How a Systematic Approach to Uncertainty Quantification Rends Molecular Simulation a Quantitative Tool in Predicting the Critical Constants for Large $n$-Alkanes

Richard Alma Messerly
Brigham Young University

Follow this and additional works at: https://scholarsarchive.byu.edu/etd
Part of the Chemical Engineering Commons

BYU ScholarsArchive Citation
Messerly, Richard Alma, "How a Systematic Approach to Uncertainty Quantification Renders Molecular Simulation a Quantitative Tool in Predicting the Critical Constants for Large n-Alkanes" (2016). All Theses and Dissertations. 6598.
https://scholarsarchive.byu.edu/etd/6598

This Dissertation is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
How a Systematic Approach to Uncertainty Quantification Renders Molecular Simulation a Quantitative Tool in Predicting the Critical Constants for Large \( n \)-Alkanes

Richard Alma Messerly

A dissertation submitted to the faculty of Brigham Young University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

W. Vincent Wilding, Chair
Thomas A. Knotts IV
Dean R. Wheeler
Larry L. Baxter
David O. Lignell

Department of Chemical Engineering
Brigham Young University
2016

Copyright © 2016 Richard Alma Messerly
All Rights Reserved
ABSTRACT

How a Systematic Approach to Uncertainty Quantification Renders Molecular Simulation a Quantitative Tool in Predicting the Critical Constants for Large $n$-Alkanes

Richard Alma Messerly
Department of Chemical Engineering, BYU
Doctor of Philosophy

Accurate thermophysical property data are crucial for designing efficient chemical processes. For this reason, the Design Institute for Physical Properties (DIPPR 801) provides evaluated experimental data and prediction of various thermophysical properties. The critical temperature ($T_c$), critical density ($\rho_c$), critical pressure ($P_c$), critical compressibility factor ($Z_c$), and normal boiling point ($T_b$) are important constants to check for thermodynamic consistency and to estimate other properties. The $n$-alkane family is of primary interest because it is generally assumed that other families of compounds behave similarly to the $n$-alkane family with increasing chain-length. Unfortunately, due to thermal decomposition, experimental measurements of $T_c$, $\rho_c$, and $P_c$ for large $n$-alkanes are scarce and potentially unreliable. For this reason, molecular simulation is an attractive alternative for estimating the critical constants. However, molecular simulation has often been viewed as a tool that is limited to providing qualitative insight. One key reason for this perceived weakness is the difficulty in quantifying the uncertainty of the simulation results. This research focuses on a systematic top-down approach to quantifying the uncertainty in Gibbs Ensemble Monte Carlo (GEMC) simulations for large $n$-alkanes.

We implemented four different methods in order to obtain quantitatively reliable molecular simulation results. First, we followed a rigorous statistical analysis to assign the uncertainty of the critical constants when obtained from GEMC. Second, we developed an improved method for predicting $P_c$ with the standard force field models in the literature. Third, we implemented an experimental design to reduce the uncertainty associated with $T_c$, $\rho_c$, $P_c$, and $Z_c$. Finally, we quantified the uncertainty associated with the Lennard-Jones 12-6 potential parameters.

This research demonstrates how uncertainty quantification renders molecular simulation a quantitative tool for thermophysical property evaluation. Specifically, by quantifying and reducing the uncertainty associated with molecular simulation results, we were able to discern between different experimental data sets and prediction models for the critical constants. In this regard, our results enabled the development of improved prediction models for $T_c$, $\rho_c$, $P_c$, and $Z_c$ for large $n$-alkanes. In addition, we developed a new $T_b$ prediction model in order to ensure thermodynamic consistency between $T_c$, $P_c$, and $T_b$.

Keywords: GEMC, force field models, monte carlo, nonlinear statistics, experimental design, propagation of errors, thermodynamic properties
I would like to acknowledge some of those that have made my PhD experience at Brigham Young University both successful and enjoyable. First and foremost, I need to thank my wife, Erica Lynne, for being my inspiration and constant companion through the highs and lows that inevitably accompany a graduate degree. Furthermore, my children, McKay Richard and Macey Lynne, gave meaning to the sacrifices required to attain a PhD. The close relationships with family and friends were key to my personal and emotional well-being. Most importantly, my faith in the Lord Jesus Christ was the rock that kept my foundation sure through the trials of life that my family and I have faced over the past five years.

I am deeply grateful that Brigham Young University faculty understand the importance of maintaining a balance between things of temporal and eternal significance. I believe my chair, Dr. W. Vincent Wilding, has been an excellent role model and example of what it means to be an academic disciple of Christ. I greatly appreciate all the guidance that I received from him both as an undergraduate and graduate student. I am also very appreciative of Dr. Thomas A. Knotts IV treating me as one of his own PhD students with the amount of time he dedicated to my research. His experience with molecular simulations was essential for my success. I was truly blessed to be assigned such a strong committee consisting of Dr. Dean R. Wheeler, Dr. Larry L. Baxter, and Dr. David O. Lignell. Each committee member brought a unique and invaluable set of skills that greatly strengthened my research. Despite not being an official committee member, Dr. Neil F. Giles made a considerable contribution to the direction of my research. As the project coordinator of the Design Institute for Physical Properties (DIPPR 801), he helped integrate my research into the goals of the DIPPR 801 project. Finally, I would be remiss if I did not acknowledge all of the expertise that Dr. Richard L. Rowley provided me. Although he was an emeritus faculty member prior to me starting my research he was an integral contributor to my project. In addition, after he was removed from my committee he continued to frequently offer his advise.
# TABLE OF CONTENTS

**LIST OF TABLES** ................................................................. viii

**LIST OF FIGURES** .............................................................. ix

**NOMENCLATURE** .................................................................. xii

**Chapter 1 Introduction** ....................................................... 1
  1.1 Background ................................................................. 1
  1.2 Objective ................................................................. 6
  1.3 Outline ......................................................................... 7

**Chapter 2 Molecular Simulation Fundamentals** ....................... 8
  2.1 Molecular Simulation Methods ........................................ 8
    2.1.1 Monte Carlo ......................................................... 8
    2.1.2 Molecular Dynamics .............................................. 9
    2.1.3 Method Selection ................................................ 10
    2.1.4 Configurational-bias ............................................ 12
    2.1.5 Simulation Package .............................................. 13
    2.1.6 Simulation Protocol .............................................. 13
  2.2 Molecular Models ......................................................... 14
    2.2.1 Comparison of Force Field Models ......................... 15
    2.2.2 Implementation in Simulation ................................. 20
    2.2.3 Model Development .............................................. 21
  2.3 Layers of Uncertainty .................................................... 23
    2.3.1 Numerical Uncertainty .......................................... 24
    2.3.2 Model Uncertainty .............................................. 25
    2.3.3 Parameter Uncertainty ......................................... 26
  2.4 Uncertainty Quantification and Propagation of Errors .......... 26
    2.4.1 Frequentist Theory .............................................. 26
    2.4.2 Uncertainty Quantification .................................... 28
    2.4.3 Propagation of Errors .......................................... 29

**Chapter 3 Numerical Uncertainty in $T_c$ and $\rho_c$ with GEMC** .... 30
  3.1 Introduction ............................................................. 30
  3.2 Determining the Critical Point through Regression ............... 31
  3.3 Determining the Uncertainties in $T_c$ and $\rho_c$ ................. 32
    3.3.1 Estimating Standard Deviations ............................. 32
    3.3.2 Blocking and Replicate Methods ............................ 33
    3.3.3 Indirect and Direct Methods ................................. 33
    3.3.4 Rigorous Nonlinear Analysis ................................. 36
  3.4 Results and Importance of Rigorous Statistical Analysis ......... 41
  3.5 Case Study: Long-Chain-Length Behavior of $n$-alkanes ........... 42
Chapter 4 Improved Estimates of $T_c$, $\rho_c$, $P_c$ and $Z_c$ with GEMC

4.1 Introduction ........................................... 46
4.2 Alternative Method to Predict $P_c$ ........................... 48
4.3 Estimating Uncertainty in $P_c$ and $Z_c$ .............. 51
4.4 Case Study: Long-Chain-Length Behavior of $n$-Alkanes .... 51
4.5 Optimal Experimental Design to Minimize the Uncertainty in $T_c$, $\rho_c$, and $P_c$ ......... 55
  4.5.1 General Theory of Experimental Designs ............. 55
  4.5.2 GEMC Experimental Design .......................... 60
4.6 Comparison of Experimental Design with Literature ...... 66
4.7 Considerations ............................................ 69
  4.7.1 Possible Modifications to Experimental Designs ....... 69
  4.7.2 Implications on Finite-Size Effects ................... 71
  4.7.3 Possible Limitations .................................. 71
4.8 Outline of Methodology ................................ 74
4.9 Simulation Specifications ................................. 74
4.10 Validation of D-optimal Design ......................... 77
4.11 Results for $C_{24}$, $C_{36}$, and $C_{48}$ ................... 78
  4.11.1 $T_c$ and $\rho_c$ Results for Larger $n$-Alkanes ......... 79
  4.11.2 $P_c$ Results for Larger $n$-Alkanes .................. 84
  4.11.3 $Z_c$ Results .................................... 86
4.12 Validation of Results for Large $n$-Alkanes .............. 88
4.13 Conclusions ............................................ 90

Chapter 5 Uncertainty Quantification and Propagation of Errors for the Lennard-Jones 12-6 Parameters

5.1 Introduction ........................................... 92
5.2 Theory of UQ+PoE ..................................... 93
  5.2.1 Type A UQ+PoE .................................. 94
  5.2.2 Type B UQ+PoE .................................. 96
5.3 UQ+PoE for CH$_4$ ..................................... 98
  5.3.1 UQ for CH$_4$ .................................. 100
  5.3.2 MCS for CH$_4$ .................................. 101
  5.3.3 PoE for CH$_4$ .................................. 102
5.4 Methods for UQ+PoE of CH$_3$ and CH$_2$ ................. 107
  5.4.1 Simulation Specifications ......................... 110
  5.4.2 Simulation Data Analysis ......................... 111
  5.4.3 Reducing Numerical Uncertainty .............. 111
5.5 Results for CH$_3$ and CH$_2$ ............................ 113
  5.5.1 UQ for CH$_3$ and CH$_2$ ................ 114
  5.5.2 MCS for CH$_3$ and CH$_2$ ................ 116
  5.5.3 PoE for CH$_3$ and CH$_2$ ................ 120
5.6 Limitations ............................................ 128
5.7 Surrogate Model ...................................... 130
## Chapter 6  New Prediction Models  

6.1 Introduction  

6.2 Theoretical Basis  
   6.2.1 Infinite Chain-Length Behavior  
   6.2.2 Internal Consistency  

6.3 Methodology  

6.4 Model Development  
   6.4.1 Experimental Data Evaluation  
   6.4.2 Model Form  
   6.4.3 Enforcing Constraints  
   6.4.4 Insight from Regression Parameters  
   6.4.5 Proposed Prediction Models  
   6.4.6 Uncertainty in Prediction Models  

6.5 Results  
   6.5.1 $P_c$, $T_c$, $\rho_c$, and $Z_c$  
   6.5.2 $\rho_l$  
   6.5.3 $T_b$  

6.6 Model Performance  

6.7 Limitations  

6.8 Conclusions  

## Chapter 7  Conclusions and Recommendations  

7.1 Conclusions  

7.2 Recommendations  
   7.2.1 Implementing Prediction Models in DIPPR 801  
   7.2.2 Predicting $\rho_l$ and $P_v$  
   7.2.3 Investigating Other Homologous Series  

7.3 Future Work  
   7.3.1 Simulating Larger $n$-Alkanes  
   7.3.2 Additional Layers of Uncertainty  
   7.3.3 Extend Mess-UP Force Field  

REFERENCES  

Appendix A  Values Used in Chapter 3  

Appendix B  Vapor-Liquid Simulation Data in Chapter 4  

Appendix C  Simulation Results in Chapter 4
<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Intermolecular potential parameters for force field models.</td>
<td>19</td>
</tr>
<tr>
<td>2.2</td>
<td>Exp-6 intermolecular potential parameters as required by Towhee force field files.</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Error model for GEMC</td>
<td>35</td>
</tr>
<tr>
<td>3.2</td>
<td>Comparison between the traditional and rigorous uncertainty regions for $T_c$ and $\rho_c$.</td>
<td>42</td>
</tr>
<tr>
<td>4.1</td>
<td>Summary of simulations performed in Chapter 4.</td>
<td>76</td>
</tr>
<tr>
<td>5.1</td>
<td>Comparison of uncertainties in critical constants for methane.</td>
<td>105</td>
</tr>
<tr>
<td>5.2</td>
<td>Comparison of root-mean-square for the Mess-UP and TraPPE models.</td>
<td>113</td>
</tr>
<tr>
<td>5.3</td>
<td>Extrema parameter sets used in Type B analysis.</td>
<td>120</td>
</tr>
<tr>
<td>6.1</td>
<td>Proposed parameters for $T_c$, $\rho_c$, $P_c$, and $T_b$ prediction models.</td>
<td>149</td>
</tr>
<tr>
<td>A.1</td>
<td>Summary of the data used in the analysis corresponding to Figure 3.5 Panel (a).</td>
<td>182</td>
</tr>
<tr>
<td>A.2</td>
<td>Summary of the data used in the analysis corresponding to Figure 3.5 Panel (b).</td>
<td>183</td>
</tr>
<tr>
<td>A.3</td>
<td>Summary of the data used in the analysis corresponding to Table 3.2.</td>
<td>184</td>
</tr>
<tr>
<td>B.1</td>
<td>Vapor-liquid coexistence data for smaller $n$-alkanes simulated in Chapter 4</td>
<td>185</td>
</tr>
<tr>
<td>B.3</td>
<td>Vapor-liquid coexistence data for larger $n$-alkanes simulated in Chapter 4</td>
<td>187</td>
</tr>
<tr>
<td>C.1</td>
<td>$T_c$, $\rho_c$, $P_c$, and $Z_c$ results for $n$-alkanes simulated in Chapter 4</td>
<td>194</td>
</tr>
<tr>
<td>D.1</td>
<td>Experimental data used to develop Mess-UP force field.</td>
<td>195</td>
</tr>
<tr>
<td>E.1</td>
<td>Parameter sets obtained with Monte Carlo Sampling.</td>
<td>196</td>
</tr>
<tr>
<td>H.1</td>
<td>Recommended $T_c$, $\rho_c$, $P_c$, and $Z_c$ values for $n$-alkanes simulated in this study.</td>
<td>224</td>
</tr>
<tr>
<td>I.1</td>
<td>Summary of simulations performed in Appendix I.</td>
<td>225</td>
</tr>
</tbody>
</table>
1.1 Comparison of simulation literature, experimental data, and prediction models for $T_c$, $\rho_c$, $P_c$, and $Z_c$ with respect to $CN$. .................................................. 5

2.1 Comparison of Lennard-Jones 12-6, Mie 16-6, and Buckingham Exponential-6. ... 17
2.2 Comparison of TraPPE, NERD, Mess-UP, Mie 16-6, and Exp-6 force fields. ....... 18

3.1 Comparison between the traditional and rigorous estimates of $s_r$ and $s_s$. ......... 35
3.2 Error model for $S_{D_r}$ and $S_{D_s}$. .......................................................... 36
3.3 Example of histograms produced from Monte Carlo Sampling propagation of errors for $T_c$ and $\rho_c$. .......................................................... 37
3.4 Example of joint confidence regions for $A$, $B$, $T_c$, and $\rho_c$. ....................... 38
3.5 Comparison between traditional and rigorous joint confidence regions for $T_c$ and $\rho_c$. 40
3.6 Impact of rigorous statistical analysis on literature results for $T_c$ and $V_c$. ......... 43

4.1 Example of histograms produced from Monte Carlo Sampling propagation of errors for $P_c$ and $Z_c$. .......................................................... 52
4.2 Comparison between experimental data, prediction models, and simulation results for $P_c$ with respect to $CN$. .................................................. 53
4.3 Contour plot of the determinant of the information matrix for the straight line model. 58
4.4 Contour plot of the inverse of the variance in the intercept of a straight-line model. 59
4.5 Contour plot of the inverse of the variance in the slope of a straight-line model. .... 60
4.6 Contour plot of the determinant for the information matrix with constant variance. ... 63
4.7 Contour plot of the determinant for the weighted information matrix. ................. 65
4.8 Contour plot of the inverse variance in $\rho_c$. .......................................... 65
4.9 Contour plot of the inverse variance in $T_c$. .......................................... 66
4.10 Comparison between experimental designs in this work and the literature. ........... 67
4.11 Dependence of the experimental designs on the lower temperature constraint. ....... 70
4.12 Effect of the lower temperature constraint on the uncertainty in $T_c$, $\rho_c$, and $P_c$. ... 70
4.13 Sensitivity of the experimental design performance. ........................................ 73
4.14 Validation of D-optimal design by comparing $T_c$ for smaller $CN$. ................. 78
4.15 Validation of D-optimal design by comparing $\rho_c$ for smaller $CN$. ................. 79
4.16 Validation of D-optimal design by comparing $P_c$ for smaller $CN$. ................. 80
4.17 Comparison of uncertainty regions of $T_c$ and $\rho_c$ for C$_{24}$, C$_{36}$, and C$_{48}$. ... 81
4.18 $T_c$ results for larger $n$-alkanes with respect to $CN$. .................................. 82
4.19 $P_c$ results for larger $n$-alkanes with respect to $CN$. .................................. 85
4.20 $Z_c$ results for larger $n$-alkanes with respect to $CN$. .................................. 88
4.21 Comparison of the vapor-liquid coexistence curves for large $n$-alkanes. ............. 89
4.22 Validation of simulation results for large $n$-alkanes. .................................... 90

5.1 Comparison of uncertainty regions for $\varepsilon_{CH_4}$-$\sigma_{CH_4}$. ....................... 101
5.2 Histograms of $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ from MCS Type A analysis. ............. 103
5.3 Comparison of $\rho_l$ and $B_2$ uncertainties for methane. ............................... 104
5.4 Comparison of uncertainties in $T_c$, $\rho_c$, and $P_c$ for CH$_4$. ..................... 106
5.5 Comparison of uncertainty regions for $\varepsilon_{CH_5}$-$\sigma_{CH_5}$ and $\varepsilon_{CH_2}$-$\sigma_{CH_2}$. 115
F.17 Numerical uncertainties in $T$
F.15 Numerical uncertainties in $T$
F.12 Histograms of $T$
F.11 Histograms of $T$
F.8 Histograms of $\rho$
F.4 Histograms of $\rho$
F.3 Histograms of $\rho$
F.2 Histograms of $\rho$
F.1 Histograms of $\rho$
5.6 Histograms of $\varepsilon_{CH_3}$, $\sigma_{CH_3}$, $\varepsilon_{CH_2}$, and $\sigma_{CH_2}$ from MCS Type A and A$_B$ analysis.
5.7 PDF contours and MCS parameter sets for $\varepsilon_{CH_3}$-$\sigma_{CH_3}$ from Type A analysis.
5.8 PDF contours and MCS parameter sets for $\varepsilon_{CH_2}$-$\sigma_{CH_2}$ and $\varepsilon_{CH_3}$-$\sigma_{CH_3}$.
5.9 Comparison of $T_c$ uncertainties for ethane.
5.10 Comparison of the $\rho_c$, $P_c$, and $Z_c$ histograms obtained using 100 and 2000 MCS parameter sets for ethane.
5.11 Comparison of $\rho_l$ simulation results for ethane and n-octane.
5.12 Comparison of the $T_c$ probability densities for ethane and n-octane.
5.13 Comparison of $\rho_l$ simulation results for larger n-alkanes.
5.14 Comparison of experimental data, prediction models, and simulation results with the Mess-UP force field for $T_c$, $\rho_c$, $P_c$, and $Z_c$ with respect to CN.
5.15 Contours of $RMS$ with respect to $\varepsilon$ and $\sigma$.
5.16 Contours of $\rho_l$ for ethane with respect to $\varepsilon_{CH_3}$ and $\sigma_{CH_3}$ (coarse grid).
5.17 Contours of $\rho_l$ for ethane with respect to $\varepsilon_{CH_3}$ and $\sigma_{CH_3}$ (refined grid).
5.18 Contours of $\rho_l$ for n-octane with respect to $\varepsilon_{CH_2}$ and $\sigma_{CH_2}$ (coarse grid).
5.19 Contours of $\rho_l$ for n-octane with respect to $\varepsilon_{CH_2}$ and $\sigma_{CH_2}$ (refined grid).
6.1 Contours of acceptable combinations of $P_c^0$, $T_c^\infty$, and $\rho_c^0$.
6.2 Performance of new $P_c$ prediction model.
6.3 Performance of new $T_c$ prediction model.
6.4 Performance of new $\rho_c$ prediction model.
6.5 Performance of new $Z_c$ prediction model.
6.6 Performance of $T_c$, $\rho_c$, and $P_c$ models for predicting $\rho_l$ with Rackett equation.
6.7 Performance of new $T_b$ prediction model.
6.8 Residual plots for $T_c$, $\rho_c$, $P_c$, and $T_b$.
6.9 Comparison of deviations for $T_c$, $\rho_c$, $P_c$, $Z_c$, and $T_b$.
6.10 Relative uncertainties for $T_c$, $\rho_c$, $P_c$, $Z_c$, and $T_b$.
F.18 Numerical uncertainties in $T_c$, $\rho_c$, $P_c$, and $Z_c$ for Type B analysis of $C_{48}$.
F.19 Effect of CH$_3$ uncertainties on $T_c$, $\rho_c$, $P_c$, and $Z_c$ of $C_{16}$.
F.20 Extrema parameter sets for CH$_3$ and CH$_2$.

G.1 Effect of CH$_2$-CH$_3$ correlation on MCS parameter sets.
G.2 Histograms for correlated and uncorrelated CH$_2$ MCS parameters.

I.1 Vapor-liquid coexistence curve of large $n$-alkanes using TEAM-CG force field.
I.2 Comparison of TEAM-CG simulation results, experimental data, and new prediction models for $T_c$, $\rho_c$, $P_c$, and $Z_c$ with respect to CN.
I.3 Comparison of TEAM-CG and Mess-UP simulation results for $\rho_l$ of $C_{24}$, $C_{36}$, $C_{48}$. 
NOMENCLATURE

A Fitting parameter in law of rectilinear diameters
AA All-atom
\( b_i \) \( i^{th} \) regression parameter
B Fitting parameter in density scaling law
\( B_2 \) Second virial coefficient
CG Coarse-grained
\( CH_i \) United-atom site with one carbon atom and \( i \) hydrogen atoms
\( C_i \) \( n \)-Alkane with \( i \) carbon atoms
CN Carbon number
\( COV \) Covariance matrix
d Order of Nikitin model
EoS Equation of State
Exp-6 Exponential-6 force field developed by Errington et al.
\( F \) F-matrix
\( F_{p,v,\alpha} \) F-statistic with \( p \) and \( v \) degrees of freedom at the \( \alpha \) confidence level
\( F_{p,v,\alpha}^{-1} \) Inverse F-statistic with \( p \) and \( v \) degrees of freedom at the \( \alpha \) confidence level
\( F^T F \) Information Matrix
\( F^T W^T W F \) Weighted Information Matrix
GCMC Grand Canonical Monte Carlo
GEMC Gibbs Ensemble Monte Carlo
GOMC GPU Optimized Monte Carlo
\( k_B \) Boltzmann constant
LB Lorentz-Berthelot combining rules
LJ Lennard-Jones
MC Monte Carlo
MCS Monte Carlo Sampling
MD Molecular Dynamics
Mess-UP Force field developed by Messerly et al. for uncertainty propagation
Mie 16-6 Mie \( \lambda \)-6 force field developed by Potoff et al.
\( M_w \) Molecular weight
\( n \) Number of samples/replicates
\( N \) Number of molecules
\( N \) Number of data points
NERD Nath, Escobedo, and de Pablo revised force field
\( NVT \) Ensemble with constant number of molecules, volume, and temperature
\( p \) Number of parameters
\( P \) Pressure
\( P_0 \) Zeroth term in Nikitin \( P_c \) model
\( P_c \) Critical pressure
\( P_c^* \) Lennard-Jones reduced critical pressure
\( PDF \) Probability density function
\( P_i \) \( i^{th} \) Type B extrema parameter set
PoE Propagation of Errors
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_v$</td>
<td>Vapor pressure</td>
</tr>
<tr>
<td>QSPR</td>
<td>Quantitative-Structure-Property-Relationship</td>
</tr>
<tr>
<td>$r$</td>
<td>Intermolecular distance</td>
</tr>
<tr>
<td>$r^*$</td>
<td>Reduced intermolecular distance ($\frac{r}{\sigma}$)</td>
</tr>
<tr>
<td>$R_g$</td>
<td>Universal gas constant</td>
</tr>
<tr>
<td>RMS</td>
<td>Root-mean-square</td>
</tr>
<tr>
<td>$s$</td>
<td>Sample standard deviation</td>
</tr>
<tr>
<td>$s^2$</td>
<td>Sample variance</td>
</tr>
<tr>
<td>$S$</td>
<td>Sum squared error</td>
</tr>
<tr>
<td>SAFT</td>
<td>Statistical Associating Fluids Theory</td>
</tr>
<tr>
<td>$S_D$</td>
<td>Population standard deviation</td>
</tr>
<tr>
<td>$S_D^2$</td>
<td>Population variance</td>
</tr>
<tr>
<td>$S_E$</td>
<td>Standard error of the mean</td>
</tr>
<tr>
<td>$T$</td>
<td>Temperature</td>
</tr>
<tr>
<td>$T_b$</td>
<td>Normal boiling point</td>
</tr>
<tr>
<td>$T_c$</td>
<td>Critical temperature</td>
</tr>
<tr>
<td>$T_c^\infty$</td>
<td>$T_c$ value for the infinite chain-length $n$-alkane</td>
</tr>
<tr>
<td>$T_c^*$</td>
<td>Lennard-Jones reduced critical temperature</td>
</tr>
<tr>
<td>$T_r$</td>
<td>Reduced temperature ($\frac{T}{T_c}$)</td>
</tr>
<tr>
<td>TraPPE</td>
<td>Transferable Potentials for Phase Equilibria</td>
</tr>
<tr>
<td>$u_{nb}$</td>
<td>Non-bonded (intermolecular) potential energy</td>
</tr>
<tr>
<td>UA</td>
<td>United-atom</td>
</tr>
<tr>
<td>UQ</td>
<td>Uncertainty Quantification</td>
</tr>
<tr>
<td>$V$</td>
<td>Volume</td>
</tr>
<tr>
<td>$V_c$</td>
<td>Critical volume</td>
</tr>
<tr>
<td>VLCC</td>
<td>Vapor-liquid coexistence curve</td>
</tr>
<tr>
<td>VLE</td>
<td>Vapor-liquid equilibrium</td>
</tr>
<tr>
<td>$W$</td>
<td>Weighting matrix</td>
</tr>
<tr>
<td>$WF$</td>
<td>Weighted F-matrix</td>
</tr>
<tr>
<td>$x$</td>
<td>Independent variable</td>
</tr>
<tr>
<td>$y$</td>
<td>Dependent variable</td>
</tr>
<tr>
<td>$Z_c$</td>
<td>Critical compressibility factor</td>
</tr>
<tr>
<td>$Z_c^\infty$</td>
<td>$Z_c$ value for the infinite chain-length $n$-alkane</td>
</tr>
</tbody>
</table>

**Greek symbols**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>Confidence level</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Critical universality exponent in density scaling law</td>
</tr>
<tr>
<td>$\delta$</td>
<td>Exponent in Nikitin model specific to property</td>
</tr>
<tr>
<td>$\varepsilon$</td>
<td>Intermolecular potential attraction parameter</td>
</tr>
<tr>
<td>$\varepsilon^*$</td>
<td>Reduced non-bonded potential energy ($\frac{u_{nb}}{\varepsilon}$)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Rackett exponent</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>Repulsive exponent in Mie $\lambda$-6 potential</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Chemical potential</td>
</tr>
<tr>
<td>$\nu$</td>
<td>Degrees of freedom</td>
</tr>
</tbody>
</table>
ω  Buckingham exponential-6 parameter
ρ_0^c  Zeroth term in Nikitin ρ_c model
ρ_c  Critical density
ρ_c^*  Lennard-Jones reduced critical density
ρ_l  Saturated liquid density
ρ_r  Rectilinear density
ρ_s  Scaling density
ρ_v  Saturated vapor density
σ  Intermolecular potential size parameter
θ  Set of parameters
θ̂  Optimal set of parameters

Subscripts, superscripts, and other indicators

\hat{\text{[ ]}}  indicates predicted value of [ ]
[ ](x)  indicates [ ] is a function of x
[ ]_c  indicates [ ] is a critical property
[ ]_i  indicates [ ] is an experimental data point
[ ]_l  indicates [ ] is a liquid phase property
[ ]_v  indicates [ ] is a vapor phase property
[ ]_∞  indicates [ ] is for the infinite chain-length n-alkane
[ ]^*  indicates [ ] is a dimensionless (reduced) Lennard-Jones property
CHAPTER 1. INTRODUCTION

1.1 Background

The critical point of a compound is defined by the temperature above which it is no longer possible to distinguish between liquid and vapor phases. The critical temperature ($T_c$), critical pressure ($P_c$), critical density ($\rho_c$), and critical compressibility factor ($Z_c$) are important thermophysical properties in science and engineering as they are required to determine equation of state parameters for PVT calculations and to predict other properties based upon the corresponding states principle [1] and other prediction methods. For example, these critical constants are needed when predicting properties such as saturated liquid density ($\rho_l$) [2], liquid heat capacity [3], second virial coefficient [4], liquid thermal conductivity [5], liquid vapor pressure ($P_v$) [6], and liquid viscosity [7]. These properties are involved in a large portion of the calculations done in science and engineering, so the accuracy of the critical constants is of high importance.

There exist three different methods for estimating the critical constants: experimentation, prediction models, and molecular simulations. Experimental measurement of critical constants is possible and reliable data exist for a large number of compounds; however, no experimental data exist for many compounds of interest for a number of reasons. For example, obtaining reliable critical constant values becomes difficult for larger compounds because they thermally decompose at temperatures well below $T_c$. In other cases, the chemical toxicity or reactivity is such that experiments are inadvisable. Situations such as these motivate alternative methods for estimating the critical constants.

Due to the importance of the critical constants, many models have been developed to predict values when experimental data are not available. These include group contribution methods like the Joback approach [8–10], Quantitative-Structure Property-Relationships (QSPR) methods like that of Kazakov et al. [11], and mathematical models for a specific homologous series [12–16]. Some of these prediction models have a strong theoretical basis, which should provide more confidence
when extrapolating to larger compounds not used in the training set; however, even the theoretical based mathematical models developed specifically for \( n \)-alkanes show strong disagreement in their extrapolation behavior.

As reliable estimates of thermophysical properties are essential for engineering design purposes, the Design Institute for Physical Properties database (DIPPR 801) requires accurate predictions of \( T_c \), \( P_c \), \( \rho_c \), and \( Z_c \) [17]. The \( n \)-alkane family is of particular interest due to its significance in all chemical industries. Also, the \( n \)-alkane family provides insight into other families of compounds because it is traditionally assumed that the critical constants for other homologous series converge to the \( n \)-alkane family with increasing chain-length [18, 19]. For these reasons, many researchers have attempted to predict or develop sophisticated experimental approaches to measure the critical constants for large \( n \)-alkanes that thermally decompose (i.e. ranging between C\(_{12}\)-C\(_{60}\)).

