

**MOLECULAR-LEVEL KINETIC MODELING OF CONVENTIONAL AND
UNCONVENTIONAL HYDROPROCESSING FEEDSTOCKS**

by

Pratyush Agarwal

A dissertation submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering

Spring 2019

© 2019 Pratyush Agarwal
All Rights Reserved

**MOLECULAR-LEVEL KINETIC MODELING OF CONVENTIONAL AND
UNCONVENTIONAL HYDROPROCESSING FEEDSTOCKS**

by

Pratyush Agarwal

Approved: _____
Eric M. Furst, Ph.D.
Chair of the Department of Chemical & Biomolecular Engineering

Approved: _____
Levi T. Thompson, Ph.D.
Dean of the College of Engineering

Approved: _____
Douglas J. Doren, Ph.D.
Interim Vice Provost for Graduate and Professional Education

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Michael T. Klein, Sc.D.
Professor in charge of dissertation

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Antony N. Beris, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Prasad S. Dhurjati, Ph.D.
Member of dissertation committee

I certify that I have read this dissertation and that in my opinion it meets the academic and professional standard required by the University as a dissertation for the degree of Doctor of Philosophy.

Signed:

Ashish B. Mhadeshwar, Ph.D.
Member of dissertation committee

ACKNOWLEDGMENTS

First and foremost, I would like to acknowledge my advisor, Professor Michael Klein, for his continued guidance and support in the completion of this dissertation work. He provided a great balance of guidance and scientific independence that allowed me to successfully complete the work while exploring my interests. He also encouraged me to pursue internship opportunities that helped me develop my skills and become a very attractive candidate during my search for employment.

I would like to express my sincere gratitude to my committee members Dr. Prasad Dhurjati, Dr. Antony Beris, and Dr. Ashish Mhadeshwar for their guidance and feedback on my dissertation work. My conversations with them inside and outside of committee meetings helped me mold my thesis work into a stronger argument for developing molecular-level kinetic models.

The financial support of Reliance Industries is gratefully acknowledged. Their data and expertise in refinery processes was invaluable in creating accurate, commercially relevant models. I especially wanted to recognize the support of Dr. Mayuresh Sahasrabudhe who acted as the main point of contact for the vacuum gas oil project.

I would like to thank my research group members, past and present, for creating a positive work environment that fostered collaboration and friendship. Without their guidance, especially in the formative stages, this dissertation would have been impossible. I am indebted to Juan Lucio-Vega, Scott Horton, Craig Bennett, Zhen Hou, and Triveni Billa for the intellectual conversations during the course of the

thesis work. I am especially glad to have maintained friendships with Juan and Scott even as they graduated and relocated to different locations. Additionally, I would like to thank Marguerite Mahoney for her friendly conversations and froyo trips.

I wanted to thank my friends and extended family in the area for sticking with me through the good and bad times. I am grateful for the trivia nights, board game nights, Philadelphia trips, and family vacations that allowed me to unwind during my dissertation. As I leave the area to move to Texas, I will miss seeing them regularly.

Lastly, I would like to express my gratitude to my family for their unconditional love. My parents, Sandeep and Pinky, made countless sacrifices to support my academic journey and always encouraged me to be my best self. My brother and sister-in-law provided guidance and support when I needed it. This thesis is dedicated to them.

TABLE OF CONTENTS

LIST OF TABLES	ix
LIST OF FIGURES	x
ABSTRACT	xv

Chapter

1	INTRODUCTION	1
1.1	Conventional Feedstocks	1
1.2	Unconventional Feedstocks	3
1.3	Hydroprocessing	5
1.4	Kinetic Modeling	6
1.5	Research Objectives	7
1.6	Dissertation Scope	9
2	THE KINETIC MODELER'S TOOLBOX	11
2.1	Introduction	11
2.2	Background of KMT	13
2.3	The Interactive Network Generator (INGen)	13
2.3.1	Computational Representation of Molecules	14
2.3.2	Network Seed Molecule Selection	15
2.3.3	Computational Representation of Reactions	16
2.3.4	Reaction Family Selection	17
2.4	The Initial Condition Generator (ICG)	20
2.5	The Dynamic Model Builder (DMB)	20
2.5.1	Kinetic Parameter Definition and Minimization	22
2.5.2	Parameter Optimization	25
2.6	Summary	27
3	MOLECULAR-LEVEL KINETIC MODELING OF TRIGLYCERIDE HYDROPROCESSING	28

3.1	Abstract.....	29
3.2	Introduction	29
3.3	Reaction Network Generation	33
3.4	Feed Specifications.....	36
3.5	Model Equations and Kinetics.....	37
3.6	Kinetic Model Evaluation.....	39
3.7	Diesel Property Calculation.....	45
	3.7.1 Cetane Number Model	45
	3.7.2 Cloud Point Model	47
	3.7.3 Predicting Diesel End-Use	49
3.8	Conclusions	50
3.9	Nomenclature	51
3.10	Acknowledgement.....	52
4	MOLECULAR-LEVEL KINETIC MODELING OF A REAL VACUUM GAS OIL HYDROPROCESSING REFINERY SYSTEM	53
4.1	Abstract.....	54
4.2	Introduction	55
4.3	Experimental Data	57
4.4	Reaction Network Generation	60
	4.4.1 Molecule Selection	61
	4.4.2 Reaction Selection and Constraints	64
	4.4.3 Coking Chemistry.....	67
	4.4.4 Network Results	69
4.5	Feed Composition Generation.....	70
	4.5.1 Feedstock PDF Definitions.....	71
	4.5.2 Parameter Optimization.....	73
	4.5.3 Feed Composition Results	74
4.6	Kinetic Model Generation	79
	4.6.1 Model Equations and Kinetics.....	81
	4.6.2 Kinetic Model Evaluation.....	85
4.7	User Interface	92
4.8	Conclusions	94
4.9	Nomenclature	95
4.10	Acknowledgement.....	97

5	GENERATING DATA-DRIVEN MODELS FROM MOLECULAR-LEVEL KINETIC MODELS: A KINETIC MODEL SPEEDUP STRATEGY	98
5.1	Abstract.....	99
5.2	Introduction	99
5.3	Molecular-Level Kinetic Model	102
5.4	Model Setup.....	103
5.5	Multilinear Regression	105
5.6	Machine Learning.....	109
5.7	Impact of Data	113
5.8	Conclusions	115
5.9	Acknowledgement.....	116
6	THE INITIAL CONDITION GENERATOR: A SOFTWARE TOOL TO STATISTICALLY DETERMINE INDIVIDUAL MOLECULE FRACTIONS FROM EXPERIMENTAL MEASUREMENTS	117
6.1	Introduction	117
6.2	Molecule Properties.....	120
6.3	Probability Density Function Trees.....	121
6.4	Bulk Property Correlations.....	124
6.5	Parameter Optimization.....	125
6.6	ICG User Interface	126
6.7	ICG Power User	136
6.8	Current and Future Development	137
6.9	Summary.....	138
7	SUMMARY, CONCLUSIONS, AND FUTURE WORK.....	139
7.1	Summary.....	139
7.2	Conclusions	142
7.3	Recommendations for Future Work	143
	REFERENCES	147

LIST OF TABLES

Table 3.1: Network representation for coconut oil hydroprocessing generated using the Interactive Network Generator (INGen)	36
Table 3.2: Fatty acid composition of coconut oil ⁴¹ and soybean oil ⁶ . Data directly from source.....	37
Table 3.3: Reaction parameters for the linear free-energy relationship that define the kinetic rate constants in Equation 2. α for all reaction families was kept constant at a value of 0.02	40
Table 3.4. Adsorption constants for the kinetic model defined in Equation 3	41
Table 3.5: Calculated average βi values in this work compared to the work of Ghosh and Jaffe ³⁸	46
Table 3.6: ε correction parameter values for the cloud point model for the molecular lumps that impacted the final result based on available experimental data	48
Table 4.1: Typical feed measurements and inlet process conditions for a vacuum gas oil hydroprocessing system	60
Table 4.2: Molecule types present in the vacuum gas oil hydroprocessing network ...	63
Table 4.3: Reaction families in the vacuum gas oil hydroprocessing network	67
Table 4.4: Summary of molecules and reactions in the vacuum gas oil hydroprocessing model	70
Table 4.5: LFER and catalyst LFER parameters for the VGO hydroprocessing model	87
Table 4.6: Adsorption parameters for the VGO hydroprocessing model for the acid and metal sites on the catalyst	88
Table 6.1: List of properties generated by the property database application	121

LIST OF FIGURES

Figure 1.1: World proven crude oil reserves. Data from the CIA World Factbook (2017 estimate) ¹	2
Figure 1.2: Typical fractionation of crude oil by boiling points. Data from Srivastava and Hancsók ⁴	3
Figure 1.3: The structure of a triglyceride molecule containing three fatty acid chains fused to a propane backbone	5
Figure 2.1: The main software tools in the Kinetic Modeler's Toolbox.....	12
Figure 2.2: Computational representations of a molecule as bond-electron matrices and adjacency lists.....	15
Figure 2.3: Limiting reaction network growth via seeding molecules (grey) along the reaction pathway in a rank 1 network.....	16
Figure 2.4: Computation representation of a reaction as a matrix addition operation .	17
Figure 2.5: a) Mechanism-level and b) pathways-level representations of a β -scission reaction	19
Figure 2.6: Workflow of the Initial Condition Generator tool	20
Figure 2.7: Figure from Horton and Klein ²¹ (a) Figure from Mochida and Yoneda ²² showing the dealkylation of the alkylaromatic reaction family. Each line represents a different catalyst, where the data points are particular reaction measurements. (b) Figure generated from reaction family parameters from Mochida and Yoneda for 10 total catalysts.....	24
Figure 3.1: Pure-component melting points for different n-paraffins and i-paraffins. Data from the NIST Chemistry Webbook ³⁹ . The lines illustrate general trends of melting temperature with carbon number for different paraffin classes.	32
Figure 3.2: Reaction network for triglyceride hydroprocessing.....	34

Figure 3.3: Comparison of the predicted and experimental ^{6,41} overall conversions of the coconut and soybean oils at a range of temperatures, pressures, and catalyst contact times.....	42
Figure 3.4: Parity plot showing the comparison between the experimental ⁶ and predicted bulk fraction concentrations of oxygenates, aromatics, cycloparaffins, olefins, isoparaffins, and n-paraffins for soybean oil hydroprocessing at 350-440°C and 4.5-12 MPa	43
Figure 3.5: Parity plot showing the comparison between the experimental ⁶ and model predictions for the gas phase concentrations of CO, CO ₂ , CH ₄ , C ₂ H ₆ , and C ₃ H ₈ for soybean oil hydroprocessing at 350-440°C and 4.5-12 MPa	43
Figure 3.6: Parity plot showing the comparison between experimental ^{6,41} results and model predictions of products from a) coconut oil hydroprocessing at 350°C and 0.8 MPa for five different contact times and b) soybean oil hydroprocessing at 350-440°C and 4.5-12.0 MPa for the same contact time.....	44
Figure 3.7: Parity plots comparing experimental and calculated values from the a) cetane number and b) cloud point models. The y=x line is displayed for comparison.....	47
Figure 3.8: Tradeoff between cetane number and cloud point with varying degrees of isomerization. Degree of isomerization is defined as the weight of isoparaffins out of the total paraffin weight	50
Figure 4.1: A simplified representation of the two-reactor hydroprocessing unit for vacuum gas oil.....	58
Figure 4.2: Average reactor inlet temperature over the lifetime of the catalysts. The (●) dots represent process data and the (--) line represents a fit with an average $\Delta T = 0.045$ K.	59
Figure 4.3: Matrix representation of a reactive subgroup and reaction matrix in INGen	61
Figure 4.4: Metal deposition from a Ni porphyrin on a catalyst site L based on the mechanism by Ware et al. ⁷⁵	68
Figure 4.5: Aromatic ring building in the coking reaction pathways via a) coupling of aromatic rings with each other and b) alkylation, ring closing, and aromatization to form larger aromatic clusters from alkyl chains.....	69

Figure 4.6: PDF tree representation of the PDFs on the VGO hydroprocessing model.	72
Figure 4.7: Overall carbon number distribution and class distribution in the feedstock model for a representative dataset.....	75
Figure 4.8: Ring number distributions for aromatics and naphthenics in the feedstock model for a representative dataset.....	75
Figure 4.9: Maximum prediction errors in the simulated distillation 5%, 10%, 30%, 50%, 70%, 90% and 95% boiling cuts for the feedstock model over the entire process range with a datapoint every seven days on stream ...	77
Figure 4.10: Error in the prediction of the a) sulfur and b) nitrogen elemental analysis for the feedstock model over the entire process range with a datapoint every seven days on stream	78
Figure 4.11: Error in the prediction of the feedstock density values over the entire process range with a datapoint every seven days on stream	79
Figure 4.12: Flowchart displaying the logic of kinetic model evaluation in DMB.....	80
Figure 4.13: Maximum prediction errors in the simulated distillation 5%, 10%, 30%, 50%, 70%, 90% and 95% boiling cuts for the reactor effluent over the entire process range where experimental data was available	90
Figure 4.14: Error in the prediction of the a) sulfur and b) nitrogen elemental analysis for the reactor effluent over the entire process range where the data was available.....	91
Figure 4.15: Error in the prediction of the product density values over the entire process range where data was available	92
Figure 4.16: The user interface of the user-friendly application designed to run the VGO hydroprocessing model.....	93
Figure 5.1: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a multilinear regression model for the testing data	106
Figure 5.2: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with four multilinear regression model partitioned by inlet temperature for the testing data.....	107

Figure 5.3: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with four multilinear regression model partitioned by inlet hydrogen pressure for the testing data	108
Figure 5.4: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a decision tree regression model for the testing data	110
Figure 5.5: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a gradient boosted regression tree model for the testing data	111
Figure 5.6: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with an artificial neural network model for the testing data	111
Figure 5.7: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a multilinear regression (red), decision tree (green), and artificial neural network (blue) model for data outside the range of the training data for the DDMs.....	113
Figure 5.8: Number of datasets needed for each of a) multilinear regression, b) decision trees, and c) neural networks to generate models of a given accuracy.....	114
Figure 6.1: A sample PDF tree representation of a basic petrochemical feedstock. ..	123
Figure 6.2: Home page of the ICG application user interface.....	127
Figure 6.3: Automatic population of the four tabs representing the ICG feedstock model generation steps after creating a new model	128
Figure 6.4: Step 1 for the ICG feedstock model construction	129
Figure 6.5: Navigation menu to the property database molecule list	130
Figure 6.6: Histogram addition window for a new custom histogram	130
Figure 6.7: Selection of pre-defined PDF tree structures in ICG	131
Figure 6.8: Exploring the molecules and PDFs after completing step 1 in ICG	132
Figure 6.9: Defining the experimental properties in step 2 of ICG	133

Figure 6.10: Adding a user-defined experiment in ICG.....	133
Figure 6.11: Defining the optimization problem in step 3 of ICG	134
Figure 6.12. Optimizing and viewing the results in step 4 of ICG.....	136
Figure 6.13: The definition of the PDF tree in the '.csv' file format	137

ABSTRACT

Hydroprocessing is a catalytic upgrading process in a hydrogen-rich environment that is commonly used to remove impurities, saturate carbon-carbon bonds, and sometimes break carbon-carbon bonds. Kinetic models are essential in the optimization of hydroprocessing reactor systems to meet strict environmental and product yield specifications. Traditional lumped kinetic models are often feedstock dependent and limited in their prediction capability. Molecular-level kinetic models can address those drawbacks by considering the fundamental chemistry and kinetics of the process. In this work, molecular-level kinetic models were developed for triglyceride and vacuum gas oil hydroprocessing. Model development parallelly also facilitated the improvement of the model building tools to accurately capture the real process and improve the user experience.

First, a molecular-level kinetic model was developed for triglyceride hydroprocessing. Triglycerides representative of the types present in coconut oil and soybean oil were defined using 8 to 22 carbon fatty acids. The triglycerides acted as seeds to a reaction network detailing three parallel deoxygenation pathways: hydrodeoxygenation, decarboxylation, and decarbonylation. Isomerization, cyclization, aromatization, and cracking reactions were iteratively added to the reaction network. The final network contained 476 species and 1709 reactions. Using the network, material balances were written for the kinetic model. The kinetic parameters were optimized based on experimental data over a range of temperatures, pressures, and catalyst contact times. The final kinetic model simulated properties had

good agreement with experimental values. To evaluate the end-use value of the diesel product, cetane number and cloud point property models were constructed and optimized based on experimental data. These property models were used to study the product diesel cetane number versus cloud point tradeoff to determine the end-use properties of the product fuel.

Then, a molecular-level kinetic model was constructed for a vacuum gas oil hydroprocessing unit. The experimental data were from an operating refinery unit over the two-year catalyst life. Feedstock molecules containing up to 45 carbons of paraffinic, olefinic, naphthenic, and aromatic type were selected to represent the molecular composition. On those molecules, the typical desulfurization, denitrogenation, saturation, cracking, ring opening, and isomerization reactions were applied. The final network included 5747 reactions and 1532 species. The species feedstock concentrations were determined by sampling the probabilities of the presence of different structural attributes using probability density functions (PDFs). PDF parameters were optimized using a simulated annealing algorithm. To minimize the optimization burden of the PDF parameters, a library containing 21 sets of PDF parameters was created and used to determine the starting point of optimization for each new dataset. For the kinetic model, the reactor system was represented as a series of 19 pseudo-PFRs. The pseudo-PFRs were individual catalyst layers with side-by-side reaction and vapor-liquid equilibrium. Quantitative structure/reactivity correlations and linear free-energy relationships (LFERs) were used to reduce the number of kinetic parameters. The activity of each type of catalyst was modeled independently using the catalyst LFER concept. After optimization using a simulated

annealing algorithm, the model showed good agreement with the experimental measurements.

Next, a strategy to generate data-driven models from molecular-level kinetic models was evaluated to greatly reduce the time required for a prediction. The triglyceride hydroprocessing model was used to generate 20,000 datasets for a small ranges of input parameters. For each dataset, the calculated cetane number, cloud point, and yield were recorded. As an initial approach, multilinear regression, decision tree regression, gradient boosting regression, and artificial neural network models were generated from the data. All data-driven models were able to predict results accurately and very quickly ($\ll 1$ second), but they only worked in the small ranges of the underlying data. As soon as the inputs exceeded the input parameter ranges, data-driven model predictions greatly diverged from the kinetic model results. In terms of data requirements, multilinear regression models needed much less data than the decision tree regression and artificial neural network models at the cost of some accuracy if the model input parameter range was too wide.

Finally, a feed concentration modeling tool, the Initial Condition Generator (ICG) was developed. The tool was designed to allow users to import a list of molecules, define the PDF structures, input the experimental data, and optimize the kinetic parameters. Special attention was paid to reducing the user requirement of computer science or kinetic modeling expertise. The tool was a direct outcome of the identified need of a simplified feedstock composition model compared to the previous modeling framework. Thus, the need for parallel model development and model builder development to improve future model development can be seen.

Chapter 1

INTRODUCTION

1.1 Conventional Feedstocks

The conventional refinery feedstock is crude oil. Crude oil is a naturally occurring source of hydrocarbons that is extracted from underground resources. While it is limited in quantity and requires millions of year to replenish naturally, there are still very large reserves of crude oil available for use in the coming decades, as displayed in Figure 1.1.¹ The hydrocarbons derived from crude oil provide fuel sources and the basic building blocks for most plastics and materials in use today. Gasoline, diesel, jet fuel, heating oil, and natural gas all derive primarily from crude oil resources.² The production of crude oil is anticipated to continue increasing through 2026, after which a period of nearly constant production is anticipated till 2040.³ Clearly, crude oil is a valuable resource that will continue to be an important part of society for the years to come.

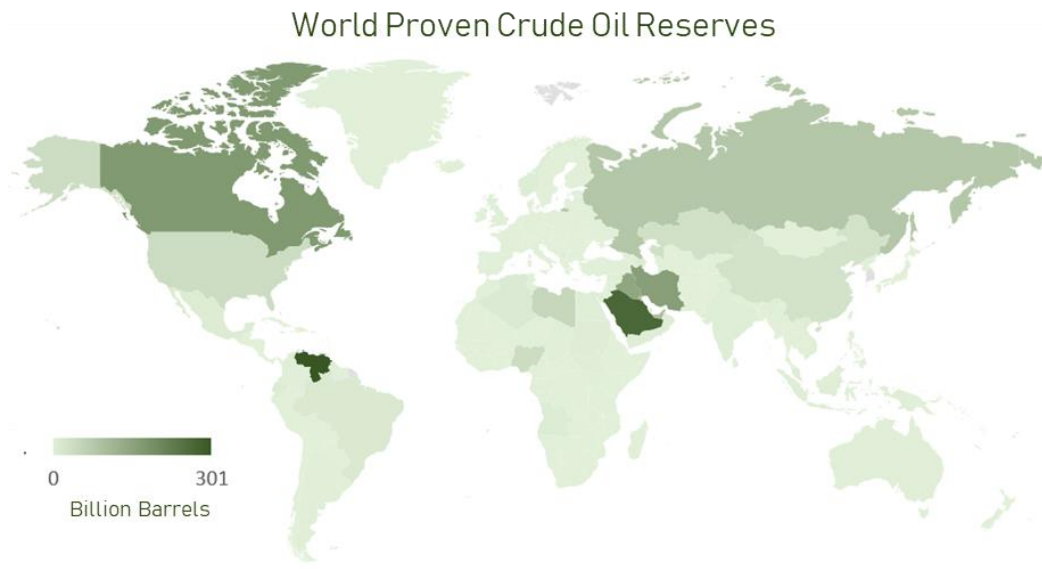


Figure 1.1: World proven crude oil reserves. Data from the CIA World Factbook (2017 estimate)¹

Crude oil is a mixture of hydrocarbons of different sizes and structures. End-products from crude oil typically lie in narrow ranges of boiling points and types of molecules. In a typical refinery, crude oil is first fractionated into a few boiling point cuts, as shown in Figure 1.2. The fractions are routed to different processing units designed to accommodate the processing needs of the specific fraction. The light gas fraction is the source of natural gas for fuel as well as a feed to cracking units designed to create monomers for plastics. Molecules in the gasoline range are typically present in the naphtha range along with molecules that lend to specialty chemical and products. The naphtha fraction can be processed in a catalytic reformer to improve the yields of desired types of molecules. Kerosene and diesel fractions typically lend directly to the corresponding fuels. Some processing is often required to remove heteroatom impurities as designated by regulatory fuel standard in the region of use.

The vacuum gas oil and vacuum residue fractions are heavier than the other fractions with limited direct uses. They also suffer from the presence of a high number of impurities like sulfur, nitrogen, and heavy metals that have detrimental environmental effects if released. However, they contain a large amount of valuable hydrocarbons. Typically, cracking in thermal or hydrogen-rich environments is needed to reduce the overall boiling point and remove impurities of these heavy fractions. The cracked products are fractionated and lend to the products of the lighter fractions of crude oil.^{2,4}

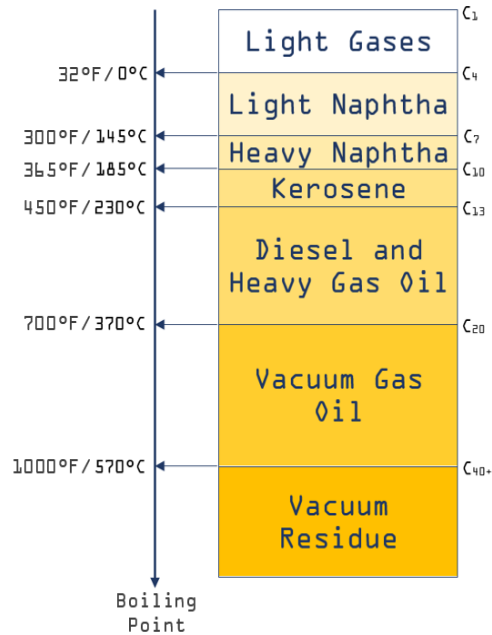


Figure 1.2: Typical fractionation of crude oil by boiling points. Data from Srivastava and Hancsók⁴

1.2 Unconventional Feedstocks

With a growing concern about the consumption of crude oil and adverse environmental impact associated with its use, there has been a great deal of interest in

renewable feedstocks. The refineries of the future will use feedstocks derived mainly from plant, waste, or algal sources. These feedstocks are either more carbon neutral than conventional feedstocks or a method of utilizing otherwise wasted material. Research in the past few decades has illuminated many pathways of tapping into these unconventional sources for energy, chemicals, and fuels.

Plant and algal sources fix carbon from the environment using solar energy into biomass. Consuming biomass-based fuels and chemical releases the carbon back into the environment, hence creating a cycle that is much more sustainable than consuming crude oil. Oils and sugars from plants and algae provide excellent building blocks for value-added fuels and chemicals. These molecules contain a much higher oxygen content than crude oil which can be undesirable in many applications. Therefore, the first step in utilizing biomass is usually a deoxygenation mechanism. Cracking, transesterification, hydroprocessing, and hydrolysis are viable options for deoxygenation, each with its merits and drawbacks for different types of oils and sugars.^{5,6}

Municipal solid waste (MSW) is a mixture of discarded plastics and biomass. Typically, this waste accrues in unsightly landfills that release greenhouse gases during microbial degradation. Additionally, leeching from landfills can contaminate groundwater. However, MSW is energy rich and can be a valuable resource for energy production. Gasification and combustion are viable methods of extracting this energy as heat or valuable gases.⁵ The gases can be used as light fuels or further processed in gas-to-liquid technologies that produce gasoline- and diesel- range fuels and chemicals.

One viable method of renewable diesel production is the hydroprocessing of triglycerides present in plant and algal oils. Triglycerides oils usually contain long carbon chain fatty acids fused to a propane backbone as shown in Figure 1.3. The lengths of these chains vary between 8 and 24 carbons in common oils. Long carbon chains are excellent hydrocarbon sources for diesel fuels. While triglycerides and fatty acids can directly be used in diesel engines, they have poor oxidative stability and poor performance in cold climates. Therefore, they need to be processed to remove the oxygen and improve the solidification properties.

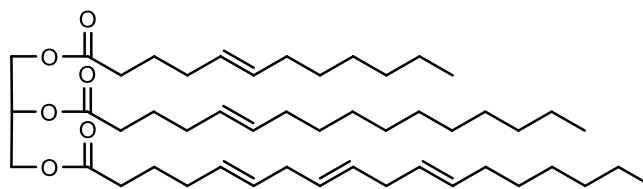


Figure 1.3: The structure of a triglyceride molecule containing three fatty acid chains fused to a propane backbone

1.3 Hydroprocessing

Hydroprocessing is a popular technology to catalytically process feedstocks in the presence of hydrogen. In general, hydroprocessing is used to remove heteroatoms and saturate carbon-carbon bonds. Depending on the hydroprocessing severity, it can also be used to break carbon-carbon bonds for a drastic reduction in the overall molecular weight. The process is very versatile in its applications and robust in its sensitivity to different feedstock components. It has been applied to the upgrading of feedstocks like naphtha, diesel, vacuum gas oil, vacuum residue, plant oils, algal oils, and pyrolysis oils. In most cases, the desired activity is the removal of sulfur, nitrogen, oxygen, and metal impurities while saturating double bonds and aromatic rings. The

severity of the process conditions, controlled mainly by temperature and catalyst type, determines the amount of cracking.

Hydroprocessing units typically employ bifunctional metal/acid catalysts. The metal in the catalyst is typically a combination of nickel, molybdenum, tungsten, and cobalt. It is mainly responsible for the heteroatom capture and double-bond saturation activity. The acid support, typically a γ -alumina carrier, provides the acid character that helps with the carbenium ion formation responsible for molecular rearrangements and cracking reactions. Many layers of catalysts are used in a process unit where the type of catalyst and size of the layer depend on the type of hydroprocessing needed. The amount and activity of the different catalysts along with the process conditions determines the specifications of the end product. An important aspect of the hydroprocessing system is the overall deactivation of the catalyst over time. Deactivation occurs mainly due to coke and metal deposition, where there is an additional impact of effects like leaching, sintering, and mechanical breaking.

1.4 Kinetic Modeling

Kinetic modeling is a useful technique of organizing kinetic information and predicting reactor results without needing extensive experimental programs that can be expensive and slow every time there is a minor change in the process. The most basic type of kinetic models are lumped models that represent feedstock transformations using lumps. Lumped models are popular because they are easy to understand, build, and solve. However, the lumps are usually defined by boiling points or structural attributes, obscuring the true chemistry and kinetics in the process. They are also feedstock dependent and lack detail beyond the lumps that may be required to answer new questions concerning specific molecules in a process.

To overcome some of the limitations of lumped models, molecular-level kinetic models can be created that model explicit molecules and their transformations in a reactor. The kinetic parameters have a fundamental basis, removing the dependence on specific feedstocks and specific reactor configurations. Additionally, the molecular output can be used to calculate any desired property using an appropriate mixing rule. Any new questions about the effluent can thus be answered by the molecular model. The major downside to molecular-level kinetic models is that they are harder to understand and more computationally expensive.

1.5 Research Objectives

This dissertation had two parallel research objectives: 1) develop molecular-level kinetic models for hydroprocessing systems and 2) advance model-building tools to improve the user experience. Developing the kinetic models answers immediate questions about the effluent of the process being studied. During the development of the kinetic models, the model-building tools were improved based on the functionality needed to effectively construct and solve the kinetic model. In this manner, both research objectives were achieved simultaneously.

The primary objective of this thesis was to model vacuum gas oil and triglyceride hydroprocessing at the molecular level. For the vacuum gas oil model, experimental data were available for a two-year catalyst life cycle in an operating refinery unit. The goal of the project was to accurately represent the data while providing a model with predictive capabilities. An important consideration in the model was a designation of the activity of the different catalysts that may be present in the system to provide information about catalyst performance in the future. For the triglyceride hydroprocessing model, experimental data were found in literature that

described to conversion of triglycerides to diesel fuels in hydroprocessing reactors. Once the product could be calculated, of further interest were the ignition and solidification properties of the product fuel. These properties allow for a better estimate of the commercial value of the product fuel.

While developing the two models, several areas of improvement were identified in the model-building tools. Firstly, a method was needed to model the vapor and liquid streams independently for the heavy feedstocks. The modeling software previously only allowed gas phase reactions and kinetics. Additionally, capabilities were needed to model the quenches and the different catalyst activities in the system. Lastly, a user-friendly tool was needed to model the feedstock mole fractions from the experimental bulk properties. The old capability for feedstock modeling lacked customization ability and was overly complicated to use. With these improvements, the current models and models of the future should be easier to develop with more available options for the types of models.

Furthermore, a need was identified to access the results of molecular-level kinetic models without the long solution times that accompany solving the complex system of equations. To this end, opportunities needed to be explored in creating data-driven models based on data generated from the molecular-level kinetic models. A good data-driven model was defined as one that accurately captured the information in the kinetic model, had a solution time of $\ll 1$ second, and could be easily created as more data were available. Machine learning regression techniques were explored for this reason.

1.6 Dissertation Scope

Chapter 2 provides an overview of the modeling tools used in the Klein research group to build molecular-level kinetic models. The chapter details the theory, functionality, and development of the main software tools. In addition, a brief discussion is provided on how the tools are used to generate the reactions, calculate the feedstock mole fractions, and represent the kinetics in the model.

Chapter 3 discusses the development of a molecular-level kinetic model for triglyceride hydroprocessing. The three parallel triglyceride deoxygenation pathways are discussed and modeled in a reaction network. Coconut oil and soybean oil feedstocks are modeled, and the simulated and experimental yields for both feedstocks are compared. Also, property models for cetane number and cloud point are discussed to calculate the end-value of the product diesel. The variations of the properties with individual molecule identities highlights the need for modeling the system at a molecular level.

Chapter 4 details the construction of the vacuum gas oil hydroprocessing kinetic model. Reaction and molecule selection are discussed to accurately represent the real chemistry in the process. Then, the setup of the kinetic model is discussed with the functionality of vapor-liquid equilibrium, multiple reactors, hydrogen quenches, and multiple catalysts. Finally, the kinetic model is evaluated by its ability to predict the product over the catalyst lifetime with various process conditions and feedstock properties.

Chapter 5 focuses on generating data-driven models from molecular-level kinetic models. Different regression algorithms are studied and evaluated on their ability to accurately represent the information in the kinetic model. Qualitative

assessments are made about the data requirements for each algorithm and the suitability of a specific algorithm in a real application.

Chapter 6 provides a detailed summary of the development of a new software tool to model the individual molecule mole fractions from experimental data. The underlying theories and program logic are discussed. Additionally, the chapter explains the capabilities of the software to model feedstocks and shows the user-interface that guides users in developing the feedstock models.

Chapter 7 summarizes the findings of this dissertation, presents the major conclusions of the research work, and provides avenues for future development of kinetic models and modeling tools.

Chapter 2

THE KINETIC MODELER'S TOOLBOX

2.1 Introduction

The Kinetic Modeler's Toolbox (KMT) is a suite of software tools that help build and solve molecular-level kinetic models.⁷⁻¹¹ There are three main components in KMT: the Interactive Network Generator (INGen), the Initial Condition Generator (ICG), and the Dynamic Model Builder (DMB). Figure 2.1 shows an overview of the software tools and their relationship to each other. INGen is a network generator that allows users to select molecule seeds and set reaction rules to systematically generate a reaction network. ICG models the composition by assigning probability density functions (PDFs) to the presence of structural moieties in the species and simulates the experimental measurements. Lastly, DMB organizes the rate equations and the forms of the material balances to solve an initial value problem that defines the kinetic model.

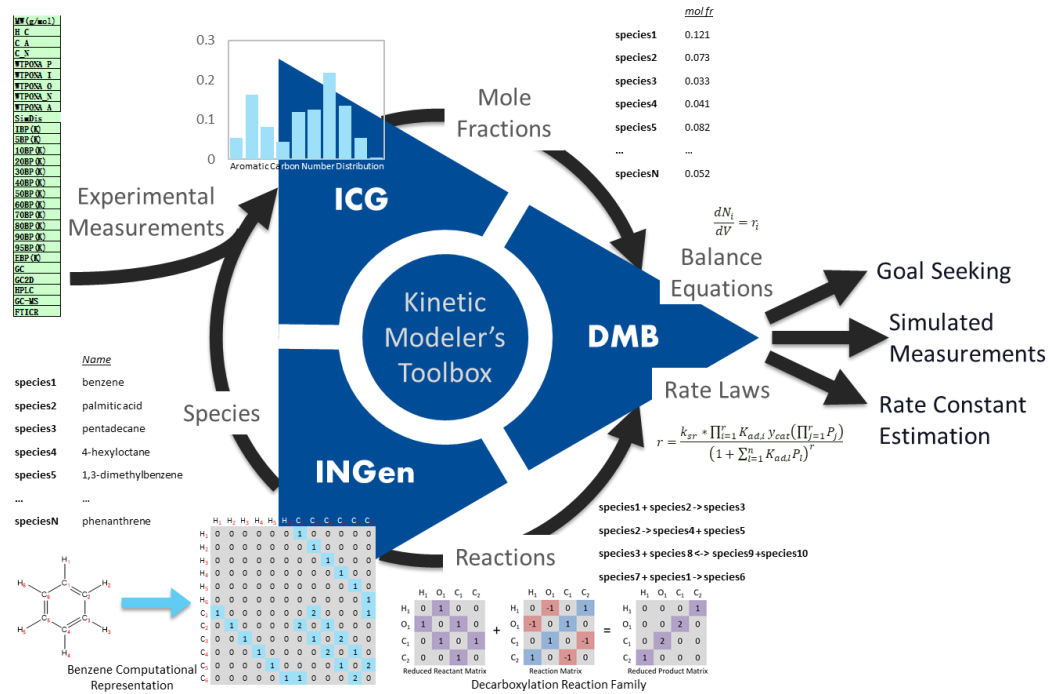


Figure 2.1: The main software tools in the Kinetic Modeler's Toolbox

The typical workflow of developing molecular-level kinetic models starts with the identification of representative feedstock structural moieties or seed molecules in INGen. Next, reactions and reaction rules are selected for the chemistry in question and INGen iteratively searches the seed molecules to generate a list of reactions and a list of species that define the composition. This list of species is an input to ICG, where the properties of the molecules can be calculated using group contribution methods. Additionally, ICG stochastically simulates the probabilities of appearance of the structural moieties in the feedstock to minimize an objective function comparing simulated and experimental properties. The result of ICG is a list of feed mole fractions for the molecule list generated by INGen. The reaction network and the mole fractions define the material balances and initial conditions set up in DMB required to

solve the kinetic model. The output of DMB is the molecular reactor effluent. The molecular effluent can be used to calculate the bulk properties and concentrations of commercial relevance to model users.

