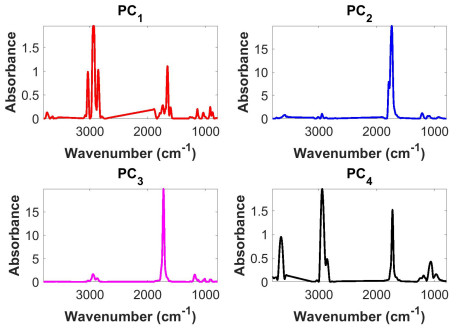
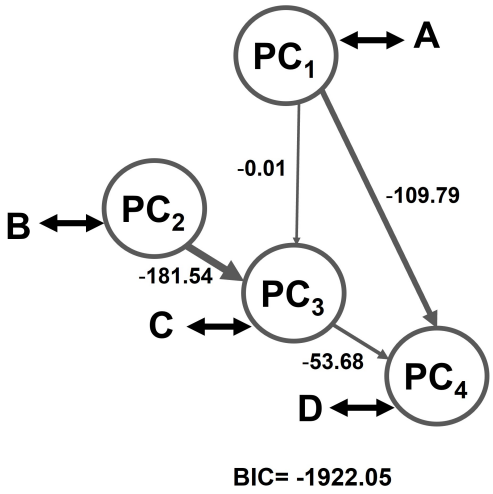


Figure D.2: Comparison of the predictions from the chemical neural ODE and constrained regression against the reconstructed data from integration of the smoothed time derivative of temporal concentration obtained by the deconvolution of synthetic spectroscopic data, at a signal to noise ratio of 35.

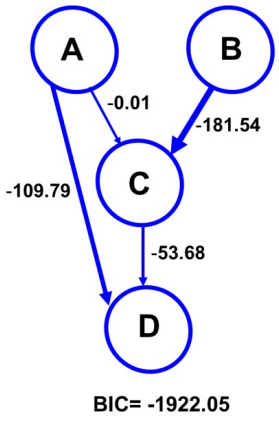
D.2 Impact of preferentially weighting synthetic spectra on the adjacency matrix



(a) Preferentially weighted pseudo-component spectra after deconvolution



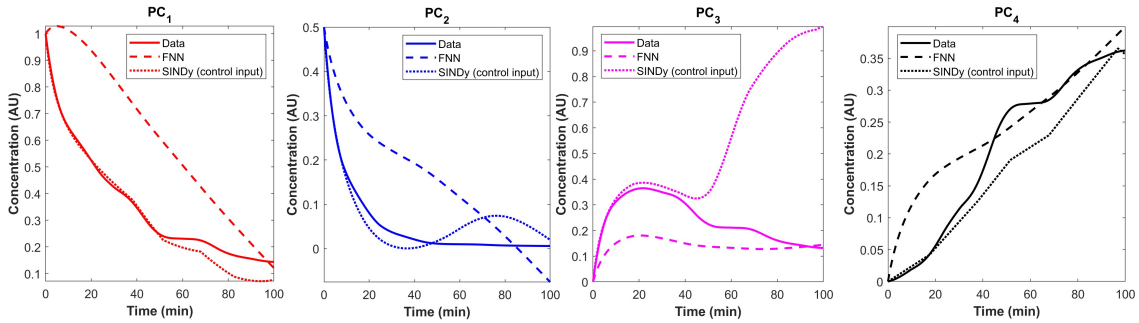
(b) Reaction network structure, arc strengths and score inferred from the preferentially weighted pseudo-component spectra



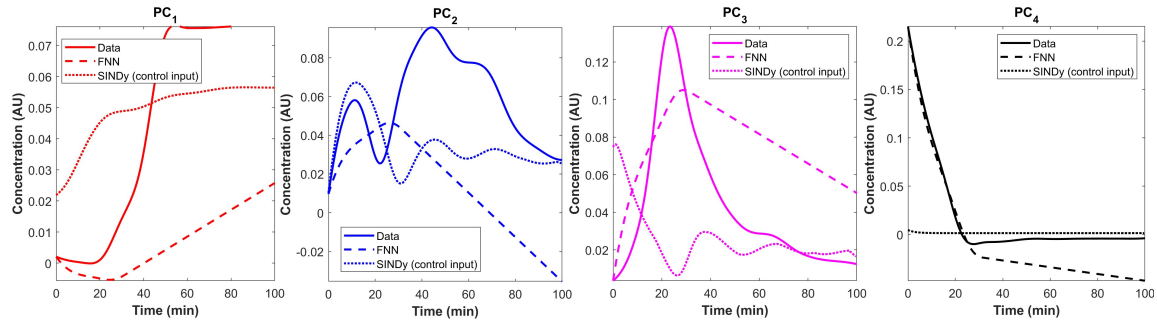
(c) Arc strengths and score inferred from the preferentially weighted spectra, given the reaction template

Figure D.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.

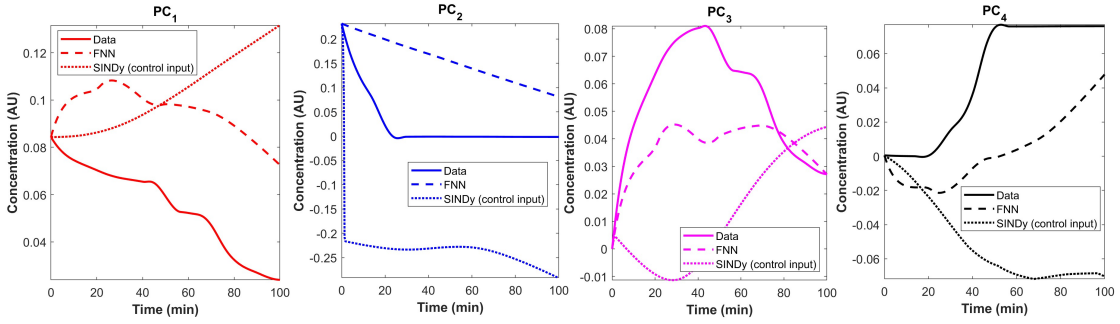
D.3 Performance assessment of the proposed frameworks against their baselines



(a) Predictions compared against the baselines for concentration profiles from latent factorization of noise-free synthetic spectroscopic data.



(b) Predictions compared against the baselines for concentration profiles from latent factorization of synthetic spectroscopic data at a SNR=35.



(c) Predictions compared against the baselines for concentration profiles from latent factorization of synthetic spectroscopic data at a SNR=100.

Figure D.4: Predictions from the chemical neural ODE and the constrained regression by ALS are compared against their baselines: a simple feed forward neural network (FFN) and a SINDy with control input regression, respectively.

Table D.1: Comparison of the performance of the neural ODE and constrained regression frameworks against their respective unconstrained baseline models.

	Root mean squared error (RMSE)		% Improvement over baseline	Root mean squared error (RMSE)		% Improvement over baseline
Type of data	Constr. regression	SINDy with control input (baseline)		Neural ODE	FNN (baseline)	
Without noise	0.0132	0.2289	94.23%	0.0276	0.1737	84.11%
SNR= 35	0.1121	0.042	-62.53%	0.0278	0.0427	34.89%
SNR= 100	0.0314	0.1473	78.68%	0.01	0.0772	87.05%