For example, the Nikitin group has developed a pulse-heating technique which has permitted experimental measurements of \( T_c \) and \( P_c \) for \( n \)-alkanes as large as C\(_{60}\) [15]. However, the \( P_c \) values reported by Nikitin et al. for C\(_{36}\) differ significantly between their original and more recent publications [20]. This discrepancy is troubling since both the original and more recent values were obtained from the same experimental data and differ only in the data analysis method. In addition, experimental \( P_c \) data measured by Teja for \( n \)-alkanes up to C\(_{18}\) show a strong discontinuity compared with Nikitin’s data for larger compounds [12]. Disagreement between the \( T_c \) data from Teja et al. and Nikitin et al. is also present, although it is more subtle than for \( P_c \).

To further complicate matters, some mathematical models for predicting \( T_c \) for large \( n \)-alkanes are more consistent with Nikitin’s original \( T_c \) values than with the most recent values. For example, Tsonopoulos’ model [13] for \( T_c \) agrees strongly with Nikitin’s original \( T_c \) values whereas Teja’s \( T_c \) model [12] agrees more closely with Nikitin’s most recent data. The situation is even more troublesome for \( P_c \) as some models (e.g. Tsonopoulos, Teja, Lemmon, Hukkerikar [10, 12–14]) propose a limiting, non-zero asymptote for \( P_c \) while others (e.g. Nannoolal, Nikitin [9,15,20]) predict a trend with \( P_c \) converging to zero for the infinite chain-length. These different models for \( P_c \) also result in very different trends for \( Z_c \). Unfortunately, experimental \( Z_c \) data are only available for \( n \)-alkanes as large as C\(_{18}\) because \( \rho_c \) values have not been reported for larger compounds. Furthermore, the \( Z_c \) data have large uncertainties and follow a peculiar trend with increasing chain-length. Therefore, experimental data alone are not sufficient to discern between different \( Z_c \) models. Due
to the experimental limitations and the dubious extrapolation of simple prediction models, molecular simulation of vapor-liquid equilibrium (VLE) has emerged as a common and often essential tool to examine the critical constants of large compounds that thermally decompose.

Multiple simulation methods exist to calculate critical constants. Historically, those obtained from Grand Canonical Monte Carlo (GCMC) finite-size scaling methods were considered superior to those obtained from Gibbs Ensemble Monte Carlo (GEMC) due to two key limitations found in the latter. First, the GEMC approach suffers from large fluctuations near the critical point. For this reason, it is necessary to perform GEMC simulations at temperatures below \( T_c \) and then extrapolate to the critical point. The second limitation is that there is no rigorous approach to correct for finite-size effects with GEMC [21]. However, it was recently shown that finite-size effects for \( T_c, \rho_c, \) and \( P_c \) are smaller than the corresponding statistical uncertainty for GEMC systems with 200 molecules of \( n \)-decane [22]. Therefore, GEMC presents an attractive alternative to GCMC when the computational demand of finite-size scaling methods is not justified. In Section 2.1 we provide a more thorough comparison of the different simulation methods along with a justification for our selection of the Gibbs Ensemble Monte Carlo approach.

Molecular simulation has the potential to help resolve the disagreements between different experimental data and prediction models. However, the molecular simulation literature contains a plethora of conflicting results for the critical constants of large \( n \)-alkanes. In Figure 1.1 we present the most reliable simulation results (i.e. accurate molecular models with widely accepted methods) for large \( n \)-alkanes found in the literature prior to this research. Specifically, we have included simulation results from Errington et al. [23], Potoff et al. [24], Nath et al. [25], Muller et al. [26], Martin et al. [27], and Zhuravlev et al. [28] Since the results from Martin et al. and Zhuravlev et al. were obtained by the same group using the same simulation code and molecular model they are referred to collectively as “Siepmann” for simplicity. In general, each study used a different force field (i.e. molecular model, see Chapter 2). Specifically, Errington et al., Potoff et al., Nath et al., and Siepmann’s group used the Exp-6, Mie 16-6, NERD, and TraPPE-UA force fields, respectively while Muller et al. used both the NERD and TraPPE-UA force fields. In addition, these studies also implemented different simulation methodologies. To be specific, Errington et al. and Potoff et al. used GCMC while Nath et al. and Siepmann’s group utilized GEMC and Muller et al. employed molecular dynamics (MD). The error bars included in Figure 1.1 represent the 95%
confidence interval as reported by the authors. In the case where uncertainties are not reported, the 95% confidence intervals are obtained from the traditional data analysis method (see Chapter 3). We compare these simulation results with the aforementioned experimental data from Teja et al. and both studies by Nikitin et al. [12, 15, 20] Only a single representative error bar is included for the experimental data. We also include several of the previously mentioned prediction models that diverge from each other with increasing carbon number [9, 12, 13, 15, 29].

There are several key observations from Figure 1.1. First, each force field appears to agree well with $T_c$ and $\rho_c$ for compounds smaller than $C_{18}$ while there exists a wide range of reported values for larger compounds. For example, in Panel (a) we see that the results for the NERD force field from Nath et al. and Muller et al. support Tsonopoulos’ $T_c$ model and Nikitin’s original $T_c$ data while the results from the Exp-6 force field reported by Errington et al. and from the TraPPE-UA force field reported by Siepmann’s group and Muller et al. support Teja’s $T_c$ model and Nikitin’s more recent $T_c$ data. In Panel (c) we see that the simulation literature are inconclusive whether Teja’s or Nikitin’s $\rho_c$ prediction model is more reliable while it is clear that Tsonopoulos’ $\rho_c$ model is less reliable for large $n$-alkanes. Specifically, Muller’s results for both the NERD and TraPPE-UA force fields support Teja’s model while the results from Errington et al. and Nath et al. agree more so with Nikitin’s $\rho_c$ model (although the uncertainties are quite large). Furthermore, we see in Panels (b) and (d) that obtaining quantitatively accurate and precise values of $P_c$ (and consequently, $Z_c$) for $n$-alkanes is something that has previously eluded the molecular simulation community. Finally, the simulation uncertainties for each critical constant typically increase with increasing chain-length. In addition, in some cases the simulation uncertainties reported in the literature are too large to provide any insight while in other cases the uncertainties are much smaller than would be expected considering the large range of literature values. For example, the uncertainties in $P_v$ reported by Muller et al. cause the $P_c$ uncertainties to be smaller than one symbol size. For this reason, the uncertainties that are assigned to molecular simulation results are often questioned by those not familiar with the field. The primary goal of this research is to provide quantitatively meaningful uncertainties for the simulation results so that they can be used to develop improved prediction models for the critical constants of large $n$-alkanes.

A significant factor when assessing the reliability of any physical property value, obtained either from prediction or experiment, is assigning the uncertainty associated with that value. De-
Figure 1.1: Comparison of critical constant trends with respect to carbon number for conflicting experimental data, prediction models, and the simulation results found in the literature. Panels (a)-(d) correspond to $T_c$, $P_c$, $\rho_c$, and $Z_c$, respectively. Carbon number is plotted on a log base 10 scale. The error bars represent the 95% confidence interval.

Determining an appropriate uncertainty is normally more straightforward for experimental measurements than for molecular simulations. This is because molecular simulations have many layers of uncertainty that propagate and are difficult to quantify. These layers of uncertainty can be listed in a bottom-up order as: the quality and type of data used for developing the model (liquid density, vapor pressure, \textit{ab initio} energy calculations, spectroscopic data, etc.), the intramolecular model
used (rigid bond lengths, flexible bonds, angles, etc.), the type of site model chosen (all-atom, united-atom, anisotropic-united-atom, etc.), the mathematical form of the model chosen (Lennard-Jones, Exponential-6, etc.), the optimization of the model parameters ($\epsilon$, $\sigma$, etc.), the simulation limitations (finite-size effects, cutoff errors, pairwise additivity, etc.), the experimental design conditions (temperature, composition, etc.), the type of simulation (GEMC, GCMC, MD, etc.), and the analysis of simulation data to yield the desired property (law of rectilinear diameters and density scaling law for $T_c$ and $\rho_c$, Clapeyron equation for $P_c$, etc.). To assess the uncertainty of each layer systematically, a top-down evaluation method is employed. This research has focused on addressing the top five layers of the uncertainty hierarchy in molecular simulation. Specifically, we quantified and reduced the uncertainties due to the data analysis, the inherent uncertainty in GEMC simulations, the choice of simulation temperatures, finite-size effects, and the Lennard-Jones 12-6 (LJ 12-6) parameters ($\epsilon$ and $\sigma$).

1.2 Objective

The purpose of the present work is six-fold. The first is to develop methods for predicting $T_c$, $\rho_c$, $P_c$, and $Z_c$ with greater accuracy and precision. The second is to obtain improved estimates of $T_c$, $\rho_c$, $P_c$, and $Z_c$ for large $n$-alkanes using different force field models. The third is to demonstrate that by reducing the critical constant uncertainties, particularly for $P_c$ and $Z_c$, we are able to elucidate the long-chain-length critical constant trends for $n$-alkanes. This will involve resolving a conflict between the $T_c$ values for large $n$-alkanes reported in the simulation literature for the NERD and Exp-6 models [23,25]. The fourth task is to verify that finite-size effects are reasonably small for the estimated critical constant values. The fifth goal is to demonstrate how to account for the uncertainties in the force field model when predicting the critical constants. The sixth and final objective is to develop new prediction models for $T_c$, $\rho_c$, $P_c$, and the normal boiling point ($T_b$) for the $n$-alkanes based on the insight that we have gained from the molecular simulation results. These models are to have a strong theoretical basis and a multi-property optimization will require that the properties are internally consistent.
1.3 Outline

The outline for this document is the following. In Chapter 2 we review the molecular simulation theory that is required for this work. In Chapter 3 we present a statistically rigorous approach for estimating the uncertainties in $T_c$ and $\rho_c$. Then, in Chapter 4 we focus our attention on $P_c$ (and, thereby, $Z_c$). In this chapter we develop an improved approach for predicting $P_c$ both accurately and precisely. We demonstrate how this approach can be used with several simple force field models. Next, in Chapter 5 we derive different methods for quantifying the uncertainty in the Lennard-Jones 12-6 parameters and how to propagate these uncertainties to the critical constants. In both Chapters 4 and 5 we predict the critical constants for C$_{16}$, C$_{24}$, C$_{36}$, and C$_{48}$ using several potential models. In Chapter 6 we apply the insight gained from the simulation results to develop new prediction models for large $n$-alkanes. Finally, in Chapter 7 we discuss our conclusions and recommendations.

The results presented in Chapter 3 were published in *The Journal of Chemical Physics* [30]. Chapter 4 contains results that were published in two separate publications. The work presented in Sections 4.2-4.8 was published in *Fluid Phase Equilibria* [31] while that from Sections 4.9-4.12 was published in the *Journal of Chemical & Engineering Data* [32]. Our intention is to publish the results from Chapter 5 in either *The Journal of Chemical Physics* or *The Journal of Physical Chemistry* by the end of January 2017. We will also submit the work found in Chapter 6 to *Fluid Phase Equilibria* by February 2017.
CHAPTER 2. MOLECULAR SIMULATION FUNDAMENTALS

In this chapter we discuss the essential concepts for understanding our molecular simulation results. First, we review the basics of several molecular simulation methodologies. Second, we discuss some more specific approaches that were implemented in our work. Next, we compare the different types of force fields found in the literature. Subsequently, we discuss a specific example for the parameterization approach of a transferable potential model. Then, we review the most significant layers of uncertainty for molecular simulations. Finally, we present the uncertainty quantification and propagation of errors approach used in this study.

2.1 Molecular Simulation Methods

The two approaches for performing molecular simulations are generally categorized as Monte Carlo (MC) and molecular dynamics (MD) [21,33]. Within the MC framework there are two popular ensembles for predicting vapor-liquid equilibria (VLE), namely, Gibbs Ensemble Monte Carlo (GEMC) [34] and Grand Canonical Monte Carlo (GCMC) [35]. Each of these methods has advantages and disadvantages, especially when simulating the two-phase dome of long carbon chains.

2.1.1 Monte Carlo

In general, the MC method involves attempting various types of MC moves to perturb and, eventually, equilibrate the system. Each attempted move is either accepted or rejected based upon the Metropolis algorithm [21,33]. Specifically, a move is accepted if a generated random number is less than the acceptance factor determined for a specific type of move. The most common MC moves can be classified as either particle displacement (i.e. translation, rotation, vibration, etc.), particle insertion/deletion, and volume exchange. Depending on the ensemble (GEMC or GCMC)
chosen, the particle insertion/deletion and volume exchange moves are performed in slightly different ways. The particle displacement moves are the same in each ensemble.

GEMC involves simulating two separate boxes that are in mechanical and chemical equilibrium with each other where one box approximates the liquid phase and the other the vapor phase. We have used the $NVT$ form of GEMC where the total number of molecules ($N$), the total volume ($V$), and the temperature ($T$) are specified. In this ensemble the particle insertion/deletion involves the simultaneous transfer of a particle from one box to the other. Particle transfer leads to the two boxes being in chemical equilibrium. The volume exchange is also between the two boxes while maintaining a constant total volume. Volume exchange assures that mechanical equilibrium exists, i.e. the two boxes are at the same pressure. Finally, particle displacement assures that each box is in local equilibrium for the prescribed temperature.

GCMC utilizes a single simulation box that is in chemical equilibrium with a reservoir. Chemical equilibrium is obtained by performing particle insertion/deletion moves where a molecule is transferred from the reservoir to the simulation box and vice versa. GCMC operates in the $\mu VT$ ensemble which requires the user to specify the chemical potential ($\mu$), volume of simulation box, and temperature. Much attention has been devoted to developing different methods for determining the chemical potential at which vapor-liquid equilibrium exists [36]. Again, displacements are required for local equilibrium but volume exchange moves are not necessary.

### 2.1.2 Molecular Dynamics

MC and MD are both feasible methods for calculations of equilibrium properties. MD has the added ability to perform simulations with respect to time. However, for the purposes of our research, this additional flexibility is not necessary since we are concerned with thermodynamic equilibrium properties. That being said, since MD is a viable approach for simulating two-phase equilibria, it merits equal consideration. The main difference between MD and MC is that MD is deterministic rather than stochastic. This means that it does not perform random moves but instead it relies upon evaluating the forces acting on a particle to determine its next position. Common difficulties with two-phase MD simulations are: accounting for the interfacial tension, determining where each phase begins, and maintaining an isothermal or isobaric simulation [37–39].
2.1.3 Method Selection

Since our research involves simulations of large molecules, this was a significant factor when deciding whether to use GEMC, GCMC, or MD. Specifically, we considered three aspects of simulating large molecules: computational cost (feasibility), inherent numerical uncertainty, and finite-size effects.

Historically, simulating long chain compounds has proven difficult for MC simulations due to the near improbability of a successful insertion step into the liquid phase. That being said, as discussed in Section 2.1.4 special algorithm advances have enabled MC to overcome the challenge of the insertion step. By contrast, MD benefits from not requiring an insertion step. In addition, parallel computing can be used in MD which renders it more computationally friendly for large molecules. For example, the longest \( n \)-alkane to be simulated in the literature (for a united-atom model) was \( \text{C}_{100} \) for MD \([26]\) compared to \( \text{C}_{48} \) for GEMC and GCMC \([23, 25]\). Although the MC studies were several years older and computational costs have decreased dramatically since that time (ca. 1998), as discussed in Chapter 7 our efforts to simulate \( \text{C}_{60} \) with GEMC proved to be futile because of the low probability of a successful insertion step.

Although some studies have implemented MD for predicting critical constants, it is not an optimal approach because MD utilizes a single heterogenous simulation box and, thus, lacks a clear distinction between the two phases. Moreover, the algorithms used to identify which molecules constitute the vapor and liquid phases are tedious, iterative in nature, and introduce a large degree of uncertainty into the orthobaric densities \([37, 38]\). These large uncertainties in \( \rho_l \) and \( \rho_v \) lead to uncertainties in the critical constants that are too large to be useful. (Although the uncertainties reported by Muller et al. are surprisingly small \([26]\).) By contrast, GCMC has proven to be the most precise of the three methods for predicting the critical constants. Historically, the uncertainties in the critical constants obtained by GEMC are around an order of magnitude larger than those from GCMC. That being said, the uncertainties for GEMC are still reasonably small. Furthermore, it is possible that the uncertainties in the force fields themselves do not justify the extremely precise results that are reported in GCMC studies.

Finite-size effects must also be considered since larger compounds will necessitate smaller system sizes. This is significant since different methods suffer from finite-size effects to varying extents. For example, GCMC is highly susceptible to finite-size effects and, therefore, finite-
size corrections must be included when using this approach [22]. Fortunately, rigorous finite-size scaling relationships exist for this ensemble that allow for extrapolation to the infinite system size [21]. This is done by performing numerous simulations at different system sizes and then regressing these results to a linear model. However, not only does this come at an increased computational cost but the extrapolation also introduces another level of uncertainty. In contrast to GCMC, a limitation for MD and GEMC is the lack of a rigorous approach to correct for finite-size effects. Instead, MD and GEMC rely upon heuristic brute force methods to determine whether or not a given system suffers from finite-size effects.

For example, Muller et al. demonstrated that finite-size effects are significant for MD by comparing the vapor-liquid coexistence curve (VLCC) obtained with two different intermolecular cutoff lengths of 13.8 and 24 Å [26]. By contrast, it was recently shown that finite-size effects for $T_c$, $\rho_c$, and $P_c$ are smaller than the corresponding statistical uncertainty for GEMC systems with 200 molecules of $n$-decane using a 14 Å cutoff [22]. Specifically, the critical constant values are equivalent, to within the statistical uncertainty, for the aforementioned system and a 1600 molecule system with cutoff lengths of 14, 20, and 30 Å. Therefore, as reported by Dinpajooh et al., “when resources are not available for a rigorous finite-size scaling study, GEMC simulations provide a straightforward route to determine fairly accurate critical properties using relatively small system sizes.” [22] Mitigating finite-size effects is particularly important in this work since the focus is large compounds where smaller system sizes are necessary to reduce the computational cost. For this reason, we have decided to employ the GEMC method in this study. However, this necessitated investigating ways to improve upon the ability for GEMC to predict “fairly accurate critical properties.” Furthermore, we considered it necessary to validate the claim that finite-size effects are small for GEMC simulations of larger compounds.

In summary, the advantage of the MD approach is the reduced computational cost. The main disadvantage is the very large uncertainties in VLE and critical constant estimates. In addition, it appears that finite-size effects play a significant role in MD and it is unclear what size of system is necessary to mitigate this effect. The main advantage in GCMC is the extremely precise estimates of the VLCC. However, the need for finite-size corrections limits the feasibility of this approach. GEMC has a similar computational cost as GCMC but without the need for expensive finite-size corrections. The primary limitation in GEMC is the inherent uncertainty, although it
is still much smaller than that from MD. These conclusions suggest that GEMC is an attractive option to obtain reliable critical constants for large molecules. Thus, since large compounds are the primary focus of this study, GEMC is our method of choice. In addition, GEMC is perhaps the most common type of molecular simulation for predicting the two-phase VLE dome. Therefore, the methods developed in this work will make a significant contribution to the simulation literature as future researchers implement our recommendations.

2.1.4 Configurational-bias

As mentioned previously, simulating large compounds with GEMC was not feasible for many years because the probability of a successful insertion move becomes infinitesimal as the carbon length becomes greater than C\textsubscript{10} [40]. Siepmann and de Pablo developed several configurational-bias moves to overcome this deficiency in MC methods by not inserting the entire molecule at once [41–43]. Instead, the molecule is inserted piecewise, one interaction site at a time. The next site is inserted by first proposing several possible locations and then accepting an energetically favorable new position. Since this method biases the configurational sampling, a correction term must be included in the Metropolis algorithm acceptance criterion [44].

The configurational-bias methodology is also very useful for performing many other types of MC moves such as the “molecular regrowth” move [45]. This move does not transfer molecules between the two boxes, rather it chooses a molecule in a single box and regrows a portion of that molecule in a different direction. Essentially, the regrowth move has the same effect as a translation and rotation move. The advantage of the regrowth move is that it does not require moving an entire molecule or rotating a large portion of a molecule, which becomes less and less probable in the liquid phase for large (and complex) molecules. The primary purpose of the regrowth move is to ensure that local equilibrium is obtained for a specific box. This is important for simulations of large compounds that can get trapped in local minima and, thus, may have difficulty sampling from the entire phase space. Although even more sophisticated moves have been developed for simulating large compounds these moves are not widely available on open source codes, e.g. rebridging and double rebridging [46]. Fortunately, the insertion/deletion and molecular regrowth configurational-bias moves are commonly found in most MC packages.
2.1.5 Simulation Package

The GEMC code that Martin and Siepmann developed at the University of Minnesota is now available online as “Towhee.” The Towhee package has been extensively used in the literature for GEMC simulations. Siepmann also believed this code could be used to predict the critical point of carbon chains even longer than \( C_{48} \). For these reasons, we decided to use Towhee for performing molecular simulations. Specifically, we used the Towhee 7.0.4 package with the DX-1597-2-7 pseudorandom number generator [47].

Recently two new MC codes were developed that have demonstrated considerable improvement compared to Towhee, namely, Cassandra [48] and GOMC [49]. Cassandra is the first parallel computing open-source MC code. GOMC is the first GPU based open-source MC code. Both of these packages are considerably faster than Towhee, but as both are in their developmental stages, they include fewer features and force field models. Thus, Towhee was chosen as the simulation package in this study.

2.1.6 Simulation Protocol

In order to perform a GEMC simulation it is necessary to specify the number of molecules \( N \), total volume \( V \), and temperature \( T \) of the system. In addition, the values of \( N \) and \( V \) that pertain to the liquid and vapor phases are strongly related to the initial conditions. As recommended by Morales et al., we initialized our systems with 20% of the molecules in the vapor phase and 80% in the liquid phase [50]. The initial volumes for the two boxes were such that the corresponding densities would be similar to the predicted densities. We performed replicate simulations that utilized different initial configurations and random seeds. This was done, in part, to obtain independent results but also to further ensure that our simulations of large compounds adequately sampled from the entire phase space rather than from a single local minimum. In each chapter we detail the specific values for the total number of molecules, the temperature, the number of replicates at each temperature, and the number of the MC moves used in the equilibration and production periods.

As mentioned previously, GEMC simulations perform several different MC moves. The probability of performing a specific MC move is set by the user prior to the simulation. Morales et al. demonstrated how a judicious selection of MC move probabilities can reduce the uncertainties
in the VLE results [50]. We took their recommendations into consideration when selecting our MC move probabilities. Specifically, the probabilities of performing a volume exchange, molecule regrowth [45], configurational bias exchange [41], translation, and rotation move were 1%, 13%, 20%, 33%, and 33%, respectively. We employed the same move probabilities unilaterally for each system.

2.2 Molecular Models

Although the efficiency and feasibility of a simulation depend heavily on the technique, the accuracy and insight come directly from the model employed. Therefore, in this section we focus the discussion on the various molecular models available to simulate \( n \)-alkanes.

The purpose of the molecular model, or force field, is to mathematically express the potential energy landscape, i.e. to calculate the potential energy for any configuration of molecules. The difference in potential energy between two states is used in MC simulations to determine whether or not a move is accepted. For MD simulations, the force field is used to calculate the forces acting upon each interaction site, hence the name. Force fields consist of both intramolecular and intermolecular interactions. Intramolecular potentials include vibrations of the bond distance, bending of the bond angle, and rotations around the torsional angles. Since intramolecular interactions have only a small effect on VLE properties, we will focus on the intermolecular forces [27].

Intermolecular interactions are those between two separate molecules. These are generally represented as a function that accounts for both repulsive and attractive forces. Some of the most common mathematical expressions for the intermolecular interactions are Lennard-Jones (LJ) 12-6, Mie \( \lambda \)-6, Exponential-6, and modified Morse. The LJ 12-6 and Mie \( \lambda \)-6 (a generalized version of the LJ 12-6) equation format is preferred over the last two because evaluating an exponential is computationally more expensive than evaluating a power. All of these equations can include a point-charge term to account for Coulombic interactions but this complicates the simulation because of the long-range interactions of point charges that require Ewald summations [33, 51]. Fortunately, this inclusion is unnecessary in cases where the potential model has been developed and properly parameterized to include the charge distribution within the molecule as part of the shorter-range potential model [52]. In addition, Coulombic interactions are not of interest in this study since our concern is with \( n \)-alkanes.
2.2.1 Comparison of Force Field Models

A plethora of force field models exist in the literature, particularly for \( n \)-alkanes. In this section we compare only the force field models simulated as a part of our research. Specifically, we simulated the TraPPE-UA (or simply TraPPE) [27], NERD [25], Exp-6 [23], and the force field developed by Messerly et al. for Uncertainty Propagation (Mess-UP). These models were selected after an extensive review of the performance of different force fields for predicting VLE and critical constants. Admittedly, the Mie 16-6 model developed by Potoff et al. for \( n \)-alkanes may be an omission since it has received considerable attention due to its ability to predict \( P_v \) [24]. However, we did not feel that it was necessary to simulate this model since it performs slightly worse than the others at predicting \( \rho_l \) and \( T_c \), which are more important than \( P_v \) for our purposes (see discussion in Chapter 4). The Mie 16-6 model is included in the discussion that follows for completeness.

Intramolecular Potentials

The TraPPE, Mie 16-6, and Exp-6 models for \( n \)-alkanes use constant bond lengths. For the TraPPE and Mie 16-6 models the length is 1.54 Å for all bond types while for the Exp-6 model the CH\(_3\)-CH\(_2\) and CH\(_2\)-CH\(_2\) bond lengths are 1.687 Å and 1.535 Å, respectively. For the NERD model, the bond lengths are subject to a harmonic stretching potential

\[
u_{\text{bond}}(r) = K_r (r - r_{\text{eq}})^2 \tag{2.1}\]

where \( u_{\text{bond}} \) is the bond stretching potential, \( r \) is the distance between neighboring sites, \( K_r \) is the bond stretching constant, and \( r_{\text{eq}} \) is the equilibrium bond length. For the NERD model, \( K_r = 48250 \) K/Å\(^2\) and \( r_{\text{eq}} = 1.54 \) Å. For each model, bond angles bend according to a harmonic potential

\[
u_{\text{bend}}(\theta) = K_\theta (\theta - \theta_{\text{eq}})^2 \tag{2.2}\]

where \( u_{\text{bend}} \) is the bond bending potential, \( \theta \) is the angle formed by three neighboring sites, \( K_\theta \) is the bond bending constant, and \( \theta_{\text{eq}} \) is the equilibrium bond angle. For each model, \( K_\theta = 31250 \) K/rad\(^2\) and \( \theta_{\text{eq}} = 114^\circ \). Bond rotations, represented the same in each model, are governed by the
following torsional potential

\[ u_{\text{tor}}(\phi) = c_1(1 + \cos(\phi)) + c_2(1 - \cos(2\phi)) + c_3(1 + \cos(3\phi)) \]  

(2.3)

where \( u_{\text{tor}} \) is the torsion potential, \( \phi \) is the dihedral angle formed by four sites, and \( c_1, c_2, \) and \( c_3 \) are Fourier coefficients. The fitting constants for the Exp-6 model are \( c_1 = 177.515 \) K, \( c_2 = -34.095 \) K, and \( c_3 = 395.66 \) K while the parameter values for the TraPPE, NERD, and Mie 16-6 models are twice those of the Exp-6 model, i.e. \( c_1 = 355.03 \) K, \( c_2 = -68.19 \) K, and \( c_3 = 791.32 \) K. Since the intramolecular interactions play a small role in predicting VLE, the Mess-UP model presented in Chapter 5 utilizes the same intramolecular potentials as the TraPPE and Mie 16-6 models.

### Intermolecular Potentials

The TraPPE, NERD, and Mess-UP force field models utilize the LJ 12-6 potential model for intermolecular site-site interactions

\[ u_{\text{nb}}(r) = 4\varepsilon \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^{6} \right] \]  

(2.4)

where \( u_{\text{nb}} \) is the non-bonding intermolecular potential, \( r \) is the interstitial distance, \(-\varepsilon\) is the minimum value for \( u_{\text{nb}} \), and \( \sigma \) is the distance where \( u_{\text{nb}} = 0 \). A more generalized form of the LJ 12-6 potential is the Mie \( \lambda \)-6 potential

\[ u_{\text{nb}}(r) = \left( \frac{\lambda}{\lambda - 6} \right) \left( \frac{\lambda}{6} \right)^{\frac{\lambda}{\lambda - 6}} \varepsilon \left[ \left( \frac{\sigma}{r} \right)^{\lambda} - \left( \frac{\sigma}{r} \right)^{6} \right] \]  

(2.5)

where the repulsive exponent (\( \lambda \)) is an additional parameter, whereas in the LJ 12-6 potential \( \lambda = 12 \) (note that for \( \lambda = 12 \) Equation 2.5 simplifies to Equation 2.4). The normalization factor (equal to 4 with \( \lambda = 12 \)) assures that the potential minimum has a value of \(-\varepsilon\) regardless of the value of \( \lambda \). The Mie 16-6 force field model developed by Potoff et al. for \( n \)-alkanes uses \( \lambda = 16 \) for CH\(_3\) and CH\(_2\) interaction sites. The Exp-6 force field model developed by Errington et al. employs
the Buckingham exponential-6 potential

\[
    u_{nb}(r) = \begin{cases} 
    \frac{\varepsilon}{1 - \frac{6}{\omega} \exp \left( \omega \left[ 1 - \frac{r}{r_m} \right] \right) - \left( \frac{r_m}{r} \right)^6} 
    & \text{for } r > r_{\text{max}} \\
    \infty 
    & \text{for } r < r_{\text{max}} 
    \end{cases}
\] (2.6)

where \( u_{nb} \), \( \varepsilon \), and \( r \) are the same as in Equation 2.4, \( r_m \) is the distance that corresponds to the minimum in the potential (i.e. \( u_{nb}(r_m) = -\varepsilon \)), \( \omega \) is a Buckingham exponential-6 parameter, and \( r_{\text{max}} \) is the smallest positive value for which \( \frac{du_{nb}}{dr} = 0 \). Figure 2.1 plots Equations 2.4-2.6 on a reduced scale, i.e. \( r^* = \frac{r}{\sigma} \) and \( \varepsilon^* = \frac{u_{nb}}{\varepsilon} \), with \( \lambda = \omega = 16 \) (to be consistent with the values reported in Refs [23] and [24]).

![Figure 2.1: Comparison of intermolecular potentials on a reduced scale, i.e. \( r^* = \frac{r}{\sigma} \) and \( \varepsilon^* = \frac{u_{nb}}{\varepsilon} \). Included are the Lennard-Jones 12-6, Mie 16-6, and Buckingham Exponential-6 with \( \omega = 16 \).](image)

Combining rules are commonly implemented in order to reduce the number of parameters optimized when developing a force field (see Section 2.2.3). The Lorentz-Berthelot (LB) combining rules are used for each of these force fields despite lacking a strong theoretical basis [53]. These rules can be expressed as

\[
    \sigma_{i,j} = \frac{\sigma_{i,i} + \sigma_{j,j}}{2}
\] (2.7)
where the subscript $i, j$ refers to cross-interaction parameters and the subscripts $i, i$ and $j, j$ refer to same site parameters. Equations 2.7-2.10 should be regarded as an integral aspect of the potential models since they act as constraints in the force field development process.