2.2 Background of KMT

KMT has been developed over the past three decades in the Klein research group. The software tools evolved out of a desire to automate the generation of reaction networks and kinetic equations for networks that can be tedious to write by hand. Moreover, manually written networks and equations were prone to typographical errors. As such, NetGen (Network Generator), the precursor to INGen, was developed in the 1990s to address automated reaction network generation.^{12,13} From NetGen, the model building capabilities of KMT were expanded with the addition of an interface to NetGen⁸, development of a composition modeling tool⁹, development of an interface for the kinetic model builder^{7,9}, increase in modeling capabilities¹⁴⁻¹⁶, development of analysis applications¹⁰, and addition of new frameworks for existing tools¹¹.

2.3 The Interactive Network Generator (INGen)

INGen is the reaction network generation tool in KMT. The reaction network is the fundamental basis of the rate equations contained in the kinetic model. It is designed as an easy-to-use Microsoft Excel interface with a VBA and C backend containing the program logic. In the interface, users can select the desired molecules in the system, the types of reactions, and any rules for the reaction network generation. For most chemistries, the user does not need to have any expertise in computer science

or any understanding of the program logic. In rare cases though, when the chemistry is fundamentally new, some knowledge and expertise may be needed to

2.3.1 Computational Representation of Molecules

To model explicit molecules, a consistent and efficient method is needed to organize the molecule structure computationally. In KMT, molecules are fundamentally represented as bond-electron matrices. Figure 2.2 shows the computation representation of a 2-butene molecule in KMT. Bond-electron matrices arrange the atoms of a molecule along both the rows and columns of the matrix in the same order. A zero entry in the matrix signifies the absence of a connection between the atoms in the corresponding row and column headings of the matrix. A non-zero entry signifies a connection and the order of the bond between the two atoms. However, storing an entire bond-electron matrix can be very inefficient due to the large number of zero entries relative to non-zero entries. As such, bond-electron matrices are converted to adjacency lists that only store the relevant connections for each atom, as shown in Figure 2.2. Every unique molecule can be represented computationally in this manner.

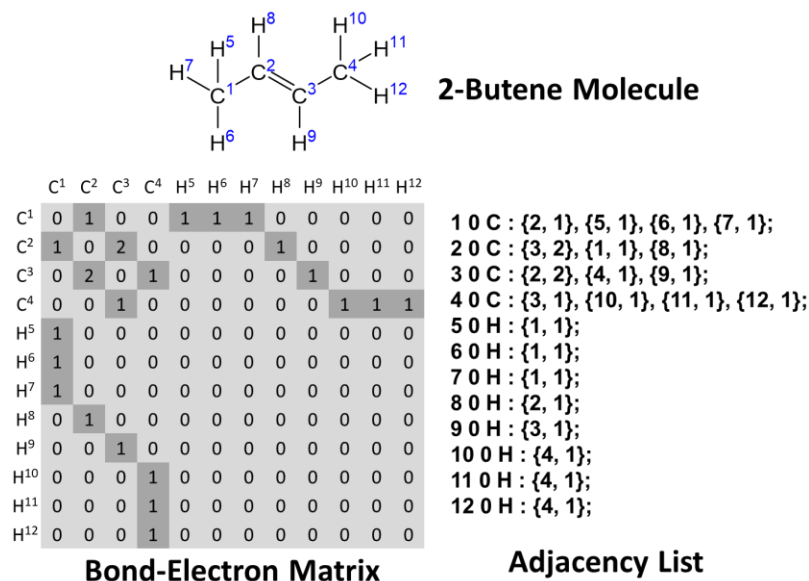


Figure 2.2: Computational representations of a molecule as bond-electron matrices and adjacency lists

2.3.2 Network Seed Molecule Selection

INGen provides a database of molecules representing the typical arrangements of structural attributes in naturally occurring and processed feedstocks. New molecules can be easily added by drawing the molecules in the ChemDraw software or writing the appropriate adjacency list file. The seed molecules should be the logical types of molecules that are present in the feedstock or can be used to create the feed and product molecules through reaction. They do not necessarily need to be the feed molecules, although that may be a good starting point. For example, the seed molecules for a naphtha reforming network can be the paraffins, olefins, naphthenics, and aromatics in the 6-12 carbon range. However, the entire network can also be created by selecting only *n*-dodecene and hydrogen. The cracking, cyclization, aromatization, hydrogenation, and isomerization reactions will lead to the creation of all other feed and product molecules in the naphtha reforming system.

In the case where network size needs to be controlled, selecting the largest molecule and allowing the reactions to generate the network may not work. Molecule seeding can be used as a network reduction technique.¹⁷ INGen allows reactions to be specified by rank, or the order of appearance of molecules in the network. A seed molecule is rank 0, new unique products of the seed molecules are rank 1, new unique products of rank 1 molecules are rank 2, etc. If the reaction rank is limited, then molecules beyond a certain rank will not be created. This can be used to trim the reaction network by seeding the desired molecules in the network and allowing the network to add the reactions between those molecules without adding too many reactions or new molecules, as shown in Figure 2.3.

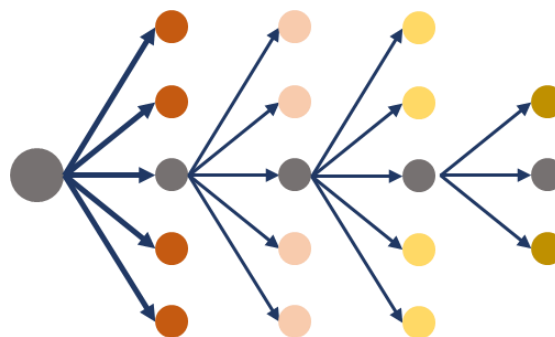


Figure 2.3: Limiting reaction network growth via seeding molecules (grey) along the reaction pathway in a rank 1 network

2.3.3 Computational Representation of Reactions

If molecules are fundamentally matrices, then reactions are matrix addition operations that change the bond connectivities in molecules. For a given type of reaction, the bond breaking and bond forming activity is the same. Thus, a reaction only effects a subset of the atoms in a molecule and the reaction matrices are written to only reflect those atoms. Since each molecule has a unique bond-electron matrix,

the first step is the identification of the subset of atoms in the molecule. The subset can be extracted from the molecule, added to the reaction matrix, and then reimplemented into the original molecule. INGen can parse the new molecule matrix to identify if two or more products exist after reaction and separate the matrix respectively. As an example, Figure 2.4 shows the hydrogenation reaction of an olefinic moiety in INGen as a matrix addition operation. The *R* groups in the molecule represent the entire set of possible molecules that can contain a double bond; therefore, the single reaction matrix in the figure is the same matrix for all double-bond hydrogenation reactions in INGen.

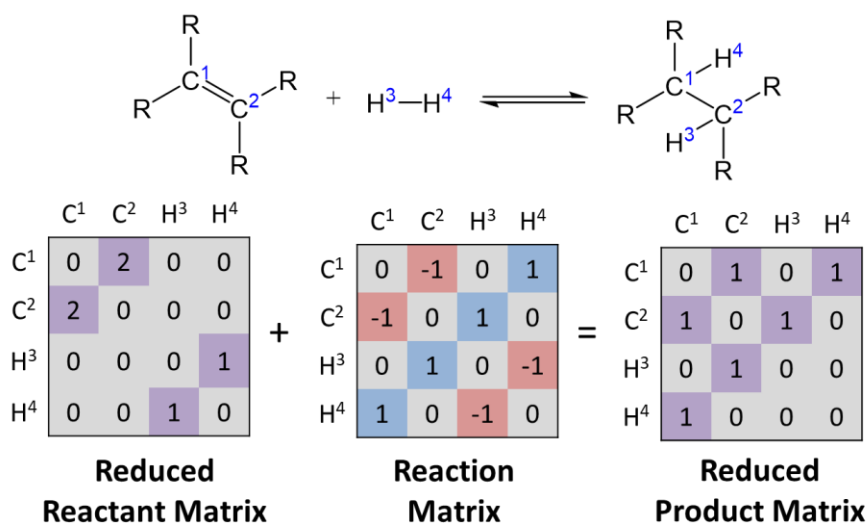


Figure 2.4: Computation representation of a reaction as a matrix addition operation

2.3.4 Reaction Family Selection

The uniqueness of the reaction matrix irrespective of the specific moiety-containing molecule lends to the concept of reaction families. A reaction family is a homologous series of reactions of the same type acting on the same moiety. For any

given process, even one with tens of thousands of individual reactions, the overall types of reactions can be organized into a short list. The specific set of reaction families for a process depend on the process. INGen provides a built-in list of reaction families that cover most traditional refinery processes like reforming, fluidized catalytic cracking, hydroprocessing, and thermal cracking. Additionally, many specialized reactions exist that cover processes like triglyceride hydroprocessing, municipal solid waste gasification, and biomass pyrolysis.

INGen has reaction families defined both mechanistically and at the pathways level. Figure 2.5 shows the difference between the two types for a β -scission reaction. At the mechanistic level, all reactive and intermediate species in a reaction pathway are included. While the mechanisms are very fundamental and truly represent the actual chemistry in a process, they add numerous species and reactions to the reaction network to represent the chemistry. They also contain intermediate radicals and ions, depending on the chemistry, that are highly reactive. The abstraction, ion shift, and radical recombination reactions of the intermediates can cause a combinatorial explosion, significantly increasing the overall size of the network. To mitigate some of the issues of mechanistic networks, pathways-level reactions can be used. Pathways-level reactions are mechanistically informed but only contain the observable reactants and products in a reactions network. This can be an effective way of managing the network size while still capturing the fundamental chemistry in the desired process. The work detailed in this dissertation focuses on pathways-level reaction networks.

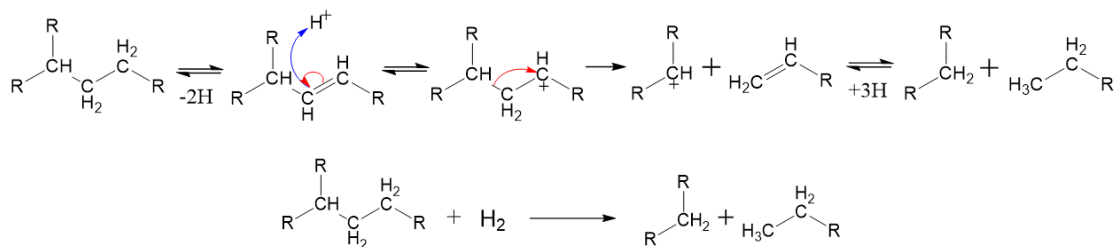


Figure 2.5: a) Mechanism-level and b) pathways-level representations of a β -scission reaction

Selection of reaction families and associated reaction rules limiting the families is the most important step in INGen. The identities of the reaction families are based on the user's knowledge of the chemistry and kinetics of the process being modeled. The reaction families selected determine the types or reactions that can occur in a network on the selected seed molecules. INGen allows users to limit reactions via rules on product rank, carbon number, ring number, double-bond equivalent, and branching. Special care should be afforded to reaction family rules to limit network size. For example, without any limitations, the isomerization of a 25-carbon paraffin can create thousands of species and reactions covering all possible branch lengths and counts. While molecular-level models have the capability of representing all possible isomers, analytical methods usually are not sophisticated enough to characterize each one. Adding that much detail increases the complexity and size of the model without providing useful additional information. A reaction rule limiting the branch length and/or branch number can easily limit the products to a few isomers that represent the isomerization reactions without adding extraneous details.

2.4 The Initial Condition Generator (ICG)

ICG is a streamlined approach to the older generation Composition Model Editor (CME)⁹ software for calculating the mole fractions of individual molecules in the model based on experimental measurements. It allows users to import a list of molecules, set up probability density functions (PDFs) for structural attributes, and juxtapose the attribute probabilities to calculate the individual molecule fractions, as shown in Figure 2.6. The application removes some of the limitations and dependencies of CME on external software packages and gives users control over the setup of the PDFs and optimization problem. This allows users without computer science knowledge or modeling expertise to quickly model the feedstocks using collected experimental data. ICG has been utilized to model feeds from naphtha to vacuum resid. As ICG was developed during the course of this dissertation, further information about the algorithms and functionality of ICG is discussed in detail in Chapter 6.

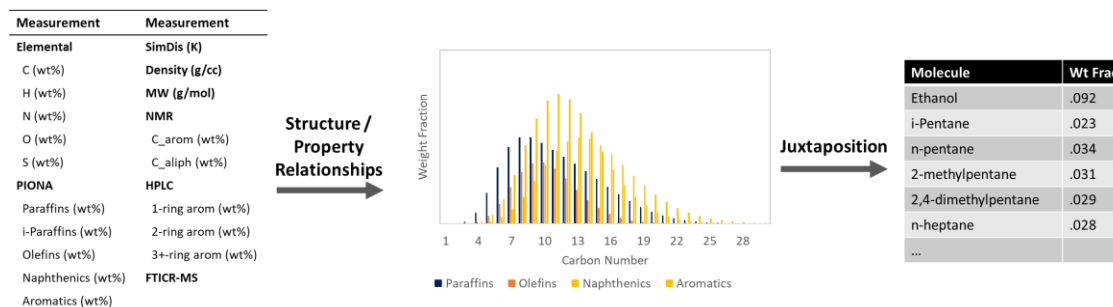


Figure 2.6: Workflow of the Initial Condition Generator tool

2.5 The Dynamic Model Builder (DMB)

DMB is the newest member of KMT designed to build and solve the kinetic model. It organizes the material balances based on a reaction network and reactor type

selection. The aim of DMB was to replace the Kinetic Model Editor (KME) software due to the limitations of KME when dealing with large networks. In terms of functionality and algorithms, DMB is essentially the same as KME. The real power of DMB is its ability to handle large networks with additional functionality for multiple reactors, multiple catalysts, quenches, and side-by-side reaction and vapor-liquid equilibrium. Additionally, DMB is a C++ software compiled in Microsoft Visual Studio, allowing it to be deployed without the Cygwin environment required by KME.¹¹ It can therefore be used as an easily deliverable executable called from different user-friendly applications that customize the user experience.

The basic kinetic model is an initial value problem defined by Equation 2.1. In the equation, \bar{y} is a vector of the molar flows of each component in the system, \bar{y}_{in} and \bar{y}_{out} are the flow in and out of the system, \bar{v}_i is the vector of stoichiometric coefficients, and $rate_i$ is the rate of reaction. \bar{y}_{in} and \bar{y}_{out} are functions of the selected reactor type, where they may be zero in the case of a batch reactor or non-zero flow rates in cases of reactors like plug flow reactors (PRFs), continuous stirred-tank reactors (CSTRs), and microactivity test (MAT) units. The stoichiometric coefficients are directly parsed from the reaction list. Lastly, the rate laws are automatically generated based on user selection of rate law type.

$$\frac{d\bar{y}}{dt} = \bar{y}_{in} - \bar{y}_{out} + \sum_{i, reactions} \bar{v}_i * rate_i \quad (2.1)$$

$$\bar{y}(0) = \textit{feed composition}$$

Different types of rate laws are built into DMB. Of main interest are the microkinetic, power law, and Langmuir-Hinshelwood-Hougen-Watson (LHHW) kinetics. Each type of rate law has its own kinetic formulation defined in DMB.^{11,18} To determine the rate laws, DMB parses the reaction network to find the participating

reactants and products along with their stoichiometric ratios. It also determines whether the reactions are reversible or irreversible. It then calculates each rate via the concentrations, flow rates, partial pressures, kinetic constants, adsorption, and thermodynamic properties as they appear in the selected rate law type.

2.5.1 Kinetic Parameter Definition and Minimization

For each rate law, one or more kinetic constants must be defined to calculate the value of the rate. There are three main kinetic constants of interest: the Arrhenius rate constant, the equilibrium constant, and the adsorption equilibrium constant. Regardless of the type of rate law, every reaction has a rate constant. Reversible reactions have equilibrium constants. For a reaction on a catalyst, a rate law of the LHHW form contains adsorption constants to define the adsorption and desorption ability of molecules participating in a reaction. When developing a kinetic model for a reaction network containing thousands of reactions and species, thousands of individual rate constants, equilibrium constants, and adsorption constants would be needed. If the reactor system contains multiple catalysts, the number of parameters would proportionally increase. However, the data requirement and optimization burden for that many parameters would make the solution intractable. Hence, DMB employs some parameter minimization techniques to reduce the overall number of kinetic parameters.

$$\ln k_i = \ln A_i - \frac{E_i^0}{RT} \quad (2.2)$$

$$\ln k_{i,f} = \ln A_f - \frac{E_{0f} + \alpha_f \Delta H_i}{RT} \quad (2.3)$$

The standard Arrhenius equation shown in Equation 2.2 contains two kinetic parameters for each equation, $\ln A_i$ and E_i^0 . Thousands of rate equations would require

twice as many parameters. To minimize the parameters, DMB uses the Arrhenius equation with the Bell-Evans-Polanyi linear free-energy relationship (LFER) for the activation energy, as shown in Equation 2.3.^{19,20} The LFER concept exploits the systematic differences in the reaction rate between member reactions of a reaction family. The systematic differences can be organized as a linear dependence based on a reaction index, which in this case is the heat of reaction calculated from pure component thermodynamic properties. Figure 2.7a displays this concept, where rates of the alkylaromatic dealkylation reactions can be linearly organized by the heat of formation of the carbenium ion. Using the LFER concept, the number of parameters for the rate constant can be reduced from two per reaction to three per type of reaction. As there are only a few types of reactions in any given process, the number of rate constant parameters is greatly reduced.

$$\ln A_{f,c} = \ln A_{f,c_{ref}} + \Delta \ln A_{f,c_{ref} \rightarrow c} \quad (2.4)$$

$$\ln k_{i,f,c} = \ln A_{f,c} + \Delta \ln A_{f,c_{ref} \rightarrow c} - \frac{E_{0f} + \alpha_f \Delta H_i}{RT} \quad (2.5)$$

But, since every catalyst will have its own kinetic rate parameters, a reactor system with multiple different catalyst layers would require the three LFER parameters for each reaction family on each catalyst. Figure 2.7: Figureb shows how the LFER concept can be extended to catalysts. The difference in activity of a reaction on two different catalysts can be modeled as a constant departure from one catalyst to another. While the major difference between different catalysts is often the reaction activation energy, the differences can be modeled by a constant departure term on the $\ln A$ factor for narrow temperature ranges.²¹ Extending the LFER concept to create catalyst LFERs then states that for all reaction families j , a $\Delta \ln A_{j,c_{ref} \rightarrow c}$ term can capture the reactivity differences of a reaction family on each catalyst c from the base

parameter $\ln A_{j,c,ref}$, as shown in Equation 2.4. The final equation for the surface rate constant is then provided in Equation 2.5 for reaction i on catalyst c that belongs to reaction family f .

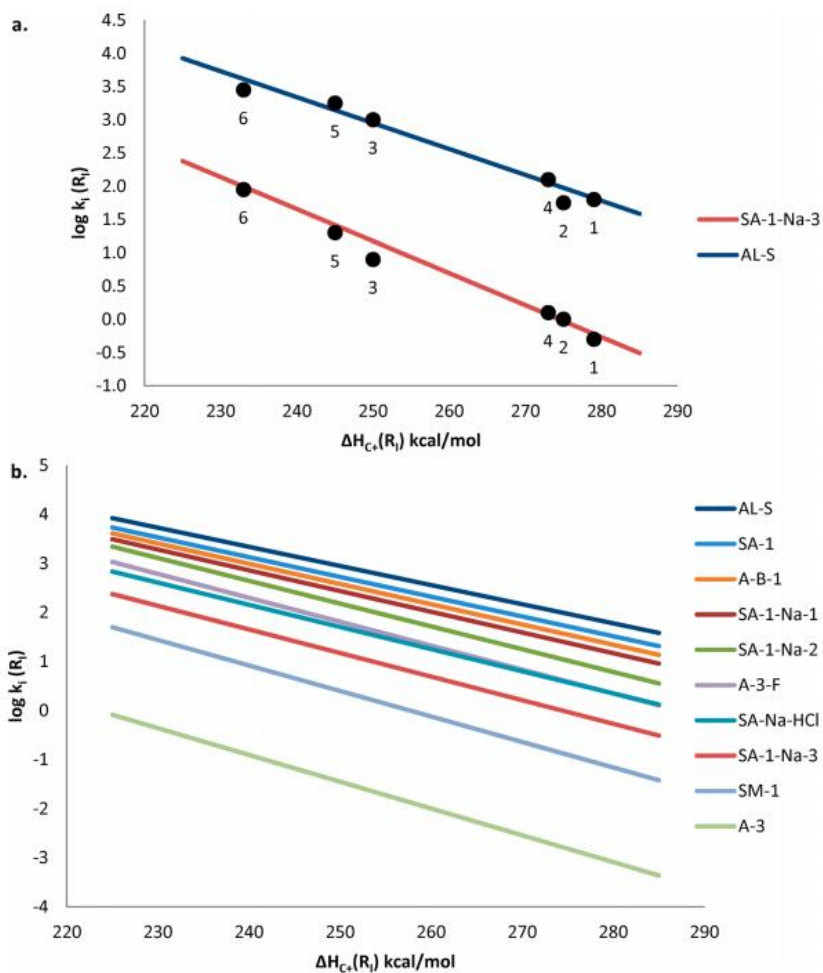


Figure 2.7: Figure from Horton and Klein²¹ (a) Figure from Mochida and Yoneda²² showing the dealkylation of the alkylaromatic reaction family. Each line represents a different catalyst, where the data points are particular reaction measurements. (b) Figure generated from reaction family parameters from Mochida and Yoneda for 10 total catalysts

Next, the equilibrium constants for each reaction are calculated using the standard thermodynamic formulation shown in Equation 2.6. The Gibbs free energy is calculated for each reaction at each temperature from the pure component properties and thermodynamic state functions. Lastly, the adsorption constants for each molecule are calculated using a quantitative structure/property relationship (QSPR) given in Equation 14. The QSPR defines adsorption based on the structural attributes of the molecules that have been shown to work well for hydroprocessing systems by Korre et al.²³ For each type of site in the system, one set of b_1 to b_7 constants are needed. In this manner, the adsorption constants for all species in the system can be defined by seven constants for each type of site in the system.

$$\ln K_{eq,i} = -\frac{\Delta G_i}{RT} \quad (2.6)$$

$$\ln K_{ad,k} = b_{1,k} + \frac{b_{2,k}N_{AR} + b_{3,k}N_{NR} + b_{4,k}N_{SC} + b_{5,k}N_S + b_{6,k}N_N + b_{7,k}N_O}{RT} \quad (2.7)$$

2.5.2 Parameter Optimization

The final step in building a kinetic model in DMB is the optimization of the kinetic parameters. An objective function of the form provided in Equation 2.8 is used to measure the model goodness of fit. The objective function is a sum-square of errors of the experimentally measured values and the model predictions for each experiment q scaled by a weighting factor. The objective function is summed for all datasets d since all datasets employ the same set of kinetic parameters being optimized. An experiment for an objective function can be bulk property measurements like density, simulated distillation, and elemental analysis or molecule yields for specific molecules in the product. The molecular-level kinetic model calculates the molecular effluent,

and any experimentally measured value can be calculated from the molecular effluent as long as an appropriate structure/property relationship is available.

$$obj = \sum_d \sum_q \left(\frac{Exp_{d,q} - Pred_{d,q}}{Weight_{d,q}} \right)^2 \quad (2.8)$$

The major optimization routine used to minimize the objective function in DMB is simulated annealing. Simulated annealing is a global optimization method designed to find the global minima of problems with a large number of parameters. It begins by randomly sampling different parameter values within the specified lower and upper bounds of the kinetic parameter ranges. As the simulated temperature in the algorithm decreases, the parameter range shrink to fine tune the answer around a small range. The major characteristic of simulated annealing is that it has a probability to allow uphill steps, or steps that increase the objective function. This gives the algorithm a chance to escape local minima to find better global minima. A solution method like gradient descent that only allows downhill steps would necessarily only ever reach the closest local minima to the initial value of the objective function.

Special consideration should be afforded to the kinetic parameters in large systems. While automated tuning algorithms like simulated annealing do provide good solution methods, if the parameter range is too broad or the initial parameter values too far from a good solution, the automated algorithm usually cannot produce any good results. For example, an initial overall reaction rate that is sufficiently close to zero with a parameter range that also results in reaction rates close to 0 will never produce meaningful changes in the objective function, especially if the real parameter value is still outside of the specified parameter range. As a general heuristic for modeling large systems, kinetic parameters are manually optimized to reach the same order of magnitude values of the product. Then, the simulated annealing algorithm is

used to optimize the final parameter values. The final kinetic parameters should provide a good agreement between the experimental data and model predictions.

2.6 Summary

KMT is a software suite designed to build and solve kinetic models. Model development occurs in three steps: 1) seed and reaction selection in INGen, 2) initial composition values from PDF sampling in ICG, and 3) writing and solving the material balance in DMB. Each software is designed to provide user-friendly interfaces that allow model builders to create models without computer science knowledge or expertise in kinetic modeling. The final output from the kinetic model is a set of kinetic parameters that can be used for the prediction of product properties over a range of process conditions and feedstock variations.

Chapter 3

MOLECULAR-LEVEL KINETIC MODELING OF TRIGLYCERIDE HYDROPROCESSING

Pratyush Agarwal¹, Sulaiman S. Al-Khattaf², and Michael T. Klein^{1,2}

*¹Department of Chemical and Biomolecular Engineering, University of Delaware,
Newark, DE 19716*

*²Center for Refining and Petrochemicals, King Fahd University of Petroleum and
Minerals, Dhahran, Saudi Arabia*

3.1 Abstract

A molecular-level kinetic model was developed for triglyceride hydroprocessing. Triglyceride molecules were defined based on the 8 to 22 carbon fatty acids commonly present in renewable diesel feeds. A reaction network detailing the hydrodeoxygenation, decarboxylation, and decarbonylation parallel pathways of the triglyceride chemistry was constructed. The final network contained 476 species and 1709 reactions. The network was used to build a kinetic model based on experimental data for coconut and soybean oil hydroprocessing at various temperatures, pressures, and catalyst contact times. Parameter optimization for the kinetic parameters was performed for two different catalysts. The final kinetic model provided good agreement with experimental results. Diesel cetane number and cloud point property models were also constructed and optimized based on experimental data. These property models were used to study the product diesel cetane number versus cloud point tradeoff to determine the end-use properties of the product fuel.

3.2 Introduction

Environmental concerns and a desire to reduce fossil fuel use have led to the implementation of policies mandating the increase of renewables like the Renewable Fuel Standard in the United States and the 2009/28/EC directive by the European Union. These policies can often strain refineries; for example, US merchant refiners that do not integrate at least equal volume production of renewable products and petroleum products pay for the price of renewable identification numbers (RINs).²⁴ Therefore, alternative fuels have been an area of significant interest for the oil industry. The production of renewable diesel and jet fuel from hydroprocessing triglycerides has been shown to be a viable renewable alternative to traditional petro-

crude oil fuels.²⁵⁻²⁷ Triglyceride-based green fuels exhibit good combustion properties, high oxidation stability, and low impurity content while having processing flexibility in the specific feedstock oil mixture.²⁸⁻³¹ Although triglycerides could directly be used as diesel engine fuels, they result in lower engine performance and increased particulate and CO production.³²

Triglycerides present in plant and algal oils are usually a combination of different straight-chain fatty acid moieties containing 8 to 22 carbons. A large majority of the hydroprocessed products from triglycerides therefore closely resemble *n*-paraffins and *i*-paraffins in the diesel range. This paraffin mixture has an excellent cetane number: an important ignition property for combustion of fuels in a diesel engine. The minimum cetane number for commercial diesel is typically 40 in North America³³ and 51 in Europe³⁴, but triglyceride-based paraffin mixtures typically have cetane numbers close to 100. However, the mixture also has a very high cloud point temperature, often above room temperature, at which solid crystals begin to form in a liquid. While the minimum cloud point standards are vague^{33,34}, solidification of fuels in automobile engines and tanks can greatly reduce marketability of the fuel. As such, a careful balance between processing costs, cetane number, and cloud point temperature must be maintained based on end-use of the product paraffin mixture. Modeling the production of the mixture with appropriate property correlations can greatly reduce the experimental burden for such an optimization.

Some attempts have been made to model triglyceride hydroprocessing via thermodynamic equilibrium³⁵ and component lumps.^{36,37} These models fail to characterize and predict carbon-number based experimental data fully. Triglyceride mixtures vary regionally and seasonally with different ratios of the fatty acid

constituents based on the sources of the oils²⁶, so, clearly, feedstock-lump independent models are needed. Additionally, each molecule has an independent reactivity that better represents the real hydroprocessing mixture than a lump. These reactivities directly correlate to the selectivities of different carbon number paraffins in the product. Since the paraffin products can be analytically classified by separate carbon numbers and molecule types, modeling the individual molecules is important for better product predictions.

Once the product distributions can be predicted, accurate cetane number and cloud point structure/property models are also needed to effectively use the models for commercial applications. Cloud point and cetane numbers of pure components both vary by molecule type and molecule size. Ghosh and Jaffe displayed the significant variation of pure component cetane numbers for different molecule types over the diesel carbon number range.³⁸ The cloud point of a mixture is also directly correlated to the melting temperature of its individual molecular constituents. Normal paraffins are thermodynamically most likely to dictate the cloud point of a diesel mixture due to their high melting points. For a paraffinic mixture, Figure 3.1 shows that pure component melting temperatures can vary not only by carbon number and branching, but also by the specific location of the branch on the paraffin. These molecule-to-molecule variations for different isomers illustrate the importance of modeling the specific molecules of the paraffinic product from triglyceride hydroprocessing to better predict the properties of commercial interest.

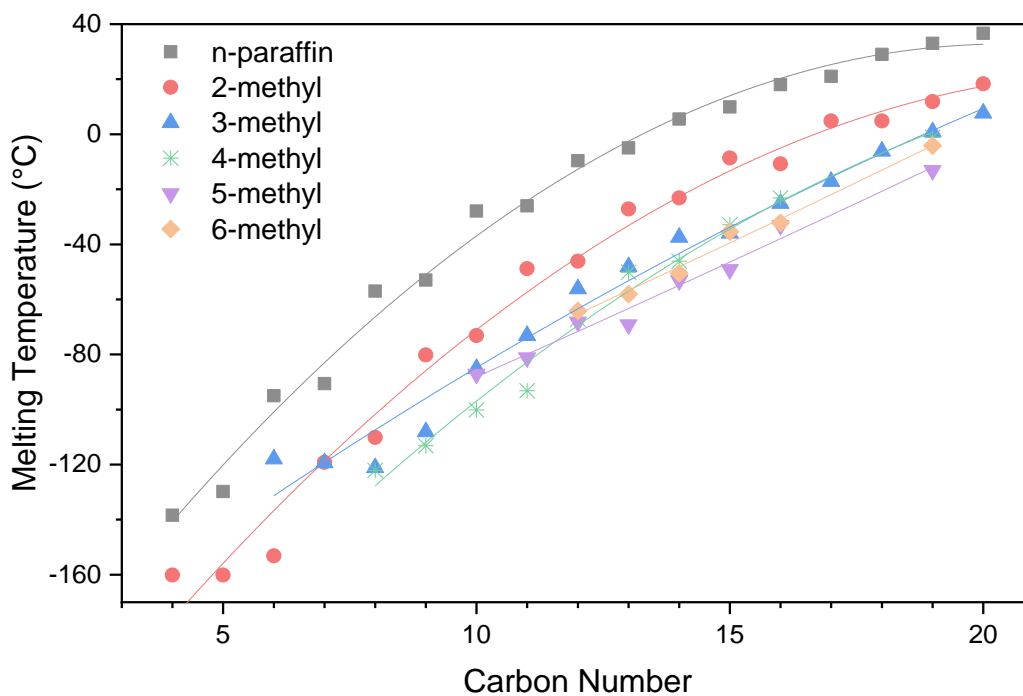


Figure 3.1: Pure-component melting points for different n-paraffins and i-paraffins. Data from the NIST Chemistry Webbook³⁹. The lines illustrate general trends of melting temperature with carbon number for different paraffin classes.

This present work builds upon previous modeling work⁴⁰ on methyl laurate, a model compound for coconut oil, to develop a molecular-level kinetic model for triglycerides based on recently published experimental data^{6,41} on coconut oil and soybean oil hydroprocessing. The experimental work used alumina supported NiMo and CoMo catalysts to hydroprocess vegetable oils at different temperatures, pressures, and liquid hourly space velocities (LHSV).^{6,41} Using an in-house software suite, the Kinetic Modeler's Toolbox (KMT)^{7,8}, a reaction network for the triglycerides present in vegetable oils was constructed and used to make a kinetic model. Rate constants were adjusted to account for the changing process conditions in

the experimental work to satisfy the product distributions. Finally, cloud point and cetane number correlations were constructed to define the end-use properties of the product fuel mixture.

3.3 Reaction Network Generation

The overall hydroprocessing chemistry and reaction routine selections have been explained in detail in previous work using methyl laurate as a model compound⁴⁰ and analytical measurements of observable molecules in the reaction network.⁶ Figure 3.2 shows a general representation of the overall reaction network. The reaction network defines the decarbonylation, decarboxylation, and hydrodeoxygenation activity of the ester bonds present in the feed molecules. In the decarbonylation and decarboxylation routes, a paraffin with one less carbon than the corresponding fatty acid chain in the triglyceride is produced directly as a carbon is lost to CO or CO₂. The hydrodeoxygenation route preserves the carbon content of the fatty acid chain with the downside of a higher hydrogen consumption. It proceeds through carboxylic acid, aldehyde, and alcohol intermediates that sequentially undergo hydrodeoxygenation. Additional pathways for paraffin isomerization, paraffin cyclization, and aromatization were added to represent the experimentally measured products.

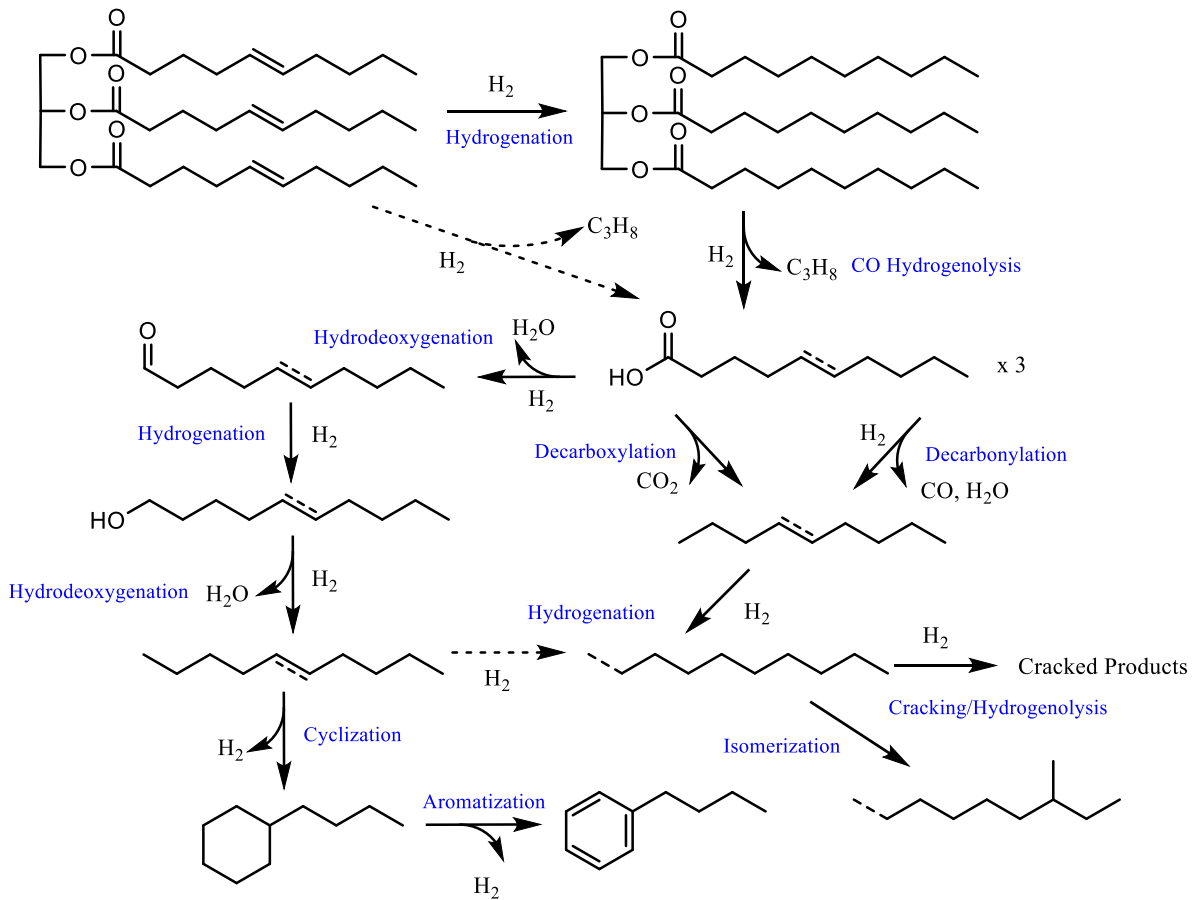


Figure 3.2: Reaction network for triglyceride hydroprocessing

The triglyceride molecules were assumed to have three identical branches where the triglyceride distribution corresponds to the reported fatty acid distribution of the oil.⁴¹ This resulted in the definition of 13 triglycerides with C₈, C₁₀, C₁₂, C₁₄, C₁₆, C_{16:1}, C₁₈, C_{18:1}, C_{18:2}, C_{18:3}, C₂₀, C_{20:1}, and C₂₂ fatty acid chains for the feed. Each of the triglyceride chains has an analogous fatty acid, aldehyde, alcohol, and paraffin in the network. Removal of chains from the propane backbone of the triglyceride was considered in only one sequence to reduce the combinatorial load of multiple routes. Hydrogenation was allowed to occur on the triglyceride and the olefin products. All

paraffin products from C₁ to C₂₂ were defined, with the *i*-paraffins being limited to two methyl branches. These paraffin products could subsequently crack to form the lighter products but cracking directly from the triglyceride or other intermediates was not allowed. The n-paraffins, but not the isoparaffins, were also allowed to cyclize and aromatize to form the observed cycloparaffins and aromatics.

Systematic generation of the reaction network was done computationally using the Interactive Network Generator (INGen).^{7,8} INGen iteratively searches for specific reactive subgroups in computational representations of molecules and applies the desired reaction algorithm to generate products. The detailed list of reaction algorithms for the triglyceride hydroprocessing network, represented as matrices in INGen, has been discussed in earlier work.⁴⁰ This model used the 13 aforementioned triglycerides along with hydrogen as seeds to the reaction network. The final network statistics based on the selected reaction scheme for triglyceride hydroprocessing are listed in Table 3.1. The network contains a total of 476 species and 1709 reactions which are used to make the kinetic model. The average network building time on a computer running Windows 10 with an Intel i7-4770 CPU (@3.40 GHz) and 16GB RAM was 42 seconds.

Table 3.1: Network representation for coconut oil hydroprocessing generated using the Interactive Network Generator (INGen)

Species Type	Number	Reaction Type	Count
Triglyceride	13	CO Hydrogenolysis	39
Diglyceride	13	Hydrodeoxygenation	65
Monoglyceride	13	Decarbonylation	26
Fatty Acid	13	Decarboxylation	13
Aldehyde	13	Hydrogenation	64
Alcohol	13	Paraffin Isomerization	658
<i>n</i> -Paraffin	22	Paraffin Cracking	121
<i>i</i> -Paraffin	271	Paraffin Cyclization	17
Olefin	67	Aromatization	17
Cycloparaffin	17	CC Hydrogenolysis	688
Aromatic	17	Methanation	1
CO, CO ₂ , H ₂ O, H ₂	4		
Total Species	476	Total Reactions	1709

3.4 Feed Specifications

The feed to the reactors is a mixture of triglycerides with fatty acid chains containing 8 to 22 carbons. Since the feed oils can change seasonally and regionally, the kinetic model should be able to account for different feeds. One of the benefits of a molecular-level mode is that each molecule is defined independently rather than having feedstock dependent lumps. Differentiating the different feeds of the triglyceride-containing oils as coconut oil, soybean oil, jatropha oil, etc. requires a change of the different fatty acid concentrations in the oil. Furthermore, variations in compositions of a specific type of oil can also be captured by changing the feed fatty acid profile. Therefore, the same reaction network applies to a wide range of possible feedstocks to a triglyceride hydroprocessing unit.^{6,41,42} Although most common vegetable oils can be well-represented by the current network, some of the rarer fatty acid chains present in oil mixtures can be encountered occasionally. The utility of the INGen tool allows for fast reconstruction of a new model in case the feed specification

changes significantly beyond the current definition. For the current work, Table 3.2 defines the fatty acid ratios for coconut oil and soybean oil.

Table 3.2: Fatty acid composition of coconut oil⁴¹ and soybean oil⁶. Data directly from source.

Fatty Acid	Designation	Coconut Oil (wt%)	Soybean Oil (wt%)
Octanoic	C8:0	9.23	0
Capric	C10:0	7.80	0
Lauric	C12:0	49.32	0
Myristic	C14:0	16.87	0.08
Palmitic	C16:0	8.19	10.98
Palmitoleic	C16:1	0	0.11
Stearic	C18:0	8.59	4.39
Oleic	C18:1	1.56	23.98
Linoleic	C18:2	5.47	52.56
Linolenic	C18:3	1.56	6.76
Arachidic	C20:0	0	0.43
Gadoleic	C20:1	0	0.22
Behenic	C22:0	0	0.49

3.5 Model Equations and Kinetics

An in-house software tool, the Kinetic Model Editor (KME)⁷, was used to set up and solve the kinetic model. KME systematically constructs material balances for each component in the reaction network that can be used to define an initial value problem. Details of the model equations and techniques have been explained in previous work.⁴⁰ In general, a hydroprocessing unit with a bifunctional catalyst was modeled using Langmuir-Hinshelwood-Hougen-Watson (LHHW) rate laws of the form displayed in Equation 3.1. To minimize the kinetic parameters, the concepts of linear free-energy relationships (LFERs) and quantitative structure-reactivity relationships (QSRRs) were applied. These LFERs, as shown in Equation 3.2, define

the Arrhenius rate parameters of reactions i belonging to reaction family j , a homologous series of reactions, as a linear correlation with the heat of reaction as the reaction index.^{19,20} This reduces the number of parameters from two per reaction to three per reaction family. QSRRs, as shown in Equation 3.3, can define adsorption equilibrium constants based on components of individual molecules, thereby requiring one set of parameters per type of site on a catalyst. The adsorption QSRRs have been shown to work well for hydroprocessing systems.²³

$$r = \frac{k_{sr} * \prod_i^{reactants} K_{ad,i} * y_{cat} \left(\prod_i^{reactants} P_i - \frac{\prod_j^{products} P_j}{K_{eq}} \right)}{P_{H_2}^{\alpha,k} (1 + \sum_l^{species} K_{ad,k} P_l)^n} \quad (3.1)$$

$$k_{sr,i,j} = \ln A_j - \frac{E_j^0 - \alpha_j \Delta H_i}{RT} \quad (3.2)$$

$$\ln K_{ad} = a + \frac{bN_{Aromatic} + cN_{Naphthenic} + dN_{Carbon} + eN_{Oxygen}}{RT} \quad (3.3)$$

As an initial evaluation, an extrapolation from the kinetic parameters of the catalyst in previous work⁴⁰ were used in the current work based on the catalyst family concept. In the catalyst family concept, the changes in reactivity between different but similar catalysts are modeled as constants. This was used to represent the reactivity of the 9.1 wt% Ni, 5 wt% Mo in the coconut oil experimental setup⁴¹ based on a departure from the rate constants of the different Ni:Mo catalysts examined in the previous study. For the soybean oil experimental setup, a new set of $\ln A$ parameters were adjusted to account for the CoMo catalyst used in the study. Adjustments of the kinetic parameters accounted for the temperature variations and process changes in the coconut and soybean oil reactor systems.

3.6 Kinetic Model Evaluation

The kinetic parameters were optimized by minimizing the differences between the kinetic model predictions and experimental data of Kimura et al. and Kim et al.^{6,41} An objective function of the form defined in Equation 3.4 was used to quantify the differences between the experimental and predicted results weighted by an approximate standard deviation of the measurement. A simulated annealing algorithm was used for the optimization of the kinetic parameters, where the α_j parameters were set to a constant value of 0.2. This value sufficiently models the differences in rate between member reactions of a reaction family. The experimental data used to optimize consist of 1) five datasets for coconut oil defined at different catalyst contact times and 2) six datasets covering three different temperatures and four different pressures for soybean oil. The average model simulation time on a computer running Windows 10 with an Intel i7-4770K CPU (@3.40 GHz) and 16GB RAM was ~0.9 s for a once-thru simulation.

$$obj = \sum_{set} \sum_{exp} \left(\frac{y_{obs} - y_{pred}}{\sigma_y} \right)^2 \quad (3.4)$$

Table 3.3: Reaction parameters for the linear free-energy relationship that define the kinetic rate constants in Equation 2. α for all reaction families was kept constant at a value of 0.02

Reaction Family	logA_{NiMo}	logA_{CoMo}	E₀ (kJ/mol)
Aromatization	-	6.48	21.2
CO Hydrogenolysis (Ester → Aldehyde)	5.77	2.05	84.0
CO Hydrogenolysis (Ester → Carboxylic Acid)	1.01	2.74	31.3
Decarbonylation (Aldehyde)	-1.07	5.89	13.3
Decarbonylation (Carboxylic Acid)	0.87	7.36	27.5
Decarbonylation (Ester)	6.87	4.45	72.8
Decarboxylation (Carboxylic Acid)	-	9.85	5.77
HDO (Alcohol → Paraffin)	5.59	5.04	16.8
HDO (Carboxylic Acid → Alcohol)	0.99	5.36	10.2
HDO (Carboxylic Acid → Aldehyde)	7.85	4.44	83.7
Hydrogenolysis (Paraffin/Olefin)	15.8	14.7	198
Methanation	-0.36	-1.02	49.4
Paraffin Isomerization (Paraffin)	-0.61	2.41	16.2
Paraffin Cracking (Paraffin)	4.57	1.01	35.9
Paraffin Cyclization	-	3.57	4.90
Hydrogenation (Triglyceride)	0.52	-1.09	47.3
Hydrogenation (Olefin)	-	3.69	36.9
Hydrogenation (Aldehyde)	7.96	4.45	76.7

The kinetic rate constants for the individual reaction families based on the LFER concept are defined in

Table 3.3. An independent set of logA parameters was used for each of the NiMo and CoMo catalysts used for coconut oil and soybean oil, respectively. While

the activation energy theoretically differs between two different catalysts, the approach of catalyst LFERs, where the $\log A$ is perturbed, has been shown to work well for different catalysts in narrow temperature ranges.²¹ Since temperature-varied experimental data were not available for the coconut oil, the activation energy was determined by the soybean oil data on the CoMo catalyst. The paraffin cyclization and aromatization reactions were unimportant in the coconut oil case because the requisite product species for those reactions were not present in the measured data. The exponents for the hydrogen adsorption term $P_{H_2}^{\alpha,k}$ in Equation 3.1 were -0.273 and -0.134 for the acid and metal sites, respectively. Table 3.4. contains the parameters for the adsorption constants defined by Equation 3.3.

Table 3.4. Adsorption constants for the kinetic model defined in Equation 3

Site	a	b	c	d	e
Acid	0.182	1.934	1.134	0.00587	0.0097
Metal	1.324	0.887	0.487	0.00523	0.0083

The model results from parameter optimization have good agreement with the experimental results in both the coconut oil and soybean oil cases. Conversion, based on the experimental^{6,41} definitions, is simulated well by the model and can be seen in 3.3 for both the coconut oil and soybean oil for the various process conditions. The correct trends are followed with increasing contact time (datasets 1-5), increasing temperature (datasets 6-8), and increasing pressure (datasets 9, 10, 7, and 11). While only paraffinic product was observed for the coconut oil case, intermediates and minor products were measured for the soybean oil case with a different catalyst. Figure 3.4 shows that the bulk oxygenate, aromatic, cycloparaffin, olefin, isoparaffin, and n-

paraffin fractions are correctly modeled. Additionally, the congruent gas phase composition is correctly modeled for the soybean oil, as shown in Figure 3.5. These figures show that the molecular model can predict results at least as well as a lumped model while being feedstock independent and having only a single set of parameters per type of catalyst for a wide range of process conditions.

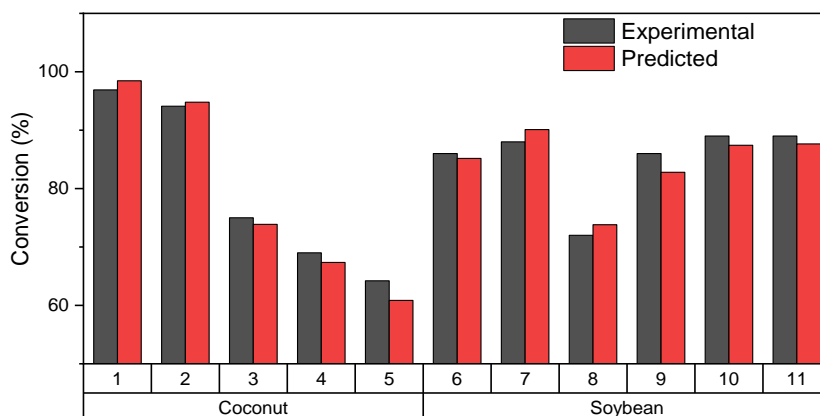


Figure 3.3: Comparison of the predicted and experimental^{6,41} overall conversions of the coconut and soybean oils at a range of temperatures, pressures, and catalyst contact times

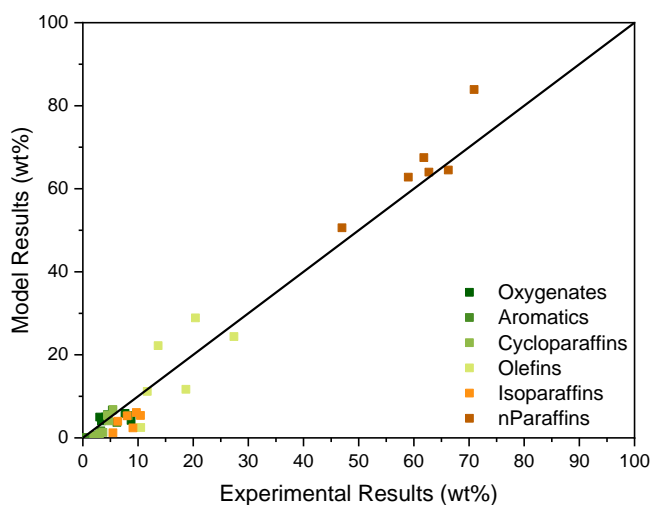


Figure 3.4: Parity plot showing the comparison between the experimental⁶ and predicted bulk fraction concentrations of oxygenates, aromatics, cycloparaffins, olefins, isoparaffins, and n-paraffins for soybean oil hydroprocessing at 350-440°C and 4.5-12 MPa

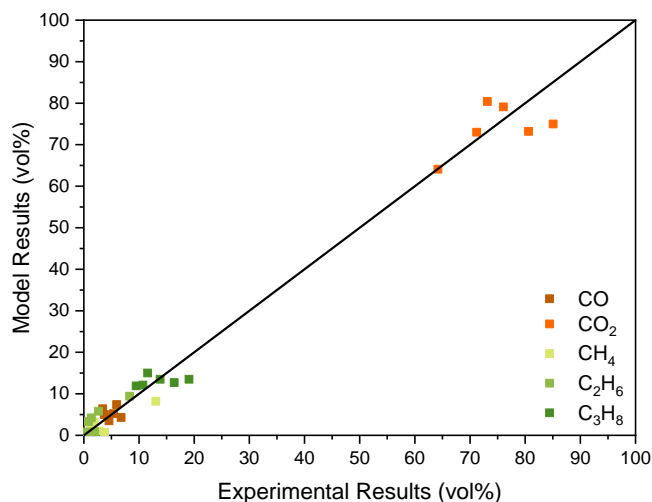


Figure 3.5: Parity plot showing the comparison between the experimental⁶ and model predictions for the gas phase concentrations of CO, CO₂, CH₄, C₂H₆, and C₃H₈ for soybean oil hydroprocessing at 350-440°C and 4.5-12 MPa

A parity plot showing the molecular product comparison can be seen in Figure 3.6. The model had some difficulty predicting experimental results that did not display a consistent trend with changing contact time, temperature, or pressure. For the coconut oil, the *n*-dodecane and propane formation did not follow a trend with increasing contact time, resulting in prediction errors. The greatest error is visible in the highest and lowest contact time datasets. A preferentially higher weighting of the dodecane and undecane formation along with conversion was applied to the objective function defined in Equation 3.4. This allowed for greater accuracy for the majority products at the expense of some accuracy in the cracked products. For the soybean oil

case, the selectivity of C17 and C18 paraffin to olefin ratios proved to be difficult to accurately predict due to inconsistencies with expected trends. The low-pressure case (4.5 MPa) contained the most error. The prediction errors may stem from both the accuracy of the measurement and the calculation of the total hydrogen moles from the provided hydrogen partial pressure.

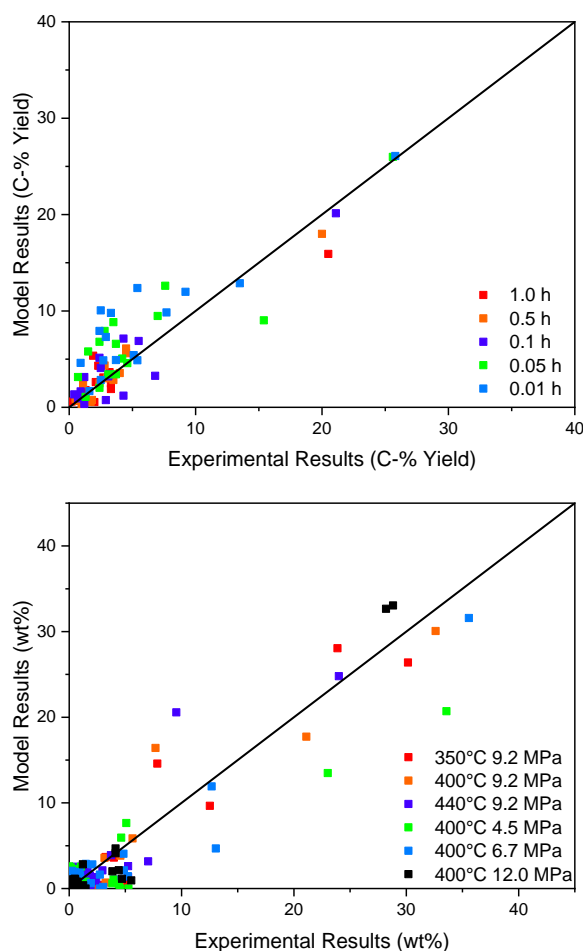


Figure 3.6: Parity plot showing the comparison between experimental^{6,41} results and model predictions of products from a) coconut oil hydroprocessing at 350°C and 0.8 MPa for five different contact times and b) soybean oil hydroprocessing at 350-440°C and 4.5-12.0 MPa for the same contact time

3.7 Diesel Property Calculation

After calculating the molecular product composition, bulk properties can be calculated using structure/property relationships. In this work, methods to calculate cetane number and cloud point were implemented and used to predict the end-use of the outlet composition. Both cetane number and cloud point can vary significantly owing to the individual molecules that are present in a mixture. Therefore, the molecular output from the kinetic model provides an excellent basis for the property calculation method.

3.7.1 Cetane Number Model

Cetane number calculations were performed using Equation 3.5 from the model developed by Ghosh and Jaffe³⁸. The CN_i descriptions from the paper were used exactly, but the set of β_i values for the individual molecule lumps were not reported by the authors for proprietary reasons. These were optimized using experimental data. The experimental data includes nine traditional diesel blends⁴³ and 11 paraffinic mixtures⁴⁴ which suitably represent the products from triglyceride hydroprocessing. A Levenberg-Marquardt-Fletcher algorithm was used to optimize the β_i values following the parameter optimization strategy of Ghosh and Jaffe. Since the Ghosh and Jaffe model assumes a derived cetane number definition with *n*-hexadecane having perfect (100) cetane number, experimental data for pure components and mixtures with higher cetane numbers were constrained to 100.

$$CN = \frac{\sum_{lumps} v_i \beta_i CN_i}{\sum_{lumps} v_i \beta_i} \quad (3.5)$$

Table 3.5: Calculated average β_i values in this work compared to the work of Ghosh and Jaffe³⁸

lump	β_i , This Work	β_i , Ghosh and Jaffe ³⁸
nParaffins	3.9831	0.5212
Isoparaffins	0.8043	7.3717
Olefins	2.4795	0.3597
Naphthenics	-0.2231	0.0727
Aromatics	1.3113	3.1967

Cetane number calculation results in Figure 3.7a based on the optimized β_i parameters show excellent agreement with the experimental data. These experimental diesel fuels did not include significant quantities of some molecular lumps; therefore, some β_i were fixed to the average values provided by Ghosh and Jaffe. Table 3.5 contains a comparison of the average calculated β_i values in this work compared to the work done by Ghosh and Jaffe. While the model may contain some error calculating the cetane number of diesel fuels containing significant quantities of some molecular lumps not well represented in the current data, those lumps are not important for the primarily paraffinic product produced via triglyceride hydroprocessing. The *n*-paraffin and *i*-paraffin lumps are well represented in the literature experimental data and predict paraffinic diesel cetane accurately. Therefore, the model can be used for the products of triglyceride hydroprocessing with the ability for improvement once more data is available for a broader range of diesel fuels.

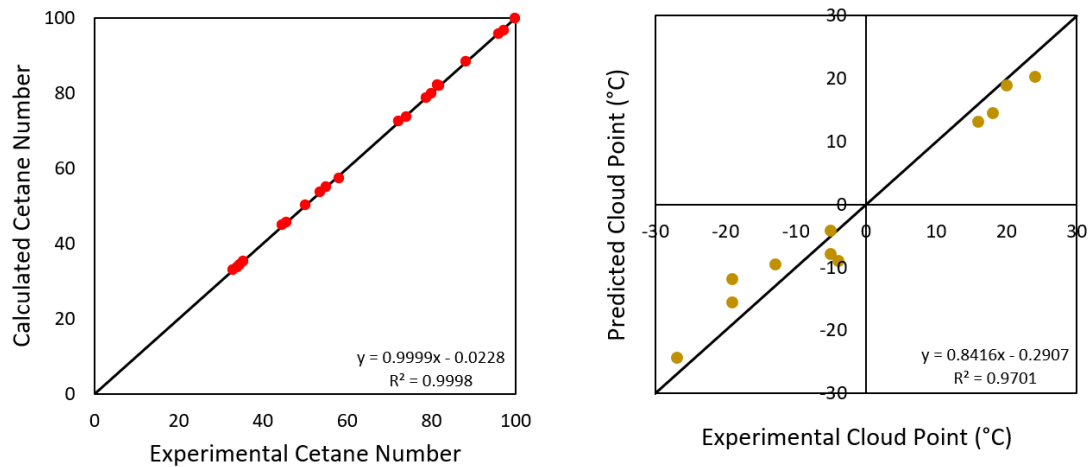


Figure 3.7: Parity plots comparing experimental and calculated values from the a) cetane number and b) cloud point models. The $y=x$ line is displayed for comparison.

3.7.2 Cloud Point Model

The cloud point model was based on solid-liquid mixture phase equilibria, as shown in Equation 3.6.⁴⁵ It accounts for paraffins that are the majority of products from triglyceride hydroprocessing. Two key assumptions were made in the model: all activity coefficients are unity and the specific heat terms can be approximated as a function of T_m and T_f . The temperature dependence of the specific heat was approximated as shown in Equation 3.7, where ε is a species-dependent correction parameter. The form of the temperature dependence was calculated from the integrated form Equation 3.6 assuming the ΔC_p is independent of temperature and using the first two terms of the Taylor expansion for $\ln x$. From the equation, the freezing temperature was calculated for each individual component in the model, with the highest freezing temperature determining the cloud point of the mixture. Work by Affens et al.⁴⁶ was used to adjust the heat of fusion values for the paraffins. ε values were optimized based on experimental data, with the non-zero results for the molecule

lumps being shown in Table 3.6. The model still performs well without the ε correction, but the correction enhances model results. The model results are in good agreement with the experimental values, as shown in Figure 3.7b.^{30,44}

$$\ln x_i \gamma_i = -\frac{\Delta_{fus}H(T_m)}{R} \left(\frac{T_m - T_f}{T_m T_f} \right) - \frac{1}{RT_f} \int_{T_m}^{T_f} \Delta C_P dT + \frac{1}{R} \int_{T_m}^{T_f} \frac{\Delta C_P}{T} dT \quad (3.6)$$

$$\ln x_i = -\frac{\Delta_{fus}H(T_m)}{R} \left(\frac{T_m - T_f}{T_m T_f} \right) - \varepsilon \left(\frac{T_m - T_f}{T_f} \right)^2 \quad (3.7)$$

Table 3.6: ε correction parameter values for the cloud point model for the molecular lumps that impacted the final result based on available experimental data

lump	ε
nC ₁₆	0.1003
nC ₁₇	0.1149
nC ₁₈	0.1861
nC ₁₉	0.1900
nC ₂₀	0.2019
iC ₁₇	0.2163
iC ₁₈	0.4200

The errors in the model can be explained due to errors in the experimental measurements and a lack of definition in the *i*-paraffin distribution. The experimental data^{30,44} did not differentiate between the set of isomers that can form during hydroprocessing. As evidenced by Figure 3.1, isomer variations can impact the cloud point significantly, especially if the end product is further isomerized in a hydroisomerization reactor. The kinetic model fully defines the single methyl-branched isomers *i*-paraffin distribution based on thermodynamic equilibrium and

isomerization rate. Additionally, there is a broader definition of the multi-branched molecules in the kinetic model compared to the experimental data. For the purposes of the cloud point calculation, the *i*-paraffin definitions were modeled as methyl-branched isomers with properties that resembled an average of all possible methyl-branched isomers for a given carbon number and branch number.

3.7.3 Predicting Diesel End-Use

Applying the cetane number and cloud point models to the product of the kinetic model can provide valuable insights into the operation of reactor. In a mostly paraffinic mixture, the ratio and types of n-paraffins and isoparaffins greatly impacts the properties. Isomerization can be controlled by the catalyst and process conditions in a hydroprocessing reactor or by further processing the effluent of the hydroprocessing reactor in a hydroisomerization reactor. Figure 3.8 shows the tradeoff between cetane number and cloud point with the degree of isomerization of a C17/C18 paraffin mixture. Increasing branching helps reduce the cloud point to more useable levels in certain climate at the cost of some cetane. The cetane number for the paraffinic mixture is usually much higher than required, so the loss can be absorbed. However, increasing the isomerization also coincides with an increased diesel yield loss to cracking and so may be undesirable. But, another application for the paraffinic mixture can be as a cetane enhancer additive to a poor cetane fuel, thereby removing the need for further processing.

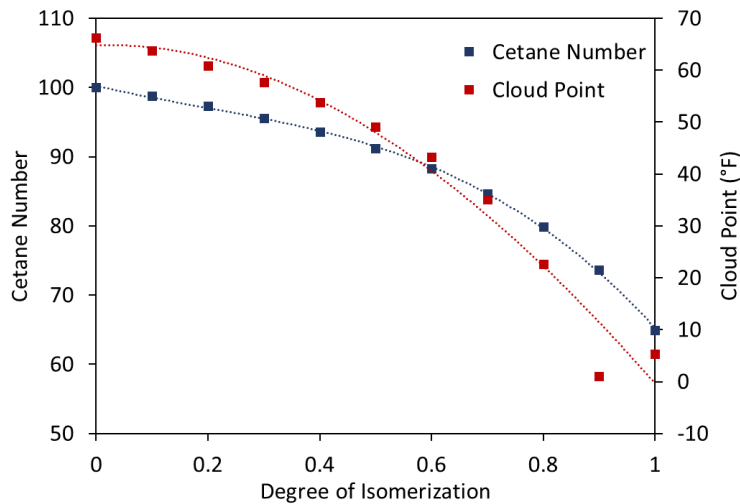


Figure 3.8: Tradeoff between cetane number and cloud point with varying degrees of isomerization. Degree of isomerization is defined as the weight of isoparaffins out of the total paraffin weight

3.8 Conclusions

A kinetic model to simulate green fuel production via triglyceride hydroprocessing predicted the experimental data for a range of process conditions well. A single set of kinetic parameters for each catalyst sufficiently captured the different temperatures, pressures, and catalyst contact times in the model. The definition and modeling of the reaction chemistry at the molecular level helped model the different oil feeds since the feed mixtures vary only in the initial fatty acid profiles. The definition of molecules is also vital in predicting the end-use properties like cetane number and cloud point that can vary significantly with each isomer in a mixture. Accurately calculating diesel properties can help reduce the burden of an experimental program for the implementation of triglyceride hydroprocessing. Using the approach of this work, modeling additional triglycerides and oil compositions should be fast with minimal adjustment of kinetic parameters.

3.9 Nomenclature

A_j = Arrhenius constant for reaction family j

α_j = reaction index factor in the Bell-Evans-Polyani LFER for reaction family j

β_i = Ghosh-Jaffe parameter 1 for cetane number³⁸

CN = cetane number

CN_i = Ghosh-Jaffe parameter 2 for cetane number³⁸

ΔC_p = change in specific heat from solid to liquid ($C_p^L - C_p^S$)

E_{0j} = activation energy factor in the Bell-Evans-Polyani LFER for reaction family j

ε = temperature dependence parameter for change in T_f

ΔH_i = enthalpy of reaction for reaction i

$\Delta_{fus}H(T_m)$ = heat of fusion for species i at the melting point of species i

$K_{ad,i}$ = adsorption constant for species i

MW_i = molecular weight of species i

ρ_i = density of species i

$P_{H_2}^{\alpha,k}$ = hydrogen partial pressure on site k with adjustable parameter α

R = universal gas constant

T = temperature

T_f = freezing point for species i in solution

T_m = melting temperature for species i

v_i = volume fraction of species i

x_i = mole fraction of species i

γ_i = activity coefficient for species i

3.10 Acknowledgement

Michael T. Klein acknowledges collaborations with and support of colleagues via the Saudi Aramco Chair Program at KFUMP and Saudi Aramco.

Chapter 4

MOLECULAR-LEVEL KINETIC MODELING OF A REAL VACUUM GAS OIL HYDROPROCESSING REFINERY SYSTEM

Pratyush Agarwal¹, Mayuresh Sahasrabudhe², Sumit Khandalkar², Chandra Saravanan², and Michael T. Klein^{1,3}

¹Department of Chemical and Biomolecular Engineering, University of Delaware, Newark, DE 19716, United States

²Reliance Industries Limited, Ghansoli, Navi Mumbai 400701, India

³Center for Refining and Petrochemicals, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia

4.1 Abstract

A molecular-level kinetic model was constructed for a vacuum gas oil hydroprocessing unit. The feedstock molecule selection was based on typical arrangements of structural attributes in crude oil. Based on the fundamental hydroprocessing chemistry, a reaction network was developed for the feedstock molecules including 5747 reactions distributed among 12 core types of reactions. The final molecular model contained 1532 unique species up to 45 carbons encompassing molecules up to five aromatic rings with heteroatoms. To determine the initial condition of the feedstock, a statistical approach was applied by using probability density functions (PDFs) characterizing the molecules in terms of their structural attributes. Experimental feed measurements were used to determine the values of the PDF parameters. A library containing 21 sets of PDF parameters representing the range of the feed measurements was established and used to determine the starting point for optimization. Simulated feed properties showed excellent agreement with experimental values. For the kinetic model, the reactor system was divided into a series of 19 pseudo-PFRs, one for each catalyst layer, interspersed with the appropriate quench streams. Each pseudo-PFR was modeled using a side-by-side reaction and vapor-liquid equilibrium approach. The activity of each type of catalyst and the deactivation due to coking and metal deposition were included in the simulation. Quantitative structure/reactivity correlations were used to greatly reduce the number of parameters in the model. The parameters were optimized using a simulated annealing algorithm so that the model results corresponded to the measured reactor effluent. The optimized model showed good agreement with the experimental measurements. To simplify the day-to-day running of the kinetic model while still

allowing developers to change and study the model in more advanced applications, a user-friendly application was developed.

4.2 Introduction

Catalytic hydroprocessing is one of the most valuable refinery processes for feedstock upgrading. It is often used in the hydrocracking of heavier feeds and in the removal of impurities like sulfur, nitrogen, and heavy metals from all feeds. Typically, molecules classified in the range of vacuum gas oil (VGO), coker gas oil, and residual oil are hydroprocessed to obtain high value liquid products from otherwise low-value feedstocks. Since a 20% growth in liquid fuels and chemicals demand is projected between 2016-2040 to support the growing population, the oil industry constantly needs to strategize investment for refinery units to meet these growing demands.⁴⁷ An important aspect of this mission is the ability to characterize the profit which is rooted in the underlying chemistry and kinetics of the processing units. Kinetic modeling has been vastly successful for this purpose.

Lumped kinetic models have historically been applied to unit operations.⁴⁸⁻⁵¹ Reactor analysis via lumped components provides plenty of information. Still, obscuring the fundamental chemistry and kinetics results in feedstock-dependent models valid in narrow process ranges. Environmental concerns like SO_x and NO_x emission reduction, political pressures, and the economic advantages of crude oil selection and routing have demonstrated the need for detailed understanding of processes independent of the feedstock. Additionally, with a growing interest in better monetizing bio-based feedstocks like algal oil and pyrolysis oil, refineries face unprecedented questions about the use of existing frameworks for new applications. While hydroprocessing is uniquely suitable to upgrade both traditional and new

feedstocks, without understanding the fundamental chemistry and kinetics of the hydroprocessing units, the best value for its current and future applications is uncertain. Clearly, kinetic models for the modern refinery need to include more detail and be more adaptable than the lumped kinetic modeling strategies of the past.

Technological advances in recent years have made molecular-level kinetic modeling possible. Better computer processor clock speeds and more efficient algorithms have allowed scientists to solve larger and more complex systems. Additionally, the advent of modern analytical techniques like FTICR-MS and 2D-GC have identified and quantified molecular components to within their respective isomer classes for feeds as heavy as vacuum residue.⁵² By characterizing the feedstock and product and studying the individual reactions and overall kinetics in a reactor, molecular-level kinetic models can be constructed. These models are more feedstock independent and adaptable to the addition different process chemistry or kinetics than the kinetic models of the past. Several research groups in academia and industry have successfully developed techniques for this type of modeling of hydroprocessing units.

Quann and Jaffe developed the structure-oriented lumping (SOL) approach defining molecules by their structural components expressed as vectors and reactions represented as vector addition operations. They applied the approach to several chemistries and processes at ExxonMobil.^{53,54} Froment and co-workers used the single-event kinetics approach for elementary reaction steps to model hydroprocessing units for complex feedstocks at the molecular level.⁵⁵ Verstraete et al. at IFP applied both the single event approach⁵⁶ and a Kinetic Monte Carlo (KMC) method with molecular reconstruction⁵⁷ to hydrocracking units. Martens et al. developed a fundamental kinetic model for the hydrocracking of hydrogenated VGO using single

event kinetics.⁵⁸ Alvarez-Majmutov et al. simulated VGO hydrocracking at the molecular level using KMC and statistical sampling for the feed reconstruction from structural attributes.⁵⁹ In the Klein research group, software tools designed to create molecular-level kinetic models based on mechanistic or pathways-level reaction networks have been applied to several hydroprocessing applications.^{7,18,40,60}

In this work, a vacuum gas oil hydroprocessing unit was modeled using the Kinetic Modeler's Toolbox (KMT) developed by the Klein research group.^{7,8,11} The suite of software tools can be loosely characterized into four main functionalities: reaction network generation, feedstock composition modeling, kinetic modeling, and user-friendly application design. This paper discusses each of these functionalities in turn as they apply to the modeling of VGO hydroprocessing. Special consideration is assigned to the model applicability over the entire catalyst life cycle with varying feeds.