Figure 2.2 is provided to help visualize the subtle differences in the TraPPE, NERD, Mess-UP, Mie 16-6, and Exp-6 force fields. Panels (a) and (b) of Figure 2.2 depict the non-bonded potentials for the CH$_3$-CH$_3$ and CH$_2$-CH$_2$ interactions, respectively. We have provided all of the intermolecular potential parameters (including cross-interactions) in Table 2.1. The distance that corresponds to the potential minimum ($r_m$) is calculated for the LJ 12-6 models (TraPPE, NERD, and Mess-UP) using $r_m = 2^{\frac{1}{6}}\sigma$ and for the Mie 16-6 model with $r_m = \frac{8}{3}\pi\sigma$. By contrast, the $r_m$ values for the Exp-6 model are obtained by numerically solving $u_{nb}(\sigma) = 0$ (because Errington et al. reported $\sigma$ instead of $r_m$).

![Figure 2.2: Comparison of TraPPE, NERD, Mess-UP, Mie 16-6, and Exp-6 force fields. Panels (a) and (b) correspond to the CH$_3$-CH$_3$ and CH$_2$-CH$_2$ interactions, respectively.](image-url)
Table 2.1: Intermolecular potential parameters for the NERD, TraPPE, Mess-UP, Mie 16-6, and Exp-6 models.

<table>
<thead>
<tr>
<th>Model</th>
<th>CH₃-CH₃ Interactions</th>
<th>CH₂-CH₂ Interactions</th>
<th>CH₃-CH₂ Cross-Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\varepsilon_{k_B}$ (K)</td>
<td>$\sigma$ (Å)</td>
<td>$r_m$ (Å)</td>
</tr>
<tr>
<td>NERD</td>
<td>104.0</td>
<td>3.91</td>
<td>4.39</td>
</tr>
<tr>
<td>TraPPE</td>
<td>98</td>
<td>3.75</td>
<td>4.21</td>
</tr>
<tr>
<td>Mie 16-6</td>
<td>121.25</td>
<td>3.783</td>
<td>4.173</td>
</tr>
<tr>
<td>Exp-6</td>
<td>129.6</td>
<td>3.679</td>
<td>4.094</td>
</tr>
</tbody>
</table>

The functional form utilized by Towhee for the Exponential-6 model is

\[ u_{nb}(r) = \frac{A}{r^6} + B\exp(Cr) \quad (2.11) \]

where \( A, B, \) and \( C \) are parameters that can be obtained from \( \varepsilon, r_m, \) and \( \omega \) with the following expressions

\[ A = -\frac{\varepsilon r_m^6}{1 - \frac{6}{\omega}} \quad (2.12) \]

\[ B = \frac{6\varepsilon \exp(\omega)}{\omega - 6} \quad (2.13) \]

\[ C = -\frac{\omega}{r_m} \quad (2.14) \]

We have provided \( A, B, \) and \( C \) in Table 2.2 to facilitate future research and validation of our results using Towhee force field files. Furthermore, since Towhee requires \( r_{\text{max}} \) values for the Exponential-6 model, we have provided \( r_{\text{max}} \) in Table 2.2.

Table 2.2: Exp-6 intermolecular potential parameters as required by Towhee force field files.

<table>
<thead>
<tr>
<th>Site-Site Interactions</th>
<th>A (x10^{-5}) (K Å^6)</th>
<th>B (x10^{-9}) (K)</th>
<th>C (Å^{-1})</th>
<th>r_{\text{max}} (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH₃-CH₃</td>
<td>-9.765</td>
<td>0.691</td>
<td>-3.908</td>
<td>0.574</td>
</tr>
<tr>
<td>CH₂-CH₂</td>
<td>-6.942</td>
<td>98.81</td>
<td>-5.046</td>
<td>0.221</td>
</tr>
<tr>
<td>CH₃-CH₂</td>
<td>-8.172</td>
<td>6.453</td>
<td>-4.440</td>
<td>0.365</td>
</tr>
</tbody>
</table>

2.2.2 Implementation in Simulation

In order to implement these equations in a simulation, it is necessary to assume pair-wise additivity. This assumption is preferable due to the fact that it is extremely difficult and practically infeasible to perform calculations for all of the different multi-body interactions. Also, pair-wise additivity often works well even at high densities due to cancellation of higher order interactions [54]. Furthermore, the non-bonded interactions are typically truncated at a distance where they are assumed to be negligible and the radial distribution function is assumed to be 1. We used a cutoff distance of 13.8 Å for the NERD potential and 14 Å for the TraPPE, Exp-6, and Mess-UP models.
The cutoff distance for the TraPPE and NERD models were chosen so as to be consistent with the values used in the development of the respective force field. However, since Errington et al. did not report a long-range cutoff distance for the Exp-6 model, we chose the larger value of 14 Å. The same reasoning was applied for selecting the cutoff distance for the Mess-UP model. For each model, we implemented standard long-range tail corrections [21]. (For an important discussion of the use of this approach with polymeric molecules see Note 18 in Ref [25].) In addition, we employed a hard inner cutoff value of 1 Å to improve the computational efficiency. Notice that our interior cutoff of 1 Å is larger than each $r_{\text{max}}$ value found in Table 2.2.

2.2.3 Model Development

Once the mathematical model has been selected, the task of determining the model parameters still remains. The two general categories of models are those that rely upon experimental data regression (empirical) and those that rely upon quantum mechanical calculations (predictive) for their parameters.

Predictive Parameters

Although predictive models are less common, they do represent one of the ultimate objectives for molecular simulations; namely, to be able to predict properties without any experimental data. These methods rely upon quantum mechanics, physical mechanics, and complex mathematics. These models are often obtained by performing energy computations between two molecules in various orientations and at several distances. For example, Patel [55] utilized a modified Morse model with parameters regressed from calculations performed by Rowley et al. with MP2/6-311+G(2df,2pd) level of theory and basis set [52, 56]. Obviously, it is impossible to calculate the energy at every orientation and distance possible. For this reason, the aforementioned mathematical model parameters are regressed to the $ab\ initio$ data in a pair-wise fashion. Although modified Morse potentials are computationally more demanding, they are more effective at fitting potential energy landscapes for predictive models [52].
Empirical Parameters

Empirical models are considerably more popular than predictive models (although intramolecular parameters are commonly obtained from \textit{ab initio} calculations). For example, the TraPPE, NERD, Mie 16-6, Exp-6, and Mess-UP models all use empirical intermolecular parameters. This is due to the fact that empirical parameters are easier to obtain (i.e. no extensive knowledge of chemistry, physics, or mathematics is needed) and because empirical models generally agree more than predictive models when compared with experimental data. The reason for their accuracy is that short-comings in the model can be overcome by the fitting of parameters to experimental data. For example, empirical parameters can compensate for neglecting three body interactions [38]. Also, it is possible, and quite common, to generate united-atom (UA) models by combining carbon-hydrogen bonds into one pseudo-atom. By contrast, predictive parameter models tend to use all-atom (AA) molecules. The UA approach is useful because it greatly reduces the number of interacting sites which decreases the computational demand. This is vital for our purposes of predicting larger \textit{n}-alkanes. However, since empirical models do have a low level of theory they may breakdown for simulations that are at conditions too far removed from where the data were tabulated. For example, Siepmann demonstrated that changing the compounds from which the parameters are regressed can affect which compounds have the most accurate results [57].

Parameterization of TraPPE Model

The Transferable Potentials for Phase Equilibria (TraPPE) family of force fields is commonly regarded as the most inclusive and accurate transferable potential model for organic compounds. In the present study, the term “transferable” simply means that the LJ 12-6 parameters for the different interaction sites are transferable between \textit{n}-alkane compounds. The primary objective of transferability is to maximize the ability to build new chemical compounds by minimizing the number of interaction sites included [27, 58]. We have chosen the TraPPE-UA (referred to simply as TraPPE) methodology as our prototype for parameterization in Chapter 5. Because CH$_3$ and CH$_2$ sites are commonly found in organic compounds, the \textit{n}-alkane family is a logical starting point for force field parameterization. The UA model is implemented because it greatly reduces
the number of interaction sites (by grouping hydrogen and carbon atoms into a single site) which enables the simulation of larger molecules.

In Chapter 5 we attempt to replicate the original work done by Martin et al. [27] Specifically, we repeat the TraPPE intermolecular parameterization by performing simulations at the same temperatures, with the same compounds (ethane and \( n \)-octane), and with the same objective function. The objective function used to optimize \( \varepsilon_{CH_3} \), \( \sigma_{CH_3} \), \( \varepsilon_{CH_2} \), and \( \sigma_{CH_2} \) by Martin et al. and in this study is

\[
RMS(\theta) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\rho_i - \hat{\rho}_i(T_i, \theta))^2}
\]

where \( RMS \) is the root-mean-square, \( \rho_i \) are the experimental liquid density data at temperature \( T_i \), \( N \) is the number of data points, and \( \hat{\rho}_i(T_i, \theta) \) is the simulation result for \( \rho_i \) at \( T_i \) using the parameters \( \theta \). Although reproducing the experimental \( T_c \) value to within a given tolerance was also a requirement for the TraPPE \( n \)-alkane model, Martin et al. included \( T_c \) more as a constraint rather than as a multi-property optimization.

A fundamental assumption in this process is that the \( \varepsilon_{CH_3} \) and \( \sigma_{CH_3} \) values obtained from ethane are transferable to \( n \)-octane. This assumption greatly simplifies the optimization procedure when compared with regressing all four parameters simultaneously (i.e. the NERD model [25]). A four parameter optimization is very difficult primarily due to the strong correlation between the \( CH_3 \) and \( CH_2 \) parameters [59]. Also, the uncertainty quantification approach outlined in Section 2.4 is considerably more difficult for a four parameter optimization. This is because scanning a 4-dimensional parameter space is more computationally expensive than a pair of 2-dimensional scans. This is another reason we have chosen the TraPPE model as our prototype in Chapter 5.

### 2.3 Layers of Uncertainty

In Chapter 1 we listed numerous sources of uncertainty in molecular simulation. However, there are three sources of uncertainty that are generally considered the most significant. In general, these uncertainties are classified as numerical, model, and parameter uncertainty (although some authors use slightly different distinctions [60, 61]).

The numerical error refers to the uncertainty in the molecular simulation output due to fluctuations, finite-time, finite-size, simulation parameters, and data analysis methods. In this
study the numerical error is the uncertainty associated with the vapor-liquid coexistence curve obtained from GEMC simulations and the extrapolation to the critical point. The size of these uncertainties is dependent upon the system size, number of MC cycles, and MC move probabilities. The model error refers to several different approximations in the force field model that simplify the computational cost for a molecular simulation. These approximations are typically related to the mathematical form for calculating intramolecular and intermolecular interactions, combining rules of unlike interaction sites, the cutoff length for truncating short and long range interactions, the grouping of an electron cloud into a single point interaction site (either with an AA or UA representation), pair-wise additivity, and the neglect of three body interactions. Finally, parameter uncertainty refers to the imprecision in the parameter values that are required for the force field model. Parameter uncertainties arise from both the uncertainty in the experimental data used in the parameterization as well as from the model limitations.

2.3.1 Numerical Uncertainty

Molecular simulation is a numerical approach for approximately solving an exact equation derived from statistical mechanics—namely, the configurational partition function \( Z \) [21, 33]. This approach is necessary because solving for \( Z \) requires integrating \( 3N \) integrals, where \( N \) is the number of molecules in the system [54]. Although molecular simulation makes solving the many-body-problem tractable it does come with the cost that the output from molecular simulation has random error. With the advances in computational power, the effect of numerical uncertainty continues to decrease. That is to say that in the early years of molecular simulation the uncertainty associated just with the simulation output was often the largest source of uncertainty. By contrast, often simulation results in the recent literature assign uncertainties that are very small. However, these results must be viewed with caution because they typically only account for numerical uncertainty. In Chapters 3-4 we focus primarily on rigorous means of quantifying and reducing numerical uncertainty. The reduction of the numerical uncertainty in the critical constants is especially important when utilizing the GEMC approach.
2.3.2 Model Uncertainty

One of the largest sources of model uncertainty is due to the inexactness of the mathematical form for intermolecular interactions. In the present study the model was a Lennard-Jones 12-6 potential (LJ 12-6) without Coulombic contributions. It is important to remember that the LJ 12-6 model is only a crude approximation of the true dispersive interactions. Fortunately, it is typically believed that neglecting electrostatic interactions is a safe assumption for \( n \)-alkanes. Another source of model error is due to the united-atom approximation. Although this is often a safe assumption, in some cases the UA model is known to perform poorly. For example, the TraPPE-EH (explicit hydrogen) model greatly outperforms the TraPPE-UA model for predicting \( P_v \) [62]. The assumption of pair-wise additivity is typically valid and reasonably accurate. Likewise, the cutoff of 2.5 \( \sigma \) generally leads to a negligible error.

Transferability is another source of model uncertainty. The assumption that the CH\(_3\) parameters obtained from ethane transfer to all \( n \)-alkanes is often considered to be a good assumption. However, since the optimal CH\(_3\) parameters for \( n \)-octane are likely slightly different, the CH\(_2\) parameters must compensate in the optimization scheme. This issue is related to the use of LB combining rules, which are an additional source of model uncertainty. Finally, the assumption that the CH\(_2\) parameters obtained for \( n \)-octane transfer to even larger compounds is another potential source of model uncertainty. We believe that the assumption of transferability leads to a consistent over-estimation of \( T_c \) for longer chain-lengths for the TraPPE and Mess-UP models. By contrast, the NERD CH\(_3\) and CH\(_2\) parameters were optimized simultaneously to \( C_{10}, C_{12}, \) and \( C_{16} \). This is likely the reason why the NERD model appears to transfer better to larger \( n \)-alkanes.

Quantifying model uncertainty is difficult, especially when the model is being extrapolated to conditions far removed from the training set used in parameterization. In our research we accounted for model uncertainty by performing simulations with various force fields. Specifically, by comparing the NERD model results with those of the TraPPE and Mess-UP models we are able to determine the effect of assuming transferability from the relatively small \( n \)-alkanes used in the TraPPE and Mess-UP parameterization. Likewise, by comparing the TraPPE/NERD/Mess-UP models with the Exp-6 model we are able to assess how sensitive the critical constant results are to the intermolecular potential model form. Specifically, we evaluate whether a LJ 12-6 model is sufficiently reliable or if an Exponential-6 model form is necessary.
2.3.3 Parameter Uncertainty

The general approach for parameterization of the intermolecular parameters ($\varepsilon$, $\sigma$, etc.) is to minimize the deviations between simulation results and experimental data of a given physical property. Although numerical and model uncertainty play a role in this process, even if the simulation results and the model were exact there would still be an uncertainty associated with the optimal parameter set. This is because the experimental data have random error (assuming all systematic error has been eliminated by meticulous experimental work). This random error leads to uncertainty in the model parameters which leads to uncertainty in any property predicted from the resulting force field. For this reason, an Uncertainty Quantification and Propagation of Errors (UQ+PoE) approach is useful when reporting an optimal set of force field parameter values. Chapter 5 is centered around the UQ+PoE method for the LJ 12-6 parameter uncertainties. In the following section we discuss the general methodology for UQ+PoE.

2.4 Uncertainty Quantification and Propagation of Errors

In this section we present a frequentist statistical approach to uncertainty quantification (UQ) and propagation of errors (PoE). First, we develop the fundamental equations of UQ+PoE. Second, we discuss how to quantify the uncertainty in the parameters. These parameters may be those of a regression model (such as a straight-line equation) or a potential model (such as the Lennard-Jones 12-6). This discussion includes both how to develop a single parameter confidence interval as well as a joint parameter confidence region. Finally, we detail the methodology for propagating the uncertainty in the parameters.

2.4.1 Frequentist Theory

The fundamental equation to the frequentist nonlinear statistical analysis states that [63]

$$S(\theta) - S(\hat{\theta}) \leq ps^2 F_{p,v,\alpha}$$  \hspace{1cm} (2.16)

where $S(\theta)$ is the sum squared error of a specific set of parameters ($\theta$), $S(\hat{\theta})$ is the minimized sum squared error of the optimal set of parameters ($\hat{\theta}$), $p$ is the number of parameters, $s^2$ is an
independent estimate of the inherent variance \( (S_D^2) \) having \( \nu \) degrees of freedom, and \( F_{p,\nu,\alpha} \) is the F-statistic at the \( \alpha \) confidence level with \( p \) and \( \nu \) degrees of freedom. This equation is frequently rearranged to obtain the probability density function

\[
PDF(\theta) = F_{p,\nu}^{-1} \left( \frac{S(\theta) - S(\hat{\theta})}{ps^2} \right)
\]

where \( PDF(\theta) \) is the probability density function for a specific set of parameters and \( F_{p,\nu}^{-1} \) is the inverse of the F-statistic with \( p \) and \( \nu \) degrees of freedom.

In order to evaluate Equations 2.16-2.17 it is necessary to approximate \( s^2 \). It is common to assume that

\[
s^2 = \frac{S(\hat{\theta})}{n - p}
\]

and \( \nu = n - p \). Although \( s^2 \) is no longer an independent estimate of \( S_D^2 \) it has been shown that this assumption is valid in most practical cases [63]. After applying Equation 2.18 to Equation 2.16 and solving for \( S(\theta) \) we obtain

\[
S(\theta) \leq S(\hat{\theta}) \left( 1 + \frac{p}{n - p} F_{p,n-p,\alpha} \right)
\]

where \( F_{p,n-p,\alpha} \) is the F-statistic at the \( \alpha \) confidence level with \( p \) and \( n - p \) degrees of freedom. Likewise, we can substitute Equation 2.18 into Equation 2.17 to obtain

\[
PDF(\theta) = F_{p,n-p}^{-1} \left( \frac{S(\theta) - S(\hat{\theta})}{S(\hat{\theta})} \times \frac{n - p}{p} \right)
\]

where \( F_{p,n-p}^{-1} \) is the inverse of the F-statistic with \( p \) and \( n - p \) degrees of freedom.

Before proceeding, it is important to understand the assumptions made in deriving Equation 2.16 and the possible limitations of Equations 2.16-2.20. The left-hand side of Equation 2.16 provides the exact shape of the joint confidence region in the parameters, \( \theta \), for both linear and non-linear systems. Furthermore, the fundamental assumption that \( \frac{S(\theta) - S(\hat{\theta})}{ps^2} \) follows an F distribution has a strong theoretical basis. However, the “critical values” (i.e. the specific values of \( F_{p,\nu,\alpha} \)) are only approximations. Therefore, the confidence level, \( \alpha \), that satisfies Equations 2.16 and 2.19 and the \( PDF \) values calculated with Equations 2.17 and 2.20 are approximations to the true confidence.
level and PDF values, respectively. Fortunately, the F-statistic critical values are an adequate approximation for most systems, especially those that are only moderately non-linear [63].

2.4.2 Uncertainty Quantification

A rigorous assessment of the joint uncertainty in the parameters involves scanning the parameter space and accepting any combination of parameters that satisfies Equation 2.16 (or Equation 2.19). In other words, a set of parameters, $\theta$, is acceptable at the $\alpha$% confidence level if Equation 2.16 (or Equation 2.19) is true. The general, though not necessarily the most efficient, algorithm for obtaining a joint confidence region contains the following sequence of steps

1. Determine the optimal set of parameters ($\hat{\theta}$) by minimizing $S(\theta)$.
2. Select a range and discretization size for each parameter.
3. Cycle through every possible set of parameters ($\theta$) in the predefined ranges.
4. Calculate $S(\theta)$ for each parameter set.
5. Accept $\theta$ at an $\alpha$ confidence level if Equation 2.19 is satisfied.
6. If a parameter extremum chosen in step 2 is accepted, increase the range for the corresponding parameter(s).
7. Repeat steps 3-5.

Other algorithms are possible, and statistical packages, such as R, may be used to perform these calculations [64]. This method yields a joint confidence region that demonstrates the correlation between each parameter. In addition, a contour plot of the PDF depicts different confidence levels as the various contours.

The rigorous confidence interval for a given parameter is approximated as the range that parameter encompasses at the $\alpha$ confidence level [65]. Alternatively, the uncertainty in a single parameter can be determined with Monte Carlo Sampling (MCS) of the parameter space. This methodology is the same as the PoE approach for other properties not included in the parameter space (i.e. the product of two parameters, etc.).
2.4.3 Propagation of Errors

An important component of an MCS+PoE analysis is a means of properly sampling from the $p$-dimensional parameter space. This is typically done with the probability density function and a random number generator [60, 61]. The MCS+PoE algorithm is

1. Create a grid of the $p$-dimensional parameter space

2. Calculate the $PDF$ for each parameter set, $\theta$

3. Generate millions of random numbers

4. Assign each random number to a set of parameters based on $PDF(\theta)$

5. Obtain the desired property value for the parameter sets sampled in Step 4

6. Create a histogram of the property values from Step 5

7. Integrate the histogram at a given confidence level

where Steps 1-2, 3-4, and 5-6 are typically categorized as the UQ, MCS, and PoE steps, respectively. To clarify, when producing a single parameter uncertainty the “desired property” in Step 5 would simply be a single parameter. However, for a PoE analysis the desired property may be any combination of the parameters.
CHAPTER 3. NUMERICAL UNCERTAINTY IN $T_c$ AND $\rho_c$ WITH GEMC

3.1 Introduction

Prediction of the critical constants with GEMC requires an indirect approach involving regression of the VLE simulation data and subsequent extrapolation to the critical point. The need for this indirect approach is due to the inherent limitations of GEMC that prevent the method from converging near the critical point. Specifically, as the simulation temperature approaches the critical temperature, the identity of the two simulation boxes becomes ambiguous as they alternate between the liquid and vapor phases [66]. Although GEMC is ill-suited for predicting the critical point directly, it is still frequently used for this purpose because it is simple, relatively fast, and produces a clear interface between the two phases.

The primary weakness of the GEMC method is the inherent uncertainty in the simulation results and the subsequent analysis of those data. The working thesis of this chapter is that improving the extrapolation of the data using rigorous statistical analysis will improve critical point estimation. Specifically, the hypothesis tested in this chapter is that the uncertainty in predicting the critical point using GEMC can be reduced using a rigorously-correct approach for evaluating the statistical uncertainty of the regression to the data.

The outline of this chapter is as follows. First, the inherent uncertainty in GEMC simulations is described and an error model is then proposed. This is followed by an explanation of the algorithm for proper nonlinear statistical analysis. Next, a comparison is drawn between the parameter uncertainties from traditional and rigorous procedures. Finally, results are presented which show the need for proper nonlinear statistical analysis when estimating $T_c$ and $\rho_c$. 
3.2 Determining the Critical Point through Regression

The common approach used for predicting $T_c$ and $\rho_c$ using GEMC begins with running several simulations at different temperatures below the critical point. A phase diagram of the orthobaric vapor and liquid densities results from such simulations. Once these data are produced, the critical constants are found by performing a weighted least-squares regression of two coupled equations, namely, the law of rectilinear diameters and the density scaling law. The law of rectilinear diameters has a statistical mechanical derivation that involves a Taylor series expansion [67]. However, for regression purposes, this is often truncated after the linear term and can be expressed as

$$\rho_r \equiv \frac{\rho_l + \rho_v}{2} = \rho_c + A(T_c - T) \quad (3.1)$$

where $\rho_r$ is the rectilinear density, $\rho_l$ is the liquid density, $\rho_v$ is the vapor density, $\rho_c$ is the critical density, $A$ is a fitting parameter, $T_c$ is the critical temperature, and $T$ is the temperature. The density scaling law is

$$\rho_s \equiv \frac{\rho_l - \rho_v}{2} = B(T_c - T)^\beta \quad (3.2)$$

where $\rho_s$ is the scaling density, $B$ is a fitting parameter, and $\beta$ is the critical universality exponent, which is theoretically the same for all fluids [68]. Therefore, $\beta$ is usually not included as a regression parameter but is set to a specific value of 0.32, 0.325, or 0.326 [22].

As mentioned previously, a weighted regression is employed due to the fact that fluctuations in the orthobaric densities increase with increasing temperature. In accordance with the maximum likelihood approach [69], the weights utilized are the inverse of the population standard deviation ($S_D$). However, since $S_D$ is generally unknown, the sample standard deviation ($s$) is typically used instead. By using $s$ as an approximation for the “true” standard deviation the objective function can be expressed as

$$\sum_i \left( \frac{\hat{\rho}_r(T_i) - \rho_{r_i}}{s_{r_i}} \right)^2 + \left( \frac{\hat{\rho}_s(T_i) - \rho_{s_i}}{s_{s_i}} \right)^2 \quad (3.3)$$

where $\hat{\rho}_r(T_i)$ is the predicted rectilinear density (Equation 3.1) at the temperature corresponding to the simulation rectilinear density $\rho_{r_i}$ with sample standard deviation $s_{r_i}$ and $\hat{\rho}_s(T_i)$ is the predicted scaling density (Equation 3.2) at the temperature corresponding to the simulation scaling density $\rho_{s_i}$ with sample standard deviation $s_{s_i}$. 
3.3 Determining the Uncertainties in $T_c$ and $\rho_c$

Two factors need to be taken into account when estimating the uncertainty in the critical constants obtained from the regression process. The first is to properly estimate the standard deviations used in Equation 3.3. The second concerns the method used to estimate the error of the parameters found in the regression equations and how the method accounts for the interdependence of the parameters themselves. The following sections explain how the traditional methods fail in both regards and how proper rigorous statistics can improve the estimates of errors in the predicted critical constants. A discussion on calculating the standard deviation is first, followed by a discussion on calculating the rigorously-correct joint confidence regions for the parameters.

3.3.1 Estimating Standard Deviations

Before discussing the proper method to estimate $S_D$ with $s$, we will briefly discuss the difference between $s$ and the standard error of the mean ($S_E \equiv \frac{s}{\sqrt{n}}$, where $n$ is the number of samples). This discussion is important because the traditional approach for estimating the uncertainty in $T_c$ and $\rho_c$ makes use of $S_E$ for the simulation data rather than $s$. Therefore, we will use $S_E$ with the traditional linearized analysis but $s$ (as an estimate of $S_D$) with the rigorous nonlinear analysis.

Since the main purpose of this work is to demonstrate the need for the rigorous nonlinear analysis, the discussion that follows will focus primarily on how to properly estimate $S_D$, and not $S_E$. Furthermore, we will present an error model that enables prediction of $S_D$ a priori. An error model to predict $S_E$ and $s$ are not feasible, since these quantities depends on $n$. On the other hand, an error model for $S_D$ is useful because it reduces the need for large amounts of simulation data to properly estimate the uncertainty. This is helpful when performing a computationally-intensive process such as a weighted optimization of potential model parameters [70]. It is also essential for developing the experimental design proposed in Chapter 4.

In Sections 3.3.2 and 3.3.3, we will discuss two important considerations when developing our error model for $S_D$. The first is whether an individual run of $n$ blocks or $n$ replicate runs at a given temperature should be used to estimate $S_D$. The second is a demonstration that the traditional propagation of error approach neglects statistically relevant factors and, therefore, improperly weighs the different points. After clarifying these two issues, we will present an error
model that accurately estimates $S_D$ for $\rho_r$, $\rho_s$, $\rho_l$, and $\rho_v$ for $n$-alkanes. This error model is used in Chapters 4-5 to calculate critical parameters without devoting extensive computational resources to calculating the various $s$ (or $S_D$) values.

3.3.2 Blocking and Replicate Methods

Historically, the most common practice for estimating $S_E$ of the liquid and vapor densities is to perform a single simulation at each temperature of interest and divide the simulation into $n$ blocks. Due to the interdependency of the blocks, a special algorithm was developed to determine the block size necessary for an accurate estimate of $S_E$ [71]. Recently, practitioners have favored approximating $S_E$ by performing a few replicate simulations at the same temperature with different starting configurations and random seeds [50, 72]. Since these replicate simulations are inherently independent, no special algorithm for calculating $S_E$ is necessary.

In order to ensure that our error model is developed using accurate estimates for $S_D$, we investigated whether the blocking or replicate simulations approach should be used for calculating $s$. In theory, the blocking approach is valid for an ergodic simulation, such as GEMC; however, since replicate simulations can be run in parallel, we observed that the replicate approach is significantly more efficient. Therefore, consistent with the practice found in the recent literature, we have adopted the method of performing replicate simulations to approximate $S_D$ for our error model.

3.3.3 Indirect and Direct Methods

The standard practice when optimizing the objective function (Equation 3.3) is to assume that the standard deviation at a given temperature is the same for both the rectilinear and the scaling density, i.e. $s_r = s_s$, which causes the rectilinear and scaling portions of the optimization to have an equal weighting [47]. The rationale for this assumption comes from a propagation of error approach which states

$$s_{r,s}^2 = \left( \frac{\partial \rho_r}{\partial \rho_l} \right)^2 s_l^2 + \left( \frac{\partial \rho_r}{\partial \rho_v} \right)^2 s_v^2$$

(3.4)

$$s_r = s_s = \sqrt{s_l^2 + s_v^2}$$

(3.5)
where $s_l$ and $s_v$ represent the sample standard deviation for the liquid and vapor phases, respectively. A key step in the derivation for Equation 3.4 assumes that $\rho_l$ and $\rho_v$ are independent [73]. Unfortunately, this assumption is fundamentally invalid in a GEMC simulation. In fact, rather than being independent, $\rho_l$ and $\rho_v$ are connected, correlated, and dependent because as one increases the other decreases and vice versa. This dependence is due to the fact that the only way that the liquid (or vapor) density changes is by a particle transfer from one box to the other or from an exchange of volume between the two boxes.

In order to properly calculate $s_r$ and $s_s$, the actual quantities ($\rho_r$ and $\rho_s$) must be recorded throughout the simulation. This permits a direct statistical analysis of $\rho_r$ and $\rho_s$ to calculate their standard deviations rather than the indirect propagation of error approach which neglects the interdependence of the two quantities. The need for the direct approach can be demonstrated by comparing the rigorous direct values of $s_r$ and $s_s$ with the traditional propagation of error values.

Figure 3.1 provides a plot of $s_r$ and $s_s$ obtained from 20 replicate simulations of n-octane using a 200 molecule system. Since these replicate simulations are inherently independent, the standard deviation for $n = 20$ can be assumed to be identical to the “true” standard deviation (i.e. $s \approx S_D$) [73]. The equilibration and production periods both consisted of $10^5$ MC cycles, where a MC cycle consisted of 200 MC moves. As seen in Figure 3.1, the traditional approach underestimates $s_r$ and overestimates $s_s$. Note that there is only one set of data for the traditional approach because this method predicts that $s_{ri} = s_{si}$.