4.3 Experimental Data

The experimental data is from an operating hydroprocessing unit. A simplified diagram of the reactor system is shown in Figure 4.1. Other process units for product separation and scrubbing are outside the scope of this work and are not included. The hydroprocessing unit consists of a two-reactor system. The first unit acts as a guard bed unit containing a single bed with multiple catalyst layers predominantly for HDM activity and some HDS activity. The second unit is a three-bed unit with interstage quenching containing catalyst layers with strong HDS, HDN, and HDA activity. All catalysts employed are commercially available hydroprocessing catalysts mainly composed of Ni, NiMo, CoMo, or NiCoMo on acid support. The quench streams are

primarily hydrogen gas with a small amount of C1-C5 hydrocarbons, hydrogen sulfide, and nitrogen gas.

The hydroprocessing catalysts have an active life of approximately two years after which the catalyst must be replaced. Over the lifetime of the catalyst, the temperature of the reaction system is slowly increased to offset the loss of activity. Other than the loss in activity, the reactor temperature also depends on the changes in feed, product specifications, product quality, overall flow rates, hydrogen quench quality, etc. Figure 4.2 shows the average inlet temperature over the life of the catalyst, where the difference between the start-of-life and the end-of-life of the catalyst is approximately 30 K. After a period of operation near the end of the catalyst life, the guard bed unit is slowly bypassed due to low activity.

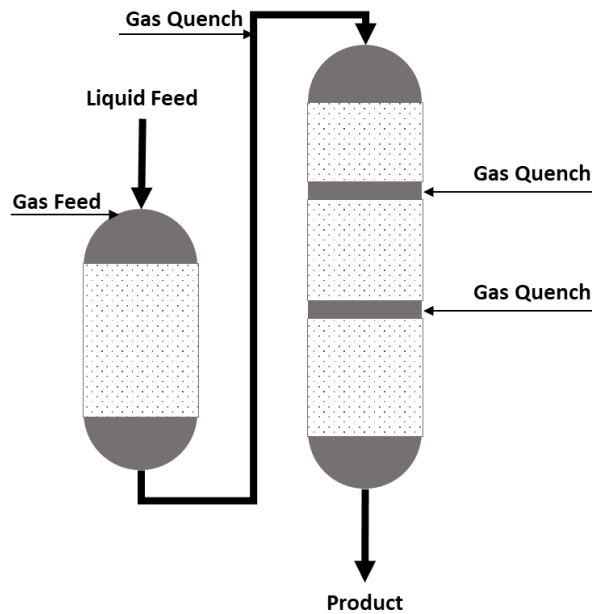


Figure 4.1: A simplified representation of the two-reactor hydroprocessing unit for vacuum gas oil

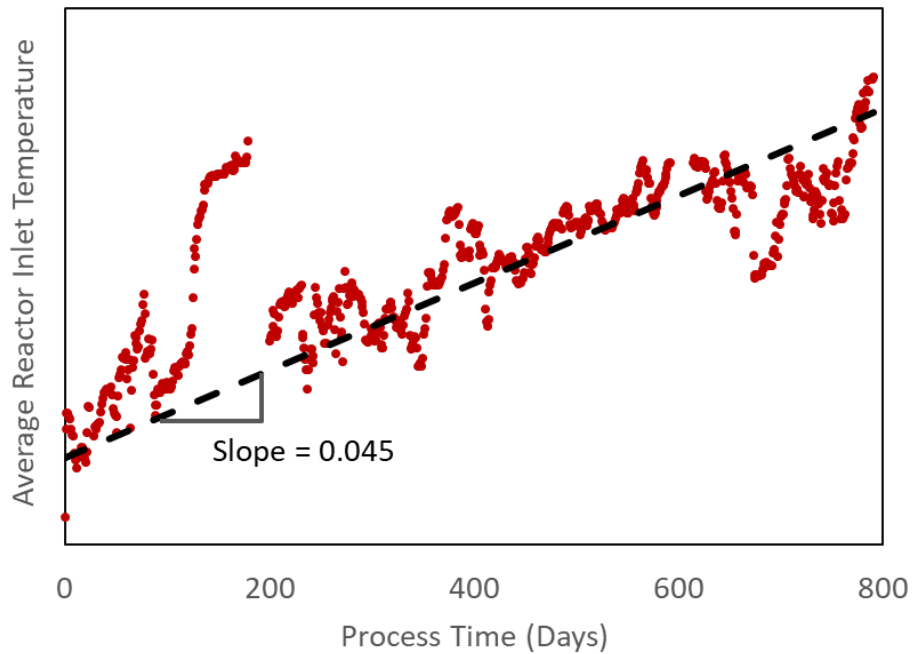


Figure 4.2: Average reactor inlet temperature over the lifetime of the catalysts. The (●) dots represent process data and the (--) line represents a fit with an average $\Delta T = 0.045$ K.

The feed to the unit is in the vacuum gas oil boiling point range of 300-540°C. It is a mixture of the vacuum gas oil cut from crude oil as well as small amounts of the high boiling fractions from the outlet units like a delayed coking unit. A representative range for the feed and process conditions for typical VGO hydroprocessing is shown in Table 4.1. For the feed specification, density, simulated distillation (modified ASTM D2887), sulfur, nitrogen, and metal contaminant data were collected. Process parameters like the temperature, pressure, flow rates, and quench gas composition were monitored based on the desired product specification. There can be large variations in both the feed specification and the process conditions over the catalyst lifetime that must be accounted for in the kinetic model for the system.

Table 4.1: Typical feed measurements and inlet process conditions for a vacuum gas oil hydroprocessing system

Parameter	Units	Value
Density	g/ml	0.90-0.95
Boiling Range	°C	300-540
Sulfur	wt%	1.5-3.0
Nitrogen	ppmw	1000-3000
Basic Nitrogen	ppmw	300-1200
Nickel	ppmw	0-2
Vanadium	ppmw	0-2
Inlet Temperature	°C	300-400
Inlet Pressure	bar	70-120

4.4 Reaction Network Generation

The overall reaction network for VGO hydroprocessing was computationally generated from the seed molecules using the Interactive Network Generator (INGen) tool.^{7,8} INGen represents each molecule computationally as a bond-electron matrix. Bond breaking and bond formation in reactions is characterized as a matrix addition operation on a reactive subgroup of the overall molecule matrix. An example of this operation is shown in Figure 4.3 for the hydrogenation of an olefin. These reaction matrices are universal for a homologous reaction series (reaction family) that acts upon a specific type of site on a molecule. Site selection is based on the desired overall chemistry, where the typical reaction families for VGO hydroprocessing are shown in Table 4.3.

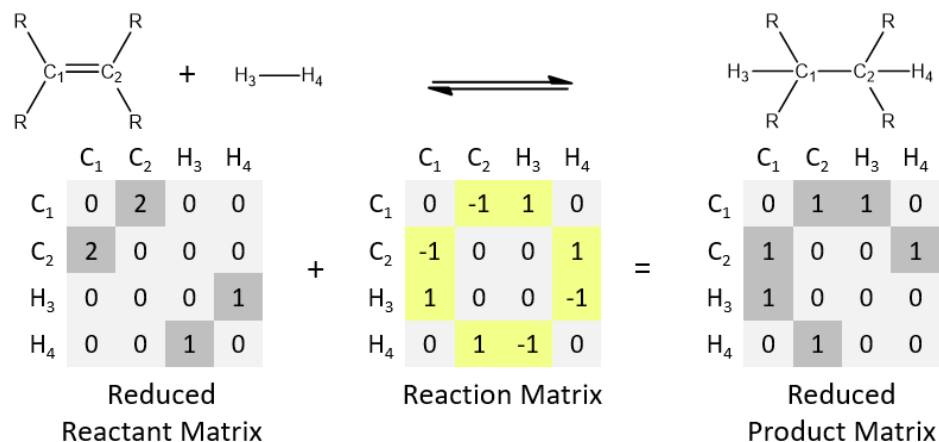


Figure 4.3: Matrix representation of a reactive subgroup and reaction matrix in INGen

4.4.1 Molecule Selection

The first step in constructing a VGO kinetic model molecular representation requires the identification of candidate seed molecules that are present in the feedstock. The reactions of these molecules generate a complete set of molecules that define the VGO hydroprocessing system. Generally, the structure of a petroleum molecule is a statistically probable arrangement of one or more structural attributes, such as a double bond, an aromatic ring, a saturated ring, a heteroatom, or a branch. Past analytical experiments have identified the most likely arrangements of these structural attributes.^{61,62} The boiling point range of the feedstock defines the carbon chain lengths of the molecules present. Table 4.2 includes a summary of the types of molecules included in the current VGO model. Each of these molecules types was extended to up to 45 carbon atoms.

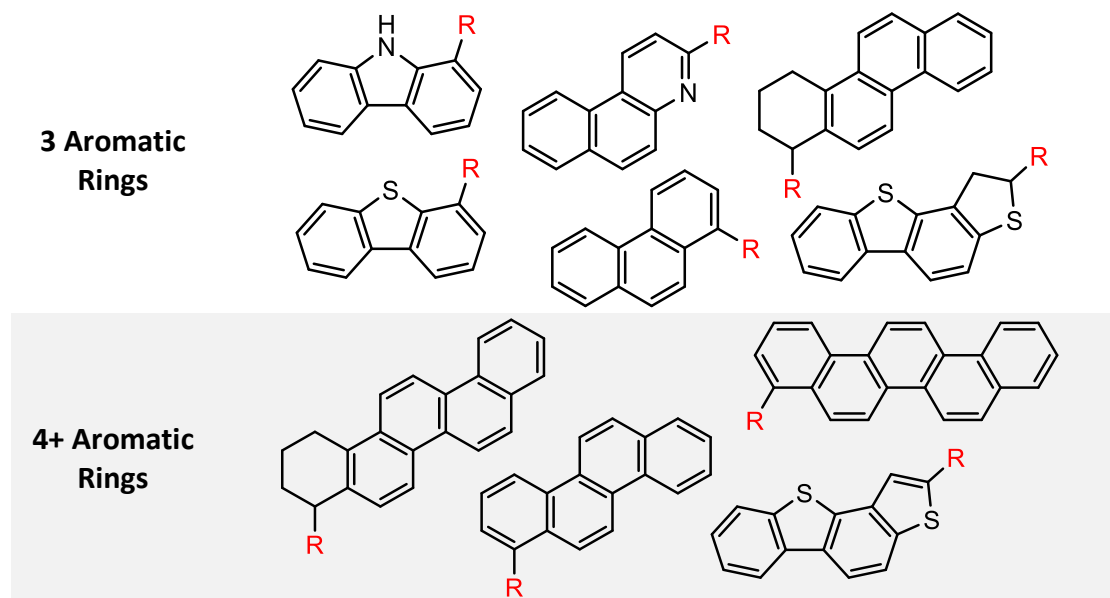
Paraffins, olefins, naphthenics, and aromatics (PONA) define the four broad classes of molecules in the system. While thousands of different isomeric arrangements are possible for the structural attributes to generate PONA molecules,

some simplifications in molecule selection were needed to limit network growth and reduce the model complexity. The paraffins were allowed to contain up to one methyl branch. Paraffinic sulfur was included as thiols and disulfides. Aliphatic amines were ignored due to their low presence in feeds.^{63,64} Oxygenates were excluded because no data was available on their concentrations and their concentrations are expected to be low in the feed.⁶² Olefins containing a single double bond were included to represent any double bond content resulting from feed mixing with a thermally cracked heavy oil feed.

The naphthenics and aromatics contained up to five rings. A staggered arrangement of rings (chrysene type) was preferred over a denser arrangement (pyrene type). Thiophene and its derivatives of up to four rings characterized the aromatic sulfur content.⁶⁵ Quinoline and benzoquinoline represented the basic nitrogen while indole and carbazole represented the neutral nitrogen.⁶⁴ Various saturated stages of the base aromatics were included as determined by the saturation reaction network. Only one side chain was allowed from rings to minimize the network size. Additionally, side chains on the molecules were not allowed to contain heteroatoms, double bonds, or branching. While different side chains are present in a real feed, the variations were captured in the paraffinic and olefinic distributions to reduce the overall network size.

Table 4.2: Molecule types present in the vacuum gas oil hydroprocessing network

Species Type	Sample Species
Paraffins	
Olefins	
Naphthenics	
1 Aromatic Ring	
2 Aromatic Rings	



4.4.2 Reaction Selection and Constraints

Girgis and Gates compiled a concise summary of the experimental evidence for most hydroprocessing reactions on model compounds.⁶⁶ While maintaining the scientific rigor of model compound studies, site selection constraints and rank limitations were implemented to limit the growth of the reaction network. Unconstrained network growth leads to large numbers of isomers and reaction pathways that do not increase the useful information of the network while greatly increasing its size and complexity. Therefore, careful pathway selection and network pruning was performed via the molecule seeding technique described by Joshi et al.¹⁷ Seeding allows the specification of set molecules in a reaction pathway, minimizing parallel reaction pathways and reducing isomer generation. A brief description of the reaction families in Table 4.3 follows.

Hydrodesulfurization (HDS) was allowed to proceed either via saturation of the sulfur-containing ring followed by the removal of sulfur or via direct removal of the sulfur from the aromatic ring. Since aliphatic double bonds should immediately

saturate in the hydrogen-rich environment, the resulting side chains during the HDS of molecules like benzothiophenes were saturated at the cost of another hydrogen molecule in the overall HDS reaction. Hydrodenitrogenation (HDN) of basic nitrogen compounds was allowed to occur only after the saturation of the nitrogen containing ring due to the relatively large carbon-nitrogen bond energy in those rings.⁶⁶ For neutral compounds like indole, saturation was still forced to be the first step due to the much faster relative rate of saturation to HDN and to avoid the formation of aliphatic double bonds.^{66,67} The formation of the intermediate amine group molecules was ignored in favor of complete HDN to ammonia.

Dealkylation was separated into two classes based on whether the dealkylation occurred from an aromatic ring or a saturated ring since dealkylation from aromatic rings is significantly faster.⁶⁸ No differentiation was made for the type of ring (5-member or 6-member) or the number of other rings fused to the side chain containing ring. Side chain cracking was allowed to occur anywhere for chains of up to 10 carbons, but methane formation was disallowed due to its high activation energy barrier. For chains longer than 10 carbons, cracking was limited to the middle of the carbon chain to limit network size.⁶⁹

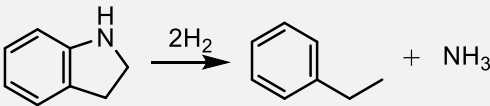
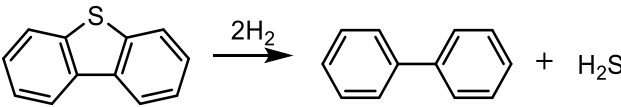
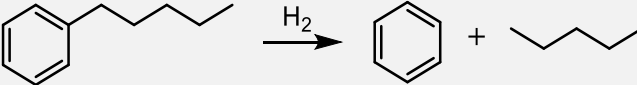
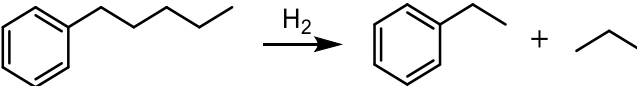
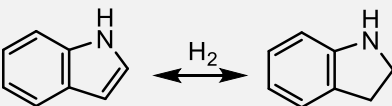
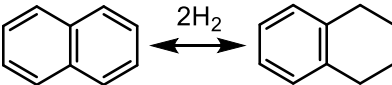
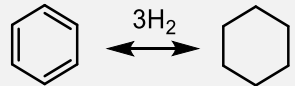
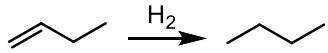
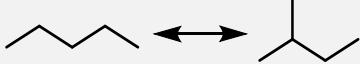
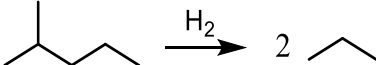
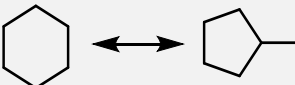
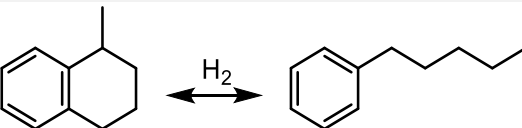
Saturation of aromatic rings was only allowed to occur in the most likely pathway. In general, this pathway involves the successive saturation of the outermost aromatic rings in an aromatic cluster. For a molecule with a side chain, the saturation of the outermost aromatic ring containing the side chain was favored to allow ring opening reactions. Although 2H saturation of the middle ring in a molecule like phenanthrene is kinetically favored, further saturation of the resulting product is hindered.^{68,70} Therefore, that pathway was ignored. Due to low operating pressures in

the current system (80 bar) for 6H saturation of isolated aromatic rings (benzene type), the HDS and HDN reactions were favored over saturation for molecules like 2,3-dihydrobenzothiophene. Additionally, saturation of biphenyl and its analogs was ignored.^{70,71}

Hydrogenation was included to model saturation of the small amounts of non-aromatic double bond content in the feed. Paraffin isomerization was limited to the formation of one methyl branch. All methyl branched isomers were included for molecules containing up to 10 carbons, and only one isomer was allowed above 10 carbons. Cracking was allowed at the branch position for the isoparaffins. For *n*-paraffins and isoparaffins with more than 10 carbons, cracking was also permitted in the middle of the carbon chain where implicit isomerization before cracking was assumed.

Ring isomerization was limited to the reversible isomerization of isolated cyclohexane rings to cyclopentane rings. Analytical information typically cannot differentiate the C5/C6 ring ratio in molecules with more than one ring, and so those isomerizations were not included. Only saturated rings attached to aromatic rings with a branch of up to four carbons were allowed to ring open since molecules with longer alkyl chains are more likely to crack. In situations where ring isomerization is the likely first step before ring opening, as in the case of tetralin, isomerization was assumed to implicitly occur before ring opening. Cyclopentanes and cyclohexanes were not allowed to ring open due to the mild conditions for hydrocracking.^{18,68,69}

Table 4.3: Reaction families in the vacuum gas oil hydroprocessing network

Reaction Family	Sample Reaction	Catalyst Site
Hydrodenitrogenation		Metal
Hydrodesulfurization		Metal
Dealkylation		Acid
Side Chain Cracking		Acid
Saturation 2H		Metal
Saturation 4H		Metal
Saturation 6H		Metal
Hydrogenation		Metal
Paraffin Isomerization		Acid
Paraffin Cracking		Acid
Ring Isomerization		Acid
Ring Opening		Acid

4.4.3 Coking Chemistry

Another important aspect of the hydroprocessing chemistry is the deactivation of the catalysts. In hydroprocessing, catalysts can deactivate via four main

mechanisms: coke formation leading to pore blockage, poisoning by strongly adsorbed species, metal deposition, and sintering.^{72,73} In this work, two types of deactivation mechanisms were considered: metal deposition and aromatic coke formation. Nickel and vanadium are typically found in porphyrins and were modeled as such.⁷⁴ The remaining elements that may be present like iron, sodium, and arsenic were considered to be free elements in the feed. Metal deposition proceeded via a direct reaction of the free metal with a catalyst site L. For Ni and V, a balance equation was written to saturate the porphyrin and free the metal, as shown in Figure 4.4. The metal in the porphyrin underwent a reaction with a catalyst site and the remaining hydrocarbon became an aromatic coke formation precursor.⁷⁵

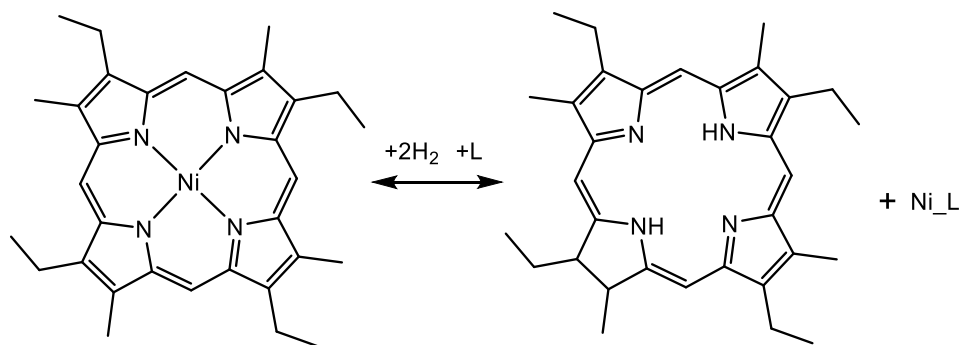


Figure 4.4: Metal deposition from a Ni porphyrin on a catalyst site L based on the mechanism by Ware et al.⁷⁵

For aromatic coke formations, two pathways were included to build large clusters of aromatic rings, as shown in Figure 4.5. Aromatic molecules without side chains were allowed to undergo coupling reactions with each other to form larger aromatic clusters.⁷² These reactions were limited to two or more identical aromatic molecules reacting together. Additionally, a pathway was added to allow alkyl chains

to undergo alkylation, ring closing, and aromatization on aromatic molecules without any branching. Molecules with up to 10 aromatic rings were allowed to form via these two mechanisms and those containing more than 7 aromatic rings were termed to be coke precursors. The formation of coke was assumed from the coupling of these aromatic coke precursors on the larger timescale of the catalyst deactivation.

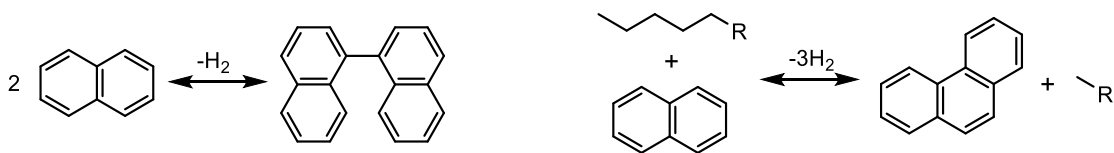


Figure 4.5: Aromatic ring building in the coking reaction pathways via a) coupling of aromatic rings with each other and b) alkylation, ring closing, and aromatization to form larger aromatic clusters from alkyl chains

4.4.4 Network Results

Table 4.4 provides a summary of the reactions in this system. Overall, the final VGO hydroprocessing network contained 1532 species and 5747 reactions. The species were distributed among the paraffin, olefin, naphthenic, and aromatic classes as shown in Table 4.2. The reactions represent the breadth of expected reactions shown in Table 4.3. An iterative approach was applied to limit network building, starting with very few limitations and adding additional constraints as necessary to reduce the number of reactions and species. While reactions do increase solution times slightly, the main determining factor for a computationally relevant model is the number of species. This is because the number of species directly corresponds to the number of material balances that need to be concurrently solved during numerical integration in the kinetic model. Therefore, special care was afforded to limiting the new species formation during network growth.

Table 4.4: Summary of molecules and reactions in the vacuum gas oil hydroprocessing model

Species Type	Number	Reaction	Number
<i>n</i> -Paraffins	45	Hydrodesulfurization	404
<i>i</i> -Paraffins	51	Hydrodenitrogenation	142
Sulfides	90	Hydrogenation	44
Olefins	44	Dealkylation	1292
Naphthenics	243	Side Chain Cracking	2590
1 Aromatic Ring	410	Saturation 2H	108
2 Aromatic Rings	316	Saturation 4H	328
3 Aromatic Rings	194	Saturation 6H	221
4 Aromatic Rings	87	Paraffin Isomerization	51
5+ Aromatic Rings	42	Paraffin Cracking	94
H ₂ , H ₂ S, NH ₃ , N ₂	4	Ring Isomerization	40
Metals and Free Elements	6	Ring Opening	35
		Coking 1	392
		Coking 2	6
All Species	1532	All Reactions	5747

4.5 Feed Composition Generation

Once a viable reaction network has been generated and the molecules in the system have been identified, the molecule fractions in the feedstock must be defined. While it is realistic to envision getting individual molecule fractions directly from analytical techniques, these techniques are still relatively expensive, time intensive, and sometimes subject to large uncertainties. For the current model, discrete molecule analytical information was not available. Therefore, the feedstock problem was reduced to a statistical one. For each structural attribute set in the system, a probability density function (PDF) can denote the probability of occurrence of each attribute in the system. A juxtaposition of the attributes into the respective molecules in the system then provides the mole fractions. Trauth et al. have demonstrated the viability of such an approach for heavy feedstocks.⁷⁶ The remaining question is simply one of

organizing and defining the PDFs. For this, the Initial Condition Generator (ICG) tool was used to define the PDFs and optimize the distributions.

4.5.1 Feedstock PDF Definitions

ICG is a tool that allows the user to define PDFs based on the properties and structural attributes of each individual molecule. In ICG, a property generator calculates the properties of each molecule based on Gani⁷⁷ and Benson⁷⁸ group contribution methods, and the ChemGraph^{7,8} routine provides a rigorous accounting of all the structural attributes of a molecule. The definition of the PDFs is then based on discriminating molecules using the properties and structural attributes. Histograms (Equation 4.1) and two-parameters gamma PDFs (Equation 4.2) were chosen due to their versatility in representing different functional forms. The most efficient way to rigorously account for the different structural attributes involved defining a PDF tree containing a set of parent-child relationships. Each new child generation allowed for the further discernment of the structural attributes of a molecule via the definition of additional constraints. Sibling distributions signify a juxtaposition of the respective PDF members so that molecules in each bin of a given PDF have the PDF distributions of sibling PDFs. Mole fractions are calculated as a product of the attribute weights of a molecule in each PDF that satisfies the constraints of a given molecule.

$$f(x) = H(x) \quad (4.1)$$

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}}{e^{\beta x} \Gamma(\alpha)} \quad (4.2)$$

Figure 4.6 shows a visual representation of the PDF tree structure used in this model. The continuous rings represent gamma PDFs while the sectioned rings

represent histograms. An overall PDF divided the molecules into n-paraffin, i-paraffin, olefin, naphthenic, and aromatic fractions. From there, further refinement was added as necessary. The n-paraffins were split into sulfur containing and non-sulfur containing molecules. Aromatics and naphthenics were split by ring number. For the aromatic ring distribution, further refinement was made based on the number of naphthenic rings also present with the aromatic rings. Parallely, the aromatic and naphthenic ring distribution was discriminated based on the presence of sulfur and nitrogen atoms. A gamma PDF discretized by the carbon number was used as a final child generation in each branch of the PDF tree to impose a carbon number distribution on the molecules. Anything not explicitly classified in the PDF tree was assumed to be equimolarly distributed.

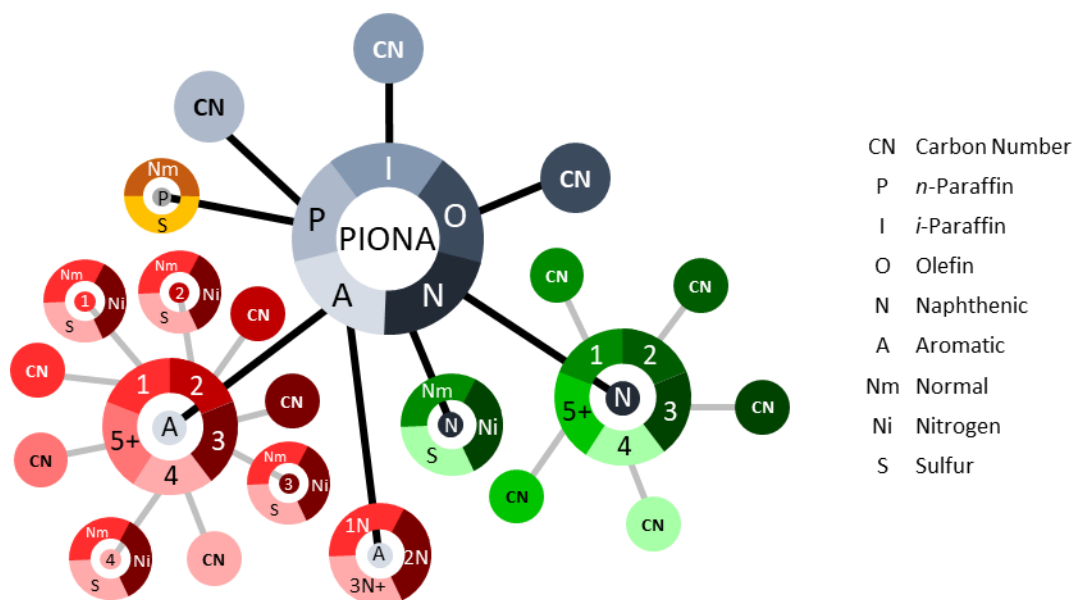


Figure 4.6: PDF tree representation of the PDFs on the VGO hydroprocessing model.

4.5.2 Parameter Optimization

In the PDF tree, each histogram has $(\# \text{ bins} - 1)$ parameters and each gamma PDF has two parameters. An objective function of the form shown in Equation 4.3 was employed to minimize the difference between the observed and predicted properties of the feedstock by perturbing the PDF parameters. For the chi-square statistic, the weighting factor is the standard deviation of the measurement. A simulated annealing algorithm was used to vary the PDF parameters and minimize Equation 4.3. However, the parameters that define the PDF tree in Figure 4.6 greatly exceed the individual data measurements, leading to an underdefined system. A number of heuristics were applied to the parameters to constrain the outputs based on detailed experiments on a representative feed. Some additional constraints were gleaned from the product profile and expected relative rates of reaction for different types of molecules.⁶⁶

$$obj = \sum_q \left(\frac{Exp_q - Pred_q}{Weight_q} \right)^2 \quad (4.3)$$

A small subset of the datasets embodying a wide range of the feedstock measurements were then optimized using the heuristics. This created a small library of representative feedstocks with excellent agreement between experimental and predicted properties with a good underlying chemistry basis. For a new dataset, Equation 4.4 defined an error value based on the datasets in the library D . The PDF parameters of the dataset producing the lowest error were selected as the starting point of optimization. An upper threshold was established for the objective function value calculated using Equation 4.3 to maintain model integrity. If the objective function value calculated for the new dataset was higher than the upper threshold, an error

warned the user that the feedstock was not satisfactorily modeled and may require some manual input.

$$\text{Given dataset } n, \forall d \in D, n_0 \ni \min \left(\sum_q \left(\frac{Exp_{q,d} - Exp_{q,n}}{Weight_q} \right)^2 \right) \quad (4.4)$$

4.5.3 Feed Composition Results

A library of 21 datasets was compiled to characterize the VGO feedstock slate to the hydroprocessing unit. The datasets in the library cover the range of the density, nitrogen, sulfur, and simulated distillation boiling cuts observed and represent 20% of the overall datasets. It is anticipated that the number of datasets in the library should not need to increase as additional datasets are introduced unless the new datasets represent significantly different feedstocks. This is not expected for the current refinery setup, and data from other hydroprocessing unit catalyst cycles confirms this assertion. As an example of the library datasets, Figure 4.7 presents the overall carbon number distribution and class distributions of a representative dataset. Further classification of the aromatic and naphthenic ring-wise distributions is shown in Figure 4.8. The variations of these distributions provide the feedstocks for the range of VGO properties.

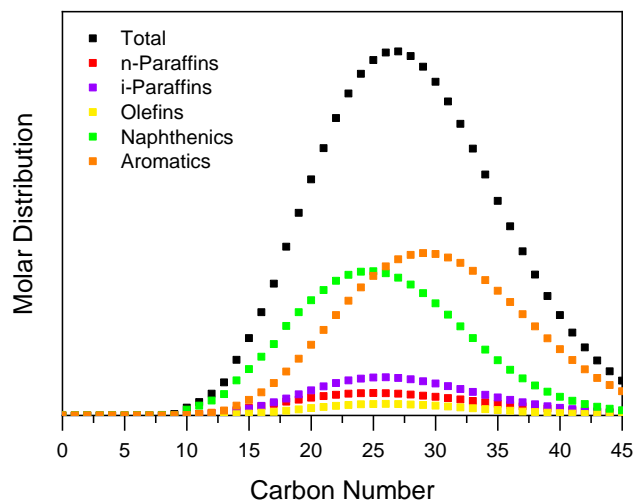


Figure 4.7: Overall carbon number distribution and class distribution in the feedstock model for a representative dataset

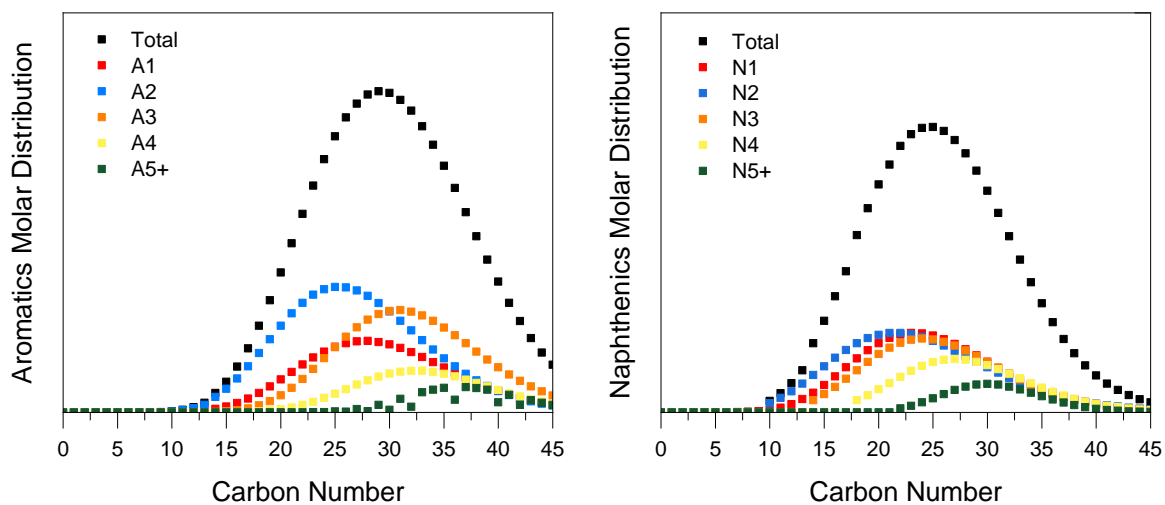


Figure 4.8: Ring number distributions for aromatics and naphthenics in the feedstock model for a representative dataset

From the library, a set of 95 datasets corresponding to weekly measurements over the two-year period were optimized. Figures 4.9-4.11 show the results of the optimization for the simulated distillation boiling cuts, the sulfur and nitrogen weight fractions, and the density. The simulated distillation boiling cuts show good agreement with the experimental feedstock measurements to within 5°C for all cutpoints and 2°C for the 50% cutpoint. The 50% cutpoint was afforded the most weight during tuning. The 5% and 95% cutpoints provided the most error in prediction but also suffer from poor analytical accuracy. Sulfur and nitrogen are also well predicted for most datasets to within 0.1 wt% and 100 ppmw, respectively, of the experimental value. Density results agree well with the experimental data to within 0.005 g/mL, but a few measurements proved to be slightly challenging to predict due to the uncertainty associated with the prediction correlation. The intermolecular interactions that impact the density measurement require an unfeasible number of parameters to incorporate into the correlation, so those effects were neglected. Additionally, it should be noted that a few datasets contained values deemed to be outliers associated with experimental or typographical errors. These datasets were included in the figures for thoroughness.