Figure 3.2 shows the estimates of $S_{Dr}$ and $S_{Ds}$ for five different n-alkanes ranging between C$_6$-C$_{16}$. Each point was generated by performing 20 independent simulations (again, so that $s \approx S_D$) at each reduced temperature ($T_r$) while tracking $\rho_r$ and $\rho_s$ directly. The simulation specifications were the same as those described previously. The simulations indicate that the standard deviation is not dependent upon the size of the compound when compared on a reduced temperature basis. This fact facilitates developing a model to predict $S_{Dr}$ and $S_{Ds}$ for the n-alkanes so that the computationally expensive process of performing several independent runs need not be repeated for other members of this family of compounds. An exponential model of the form $S_D = b_0 + b_1 e^{(b_2 \times T_r)}$ fits the $S_D$ data well, as indicated by the solid lines in Figure 3.2. The parameters for the model are listed in Table 3.1, along with similar parameters to estimate the liquid and vapor standard deviations ($S_{Dl}$ and $S_{Dv}$), although these are not displayed in Figure 3.2. In
addition, the dashed lines in Figure 3.2 represent the 95% confidence intervals of regression for the two error models and indicate the certainty with which the model describes the data. As seen in Figure 3.2, $s_{ri}$ and $s_{si}$ are statistically different at the 95% confidence level for $T_r > 0.825$. This further demonstrates that, because $\rho_l$ and $\rho_v$ are dependent, Equation 3.5 is not valid as it predicts that $s_{ri} = s_{si}$. Since $s_{ri} > s_{si}$, the density scaling law portion of the optimization should have a greater weighting than the rectilinear diameters portion (see Equation 3.3), especially near $T_c$.

Table 3.1: Results from a least-squares fit to the standard deviation data presented in Figure 3.2 using an exponential error model:

$$S_D = b_0 + b_1 e^{(b_2 \times T_r)}$$

Units are $\left( \frac{kg}{L} \right)$.

<table>
<thead>
<tr>
<th>Property</th>
<th>$b_0 \times 10^4$</th>
<th>$b_1 \times 10^{14}$</th>
<th>$b_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_{Dr}$</td>
<td>5.00</td>
<td>225</td>
<td>22.48</td>
</tr>
<tr>
<td>$S_{Ds}$</td>
<td>3.99</td>
<td>2.23</td>
<td>26.55</td>
</tr>
<tr>
<td>$S_{Dr}$</td>
<td>8.27</td>
<td>837</td>
<td>20.98</td>
</tr>
<tr>
<td>$S_{Dv}$</td>
<td>3.51</td>
<td>32.3</td>
<td>24.86</td>
</tr>
</tbody>
</table>
Figure 3.2: Error models for the standard deviation of the rectilinear and scaling densities. The solid lines and dashed lines correspond to the model best fits and the uncertainties in the model fits at the 95% confidence level, respectively. Symbols were obtained by analyzing 20 replicate simulations of 200 molecules for \(n\)-alkanes with carbon numbers of 6, 8, 10, 12, and 16.

3.3.4 Rigorous Nonlinear Analysis

Having demonstrated the correct way to estimate the standard deviations used in Equation 3.3 we will now explore the correct way to determine the uncertainties in parameter regression associated with the minimization of Equation 3.3. Specifically, for the present purpose we are concerned with the parameter uncertainties in \(T_c\) and \(\rho_c\). The most common way of reporting the uncertainty of \(T_c\) and \(\rho_c\) is to linearize the coupled equations (Equations 3.1 and 3.2), to use \(S_E\) as an estimate of the error in the simulation results, and to propagate the standard error associated with the corresponding slope and intercept. Although this method has been widely accepted, it lacks statistical rigor for nonlinear equations. As mentioned previously, for propagation of error to be appropriate, the standard deviations must be independent and, unfortunately, the slope and intercept parameters are highly correlated. In addition, linearizing an equation distorts the assumed normal error distribution so the traditional error analysis that uses a t-statistic and standard error is no longer valid.

Section 2.4 outlines rigorous algorithms for determining the single parameter uncertainty as well as a joint confidence region for the parameters \((A, B, T_c, \text{ and } \rho_c)\). Specifically, the joint
confidence regions are obtained following the algorithm found in Section 2.4.2 with Equation 2.19 while the single parameter uncertainties are obtained using the algorithm presented in Section 2.4.3 with Equation 2.20. (The results from this analysis are assumed to provide appropriate estimates of uncertainty since Equations 3.1-3.2 are nearly linear in the parameters.) Figures 3.3-3.4 are provided to demonstrate the expected results from this analysis. The data used to produce Figures 3.3-3.4 are the same as those discussed in Section 3.3.4.

In Figure 3.3 we present an example of the histograms produced for a single parameter uncertainty of $T_c$ and $\rho_c$ by MC Sampling of the $A, B, T_c,$ and $\rho_c$ parameter space. The 95% confidence interval is approximated by integrating the histogram such that 5% of the area is evenly distributed between the left and right tails. Also, the standard deviation in $T_c$ and $\rho_c$ is estimated by fitting a normal distribution model to the probability density, where the probability density is equal to the number of counts in a single bin divided by both the total number of counts and the bin width. This is useful for quantifying the numerical uncertainty in $T_c$ and $\rho_c$.

![Figure 3.3: Examples of the histograms produced from a Monte Carlo Sampling propagation of errors approach. Panels (a) and (b) correspond to $T_c$ and $\rho_c$, respectively. Probability density is defined as the number of counts in a single bin divided by both the total number of counts and the bin width. The data used are the same as those in Figure 3.5 Panel (b).]
Figure 3.4 contains an example of the resulting 95% joint confidence intervals for the six different pairs of parameters. Each plot is a 2-dimensional cross-section of the 4-dimensional parameter space. For graphical purposes, the two parameters that are not included in a given graph are set to their optimal values. As seen in Panels (c) and (d), the strongest correlation is found between $T_c$-$B$ and $\rho_c$-$A$, respectively.

**Comparison with Traditional Approach**

Technically, four parameters are fit when regressing Equations 3.1 and 3.2 against the simulation data — $A$, $B$, $T_c$ and $\rho_c$. Since $T_c$ and $\rho_c$ are the only parameters that have physical meaning, only their joint confidence region is discussed below and the other joint confidence regions will not be displayed. Recall that in Figure 3.4 the two-dimensional plots are depicted as cross-sections from the four-dimensional parameter space (i.e. the other two parameters are constant). However,
to make a valid comparison with the traditional confidence intervals for $T_c$ and $\rho_c$, the rigorous joint confidence regions are plotted as projections from the four-dimensional ($A$, $B$, $T_c$ and $\rho_c$) parameter space to the two-dimensional ($T_c$ and $\rho_c$) parameter space. Furthermore, this comparison requires that the rigorous joint confidence regions are plotted at the $\alpha^2$ confidence level while the traditional confidence intervals for $T_c$ and $\rho_c$ are plotted at the $\alpha$ confidence level.

Figure 3.5 demonstrates the insight gained by properly mapping out the joint confidence regions. The circles are data from GEMC simulations for ethane while the line is obtained by optimizing Equation 3.3. The simulation data in Figure 3.5 Panel (a) were obtained from the original publication of the TraPPE model [27] while the data in Panel (b) were from a more recent validation of the TraPPE model [72]. The main difference between the two data sets is that $S_E$ for the original data was estimated by blocking a single run whereas it was estimated by eight independent replicates for the validation data. The error bars for the data in (a) represent the 99.5% confidence intervals using the reported uncertainties. The diamond is the predicted critical point and the oval-like region is the 99.5% confidence region for the critical point as determined using the aforementioned rigorous algorithm. The rectangular region is the 99.5% confidence interval produced using the traditional method of regression analysis. Strictly speaking, the traditional and rigorous methods each provide a best fit estimate of $T_c$ and $\rho_c$. However, only a single critical point is presented in Figure 3.5 because both methods yield essentially identical results for $T_c$ and $\rho_c$. This is expected since the fundamental shortcoming in the traditional regression approach is in the assessment of parameter uncertainty, not in the estimate of the parameters. (The data used to generate the results in Figure 3.5 can be found in Tables A.1-A.2.)

The most significant result from Figure 3.5 is that the rigorous nonlinear analysis reduces the uncertainty in $T_c$ and $\rho_c$, even for the more recent data. Another important observation is that the rigorous approach does not result in nonphysical confidence regions for $T_c$ and $\rho_c$. Notice that in Figure 3.5 Panel (a) the traditional uncertainty analysis results in the lower limit of $T_c$ (264 K) being below that of the highest temperature used in simulation (275 K). Since the simulation produced a 2-phase system at 275 K, extending the $T_c$ confidence interval to 264 K is not sensible. By contrast, the rigorous statistical analysis produces a $T_c$ lower bound of 290 K — a value that is much more reasonable. The difference between the two confidence regions occurs because the traditional analysis assumes a normal distribution centered on the best estimate of $T_c$, when in fact,
Figure 3.5: Comparison between the traditional confidence interval and the rigorous joint confidence region for $T_c$ and $\rho_c$ at the 99.5% confidence level. Simulation data were obtained from the literature [27, 72]. The data used in (a) are from the original source where blocking of a single run was used, whereas, the data in (b) are from a more recent study where eight replicate runs were performed for each temperature. Error bars in (a) represent the 99.5% confidence intervals found from the uncertainties reported in the literature.

The confidence region is skewed towards higher temperatures. However, it should be noted that Figure 3.5 Panel (a) illustrates an extreme case simply to demonstrate that the traditional approach is invalid. The excessively large uncertainties in Panel (a) for $T_c$ and $\rho_c$ are primarily due to the fact that only three data points were used and $S_E$ was approximated from only five blocks. A more typical confidence region, both in size and symmetry, is that in Panel (b) (notice the difference in scale) where only the four highest temperatures (275, 279, 283, and 288 K) were included in the analysis, as recommended in Ref [72]. Although the rigorous confidence region in Panel (b) is still asymmetric, the degree of asymmetry is reduced because the uncertainty is small enough that the lower portion is less affected by the unfeasible region. (Note that the ellipticity and size are very different for the joint confidence regions presented in Figure 3.5 Panel (b) and Figure 3.4 Panel (a) because the former is a two-dimensional projection while the later is a cross-section.)

In review, there are two steps in the traditional error analysis approach that lead to improper predictions of uncertainty for $T_c$ and $\rho_c$ which we have corrected in this work. The first improvement is to determine $s_r$ and $s_s$ by directly analyzing $\rho_r$ and $\rho_s$, rather than an indirect propagation of
error approach. The second, and most important, improvement is to perform a rigorous regression analysis by scanning the parameter space instead of linearizing the equations, propagating $S_E$ for the densities, and propagating the standard error of the slope and intercept.

### 3.4 Results and Importance of Rigorous Statistical Analysis

In order to further demonstrate the importance of employing rigorous statistics when regressing $T_c$ and $\rho_c$ from GEMC data, this section describes a case study for $n$-octane. Results are given to quantify how much each of the improvements explained previously affects the parameter uncertainties. The simulation data were generated using the TraPPE potential model with 200 molecules by performing a single simulation at temperatures of 390, 440, 490, 515, and 543 K. The equilibration and production period both consisted of $10^5$ MC cycles (consisting of the same move probabilities as described in Section 2.1.6).

Table 3.2 contains the best fit values and 95% confidence intervals for $T_c$ and $\rho_c$ for $n$-octane using eight different combinations of traditional and statistically correct procedures. More specifically, the following routes are compared: blocking the individual run vs using an error model developed previously from replicates, propagation of error vs direct calculation for determining $s_r$ and $s_s$, and the linearized propagation of the standard error of the mean vs the rigorous nonlinear algorithm for regression uncertainties. To clarify, when propagating the error from the GEMC data, the linearized regression used $S_E$ for the top two rows whereas the bottom two rows used the error models proposed in Section 3.3.3. (The data used to generate the values in Table 3.2 can be found in Table A.3.)

Multiple significant conclusions are drawn from the results presented in Table 3.2. First, the completely rigorous uncertainties (Table 3.1/Direct/Nonlinear — bottom right uncertainties in Table 3.2) are much smaller than the completely traditional uncertainties (Blocking/Equation 3.5/Linearized — upper left uncertainties in Table 3.2). The significance of this fact will be magnified in the case study that follows in Section 3.5. Second, the largest reduction in uncertainty is observed when the rigorous nonlinear algorithm is utilized. Therefore, the key improvement to estimating the uncertainty in $T_c$ and $\rho_c$ is to use Equation 2.19 and the associated algorithm. Finally, although the different methods for estimating the standard deviations affect the weighting of Equation 3.3 their effect on the best fit estimates is small.
Table 3.2: The effect of the improvements proposed in this work on the uncertainty in the critical point presented at the 95% confidence level. An average value is reported for the rigorous confidence intervals since they are nearly symmetric. Units for $T_c$ and $\rho_c$ are K and $\frac{kg}{L}$, respectively.

<table>
<thead>
<tr>
<th>GEMC data analysis</th>
<th>Best fit estimates</th>
<th>Linearized Regression</th>
<th>Nonlinear Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertainty</td>
<td>s_r and s_s</td>
<td>$T_c$</td>
<td>$\rho_c$</td>
</tr>
<tr>
<td>Blocking</td>
<td>Equation 3.5</td>
<td>569.31</td>
<td>0.2391</td>
</tr>
<tr>
<td>Blocking</td>
<td>Direct</td>
<td>569.36</td>
<td>0.2390</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Equation 3.5</td>
<td>569.45</td>
<td>0.2394</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Direct</td>
<td>569.42</td>
<td>0.2390</td>
</tr>
</tbody>
</table>

3.5 Case Study: Long-Chain-Length Behavior of $n$-alkanes

The benefit of reducing the error associated with estimates of the critical constants can be appreciated by examining a case study from the literature. As mentioned in Chapter 1, two common but conflicting models for predicting $T_c$ and $V_c$ have been developed by Teja [12] and by Tsonopoulos [13]. These models are equally reliable for predicting $T_c$ and $V_c$ in the range where experimental data are available. The main difference between these models is found in their limiting behavior (see Figure 3.6). In other words, discerning between these two models is not possible when one resorts solely to experimental data. For this reason, molecular simulation has proven to be an attractive alternative [66].

For example, Nath et al. tried to elucidate which model was better by performing GEMC simulations with the NERD model for $n$-alkanes up to C_{48} to estimate $T_c$ and $V_c$ [25]. Based upon their results, they asserted that Tsonopoulos’ model was superior for $T_c$ and Teja’s model was better for $V_c$. Figure 3.6 demonstrates the results of Nath et al. along with the predictions of the two models in the region where they become distinguishable (carbon number greater than 20). Critical volume is on the left and critical temperature is on the right. The error bars are the 95% confidence intervals calculated using the traditional (Blocking/Equation 3.5/Linearized Propagation of Error) and rigorous (Table 3.1/Direct/Nonlinear) methods. To clarify, the rigorous
method uses the aforementioned error models for $S_{D_r}$ and $S_{D_s}$ whereas the traditional method utilizes the uncertainties reported by Nath et al. and Equation 3.5.

Although Nath et al. reported relatively small error bars for compounds up to C$_{16}$, the traditional uncertainty grows rapidly with respect to carbon number. This increasing uncertainty is problematic for the task at hand — determining whether the model of Teja or Tsonopoulos correctly predicts the long-chain-length behavior of $n$-alkanes. In fact, if the simulation results did indeed have this degree of uncertainty, the ability of GEMC to settle the debate would need to be called into question. However, the rigorous approach finds that Nath’s simulations are more reliable statistically than originally reported. Figure 3.6 shows that molecular simulation is capable of identifying which model correctly describes the long-chain-length behavior of $n$-alkanes. Specifically, Nath et al. are justified, at the 95% confidence level, in the conclusion that Teja’s model performs better for $V_c$ while that of Tsonopoulos is more reliable for $T_c$. The resulting joint confidence regions for $T_c$ and $\rho_c$ can be found in Chapter 4.

Figure 3.6: Comparisons between Nath’s reported simulation results with Teja’s model and Tsonopoulos’ model. Panels (a) and (b) plot $V_c$ and $T_c$, respectively, with respect to carbon number for $n$-alkanes. Note that the error bars are the 95% confidence intervals, whereas the error bars reported by Nath et al. represent the standard error.
To clarify, the results presented in Figure 3.6 were obtained by including all the simulation orthobaric densities for each compound reported by Nath et al. This is significant because the traditional lower bound for $T_c$ approaches a nonphysical value for $C_{36}$ and $C_{48}$ when compared with the range of temperatures simulated (see Ref [25]). Specifically, a nonphysical lower bound is observed with the traditional method for $C_{36}$ and $C_{48}$ at the 97% and 98% confidence levels, respectively. Therefore, although Figure 3.5 Panel (a) demonstrates an extreme case for ethane, the results of Figure 3.6 illustrate that for very large compounds this phenomenon is observed at a moderate confidence level even when all the data are included.

3.6 Conclusions

This chapter demonstrates two improvements for the statistical analysis of the critical point estimation from GEMC simulations. First, direct analysis of the rectilinear and scaling densities is necessary to estimate their corresponding standard deviation, rather than the traditional propagation of error approach. To facilitate future use, an error model was developed to predict the standard deviation for the liquid, vapor, rectilinear, and scaling densities. Second, the uncertainty in the parameters is decreased by performing a rigorous nonlinear analysis. In addition, we gain more insight into the acceptable range of $T_c$ and $\rho_c$ by scanning the parameter space and finding a joint confidence region. In brief, the uncertainty attributed to analyzing the simulation data is much lower than traditionally believed, and the regression process should no longer be viewed as a significant contribution to the uncertainty in predicting the critical point. Any inconsistency between the estimated critical constants from simulation and the experimental values should be attributed to a deeper layer in the uncertainty hierarchy (such as model uncertainty, parameter uncertainty, or finite-size effects).

Up to this point, the entire discussion has focused only on the numerical uncertainty associated with the inherent error in GEMC simulations and in the subsequent extrapolation to the critical point. As outlined in Chapter 1, the next layer of uncertainty propagation is the simulation limitations, where finite-size effects are most likely the largest contributor. Since Nath et al. did not correct for finite-size effects they qualified their results with the statement that “GEMC simulations suffer from significant size effects, particularly near the critical point. Our results should therefore be regarded with caution.” In other words, even though we have demonstrated
that Nath’s results have a stronger statistical certainty than originally reported, it is possible that these results deviate considerably from the true infinite-size system. Therefore, although we have mitigated the uncertainty associated with data analysis, more fundamental levels of error still need to be addressed. In Chapter 4 we verify the supposition of Nath et al. that finite-size effects are significant for the systems they simulated. In addition, we develop an improved methodology for predicting $P_c$ with simple (UA, LJ 12-6) force fields that mitigate finite-size effects and reduce uncertainties. Furthermore, in Chapter 5 we address the uncertainty associated with the LJ 12-6 force field parameters.
CHAPTER 4. IMPROVED ESTIMATES OF $T_c$, $\rho_c$, $P_c$ AND $Z_c$ WITH GEMC

4.1 Introduction

In Chapter 3 we detailed how to rigorously quantify the uncertainty in $T_c$ and $\rho_c$ due to the extrapolation of sub-critical GEMC simulations to the critical point. In this chapter, we demonstrate how to reduce the uncertainties in the critical point constants. Since GEMC simulation results typically have larger uncertainties than those from GCMC simulations, the reduction of the critical constant uncertainties is particularly important when utilizing GEMC simulations, as in this study. In particular, we are concerned with reducing both the uncertainty and the systematic error in $P_c$ and $Z_c$.

Furthermore, in this chapter we attempt to verify that finite-size scaling of the critical constant values is not necessary for a 200 molecule system of compounds as large as $C_{48}$. The literature is not clear as to the degree of finite-size effects for GEMC simulations of larger compounds. Specifically, Nath et al. suggested that finite-size effects may be significant for simulations of $C_{48}$ using 120 molecules [25]. By contrast, Dinpajooh et al. demonstrated that finite-size effects were negligible for a 200 molecule system of $n$-decane [22]. If system size dependence is indeed less for GEMC than GCMC, GEMC would be an attractive option to obtain reliable critical constants because the computational demand is much less than that for GCMC finite-size scaling methods.

The magnitude of the uncertainties and the degree of finite-size effects are the primary two concerns when estimating $T_c$ and $\rho_c$ with GEMC. The situation is even worse for $P_c$. Specifically, the traditional data analysis approach for predicting $P_c$ (and subsequently $Z_c$) from GEMC simulation data yields inaccurate results with large statistical uncertainties. This limitation is the primary reason why molecular simulation studies have not been able to elucidate the correct $P_c$ and $Z_c$ trends for large $n$-alkanes. We aim to obtain accurate predictions of each of the critical constants with correspondingly small uncertainties. The proposed methodology to achieve this goal is composed of two key aspects. First, as Vetere proposed [74], we utilize the Rackett equation with
liquid density \((\rho_l)\) to predict \(P_c\) more accurately than the traditional approach that uses the Antoine equation with vapor pressure \((P_v)\) [22]. Second, we implement an experimental design to predict the ideal temperatures for the GEMC simulations that will minimize the uncertainty in \(T_c, \rho_c, P_c,\) and \(Z_c\).

In addition to reducing the statistical uncertainty, we show that this experimental design is less susceptible to finite-size effects which have historically posed a serious limitation for predicting the critical point of large molecules (greater than about \(C_{26}\)) [75]. Since this experimental design can reduce uncertainty and finite-size effects without increasing computational time, its use makes GEMC simulation a viable option for predicting the critical constants of large compounds and should become a valuable tool in this regard. In brief, the proposed method is equally rigorous as the approach traditionally found in the literature yet it provides some significant advantages, especially for large compounds, as will be shown.

The purpose of this chapter is six-fold. The first is to demonstrate how the Vetere method provides more accurate estimates of \(P_c\) from GEMC simulations than the traditional approach. The second is to develop an experimental design to reduce the uncertainties in the critical constants. The third objective is to present improved estimates of \(T_c, \rho_c, P_c,\) and \(Z_c\) using different potential models for large \(n\)-alkanes. The fourth purpose is to demonstrate that by reducing the critical constant uncertainties, particularly for \(P_c\) and \(Z_c\), we are able to elucidate the long-chain-length \(P_c\) and \(Z_c\) trends for \(n\)-alkanes. The fifth goal is to resolve a conflict between the \(T_c\) values for large \(n\)-alkanes reported in the simulation literature for the NERD and Exp-6 models [23,25]. The sixth and final task is to verify that finite-size effects are reasonably small for the estimated critical constant values.

The outline for this chapter is the following. In Section 4.2 we explain the benefits of utilizing Vetere's method for predicting \(P_c\) from the Rackett equation and \(\rho_l\) simulation data. In Section 4.3 we explain how the uncertainty in \(P_c\) can be rigorously estimated from this method. Then, in Section 4.4 we demonstrate the need for a D-optimal experimental design to minimize the uncertainty in \(P_c\). In Section 4.5 the experimental design for GEMC simulations is derived in detail. Next, in Section 4.6 we present a quantitative comparison of the uncertainties in \(T_c, \rho_c,\) and \(P_c\) between the experimental design proposed in this work and those found in the literature. This is followed in Section 4.7 by a brief consideration of modifications to the experimental design.
the implications of this experimental design on finite-size effects, and possible limitations of this approach. A step-by-step outline of the proposed methodology is then presented in Section 4.8. This methodology is then implemented by performing molecular simulations with the systems described in Section 4.9. Then, in Section 4.10 we present simulation results at the D-optimal conditions for C\textsubscript{6}, C\textsubscript{8}, C\textsubscript{10}, C\textsubscript{12}, C\textsubscript{16} to validate the D-optimal design. Next, in Section 4.11 we present results for several potential models for C\textsubscript{24}, C\textsubscript{36}, and C\textsubscript{48}. Finally, we discuss our conclusions and how they have affected recommendations for predicting the critical constant trends of large n-alkanes.

4.2 Alternative Method to Predict \( P_c \)

A primary reason for the present work is to develop a more effective method for estimating \( P_c \) from GEMC simulations of larger compounds. The traditional approach to predicting \( P_c \) from GEMC simulation results makes use of the Antoine equation that relates vapor pressure (\( P_v \)) to temperature (\( T \)) [22]. The form typically used in the simulation literature is

\[
\ln(P_v) = A_0 + \frac{A_1}{T} \tag{4.1}
\]

where \( A_0 \) and \( A_1 \) are fitting parameters. The critical pressure is obtained by first regressing the \( P_v \) GEMC data to Equation 4.1. Then, \( T_c \) is obtained by regressing the GEMC results to the law of rectilinear diameters and density scaling law (Equations 3.1-3.2). Finally, \( P_c \) is calculated using \( T_c \) and the regression to Equation 4.1.

The traditional approach for predicting \( P_c \) is not ideal for larger compounds for several reasons. The first is that it is very difficult to achieve accurate vapor pressure results from molecular simulation. This is partially due to the fact that \( P_v \) is an inherently noisy property when obtained from GEMC simulations. For better statistics, more vapor phase molecules are required which necessitates an increase in the overall number of molecules. However, this comes at a great computational cost for longer chains (although the vapor phase is less expensive than the liquid phase due to fewer neighbors within the non-bonded cutoff distance).

Furthermore, even without increasing the number of molecules in the vapor phase, obtaining \( P_v \) becomes computationally intensive for larger molecules because of the “virial” force.
calculation. The virial forces require an individual calculation for each unique pair of sites in a simulation [21]. Thus, the calculation cost for the virial forces scales as $M^2$, where $M$ is the number of sites, not the number of molecules. Therefore, assuming the same number of molecules, the computational cost for the virial forces is about $24^2$ times more for C$_{48}$H$_{98}$ than ethane when using a united-atom model. By contrast with molecular dynamics, the virial forces are not necessarily computed in Monte Carlo simulations [21]. Therefore, the computationally expensive virial calculation can be completely eliminated by not calculating $P_v$.

The primary reason accurate $P_v$ values are difficult to obtain is because most intermolecular potential model functions, for example the Lennard-Jones 12-6 model (LJ 12-6), do not have the right form to accurately model both $\rho_l$ and $P_v$ [76]. For this reason, most potential models (e.g. OPLS, TraPPE, NERD, etc. [25, 27, 77]) are optimized to reproduce $\rho_l$ data. Because the extant models were not designed to yield accurate $P_v$ values the traditional approach for predicting $P_c$ from $P_v$ is inherently limited. On the other hand, the Exponential-6 model developed by Errington et al. [23] and the Mie 16-6 model developed by Potoff et al. [24] are capable of reliably predicting $P_v$ within 3% for $n$-alkanes ranging in carbon number between C$_2$-C$_{12}$. Despite this fact, both the Exp-6 and Mie 16-6 models, widely considered two of the best for $n$-alkanes, show systematic deviations in $P_c$, especially for larger compounds. This poor performance for $P_c$ is rationalized by Potoff et al. as a “result of compound errors in the predictions of critical temperatures and vapor pressures.” In other words, because of the exponential relationship between $P_v$ and $T$ in Equation 4.1, any deviation or uncertainty in $P_v$ and $T_c$ will be magnified for $P_c$. To our knowledge, there is not a force field model in existence that is capable of predicting both $P_v$ and $T_c$ to such a high degree of accuracy. Thus, although the extrapolation of $P_v$ to predict $P_c$ has a strong theoretical basis, the practical use of this approach is limited due to the inadequacies of the extant force field models.

The final limitation of the traditional method is that Equation 4.1 is often not reliable over a large temperature range. The traditional approach seeks to alleviate this deficiency by obtaining several $P_v$ values at temperatures near $T_c$. However, since fluctuations in simulation results increase near the critical point this only exacerbates the uncertainty in $P_v$. Therefore, although it should be expected that more accurate $P_v$ values from simulation will be achievable in the future, the shortcomings of the Antoine equation extrapolation will remain relevant. More flexible models
that utilize additional terms, such as the Riedel equation, do exist but they are traditionally not 
used considering the noise in the $P_v$ simulation results [78].

Since most potential models do well modeling $\rho_l$, Vetere’s approach to estimating $P_c$ is a 
logical alternative. This method rearranges the Rackett equation [79] to solve for $P_c$ according to

$$P_c = \frac{R_g T_c \rho_c}{M_w} \left( \frac{\rho_c}{\rho_l} \right) \left( 1 - \frac{T_i}{T_c} \right)^{-\gamma}$$  \hspace{1cm} (4.2)

where $R_g$ is the universal gas constant, $M_w$ is the molecular weight, $T_i$ is the temperature corre-
sponding to $\rho_l$, and $\gamma$ is an empirical parameter estimated to be around $\frac{2}{7}$. By developing a more 
reliable approach for predicting $P_c$ we can also obtain more accurate estimates for $Z_c$—the critical 
compressibility factor that is defined as

$$Z_c \equiv \frac{P_c M_w}{R_g T_c \rho_c}$$  \hspace{1cm} (4.3)

Although Vetere’s method has existed for nearly two decades, it has not been utilized with data 
from molecular simulations [74].

The Vetere approach to predict $P_c$ is not the most useful of methodologies when $T_c$ and $\rho_c$ 
are obtained from experiment. The reason is that it is much easier to accurately measure $T_c$ and 
$P_c$ than $\rho_c$. For example, Nikitin et al. did not report $\rho_c$ values for any of the large $n$-alkanes they 
studied [15, 20]. The situation is quite different for GEMC simulations where obtaining $T_c$ and $\rho_c$ 
from the density scaling law and the law of rectilinear diameters is straightforward. In addition, $\rho_l$ 
has a much lower inherent uncertainty than $P_v$ when obtained from GEMC simulations. Finally, 
while the traditional approach performs poorly for force field models that do not accurately predict 
$P_v$ or $T_c$, the Vetere method is capable of overcoming these model deficiencies. Therefore, the 
Vetere method is an ideal approach for predicting $P_c$ with the force field models utilized in this 
study (NERD [25], Exp-6 [23], TraPPE-UA (referred to simply as TraPPE) [27], and Mess-UP) as 
they predict $\rho_l$ more accurately than $P_v$. 

50
4.3 Estimating Uncertainty in $P_c$ and $Z_c$

The statistical uncertainty of $P_c$ and $Z_c$ predicted from Equations 4.2-4.3, respectively, depends on the uncertainty in $\rho_l$, $T_c$, and $\rho_c$. Here we will discuss how to propagate the uncertainty in $\rho_l$, $T_c$, and $\rho_c$ to $P_c$ and $Z_c$. In Chapter 3 we demonstrated how to estimate the uncertainty in $\rho_l$ by performing replicate simulations at the same temperature. However, the replicate simulation technique cannot be used to independently calculate the uncertainties of $\rho_l$, $T_c$, and $\rho_c$ because they are correlated. In other words, any deviation in $\rho_l$ leads to a deviation in $T_c$ and $\rho_c$ because $T_c$ and $\rho_c$ are regressed to $\rho_r$ and $\rho_s$ that depend upon $\rho_l$ (see Equations 3.1 and 3.2). To account for this interdependence the law of rectilinear diameters and the density scaling law may be combined to obtain

$$\rho_l = \rho_c + A(T_c - T) + \frac{B}{2}(T_c - T)^\beta$$

where $A$, $B$, and $\beta$ are the same parameters found in Equations 3.1 and 3.2. This expression is beneficial because the uncertainty in $\rho_l$ can be obtained from the uncertainties in the parameters $A$, $B$, $\rho_c$, and $T_c$, which are all fit at the same time and thus account for the correlation. Specifically, Equation 2.20 is used to calculate the probability density function (PDF) for the 4-dimensional parameter space of $A$, $B$, $\rho_c$, and $T_c$. Subsequently, $\rho_l$, $P_c$, and $Z_c$ are calculated by implementing Equations 4.4 and 4.2-4.3, respectively, for a given parameter set. As outlined in Chapter 2, the next step in the MCS propagation of errors is to randomly sample from this parameter space using the PDF to weigh the probability of each parameter set, $\theta$. By sampling millions of different parameter sets we generate a histogram of $P_c$ and $Z_c$ values that can be integrated at the desired confidence level. In short, Equations 2.20, 4.2-4.3, and 4.4 allow for a rigorous estimate of the uncertainty in a predicted value of $P_c$ and $Z_c$ obtained with the Vetere approach. In Figure 4.1 we have provided an example of the $P_c$ and $Z_c$ uncertainties for the same system used in Section 3.3.4.

4.4 Case Study: Long-Chain-Length Behavior of $n$-Alkanes

We have chosen the $n$-alkane family for a case study of the Vetere approach. There are two objectives for this section. First, to illustrate the effectiveness of the Vetere method for predicting $P_c$ from GEMC simulations. Second, to demonstrate the need for an optimal experimental design to reduce the uncertainty in $P_c$. For the former, the analysis will include using existing simulation data
Figure 4.1: Examples of the histograms produced from a Monte Carlo Sampling propagation of errors approach. Panels (a) and (b) correspond to $P_c$ and $Z_c$, respectively. Probability density is defined as the number of counts in a single bin divided by both the total number of counts and the bin width. The data used are the same as those in Figure 3.5 Panel (b).

but applying the Vetere method to predict $P_c$ rather than the Antoine equation. This comparison provides a rigorous test of the abilities of the method by eliminating biases that could be introduced by using simulation codes or methods different than those of the original authors.

In this case study we compare the $P_c$ results obtained from four different force fields (TraPPE [27], NERD [25], Exp-6 [23], and Mie 16-6 [24]). Nath et al. used GEMC simulations to predict $T_c$ and $\rho_c$ for $n$-alkanes up to C$_{48}$ using the NERD force field [25]; however, they did not report $P_c$ or even $P_v$ values. To analyze this force field, we used their orthobaric density results to calculate $P_c$ using Vetere’s method. We refer to these $P_c$ values as Nath-Vetere. Likewise, we predicted $P_c$ with Equation 4.2 for the TraPPE model by using the GEMC data reported by two different articles from Siepmann’s group, Martin et al. for C$_2$-C$_{12}$ and Zhuravlev et al. for C$_{30}$ [27, 28]. For simplicity, we refer to these values collectively as Siepmann-Vetere. In addition, we employed Vetere’s method with the orthobaric densities reported by Muller et al. to obtain the Muller-Vetere values for the NERD and TraPPE force fields [26]. The Siepmann-Vetere and Muller-Vetere results provide a useful benchmark when assessing the performance of Vetere’s
method because Martin et al., Zhuravlev et al., and Muller et al. also reported $P_c$ values that were obtained from the traditional method using Equation 4.1.

Figure 4.2 depicts the dependence of $P_c$ upon carbon number ($CN$) for $n$-alkanes. Panel (a) shows the data for chain-lengths up to $C_{20}$ and Panel (b) from $C_{15}$ to $C_{100}$ (the longest chain-length simulated by Muller et al.). Included in Figure 4.2 are experimental measurements from three different sources [12,15,20], three different empirical prediction models [9,14,15], molecular simulation results from five different sources [23,24,26–28], and the aforementioned Nath-Vetere, Siepmann-Vetere, and Muller-Vetere results.

Figure 4.2: The dependence of $P_c$ on $CN$. Panels (a) and (b) focus on the lower and higher $CN$ ranges, respectively. Three conflicting experimental data sets are included. Simulation results are as reported by Siepmann’s group (Martin et al. and Zhuravlev et al.), Potoff et al., Errington et al., and Muller et al. The Nath-Vetere, Siepmann-Vetere, and Muller-Vetere results are also shown. Panel (b) includes three different prediction models for the long-chain-length behavior. Error bars correspond to 95% confidence intervals. A slight offset in $CN$ for the simulation results is used for visual clarity.
There are several key conclusions from Figure 4.2. First, the Nath-Vetere results in Panel (a) appear to be equally reliable as those reported by Errington et al. and Potoff et al. This is significant because Errington et al. and Potoff et al. used respectively the Exp-6 and Mie 16-6 models (which reproduce $P_v$ more accurately than the LJ 12-6 model) with a more computationally demanding Grand Canonical Monte Carlo (GCMC) approach [23, 24]. We also observe a considerable improvement between the Muller-Vetere $P_c$ values and those reported by Muller et al., which utilized Equation 4.1 with $P_v$ values obtained from two-phase molecular dynamics. Furthermore, we see a slight improvement when comparing Siepmann-Vetere to the $P_c$ values reported by Siepmann’s group, which utilized $P_v$ GEMC results and Equation 4.1.

The results in Panel (a) demonstrate that simply changing the data analysis method can produce more accurate results. However, the data in Panel (b) demonstrate that the degree of improvement is not large enough to elucidate the correct trend of $P_c$ for larger n-alkanes. For example, the Nath-Vetere result for C$_{36}$ appears to support Nikitin’s old data whereas the Siepmann-Vetere value at C$_{30}$ agrees more strongly with the recent Nikitin data. It is unclear whether this discrepancy between the Nath-Vetere and Siepmann-Vetere results is due to differences in the NERD and TraPPE force fields because the Muller-Vetere results for the NERD and TraPPE force fields are in good agreement with each other and both agree with the most recent Nikitin data.

Furthermore, notice that the uncertainties in $P_c$ are too large to offer insight into which prediction model (that of Nannoolal et al. [9], Lemmon et al. [14], or Nikitin et al. [15]) correctly describes the long-chain-length behavior of the n-alkanes. For example, for C$_{24}$, the Nath-Vetere results seem to support either the Nikitin or Nannoolal model while those of Errington et al. more closely follow the Lemmon model. The situation is even worse for C$_{48}$ where the results from Errington et al. do not agree well with any prediction model and those of Nath-Vetere only moderately agree with Nannoolal’s model. Notably, the uncertainty for C$_{30}$ reported by Zhuravlev et al. spans nearly all three prediction models while the Siepmann-Vetere uncertainty is considerably smaller. However, minimizing the uncertainty further is necessary to give statistically-meaningful answers regarding the true behavior of $P_c$ for larger n-alkanes. Doing so requires that the uncertainty in $\rho_l$, $\rho_c$, and $T_c$ be reduced – something that changing the analysis method from the Antoine Equation to the Rackett Equation cannot address. In Section 4.5 we propose an experimental design that will minimize the uncertainty in $P_c$ by minimizing the uncertainty in $A$, $B$, $\rho_c$ and $T_c$. 

54
4.5 Optimal Experimental Design to Minimize the Uncertainty in $T_c$, $\rho_c$, and $P_c$

In this section we determine the optimal reduced temperatures ($T_r$) for performing the GEMC simulations in two different ways. First, a so-called D-optimal experimental design is developed to minimize the uncertainty in $T_c$, $\rho_c$, $P_c$, and $Z_c$. A D-optimal design has the advantage that the uncertainties for all three critical constants are reduced simultaneously. Second, a parameter-specific optimization is performed by finding the two temperatures that minimize the variance in just $T_c$ or $\rho_c$. This approach is ideal in the case where only one critical constant is of interest. For each of the experimental designs, we demonstrate that only two temperatures are needed while most GEMC studies perform simulations at several temperatures. To establish a common basis for the reader, we present the theory of D-optimal and parameter specific designs before deriving the different experimental designs for GEMC simulations.

4.5.1 General Theory of Experimental Designs

In general, an optimal experimental design determines the experimental (or, in this case, the simulation) conditions that provide the most precise estimates of the desired property. Although many different experimental designs exist, here we focus on designs that reduce the uncertainty in one or all parameters obtained from regression. This type of experimental design is ideal for the task at hand because our primary goal is to minimize the uncertainties in $T_c$, $\rho_c$, $P_c$, and $Z_c$ obtained from Equations 3.1-3.2 and 4.2-4.3. The uncertainties for $T_c$ and $\rho_c$ can be minimized by using a parameter-specific design for Equations 3.1-3.2 because they appear explicitly as regression parameters. The parameter-specific designs for $T_c$ and $\rho_c$ will be different and, therefore, we refer to these two designs as $T_c$-optimal and $\rho_c$-optimal designs. However, since $P_c$ depends upon $\rho_l$, $\rho_c$, and $T_c$ it also depends upon parameters $A$ and $B$ through Equation 4.4. Therefore, to minimize the uncertainty in $P_c$ a so-called D-optimal design is necessary to simultaneously reduce the uncertainty in $A$, $B$, $\rho_c$, and $T_c$. Since $Z_c$ depends on $T_c$, $\rho_c$, and $P_c$ a D-optimal design is also ideal for minimizing the uncertainty in $Z_c$. An essential assumption to the D-optimal and parameter-specific experimental designs is that the model is valid over the range considered. For this reason, in Section 4.12 we verify that the law of rectilinear diameters and density scaling law are reliable over a large temperature range.
D-optimal Design

A D-optimal design is a type of experimental design that minimizes the parameter imprecision by determining the optimal values for the independent variables [80]. The general approach for a D-optimal design is to find the independent variable values that maximize the determinant of the information matrix (which will be explained later). The complete theoretical derivation of the D-optimal design can be found in the literature [80]. For the present purposes it is sufficient to begin by defining the covariance matrix of the parameters as

\[
\text{COV}(b) = (F^T F)^{-1} S_D^2
\]  
(4.5)

where \(\text{COV}\) is the covariance matrix, \(b\) is a vector of parameters (in our case \(A, B, \rho_c,\) and \(T_c\)), \(F\) is the “F-matrix”, \(S_D^2\) is the inherent variance in the data, and \(F^T F\) is the information matrix. The F-matrix is obtained as follows. For a function, \(f\), that depends upon the vector of parameters, \(b\), and the independent variable, \(x\), the elements for the F-matrix are defined as

\[
F_{i,j} \equiv \frac{\partial f}{\partial b_j} \bigg|_{x_i}
\]  
(4.6)

where \(i\) and \(j\) correspond to the respective rows and columns in the F-matrix and \(|_{x_i}\) denotes evaluating the expression at \(x = x_i\).

Since a D-optimal design aims to reduce the uncertainty in all the parameters it must decrease each element of the covariance matrix. In the assumption of constant \(S_D^2\), the only way to minimize \(\text{COV}\) is to minimize \((F^T F)^{-1}\). Since the inverse of a matrix is inversely proportional to the determinant of that matrix, minimization of \((F^T F)^{-1}\) requires maximizing \(|F^T F|\). Hence the “D” refers to the determinant \(|F^T F|\) that is key to this analysis.

A simple example can help demonstrate the D-optimal method. Consider a linear model of the form

\[
y = b_1 + xb_2
\]  
(4.7)

where \(b_1\) is the y-intercept, \(b_2\) is the slope, \(x\) is the independent variable, and \(y\) is the dependent variable. Since there are only two parameters a D-optimal design will provide the best two values for \(x\) so as to minimize the uncertainty in \(b_1\) and \(b_2\). The F-matrix is found by applying Equation

56
4.6 with Equation 4.7 to yield

\[ F \equiv \begin{bmatrix} \frac{\partial y}{\partial b_1} |_{x_1} & \frac{\partial y}{\partial b_2} |_{x_1} \\ \frac{\partial y}{\partial b_1} |_{x_2} & \frac{\partial y}{\partial b_2} |_{x_2} \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \end{bmatrix} \]  

(4.8)

where \( x_1 \) and \( x_2 \) denote the two different values for \( x \). The information matrix is

\[ F^T F = \begin{bmatrix} 2 & x_1 + x_2 \\ x_1 + x_2 & x_1^2 + x_2^2 \end{bmatrix} \]  

(4.9)

and the determinant of the information matrix is

\[ |F^T F| = (x_1 - x_2)^2 \]  

(4.10)

Clearly, to maximize \(|F^T F|\), \( x_1 \) and \( x_2 \) must be as far apart as possible. This means that to minimize the uncertainty in the slope and intercept for a straight line model the best selection of independent values is \( \pm \infty \). Obviously, the physical constraints will determine how close \( x_1 \) and \( x_2 \) can approach \( \pm \infty \).

Visually, this analysis may be represented through a contour plot of \(|F^T F|\) where the two axis constitute \( x_1 \) and \( x_2 \). Such a plot is shown in Figure 4.3 where the blue region corresponds to the maximum in \(|F^T F|\). Notice that \(|F^T F|\) is largest when \( x_1 \) is small and \( x_2 \) is large, and vice versa. In other words, the greater the difference in \( x_1 \) and \( x_2 \) the greater the value of \(|F^T F|\). Also notice that contour plots of this nature should always be symmetric about the line \( x_1 = x_2 \) because the assignment of which \( x_i \) is greater is arbitrary.

In the case of a non-constant variance, the \( S_D^2 \) term in Equation 4.5 must be included in the information matrix. This is done by including a weighting matrix \( W \) such that the “weighted information matrix” for non-constant variance can be expressed as

\[ F^T W^T W F \]  

(4.11)

where \( W \) is a diagonal matrix with \( W_{i,i} = S_D^2 \). Since \( W F \) effectively acts as a new F-matrix, we define this as the “weighted F-matrix”. For our simple example of a straight-line model, the
Figure 4.3: A contour plot of the determinant of the information matrix $|F^T F|$ for the straight line model found using Equation 4.10. The best experimental design conditions correspond to the maxima which are at $x_1 = \pm \infty$ and $x_2 = \mp \infty$. The origin is located at the center.

The weighted F-matrix becomes

$$WF = \begin{bmatrix}
\frac{1}{S_{D_1}} & \frac{x_1}{S_{D_1}} \\
\frac{1}{S_{D_2}} & \frac{x_2}{S_{D_2}}
\end{bmatrix}$$

(4.12)

where $S_{D_i}$ represents the standard deviation evaluated at $x_i$.

The covariance matrix, $COV$, for non-constant variance is obtained from the inverse of Equation 4.11. Therefore, the D-optimal design is obtained by maximizing $|F^T W^T W F|$. As $S_D$ is included in the weighted F-matrix, a model for $S_D$ is necessary to develop an experimental design with a non-constant variance. Normally, this requirement might hinder a D-optimal design because an error model is not usually available a priori. For this reason, in Chapter 3 we developed an error model for GEMC simulations of $n$-alkanes that is accurate for a large range of compound sizes.

**Parameter-Specific Design**

An experimental design that is similar to a D-optimal design is a parameter-specific design, that is, an experimental design that minimizes the uncertainty in a specific parameter rather than the overall uncertainty in all of the parameters. This can be accomplished by minimizing the variance of the element of the covariance matrix, $COV$, that corresponds to that parameter. In other words,
to minimize the uncertainty in the parameter that corresponds to the \( j^{th} \) column of \( F \) we minimize the \( j^{th} \) main diagonal element of \((F^T F)^{-1}\) or of \((F^T W^T W F)^{-1}\).

To demonstrate this we will continue our example for the linear model. If we are only concerned about minimizing the uncertainty in \( b_1 \) we must minimize

\[
(F^T F)^{-1}_{1,1} = \frac{x_1^2 + x_2^2}{(x_1 - x_2)^2}
\]  

(4.13)

Likewise, if we want to minimize the error in \( b_2 \) we must minimize

\[
(F^T F)^{-1}_{2,2} = \frac{2}{(x_1 - x_2)^2}
\]

(4.14)

In this particular case the optimal design for \( b_1 \) is the same as that for \( b_2 \), namely \( \pm \infty \). Contour plots can also be used to demonstrate these results. In Figures 4.4 and 4.5 we present \( \frac{1}{(F^T F)^{-1}_{1,1}} \) and \( \frac{1}{(F^T F)^{-1}_{2,2}} \), respectively. Since we have plotted the inverse of \((F^T F)^{-1}_{j,j}\) the optimal values of \( x_1 \) and \( x_2 \) again correspond to the blue region. We see that the optimal selection of \( x_1 \) and \( x_2 \) is as far apart as possible for both parameters.

Figure 4.4: Contour plot of the inverse of the variance in the intercept \( \frac{1}{(F^T F)_{1,1}^{-1}} \) found using Equation 4.13. The best experimental design conditions correspond to the maxima which are at \( x_1 = \pm \infty \) and \( x_2 = \mp \infty \). The origin is located at the center.
Figure 4.5: Contour plot of the inverse of the variance in the slope $\frac{1}{(F^TF)^{-1}_{2,2}}$ found using Equation 4.14. The best experimental design conditions correspond to the maxima which are at $x_1 = \pm \infty$ and $x_2 = \mp \infty$. The origin is located at the center.

4.5.2 GEMC Experimental Design

In the specific case of performing an experimental design for GEMC simulations, the two equations used in regression are the law of rectilinear diameters and the density scaling law (Equations 3.1-3.2). Between these two coupled equations there are four parameters ($A, B, \rho_c,$ and $T_c$) and so it is common to think that four different simulation temperatures are required. However, since each simulation renders two data points ($\rho_r$ and $\rho_s$) two temperatures are sufficient. Therefore, our D-optimal design will provide the two temperature values that will maximize $|F^TW^TF|$. We will also develop an experimental design to minimize the uncertainty in either $T_c$ or $\rho_c$ (since these are the only two parameters with physical meaning in Equations 3.1-3.2) by minimizing the respective COV element (see Equation 4.5). It is important to reiterate that the $P_c$-optimal design is the same as the D-optimal design since minimizing the uncertainty in $P_c$ obtained from Equation 4.2 requires minimizing the uncertainty in $\rho_l$, $\rho_c$, and $T_c$. If $\rho_l$ is obtained from Equation 4.4 then $\rho_l$, and subsequently $P_c$, also depends upon the uncertainties in $A$ and $B$. Therefore, the $P_c$-optimal design should be the same as the D-optimal design that minimizes the uncertainty in $A$, $B$, $\rho_c$, and
 Likewise, the $Z_c$-optimal design should be the same as the D-optimal design since $Z_c$ depends on $T_c$, $\rho_c$, and $P_c$.

**GEMC Temperature Constraints**

It is common in an experimental design that the optimal value for an independent variable is unattainable due to physical or practical limits. This is the case for GEMC since there are limitations that affect the range of feasible temperatures for molecular simulations. The lowest feasible temperature is limited by the difficulty of a successful particle insertion step into the liquid phase. A successful insertion move is less probable when the liquid density and/or the molecular size increases. If too few insertion moves are accepted the two simulation boxes representing the liquid and vapor phases will not be in chemical equilibrium [21]. In addition, the lower constraint must consider the temperature range over which the law of rectilinear diameters and density scaling law are valid (see Sections 4.7.3 and 4.12). For our purposes, the lowest possible $T_r$ was assumed to be 0.7. On the other extreme, the highest possible $T_r$ was assumed to be 0.95. This upper constraint is imposed for two reasons. First, as the critical point is approached the two simulation boxes often alternate between the vapor and liquid phases [66], so that very long simulation times and more advanced analysis methods are required to obtain reliable estimates of $\rho_l$ and $\rho_v$ [22]. Second, near the critical point a crossover from Ising-like to mean field behavior causes the optimal value of $\beta$ in Equation 3.2 to change [21]. Fortunately, this is not observed for GEMC simulations of small $n$-alkanes below 0.95 $T_c$ [27].

We will show in Section 4.5.2 that when the variance is assumed to be constant with respect to temperature, the D-optimal design recommends performing simulations at these two extreme values to minimize the uncertainty in the prediction of $T_c$, $\rho_c$, $P_c$ and $Z_c$. However, in Section 4.5.2 we will also show that when the variance is not considered constant, as is the case with GEMC simulations, it is undesirable to perform simulations at the upper limit because the inherent error in the GEMC results increases exponentially with respect to reduced temperature. Therefore, the somewhat arbitrary choice of 0.95 as the upper bound does not affect any of the conclusions of this work. By contrast, the value of the lower bound does affect the experimental design. There are several cases where the $T_r$ lower bound of 0.7 may not be practical and a higher value will be necessary. For example, particle insertion is more difficult with non-linear molecules, such as
branched or cyclic compounds. In Section 4.7.1 we demonstrate how to account for variability in this lower bound value.

**Constant Variance**

A good starting point for a discussion of the proposed experimental design is the case of constant variance. In other words, if the uncertainty in the GEMC data were constant with respect to temperature what would be the optimal simulation temperatures? We can answer this question by deriving the D-optimal design conditions for the law of rectilinear diameters and density scaling law.

The first step in a D-optimal design is to construct the F-matrix. In this case, we must construct the F-matrix for Equations 3.1-3.2. When two equations are coupled together, as are Equations 3.1-3.2, the definition for the F-matrix is not as straightforward as for a single equation. The F-matrix for the law of rectilinear diameters and the density scaling law coupled together is

\[
F = \begin{bmatrix}
\frac{\partial \rho}{\partial r} & \frac{\partial \rho}{\partial r} & \frac{\partial \rho}{\partial A} & \frac{\partial \rho}{\partial B} \\
\frac{\partial \rho}{\partial c} & \frac{\partial \rho}{\partial c} & \frac{\partial \rho}{\partial A} & \frac{\partial \rho}{\partial B} \\
\frac{\partial \rho}{\partial c} & \frac{\partial \rho}{\partial c} & \frac{\partial \rho}{\partial T} & \frac{\partial \rho}{\partial T} \\
\frac{\partial \rho}{\partial c} & \frac{\partial \rho}{\partial c} & \frac{\partial \rho}{\partial T} & \frac{\partial \rho}{\partial T}
\end{bmatrix} = \begin{bmatrix}
1 & T_c - T_1 & A & 0 \\
1 & T_c - T_2 & A & 0 \\
0 & 0 & \beta B(T_c - T_1)^{\beta - 1} & (T_c - T_1)^\beta \\
0 & 0 & \beta B(T_c - T_2)^{\beta - 1} & (T_c - T_2)^\beta
\end{bmatrix}
\] (4.15)

where \( T_1 \) and \( T_2 \) are the lower and upper temperatures, respectively. From the F-matrix we can express the determinant of the information matrix as

\[
|F^TF| = \left( \beta BT_c^{2\beta} (T_{r_1} - T_{r_2})^2 (1 - T_{r_1})^{\beta - 1} (1 - T_{r_2})^{\beta - 1} \right)^2
\] (4.16)

where \( T_{r_1} \) and \( T_{r_2} \) are the reduced temperature values that correspond to the lower and upper temperatures, respectively. From this expression it may not be obvious which values of \( T_{r_1} \) and \( T_{r_2} \) maximize \( |F^TF| \). For this reason, in Figure 4.6 we present a contour plot of \( |F^TF| \) with respect to \( T_{r_1} \) and \( T_{r_2} \). (Recall that these plots are symmetric about the line \( T_{r_1} = T_{r_2} \) because the assignment of \( T_{r_2} > T_{r_1} \) is arbitrary.) The blue region corresponds to the maximum in \( |F^TF| \) and thus the optimal set of reduced temperature values. As seen in Figure 4.6, if the variance in the simulation
results is constant with respect to temperature the resulting D-optimal design suggests performing simulations at the two extreme \( T_r \) values, 0.7 and 0.95. Notice that this is already very different from most GEMC studies since the common assumption is that the low temperature data are not as useful when extrapolating to the critical point.

Figure 4.6: Contour plot of \( |F^T F| \) for the coupled law of rectilinear diameters and density scaling law with constant variance (see Equation 4.16). The best experimental design condition is \( T_{r1} = 0.7 \) and \( T_{r2} = 0.95 \).

The conclusions obtained from Figure 4.6 are best understood by linearizing both the law of rectilinear diameters and the density scaling law. The law of rectilinear diameters can be rewritten as a straight line according to

\[
\rho_r = (\rho_c + AT_c) - AT = b_1 + b_2 T
\]

where \( b_1 = \rho_c + AT_c \) and \( b_2 = -A \). Similarly, the linearized form of the density scaling law gives

\[
\rho_s^{\frac{1}{\beta}} = B^{\frac{1}{\beta}} T_c - B^{\frac{1}{\beta}} T = b_1 + b_2 T
\]

where \( b_1 = B^{\frac{1}{\beta}} T_c \) and \( b_2 = -B^{\frac{1}{\beta}} \). As discussed in Section 4.5.1, the D-optimal design for a straight-line model suggests simulating at the lower and upper constraints. Therefore, it seems reasonable...
that the D-optimal design with constant variance supports performing simulations at reduced temperatures of 0.7 and 0.95.

Though the derivations are not provided, the $T_c$-optimal and $\rho_c$-optimal designs also support using 0.7 and 0.95 for the $T_r$ values when a constant variance is assumed. This can be explained by noticing that $T_c$ and $\rho_c$ are only found in the intercept terms ($b_1$) in Equations 4.17-4.18. Recall that the parameter specific design for the intercept is to have $x_1$ and $x_2$ as far apart as possible (see Figure 4.4). Therefore, it can be expected that the $T_c$-optimal and $\rho_c$-optimal designs should be at the $T_r$ extrema.

**Non-constant Variance**

In Chapter 3 we demonstrated that the variance in GEMC simulation data is not constant with respect to temperature. In fact, the uncertainty in the rectilinear and scaling densities increase exponentially with respect to $T_r$. We will see that this exponential dependence is significant for the experimental designs.

To account for the non-constant variance the information matrix must include the weighted F-matrix (see Equation 4.11). For Equations 3.1-3.2 the weighted F-matrix becomes

$$WF = \begin{bmatrix} \frac{1}{S_{Dr}(T_1)} & \frac{T_c-T_1}{S_{Dr}(T_1)} & \frac{A}{S_{Dr}(T_1)} & 0 \\ \frac{1}{S_{Dr}(T_2)} & \frac{T_c-T_2}{S_{Dr}(T_2)} & \frac{A}{S_{Dr}(T_2)} & 0 \\ 0 & 0 & \beta B(T_c-T_1)^{\beta-1} S_{Dr}(T_1) & (T_c-T_1)^{\beta} S_{Dr}(T_1) \\ 0 & 0 & \beta B(T_c-T_2)^{\beta-1} S_{Dr}(T_2) & (T_c-T_2)^{\beta} S_{Dr}(T_2) \end{bmatrix}$$

(4.19)

where $S_{Dr}(T_i)$ and $S_{Ds}(T_i)$ are the standard deviations for $\rho_r$ and $\rho_s$, respectively, evaluated at $T_i$ with the corresponding parameters found in Table 3.1.

The D-optimal design for Equations 3.1-3.2 is obtained by finding the $T_1$ and $T_2$ values that maximize $|F^TW^TWF|$ where $WF$ is found from Equation 4.19. The parameter-specific design for $\rho_c$ and $T_c$ are the $T_1$ and $T_2$ values that minimize $(F^TW^TWF)^{-1}_{1,1}$ and $(F^TW^TWF)^{-1}_{3,3}$, respectively. Since analytical expressions for $|F^TW^TWF|$ and $(F^TW^TWF)^{-1}$ are cumbersome and provide little insight, we have provided contour plots of $|F^TW^TWF|$, $\frac{1}{(F^TF)_{1,1}^{-1}}$, and $\frac{1}{(F^TF)_{3,3}^{-1}}$ in Figures 4.7-
4.9, respectively. We have plotted \(\frac{1}{(F^TF)^{-1}_{j,j}}\) so that, in each case, the optimal design is found by maximizing the appropriate quantity. Therefore, the D-optimal, \(\rho_c\)-optimal, and \(T_c\)-optimal designs correspond to the blue region in Figures 4.7-4.9, respectively.

Figure 4.7: Contour plot of \(|F^TWF|\) where \(WF\) is defined by Equation 4.19. The D-optimal experimental design condition corresponds to the maximum at \(T_{r_1} = 0.7\) and \(T_{r_2} = 0.85\).

Figure 4.8: Contour plot of \(\frac{1}{(F^TWF^TW)^{-1}}\) where \(WF\) is defined by Equation 4.19. The \(\rho_c\)-optimal experimental design condition corresponds to the maximum at \(T_{r_1} = 0.7\) and \(T_{r_2} = 0.84\).
There are two main implications from these figures. First, we see that all three designs support having $T_{r_1}$ at the lower constraint of 0.7. Second, in all three designs the value of $T_{r_2}$ is far removed from the upper constraint of 0.95. Specifically, the D-optimal, $\rho_c$-optimal, and $T_c$-optimal designs result in a $T_{r_2}$ value of 0.85, 0.84, and 0.89, respectively. These three designs are very different from conventional wisdom that places less emphasis on the low temperature regime. Indeed, these results seem counter intuitive since one might think that the closer the simulations are to the critical point the more beneficial they will be due to reducing the degree of extrapolation needed. However, because the error increases exponentially with respect to $T_r$ for GEMC orthobaric densities, simulations near the upper limit actually become less beneficial. We will substantiate these conclusions in the following section.

### 4.6 Comparison of Experimental Design with Literature

In this section, we compare the uncertainties in $T_c$, $\rho_c$, and $P_c$ using the three experimental designs explained above (D-optimal, $\rho_c$-optimal, and $T_c$-optimal designs) with two designs that are more representative of those found in the literature [25, 72]. The errors in the three critical constants for each of the five methods are presented in Figure 4.10. The “percent uncertainty” in
Figure 4.10 corresponds to the two-sided 95% confidence interval divided by the best estimate of the critical property multiplied by 100%. The uncertainties displayed in Figure 4.10 were obtained by performing a Monte Carlo sampling propagation of errors analysis for $\rho_c$, $T_c$, and $P_c$. For each design, $P_c$ is obtained from Equation 4.2. The standard deviations needed for the Monte Carlo analysis came from the error models developed in our previous study to predict $S_{D_r}$ and $S_{D_s}$. As such, the quantitative values in Figure 4.10 are only valid when the simulation conditions are the same as those used to develop the error models, namely, a GEMC simulation with 200 molecules, roughly 20% in the vapor phase, and the same move probabilities (see Section 2.1.6). That being said, it is expected that the qualitative nature of Figure 4.10 is maintained for different system specifications.

Each “literature” design consists of four equally spaced temperatures between a maximum temperature of $T_r = 0.95$ and a minimum temperature. The two “literature” cases differ in the value used for the minimum temperature – $T_r = 0.8$ or 0.9. The lower $T_r$ of 0.8 is more indicative of what the common practice was prior to 2013 while the value of 0.9 follows the more recent practice of only including simulations with $T_r \geq 0.9$ in the regression to the law of rectilinear diameters.
and the density scaling law. The pre-2013 literature design uses only a single simulation at each temperature, whereas the post-2013 literature design uses eight replicates at each temperature [25, 72]. By contrast, the three experimental designs developed in this work perform two replicates at two temperatures. Therefore, the uncertainties displayed in Figure 4.10 for the D-optimal, \( \rho_c \)-optimal, and \( T_c \)-optimal designs use the same number of simulations as the pre-2013 literature and only \( \frac{1}{8} \) as many simulations as the post-2013 literature. If more replicate simulations are performed at each temperature the uncertainties will be less than those reported in Figure 4.10.