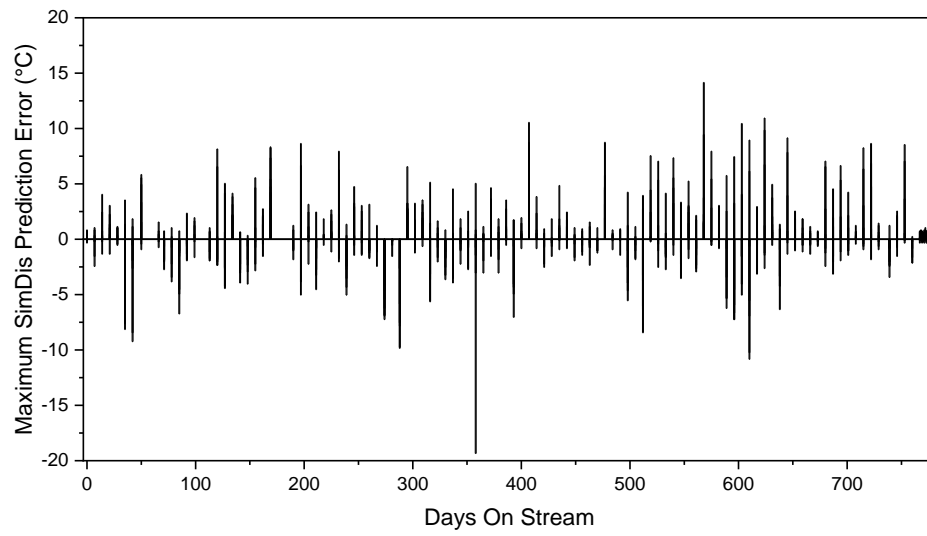


Figure 4.9: Maximum prediction errors in the simulated distillation 5%, 10%, 30%, 50%, 70%, 90% and 95% boiling cuts for the feedstock model over the entire process range with a datapoint every seven days on stream

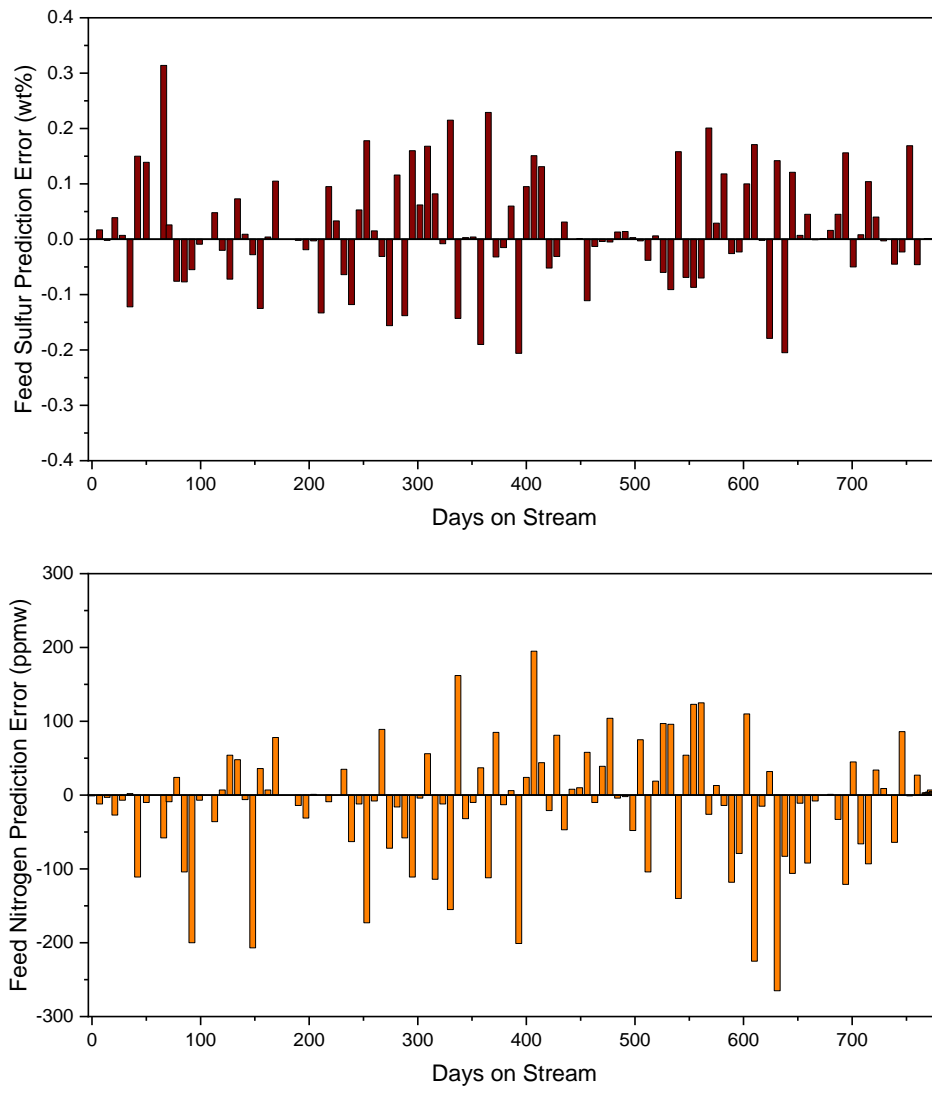


Figure 4.10: Error in the prediction of the a) sulfur and b) nitrogen elemental analysis for the feedstock model over the entire process range with a datapoint every seven days on stream

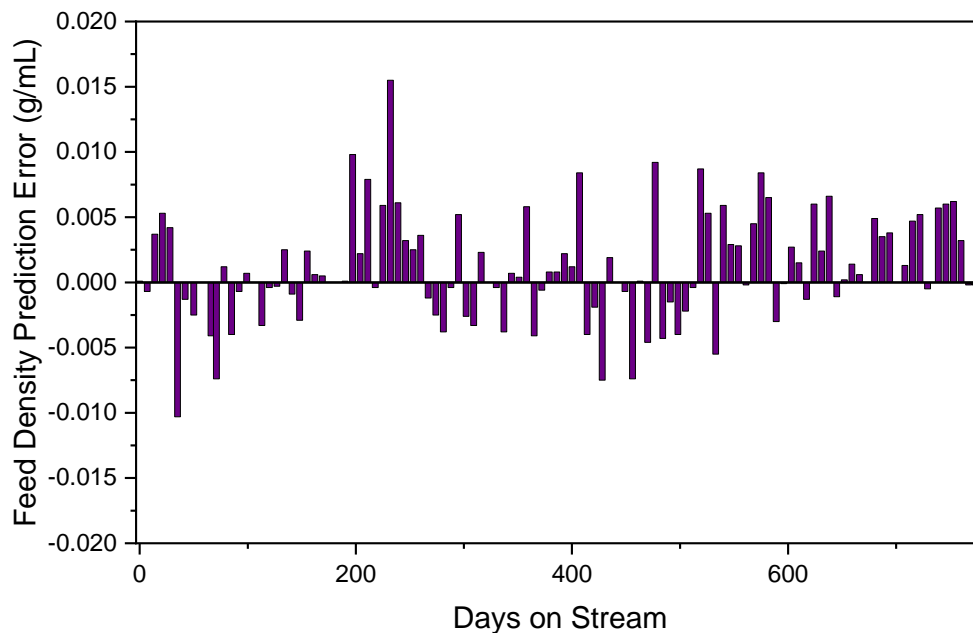


Figure 4.11: Error in the prediction of the feedstock density values over the entire process range with a datapoint every seven days on stream

4.6 Kinetic Model Generation

An in-house software, the Dynamic Model Builder¹¹ (DMB) was used to create the kinetic model. DMB is a C++ executable that can parse a reaction network, generate the model equations based on a specified rate law type, and numerically integrate the system of ordinary differential equations using the CVODES solver⁷⁹ from Lawrence Livermore National Laboratory. The application can dynamically adapt to changing reactions, species, and model equations via user input, and it works independently on Microsoft Windows-based system without additional software requirements. For the VGO hydroprocessing system, functionality was added to the software to simulate two-phase flow, reactors in series, different catalyst reactivities,

and side-stream additions based on user specified options. The current system was simulated as a side-by-side reaction and vapor-liquid equilibrium fixed-bed hydroprocessing system containing a series of plug flow reactors (PFRs) with multiple catalysts and interstage quenching.

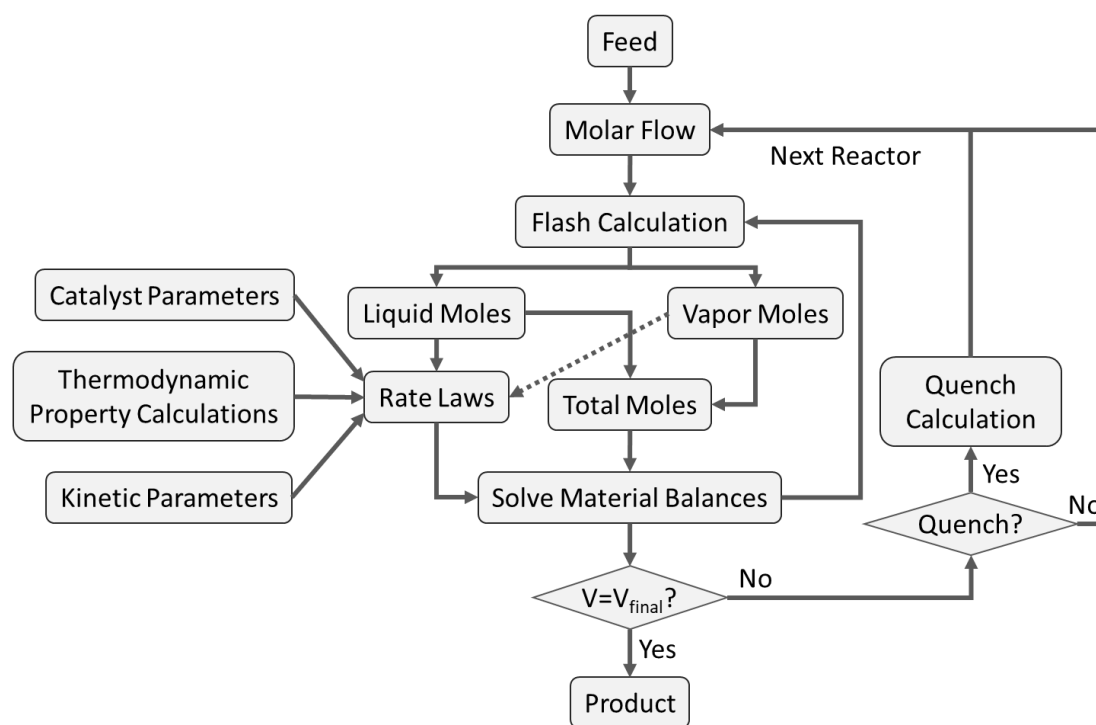


Figure 4.12: Flowchart displaying the logic of kinetic model evaluation in DMB

Figure 4.12 displays the logic of the kinetic model evaluation in DMB. Each catalyst layer that forms a reactor bed in the system was considered to be a pseudo-PFR. The reactor system was then simulated as a series of pseudo-PFRs. For each pseudo-PFR, a flash calculation provided an accounting of the liquid and vapor fraction of the feed. The liquid and vapor molar concentrations along with the thermodynamic, kinetic, and catalyst parameters defined the set of rate-law-based

material balances in the system. Numerically integrating these material balances provided the pseudo- PFR output. It should be noted that the flash calculation and subsequent calculation of the kinetic and thermodynamic parameters was performed at each step of the numerical integration to accurately represent their evolution in the catalyst layers. If the pseudo- PFR represented the end of the reactor bed and there was a quench stream addition after the bed, a mass and energy balance calculated the new feed condition for the next pseudo- PFR. The methodology was repeated for each defined pseudo- PFR until the outlet of the entire reactor system was reached.

4.6.1 Model Equations and Kinetics

From the reaction network, DMB determines an overall list of species and equations. One rate law-based material balance equation (Equation 4.5) was needed for every species in the system along with an overall energy balance (Equation 4.6) and a simplified pressure balance (Equation 4.7). A PFR balance equation was used for the material balances through the fixed bed. The energy balance in the system was tracked as the overall change in the energy of both the liquid and vapor states assuming that they are in constant equilibrium. For the pressure balance, the lack of a viable viscosity correlation and some catalyst data made use of the Ergun equation unfeasible for the current system. Therefore, an empirical relationship was assumed based on the temperature, density, superficial velocity, and coking along with six adjustable parameters.

$$\frac{dF_s}{dL} = -r_s * A_c \quad (4.5)$$

$$\frac{dT}{dL} = \frac{A_c \sum_i^{rxns} r_i * \Delta H_i}{\sum_S^{species} (c_{P,S,L} F_{S,L} + c_{P,S,V} F_{S,V})} \quad (4.6)$$

$$\frac{dP}{dL} = -(a_1(T - a_2) + a_3\rho)v_s \left(\frac{a_4}{1 - L_k y_{cat,k}} - a_5 \right) - a_6 \quad (4.7)$$

To model the aforementioned equations, thermodynamic properties and states were needed for each molecule in the feed. The gas phase properties were calculated from Benson⁷⁸ and Gani⁷⁷ group contribution methods for each molecule. Liquid phase thermodynamic properties were calculated using corresponding state functions from the vapor phase values using the methods explained by Poling et al.⁸⁰ Since components in a mixture may not be close to their critical even though the pure components may be, the properties were linearly extrapolated beyond $T_r = T/T_c = 0.8$ to avoid the asymptote in the corresponding state equations at the critical point. The reacting system was assumed to be in constant vapor-liquid equilibrium, and the vapor fraction was calculated using the Rachford-Rice equation.⁸¹ Partition coefficients were estimated from Raoult's Law with the saturated vapor pressures from the method of Lee and Kesler.⁸² For hydrogen, Henry's law was used to calculate the dissolved concentration in the liquid phase from Equation 4.8.⁸³

$$C_{aq}(H_2) = 0.00078 \exp \left[-500 \left(\frac{1}{T} - \frac{1}{298} \right) \right] P_{H_2} \quad (4.8)$$

Once the thermodynamic properties and vapor-liquid fractions are known, the rate laws for each reaction can be tabulated. The catalytic liquid phase reactions were considered to be of the Langmuir-Hinshelwood-Hougen-Watson (LHHW) form assuming surface rate control.^{18,84} Equation 4.9 shows the form of the LHHW rate law for reaction i on site k for reactor x . Each rate equation requires a surface rate reaction constant, an equilibrium constant, and adsorption constants for all species. Deactivation (L_k) and catalyst dependent ($y_{cat,k}$) parameters modeled the differences in reactivity over the lifetime of the catalyst. The reactions in Table 4.3 were

designated as either metal site or acid site reactions for each reaction family to differentiate between the different types of catalysts sites k . Additionally, an explicit partial pressure dependence $P_{H_2}^\mu$ was added to model the impact of the hydrogen gas partial pressure.

$$r_{i,k,x} = \frac{k_{sr} * \prod_r^{reactants} K_{ad,r} * L_k * y_{cat,k} \left(\prod_r^{reactants} C_r - \frac{\prod_p^{products} C_p}{K_{eq,i}} \right)}{P_{H_2}^\mu \left(1 + \sum_s^{species} K_{ad,s} C_s \right)^n} \quad (4.9)$$

For the surface rate constants, the Arrhenius equation was used with the Bell-Evans-Polanyi linear free-energy relationship (LFER) for the activation energy, as shown in Equation 4.10.^{19,20} Overall, nine different types of catalysts were utilized in the reactor system. Since the design of each catalyst is optimized for certain types of reactions, independent surface rate constants would be needed for each catalyst. However, the LFER concept can be extended to catalyst families, minimizing the number of additional parameters. While the reaction activation energy is often the major difference between two different catalysts, Horton et al. demonstrated that the differences in reactivity between different catalysts can be modeled by a constant departure term on the $\ln A$ factor for narrow temperature ranges.²¹ Extending that concept to all reaction families j , a base parameter $\ln A_{j,c_{ref}}$ was determined, and the reactivity differences were modeled as a $\Delta \ln A_{j,c_{ref} \rightarrow c}$ term for each catalyst c , as shown in Equation 4.11. The final equation for the surface rate constant is given in Equation 4.12 for reaction i on catalyst c that belongs to reaction family f .

$$\ln k_{sr,i,f} = \ln A_f - \frac{E_{0f} + \alpha_f \Delta H_i}{RT} \quad (4.10)$$

$$\ln A_{f,c} = \ln A_{f,c_{ref}} + \Delta \ln A_{f,c_{ref} \rightarrow c} \quad (4.11)$$

$$\ln k_{sr,i,f,c} = \ln A_{f,c} + \Delta \ln A_{f,c_{ref} \rightarrow c} - \frac{E_{0f} + \alpha_f \Delta H_i}{RT} \quad (4.12)$$

The equilibrium constants for each reaction were calculated using the standard thermodynamic formulation shown in Equation 4.13. For the adsorption constants, a quantitative structure/property relationship given in Equation 4.14 was applied based on the structural attributes of the molecules. These correlations have been shown to work well for hydroprocessing systems by Korre et al.²³ One set of parameters is needed for each type of site in the system. Since the number of reactions i (10^3) and species s (10^3) greatly outnumber the number of reaction families f (10^1) and sites k (10^0), using Equations 4.12, 4.13, and 4.14, the overall number of parameters in the model can be greatly reduced from $\mathcal{O}(10^4)$ to $\mathcal{O}(10^2)$. This parameter reduction is essential for a feasible optimization problem based on the experimental data.

$$\ln K_{eq,i} = -\frac{\Delta G_i}{RT} \quad (4.13)$$

$$\ln K_{ad,k} = b_{1,k} + \frac{b_{2,k}N_{AR} + b_{3,k}N_{NR} + b_{4,k}N_{SC} + b_{5,k}N_S + b_{6,k}N_N}{RT} \quad (4.14)$$

Finally, the catalyst layers undergo deactivation over the two-year period where the catalyst is operational. Deactivation for a once-thru simulation representing the reactor state at any given day was assumed to be a constant. Over time, a deactivation parameter given in Equation 4.15 was calculated and adjusted based on the aggregate metal deposition and coke formation from all previous days where the system was run. One parameter was calculated for each overall reactor as defined by Figure 4.1. Since coke structure and composition are difficult to determine, coke was tracked as an elemental mixture of carbon, hydrogen, nitrogen, and sulfur. Each coke precursor contributed to the coke formation by a lumped parameter c_1 that accounted

for both the rate of coke formation and the subsequent effect of that formation on the deactivation parameter. It was assumed that the coke formation would not be negative for any given day, and the reversibility and impact of the hydrogen partial pressure is captured by the coke precursor formation. Metal deposition was tracked directly from the kinetic model.

$$L = 2 - \exp(c_1 wt_{cokeP} + c_2 wt_{metal} + c_3 D) \quad (4.15)$$

4.6.2 Kinetic Model Evaluation

The kinetic model was constructed in DMB from the developed reaction network for VGO hydroprocessing. The average simulation time for the complete system of 19 reactor in series on a i7-4770 (@3.40 GHz) processor with 16GB RAM was 490 seconds. The requisite kinetic parameters were optimized using a simulated annealing algorithm. An objective function of the form in Equation 4.16 calculated the error for each property q for each dataset d . Due to the large number of parameters and long simulation time, two datasets were selected for the optimization from the start, middle, and end of the catalyst cycle for the optimization. First, manual parameter adjustments were used to narrow the parameter bounds to regimes of interest. The catalyst LFER parameters were fixed during this stage based on qualitative catalyst information. Then, adjustments to the other parameter values were performed using the simulated annealing algorithm.

$$obj = \sum_d \sum_q \left(\frac{Exp_{d,q} - Pred_{d,q}}{Weight_{d,q}} \right)^2 \quad (4.16)$$

Table 4.5 provides a summary of the kinetic LFER and catalyst LFER parameters. The reaction families were defined based on the specific reaction types believed to have different rates since independent molecules can have significantly

different reactivity. For example, rather than desulfurization being one reaction family, desulfurization was separated into distinct classes like dibenzothiophene, benzothiophene, and sulfide that tend to have very different reaction rates. A base $\log_{10}A$ value and E_0 value was optimized for each of the 39 reaction families in the model. The α_f values were all constrained to 0.1 to sufficiently capture the individual reaction deviations in a reaction family. Some parameters were forced to have certain relationships to others. For example, without more detailed experimental data, the overall cracking activity can be modeled by either paraffin cracking, side chain cracking, aromatic dealkylation, or saturate dealkylation. In this case, the cracking reactions were given similar rates that were slightly slower than the dealkylation rates. Aromatic dealkylation was given a faster rate than saturate dealkylation due to the stability of the underlying benzylic carbenium ion formed in aromatic dealkylation.

For the catalyst LFER parameters, manual increments were made based on the qualitative catalyst definitions and the deviations from the experimental reactor temperature profile. It should be noted that many of the catalyst LFER parameters are degenerate. The $\Delta \log_{10} A_f$ values were classified by the four main types of reactions: demetallization, desulfurization, denitrogenation, and hydrocracking. For a given catalyst, qualitative information available provided the relative activity of those four reaction types. Within those broad classifications, most of the $\Delta \log_{10} A_f$ parameters for a given catalyst type have the same value. This greatly reduced the burden of parameter optimization. Nevertheless, once experimental data is available for individual catalysts, better quantification can be captured in the catalyst LFER parameters. As the parameters are defined relative to one another, one controlled experiment per catalyst type should provide sufficient information.

Table 4.5: LFER and catalyst LFER parameters for the VGO hydroprocessing model

Reaction Family	$\log_{10}A$	E_0 (kJ/mol)	$\Delta \log_{10} A_{f,ref \rightarrow c}$ Values for Catalyst C									
			1	2	3	4	5	6	7	8	9	10
Cyclization-Coke	1.14	89.3	-2	-2	-2	0	0	-2	0	0	0	0
Alkylation-Coke	2.18	84.7	-1	-1	-1	0	0	-1	0	0	0	0
Aromatization-Coke	4.70	59.6	-1	-1	-1	0	0	-1	0	0	0	0
DeNitrogenation-BQN4H	6.48	43.9	-1.8	-1.8	-1.8	0	1.1	-1.8	0.1	1.2	1.6	1.1
DeNitrogenation-CBZ6H	6.19	51.0	-1.8	-1.8	-1.8	0	1.1	-1.8	0.1	1.2	1.6	1.1
DeNitrogenation-QN4H	6.13	35.2	-1.8	-1.8	-1.8	0	1.1	-1.8	0.1	1.2	1.6	1.1
DeNitrogenation-Indole2H	3.70	32.5	-1.8	-1.8	-1.8	0	1.1	-1.8	0.1	1.2	1.6	1.1
HDM	2.43	51.5	0.5	0.5	0.8	1.5	1.5	1.5	0	0	0	0
HDM-Fe	2.43	51.5	0	0	1.7	1.2	1.2	0	0	0	0	0
Aromatic Dealkylation	-4.27	34.0	-1.6	-1.6	-1.6	0	0	-1.6	0	0	0	0
Desulfurization-BT	5.30	50.0	-1.8	-1.8	-1.8	0.3	0.6	-1.8	0.8	0.3	0.5	0.6
Desulfurization-DBT	6.22	56.1	-2.8	-2.8	-2.8	0.3	0.7	-2.8	1	0.3	0.5	0.6
Desulfurization-NBT	7.34	57.4	-1.8	-1.8	-1.8	0.3	0.9	-1.8	1.2	0.3	0.6	0.8
Desulfurization-DBT6H	6.61	47.7	-2.8	-2.8	-2.8	0.3	0.7	-2.8	0.9	0.3	0.5	0.6
Desulfurization-BT2H	6.71	38.6	-1.8	-1.8	-1.8	0.3	0.7	-1.8	1.1	0.3	0.5	0.6
Desulfurization-TDBT	6.89	56.8	-1.4	-1.4	-1.4	0.3	0.8	-1	1.1	0.4	0.5	0.6
Desulfurization-T4H	5.78	26.5	-0.9	-0.9	-0.9	0.3	0.7	-1.2	0.7	0.4	0.6	0.5
Desulfurization-PhenylBT2H	5.79	38.2	-1.8	-1.8	-1.8	0.3	0.7	-1.8	1	0.3	0.6	0.5
Desulfurization-Thiol	4.05	18.0	-0.4	-0.4	-0.4	0.2	0.9	-0.8	1	0.4	0.5	0.4
Desulfurization-Disulfide	4.64	19.1	-0.4	-0.4	-0.4	0.2	0.9	-0.8	1	0.4	0.5	0.4
Mid-chain Cracking	0.54	31.8	-1	-1	-1	0	0	-1	0	0	0	0
Mid-chain Cracking-Paraffin	1.28	28.8	-1	-1	-1	0	0	-1	0	0	0	0
Saturate Dealkylation	-4.06	36.5	-1.8	-1.8	-1.8	0	0	-1.8	0	0	0	0
Sidechain Cracking	-4.46	31.8	-1	-1	-1	0	0	-1	0	0	0	0
HSaturation2H-BT	5.11	48.9	-1.4	-1.4	-1.4	0	0.3	-1.4	0.5	1	1.2	0.7
HSaturation2H-NBT	0.08	26.5	-1.4	-1.4	-1.4	0	0.3	-1.4	0	1	1.2	0
HSaturation2H-Nitrogen	3.10	49.9	-1.4	-1.4	-1.4	0	0.4	-1.4	0.5	1	1.2	0.9
HSaturation4H-Thiophene	3.97	29.8	0.1	0.1	0.1	0.5	0.8	0.1	0.8	1	1.2	0.7
Hydrogenation	-5.54	19.3	0	0	0	0	0	0	0	0	0	0
Paraffin Isomerization	-3.33	21.8	0	0	0	0	0	0	0	0	0	0
Paraffin Cracking	-4.89	26.8	-0.5	-0.5	-0.5	0	0	-0.5	0	0	0	0
Ring Isomerization	-5.05	24.9	-2	-2	-2	0	0	-2	0	0	0	0
Ring Opening	-3.48	43.2	-2	-2	-2	0	0	-2	0	0	0	0
Saturation4H-Other	5.27	60.0	-2	-2	-2	-1	1	-2	0.4	1	1.2	0.7
Saturation4H-Nitrogen	4.36	56.5	-2	-2	-2	-1	1	-2	0.4	1	1.2	0.7
Saturation6H-Aromatic	-2.84	50.7	-2.3	-2.3	-2.3	-1.3	-0.3	-2.3	0.2	1	1.2	0.3
Saturation6H-DBT	1.68	56.7	-2.3	-2.3	-2.3	-1.3	-0.3	-2.3	0.2	1	1.2	0.3

Saturation6H-Other	-1.95	61.1	-2.3	-2.3	-2.3	-1.3	-0.3	-2.3	0.2	0.4	0.6	0.3
Saturation6H-Nitrogen	11.18	47.4	-2	-2	-2	-1	0	-2	0.4	1	1.2	0.7

Several of the other kinetic parameters were also manually optimized to fit the experimental data. Table 4.6 provides a summary of the adsorption parameters. These parameters were directly from the work of Korre et al. with minor modifications.²³ The catalyst deactivation c_1 , c_2 , and c_3 parameters for Equation 4.15 were 4.2E-7, 5.9E-4, and 2.1E-5, respectively. The pressure parameters a_1 , a_2 , a_3 , a_4 , a_5 , and a_6 were 1.67E-4, 6.0E2, 3.33E-4, 1.0E2, 9.71E1, and 1.0E-3. Since the catalyst and pressure parameters are empirically based, they do not have strong theoretical basis. However, they provide valuable quantifications of changes in the process that impact the kinetics.

Table 4.6: Adsorption parameters for the VGO hydroprocessing model for the acid and metal sites on the catalyst

Site	b_1	b_2	b_3	b_4	b_5	b_6
Acid	0.182	1.934	0.187	0.102	0.500	0.800
Metal	1.324	0.887	0.123	0.102	0.500	0.800

Figures 4.13, 4.14, and 4.15 show the prediction errors of the simulated distillation, sulfur, nitrogen, and density over the catalyst life cycle. Product data were only available in small ranges of the start, middle, and end of the catalyst cycle. The error results show that the model performs well over the entire catalyst cycle in predicting the effluent. The errors do not show an obvious trend with the days on catalyst stream. Therefore, the model is not missing a systematic trend but rather the small variabilities in day-to-day operation. Variations often stem from the standard

deviation of the measurements and small variabilities in process conditions not recorded in the data. Additionally, real-time variations in process parameters considered to be constant due to missing data can cause significant changes in the model results.

Errors in the feed composition model can also propagate to the kinetic model results. Modeling the initial conditions better can improve the results of the kinetic model. On the other hand, overfitting the feed model may actually lead to errors in the predictions, especially in the case of high analytical uncertainty. For example, the 5% and 95% distillation cutpoints generally tend to have poor accuracy and precision. Fitting those cutpoints well could lead to errors in predicting the reactor output if the measured cutpoints were inaccurate. Therefore, a careful balance informed by the real measurement errors is needed to improve the model. Beyond the analytical measurements, the underlying structure/property correlations used to calculate the bulk properties in the model could be improved to better represent the kinetic model results.

It should be mentioned that the data represent measurements of a real refinery unit at run time that were not collected with the motivation of building detailed kinetic models. Inconsistencies in measured values and predicted values often stem from irregularities in the data that propagate through the simulation run. Therefore, the data are not always mutually consistent. The results presented show a good foundation for a detailed kinetic model that captures most of the fundamental chemistry and kinetics. Errors in predictions can be explained due to the assumptions made during model building and data evaluation, as mentioned above. However, the model results provide good guidelines for the type of data needed to improve the model results in the future.

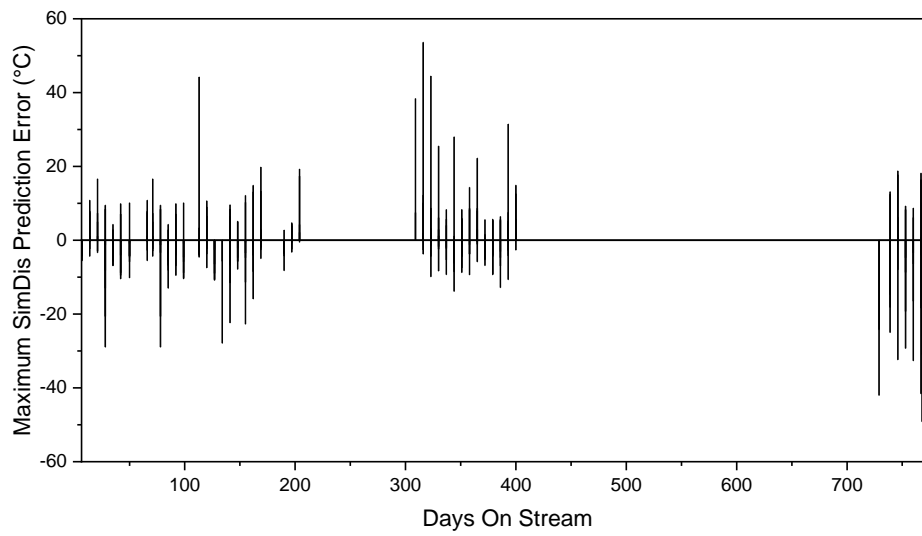


Figure 4.13: Maximum prediction errors in the simulated distillation 5%, 10%, 30%, 50%, 70%, 90% and 95% boiling cuts for the reactor effluent over the entire process range where experimental data was available

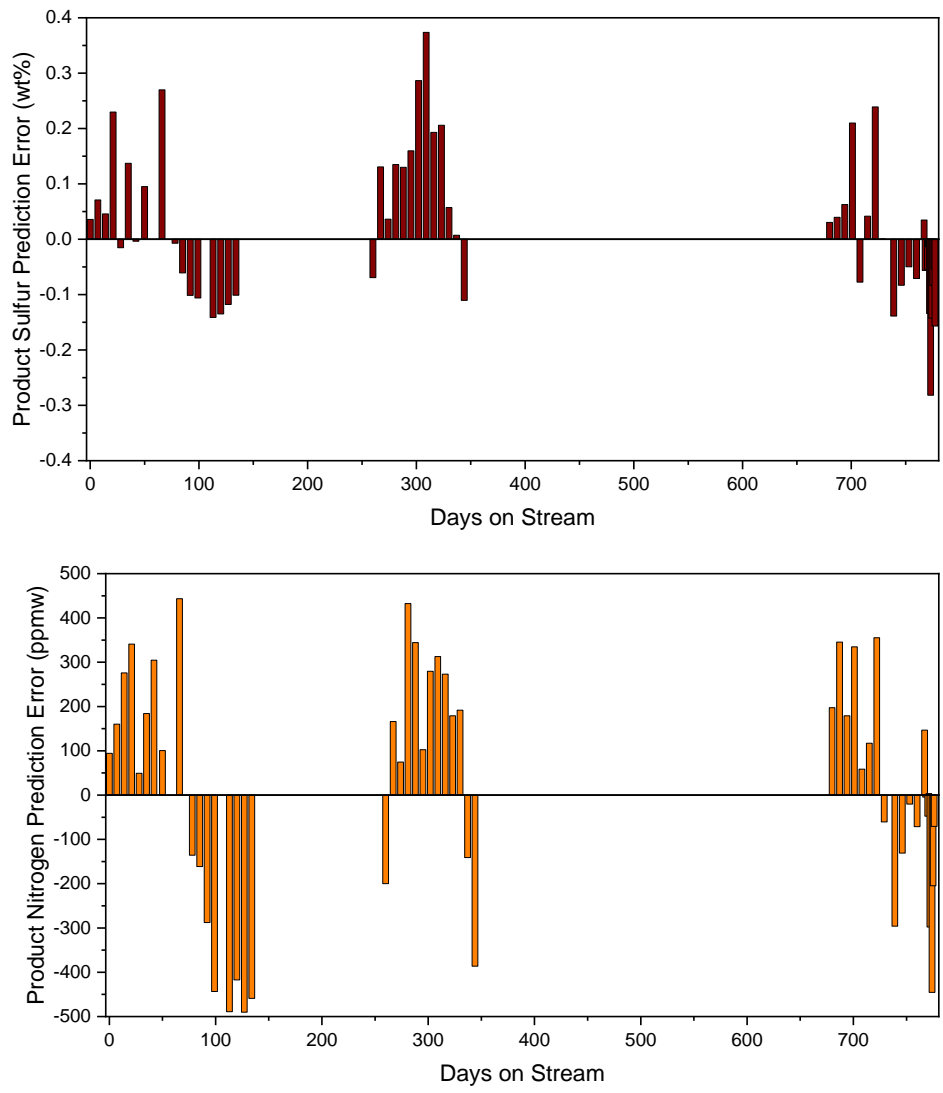


Figure 4.14: Error in the prediction of the a) sulfur and b) nitrogen elemental analysis for the reactor effluent over the entire process range where the data was available

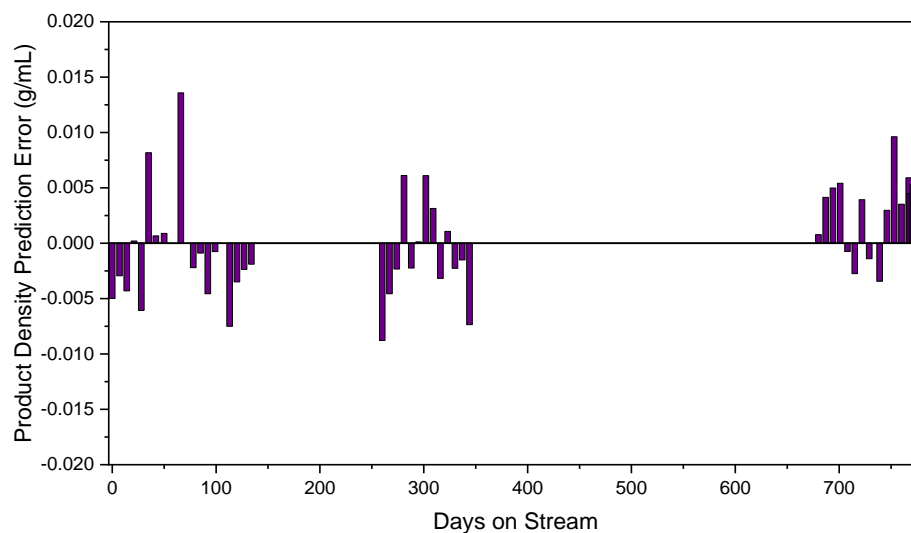


Figure 4.15: Error in the prediction of the product density values over the entire process range where data was available

4.7 User Interface

A final consideration while building kinetic models is the audience using the models. Kinetic model developers and users often have distinct modeling requirements: developers usually care about the modeling techniques and fundamentals while users usually care about the model results. As such, running the kinetic model should be as user-friendly as possible while allowing developers to modify, develop, and explore the model. Billa et al. discussed the idea of tool design for this purpose in detail.¹⁵ In this work, users had access to the entire KMT^{7,8,11} suite for model development purposes. A lighter and easier-to-use user interface was designed in the C# programming language to plainly convey model results for a given set of model inputs without the intricate details. The design was intended to require minimal user input and knowledge. Figure 4.16 shows an image of the user interface

designed for this model. The specific values and catalyst layer definitions have been obscured for proprietary reasons.

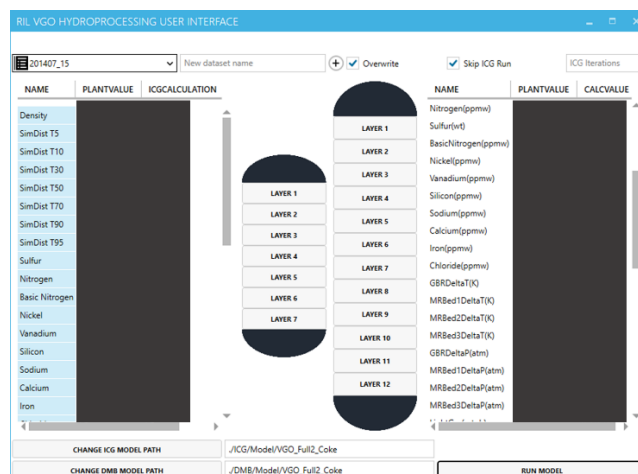


Figure 4.16: The user interface of the user-friendly application designed to run the VGO hydroprocessing model

The interface allows user to change datasets, change values in datasets, add new dataset, choose the models for both ICG and DMB, and run the model. The reactor setup and catalysts layers are displayed in the middle; however, the exact number and catalyst identities have been obscured for proprietary reasons. A list of the parsed and calculated outputs is presented after the model run, where the user has the ability to check the output after every catalyst layer in the system via the catalyst layer buttons in the middle. The main input and output data are stored in a Microsoft Excel comma-separated values (‘.csv’) format. The user can change the ‘.csv’ files when adding multiple datasets or looking for trends in the output. For developers and more advanced users, since the model folders are linked, modifications and updates to the

model in the respective INGen, ICG, or DMB softwares will automatically be reflected in the user interface.

4.8 Conclusions

A molecular-level kinetic model of a real refinery unit was successfully demonstrated using the KMT software suite. The model can account for the feed variability and the changing catalyst condition to apply a singular set of parameters to model VGO hydroprocessing. VGO feed variations can be accurately modeled by a set of 1532 species with typical crude oil molecular arrangements. A library of datasets that informs new feed conditions assists in the feed modeling by reducing the PDF parameter optimization burden since similar answers are known. From the representative feed, a reaction network containing 5747 reactions can well characterize the reactivity of the typical hydroprocessing reactions on the catalysts. Building a kinetic model from the reaction network can be successfully performed using modeling tools that allow for the setup of each individual catalyst as its own layer. It was found that modeling each catalyst layer as an independent pseudo-reactor is necessary to capture the different reactivities of the individual catalysts. The final kinetic model prediction provides an accurate representation of the product with variations in feed condition, process conditions, and catalyst condition. As more datapoints and more detailed data become available, the model can be adapted to reflect the new information. The benefit of modeling at the molecular level is that any property can be calculated if an appropriate structure/property correlation is available. Detailed data from more advanced analytical techniques like GCxGC and FTICR-MS can therefore be easily incorporated to inform the model.⁵² In the future, as these analytical techniques become more common, molecular-level modeling will need to

become the standard modeling practice as lumped models can no longer sufficiently represent the analytical information.

4.9 Nomenclature

α : gamma PDF shape parameter

α_f : LFER reaction parameter for reaction family f

β : gamma PDF rate parameter

μ : hydrogen partial pressure dependence parameter

ρ : density, g/cm³

a_1, a_2, a_3 : pressure dependence parameters

$b_{1,k}, b_{2,k}, b_{3,k}, b_{4,k}, b_{5,k}, b_{6,k}$: adsorption parameters for site type k

A_c : cross-sectional area for the reactor, dm²

$\ln A_f$: Arrhenius pre-exponential factor for reaction family f

c : catalyst number

c_1, c_2, c_3 : catalyst deactivation parameters

$c_{P,s,V}$ and $c_{P,s,L}$: vapor (V) and liquid (L) specific heat for species s , kcal/mol/K

C_s : concentration of species s , mol/L

D : days on stream, day

E_{0f} : activation energy LFER parameter for reaction family f , kcal/mol

f : reaction family number

$F_s, F_{s,V}, F_{s,L}$: total, vapor (V), and liquid (L) molar flow rates for species s , mol/s

ΔG_i : Gibbs free energy of reaction for reaction i , kcal/mol

ΔH_i : heat of reaction for reaction i , kcal/mol

$H(x)$: histogram bin probability for bin x

i : reaction number

j : site type number

k_{sr} : surface rate reaction parameter, $M^{-(r-1)} \cdot \text{atm}^\mu / \text{s}$

$K_{ad,k}$: adsorption equilibrium parameter for site type k

$K_{eq,i}$: reaction equilibrium for reaction i

L : reactor length

L_k : deactivation parameter for site k

N_{Ar} : number of aromatic rings

N_{NR} : number of aromatic rings

N_{SC} : number of non-ring saturated carbons

N_S : number of sulfur atoms

N_N : number of nitrogen atoms

P : pressure, atm

P_{H_2} : hydrogen partial pressure, atm

q : number of properties

r_i : rate of reaction for reaction i , mol/L/s

R : universal gas constant, kcal/mol/K

s : species number

T : temperature, K

v_s : superficial velocity, dm/s

wt_{cokeP}, wt_{metal} : weight of coke precursor and deposited metal up to current catalyst cycle day, kg

x : reactor number

$y_{cat,k}$: catalyst site concentration parameter for site type k , mol/L

4.10 Acknowledgement

This work was fully funded by Reliance Industries Limited. Michael T. Klein acknowledges collaborations with and support of colleagues at Reliance Industries Limited.

Chapter 5

GENERATING DATA-DRIVEN MODELS FROM MOLECULAR-LEVEL KINETIC MODELS: A KINETIC MODEL SPEEDUP STRATEGY

Pratyush Agarwal¹, and Michael T. Klein^{1,2}

*¹Department of Chemical and Biomolecular Engineering, University of Delaware,
Newark, DE 19716*

*²Center for Refining and Petrochemicals, King Fahd University of Petroleum and
Minerals, Dhahran, Saudi Arabia*

5.1 Abstract

Strategies to reduce the computer time to access the information in molecular-level kinetic models (MLKMs) were evaluated. A triglyceride hydroprocessing MLKM was used to generate datasets for small ranges of input parameters simulating three output parameters. The datasets were used to generate multilinear regression, decision tree regression, gradient boosting regression, and artificial neural network data-driven model (DDM) representations of the MLKM. All the DDMs were able to predict results very quickly ($\ll 1$ second). The predictive accuracy for the DDMs was compared, with the gradient boosting regression and artificial neural network models providing the best models over the entire range of the input parameters selected. However, in narrow input parameter ranges, multiple multilinear models and decision tree models also provide good accuracy with the added benefit of easily understood parameters and faster solution times. Additionally, multilinear regression models had much lower data requirements than the decision tree regression and artificial neural network models. The major downside to all the DDMs was shown to be the great loss in accuracy once the input parameters exceed the range of the input parameters in the datasets used to optimize the DDMs. This suggests that the predictive accuracy of the DDMs is very low, and as such, new data should be generated from the MLKM every time predictions are required outside the range of the underlying DDM data.

5.2 Introduction

Developing kinetic models is now a critical resource for the optimized operation of process units. These models can assist operators in determining the best process conditions for optimal yields. However, these optimizations often require real-time modifications to the process conditions. To inform the change in conditions, a

kinetic model should be able to produce results very quickly (<1 second) on a standard computer for the underlying optimization algorithm. While lumped model kinetics have proven to satisfy this function due to the simplified kinetics and equations, the state-of-the-art in kinetic modeling uses molecules instead of lumps. Molecular-level kinetic models (MLKMs) simulate the changes in the individual molecules through a reactor system to provide better chemical accounting and answer new questions about the effluent molecular composition. These models are better suited to predict the effect of varying feedstocks, process conditions, and deactivation because the kinetic parameters have a fundamental basis and are not feedstock dependent. The challenge of the molecular model, however, is a marked increase in the solution time for a single run.

Various strategies have been proposed to reduce the solution time to access the information in the MLKM. A first strategy would be to lump the chemistry or kinetics to reduce the MLKM, but these models can be difficult to automatically create since they require a lot of user knowledge and understanding. Additionally, creating a fundamental model with fewer features and lower accuracy may not be useful given the availability of the MLKM. Another approach is to improve computer hardware to increase the basic speed of computations. This approach can be costly and is very limited in the amount of solution time reduction possible. Processing speed for a single processor and future modification are predicted to reach saturation and have minimal effects in the near future.⁸⁵ Therefore, the approximate order of magnitude reduction in solution time is not possible via hardware resources alone. The new era in computer technology involves the use of parallelization to solve the models faster. While parallelization can effectively reduce solution times, the cost of the computing

resources along with their availability in an information-protected environment can be major disincentive. Moreover, parallelization of code requires advanced computer science information and a radical revision of the underlying MLKM algorithms that can be difficult to implement. A final strategy is to use the MLKM to generate data that can be used to create a data-driven parameterized model.

Data-driven models (DDMs) rely on the availability of large amounts of data to optimize parameters of simplified relationships. The easiest relationship is a linear regression where a straight line can be used to model the effect of one variable on a desired output. However, not all underlying relationships are sufficiently linear for a linear regression to provide accurate results. Non-linear terms are often required to model real variables but determining the higher order terms can prove to be challenging and require *a priori* information about the form of the underlying relationship. One technique of capturing the higher order relationships is to use machine learning algorithms. Machine learning models are heavily data-driven representations of the relationships between the input parameters and the output parameters using different algorithms. They have been used to solve some of the most complex problems, but often come at the cost of a large data requirement. While even a linear regression is technically a simple machine learning model, more advanced techniques like decision trees, gradient boosting regression, and artificial neural networks can provide better results for inherently non-linear models.^{86,87}

Some approaches have been considered to reduce kinetic models in literature. Many methods^{88,89} are motivated from computational fluid dynamic (CFD) model applications, where the original model is too complex for CFD models but not sufficiently large to require a speedup in the current context. Other methods are

designed to remove extraneous detail from networks^{90,91} but may not necessarily produce the desired reduction in solution time if most of the information in the kinetic model is important. Approaches like those of Nigam et al.⁹² and Fake et al.⁹³ successfully lumped pyrolysis models of a complex feedstock to represent $\sim 10^5$ radicals with $\sim 10^1$ lumps, but their approaches required detailed pyrolysis chemistry and kinetics knowledge that is difficult to generalize for all chemical systems. As a general note, lumped models remove information captured in MLKMs that is considered to be of minor importance, where it may be difficult to determine when the lumping assumptions have caused significant deviations in results from the MLKM. Furthermore, the property calculation methods^{38,94} that rely on the molecular information for accuracy suffer greatly from the lack of information in lumped models. Directly employing the MLKM output avoids these issues.

In this work, DDMs were created for MLKMs to decrease the solution time for a single once-thru simulation. Data were generated from the MLKMs that could be used to optimize the DDMs. Two parallel strategies were tested: (1) generating a linear DDM in a very narrow range and (2) generating machine learning models over larger ranges. With each strategy, as soon as the model is out of range of the data or fails, the DDM can be discarded, and a new model can be constructed with new data generated from the MLKM.

5.3 Molecular-Level Kinetic Model

A small MLKM containing 476 ordinary differential equations (ODEs) that simulates a single hydroprocessing reactor with a triglyceride feed was utilized in this study for illustration purposes. The details of the MLKM have been discussed in previous work.⁹⁴ Primarily, the product from the process is a paraffin mixture that can

be directly used in diesel engines as a substitute for diesel from traditional crude oil sources. The value of the product mixture is determined by the cetane number, the cloud point, and the yield of the diesel-range paraffinic component. The cetane number quantifies the ignition quality of the diesel fuel between 0-100, with 100 representing perfect cetane. The cloud point represents the temperature at which solid crystals start appearing in a liquid mixture. When the product has nearly perfect cetane, it also has a high cloud point that cause solidification issues in certain climates. Therefore, increasing the processing to isomerize the paraffins can improve diesel quality at the expense of some yield loss to lighter products due to the associated increase in cracking. In the MLKM, each of these properties is calculated for a simulation run based on feed specifications and process conditions. For an operating unit, predictions of the effluent for different process conditions can be used to make on-line adjustments to maximize the profit from the unit.

5.4 Model Setup

Generating data is the first step in optimizing a DDM. Ideally, experimental data would provide the necessary information to generate DDMs. However, generating the data needed to capture wide ranges of process conditions and feed specifications is generally expensive and time consuming. The MLKM, optimized on a range of process conditions, has excellent predictive capability within the range of the experimental data and good predictive capability for process conditions outside the experimental range. Results from model simulations thus provide a good method for data generation for a DDM. For a dataset in the triglyceride hydroprocessing system, the cetane number, the cloud point, and the yield of diesel product from the original triglyceride feed was recorded for each set of input parameters.

Data generation from the MLKM focused on the typical parameters process operators can manipulate to improve the process effluent and yield. For a hydroprocessing system, temperature, hydrogen pressure, and flow rate are the most common options that an operator can change. Additionally, for the triglyceride hydroprocessing system, the ratio of the fatty acid profile can vary regionally and seasonally. From these guidelines, a uniform distribution was used to randomly perturb the temperature between 350 and 440 °C, the hydrogen pressure between 4.5 and 12.0 MPa, and the feed between 0.1 to 2 times the overall mass of each individual triglyceride molecule in the soybean oil feed described in previous work.⁹⁴ The MLKM simulated the reactor effluent for each new set of process conditions and feed specifications. In the current case, the MLKM solution time was 1.039 seconds on average over the entire range of input parameters on a computer with an i7-4770 (@3.4 GHz) processor and 16 GB of RAM.

20000 datasets were generated from the MLKM. For the DDMs, the algorithms in the scikit-learn package in python were employed.⁸⁷ One fourth of the datasets were randomly selected to be part of a testing set with the remainder being part of the training set. The training set was used to optimize the DDMs while the testing set was used to test the predictions of the model. For each machine learning algorithm, some variables are available to adjust the algorithm. These variables were optimized by running the model over a range of the variables and selecting the set with the best predictive accuracy. A coefficient of determination, as defined in Equation 5.1, measured the predictive accuracy of the model outputs y_i .⁸⁷ The model improves as the R^2 value approaches 1.0, with a value of 0.0 signifying a constant mean model for the output and a negative value signifying that the model performs arbitrarily

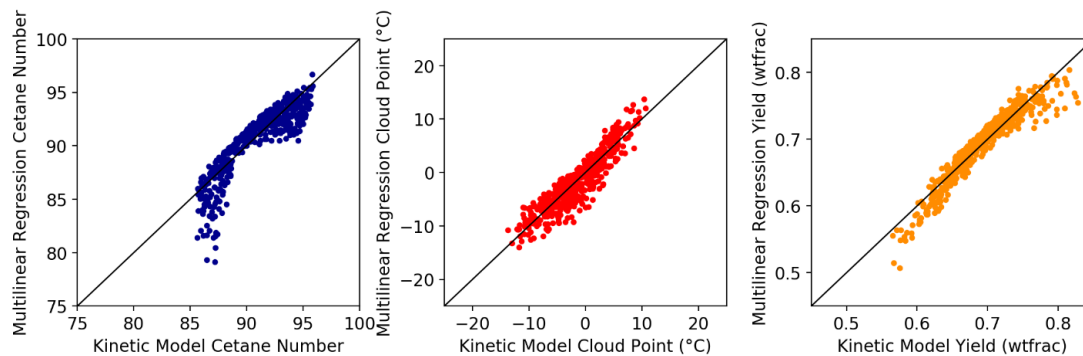
worse than a constant mean model. For the three output properties calculated, the R^2 value was averaged to provide a representative value.

$$R^2 = 1 - \frac{\sum (y_{i,actual} - y_{i,pred})^2}{\sum (y_{i,actual} - y_{mean})^2} \quad (5.1)$$

5.5 Multilinear Regression

One of the simplest approaches to generating a model from data is a multilinear regression. In a multilinear regression, dependent variables are modeled with linear relationships to one or more independent variables in a system. The linear relationship for each dependent variable y contingent on i independent variables x_i scaled by corresponding parameters α_i is shown in Equation 5.2. The parameters for the line of best-fit can be calculated using a least-squares regression that minimizes the sum of squares of the deviation of each data point to the line. Since multilinear regression has at most $i + 1$ parameters, these models are simple to understand, fast to solve, and easy to change. Additionally, the models are easily portable and do not require large software packages or tools to optimize or use. The major downside to this type of regression is that not all variable relationships are inherently linear.⁸⁶ Figure 5.1 shows the results of the multilinear regression over the entire range of temperatures, hydrogen pressures, and feed flows. There are clearly regions of large deviations that are not well captured by the linear relationships. Since it can be difficult to estimate whether a product property will fall in the range that is well modeled as opposed to one of the tails where there are large deviations, using the multilinear model over the entire range of conditions is risky.

$$y = \alpha_c + \sum_0^i \alpha_i x_i \quad (5.2)$$

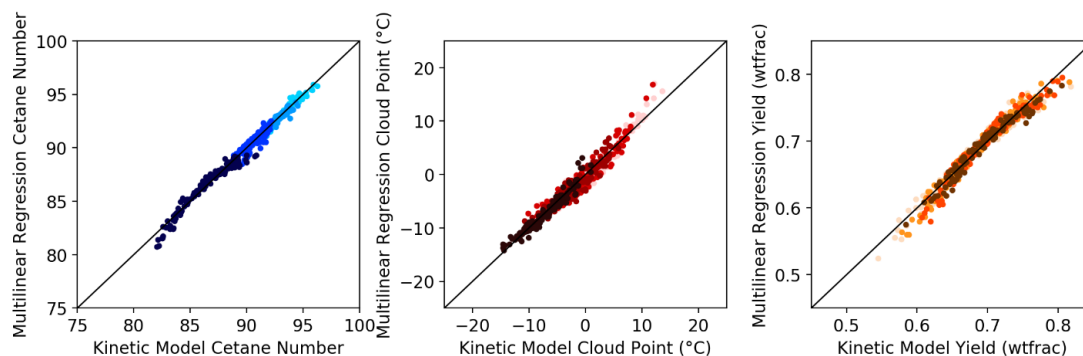


Training time (s): 0.0270
 Prediction time (s): 0.0010
 Average coefficient of determination, R^2 : 0.8533

Figure 5.1: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a multilinear regression model for the testing data

An approach to improve the multilinear model is to partition the data by one or more independent variables and generate separate models for each partition. Most relationships can be modeled well with a linear relationship in sufficiently small ranges. As an example, the data were partitioned into four equal sections representing $\sim 22^\circ\text{C}$ increments in the reactor temperature. Figure 5.2 shows the results of the four partition models, with the four temperature increments displayed from light to dark of the same color for a given product property. Each multilinear model performs very well in its respective range. Since most process usually operate in narrow working ranges, creating multilinear models with small independent variable ranges can sufficiently represent the results of the MLKM in a DDM. Once the operating regime changes, a new model can be optimized and used based on the output of the MLKM. The individual model parameter optimizations are very fast, so, optimizing a new

model should not be a problem even in a real-time optimization environment as long as the data have been previously generated.

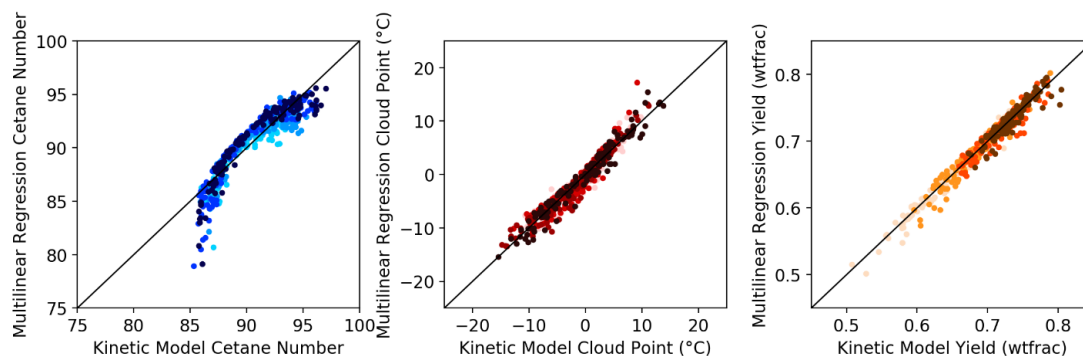


Training time (s): 0.0170
Prediction time (s): 0.0134
Average coefficient of determination, R^2 : 0.9337

Figure 5.2: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with four multilinear regression model partitioned by inlet temperature for the testing data

However, this approach does require knowledge and input from the user to define the parameters with the largest sensitivity for the desired product properties and the sufficiently small ranges needed. Figure 5.3 shows the results of repeating the example with four equal hydrogen pressure partitions of ~ 1.8 MPa rather than the temperature partitions. In this case, the group of models only perform slightly better than the single model over the entire range. Utilizing a hydrogen partition is a poor choice for the current system, but it may be the right decision for a different system or set of product properties. Therefore, the model developer must explore other independent parameter ranges or further narrow the working ranges of the models. This can greatly increase the overall time for developing and using the DDM. So, it

may be beneficial to search for an altogether different modeling approach in some cases.



Training time (s): 0.0140

Prediction time (s): 0.0010

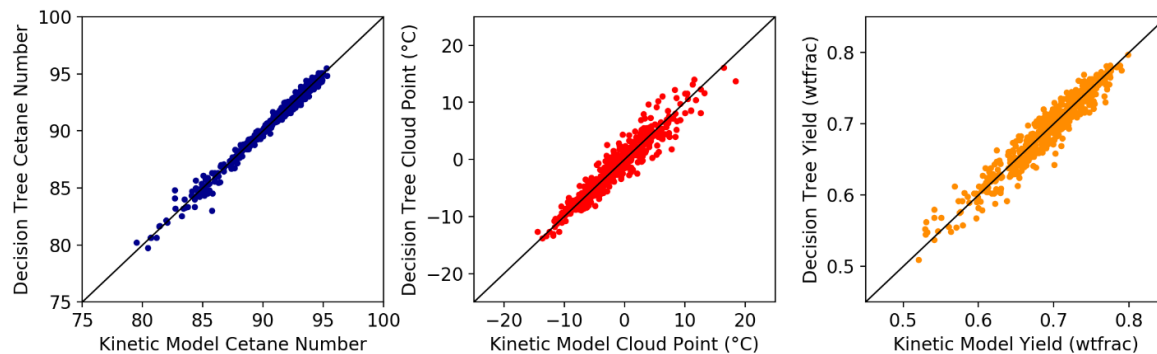
Average coefficient of determination, R^2 : 0.8732

Figure 5.3: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with four multilinear regression model partitioned by inlet hydrogen pressure for the testing data

Multilinear regression models can be extended to polynomial regression models to better capture the non-linearities in the data. However, the number of parameters can grow rapidly as the number of independent variables increases due to the associated combinatorial terms. While the model developer can greatly reduce the parameter terms by selecting the most significant terms in a design of experiments approach, it is time consuming and requires a lot of user input. Additionally, inverse and logarithmic dependencies in the product properties may still be difficult to identify and capture. Since the DDMs are generally disposable as soon as process conditions perturb outside the range of the data used to optimize the model, an approach with fewer user inputs can be more useful over time.

5.6 Machine Learning

Machine learning techniques are designed to automate analytical model building by learning from the underlying patterns in data. While multilinear regression is fundamentally a form of machine learning, several other algorithms exist for regression analysis that can perform better under certain conditions. One popular algorithm is a decision tree. Decision trees represent the data as a graph of nodes that perform binary splits by independent variables values to reach the prediction values.^{86,87} Figure 5.4 shows that the results from the decision tree regression perform just as well as results of the temperature-partitioned multilinear regression models in Figure 5.2. A benefit of decision tree regression models over the multilinear regression model is that only a single model is required to predict the entire process range of the data well. Furthermore, these models are almost equally as fast to train and provide parameters that can be visualized as a tree. The major downside is that decision tree models and most other machine learning models require advanced algorithms that can be difficult to implement or require external software packages like scikit-learn.

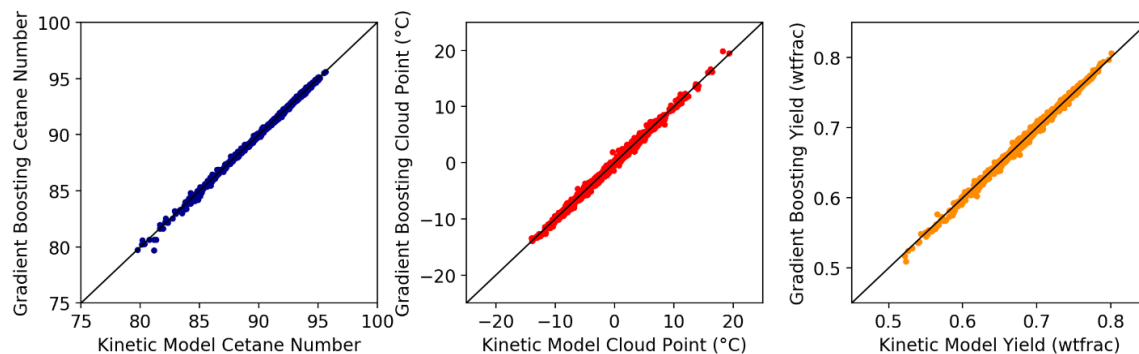


Training time (s): 0.5367
 Prediction time (s): 0.0030
 Average coefficient of determination, R^2 : 0.9398

Figure 5.4: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a decision tree regression model for the testing data

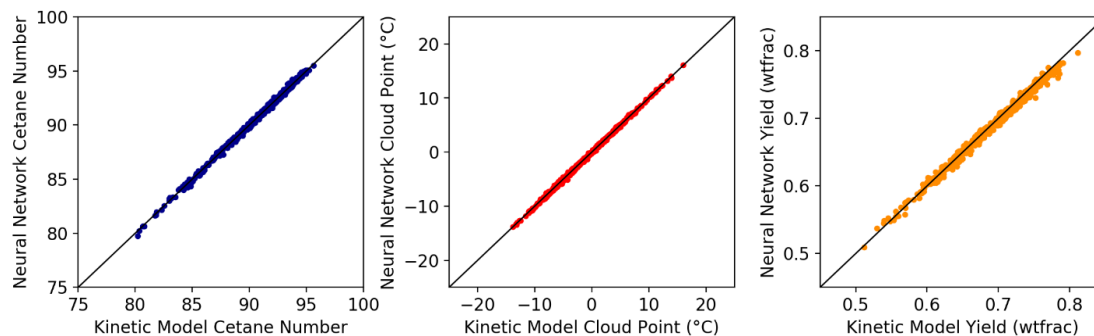
Beyond decision trees, several other types of machine learning regression algorithms exist that may be able to better capture the underlying relationships of the output properties of interest. Figures 5.5 and 5.6 show the results of using a gradient boosted regression tree (GBRT) and a single-layer artificial neural network, respectively, for the regression model. GBRTs reduce the error by using an ensemble of decision tree models that each improve the prediction. Artificial neural networks, specifically a multi-layer perceptron in this case, simulate artificial neurons containing one or more hidden layers that transform an input layer into an output layer using activation functions.^{86,87} Both models can capture the results of the MLKM extremely well over the entire data simulation range. The largest downside to these models is that their large number of model parameters are difficult to understand or obscured in the decision-making process. Training time can also be significantly longer than other

regression algorithms. Furthermore, they may require more user input due to more complex algorithms that can have sensitive algorithm variables.



Training time (s): 41.12
Prediction time (s): 0.1569
Average coefficient of determination, R^2 : 0.9953

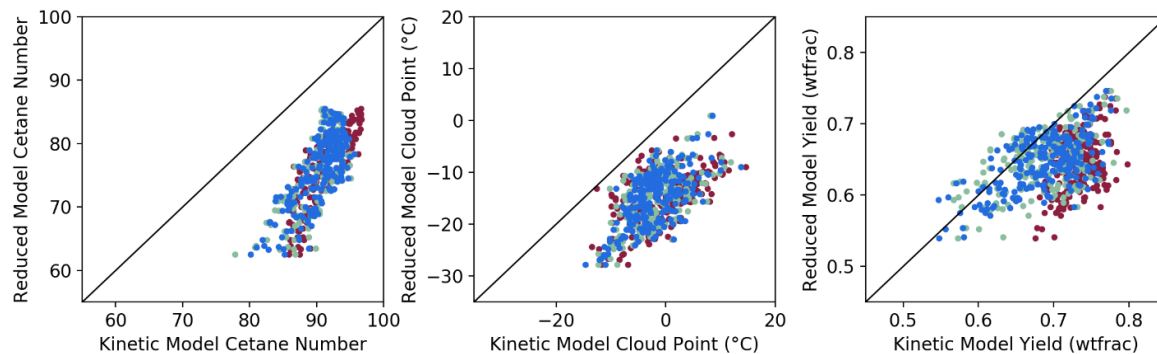
Figure 5.5: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a gradient boosted regression tree model for the testing data



Training time (s): 18.33
Prediction time (s): 0.0090
Average coefficient of determination, R^2 : 0.9955

Figure 5.6: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with an artificial neural network model for the testing data

For all DDMs, the training time is always short (less than one minute). The prediction times are also very short ($\ll 1$ second) compared to the time required to solve a MLKM. Therefore, any model can be a good candidate for an application that requires fast solution times like real-time optimization of a process unit. However, the DDMs provide very poor results if the parameter range of the input data exceeds the range of the model tuning. The MLKM was used to generate 2500 datasets for an input temperature range of 440 to 470 °C and the same ranges for the other input parameters. Figure 5.7 shows the prediction results of the new data for the multilinear regression, decision tree, and artificial neural network models trained on the original 20000 datasets. Since the original data did not contain any information about the 440 to 470 °C temperature range, the results are very poor. The negative value of the average coefficients of determination signify that the mean constant value models of the outputs greatly outperform any of the regression models. Consequently, the DDMs should not be utilized when the input process parameters exceed the bounds of the data used to optimize the models.



Multilinear average R^2 : -6.1068
 Decision tree average R^2 : -6.2262
 Artificial neural network average R^2 : -4.4838

Figure 5.7: Parity plots comparing the results of a) cetane number, b) cloud point, and c) yield in the triglyceride hydroprocessing MLKM with a multilinear regression (red), decision tree (green), and artificial neural network (blue) model for data outside the range of the training data for the DDMs

5.7 Impact of Data

An important consideration for generating data-driven models is the amount of data needed to create a good model. 5.8 displays the overall R^2 values for each of the multilinear regression, decision tree, and neural network models. Multilinear regression, where one model is used over the entire range, requires the least amount of data in order to reach an optimal prediction. However, the overall accuracy of the model may be lower than desired for highly non-linear systems over a broad range. Decision tree accuracy increases as the amount of data increases, up to more than 10^4 datasets, but at least 10^3 datasets are required to provide better accuracy than a multilinear regression model. The artificial neural network requires at least 10^3 datasets to have any accuracy but reaches a very high degree of accuracy soon after. While it is difficult to make an *a priori* estimate of the amount of data needed for a good fit for any of the types of regression models, the relative differences between the

three provide a qualitative assessment of the type of regression model that may be best suited to develop a DDM from a MLKM with a given number of datasets.



Figure 5.8: Number of datasets needed for each of a) multilinear regression, b) decision tree, and c) neural networks to generate models of a given accuracy

Data collection is the most likely bottleneck for generating data-driven models. In the present case, a simple MLKM was chosen for illustration purposes (1 reactor, ~500 model ODEs) that has a fast solution time (0.5-3 seconds). Therefore, generating 10^5 data points could be achieved in $\sim 10^2$ minutes, or proportionally less time with parallelization. With the triglyceride hydroprocessing MLKM, using one thread required 272 minutes and using four threads required 89.1 minutes to generate 10^5 datasets. When attempting the same strategy for a more complex model like a vacuum gas oil hydroprocessing unit with multiple reactors, multiple catalysts, and ~2000 model ODEs, the solution time can be 10^2 - 10^3 seconds per model run.⁹⁵ Other complex kinetic models built using KMT have once-thru simulation times ranging from 30 seconds to 30 minutes depending on the applications.^{7,15,60,96,97} Even with parallelization over the 4-8 threads available on standard CPUs, data generation can

take several days to weeks. But, generally, data generation would need to be done very infrequently due to unexpected changes in feed or process conditions. Most process and feed parameter ranges for established process units are well known, so generating a broad range of data at one point should allow the DDMs to be created to represent the months to years that a typical unit will operate.

5.8 Conclusions

DDMs for the representation of the information in MLKMs can effectively capture the relationships between the inputs and outputs of the model. Multilinear regression models are fast and easy to understand, but they can suffer from low accuracy if the input parameter range is too wide. Since determining whether a parameter range is too wide can be difficult *a priori*, a trial-and-error approach may be necessary to create accurate multilinear regression models. Decision trees can accurately capture larger parameter ranges than multilinear regression models but have much larger data requirements. GBRT and artificial neural networks both provide excellent predictive accuracy, but the models are difficult to understand and have large data requirements for good accuracy. Overall, multiple approaches can be used to generate DDMs from datasets generated from MLKMs. The most accurate specific algorithm is model dependent, but most algorithms can effectively work by varying model input parameter ranges and/or the algorithm variables until a good model is found. Once the input parameters go beyond the input parameter ranges, the DDM should be abandoned in favor of a new model from newly generated MLKM data.

5.9 Acknowledgement

Michael T. Klein acknowledges collaborations with and support of colleagues via the Saudi Aramco Chair Program at KFUMP and Saudi Aramco.

Chapter 6

THE INITIAL CONDITION GENERATOR: A SOFTWARE TOOL TO STATISTICALLY DETERMINE INDIVIDUAL MOLECULE FRACTIONS FROM EXPERIMENTAL MEASUREMENTS

6.1 Introduction

The Kinetic Modeler's Toolbox (KMT)^{7,8,11} developed in the Klein research group is capable of quickly generating kinetic models for a wide array of applications. A kinetic model, in its most basic form, is an organization of an initial value problem composed of a system of material balances with corresponding initial conditions as shown in Equation 6.1. In the equation, \bar{y} is a vector denoting the molar flow of each component in the system, \bar{y}_{in} and \bar{y}_{out} denote the flow into and out of the system, and $\bar{v}_i * rate_i$ represent the vector of stoichiometric coefficients multiplying the rates of reaction. The material balances depend on the specific type of reactor and kinetics of the process in question. The initial conditions refer to the initial flow of components in the system. For a molecular-level kinetic model, each molecule requires a corresponding mole or mass fraction associated with it along with an overall flow rate or feed charge amount. The focus of this chapter is the generation of these mole fractions for each molecule in the system in KMT.

$$\frac{d\bar{y}}{dt} = \bar{y}_{in} - \bar{y}_{out} + \sum_{i, reactions} \bar{v}_i * rate_i \quad (6.1)$$

$$\bar{y}(0) = \textit{feed composition}$$

Previously, a software tool called the Composition Model Editor (CME)⁹ was the basis for generating both the molecule identities and mole fractions for a desired

process in KMT. CME stochastically sampled possible molecular structures in a feedstock given some constraints like the boiling range of the molecules. The user could select several features that should be present in the feedstock like aromatic rings, heteroatoms, and double bonds. It then arranged a set of probability density functions (PDFs) for the generated molecules based on the ringed cores, intercore linkages, and side chain carbons of those molecules. These probability density function parameters could be juxtaposed to generate the individual mole fractions of each molecule. But, as kinetic model generation evolved over the years, CME became a less attractive option.