Several significant conclusions may be drawn from Figure 4.10. First, the D-optimal design performed the best for \( P_c \) (producing the lowest uncertainty) and was nearly the best for \( \rho_c \) and \( T_c \). The \( \rho_c \) and \( T_c \) performance is surprising as Designs 2 ( \( \rho_c \)-optimal) and 3 ( \( T_c \)-optimal) specifically minimize the errors in these two parameters. Second, Figure 4.10 shows that the uncertainty in \( T_c \) is nearly the same for all five of the experimental designs while the uncertainty in \( \rho_c \) and \( P_c \) are very dependent on the type of experimental design. Third, these results are consistent with what Panagiotopoulos originally reported when he stated that “critical densities obtained from the rectilinear diameter rule are subject to significantly larger uncertainties than the critical temperatures” [81]. Finally, it is worth restating that the post-2013 literature uses eight times as many simulations as the D-optimal, \( \rho_c \)-optimal, and \( T_c \)-optimal experimental designs but the uncertainty in \( \rho_c \) and \( P_c \) are still much greater. This demonstrates that to lower the uncertainty in \( \rho_c \) from the law of rectilinear diameters, it is better to use lower temperatures.

In conclusion, the D-optimal experimental design greatly improves the uncertainty in \( P_c \) and \( \rho_c \) without sacrificing precision in \( T_c \) and demands fewer computational resources than the current practice. Therefore, our recommendation is to perform several simulations at reduced temperatures of 0.7 and 0.85, the D-optimal conditions. Although in some cases \( T_c \) might be the only parameter of interest, the improvement in \( T_c \) when using the \( T_c \)-optimal design is minimal while the depreciation in \( \rho_c \) and \( P_c \) is significant. The \( \rho_c \)-optimal design is very similar to the D-optimal design and, therefore, it is a reasonable alternative. In any case, to reduce the uncertainties in the critical constants it is more beneficial to have replicates at two temperatures rather than the common practice of spreading out the replicates among several equally dispersed temperature values. (However, as discussed in Section 4.7.3, simulations at additional temperatures should be performed to validate the assumed models, i.e. the law of rectilinear diameters and the density scaling.
Furthermore, the $T_c$-optimal, $\rho_c$-optimal, and D-optimal experimental designs suggest that simulations near the critical point are not necessary for precise estimates of the critical point constants. This is important because these simulations require longer computer time to obtain reliable averages due to the high degree of fluctuations when the simulation temperature approaches the critical temperature.

### 4.7 Considerations

#### 4.7.1 Possible Modifications to Experimental Designs

In Section 4.5.2 we explained why we assumed that the lower temperature constraint was $T_r = 0.7$. However, in some cases the value of this constraint may change which would not only cause $T_{r1}$ to change but also $T_{r2}$. The optimal value for $T_{r1}$ will always equal that of the lower constraint but the value of $T_{r2}$ must be obtained by maximizing $|F^T W^T W F|$, $\frac{1}{(F^T F)^{-1}_{1,1}}$, or $\frac{1}{(F^T F)^{-1}_{3,3}}$.

To demonstrate the effect of a different value for the lower constraint, in Figure 4.11 we have plotted the dependence of $T_{r2}$ on $T_{r1}$ for the D-optimal, $\rho_c$-optimal, and $T_c$-optimal designs. The most important point from this figure is that the optimal upper temperature value is typically much lower than the upper constraint of 0.95 $T_c$ regardless of the value of $T_{r1}$. The significance of this will be shown in Section 4.7.2 concerning finite-size effects. In brief, since finite-size effects increase near the critical point, selecting the lowest practical value for $T_{r1}$ will reduce $T_{r2}$ and thereby help to mitigate finite-size effects.

With a different value for the lower constraint, and thus for $T_{r1}$, the uncertainty in $T_c$, $P_c$, and $\rho_c$ will change. In Figure 4.12 we plot the uncertainty in $T_c$, $P_c$, and $\rho_c$ with respect to $T_{r1}$. These uncertainties were obtained in the same was as those in Figure 4.10, namely from a Monte Carlo sampling propagation of errors. Since we propose use of the D-optimal design, the uncertainties shown in Figure 4.12 correspond to the D-optimal design. In other words, for each value of $T_{r1}$ the value for $T_{r2}$ was determined from the D-optimal design curve in Figure 4.11. Notice that the uncertainties in $T_c$, $P_c$, and $\rho_c$ increase with increasing values of $T_{r1}$. The increase in uncertainty is most pronounced for $P_c$ while for $T_c$ the error remains relatively small.
Figure 4.11: Dependence of $T_r_2$ on $T_r_1$ for the three different experimental designs. The value for the lower reduced temperature is determined by the GEMC limitation for particle insertion of high density liquids.

Figure 4.12: Effect of $T_r_1$ upon the uncertainty in $T_c$, $\rho_c$, and $P_c$ when using the D-optimal design (see Figure 4.11 for the values used for $T_r_2$). $P_c$ was obtained using Equation 4.2. The percent uncertainty for $\rho_c$, $T_c$, and $P_c$ are presented at the 95% confidence level.
4.7.2 Implications on Finite-Size Effects

Finite-size effects arise as the ratio of molecules on the periodic boundary of a box to total molecules increases [21]. This effect is prevalent in simulations of larger compounds \((CN > 26)\) because as the size of the compound increases the number of molecules required to alleviate finite-size effects increases while the amount of molecules that can be simulated in a feasible length of time decreases considerably. Also, it is a well established fact that finite-size effects become more significant as \(T\) approaches \(T_c\) [21, 25, 81]. Fortunately, the D-optimal design suggests that it is unnecessary to perform simulations so close to the critical point.

The initial objective for performing an experimental design was simply to reduce the statistical uncertainty of the parameters in Equations 3.1 and 3.2 and thereby reduce the statistical uncertainties in \(T_c, \rho_c, P_c,\) and \(Z_c\). In other words, typically an experimental design only reduces the uncertainty associated with regressing the parameters. However, as explained previously, the experimental designs presented can also be used to reduce finite-size effects. Although it is unclear whether the D-optimal design significantly reduces finite-size effects, it certainly does not contribute to them. Indeed, the amount to which finite-size effects are decreased still depends upon the size of the compound and number of molecules used in simulation.

It is important to mention that the error model mentioned previously was developed using a relatively small number of molecules (200) so that it should be accurate for the size of simulations most common for large molecules. Therefore, the results presented previously in Figures 4.10, 4.11, and 4.12 are quantitatively valid for the size of systems typically used for larger molecules. In Section 4.10 we validate the performance of the D-optimal design by performing GEMC simulations with 200 molecules at the D-optimal design conditions for small \(n\)-alkanes. Furthermore, in Section 4.11.1 we verify that finite-size effects are reasonably small for long-chain-length compounds when using the D-optimal design.

4.7.3 Possible Limitations

Although we have discussed some significant advantages for this methodology, in this section we discuss some possible limitations. For example, the proposed D-optimal design does not consider whether the law of rectilinear diameters and the density scaling law are valid at low tem-
temperatures. In fact, one of the primary reasons why the practice found in the most recent literature excludes low temperature data is to avoid potentially poor performance of the law of rectilinear diameters [22]. Therefore, it is important to demonstrate that Equations 3.1 and 3.2 are reliable over the entire temperature range where simulations are performed. For this reason, we recommend performing a few validation simulations at temperatures other than those from the D-optimal design. For example, this validation can be achieved easily by performing a single simulation at $T_r = 0.75, 0.8, 0.9$ and 0.95. Specifically, if a plot of $\rho_r$ and $\rho_c^{1/\gamma}$ with respect to $T$ yields a straight line then Equations 3.1-3.2 are valid over the temperature range in question. As we will see in Section 4.12, we have found this to be the case for $T_r$ as low as 0.7 for $n$-alkanes.

Another issue that has not been discussed in detail until now is the fact that $T_r$ is not known \textit{a priori} because $T_r$ depends upon the value of $T_c$. This means that determining the temperature for which $T_r = 0.7$ and $0.85$ would be an iterative process where a new estimate for $T_c$ would be obtained after each iteration. At first glance this appears to be problematic; however, this is of little practical importance as $T_c$ can be estimated reasonably well with several prediction models [8]. So the question is simply how sensitive is the experimental design to variations in $T_r$? Figures 4.7-4.9 show how much the uncertainties depend on locating the exact optimal values. When the contours are closer together the uncertainty increases more rapidly with respect to $T_r$. This can be visualized more easily by slicing through a contour at a fixed value of $T_{r1}$. Figure 4.13 is provided to demonstrate how much the 95% confidence intervals in $T_c$, $\rho_c$, and $P_c$ increase with respect to $T_{r2}$ ($T_{r1}$ is assumed to be 0.7). The results in Figure 4.13 were obtained in the same way as those in Figure 4.10. This figure demonstrates that locating the exact value for $T_{r2}$ is not crucial because the wells for $\rho_c$, $T_c$, and $P_c$ are fairly wide. When we compare Figure 4.13 with Figure 4.12, we see that having $T_{r1}$ as low as feasibly possible is more important than having $T_{r2}$ at the exact design conditions.

Another consideration is that when the Rackett approach was originally proposed by Vetere he stated that it was inadequate for compounds with $CN > 12$. However, Vetere’s article was published in 1988, two years before Teja’s experimental critical point data became available for $n$-alkanes ranging from C$_{12}$-C$_{18}$ [12]. When we repeated Vetere’s analysis with Teja’s data the results did not show any depreciation with increasing chain length. However, the fact remains that the Rackett exponent, $\gamma$, is simply an empirical parameter. Vetere demonstrated that $\gamma$ is not a
universal constant for all compounds and, in fact, it is not even constant with respect to temperature for a given compound [82]. A sensitivity analysis of Equation 4.2 predicts that $P_c$ will depend less on $\gamma$ when \((1 - \frac{T_i}{T_c}) \approx 1\). Therefore, based solely on the Rackett equation, $T_i$ in Equation 4.2 should be as low as possible to mitigate the effect of $\gamma$ on the estimate for $P_c$. Fortunately, our experimental design is consistent with this concept of using temperatures far removed from $T_c$. In addition, the uncertainty in $P_c$ is lowest when $T_i$ is much lower than $T_c$ because the uncertainty in $\rho_l$ decreases exponentially with respect to temperature.

Furthermore, the algorithm for estimating the uncertainty in $P_c$ outlined in Section 4.3 and the results presented in Figures 4.10, 4.12, and 4.13 correspond only to the statistical uncertainty. In other words, we have neglected possible systematic deviations due to $\gamma$ and $\beta$ deviating from $\frac{2}{7}$ and 0.326, respectively. Fortunately, our experience is that possible systematic deviations are on the order of 1%, which is usually well within the 95% statistical uncertainties for larger compounds. This was validated by varying $\gamma$ between 0.275-0.295 (the range Rackett originally reported [79]) and $\beta$ between 0.32 and 0.33.

Finally, as we mentioned in Section 4.5.2, the error model presented in Chapter 3 is pivotal to the experimental designs. Since this error model was developed assuming a constant set of
Monte Carlo move probabilities it is reasonable that different sets of moves would yield a different
dependence of $S_D$ on $T_r$. Indeed, some studies have attempted to determine the dependence of
$S_D$ on the Monte Carlo move probabilities [50]. Therefore, an alternative approach to what we
have presented for reducing the uncertainties in $T_c$, $\rho_c$, $P_c$, and $Z_c$ is to optimize the Monte Carlo
move probabilities at different temperatures. However, in Section 4.5.2, we observed that even if
the error were constant with respect to temperature the experimental designs would still support
performing simulations at the lower bound. For this reason, future attention should be given to
optimizing the Monte Carlo move probabilities in both the high and low temperature regimes.

4.8 Outline of Methodology

In this section we provide an outline of the methodology to assist the reader:

1. Estimate $T_c$ from previous simulation results or a prediction model.

2. Determine the lowest practical simulation temperature ($T_1$) with $T = 0.7 \times T_c$ as an initial
   estimate.

3. Obtain $T_{r2}$ from the D-optimal design curve in Figure 4.11 and calculate $T_2 = T_{r2} \times T_c$.

4. Perform several replicate GEMC simulations at $T_1$ and $T_2$ to obtain $\rho_l$ and $\rho_v$.

5. Regress Equations 3.1-3.2 to the GEMC data to obtain $A$, $B$, $\rho_c$, and $T_c$.

6. Calculate $\rho_l$ from Equation 4.4 with $T_i = T_1$.

7. Use Equation 4.2 to calculate $P_c$ from $\rho_c$, $T_c$, and $\rho_l$ obtained in Steps 5 and 6.

8. Use Equation 4.3 to calculate $Z_c$ from $\rho_c$, $T_c$, and $P_c$ obtained in Steps 5 and 7.

9. Follow the uncertainty algorithm outlined in Section 4.3.

4.9 Simulation Specifications

In the remainder of this chapter we demonstrate how the Vetere method and D-optimal
design can provide improved estimates of the critical constants. Specifically, we simulated $C_6$, $C_8$,
with the NERD [25] and TraPPE [27] models. Furthermore, we performed simulations with the Exp-6 model [23] for C_{24} and C_{48}. This was done in an attempt to quantify finite-size effects for larger compounds by comparing our results with those reported for C_{24} and C_{48} by Errington et al. who used a more precise GCMC histogram reweighting method. Furthermore, by comparing the LJ 12-6 based models (TraPPE and NERD) with the Exp-6 model we can assess the need for a more precise intermolecular model function.

We have provided a summary of the simulation specifications in Table 4.1. In this table we have included each compound that was simulated, the force field model used, the total number of molecules between the two simulation boxes (approximately 20% in the vapor phase), the two temperatures (D-optimal conditions), the number of replicates at each temperature, and the number of Monte Carlo cycles in the equilibration and production periods.

As it was previously demonstrated that GEMC simulations with 200 molecules for n-decane do not display significant finite-size effects [22], we limited the system size to 200 molecules for C_{6} to C_{16}. To assist in our analysis of finite-size effects for the longer chain-lengths, we performed simulations with 200 and 1600 molecules for C_{24}, C_{36}, and C_{48}. Due to inordinate computational costs, the equilibration and production periods for the 1600 molecule systems used fewer Monte Carlo cycles (although the number of Monte Carlo moves is approximately the same). Although we verified that the equilibration periods were sufficiently long, even for C_{48}, the resulting uncertainties in the 1600 molecule simulations are much larger for C_{48} than for the shorter molecules because of the relatively short production periods. The implications of these increased uncertainties are discussed later.

An initial estimate for T_{c} is required to calculate T_{1} = 0.7 \times T_{c} and T_{2} = 0.85 \times T_{c} (the D-optimal design conditions). The initial estimate for T_{c} was obtained from previous studies for the NERD, Exp-6, and TraPPE models [23, 25, 27]; however, since reliable simulations of C_{24}, C_{36}, and C_{48} have not been performed for the TraPPE model, we estimated the TraPPE T_{c} to be the same as that found in the literature for the NERD model. These initial approximations for T_{c} do not affect the final values reported for T_{c}, \rho_{c}, P_{c}, and Z_{c}. We did not use the T_{c}-optimal or \rho_{c}-optimal designs because the improvement in the uncertainty in T_{c} and \rho_{c} is small when compared with the D-optimal design uncertainties (see Figure 4.10). It is important to note that we used the same
Table 4.1: Summary of compounds, force field models, total number of molecules \((N)\), state points (temperatures \((T_1\) and \(T_2\)), number of replicates at each temperature \((n)\), and the number of Monte Carlo cycles in the equilibration \((MC_E)\) and production \((MC_P)\) periods.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Model</th>
<th>(N)</th>
<th>(T_1) (K)</th>
<th>(T_2) (K)</th>
<th>(n)</th>
<th>(MC_E)</th>
<th>(MC_P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n)-Hexane ((C_6H_{14}))</td>
<td>NERD</td>
<td>200</td>
<td>360</td>
<td>440</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Hexane ((C_6H_{14}))</td>
<td>TraPPE</td>
<td>200</td>
<td>360</td>
<td>440</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Octane ((C_8H_{18}))</td>
<td>NERD</td>
<td>200</td>
<td>390</td>
<td>490</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Octane ((C_8H_{18}))</td>
<td>TraPPE</td>
<td>200</td>
<td>390</td>
<td>490</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Decane ((C_{10}H_{22}))</td>
<td>NERD</td>
<td>200</td>
<td>435</td>
<td>535</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Decane ((C_{10}H_{22}))</td>
<td>TraPPE</td>
<td>200</td>
<td>435</td>
<td>535</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Dodecane ((C_{12}H_{26}))</td>
<td>NERD</td>
<td>200</td>
<td>465</td>
<td>575</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Dodecane ((C_{12}H_{26}))</td>
<td>TraPPE</td>
<td>200</td>
<td>465</td>
<td>575</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Hexadecane ((C_{16}H_{34}))</td>
<td>NERD</td>
<td>200</td>
<td>510</td>
<td>625</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Hexadecane ((C_{16}H_{34}))</td>
<td>TraPPE</td>
<td>200</td>
<td>510</td>
<td>625</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Tetracosane ((C_{24}H_{50}))</td>
<td>NERD</td>
<td>200</td>
<td>580</td>
<td>700</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Tetracosane ((C_{24}H_{50}))</td>
<td>TraPPE</td>
<td>200</td>
<td>580</td>
<td>700</td>
<td>3</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Tetracosane ((C_{24}H_{50}))</td>
<td>Exp-6</td>
<td>200</td>
<td>560</td>
<td>680</td>
<td>5</td>
<td>200,000</td>
<td>200,000</td>
</tr>
<tr>
<td>(n)-Tetracosane ((C_{24}H_{50}))</td>
<td>Exp-6</td>
<td>1600</td>
<td>560</td>
<td>680</td>
<td>5</td>
<td>50,000</td>
<td>100,000</td>
</tr>
<tr>
<td>(n)-Hexatriacontane ((C_{36}H_{74}))</td>
<td>NERD</td>
<td>200</td>
<td>650</td>
<td>765</td>
<td>5</td>
<td>200,000</td>
<td>400,000</td>
</tr>
<tr>
<td>(n)-Hexatriacontane ((C_{36}H_{74}))</td>
<td>NERD</td>
<td>1600</td>
<td>650</td>
<td>765</td>
<td>5</td>
<td>30,000</td>
<td>70,000</td>
</tr>
<tr>
<td>(n)-Hexatriacontane ((C_{36}H_{74}))</td>
<td>TraPPE</td>
<td>200</td>
<td>650</td>
<td>775</td>
<td>5</td>
<td>200,000</td>
<td>400,000</td>
</tr>
<tr>
<td>(n)-Hexatriacontane ((C_{36}H_{74}))</td>
<td>TraPPE</td>
<td>1600</td>
<td>650</td>
<td>775</td>
<td>5</td>
<td>30,000</td>
<td>70,000</td>
</tr>
<tr>
<td>(n)-Octatetracontane ((C_{48}H_{98}))</td>
<td>Exp-6</td>
<td>200</td>
<td>700</td>
<td>800</td>
<td>5</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>(n)-Octatetracontane ((C_{48}H_{98}))</td>
<td>Exp-6</td>
<td>1600</td>
<td>700</td>
<td>800</td>
<td>5</td>
<td>20,000</td>
<td>40,000</td>
</tr>
<tr>
<td>(n)-Octatetracontane ((C_{48}H_{98}))</td>
<td>NERD</td>
<td>200</td>
<td>730</td>
<td>830</td>
<td>5</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>(n)-Octatetracontane ((C_{48}H_{98}))</td>
<td>TraPPE</td>
<td>200</td>
<td>730</td>
<td>830</td>
<td>5</td>
<td>200,000</td>
<td>300,000</td>
</tr>
<tr>
<td>(n)-Octatetracontane ((C_{48}H_{98}))</td>
<td>TraPPE</td>
<td>1600</td>
<td>730</td>
<td>830</td>
<td>5</td>
<td>20,000</td>
<td>40,000</td>
</tr>
</tbody>
</table>

Monte Carlo move probabilities as those used to develop our error model in Chapter 3 so that the D-optimal design conditions would be valid.

In the discussion that follows we compare our simulation results with 1) experimental data, 2) prediction models, and 3) existing simulation literature. We first focus on validating the D-optimal design approach by simulating the smaller \(n\)-alkanes, for which existing simulation results can be found in the literature for the force field models in question. We then direct our attention to the larger compounds which have fewer reliable simulation data. The majority of the discussion

76
will focus on these larger systems since a primary goal of this study is to obtain more accurate estimates of the critical constants for large \( n \)-alkanes. For reference, see Tables B.1, B.3 and C.1 for the vapor-liquid coexistence data and critical constant values of each system under consideration.

### 4.10 Validation of D-optimal Design

The primary purpose of this section is to demonstrate that the predicted \( T_c, \rho_c \), and \( P_c \) values obtained with the D-optimal design are reliable (i.e. accurate with small uncertainties). For this purpose, we simulated \( n \)-alkanes with a \( CN \) range between 6 and 24 with the NERD and TraPPE models, and compared these results with those found in the literature. Specifically, our results for \( T_c \) and \( \rho_c \) can be directly compared to those reported by Nath et al. [25] and Martin et al. [27] because they were obtained through the same data analysis method, namely by using Equations 3.1-3.2. We have also compared our \( P_c \) results to the Nath-Vetere and Siepmann-Vetere values (see Section 4.4). By using the same data analysis method for \( P_c \) on both our data and those found in the literature, we are able to solely investigate the performance of the D-optimal design.

Figures 4.14 and 4.15 provide a comparison between our \( T_c \) and \( \rho_c \) results and those found in the literature. In each of these graphs, the critical property is plotted as a function of carbon number. The uncertainties for the literature values of the TraPPE model are those reported by Martin et al. However, the uncertainties for Nath et al. are those obtained from the rigorous nonlinear analysis found in Section 3.5. The key conclusion of Figure 4.14 is that our \( T_c \) values are consistent with those reported in the literature. As we discussed in our previous study, the uncertainty in \( T_c \) is not significantly reduced with the D-optimal design. However, in Figure 4.15 we see that, in general, the D-optimal design greatly reduces the uncertainty for \( \rho_c \), as predicted. Furthermore, our \( \rho_c \) results are consistent with the literature values for most of the smaller \( n \)-alkanes.

In Figure 4.16 we compare our \( P_c \) results for the NERD and TraPPE models with the Nath-Vetere and Siepmann-Vetere values. Again, this comparison provides insight into the performance of the D-optimal design because the literature values were obtained from the same potential models and used the same data analysis method as our results, i.e. Equation 4.2. There are three significant conclusions from Figure 4.16. First, our estimates for \( P_c \) agree well with the experimental data. Second, our results for the NERD model agree with the Nath-Vetere values and our results for the
Figure 4.14: The trend for $T_c$ with respect to $CN$ for lower $CN$. Three experimental data sets are included [12, 15, 20]. The values reported by Nath et al. [25] and Martin et al. [27] are provided for comparison with our simulation results using the NERD and TraPPE models with a D-optimal design. Error bars correspond to 95% confidence intervals. A slight offset in $CN$ for the simulation results is used for visual clarity.

The TraPPE model agree well with the Siepmann-Vetere values. Finally, by using a D-optimal design, the corresponding uncertainties are small enough to provide quantitative insight (which will be important for the discussion of large $n$-alkanes). Notice that our uncertainties are typically much smaller than those for Nath-Vetere and Siepmann-Vetere.

4.11 Results for $C_{24}$, $C_{36}$, and $C_{48}$

Having demonstrated that the D-optimal design predicts properties well for smaller $n$-alkanes the remainder of the discussion focuses on larger $n$-alkanes. First, we analyze $T_c$ and $\rho_c$ as this provides a useful comparison with existing simulation results because our data analysis method for these two properties is the same as that found in the literature (Equations 3.1 and 3.2). To provide further insight, we also implemented the rigorous uncertainty analysis to obtain a joint confidence region for $T_c$ and $\rho_c$. Our $T_c$ results resolve a conflict in the simulation literature as to which data and prediction models are most reliable for large $n$-alkanes. We will not discuss
in detail our results for $\rho_c$ as they are consistent with those previously reported in the simulation literature [23, 25]. In short, Teja’s $\rho_c$ model is more accurate than that of Tsonopoulos for large $n$-alkanes [12, 13].

After the description of the results for $T_c$ and $\rho_c$, we shift our attention to $P_c$ to demonstrate the strength of using the Rackett equation and a D-optimal design for predicting $P_c$ both accurately and precisely. *The results suggest a definitive answer as to which experimental data and prediction models are most reliable for $P_c$ of large $n$-alkanes.* Finally, we present our $Z_c$ values with respect to $CN$ to validate the internal consistency of our $T_c$, $\rho_c$, and $P_c$ results. Our $Z_c$ results also help discern between the different prediction models and their trends with increasing chain-length.

### 4.11.1 $T_c$ and $\rho_c$ Results for Larger $n$-Alkanes

Figure 4.17 contains an uncertainty analysis from several different methods for estimating $T_c$ and $\rho_c$. Panel (a) is for $C_{24}$, Panel (b) for $C_{36}$, and Panel (c) for $C_{48}$. The ordinate is the criti-
Figure 4.16: The trend for $P_c$ with respect to $CN$ for lower $CN$. Three conflicting experimental data sets are included. The Nath-Vetere and Siepmann-Vetere results were obtained by analyzing the values from Nath et al. and from Martin et al. with Equation 4.2. Included are simulation results from this work using the NERD and TraPPE models with a D-optimal design. Error bars correspond to 95% confidence intervals. A slight offset in $CN$ for the simulation results is used for visual clarity.

critical temperature while the abscissa is the critical density. The results are presented as uncertainty regions at the 95% confidence level. The confidence region for the experimental data is presented as two lines rather than boxes or ovals (denoting joint confidence regions) because experimental data only exist for $T_c$ for these compounds. The lines correspond to the 95% confidence intervals in $T_c$ reported by Nikitin et al., where we have included both the values reported in 1997 and 2014 [15, 20]. However, notice that no experimental data exist for C$_{48}$. We have also provided uncertainty regions for two different prediction models, that of Teja and Tsonopoulos [12, 13]. These joint confidence regions are represented as boxes because the $T_c$ and $\rho_c$ prediction models were regressed independently and, therefore, the correlation should be zero. Although some correlation is expected, the molecular simulation results from Errington et al. are also reported as boxes because the author only reported the standard error for $T_c$ and $\rho_c$. The oval shaped joint confidence regions
were obtained using the algorithm outlined in Section 3.3.4. This algorithm was performed not only on the data from this work but also on the data of Nath et al [25].

Figure 4.17: Comparison between uncertainty regions of experimental data from Nikitin, prediction models from Tsonopoulos and Teja, literature simulation results from Nath and Errington (with $T_c$ and $\rho_c$ shifted by +1% to correct for finite-size effects), and simulation results from this work using the Exp-6, NERD, and TraPPE models. Panels (a)-(c) correspond to $C_{24}$, $C_{36}$, and $C_{48}$, respectively. Each region represents the 95% joint confidence region for $T_c$ and $\rho_c$.

The results presented in Figure 4.17 support several key conclusions. For example, the confidence regions from the D-optimal design (denoted as “This Work” in Figure 4.17) are very small considering only 10 GEMC simulations were performed. Notice that the relatively simple D-optimal design yielded a smaller confidence region than the computationally intensive GCMC histogram reweighting approach employed by Errington et al. (see Figure 4.17 Panels (a) and (c)). This is significant when trying to assess which prediction model and experimental data are most reliable.

**Elucidating Correct $T_c$ Trend**

A significant observation from Figure 4.17 Panel (c) is that the $C_{48}$ results from this work for the NERD and Exp-6 models strongly agree (dashed/solid dark blue and dashed/solid dark
green ovals). By contrast, the confidence regions for Nath et al. using the NERD model (dashed-dotted light blue oval) and Errington et al. using the Exp-6 model (dashed-dotted purple rectangle) differ considerably, particularly for $T_c$. The importance of this observation is best demonstrated by comparing the different methods of estimating $T_c$ with respect to carbon number for larger $n$-alkanes.

![Graph showing the trend for $T_c$ with respect to CN for higher CN](image)

**Figure 4.18:** The trend for $T_c$ with respect to $CN$ for higher $CN$. Included are two different sets of experimental data from Nikitin, two different prediction models from Tsonopoulos and Teja, literature simulation results from Errington et al., Nath et al., and Siepmann’s group (Zhuravlev et al.), and simulation results from this work using the Exp-6, NERD, and TraPPE models. Only results from the 200 molecule simulations were included for clarity. The 1600 molecule simulation results overlap considerably with those shown for 200 molecules. Error bars correspond to 95% confidence intervals. For clarity, only a single representative error bar is presented for Nikitin’s experimental data. A slight offset is used in $CN$ for the Exp-6 and NERD results from this work for visual clarity.

In Figure 4.18 we have replotted the experimental data, prediction models, and simulation results as found in Figure 4.17 but have focused on the dependence of $T_c$ on $CN$ for longer chain-lengths. As depicted in Figure 4.18, the $T_c$ value for $C_{48}$ reported by Nath et al. for the NERD model was much lower than that reported by Errington et al. for the Exp-6 model. This is important because the results from Nath et al. agree more with Tsonopoulos’ $T_c$ model, whereas, the results
from Errington et al. support Teja’s $T_c$ model. Our results resolve the conflict between the NERD and Exp-6 models in that both conclude that Teja’s $T_c$ model is more accurate than Tsonopoulos’ $T_c$ model for larger $n$-alkanes. In addition, although the TraPPE model is known to over predict $T_c$ for larger $n$-alkanes, the TraPPE model also agrees more closely with Teja’s model than Tsonopoulos’.

Our simulation data presented in Figure 4.18 also help discern between the conflicting experimental $T_c$ data. Notice that our simulation results for $C_{24}$, $C_{36}$, and $C_{48}$ suggest that the more recent experimental values from Nikitin et al. are more reliable than their original values. Interestingly, the $T_c$ values from Nath et al. are in better agreement with the original values reported by Nikitin et al. As explained in the next section, we attribute the discrepancy between our results and those of Nath et al. to finite-size effects.

**Quantifying Finite-Size Effects**

Finite-size effects can become significant when simulating large compounds because the number of molecules typically simulated are relatively small. Unfortunately, quantifying finite-size effects with GEMC is not a straight-forward task [22]. We implemented two different methods to quantify the finite-size effects found in the results presented in Figure 4.17. First, we employed a common approach found in the literature of simply comparing the results from simulations using 200 and 1600 molecules [22]. Second, we compared our results for the Exp-6 model to those of Errington et al. Both approaches are now explained with the former being presented first.

Recently, it was shown that for GEMC systems as large as $n$-decane, finite-size effects for $T_c$ and $\rho_c$ were smaller than the corresponding statistical uncertainty [22]. Following this methodology we arrive at the same conclusion for the large $n$-alkanes simulated in this study. Notice in Figure 4.17 Panel (c) that even for $C_{48}$ the 95% confidence regions overlap between the 200 and 1600 molecule systems sizes for a given force field model. On the other hand, our results for the NERD model differ considerably from the results of Nath et al. for $C_{36}$ and $C_{48}$ (see Figure 4.17 Panels (b) and (c)). In the prior work, Nath et al. state that due to finite-size effects their “results should therefore be regarded with caution” as they were obtained with a system size of only 120 molecules. The results in Figure 4.17 Panels (b) and (c) seem to support this statement. In short, the fact that the results for 120 and 200 molecules are significantly different while those for 200 and 1600 molecules are comparable suggests that the 200 molecule simulations
have converged to the infinite system value. This conclusion is consistent with what was previously observed for GEMC simulations of \(n\)-decane [22].