The historical route of model development using KMT was generating the molecule identities and mole fractions in CME, systematically writing the reactions of the generated molecules in the Interactive Network Generator (INGen)⁸, and solving the kinetic model in the Kinetic Model Editor (KME)⁷. However, over time, INGen was identified as the single piece of software that controls the number of molecules in the system, and consequently the speed of model solution on a computer. INGen specifies the seed molecules and reaction rules that allow molecules to be created in the system. Fewer or more molecules could be created by changing reaction rules and seeding networks with different rank limitations.¹⁷ The new modeling route was thus one of generating the entire reaction network and molecule set using INGen first and then generating the mole fractions for the molecules in the reaction network. Molecules that are not in the feed could be adjusted to have zero moles in the composition model. As more models were created in INGen, seed libraries were created that allowed the direct selection of typical molecule identities in most major refinery processes. Even in cases where a library did not exist, molecule identity

generation through reactions was easy to accomplish. Thus, the molecule identity generation function of CME became obsolete.

Additionally, a few limitations were identified in CME that prevented it to effectively function in certain cases. First, many cases were identified where the feedstock molecule identities were well known and only a customized representation was required to generate the mole fractions of the molecules. CME did not allow the user to easily customize the PDF structure of the model leading to poor results when optimizing the model. Secondly, when INGen was used to create a reaction network from the CME-generated molecule list, extraneous reactions and molecules were identified that did not add information to the model but increased its size. Larger than necessary models are undesirable because they are difficult to understand and slow to computationally solve. Finally, the CME Microsoft Excel interface with the Visual Basic for Applications (VBA) and C backend was slow, difficult to edit, and prone to errors due from software updates. An external software package to simulate a LINUX environment was also required that could make transferring the software to collaborators difficult.

Clearly, a better solution was needed for new model development that addressed these limitations. A C#-based application called the Initial Condition Generator (ICG) was developed that accepted a list of molecules, allowed the user to define a set of custom PDFs or choose an existing PDF representation, define the experiments used to generate feedstock data, and optimize the PDF parameters to match the simulated composition properties to the experimental properties. The application is designed to be easy to use, edit, and transfer. This chapter discusses the theory and the interface of ICG.

6.2 Molecule Properties

Once the molecule identities have been defined from the INGen reaction network, the next step in generating the mole fractions of each molecule is determining the properties of each molecule. The output of the property database application that is a part of KMT was linked to ICG to determine the properties for each molecule.¹⁰ Table 6.1 displays the list of properties generated by the property database. The properties represent a mixture of thermodynamic properties calculated using Gani⁷⁷ and Benson⁷⁸ group contribution methods. Structural properties like molecular weight, number of aromatic rings, number of side chains etc. are calculated using the ChemGraph⁸ routine to parse the structure of the molecules. ICG uses the properties as the basis for constraints used to limit the PDFs and the experimental bulk property calculations. Some customizations can also be added by assigning user-defined property values for the three “UserClass” properties if enough detail cannot be generated from the available properties.

Table 6.1: List of properties generated by the property database application

Tc	TotalHydrogenNum	SigmaD	DBE
Pc	Nap5RingNum	SigmaP	AlkylAroCNum
Vc	Nap6RingNum	SigmaH	PhnolAroCNum
Tb	SideChainNum	Sigma	FusdAroCNum
Tm	ThphRingNum	Omega	AmineNum
Hform	TotalSulfurNum	Vm	AmideNum
Gform	TotalNitrogenNum	Visa	AlcoholNum
CPa	Quan_HF	Visb	EtherNum
CPb	AromHydrogenNum	CP	CarboxylicAcidNum
CPc	AlphaHydrogenNum	VIS	EsterNum
CPd	ThioSulfurNum	TotalCarbonNum	CarbonylNum
Hfusion	SulfidSulfurNum	AromCNum	AldehydeNum
Hvap	MrcaptanSulfurNum	NaphCNum	MethylNum
LogKw	PyridineNitrogenNum	AromRingNum	UserClass1st
Fp	PyrroleNitrogenNum	NaphRingNum	UserClass2nd
Hvb	TotalOxygenNum	MW	UserClass3rd
Svb	ZNum	Density	

6.3 Probability Density Function Trees

After generating properties, a PDF representation is required to specify the mole fractions of individual molecules in the model. The idea behind the PDFs is to isolate structural moieties in a feedstock and apply distinct probabilities to the rate of appearance of those structural moieties in the feedstock given a set of experimentally measured properties. ICG uses a tree representation of the PDFs where the PDFs have parent-child relationships to efficiently discretize the desired structural moieties. Each generation of child PDFs further classifies molecules of the parent PDF. Sibling PDFs parallelly apply constraints to the molecules, where the overall molecule fraction is a juxtaposition of the individual bins of the PDFs. Any structural moiety not explicitly discretized by the PDFs is assumed to be equimolar. The individual mole fractions of a

molecule m is a juxtaposition of the probability of each PDF that is satisfied by the molecule properties, as shown in Equation 6.2.

$$m = \prod_{i.PDFs} PDF_i \text{ if } m \subseteq PDF_i \quad (6.2)$$

Structural features of the molecules can be individually discriminated in ICG using two types of PDFs: histograms (Equation 6.3) and gamma PDFs (Equation 6.4). Histograms are discrete bins of probabilities that are very versatile to define. The definition of each bin is independent of the other bins with only the requirement that the sum of all probabilities must be unity. These can be useful for defining evidently discrete feedstock properties like paraffin, olefin, naphthenic, and aromatic (PONA) distributions. Other applications can be the ring-number distributions for aromatic and naphthenic rings, heteroatom distributions, and side chain distributions. Each histogram has as many parameters as the number of bins less one. Gamma PDFs are continuous distributions that are discretized in ICG to represent the overall spread of a moiety. A key benefit of the two-parameter gamma is that it is versatile and so can assume many distribution shapes. Most importantly, gamma PDFs can effectively represent the carbon number distribution of the molecules in the system.

$$f(x) = H(x) \quad (6.3)$$

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}}{e^{\beta x} \Gamma(\alpha)} \quad (6.4)$$

For example, Figure 6.1 shows an example PDF structure of a basic petrochemical feedstock. The feed can be separated into PONA categories using a histogram. The paraffins can be further discretized by the number of branches, heteroatom content, and the carbon number. The aromatics can be classified by number of aromatic rings, number of heteroatoms, and number of fused naphthenic rings. Olefins and naphthenics can be similarly discretized. For the paraffins, if the

carbon number PDF is on the same level (sibling) as the branch number PDF, then all paraffins will have the same carbon number distribution irrespective of the number of branches a molecule has. Another way to represent the paraffins could be to assign a child PDF to each branch number. This would mean that each branch number can have an independent PDF designation of the carbon number, so that they are not all constrained equally.

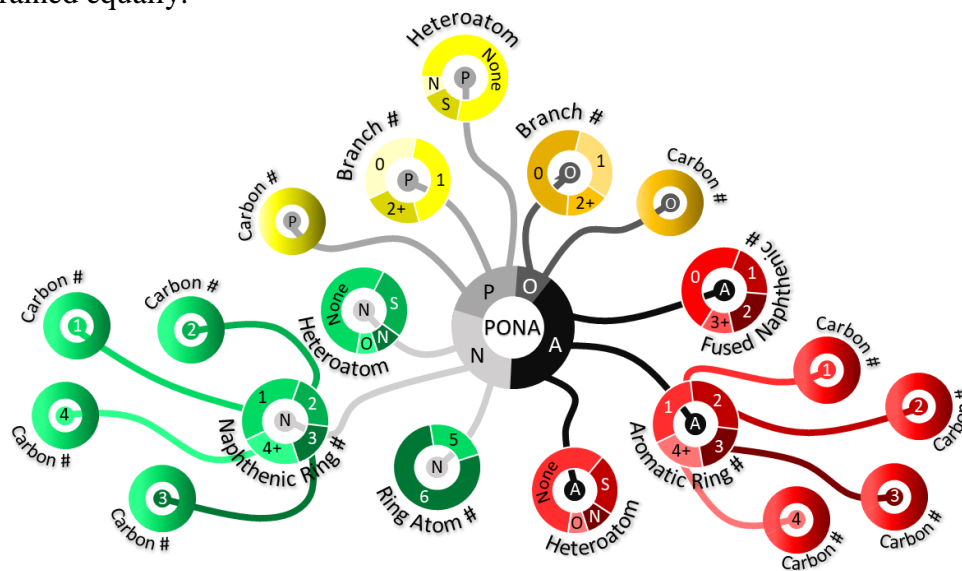


Figure 6.1: A sample PDF tree representation of a basic petrochemical feedstock.

The decision between assigning PDFs as child versus sibling PDFs can be very important. Assigning sibling PDFs reduces the overall number of PDFs, thereby simplifying the parameter optimization problem and making the PDF tree representation easier to comprehend. Child PDFs allow for better discretization of the pdfs. In the paraffin case, the argument that all paraffins have the same carbon number distribution is easily justifiable, and so a sibling carbon number distribution to the branch number distribution can be used. However, the same cannot be said for

aromatics. Aromatics will inherently have different overall carbon number (and even side chain carbon number) distributions owing to the underlying minimum carbon number requirements of having a certain number of aromatic rings. Therefore, the carbon number distribution of 1-ring aromatics and 4-ring aromatics can be very different. Assigning a sibling carbon number PDF to the aromatic ring number PDF would lead to poor results, and an independent gamma for each aromatic ring number should be used.

6.4 Bulk Property Correlations

The final definition required to model the composition is the experimental measurements. A structure/property correlation is needed to simulate the measured properties using the individual mole fractions of the molecules. ICG allows the user to define each experimental measurement using the properties of each molecule, where the bulk property is a mixing rule. Most experimental data can be represented in this manner. For example, the experimental density can be calculated as a weighted average of the individual molecule densities. Some experiments like the simulated distillation cuts, ultimate analysis, and H/C ratios that cannot be easily described by mixing rules have been pre-defined in ICG. It is anticipated that most properties of this nature already exist in the software, but if the properties do not, then some programming knowledge may be required to add them directly into the C# code.

The minimum data requirement for ICG is one experimental data point. Although, the feedstock answer with one data point has infinitely many good solutions with very low confidence in the output mole fractions. Additional experimental values defined in the bulk properties of ICG increase the size of the objective function used in parameter optimization. Consequently, additional experimental data increase the

confidence in the PDF parameters that define the mole fractions of the molecules. With the advent of advanced analytical techniques like GC-MS, GCxGC, and FTICR-MS that provide more detailed experimental classifications of feedstocks than ever before, the bulk property definitions in ICG can be expanded.⁵² This increases the overall accuracy of the mole fractions of the molecules in the system, providing a better overall kinetic model.

6.5 Parameter Optimization

Once the PDF structure and experimental properties have been defined, the PDF parameters must be optimized so that the simulated properties match the experimental properties. ICG uses a simulated annealing algorithm to perturb the parameter values and minimize an objective function of the form described in Equation 6.5. The core of ICG is the objective function designed to minimize the difference between the experimental and predicted values for each property in the system scaled by a weight. For the chi-square statistic, the weight is the standard deviation of the experimental measurement. However, often, the standard deviation of the experiment is not known though. A good heuristic for the weight is 1/10 to 1/100 the experimental value in that case.

$$obj = \sum_q \left(\frac{Exp_q - Pred_q}{Weight_q} \right)^2 \quad (6.5)$$

The starting point for parameter optimization is arbitrary. If experimental data is available for PDFs like the PONA split, then the data can directly inform the PDF probabilities. When data is not available, an equiprobable assumption is as good as any. Some heuristics can help lower the burden of parameter optimization for the simulated annealing algorithm. Firstly, the average and standard deviations of the

carbon number gamma PDFs of a feedstock can be determined from the simulated distillation cuts. Libraries like NIST Chemistry Webbook³⁹ provide a good compilation of pure component boiling points that can be used to determine the corresponding carbon number gamma PDF parameters. Secondly, the impact of PDF parameters on certain bulk properties should be considered. For example, aromatics tend to be denser than paraffins. Therefore, if the density prediction is poor, the ratio of the paraffin to aromatic probability can be adjusted. Finally, general knowledge about the feedstock and products provides information about the PDF parameters. As an example, it is known that crude oil generally contains a low amount of olefinic content, so unprocessed feeds can be assumed to have a low olefin PDF probability. Conversely, a feedstock that is an effluent of a hydrogen deficient cracking unit probably contains an amount of olefins proportional to the cracking severity. The PDF probability of olefins can thus be adjusted to reflect the cracking severity.

6.6 ICG User Interface

One of the main ideas behind the ICG application was the ability for users that are not experts in the ICG software or in kinetic model building to easily be able to create and optimize the feed models. Careful consideration was placed in enhancing the user experience and guiding the user through the steps of the software. Figure 6.2 shows an image of the home page of the ICG software. The home page allows the user to load a pre-existing model, create a new model, delete an existing model, select a different dataset in a model, and exit the program. A list of the existing models is provided, where an entry can be selected to be loaded. The buttons are designed to become mouse-clickable based on certain conditions. The 'load', 'save', and 'delete' model buttons become available once a model is selected in the composition model

list. The ‘create new model’ button becomes available if a valid model name is typed in the preceding box. The dataset buttons become available once a model is loaded and datasets are available.



Figure 6.2: Home page of the ICG application user interface

Creating a new model or loading an existing model populates four tabs at the top of the interface, as shown in Figure 6.3. For the purposes of this illustration, a new model was created. The four steps are 1) molecule list setup, 2) experimental data setup, 3) optimization setup, and 4) run optimization and view results. In the first step, molecules and properties are defined from the property database along with the PDF representation of the molecules. The second step allows the user to define the experiments and add the measured experimental values. The third step allows the user to set the PDF parameter bounds and limit the steps of the simulated annealing

algorithm. The final step allows the user to test the model, optimize the PDF parameters, and see the property and mole fraction results. The functionality of each step will be explored in turn.

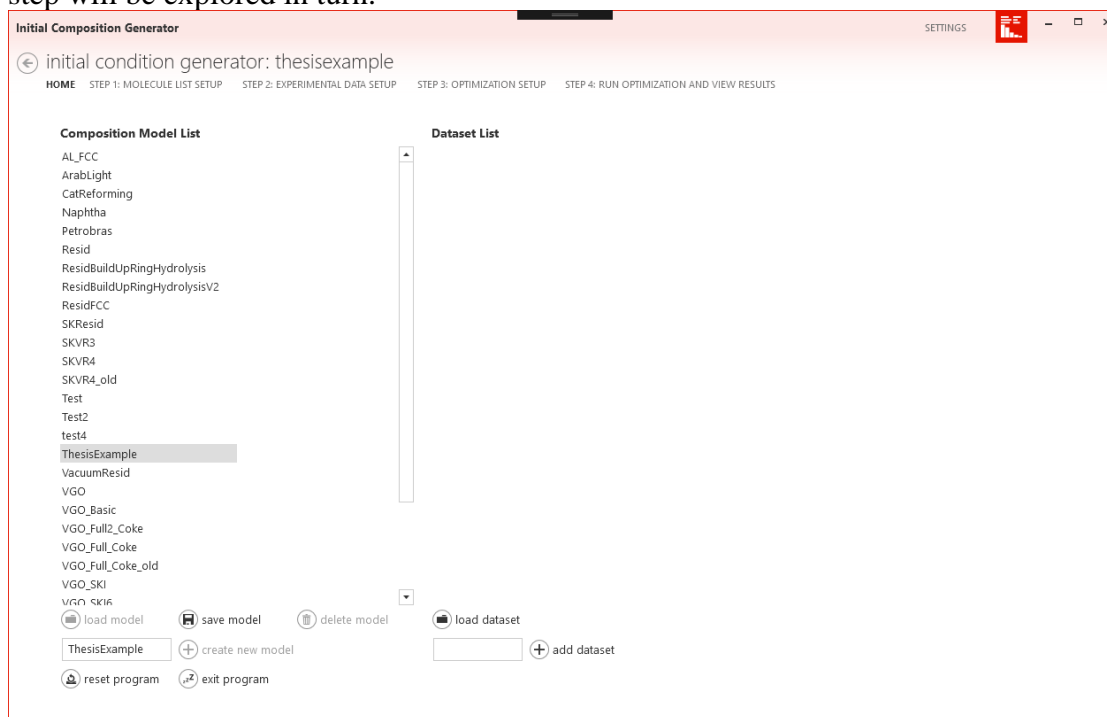


Figure 6.3: Automatic population of the four tabs representing the ICG feedstock model generation steps after creating a new model

Figure 6.4 shows the initial page for step 1. After clicking on the ‘select molecule list from database’ button, a window prompts the user to navigate to the desired list of molecules in the property database, as shown in Figure 6.5. These molecules lists have been populated from INGen models. Selecting a molecule list adds the molecules and properties to the ICG model. The next action on the step 1 page is the setup of the PDF list structure. The PDF list structure can be manually defined by choosing the appropriate type in the dropdown menu and clicking on the

'add' button. This opens a window depending on the type of PDF where the PDF can be defined based on desired constraints, as shown in Figure 6.6. The PONA PDF defined in the figure has four bins, each with its own constraints based on the available properties. For the naphthenics bin for example, the constraints are that the naphthenic ring number must be greater than zero and the aromatic ring number must be equal to zero. Additional or fewer constraints can be added as needed. Additionally, a parent PDF and associated bin can be specified in this window. Once the 'ok' button is pressed, the PDF is added to the ICG model.

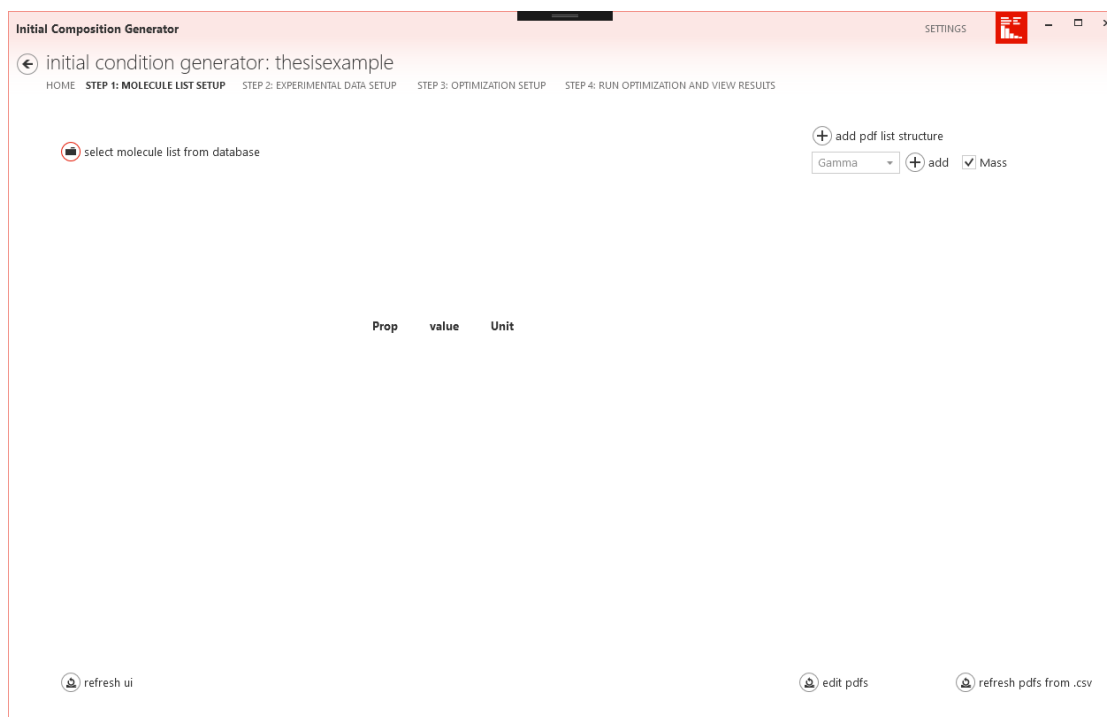


Figure 6.4: Step 1 for the ICG feedstock model construction

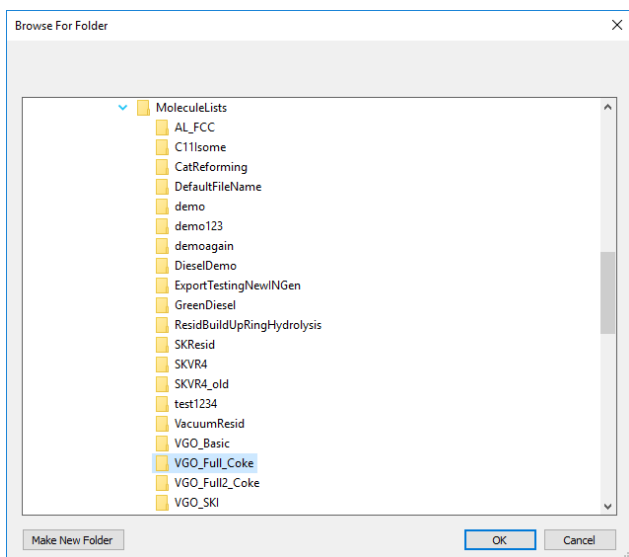


Figure 6.5: Navigation menu to the property database molecule list

ADD HISTOGRAM

PONA

load prebuilt histogram

4 adjust parameter #

X-Axis Category Fraction

P	0.25
O	0.25
N	0.25
A	0.25

Add Constraints

NaphRingNum > 0

add constraint

Property	Operator	Value
AromRingNum	==	0
NaphRingNum	>	0

Parent PDF Selection: No Parent

Select Associated Parent PDF Independent Variables

isAssociated	IndependentVariable

ok cancel

Figure 6.6: Histogram addition window for a new custom histogram

Adding each PDF manually may be needed for a new type of model, but some typical PDF tree structures have been pre-defined in the software. This allows for the quick specification of the PDF tree. In this case, a vacuum gas oil (VGO) PDF structure was chosen by clicking on the ‘add pdf list structure’ button and selecting the VGO list in the window that appears, as shown in Figure 6.7. The user also has the ability to define and add new PDF templates by adding more ‘.csv’ containing PDF definitions. These PDF ‘.csv’ files can be generated by manually adding the PDFs in the interface and navigating to the model folder, where all the PDF files exist in the PDF folder. Once the molecule list and PDF structure have been defined, the molecules, their properties, the PDFs, and the PDFs each molecule belongs to can be explored on the step 1 page, as shown in Figure 6.8. The user is now ready to proceed to the next step.

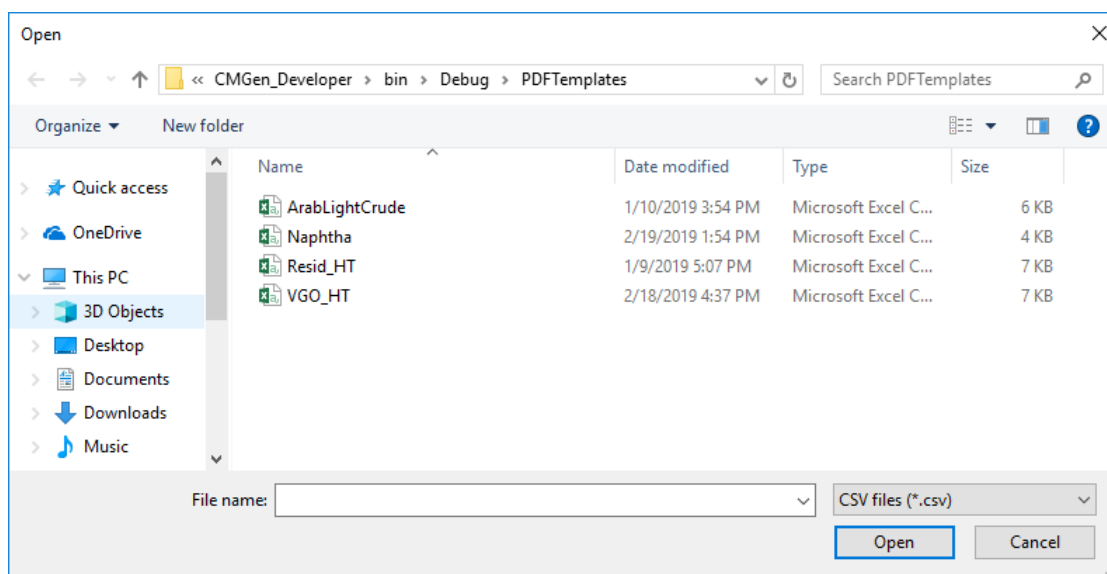


Figure 6.7: Selection of pre-defined PDF tree structures in ICG

The screenshot shows the 'Initial Composition Generator' software interface. The main window title is 'Initial Composition Generator' and the subtitle is 'initial condition generator: thesisexample'. The interface is divided into several sections:

- Species List:** A list of species from Species1 to Species25. 'Species9 non-1-ene' is highlighted in red.
- Molecule Details:** Displays a skeletal structure of 1-nonene (C₉H₁₈). Below the structure is a table of properties:

Prop	value	Unit
Tc	597.635503	K
Pc	24.294553	bar
Vc	527.5687	cm ³ /mol
Tb	427.648685	K
Tm	174.834155	K
Hform	-102.7554	kJ/mol
Gform	113.2613	kJ/mol
CPa	-8.284	J/mol/K
- PDFs:** A list of PDFs with columns 'pdfName' and 'pdfType'. 'PONA' is selected as the 'pdfName' and 'Histogram' as the 'pdfType'. Below this is a table of parameters and values:

Parameter	Value
P	0.056625057
P:Constraint 0	ZNum > 0
P:Constraint 1	SideChainNum == 0
I	0.080624081
I:Constraint 0	ZNum > 0
I:Constraint 1	SideChainNum > 0
O	0.023800024
O:Constraint 0	ZNum <= 0

Figure 6.8: Exploring the molecules and PDFs after completing step 1 in ICG

In the step 2 page, the user can define the desired bulk properties as shown in Figure 6.9. Experiments can be added from the built-in experiments list or custom experiments can be designed. The built-in experiments can simply be selected and added by clicking on the ‘add selected experiments’ button. When a custom experiment is desired, the ‘add single-user-defined experiment’ button can be clicked to open a new window shown in Figure 6.10. New experiments can be defined by setting constraints; for example, if the experimental value for the fraction of 1-ring aromatics is known, the property can be calculated by setting a constraint of ‘AromRingNum’ equal to one. This adds the experiment to the list. The user must also assign the numerical values of the experimental data on this page for the feedstock being modeled.

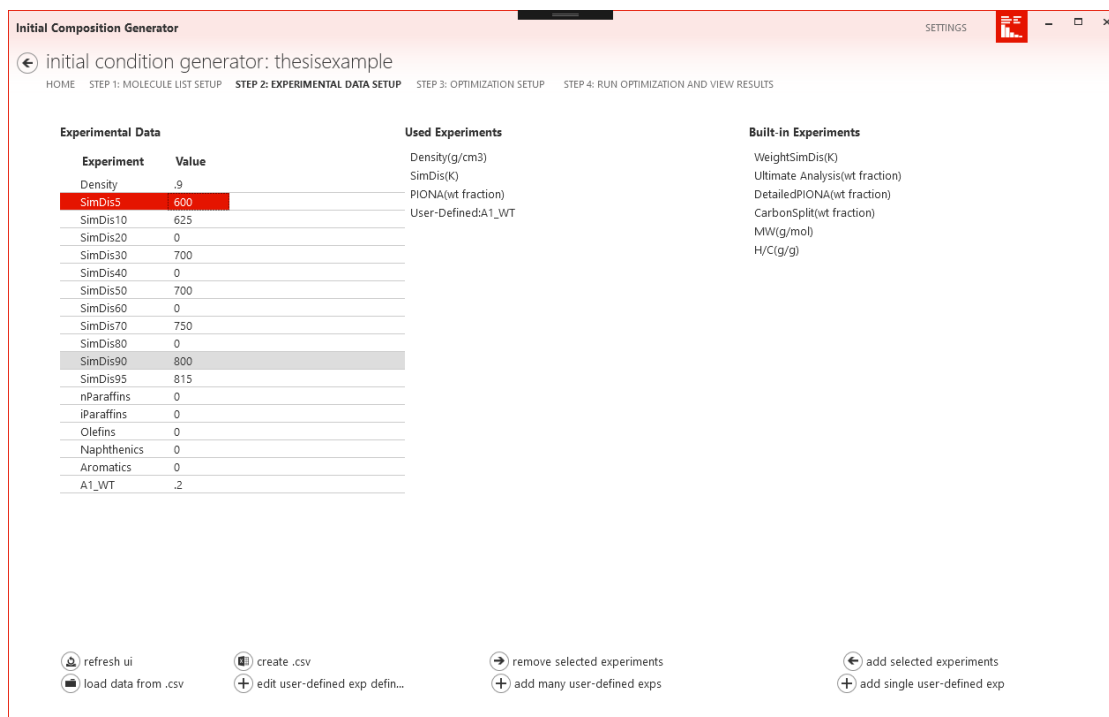


Figure 6.9: Defining the experimental properties in step 2 of ICG

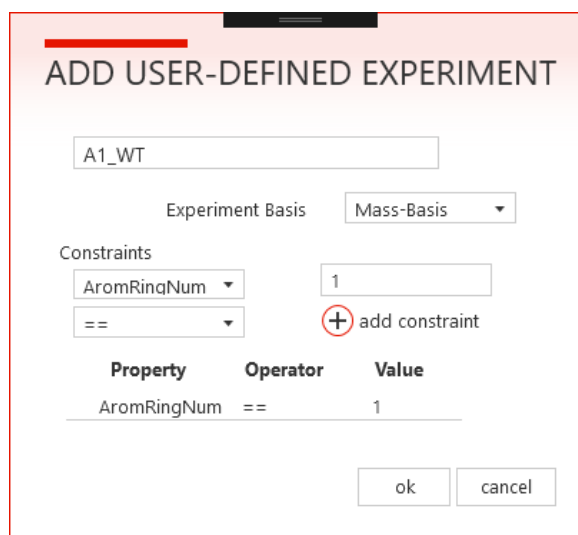


Figure 6.10: Adding a user-defined experiment in ICG

Step 3 in ICG allows the user to set the weights of each experiment in the objective function as well as guesses for the PDF parameters values, as shown in Figure 6.11. A zero weight for a property signifies that the property is not used in the objective function. Otherwise, the smaller the weight, the larger the contribution of a given predicted and experimental property differential. Initial guesses can be provided for the PDF parameter values along with the lower and upper bounds of the parameters. Narrow bounds can help the optimization algorithm perform better, but if the bounds are too narrow, they may not contain the requisite PDF parameter value for good results. If no information is available, the ‘10% bounds’ button automatically sets the bounds to +/- 10% of the PDF parameter value.

Initial Composition Generator

initial condition generator: thisisexample

HOME STEP 1: MOLECULE LIST SETUP STEP 2: EXPERIMENTAL DATA SETUP **STEP 3: OPTIMIZATION SETUP** STEP 4: RUN OPTIMIZATION AND VIEW RESULTS

Objective Function Weights

ExpName	Value	Weight
Density	0.9	0.01
SimDis5	600	5
SimDis10	625	1
SimDis20	0	0
SimDis30	700	1
SimDis40	0	0
SimDis50	700	1
SimDis60	0	0
SimDis70	750	1
SimDis80	0	0
SimDis90	800	1
SimDis95	815	5
nParaffins	0	0
iParaffins	0	0
Olefins	0	0
Naphtthenics	0	0
Aromatics	0	0
A1_WT	0.2	0.01

Parameter Setup

isOpt	Name	Value	LB	UB
<input checked="" type="checkbox"/>	PONA: P	0.056625	0.054360054	0.058890059
<input checked="" type="checkbox"/>	PONA: I	0.080624	0.077399117	0.083849044
<input checked="" type="checkbox"/>	PONA: O	0.0238	0.022848023	0.024752025
<input checked="" type="checkbox"/>	PONA: N	0.41609	0.399446799	0.432734033
<input checked="" type="checkbox"/>	PONA: A	0.42286	0.405946006	0.43977484
<input checked="" type="checkbox"/>	CycleCount_N: 1	0.20558	0.197357589	0.213804055
<input checked="" type="checkbox"/>	CycleCount_N: 2	0.28795	0.276433106	0.299469198
<input checked="" type="checkbox"/>	CycleCount_N: 3	0.23559	0.226167305	0.24501458
<input checked="" type="checkbox"/>	CycleCount_N: 4	0.18834	0.180807123	0.195874383
<input checked="" type="checkbox"/>	CycleCount_N: 4plus	0.082536	0.079234877	0.085837783
<input checked="" type="checkbox"/>	CycleCount_A: 1	0.2392	0.229633148	0.248769244
<input checked="" type="checkbox"/>	CycleCount_A: 2	0.41254	0.39604038	0.429043745
<input checked="" type="checkbox"/>	CycleCount_A: 3	0.22722	0.218132291	0.236309982
<input checked="" type="checkbox"/>	CycleCount_A: 4	0.097811	0.093899029	0.101723949
<input checked="" type="checkbox"/>	CycleCount_A: 4plus	0.023224	0.022295151	0.024153081
<input checked="" type="checkbox"/>	P_CarbonNum: Average	28.672	27.52512	29.81888
<input checked="" type="checkbox"/>	P_CarbonNum: Standard	7.9567	7.638432	8.274968
<input checked="" type="checkbox"/>	I_CarbonNum: Average	28.62	27.4752	29.7648
<input checked="" type="checkbox"/>	I_CarbonNum: Standard	6.8901	6.614496	7.165704
<input checked="" type="checkbox"/>	O_CarbonNum: Average	28.975	27.816	30.134
<input checked="" type="checkbox"/>	O_CarbonNum: Standard	6.7747	6.503712	7.045688
<input checked="" type="checkbox"/>	N1_CarbonNum: Average	26.337	25.28352	27.39048
<input checked="" type="checkbox"/>	N1_CarbonNum: Standard	6.4576	6.199296	6.715904
<input checked="" type="checkbox"/>	N2_CarbonNum: Average	27.101	26.01696	28.18504
<input checked="" type="checkbox"/>	N2_CarbonNum: Standard	7.7405	7.43088	8.05012
<input checked="" type="checkbox"/>	N3_CarbonNum: Average	28.326	27.19296	29.45904
<input checked="" type="checkbox"/>	N3_CarbonNum: Standard	6.3588	6.104448	6.613152
<input checked="" type="checkbox"/>	N4_CarbonNum: Average	30.512	29.29152	31.73248

10% bounds

$$F = \sum_{i=1}^{n_{\text{Experiments}}} \left(\frac{\text{obs}_i - \text{pred}_i}{\text{weight}_i} \right)^2$$

Advanced Optimizer Settings

Simulated Annealing

Steps per Temperature: 500

refresh ui open .csv load .csv open .csv load .csv

Figure 6.11: Defining the optimization problem in step 3 of ICG

The final step in ICG allows the user to test the model with the current parameters, run optimization, and view the results, as shown in Figure 6.12. The ‘test model’ runs the model with the current parameters to predict the calculated properties and mole fractions. The ‘tune model’ button optimizes the model to minimize the objective function value displayed near the top-left corner. A graph shows the evolution of the objective function through the optimization run. Once good simulated property results are achieved, the mass flow of the feedstock can be specified near the bottom left of the screen. Running the model and exporting the results to KME will thus connect the results of ICG with the kinetic model. At this point, the initial conditions of the initial value problem setup by the kinetic model are defined. The user can return to the home screen and add additional dataset if more experimental datasets are available.