While the previous comparison of 200 and 1600 molecule systems is promising, another approach to quantify finite-size effects is to compare our results for the Exp-6 model with those originally reported by Errington et al. This comparison was performed for \(C_{24}\) and \(C_{48}\), which are the two largest molecules previously simulated with the Exp-6 model. Errington et al. demonstrated that finite-size effects resulted in a systematic under prediction of \(T_c\) and \(\rho_c\) of approximately 1\% for ethane and \(n\)-octane with the system sizes used in developing the Exp-6 model [23]. Because Errington et al. implemented the same system size for the larger \(n\)-alkanes, it is assumed that \(C_{24}\) and \(C_{48}\) also suffer from this systematic deviation. To account for finite-size effects and to allow for a useful comparison with our results, we increased the values that Errington et al. reported for \(T_c\) and \(\rho_c\) by 1\% (see Figure 4.17). As Panels (a) and (c) of Figure 4.17 depict, the results from this study for \(C_{24}\) and \(C_{48}\) overlap considerably at the 95\% confidence level with the shifted uncertainty regions for the Errington et al. values. It is therefore reasonable to assume that finite-size effects do not affect our results to a large degree. In brief, the results for long-chain-length molecules agree with the conclusion reported by Dinpajooh et al. that “the finite-size dependence of the critical properties obtained from GEMC simulations is significantly smaller than those from grand-canonical ensemble simulations.” [22]

### 4.11.2 \(P_c\) Results for Larger \(n\)-Alkanes

Having compared our results with the literature for \(T_c\) and \(\rho_c\) we now turn our attention to \(P_c\). Figure 4.19 depicts the dependence of \(P_c\) upon carbon number for large \(n\)-alkanes. Included in Figure 4.19 are both sets of experimental values reported by Nikitin et al. [15, 20], three different prediction models [9, 14, 15], molecular simulation results from three different sources [23, 27, 28], and the results we obtained for the TraPPE, NERD, and Exp-6 models. Remember that our results utilize Vetere’s method to predict \(P_c\) from \(\rho_l\) while those from the literature used \(P_v\) to predict \(P_c\). However, we have also implemented the Vetere method with the orthobaric density results reported by Nath et al. for the NERD model, those reported by Zhuravlev et al. for the TraPPE model, and those reported by Muller et al. for the NERD and TraPPE models. These \(P_c\) values are referred to as Nath-Vetere, Siepmann-Vetere, and Muller-Vetere in Figure 4.19. The uncertainties for Nath-
Vetere, Siepmann-Vetere, Muller-Vetere, and the results from this work were obtained with the algorithm found in Section 4.3 while the uncertainties for the simulation literature values were those reported by Errington et al. [23] and Zhuravlev et al. [28] Also, we have included a single error bar for each set of experimental data that is representative of the 95% confidence intervals that were reported by Nikitin et al [15, 20].

Figure 4.19: The trend for $P_c$ with respect to $CN$ for higher $CN$. Two conflicting experimental data sets are included as well as three prediction models regressed to experimental data. Simulation results for the TraPPE and Exp-6 models are as reported by Zhuravlev et al. and Errington et al., respectively. The Nath-Vetere, Siepmann-Vetere, and Muller-Vetere results are also shown. Included are simulation results from this work using the Exp-6, NERD, and TraPPE models with a D-optimal experimental design. Only results from the 200 molecule simulations were included for clarity. The 1600 molecule simulation results overlap considerably with those shown for 200 molecules. Error bars correspond to 95% confidence intervals. For clarity, only a single representative error bar is presented for Nikitin’s experimental data. A slight offset is used in $CN$ for the simulation results from this work for visual clarity.
There are at least two important conclusions from the results presented in Figure 4.19. First, our simulation results for all three potential models (Exp-6, NERD, and TraPPE) agree with Nikitin’s most recently reported values for $P_c$. Again, although Nath-Vetere agrees most with Nikitin’s older set of data and with Nannoolal’s model, we attribute this to finite-size effects. The deviation in Errington’s results is likely due to the traditional approach of obtaining $P_c$ from $P_v$. Even though the Exp-6 model predicts $P_v$ and $T_c$ fairly accurately, the over prediction of $P_c$ is probably “the result of compound errors in the predictions of critical temperatures and vapor pressure.” [24] By contrast, we are able to obtain consistent predictions of $P_c$ for the simple Lennard-Jones 12-6 models (NERD and TraPPE) and the more advanced Exp-6 model by utilizing the Vetere method. Second, by using a D-optimal design with the Vetere method the uncertainties in our results are small enough to elucidate the correct long-chain-length trend. Notice that the uncertainties in $P_c$ from simulation literature data are typically very large even when the Vetere method is employed.

We believe that these are the first quantitatively accurate and precise estimates of $P_c$ from molecular simulation for large $n$-alkanes. These results are significant because they demonstrate that Nikitin’s most recent $P_c$ data should be accepted over his previous data. Our results also suggest that the models developed by Lemmon et al. and Nannoolal et al. are not as reliable as the model reported by Nikitin et al. for predicting $P_c$ of long-chain-length $n$-alkanes.

4.11.3 $Z_c$ Results

The critical compressibility factor is a key constant when assessing the reliability of $T_c$, $\rho_c$, and $P_c$ values collectively. Figure 4.20 provides the $Z_c$ trend with respect to $CN$ for the entire range of compounds considered in this work. Included in Figure 4.20 are the experimental values reported by Teja et al. [12] (since Nikitin et al. did not provide any $\rho_c$ values $Z_c$ cannot be calculated for their data), four different prediction models [9, 12, 13, 15, 29], molecular simulation results from three different sources [23, 27, 28], Siepmann-Vetere, Nath-Vetere, Muller-Vetere, and the results we obtained for the TraPPE, NERD, and Exp-6 models. The model labeled “Nikitin” is a composite of two different studies. Specifically, the $T_c$ and $P_c$ models are from the 2014 publication [15] while the $\rho_c$ model is from a 1998 publication [29], which is the most recent $\rho_c$ model reported by this group. We have included a single representative experimental uncertainty.
for $C_{11}$ that was obtained from the DIPPR 801 database [17]. The uncertainties corresponding to
the literature simulation values for the Exp-6 and TraPPE models were those reported by Errington et al. and Zhuravlev et al., respectively [23, 28]. Notice that for the Siepmann values we have only provided a single error bar for clarity since the uncertainties reported by Martin et al. for $C_8$ and $C_{12}$ are of the same size as that reported by Zhuravlev et al. It should be mentioned that
the propagation of errors derivation used to calculate the literature uncertainties assumes that the uncertainties in $T_c$, $\rho_c$, and $P_c$ are not correlated, which is clearly not a valid assumption. However, this is the standard approach found in the literature. The uncertainties for Siepmann-Vetere, Nath-Vetere, Muller-Vetere, and our results were obtained using the algorithm detailed in Section 4.3.

There are several key observations for the results presented in Figure 4.20. First, our results appear to be the most internally consistent set. Notice that the values from Errington et al., Siepmann, and Nath-Vetere have large oscillations between different $CN$. The Siepmann-Vetere and Muller-Vetere results agree surprisingly well with the Nikitin model, however, the uncertainties are too large to provide much confidence. Since $Z_c$ depends on all three of the critical constants the uncertainties in $Z_c$ are typically quite large. By contrast, the uncertainties for our results are small enough to discern between the different prediction models. It is also worth noting that our simulation results appear to be approaching an asymptotic value around 0.2. Recently, the statistical associating fluid theory (SAFT) was used to predict that $Z_c \rightarrow \frac{1}{3}$ for the infinite chain-length $n$-alkane [83]. Notice that the Nath-Vetere results are much lower than this value and those of Errington et al. appear to be approaching a much higher asymptote. (Although we will show in Chapter 6 that the fully theoretical SAFT model for polymer’s predicts that $Z_c \rightarrow \frac{1}{5}$.) Furthermore, the models from Teja et al. and Tsonopolous et al. diverge since $Z_c \rightarrow \infty$ while the Nannoolal model predicts that $Z_c \rightarrow 0$ [9, 12, 13]. It is worth mentioning that the asymptotic $Z_c$ value for the Nikitin composite model is 0.15. Although the Nikitin model disagrees with the limiting value obtained from both the theoretical and soft-SAFT equation of state, it does agree well with our simulation results. Therefore, we believe that the Nikitin composite model is the most reliable approach for predicting $Z_c$ for larger $n$-alkanes, at least as large as $C_{48}$. In Chapter 6 we present an improved set of critical constant models based on the initial work of Nikitin et al.
Figure 4.20: The trend for $Z_c$ with respect to $CN$. Included are experimental data, four prediction models, and simulation results from the literature and from this work using the Exp-6, NERD, and TraPPE models with a D-optimal experimental design. Only results from the 200 molecule simulations were included for clarity. The 1600 molecule simulation results overlap considerably with those shown for 200 molecules. Error bars correspond to 95% confidence intervals. For clarity, only a single representative error bar is presented for the Siepmann values and experimental data. A slight offset is used in $CN$ for the simulation results for visual clarity.

4.12 Validation of Results for Large $n$-Alkanes

In Figure 4.21 we present the vapor-liquid coexistence curves (VLCC) for the compounds and force field models discussed in this section. Each force field model corresponds to a different color: Exp-6 in blue, NERD in green, and TraPPE in red. Panels (a)-(c) contain simulation results for the $n$-alkanes $C_{24}$, $C_{36}$, and $C_{48}$, respectively. The filled circles correspond to the D-optimal design temperatures that are used to obtain the estimates of $T_c$, $\rho_c$, $P_c$, and $Z_c$ found in Section 4.11. The open circles are simply validation points that were not used in regression. Only the simulation results using 200 molecules are shown for clarity. Each circle is an average of five independent...
simulations. Uncertainties in $\rho_l$ and $\rho_v$ are not displayed since the 95% confidence intervals are approximately one symbol size.

Figure 4.21: Comparison of the different force field models used for each compound presented in this work. Panels (a)-(c) correspond to $C_{24}$, $C_{36}$, and $C_{48}$, respectively. The filled circles are the points used to obtain the estimates of $T_c$, $\rho_c$, $P_c$, and $Z_c$. The 95% confidence intervals for the circles are smaller than one symbol size. The dashed lines are the best fit when using only the filled circles in a regression. The diamond symbols represent the critical point estimates.

Having demonstrated that our VLCC curves have the correct general shape, we turn our attention to the validity of the regression equations. As mentioned in Section 4.7, some limitations exist for using the Vetere approach with the D-optimal design. One such limitation is whether or not the law of rectilinear diameters and the density scaling law are valid at temperatures far removed from $T_c$. For this reason, in Figure 4.22 we present the linearized forms of the law of rectilinear diameters and the density scaling law (Equations 4.17-4.18) with respect to $T_r$ for $C_{24}$, $C_{36}$, and $C_{48}$. For clarity, only the results for the 200 molecule systems are shown in Figure 4.22. The main point of this figure is that the low $T_r$ values are consistent with a straight-line regression.

It is worth clarifying that the results presented previously in Figures 4.17-4.20 were obtained from only two temperature values ($T_r \approx 0.7$ and 0.85). The other temperatures shown in Figures 4.21-4.22 are simply to demonstrate the validity of Equations 3.1-3.2 over a large temperature range. Also, the same simulation specifications and number of replicates were used at each temperature.
Figure 4.22: Validation of the regression equations over a large range in $T_r$ for $C_{24}$, $C_{36}$, and $C_{48}$ using 200 molecules. In Panels (a) and (b) we see that, respectively, $\rho_r$ and $\rho_c^\frac{1}{\beta}$ follow a linear trend over this entire $T_r$ range.

### 4.13 Conclusions

In this work we have demonstrated that an alternative method, namely the Vetere approach, has many advantages when predicting $P_c$ from GEMC for larger compounds. However, this method is very sensitive to the uncertainty in $\rho_c$. Although the uncertainty in $T_c$ is relatively small for the experimental designs found in the literature, the uncertainty in $\rho_c$ can be quite large. For this reason, we have developed an experimental design that minimizes the uncertainty in $T_c$, $\rho_c$, and $P_c$ (and, thereby, $Z_c$).

The proposed experimental design has several significant implications that run contrary to the common practice found in the simulation literature. First, it is essential to perform simulations at the lowest temperature feasible. In the literature it is quite common to undervalue the low temperature regime and over emphasize the high temperature regime. Second, only two temperatures are necessary for accurate predictions of the critical constants. Third, it is better to perform several replicate simulations at the two optimal temperatures rather than running simulations at several temperatures. Finally, it is not necessary to perform simulations at the upper temperature limit because the inherent uncertainty associated with the simulation results in this region increases ex-
ponentially with respect to temperature. This last conclusion is particularly significant because the required simulation length increases substantially near the critical point. This additional computational cost is actually not as beneficial for predicting the critical point as most have assumed. Furthermore, finite-size effects are reduced by performing simulations farther away from the critical point.

We validated the performance of the D-optimal design by performing simulations with the NERD and TraPPE models for C_{6}-C_{16}. We observed good agreement between our $T_c$ and $\rho_c$ values and those reported by Nath et al. and Martin et al. The primary difference is that the $\rho_c$ uncertainties are significantly smaller for our results compared to those found in the literature. By reducing the uncertainty in $\rho_c$, the Vetere approach yields reliable estimates for $P_c$. A reduction in the uncertainties for $\rho_c$ and $P_c$ leads to significantly smaller uncertainties in $Z_c$ than typically found in the literature. We then applied the Vetere method and D-optimal design to C_{24}, C_{36}, and C_{48}. We believe our results are the most accurate and precise estimates of $P_c$ and $Z_c$ for larger $n$-alkanes. Finally, we verified that finite-size effects are of approximately the same magnitude as the statistical uncertainty.

There are three main conclusions from our simulation results. First, our results resolved the discrepancy in the simulation literature between the NERD and Exp-6 models for the $T_c$ trend of large $n$-alkanes. Both models support Teja’s $T_c$ model over Tsonopoulos’ model. Second, we obtained accurate and precise results of $P_c$ using the TraPPE, NERD, and Exp-6 models for large $n$-alkanes. These results helped us evaluate which experimental data were most reliable for the critical constants. Specifically, it was found that the most recent $T_c$ and $P_c$ values are indeed more reliable than the original values reported by Nikitin et al. Third, our results appear to be approaching an asymptotic value of 0.2 for $Z_c$, in agreement with the soft-SAFT model. This suggests that an acceptable set of $T_c$, $\rho_c$, and $P_c$ prediction models should result in a non-infinite and non-zero $Z_c$ value for the infinite chain-length $n$-alkane. In Chapter 6 we will discuss in detail what the infinite chain-length value is for $Z_c$. In total, our results demonstrate how reducing the uncertainty in molecular simulation results can assist in analyzing the validity of experimental data and prediction methods.
5.1 Introduction

Since the reliability of molecular simulation results is intimately connected with the intermolecular potential model chosen and the associated model parameters ($\varepsilon$, $\sigma$, etc.), one of the most important layers of uncertainty is related to the optimization of the model parameters. Historically, most studies that report model parameters (i.e. TraPPE [27], NERD [25], Exp-6 [23], Mie $\lambda$-6 [24], OPLS [77], etc.) neglect to perform a statistical analysis of their “best fit” parameters. In 2000, Ray Mountain from the Physical and Chemical Properties Division at the National Institute of Standards and Technology (NIST) detailed a rigorous sensitivity analysis approach and proposed that “sensitivity analysis should be an integral part of potential function/force field development.” [58] Despite this endorsement, most simulation studies do not dedicate the effort needed for a useful sensitivity analysis.

Occasionally, the literature provides a sensitivity analysis where the model parameters are varied by an arbitrary amount and the effect on the simulation results is reported [84,85]. However, a sensitivity analysis of this nature actually provides little insight because the region of parameter space being sampled usually lands well outside the statistically-acceptable range of parameter sets. This is due to two factors. First, the percent that a parameter is varied is often too large to be realistic. Second, the model parameters are correlated since changing one parameter often necessitates a change in another for the parameter set to be statistically-acceptable. A more rigorous approach is to examine what combinations of parameters provide a fit to the data that is acceptable at a given statistical level. Recently, Bayesian statistics have been used for such a purpose [60,61,86]. These studies have pioneered the field of force field parameterization uncertainty quantification and propagation of errors (UQ+PoE). Although these studies have focused on small molecules, such as argon and water, their results have been quite promising. In the present study, we employ a
more intuitive method of classical (frequentist) statistics UQ+PoE, which NIST classifies as a Type A approach [87]. In addition, we develop a non-statistical approach to UQ+PoE, referred to as a Type B approach. Furthermore, this study is significant in that it is the first attempt to quantify the uncertainties due to the intermolecular potential parameters for industrially relevant compounds (n-alkanes as large as C_{48}H_{98}).

The primary purpose of this chapter is four-fold. The first task is to develop the Type A (using frequentist statistics) and Type B (using engineering intuition) methodologies of uncertainty quantification in the intermolecular model parameters. Specifically, this analysis will be performed for the Lennard-Jones (LJ) 12-6 parameters (ε and σ) for the united-atom (UA) methane (CH_{4}), methyl (CH_{3}), and methylene (CH_{2}) groups. The second objective is to demonstrate the propagation of errors methodology for properties and compounds not included in the parameterization process. For example, we will quantify how the uncertainties in the CH_{3} and CH_{2} LJ 12-6 parameters propagate to the critical constants for larger n-alkanes. The third goal is to compare the uncertainty quantification and propagation of errors results for the different uncertainty analysis methods. The final purpose is to verify that the uncertainties in the critical constants are small enough to elucidate the correct trends for large n-alkanes.

The outline for this chapter is the following. First, we develop the Type A and Type B uncertainty quantification and propagation of errors approaches. Second, we demonstrate how these analysis methods can be implemented for the simple example of a single-site LJ 12-6 molecule (a CH_{4} UA group). Next, we demonstrate how to extend this approach to CH_{3} and CH_{2} groups using the same parameterization methodology as that of the TraPPE-UA model explained in Section 2.2.3. This is followed by a force field validation for ethane and n-octane. Subsequently, we present the results for the critical constants of long-chain-length n-alkanes. Finally, we discuss some limitations and the primary conclusions of this work.

5.2 Theory of UQ+PoE

Although numerical and model uncertainties can be significant for molecular simulation results, in this section the discussion will focus upon parameter uncertainty. In general, there are two different approaches for estimating uncertainty that are referred to as “Type A” and “Type B” in the Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results [87].
A Type A uncertainty analysis utilizes statistical methods while a Type B method employs some other means. Typically a Type A analysis only accounts for random deviations in the experimental data while a Type B analysis also accounts for systematic deviations. A Type B analysis is not as straightforward since it requires an expert evaluation which must consider several factors such as the methodology, the apparatus, the measurement devices, the purity of the sample, the type of data being measured, and the degree of data analysis required. In fact, Cailliez et al. recognized systematic deviations and conflicting data as the largest obstacle in UQ of force field models [61]. Fortunately, at the Design Institute for Physical Properties (DIPPR 801) this type of analysis is performed by thermodynamic experts that are not only familiar with various experimental methods but have years of experience in assigning uncertainties. For this reason, we have relied upon the uncertainties reported in the DIPPR 801 database for our Type B analysis [17].

Although a Type A and Type B analysis typically refer to uncertainties in data, we will use the same terminology for quantifying parameter uncertainty. These terms are useful and appropriate since the parameter uncertainty is primarily a result of the uncertainty in the data used in parameterization. In Section 2.4 we developed the general Type A analysis method. In the following sections we present the Type A and Type B UQ+PoE approaches for the specific case of the intermolecular potential parameters uncertainty. Within the Type A framework are two subsets which we call Type A and Type $A_B$. Type A relies completely upon statistical methods while Type $A_B$ is a hybrid between the two analysis types. After discussing the Type A and Type $A_B$ approaches we focus on the Type B method.

## 5.2.1 Type A UQ+PoE

There are two main approaches to a Type A (statistical) uncertainty quantification, classical (frequentist) and Bayesian. Detailed derivations of the Bayesian approach for molecular simulation can be found in the literature [60,61,86]. The frequentist approach is an attractive alternative since it is more intuitive and familiar to most researchers in the field. For this reason, we employ classical frequentist statistics in this chapter.
Parameter Uncertainty Quantification

In Chapter 2 we derived the frequentist statistical (Type A) approach to parameter uncertainty quantification. In this section we demonstrate how this methodology is applied for determining the uncertainties in the force field parameters. It is important to note that Equations 2.16-2.17 are compatible with the TraPPE objective function (Equation 2.15) since \( \text{RMS} = \sqrt{\frac{S}{N}} \), where \( \text{RMS} \) is the root-mean-square, \( S \) is the sum squared error, and \( N \) is the number of data points.

Recall that an estimate for \( s^2 \) is necessary for the UQ+PoE methodology. In this chapter we have estimated \( s^2 \) with both a Type A and Type B approach. Since the overall methodology is a Type A analysis (i.e. derived from Equations 2.16-2.17) we have distinguished these two approaches as Type A and Type A\(_B\). A Type A approximation of \( s^2 \) utilizes the statistically based expression found in Equation 2.18 (which is used in deriving Equations 2.19-2.20). For the Type A\(_B\) analysis, \( s^2 \) is approximated from the reported uncertainties in the DIPPR 801 database \( (u_{\text{DIPPR}}) \) [17]. Specifically, according to Subsection 4.3 in Ref [87], we interpreted the DIPPR 801 uncertainties as the single measurement 95\% confidence interval for a known standard deviation (i.e. \( u_{\text{DIPPR}} = 1.96 \times S_D \), the degrees of freedom \( (v) = \infty \), and, thus, \( s^2 = S_D^2 = (\frac{u_{\text{DIPPR}}}{1.96})^2 \)). Note that in both cases the numerical uncertainty in the simulation results can be included in \( s^2 \).

Since both the Type A and Type A\(_B\) analysis methods accept parameter sets based on how they compare with the optimal fit \( (S(\hat{\theta})) \) they do not consider the possibility of a bias in the data. For this purpose, in Section 5.2.2 we discuss how to account for systematic deviations with a Type B approach.

Propagation of Errors

The uncertainty in \( \varepsilon \) and \( \sigma \) leads to an uncertainty in the simulation results. This propagation of errors approach is fundamentally the same as that in Section 2.4.3. The general outline of the Type A UQ+PoE is repeated here to aid in the discussion:

1. Create a grid of the \( p \)-dimensional parameter space
2. Calculate the \( PDF \) for each parameter set, \( \theta \)
3. Generate hundreds, thousands, or millions of random numbers
4. Assign each random number to a parameter set (via the PDF values)

5. Obtain the desired property value for the parameter sets sampled in Step 4

6. Create a histogram of the property values

Step 2 is the prohibitively expensive step since calculating the PDF for a single parameter set can require tens of molecular simulations. For a refined grid, of say 500x500 parameter sets, this necessitates millions of simulations. For this reason, a surrogate model is essential for the UQ portion [60]. A surrogate model reduces the number of molecular simulations required by predicting the PDF and/or physical property values for each parameter set. Fortunately, it is not necessary that the surrogate model accurately predict every physical property, only the properties included in the objective function used to calculate the PDF. An additional advantage of a surrogate model is that the numerical uncertainty is reduced by smoothing the simulation output.

An adequate pseudo-random number generator is an essential aspect to a Monte Carlo Sampling (MCS) propagation of errors (PoE) approach. The random numbers are used to obtain a properly weighted distribution of parameter sets from the PDF. The amount of random numbers generated in Step 3 is determined by the property obtained in Step 5. Specifically, if the property has an analytic expression (i.e. a theoretical derivation or surrogate model) it is possible to use on the order of a million random numbers. However, if the property requires performing molecular simulations then the amount of random numbers generated is limited by computational costs.

5.2.2 Type B UQ+PoE

The Type B acceptance criterion for a given parameter set depends solely on whether the predicted property values are within the DIPPR 801 correlation uncertainties. The essential difference between the Type A and Type B methods is that the former assigns the DIPPR 801 uncertainties to each data point while the later applies the uncertainties to the entire correlation. Notice that the uncertainties reported by DIPPR 801 are not necessarily uncertainties in the data (s^2), rather they are uncertainties in the correlation. In other words, it is possible that a bias in the data has caused the entire correlation to be shifted one direction. In contrast with both the Type A and Type A_B methods, a Type B analysis can account for a potential bias in the data. Therefore, although a
Type B approach is not as developed and refined as the Type A (or Type A\textsubscript{B}) method, it has some significant practical advantages for molecular simulation. First, it greatly reduces the complexity as well as the computational cost. Second, it may provide a more realistic, conservative, and useful assessment of the reliability of the simulation results.

**Parameter Uncertainty Quantification**

As the Type B analysis does not have a rigorous statistical derivation the methodology is much simpler than that described in Section 2.4 for the Type A approach. A parameter set is considered acceptable if it satisfies the following expression over the entire temperature range

\[
|\hat{y}(T; \theta) - y\text{DIPPR}(T)| \leq u\text{DIPPR}(T) \tag{5.1}
\]

where \(\hat{y}(T; \theta)\) is the predicted property value from the molecular simulation using the parameter set \(\theta\), \(y\text{DIPPR}\) is the property value from the DIPPR correlation, and \(u\text{DIPPR}\) is the uncertainty assigned by DIPPR 801. If desired, it is also possible to include the simulation (numerical) uncertainty on the right hand side of the above equation.

**Propagation of Errors**

Since the Type B approach does not have a rigorous derivation for generating \(\text{PDF}(\theta)\), we believe that a sufficiently reliable propagation of errors approach is to simply evaluate the predicted property values at the extrema of the accepted parameter region. The reason for this assumption is that \(u\text{DIPPR}\) represents the range in which the “true” value is found 95 out of 100 times, while saying nothing about the distribution of values (analogous to the case described in Subsection 4.6 of Ref [87]). Therefore, since the maximum and minimum values (i.e. \(y\text{DIPPR} \pm u\text{DIPPR}\)) correspond to the extrema parameter sets (as demonstrated in Section 5.5.3) there is no need to sample from the unknown distribution of interior parameter sets.

This propagation of errors approach has some significant advantages. For example, by only performing simulations with a few parameter sets at the extrema, rather than sampling hundreds of parameter sets, the number of molecular simulations and, thereby, the computational costs are...
reduced significantly. Furthermore, by contrast with the Type A and Type A_B methods, the Type B approach (Equation 5.1) does not require an estimate of the optimal parameter set \( \hat{\theta} \), which is typically a very time consuming and computationally expensive process.

Another advantage of the Type B approach is that it is more compatible with a multi-property optimization than the Type A approach. This is because a Type A approach requires an objective function for calculating the PDF. Multi-property objective functions tend to use complicated and somewhat arbitrary weighting of different data types [88]. By contrast, the Type B approach does not require a PDF, which are very tedious to develop for an entire parameter space. Instead, the multi-property Type B analysis simply requires that the different properties are predicted within a certain tolerance. Recently, some simulation studies have attempted to use advanced multi-property methods such as the Pareto front [89]. We believe that the multi-property Type B approach can reduce the need for an arduous Pareto front analysis while still providing a physically meaningful quantification of uncertainty in molecular simulation results. In Section 5.5, we demonstrate the \( \rho_l-T_c \) multi-property Type B analysis for ethane and \( n \)-octane.

### 5.3 UQ+PoE for CH\(_4\)

In this section we provide an example of the UQ+PoE methods outlined previously. The simplest example is that of the Lennard-Jones fluid, that is, a single-site molecule with a LJ 12-6 potential and no polar interactions. Since we are focusing on the \( n \)-alkanes in this study, the logical choice is the united-atom methane molecule. Again, we use the RMS as the objective function to obtain \( \epsilon_{CH_4} \) and \( \sigma_{CH_4} \).

In the case of the single-site LJ fluid some thermophysical properties have analytic expressions (such as the second virial coefficient \( B_2 \)) while others require molecular simulation (such as the critical point constants). Fortunately, molecular simulations may not be necessary even for properties that cannot be predicted by rigorous theoretical expressions because the LJ fluid has been extensively studied in the literature. For example, correlations (surrogate models) exist for the LJ fluid to relate \( \epsilon \) and \( \sigma \) to many properties, such as saturated liquid density \( \rho_l \). This is advantageous because an analytic expression to predict \( \gamma(T,\theta) \) greatly simplifies the UQ+PoE process. Likewise, molecular simulation results for the LJ coexistence curve have provided approximate estimates for the reduced critical temperature \( T_c^* \), reduced critical density \( \rho_c^* \), and
reduced critical pressure \((P^*_c)\). As different studies of the LJ fluid have predicted slightly different values for \(T^*_c, \rho^*_c,\) and \(P^*_c\) we will use those most recently reported by Dinpajooh et al. [22].

Since the main purpose of this section is to demonstrate the UQ+PoE methodology, we did not attempt to follow the TraPPE parameterization method for methane that utilizes a perturbation theory approach to relate pressure \((P)\) to \(\epsilon\) and \(\sigma\). Instead, we have chosen \(B_2\) and \(\rho_l\) to optimize the LJ parameters because these two properties have simple analytic expressions (i.e. theoretical relationships or correlations). The second virial coefficient of the single-site LJ fluid is [54]

\[
B_2(T; \epsilon, \sigma) = \frac{-2\pi \sigma^3}{3} \sum_{i=0}^{\infty} \frac{2^{(2n+1)/2}}{4n!} \left( \frac{\epsilon}{k_B T} \right)^{\frac{2n+1}{4}} \Gamma \left( \frac{2n-1}{4} \right)
\]  

(5.2)

where \(k_B\) is the Boltzmann constant and \(\Gamma\) is the gamma function. Equation 5.2 is a completely rigorous expression to calculate \(B_2\) from \(\epsilon\) and \(\sigma\) (note that Equation 5.2 is mathematically equivalent to the equation found in Ref [61]). We developed a new correlation for predicting \(\rho_l\) from \(\epsilon\) and \(\sigma\) for the LJ fluid by fitting GEMC simulation results in reduced units to the expression

\[
\rho_l^*(T^*) = b_0 + b_1(b_2 - T^*) + b_3(b_2 - T^*)^{b_4}
\]  

(5.3)

where \(\rho_l^* = \rho_l \sigma^3\), \(T^* = \frac{T k_B}{\epsilon}\), and \(b_i\) are fitting coefficients. The optimal fit was obtained using coefficients of \(b_0 = 0.3144, b_1 = 0.1741, b_2 = 1.2996, b_3 = 0.5002, b_4 = 0.3333\). We use a similar approach when developing our surrogate model in Section 5.7.

Again, since the experimental data used in parameterization have random error the regressed values for \(\epsilon_{CH_4}\) and \(\sigma_{CH_4}\) inherit this uncertainty from the data. Furthermore, it has been shown that the single-site LJ 12-6 model is not flexible enough to predict \(B_2\) over a large temperature range [90]. Therefore, the model inadequacies will also lead to an enlargement in the parameter uncertainties. For this reason, we have only used \(B_2\) data in a limited temperature range.

In Section 5.3.1 we demonstrate how the uncertainty in \(\epsilon_{CH_4}\) and \(\sigma_{CH_4}\) is quantified. In Section 5.3.2 we use Monte Carlo Sampling (MCS) to generate millions of statistically acceptable \(\epsilon_{CH_4}\) and \(\sigma_{CH_4}\) parameter sets. In Section 5.3.3 we perform a propagation of errors analysis to show how the uncertainty in \(\epsilon_{CH_4}\) and \(\sigma_{CH_4}\) leads to uncertainties in \(\rho_l, B_2, T_c, \rho_c,\) and \(P_c\). As a proof of concept, we only present a Type A and Type B analysis. One purpose of this section is to compare
the uncertainty regions for the Type A and Type B approaches. Another purpose is to compare the

effect of using two different types of data in the expression for RMS.