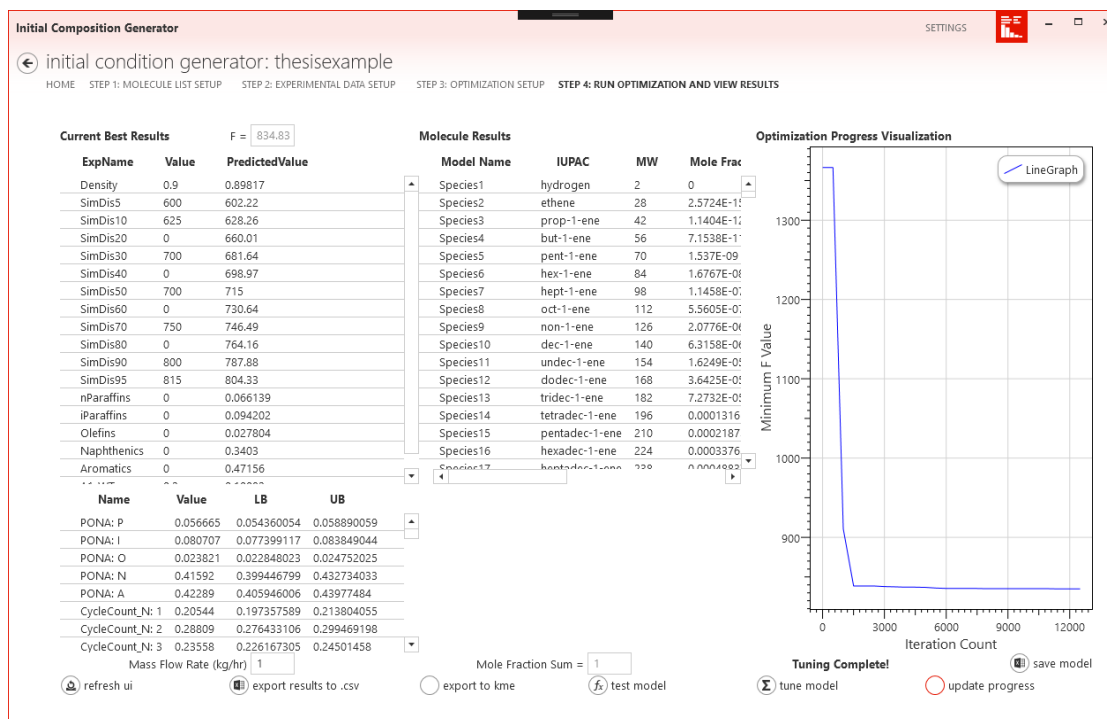


Figure 6.12. Optimizing and viewing the results in step 4 of ICG

6.7 ICG Power User

While the ICG interface provides a user-friendly environment for ICG model development, advanced users may wish to employ some shortcuts to reduce model building time. Most of the model information is stored in '.csv' in the model folder. By studying the format of the files, definitions of new PDFs and experimental properties can be easily and quickly achieved in the files. The interface is linked to the files, and any changes to the files can be reflected in the interface by clicking on the appropriate load button in steps 1, 2, and 3. This can greatly speed up model building time. For example, the PDFs file is shown in Figure 6.13. The structure of the PDFs can be studied and replicated to define the entire PDF tree in the file rather than adding the PDFs one-by-one in the interface. The file can then be loaded into the

interface by clicking on the ‘refresh pdfs from .csv’ button. In the same manner, the experimental properties and data can be specified in the ‘.csv’ files in the model folder to avoid having to add them manually one at a time.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	HISTOGRAM	PONA	PARENT	No Parent									
2	P	0.056625057	0.05436	0.0588901	ZNum	>	0	SideChainNum	==	0			
3	O	0.023800024	0.022848	0.024752	ZNum	<=	0	AromRingNum	==	0	NaphRingNum	==	0
4	N	0.416090416	0.399447	0.432734	ZNum	<=	0	AromRingNum	==	0	NaphRingNum	>	0
5	A	0.422860423	0.405946	0.4397748	ZNum	<=	0	AromRingNum	>	0			
6	END	HISTOGRAM	PONA										
7	HISTOGRAM	CycleCount_N	PARENT	PONA	N								
8	1	0.205580822	0.197358	0.2138041	AromRingNum	==	0	NaphRingNum	==	1			
9	2	0.287951152	0.276433	0.2994692	AromRingNum	==	0	NaphRingNum	==	2			
10	3	0.235590942	0.226167	0.2450146	AromRingNum	==	0	NaphRingNum	==	3			
11	4	0.188340753	0.180807	0.1958744	AromRingNum	==	0	NaphRingNum	==	4			
12	4plus	0.08253633	0.079235	0.0858378	AromRingNum	==	0	NaphRingNum	>	4			
13	END	HISTOGRAM	CycleCount_N										
14	HISTOGRAM	CycleCount_A	PARENT	PONA	A								
15	1	0.239201196	0.229633	0.2487692	AromRingNum	==	1						
16	2	0.412542063	0.39604	0.4290437	AromRingNum	==	2						
17	3	0.227221136	0.218132	0.23631	AromRingNum	==	3						
18	4	0.097811489	0.093899	0.1017239	AromRingNum	==	4						
19	4plus	0.023224116	0.022295	0.0241531	AromRingNum	>	4						
20	END	HISTOGRAM	CycleCount_A										
21	GAMMA	P_CarbonNum	TotalCarb	PARENT	PONA	P							
22	Average	28.672	27.52512	29.81888									
23	SD	7.9567	7.638432	8.274968									
24	END	GAMMA	P_CarbonNum										

Figure 6.13: The definition of the PDF tree in the '.csv' file format

6.8 Current and Future Development

As with all software, development of ICG is ongoing. Features can be added or removed as needed. There are a few areas that could be developed in the future to improve the overall user experience. First, it may be useful to allow the user to edit molecule properties in the interface. The specific need for this is usually when the custom “UserClass” properties need to be adjusted based on PDF definitions. It may be beneficial to add another type of optimization algorithm like a genetic algorithm

that could perform better in certain models. Analogous to the export to KME button, a new export to DMB button can be added as the Dynamic Model Builder (DMB) becomes more integral to KMT.¹¹ Finally, built-in models can be delivered for common feedstocks like naphtha, diesel, and vacuum gas oil that only require the user to input experimental data. These models would greatly simplify the ICG model building process for users not interested in customizing the models.

6.9 Summary

ICG is a user-friendly application that can allow the user to quickly determine the mole fractions of molecules in a feedstock model based on the available experimental data. The user follows four distinct steps to setup the molecules, define the PDFs, define the experimental data, and optimize the PDF parameters. The molecules are defined in the INGen model and the properties are calculated using the property database. PDF definitions can be customized based on the specific model and user needs or the user can select a PDF structure from the pre-existing list. Experimental properties in ICG can be setup based on the available experimental data for the model. An automatic optimization algorithm then minimizes the difference between the properties simulated by ICG and the experimental data based on the objective function weights for each property. The final output from ICG is a mole fraction value for each molecule defined in an INGen-created molecule list.

Chapter 7

SUMMARY, CONCLUSIONS, AND FUTURE WORK

7.1 Summary

This dissertation presented two parallel goals: 1) modeling conventional and unconventional hydroprocessing feedstocks and 2) further developing the Kinetic Modeler's Toolbox (KMT) to improve the model building process. The goals of this project were achieved by modeling triglyceride and vacuum gas oil hydroprocessing. The needs of the models dictated the development of the software tools used to build and solve model. The modeling tools effectively allowed for models to be built and be presented in a user-friendly manner to collaborators who can employ the models to make important processing decisions.

First, a molecular-level kinetic model was created for the hydroprocessing of triglycerides, which are an unconventional hydroprocessing feedstock. Triglycerides have three 8 to 24 carbon containing fatty acid chains on a propane backbone. They are processed to remove the oxygen atoms in the molecule, producing mostly straight chain paraffins that are a renewable diesel fuel alternative. In a hydroprocessing reactor, triglycerides can undergo three parallel reaction pathways: decarboxylation, decarbonylation, and hydrodeoxygenation. In the decarbonylation and decarboxylation pathways, the oxygen is lost as CO or CO₂, respectively, requiring little to no oxygen at the loss of some carbon yield in the product. The hydrodeoxygenation pathway removes the oxygen as water at a higher H₂ consumption cost but without carbon loss

to greenhouse gases. The final reaction network contained 476 species that underwent 1709 reactions.

The reaction network and species were used to set up the material balances in the kinetic model. Kinetic parameters were optimized for soybean oil and coconut oil feeds that represented various temperatures, pressures, and catalyst contact times. The final model represented the product yields well for all conditions. After predicting the molecular answer, cetane number and cloud point property models were optimized and used to estimate the value of the product diesel. Both property models predicted the experimental data well. The product renewable diesel had high cetane numbers, which signify good ignition quality, but also high clouds points, which can cause solidification issues in certain climates.

Next, a molecular-level kinetic model was created for a vacuum gas oil (VGO) hydroprocessing unit based on real refinery data. Vacuum gas oil is a conventional, heavy fraction of crude oil that need to be processed to remove impurities and reduce the overall molecular weight. First, molecules were selected that are representative of the typical structural attribute arrangements in crude oil. Literature studies and experimental results were used to select the typical cracking, saturation, hydrodesulfurization, hydrodenitrogenation, hydrodemetallization, dealkylation, and isomerization reactions that can occur in a hydroprocessing reactor. The final reaction network contained 5747 reactions and 1532 unique species up to 45 carbons covering molecules up to five aromatic rings with some heteroatoms.

The individual mole fractions of the species in VGO were calculated based on experimental bulk properties using the Initial Conditions Generator (ICG) tool described in detail in Chapter 6. A library of probability density function (PDF)

parameters was generated containing 21 datasets. The PDF parameters of the closest representative experimental dataset in the PDF library were used for each new dataset to minimize the burden of the optimization problem. The ICG simulated results of the feedstock showed excellent agreement with the experimental values.

In the kinetic model for VGO hydroprocessing, the reactor system was divided into a series of 19 pseudo-PFRs representing each catalyst layer in the system. Quench streams were added at the appropriate locations representing the beginning of a new reactor bed. Each pseudo-PFR was modeled as undergoing side-by-side reaction and vapor-liquid equilibrium. Due to the difference in catalyst activities, an independent parameter was used for each reaction family to model the difference in activity of one catalyst from another. Additionally, the deactivation due to coking and metal deposition were included in the simulation as a deactivation parameter scaling the rate laws. The application of quantitative structure/reactivity correlations reduced the number of parameters in the model. The kinetic parameters were then optimized using a simulated annealing algorithm based on the experimental data. The model results show good agreement with the experimental data.

A weakness of molecular-level kinetic model is that the solution time of the model on a standard computer can range from 10^{-1} to 10^3 seconds. The more the number of differential equation in the model, the slower the solution time, which may be too slow for some applications of the model like real-time optimization. A like-kind model was needed to access the information in molecular-level kinetic models that could be consistently solved in <1 second regardless of the number of differential equations in the kinetic model. To address this, data-driven models were created. Multilinear regression models provided a fast and easy to understand approach to

model data simulated by the kinetic model. However, they only worked well in narrow input parameter ranges. Various machine learning algorithms were used to regress the data and showed excellent agreement with the results of the molecular-level kinetic model. However, the machine learning models had high data requirements and required more advanced algorithms that can be difficult to implement or require external software packages.

7.2 Conclusions

KMT is an effective tool for modeling hydroprocessing systems at the molecular-level for both conventional and unconventional feedstocks. The tool can scale between small (~500 species, 1 reactor) and large (~1500 species, 19 pseudo-reactors) systems with different feeds, reactor configurations, and catalysts. For a system with different catalysts, it is possible to effectively capture the activity of multiple different catalysts in the model using the concept of catalyst linear-free energy relationships (LFERs). For the product, bulk property correlations can effectively determine the value of the product based on end-use. Additionally, molecular-level kinetic models can effectively be used in applications that require fast solution times via data driven models. While the specific algorithm for generating the data-driven model can vary, all algorithms produce results very quickly. The only limitation in developing these models can be the time required to generate the data, but since data only need to be generated infrequently, the data-driven models can function well over the long timescale of the process run.

7.3 Recommendations for Future Work

Future work may focus on both developing KMT model building capabilities and developing molecular-level kinetic models. Other than some quality-of-life improvements that can be made to improve the user experience, the greatest area for the improvement of KMT would be in removing its dependence on external software packages like Cygwin. In this dissertation, the Cygwin dependence of the composition modeling tool CME was removed with ICG. Previously, work by Lucio-Vega¹¹ removed the Cygwin dependence of KME with the DMB tool. The final major dependence on Cygwin is that of INGen. Removing that dependence can allow for the installation of the entire software suite as a 'clickable' installation file and users will not need to install external software packages. This will also avoid security issues with industrial collaborator.

A tool for chemically-based and kinetically-informed experimental data analysis can be very useful in reducing the overall model development time. Identifying possible issues in data prevents time from being used to solve an impossible problem. Firstly, typographical errors can occasionally be present in large data sets. Outlier identification can easily identify those problems. Then, there may be material balance errors in the data. A kinetic model is necessarily mass balanced and atom balanced if all material balances are written properly. However, experimental data may not necessarily be mass balanced stemming from experimental errors. An accounting of the additional or missing weight should be implemented based on the accuracy of the individual measurements. Atom balances are more difficult to identify. For example, in a hydroprocessing system, the amount of H₂ gas consumed should be evident by a proportional increase in the H/C ratio of the product. If there is more or less than expected hydrogen in the experimentally measured product, the model results

cannot match the experimental data. Finally, an outlier identification procedure may be useful for a collection of datasets. For example, in a series of vacuum gas oil hydroprocessing datasets with increasing inlet temperatures and everything else constant, there should be evident trends in the product results. Cracking activity should increase with increasing temperature, resulting in larger diesel and naphtha fractions. If a uniform trend of increasing diesel and naphtha fractions is not observed, then the datasets need to be analyzed to determine possible inconsistencies in the operation. In this manner, experimental data can be analyzed to determine if a kinetic model can possibly accurately represent the data.

Further model development may focus on increasing the feedstock variability of the process. For triglyceride hydroprocessing, additional fatty acid chains can be considered that sometimes appear in vegetable oil mixtures. Some special fatty acid chains that appear in mixtures derived from animal products or algae can also be considered. This would allow the model to truly be able to represent any oil mixture rather than common plant-based oil mixtures. In a real process, the triglyceride feedstock must depend on the regional availability of oils, the seasonal dependence of plant growth, and the base costs. Having a single model for all cases would greatly improve the usefulness of the model. The underlying chemistry and kinetics are very similar for most fatty acid chains, so the modifications need not be large.

The vacuum gas oil hydroprocessing network has already been extended to an atmospheric resid (AR) feed to simulate a resid hydrodesulfurization unit during the course of this dissertation. However, the results of that study cannot be disclosed for proprietary reasons. In general, AR is a mixture of vacuum gas oil and vacuum resid fractions that represent the “bottom-of-the-barrel” feeds. The vacuum resid portion of

AR is characterized by heavy, highly condensed molecules with a larger heteroatom presence than in vacuum gas oil. AR typically contains a heteroatom presence of 4-5 wt% sulfur, 0.2-0.4 wt% nitrogen, and a large fraction of heavy metals like nickel, vanadium, and arsenic. Heteroatoms greatly reduce the value of the feed and also cause significant deactivation and inhibition during hydroprocessing. Due to the variability of resid feeds by the crude oil source and the complexity of the asphaltene-, resin-, and polar-type molecules in heavy feeds, a study of structure variability can improve the model representation of the real system.

A vertical extension of this work would be to combine the conventional hydroprocessing feeds like vacuum gas oil with unconventional feeds like triglycerides. The processing system and catalysts are the same, so the extension is viable. The major drawback of adding the oxygen-rich triglycerides is the inhibition effect of the oxygen on the hydrodesulfurization and hydrodenitrogenation activity of the process. Additionally, the presence of sulfur and nitrogen can inhibit the hydrodeoxygenation activity. Therefore, experimental data are needed to capture the inhibition effect in the model. The first order model for inhibition would rely on the adsorption parameters for nitrogen, oxygen, and sulfur to simulate inhibition. Further refinement can be added by introducing specific terms like a Langmuir dependence in a heteroatom-based inhibition parameter.

A final area of future exploration may be employing the data-driven models in a real-time optimization environment. The data generation and data-driven model development can be automated for a given system. The fast solution times ($\ll 1$ second) of the data-driven models would allow them to be used effectively to manage process conditions. Data generation can continue in the background to build models

for future process ranges or to improve the current data-driven model. Additionally, a feedback loop can be established that informs improvements in the kinetic model parameters based on the failures of the data-driven model during real-time optimization.

REFERENCES

1. Crude Oil - Proved Reserves. In: *The World Factbook*. Central Intelligence Agency. <https://www.cia.gov/library/publications/the-world-factbook/rankorder/2244rank.html>.
2. *Use of Oil.*; 2018. https://www.eia.gov/energyexplained/index.php?page=oil_use. Accessed January 5, 2018.
3. *Annual Energy Outlook 2019.*; 2019. <https://www.eia.gov/outlooks/aeo/>.
4. Srivastava SP, Hancsók J. *Fuels and Fuel-Additives*. Hoboken, NJ: John Wiley & Sons, Inc; 2014. doi:10.1002/9781118796214
5. Klass DL. *Biomass for Renewable Energy, Fuels, and Chemicals*. Elsevier; 1998. doi:10.1016/B978-0-12-410950-6.X5000-4
6. Kim SK, Han JY, Lee H shik, Yum T, Kim Y, Kim J. Production of renewable diesel via catalytic deoxygenation of natural triglycerides: Comprehensive understanding of reaction intermediates and hydrocarbons. *Appl Energy*. 2014;116:199-205. doi:10.1016/j.apenergy.2013.11.062
7. Wei W, Bennett CA, Tanaka R, Hou G, Klein MT. Computer aided kinetic modeling with KMT and KME. *Fuel Process Technol*. 2008;89(4):350-363. doi:10.1016/j.fuproc.2007.11.015
8. Bennett CA. User-Controlled Kinetic Network Generation With INGen. 2009.
9. Hou Z. Software Tools for Molecule-Based Kinetic Modeling of Complex Systems. 2011.
10. Horton SR. Modeling Municipal Solid Waste Gasification: Molecular-Level Kinetics and Software Tools. 2016.
11. Lucio-Vega JC. Software tools to resolve the unique challenges of mega-molecular models. 2019.
12. Broadbelt LJ, Stark SM, Klein MT. Computer-generated pyrolysis modeling: On-the-fly generation of species, reactions, and rates. *Ind Eng Chem Res*.

- 1994;33(4):790-799. doi:10.1021/ie00028a003
13. Broadbelt LJ, Stark SM, Klein MT. Computer generated reaction modelling: Decomposition and encoding algorithms for determining species uniqueness. *Comput Chem Eng.* 1996;20(2):113-129. doi:10.1016/0098-1354(94)00009-D
 14. Moreno BM. Thermochemical Conversion of Biomass: Models and Modeling Approaches. 2014.
 15. Billa T, Horton SR, Sahasrabudhe M, et al. Enhancing the value of detailed kinetic models through the development of interrogative software applications. *Comput Chem Eng.* 2017;106:512-528. doi:10.1016/j.compchemeng.2017.07.009
 16. Zhang L, Hou Z, Horton SR, et al. Molecular Representation of Petroleum Vacuum Resid. *Energy & Fuels.* 2014;28(3):1736-1749. doi:10.1021/ef402081x
 17. Joshi P V., Freund H, Klein MT. Directed Kinetic Model Building: Seeding as a Model Reduction Tool. *Energy & Fuels.* 1999;13(4):877-880. doi:10.1021/ef980259r
 18. Klein MT, Hou G, Bertolacini RJ, Broadbelt LJ, Kumar A. *Molecular Modeling in Heavy Hydrocarbon Conversions.* Boca Raton, FL, FL: CRC Press; 2006.
 19. Bell R. The theory of reactions involving proton transfers. *Proc R Soc London Ser A.* 1936.
 20. Evans MG, Polanyi M. Further considerations on the thermodynamics of chemical equilibria and reaction rates. *Trans Faraday Soc.* 1936;32:1333. doi:10.1039/tf9363201333
 21. Horton SR, Klein MT. Reaction and catalyst families in the modeling of coal and biomass hydroprocessing kinetics. *Energy and Fuels.* 2014;28(1):37-40. doi:10.1021/ef401582c
 22. Mochida I, Yoneda Y. Linear free energy relationships in heterogeneous catalysis: II. Dealkylation and isomerization reactions on various solid acid catalysts. *J Catal.* 1967;7:393-396.
 23. Korre SC, Klein MT. Development of temperature-independent quantitative structure/reactivity relationships for metal- and acid-catalyzed reactions. *Catal Today.* 1996;31(1):79-91. doi:10.1016/0920-5861(96)00022-3

24. Burkholder D. A Preliminary Assessment of RIN Market Dynamics, RIN Prices, and Their Effects. *US EPA*. 2015.
<http://www.regulations.gov/#!documentDetail;D=EPA-HQ-OAR-2015-0111-0062>.
25. Mikulec J, Cvengroš J, Joríková L, Banič M, Kleinová A. Second generation diesel fuel from renewable sources. *J Clean Prod*. 2010;18(9):917-926.
doi:10.1016/j.jclepro.2010.01.018
26. Huber GW, Iborra S, Corma A. Synthesis of transportation fuels from biomass: Chemistry, catalysts, and engineering. *Chem Rev*. 2006;106(9):4044-4098.
doi:10.1021/cr068360d
27. Choudhary TV, Phillips CB. Renewable fuels via catalytic hydrodeoxygenation. *Appl Catal A Gen*. 2011;397(1-2):1-12. doi:10.1016/j.apcata.2011.02.025
28. Kalnes T, Marker T, Shonnard DR. Green Diesel: A Second Generation Biofuel. *Int J Chem React Eng*. 2007;5(1):1-15. doi:10.2202/1542-6580.1554
29. Sotelo-Boyas R, Liu Y, Minowa T. Renewable Diesel Production from the Hydrotreating of Rapeseed Oil with Pt / Zeolite and NiMo / Al₂O₃ Catalysts. *Ind Eng Chem Res*. 2011;50:2791-2799. doi:10.1021/ie100824d
30. Gong S, Shinozaki A, Shi M, Qian EW. Hydrotreating of Jatropha Oil over Alumina Based Catalysts. *Energy & Fuels*. 2012;26(4):2394-2399.
doi:10.1021/ef300047a
31. Kubička D, Kaluža L. Deoxygenation of vegetable oils over sulfided Ni, Mo and NiMo catalysts. *Appl Catal A Gen*. 2010;372(2):199-208.
doi:10.1016/j.apcata.2009.10.034
32. Altın R, Çetinkaya S, Yücesu HS. The potential of using vegetable oil fuels as fuel for diesel engines. *Energy Convers Manag*. 2001;42(5):529-538.
doi:10.1016/S0196-8904(00)00080-7
33. ASTM International. Standard Specification for Diesel Fuel Oils. 2019:1-25.
doi:10.1520/D0975-10.2
34. European Committee for Standardization. European Standard. EN 590. Automotive fuels - Diesel - Requirements and test method. 2009:1-12.
35. Azizan MT, Jais KA, Sa'aid MH, et al. Thermodynamic Equilibrium Analysis of Triolein Hydrodeoxygenation for Green Diesel Production. *Procedia Eng*. 2016;148:1369-1376. doi:10.1016/j.proeng.2016.06.603

36. Forghani AA, Jafarian M, Pendleton P, Lewis DM. Mathematical modelling of a hydrocracking reactor for triglyceride conversion to biofuel: model establishment and validation. *Int J Energy Res.* 2014;38(12):1624-1634. doi:10.1002/er.3244
37. Anand M, Sinha AK. Temperature-dependent reaction pathways for the anomalous hydrocracking of triglycerides in the presence of sulfided Co–Mo-catalyst. *Bioresour Technol.* 2012;126:148-155. doi:10.1016/j.biortech.2012.08.105
38. Ghosh P, Jaffe SB. Detailed Composition-Based Model for Predicting the Cetane Number of Diesel Fuels. *Ind Eng Chem Res.* 2006;45(1):346-351. doi:10.1021/ie0508132
39. Thermodynamics Research Center, NIST Boulder Laboratories MF director. NIST Chemistry WebBook, NIST Standard Reference Database Number 69. In: Linstrom PJ, Mallard WG, eds. Gaithersburg, MD. doi:10.18434/T4D303
40. Agarwal P, Evenepoel N, Al-Khattaf SS, Klein MT. Molecular-Level Kinetic Modeling of Methyl Laurate: The Intrinsic Kinetics of Triglyceride Hydroprocessing. *Energy & Fuels.* 2018;32(4):5264-5270. doi:10.1021/acs.energyfuels.8b00647
41. Kimura T, Imai H, Li X, Sakashita K, Asaoka S, Al-Khattaf SS. Hydroconversion of triglycerides to hydrocarbons over Mo-Ni/γ-Al₂O₃ catalyst under low hydrogen pressure. *Catal Letters.* 2013;143(11):1175-1181. doi:10.1007/s10562-013-1047-x
42. Ramos MJ, Fernández CM, Casas A, Rodríguez L, Pérez Á. Influence of fatty acid composition of raw materials on biodiesel properties. *Bioresour Technol.* 2009;100(1):261-268. doi:10.1016/j.biortech.2008.06.039
43. Alnajjar M, Cannella B, Dettman H, et al. CRC Report No . FACE-1 Chemical and physical properties of the fuels for advanced combustion engines (FACE) research diesel fuel. *CRC Rep.* 2010;(July).
44. Smagala TG, Christensen E, Christison KM, Mohler RE, Gjersing E, McCormick RL. Hydrocarbon renewable and synthetic diesel fuel blendstocks: Composition and properties. *Energy and Fuels.* 2013;27(1):237-246. doi:10.1021/ef3012849
45. Sandler SI. *Chemical, Biochemical, and Engineering Thermodynamics.* 4th ed. Hoboken, NJ: John Wiley & Sons; 2006.

46. Affens WA, Hall JM, Holt S, Hazlett RN. Effect of composition on freezing points of model hydrocarbon fuels. *Fuel*. 1984;63(4):543-547. doi:10.1016/0016-2361(84)90294-1
47. ExxonMobil. *2018 Outlook for Energy: A View to 2040.*; 2018. <https://exxonmobil.com/energyoutlook>.
48. Stangeland BE. A Kinetic Model for the Prediction of Hydrocracker Yields. *Ind Eng Chem Process Des Dev*. 1974;13(1):71-76. doi:10.1021/i260049a013
49. Jacob SM, Gross B, Voltz SE, Weekman VW. A lumping and reaction scheme for catalytic cracking. *AIChE J*. 1976;22(4):701-713. doi:10.1002/aic.690220412
50. Khang SJ, Mosby JF. Catalyst deactivation due to deposition of reaction products in macropores during hydroprocessing of petroleum residuals. *Ind Eng Chem Process Des Dev*. 1986;25(2):437-442. doi:10.1021/i200033a015
51. Laxminarasimhan CS, Verma RP, Ramachandran PA. Continuous lumping model for simulation of hydrocracking. *AIChE J*. 1996;42(9):2645-2653. doi:10.1002/aic.690420925
52. Rodgers RP, McKenna AM. Petroleum Analysis. *Anal Chem*. 2011;83(12):4665-4687. doi:10.1021/ac201080e
53. Quann RJ, Jaffe SB. Structure-Oriented Lumping: Describing the Chemistry of Complex Hydrocarbon Mixtures. *Ind Eng Chem Res*. 1992;31(11):2483-2497. doi:10.1021/ie00011a013
54. Jaffe SB, Freund H, Olmstead WN. Extension of Structure-Oriented Lumping to Vacuum Residua. *Ind Eng Chem Res*. 2005;44(26):9840-9852. doi:10.1021/ie058048e
55. Kumar H, Froment GF. Mechanistic Kinetic Modeling of the Hydrocracking of Complex Feedstocks, such as Vacuum Gas Oils. *Ind Eng Chem Res*. 2007;46(18):5881-5897. doi:10.1021/ie0704290
56. Guillaume D, Valéry E, Verstraete JJ, Surla K, Galtier P, Schweich D. Single Event Kinetic Modelling without Explicit Generation of Large Networks: Application to Hydrocracking of Long Paraffins. *Oil Gas Sci Technol – Rev d'IFP Energies Nouv*. 2011;66(3):399-422. doi:10.2516/ogst/2011118
57. de Oliveira LP, Verstraete JJ, Kolb M. Molecule-based kinetic modeling by Monte Carlo methods for heavy petroleum conversion. *Sci China Chem*.

2013;56(11):1608-1622. doi:10.1007/s11426-013-4989-3

58. Martens GG, Marin GB. Kinetics for hydrocracking based on structural classes: Model development and application. *AIChE J.* 2001;47(7):1607-1622. doi:10.1002/aic.690470713
59. Alvarez-Majmutov A, Chen J, Gieleciak R. Molecular-Level Modeling and Simulation of Vacuum Gas Oil Hydrocracking. *Energy and Fuels.* 2016;30(1):138-148. doi:10.1021/acs.energyfuels.5b02084
60. Evenepoel N, Agarwal P, Klein MT. Molecular-Level Kinetic Modeling of Lube Base Oil Hydroisomerization. *Energy & Fuels.* 2018;32(9):9804-9812. doi:10.1021/acs.energyfuels.8b02266
61. Lundanes E, Greibrokk T. Separation of fuels, heavy fractions, and crude oils into compound classes: A review. *J High Resolut Chromatogr.* 1994;17(4):197-202. doi:10.1002/jhrc.1240170403
62. Allen D, Grandy D, Jeong KM, Petrakls L. Heavier Fractions of Shale Oils, Heavy Crudes, Tar Sands, and Coal Liquids: Comparison of Structural Profiles. *Ind Eng Chem Process Des Dev.* 1985;24(3):737-742. doi:10.1021/i200030a036
63. Katzer JR, Sivasubramanian R. Process and Catalyst Needs for Hydrodenitrogenation. *Catal Rev.* 1979;20(2):155-208. doi:10.1080/01614947908062262
64. Snyder LR, Buell BE, Howard HE. Nitrogen and Oxygen Compound Types in Petroleum: Total Analysis of a 700–850 °F Distillate from a California Crude Oil. *Anal Chem.* 1968;40(8):1303-1317. doi:10.1021/ac60264a005
65. Martin RL, Grant JA. Determination of Sulfur-Compound Distributions in Petroleum Samples by Gas Chromatography with a Coulometric Detector. *Anal Chem.* 1965;37(6):644-649. doi:10.1021/ac60225a005
66. Girgis MJ, Gates BC. Reactivities, reaction networks, and kinetics in high-pressure catalytic hydroprocessing. *Ind Eng Chem Res.* 1991;30:2021-2058. doi:10.1021/ie00057a001
67. Bhide M. Quinoline hydrodenitrogenation kinetics and reaction inhibition. 1979.
68. Korre SC, Klein MT, Quann RJ. Hydrocracking of Polynuclear Aromatic Hydrocarbons. Development of Rate Laws through Inhibition Studies. *Ind Eng*

- Chem Res.* 1997;36(6):2041-2050. doi:10.1021/ie9606808
69. Russell CL, Klein MT, Quann RJ, Trewella J. Catalytic Hydrocracking Reaction Pathways, Kinetics, and Mechanisms of n-Alkylbenzenes. *Energy and Fuels.* 1994;8(6):1394-1400. doi:10.1021/ef00048a031
 70. Korre SC, Klein MT, Quann RJ. Polynuclear Aromatic Hydrocarbons Hydrogenation. 1. Experimental Reaction Pathways and Kinetics. *Ind Eng Chem Res.* 1995;34(1):101-117. doi:10.1021/ie00040a008
 71. Houalla M, Nag NK, Sapre A V., Broderick DH, Gates BC. Hydrodesulfurization of dibenzothiophene catalyzed by sulfided CoO-MoO₃-Al₂O₃: The reaction network. *AIChE J.* 1978;24(6):1015-1021. doi:10.1002/aic.690240611
 72. Furimsky E. Deactivation of hydroprocessing catalysts. *Catal Today.* 1999;52(4):381-495. doi:10.1016/S0920-5861(99)00096-6
 73. Vogelaar BM, Eijsbouts S, Bergwerff JA, Heiszwolf JJ. Hydroprocessing catalyst deactivation in commercial practice. *Catal Today.* 2010;154(3-4):256-263. doi:10.1016/j.cattod.2010.03.039
 74. Ali MF, Perzanowski H, Bukhari A, Al-Haji AA. Nickel and vanadyl porphyrins in Saudi Arabian crude oils. *Energy & Fuels.* 1993;7(2):179-184. doi:10.1021/ef00038a003
 75. Ware RA, Wei J. Catalytic hydrodemetallation of nickel porphyrins. I. Porphyrin structure and reactivity. *J Catal.* 1985;93(1):100-121. doi:10.1016/0021-9517(85)90155-1
 76. Trauth DM, Stark SM, Petti TF, Neurock M, Klein MT. Representation of the Molecular Structure of Petroleum Resid through Characterization and Monte Carlo Modeling. *Energy and Fuels.* 1994;8(3):576-580. doi:10.1021/ef00045a010
 77. Marrero J, Gani R. Group-contribution based estimation of pure component properties. *Fluid Phase Equilib.* 2001;183-184:183-208. doi:10.1016/S0378-3812(01)00431-9
 78. Benson SW, Cruickshank FR, Golden DM, et al. Additivity rules for the estimation of thermochemical properties. *Chem Rev.* 1969;69(3):279-324. doi:10.1021/cr60259a002
 79. Hindmarsh AC, Brown PN, Grant KE, et al. SUNDIALS : Suite of Nonlinear

- and Differential / Algebraic Equation Solvers. *Trans Math Softw.* 2005;31(3):363-396. doi:10.1145/1089014
80. Poling BE, Prausnitz JM, O'Connell JP. *The Properties of Gases and Liquids*. 5th ed. McGraw-Hill; 2000.
 81. Rachford HH, Rice JD. Procedure for Use of Electronic Digital Computers in Calculating Flash Vaporization Hydrocarbon Equilibrium. *J Pet Technol.* 1952;4(10):19-3. doi:10.2118/952327-G
 82. Lee BI, Kesler MG. A generalized thermodynamic correlation based on three-parameter corresponding states. *AIChE J.* 1975;21(3):510-527. doi:10.1002/aic.690210313
 83. Sander R. Compilation of Henry's law constants (version 4.0) for water as solvent. *Atmos Chem Phys.* 2015;15(8):4399-4981. doi:10.5194/acp-15-4399-2015
 84. Froment GF, Bischoff KB, De Wilde J. *Chemical Reactor Analysis and Design*. 3rd ed. Wiley; 2010.
 85. The Law That's Not A Law. *IEEE Spectr.* 2015;52(4):38-57. doi:10.1109/MSPEC.2015.7065416
 86. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. Vol 103. New York, NY: Springer New York; 2013. doi:10.1007/978-1-4614-7138-7
 87. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res.* 2011;12:2825-2830.
 88. Huang H, Fairweather M, Griffiths JF, Tomlin AS, Brad RB. A systematic lumping approach for the reduction of comprehensive kinetic models. *Proc Combust Inst.* 2005;30(1):1309-1316. doi:10.1016/j.proci.2004.08.001
 89. Frenklach M. Reduction of Chemical Reaction Models. In: *Numerical Approaches to Combustion Modeling*. Washington DC: American Institute of Aeronautics and Astronautics; 1991:129-154. doi:10.2514/5.9781600866081.0129.0154
 90. Pepiot-Desjardins P, Pitsch H. An automatic chemical lumping method for the reduction of large chemical kinetic mechanisms. *Combust Theory Model.* 2008;12(6):1089-1108. doi:10.1080/13647830802245177

91. Bhattacharjee B, Schwer DA, Barton PI, Green WH. Optimally-reduced kinetic models: reaction elimination in large-scale kinetic mechanisms. *Combust Flame*. 2003;135(3):191-208. doi:10.1016/S0010-2180(03)00159-7
92. Nigam A, Klein MT. A mechanism-oriented lumping strategy for heavy hydrocarbon pyrolysis: imposition of quantitative structure-reactivity relationships for pure components. *Ind Eng Chem Res*. 1993;32(7):1297-1303. doi:10.1021/ie00019a003
93. Fake DM, Nigam A, Klein MT. Mechanism based lumping of pyrolysis reactions: Lumping by reactive intermediates. *Appl Catal A Gen*. 1997;160(1):191-221. doi:10.1016/S0926-860X(97)00136-1
94. Agarwal P, Al-Khattaf SS, Klein MT. Molecular-Level Kinetic Modeling of Triglyceride Hydroprocessing. *Energy and Fuels*. 2019;Submitted.
95. Agarwal P, Sahasrabudhe M, Khandalkar S, Saravanan C, Klein MT. Molecular-Level Kinetic Modeling of a Real Vacuum Gas Oil Hydroprocessing Refinery System. *Energy and Fuels*. 2019;In prepara.
96. Horton SR, Hou Z, Moreno BM, Bennett CA, Klein MT. Molecule-based modeling of heavy oil. *Sci China Chem*. 2013;56(7):840-847. doi:10.1007/s11426-013-4895-8
97. Horton SR, Zhang Y, Mohr R, Petrocelli F, Klein MT. Implementation of a Molecular-Level Kinetic Model for Plasma-Arc Municipal Solid Waste Gasification. *Energy & Fuels*. 2016;30(10):7904-7915. doi:10.1021/acs.energyfuels.6b00899