5.3.1 UQ for CH$_4$

The first step in determining the uncertainty in $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ is to create a 2-dimensional
$(p = 2)$ grid. For every combination of $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ the $RMS(\varepsilon_{CH_4}, \sigma_{CH_4})$ is calculated. The
Type A 95% joint confidence region consists of all the sets of $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ that satisfy Equation
2.16 with $\alpha = 0.95$ and $p = 2$. The Type B acceptable parameter sets are determined using the
criterion in Equation 5.1.

It has long been known that different types of data yield different optimal values for the
LJ parameters [89, 91, 92]. Historically, this discrepancy is seen when comparing thermodynamic
and transport properties, liquid and vapor phase properties, as well as different temperature ranges
for a property such as $B_2$. However, rarely are confidence regions reported for these parameters
to demonstrate that they are indeed statistically different. Recently, Cailliez et al. performed
such a study for Argon [61]. They observed that the confidence regions did not overlap when the
parameters were obtained from $B_2$ or low-pressure gas viscosity data. Here we present a similar
analysis for methane.

Figure 5.1 provides a comparison between the Type A and Type B 95% confidence regions
for the two types of data ($B_2$ and $\rho_l$), two common values reported in the literature [93], the
TraPPE parameters [27], and the optimal set for predicting $T_c$, $\rho_c$, and $P_c$ (obtained using the
DIPPR values and simulation results reported by Ref [22]). Figure 5.1 demonstrates that $B_2$ and $\rho_l$
yield parameter sets that are statistically different at the 95% confidence level for both uncertainty
analysis methods. This supports the notion that the LJ 12-6 model is simply an approximation and,
therefore, one should proceed with caution when attempting to predict properties far removed from
those used in the parameterization. For example, the $\rho_l$ uncertainty region does not intersect with
either the $T_c$ or $P_c$ optimal lines.

Another significant observation from Figure 5.1 is that the confidence region for $B_2$ is con-
siderably larger than that for $\rho_l$. This suggests that utilizing $\rho_l$ data can yield lower uncertainties
in the optimized LJ parameters. However, it should be noted that the discrepancy between the $\rho_l$
and $B_2$ optimizations is attributed to the limitations in the model (i.e. UA, LJ 12-6, pair-wise addi-
tivity). In other words, if the model uncertainty were negligible the $\rho_l$ and $B_2$ uncertainty regions would overlap considerably. This demonstrates that this UQ+PoE analysis does not rigorously account for model uncertainties. Finally, it is worth noting the difference in shape between the Type A and Type B approaches. The Type A approach has an elliptical shape, similar to the results from Cailliez et al. [61], while the Type B approach has more linear borders that arise from accounting for bias with a hard constraint (Equation 5.1).

![Figure 5.1: Comparison of the UA, LJ 12-6, parameters for methane when regressed to different types of experimental data. Literature (“Lit.”) values are from Ref [93]. Optimal parameters for predicting $T_c$, $\rho_c$, and $P_c$ were obtained using the DIPPR values and simulation results reported by Ref [22]. The confidence regions for the Type A and Type B analysis are obtained at the 95% confidence level.](image)

**5.3.2 MCS for CH$_4$**

In Section 5.3.1 we obtained the $PDF$ for every set of $\epsilon_{CH_4}$ and $\sigma_{CH_4}$ in our 2-dimensional grid. We will demonstrate the Monte Carlo Sampling (MCS) approach for the Type A analysis. The first step for MCS is to generate millions of random numbers. Each random number corre-
sponds to a different set of $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$. The assignment of a random number to a parameter set is determined by the PDF values for each parameter set. We are able to assign a large number of random numbers because we have an analytic expression to calculate $\rho_l$ and $B_2$ and, thereby, PDF. We generate millions of random numbers since increasing the amount of random numbers will result in more accurate estimates of uncertainty. From this set of random numbers we obtain a properly weighted sample of millions of different $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ sets. Figure 5.2 contains histograms of the $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ values obtained from the MCS Type A analysis. Panels (a,c) and (b,d) correspond to $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$, respectively. Panels (a)-(b) were obtained using $\rho_l$ data, i.e. the MCS parameter sets are sampled from a PDF that depends on $\rho_l$, while Panels (c)-(d) utilized $B_2$ data. Probability density is defined as the number of counts in a single bin divided by both the total number of counts and the bin width. Note that the MCS parameter sets appear to follow a normal distribution.

5.3.3 PoE for CH$_4$

In this section, we demonstrate how the uncertainty in $\varepsilon$ and $\sigma$ propagate when predicting physical properties. Figure 5.3 compares the Type A and Type B uncertainties with the experimental data included in the analysis and the DIPPR uncertainties. Panels (a) and (b) correspond to $\rho_l$ and $B_2$, respectively. Since it is known that the optimal $B_2$ parameters are not capable of predicting $\rho_l$ accurately and vice versa, we did not consider predicting $\rho_l$ or $B_2$ with the parameter sets obtained from the other data type.

There are two key conclusions from Figure 5.3. First, the Type A approach results in considerably smaller uncertainties than the Type B approach. When compared to the DIPPR uncertainties the size of the Type A uncertainties does not seem justified. Remember that this is primarily because the Type A analysis does not account for systematic errors (bias) in the data. Second, the Type B uncertainties agree very well with the DIPPR uncertainties. This is significant since only the extrema of the acceptable parameter sets (those that satisfy Equation 5.1) were used in obtaining the Type B uncertainties. Therefore, the results presented in Figure 5.3 validate the assumption that the Type B uncertainties (for the properties included in Equation 5.1) can be adequately represented by sampling only at the extrema.
We will now focus our attention on $T_c$, $\rho_c$, and $P_c$ since quantifying the uncertainty in the critical constants is our primary goal for the larger $n$-alkanes. For each set of $\varepsilon_{CH_4}$ and $\sigma_{CH_4}$ we evaluate the critical constant expressions. We are able to evaluate the critical constants at millions of parameter sets because no further simulations are required as the critical constants are reported in
Figure 5.3: Comparison of \( \rho_l \) and \( B_2 \) experimental data, the DIPPR 801 correlation uncertainty, the Type A uncertainty, and the Type B uncertainty. Panels (a) and (b) correspond to \( \rho_l \) and \( B_2 \), respectively. The Type A uncertainties represent the 95% confidence interval.

Ref [22] in reduced units for the UA, LJ, single-site molecule. Specifically, \( T_c = \frac{\varepsilon T^*_c}{k_B} \), \( \rho_c = \frac{\rho^*_c}{\sigma^3} \), and \( P_c = \frac{\varepsilon P^*_c}{\sigma^3} \). In general, there is not a simple mathematical expression for propagating the uncertainty from the LJ parameters to the desired property. Typically, determining the desired property for a given set of LJ parameters requires between 1-20 molecular simulations. Such is the case for larger \( n \)-alkanes, as we will demonstrate following this discussion for methane. Again, when molecular simulations are required the number of \( \varepsilon \) and \( \sigma \) parameter sets that can be evaluated is greatly reduced.

For the Type A analysis, we create a histogram of the different values of \( T_c \), \( \rho_c \), and \( P_c \). These histograms can be integrated such that 5% of the area is found equally distributed between the left and right tails. This yields an estimate of the 95% confidence interval for \( T_c \), \( \rho_c \), and \( P_c \) due to the parameter uncertainty. The Type B analysis is much simpler because only the minimum and maximum \( \varepsilon \), \( \sigma \), and \( \frac{\varepsilon}{\sigma} \) acceptable values are required to estimate the 95% confidence regions for \( T_c \), \( \rho_c \), and \( P_c \), respectively. In Table 5.1 we compare the size of the 95% confidence intervals due to numerical, parameter, and overall uncertainties. We compare the parameter and overall uncertainties obtained from \( B_2 \) and \( \rho_l \) data for both the Type A and Type B analysis.
Table 5.1: Comparison of numerical, parameter, and overall uncertainties for the critical constants for UA, LJ 12-6, methane molecule. Uncertainties are presented as a relative combined expanded uncertainty (at the 95% confidence level) multiplied by 100%.

<table>
<thead>
<tr>
<th>Uncertainty</th>
<th>Data</th>
<th>Analysis</th>
<th>$T_c$</th>
<th>$\rho_c$</th>
<th>$P_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerical</td>
<td>VLCC</td>
<td>N/A</td>
<td>0.122</td>
<td>1.269</td>
<td>1.020</td>
</tr>
<tr>
<td>Parameter</td>
<td>$B_2$</td>
<td>Type A</td>
<td>0.264</td>
<td>0.860</td>
<td>0.229</td>
</tr>
<tr>
<td>Parameter</td>
<td>$B_2$</td>
<td>Type B</td>
<td>2.40</td>
<td>9.71</td>
<td>12.10</td>
</tr>
<tr>
<td>Parameter</td>
<td>$\rho_l$</td>
<td>Type A</td>
<td>0.036</td>
<td>0.091</td>
<td>0.010</td>
</tr>
<tr>
<td>Parameter</td>
<td>$\rho_l$</td>
<td>Type B</td>
<td>0.231</td>
<td>0.947</td>
<td>0.799</td>
</tr>
<tr>
<td>Overall</td>
<td>$B_2$</td>
<td>Type A</td>
<td>0.291</td>
<td>1.530</td>
<td>1.046</td>
</tr>
<tr>
<td>Overall</td>
<td>$B_2$</td>
<td>Type B</td>
<td>2.52</td>
<td>10.98</td>
<td>13.10</td>
</tr>
<tr>
<td>Overall</td>
<td>$\rho_l$</td>
<td>Type A</td>
<td>0.127</td>
<td>1.269</td>
<td>1.020</td>
</tr>
<tr>
<td>Overall</td>
<td>$\rho_l$</td>
<td>Type B</td>
<td>0.353</td>
<td>2.213</td>
<td>1.819</td>
</tr>
</tbody>
</table>

The standard deviation from numerical uncertainty ($s_{NU}$) was calculated as $s_{NU} = \frac{CI}{1.96}$, where $CI$ is the 95% confidence interval reported by Dinpajooh et al. [22]. In other words, we assumed that the numerical uncertainty follows a normal distribution (as we observed in Figures 3.4 and 4.1). We also discovered that the histograms obtained from MCS PoE for the critical constants follow a normal distribution, similar to the results presented in Figure 5.2. We obtained an estimate for the standard deviation caused by the parameter uncertainty ($s_{PU}$) for each critical constant by fitting a normal distribution model to the histograms produced from the MCS PoE results. One advantage of obtaining an estimate of $s_{NU}$ and $s_{PU}$ is that an estimate of the overall uncertainty ($s_{OU}$) can be obtained from the traditional propagation of error expression

$$s_{OU}^2 = s_{NU}^2 + s_{PU}^2 \quad (5.4)$$

Equation 5.4 is a rigorous approach for estimating the overall uncertainty in the critical constants because the numerical simulations and the parameterization were performed independently.

In Figure 5.4 we provide a graphical comparison between the numerical, parameter, and overall uncertainties for $T_c$, $\rho_c$, and $P_c$ for the UA, LJ 12-6, CH$_4$ molecule. We have included the results for both the $B_2$ and $\rho_l$ parameterization methods explained previously. Panels (a,d), (b,e), and (c,f) contain the uncertainties for $T_c$, $\rho_c$, and $P_c$, respectively. For clarity, we have only
included the Type A analysis results in Figure 5.4 Panels (a)-(c). The Type A and Type B overall uncertainties are compared in Panels (d)-(f).

Figure 5.4: Comparison of the uncertainties for the critical constants for the UA, LJ 12-6, CH$_4$ molecule. Panels (a,d), (b,e), and (c,f) contain the uncertainties for $T_c$, $\rho_c$, and $P_c$, respectively. The numerical uncertainties were obtained from Dinpajooh et al. while the parameter uncertainties were obtained using the Type A analysis discussed in Section 5.3.3 for the two different types of experimental data. Experimental uncertainties are those found in the DIPPR 801 database (assumed to be reported at the 95% confidence level for a normal distribution).

There are several important conclusions from the results presented in Table 5.1 and Figure 5.4. In Table 5.1 and Figure 5.4 Panels (a)-(c) we see that the parameter uncertainty is much greater when $B_2$ data are used in the objective function rather than $\rho_l$ data (in agreement with Figure 5.1). For the $\rho_l$ optimization, the numerical uncertainty is the dominant term for $T_c$, $\rho_c$, and $P_c$ as the overall uncertainty is almost identical to the numerical uncertainty. By contrast, for the
$B_2$ parameterization, both the parameter and numerical uncertainty contribute significantly to the overall uncertainties in $T_c$, $\rho_c$, and $P_c$. Also, notice that the $\rho_l$ parameterization approach yields values for $T_c$ and $\rho_c$ that agree more strongly with the experimental value. By contrast, the $B_2$ optimization results for $P_c$ are in better agreement with the experimental value. This is primarily due to the fact that $T_c$ and $\rho_c$ are predicted from the law of rectilinear diameters and the density scaling law, both of which depend strongly on $\rho_l$. On the other hand, $P_c$ is obtained from vapor pressure data (Ref [22] utilized the traditional approach with the Antoine equation) which is more closely related to a vapor phase property, such as $B_2$. However, Panels (d)-(f) demonstrate that only the $B_2$ Type B uncertainties are large enough to overlap with the experimental uncertainties for all three critical constants. By contrast, the Type A and Type B uncertainties for $\rho_l$ are approximately the same size and do not overlap with the $T_c$ and $P_c$ experimental uncertainties. As we will see in the following sections for the CH$_3$ and CH$_2$ parameters, the $\rho_l$ Type B uncertainties for longer $n$-alkanes are large enough to resolve the discrepancy with the experimental values for $T_c$ and $\rho_c$ (and $P_c$ if the Vetere approach is utilized).

5.4 Methods for UQ+PoE of CH$_3$ and CH$_2$

Since the UA CH$_4$ group does not contribute to any other compounds other than methane it does not aid in our development of a transferable potential model. The purpose for demonstrating the UQ+PoE approach for the simple case of a UA, LJ 12-6, CH$_4$ molecule is that the entire parameter space can be scanned analytically. Specifically, $\rho_l$, $B_2$, $T_c$, $P_c$ and $\rho_c$ can be calculated without significant computational costs. This facilitates calculating the PDF, generating millions of MCS parameter sets, and propagating the errors in $\varepsilon$ and $\sigma$ to $\rho_l$, $B_2$, and the critical constants. Furthermore, since analytic expressions relate the critical constants to $\varepsilon$ and $\sigma$, the parameter and numerical uncertainties for $T_c$, $P_c$, and $\rho_c$ are easily separable.

By contrast, the parameterization approach utilized for the CH$_3$ and CH$_2$ UA sites is subject to numerical uncertainty, i.e. uncertainty in the molecular simulation results for both $\rho_l$ and the critical constants. For this reason, a UQ+PoE analysis requires a slightly different approach than the general methodology presented in Section 5.2.1. The specific steps used in the UQ+PoE analysis for the CH$_3$ and CH$_2$ parameters are:
1. Create a 10x14 grid of the 2-dimensional ($\varepsilon$ by $\sigma$) parameter space

2. Perform GEMC simulations at several temperatures ($T_{\text{sim}}$) for each parameter set

3. Develop surrogate model from simulation $\rho_l$ values to predict $\hat{\rho}_l(T; \varepsilon, \sigma)$

4. Calculate $PDF$ for a refined parameter grid (500x500)

5. Generate hundreds of random numbers

6. Assign each random number to a parameter set (via the $PDF$ values)

7. Perform GEMC simulations using parameter sets sampled in Step 6

8. Obtain property values for each sampled parameter set

9. Create histogram of property values obtained in Step 8

10. Integrate histogram (or normal distribution fit) at desired confidence level

Step 3 is a very useful step for sampling from the parameter space. The surrogate model is used to interpolate between the 10x14 grid of $\varepsilon$ and $\sigma$ values that are simulated in Step 2. It is also used to interpolate between the simulated temperatures ($T_{\text{sim}}$) so that $\rho_l$ can be predicted at the experimental temperatures ($T_{\text{exp}}$). In addition, the surrogate model assures that the entire parameter space is internally consistent by smoothing out the simulation results. Essentially, the surrogate model is simply a means of reducing the number of simulations required to create the $PDF$ contours for the entire parameter space [94]. Although some analytic models have been developed for the two-site LJ 12-6 model [92], we preferred to utilize GEMC simulations to develop our surrogate model. We chose GEMC so as to be consistent with the work of Martin et al. [27] and, as discussed in Section 2.3.1, because the precision in $\rho_l$ obtained from GEMC simulations is equal to or superior to that of GCMC. Another reason for this choice is that we wanted to be able to use the same surrogate model for ethane and $n$-octane. Although some success has been found using radial basis functions [94] and the Statistical Associating Fluid Theory (SAFT) [95] as surrogate models for chain compounds, we observed that a simpler surrogate model was reliable enough to allow prediction of $\hat{\rho}_l(T; \varepsilon, \sigma)$ for ethane and $n$-octane. We have provided a detailed validation of the surrogate model in Section 5.7.
For the Type B analysis of CH₃ and CH₂ groups, Steps 1-3 and 7-8 are the same while 4-6 and 9-10 are different. Specifically, it is still necessary to predict \( \hat{\rho}_l(T_{\text{exp}}; \varepsilon, \sigma) \) for ethane and \( n \)-octane and to propagate the uncertainty in the parameters by means of molecular simulation. The difference is that, since the Type B analysis does not have a rigorous statistical derivation, there is no need to sample hundreds of parameter sets. Instead, only a few parameter sets need to be simulated to approximate the propagation of errors. Specifically, we recommend performing simulations with the extrema parameter sets that satisfy Equation 5.1, i.e. the corners and edges.

Steps 9-10 combine the numerical and parameter uncertainties to obtain the overall uncertainty. For the Type A analysis, a PDF of the overall uncertainty is obtained by fitting a normal distribution to the histogram produced in Step 9. For the Type B analysis, the numerical uncertainty is simply added to the minimum and maximum property estimates obtained with the extrema parameter sets. Specifically, the numerical uncertainties for each different parameter set are expressed as probability density functions for the desired property that can be integrated at a given confidence level. Since Steps 9-10 are very important, we outline this methodology in Section 5.5.3. We also provide the resultant histograms and PDFs in Figures F.7-F.18.

The Type B analysis greatly increases the number of acceptable parameter sets because it is no longer assumed that the optimal fit of the data is the “correct” parameter set. Instead, since a bias may exist in the data it is possible that the “correct” parameter set is one that predicts a consistently lower or higher \( \rho_l \). Furthermore, a parameter set that predicts a \( \rho_l \) curve that cuts through the experimental data is also considered acceptable as long as the curve remains within the uncertainty bands. For this reason, certain parameter sets are accepted that predict a \( \rho_l \) curve that extrapolates to a much different \( T_c \) value. Therefore, a multi-property optimization appears to be necessary for the Type B analysis to ensure that \( T_c \) is also consistent with the experimental value. It should be noted that this was the original intent of the TraPPE parameterization. That is, Martin et al. optimized their parameters such that it would minimize the deviation in \( \rho_l \) while also predicting \( T_c \) accurately [27]. The Type A analysis results in such small uncertainties in \( \rho_l \) that including \( T_c \) was unnecessary. In addition, we did not wish to arbitrarily select a weighting factor in order to develop a multi-property objective function for the PDF equation in the Type A (and A₈) analysis. Fortunately, since the Type B analysis uses a cruder uncertainty quantification approach it is much easier to include additional properties. In Section 5.5 we see that the Type
UQ analysis results in a large overlap in parameters that predict both \( \rho_l \) and \( T_c \) to within the uncertainty.

In summary, we will assess the uncertainty in the CH\(_3\) and CH\(_2\) parameters using three different approaches. The first two methods utilize the statistically rigorous Type A analysis approach. The distinction between the two methods is found in how the inherent uncertainty in the data, \( s^2 \), is assigned. We have assigned Type A to the analysis that utilizes a statistical approach for determining \( s^2 \) while Type A\(_B\) uses a Type B approach for assigning the inherent uncertainty in the data. Note that neither Type A nor Type A\(_B\) can account for bias in the data, i.e. they assume that the optimal fit of the data is the “best” model. The Type B method does not rely upon a statistical derivation but can account for bias in the data. We will demonstrate that the Type B method facilitates the multi-property aspect of obtaining a force field that predicts both \( \rho_l \) and \( T_c \) to within the DIPPR 801 uncertainties.

5.4.1 Simulation Specifications

As mentioned previously, we have replicated the work originally done by Martin et al. to parameterize \( \varepsilon_{CH_3} \), \( \sigma_{CH_3} \), \( \varepsilon_{CH_2} \), and \( \sigma_{CH_2} \) [27]. Namely, we simulated ethane at 178, 197, 217, 236, 256, and 275 K and \( n \)-octane at 390, 440, 490, 515, and 540 K. When developing the surrogate model for ethane and \( n \)-octane we performed, respectively, 12 and 9 replicate simulations at each temperature in an attempt to reduce the numerical uncertainties. By contrast, we only used two and six replicate simulations for ethane and \( n \)-octane, respectively, when simulating the MCS and extrema parameter sets. Furthermore, we simulated the MCS and extrema parameter sets for ethane at the temperatures that correspond to the experimental data. We did this to validate that our surrogate model predicted \( \rho_l \) reliably at the temperatures used in calculating RMS, i.e. \( T_{exp} \) rather than \( T_{sim} \) (see Section 5.8).

In addition, we have simulated larger \( n \)-alkanes not used in the parameterization procedure using the PoE approach. Since our primary focus is the critical constants of large \( n \)-alkanes, for these compounds we implemented the D-optimal design to minimize the numerical uncertainties in the critical constants. Because our model parameters are similar to those of the TraPPE model we used the same D-optimal temperatures as reported in Chapter 4. Specifically, we simulated \( C_{16} \) at 510 and 625 K, \( C_{24} \) at 600 and 700 K, \( C_{36} \) at 650 and 775 K, and \( C_{48} \) at 730 and 830 K.
performed 10 replicate simulations at each temperature to help decrease the numerical uncertainty for these large \( n \)-alkanes.

The equilibration periods for ethane and \( n \)-octane are 20,000 MC cycles while those for larger \( n \)-alkanes (\( C_{16} \), \( C_{24} \), \( C_{36} \), and \( C_{48} \)) are 40,000 MC cycles. The production periods are 80,000 MC cycles for ethane and \( n \)-octane and 100,000 MC cycles for larger molecules. Recall that an MC cycle consists of \( N \) moves, where \( N \) is the number of molecules. The number of molecules simulated was 800, 400, and 200 for ethane, \( n \)-octane, and larger \( n \)-alkanes, respectively.

5.4.2 Simulation Data Analysis

Estimates for \( \rho_c, T_c, P_c \), and \( Z_c \) are obtained by regressing the GEMC orthobaric densities to Equations 3.1, 3.2, 4.2, and 4.3, respectively. The numerical uncertainties for the critical constants are obtained using the nonlinear approach explained in Sections 3.3.4 and 4.3. The numerical uncertainties in the critical constants are expressed as probability density functions. For an example of these distributions, see Figures 3.3 and 4.1.

5.4.3 Reducing Numerical Uncertainty

The numerical uncertainty associated with GEMC simulations may have a significant impact on the PDF, especially close to the optimal where the contours are nearly flat. Fortunately, there are clear ways to mitigate numerical uncertainty. For example, due to advances in computational power, we have updated some of the simulation specifications reported by Martin et al. [27] to more modern values [50]. Specifically, we increased the number of molecules in the ethane simulations from 400 to 800 and in the \( n \)-octane simulations from 200 to 400. We also increased the number of MC cycles in the production period from 25,000 to 100,000, and we performed several replicate simulations at each temperature rather than a single simulation. However, numerical uncertainty remained larger than we desired.

Even a small deviation in the simulation results can lead to abnormalities in the smoothness of the minimum well because the parameter space for \( \varepsilon \) and \( \sigma \) is known to be very flat near the minimum. In fact, for this reason the optimal parameter values presented by Martin et al. were statistically the most reliable that were obtainable at that time. That is, the optimal RMS reported
by Martin et al. was approximately the same as the standard deviation in their simulation results. However, because computational power has increased dramatically since the work of Martin et al. (ca. 1998) performing a more rigorous scan of the parameter space is feasible.

For example, Mick et al. recently utilized the GOMC simulation package to produce very precise molecular simulation results [88, 96]. However, even with these incredibly reliable results the “heat-maps” they reported for the $\varepsilon$ and $\sigma$ parameter space still appear to have some small abnormalities. In other words, the contours do not have the nice smooth elliptical/spherical shape as expected. This may be due to the multi-property objective function they implement or it may demonstrate that the numerical uncertainty is still significant.

In addition to adjusting the simulation specifications (i.e. the size of simulation, number of MC cycles and MC move probabilities [50]), greater precision in vapor-liquid coexistence data (and subsequently the critical constants) is generally observed when using a more sophisticated approach known as histogram reweighting Grand Canonical Monte Carlo (GCMC) [97]. The largest improvement when comparing GCMC with GEMC is the significant decrease in uncertainty for vapor phase properties (such as vapor pressure ($P_v$) and vapor density ($\rho_v$)). However, the uncertainty in the saturated liquid density ($\rho_l$) is often very similar for the two approaches. In fact, GEMC appears to outperform GCMC in the low temperature regime [97]. Since $\rho_l$ is the only property included in the TraPPE parameterization for $n$-alkanes, we feel that there is no justification for utilizing the GCMC method over the simpler GEMC approach.

Moreover, if we allow numerical uncertainty to contribute to our results we will not be able to elucidate the uncertainty that is strictly related to the LJ parameters, and our results would quickly become outdated as future increases in computing power will continue to reduce numerical uncertainty. For these reasons, our objective is to estimate the uncertainty that is strictly associated with $\varepsilon$ and $\sigma$ in the ideal case that numerical uncertainty is negligible compared to the parameter uncertainty. As we explained previously, we utilize a surrogate model to effectively eliminate the numerical uncertainty.

Finally, it is very important to consider the numerical error introduced by extrapolating the simulation results from sub-critical temperatures to the critical point. In Chapters 3-4, we developed rigorous nonlinear approaches for quantifying and reducing the numerical uncertainty associated with predicting the critical constants obtained by regressing orthobaric densities to the
law of rectilinear diameters, the density scaling law, and the Rackett equation. This is important because the regression process typically increases the uncertainty significantly and these uncertainties can cause abnormalities in the Type A and A_B histograms. In fact, the results presented in Section 5.5 suggest that this uncertainty is at least as large as the uncertainty associated with the force field parameters.

5.5 Results for CH$_3$ and CH$_2$

In Table 5.2 we present the $RMS$ values for the optimal parameter set obtained in this work, for the 95% confidence level with the Type A, Type A_B, and Type B uncertainty analysis, and for the TraPPE parameters. There are several key observations from these results. First, the optimal parameter set obtained in this work (Mess-UP) yields $RMS$ values that are nearly an order of magnitude smaller than those of the TraPPE parameter set. Although our optimal values for $\varepsilon_{CH_3}$, $\sigma_{CH_3}$, $\varepsilon_{CH_2}$ and $\sigma_{CH_2}$ deviate from those reported by Martin et al. they are consistent considering the degree of precision with which the TraPPE parameters were reported (two and three significant figures for $\varepsilon$ and $\sigma$, respectively) [27]. As such, we do not attribute this deviation to the difference in simulation specifications but to the improved precision possible due to advances in computational power.

Table 5.2: Comparison of the root-mean-square ($RMS$, gm/mL) for the optimal parameter set obtained in this work ($\hat{\theta}_{Mess-UP}$), the 95% confidence level ($\theta$, $\alpha = 0.95$) for the Type A, Type A_B, and Type B uncertainty analysis, and the TraPPE parameters ($\theta_{TraPPE}$).

<table>
<thead>
<tr>
<th>Compound</th>
<th>$RMS(\hat{\theta}_{Mess-UP})$</th>
<th>$RMS(\theta, \alpha = 0.95)$</th>
<th>$RMS(\theta_{TraPPE})$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type A</td>
<td>Type A_B</td>
<td>Type B</td>
</tr>
<tr>
<td>Ethane</td>
<td>0.00029</td>
<td>0.00039</td>
<td>0.00172</td>
</tr>
<tr>
<td>n-Octane</td>
<td>0.00051</td>
<td>0.00068</td>
<td>0.00203</td>
</tr>
</tbody>
</table>

$\hat{\theta}_{Mess-UP}$: $\varepsilon_{CH_3} = 98.4966$ K, $\sigma_{CH_3} = 3.7491$ Å, $\varepsilon_{CH_2} = 45.3764$ K, and $\sigma_{CH_2} = 3.9724$ Å

$\theta_{TraPPE}$: $\varepsilon_{CH_3} = 98$ K, $\sigma_{CH_3} = 3.75$ Å, $\varepsilon_{CH_2} = 46$ K, and $\sigma_{CH_2} = 3.95$ Å
Second, the RMS values at the 95% confidence level for both the Type A and Type A\textsubscript{B} analysis are smaller than the RMS value for the TraPPE parameter set (although the Type A\textsubscript{B} RMS value is much larger, as expected). This suggests that the TraPPE parameter set is statistically different than the optimal parameters we obtained. However, it should be noted that Martin et al. were justified in reporting their optimal parameter set because, as they explained, the RMS was close to the average standard deviation from their simulation results. Therefore, with the simulation resources available at the time (ca. 1998) it was not possible to refine the optimal values or to obtain more significant figures.

Finally, the Type B analysis results in the largest RMS value which is larger than the TraPPE RMS values. Therefore, only by including the possibility of systematic deviations are the TraPPE parameter values considered an acceptable parameter set at the 95% confidence level. To validate these conclusions, we obtained nearly the same RMS value for the TraPPE parameter set as that reported by Martin et al. This demonstrates that our methodology, experimental data, and results are consistent with those of Martin et al. In addition, this provides a validation of the surrogate model used in this study since the TraPPE parameters were actually not simulated directly, rather the RMS values were obtained from the surrogate model. It is also worth mentioning that the Type B RMS values are simply the average uncertainties in $\rho_l$ reported by DIPPR 801.

### 5.5.1 UQ for CH\textsubscript{3} and CH\textsubscript{2}

In this section we present the uncertainty in the LJ 12-6 parameters for CH\textsubscript{3} and CH\textsubscript{2} UA groups. Figure 5.5 depicts the 95% confidence regions for the Type A, A\textsubscript{B}, and B analysis. For the Type B analysis, we present two acceptable parameter sets. The first predicts $\rho_l$ within the DIPPR uncertainty while the second predicts both $\rho_l$ and $T_c$ to within the DIPPR uncertainties. The extrema points presented in Figure 5.5 are only those for the Type B analysis with both $\rho_l$ and $T_c$. Notice that the Type A and A\textsubscript{B} confidence regions have an elliptical shape while the shape from the Type B analysis is more abnormal. An elliptical shape that is diagonal signifies that the $\varepsilon$ and $\sigma$ parameters are strongly correlated while a rectangular shape parallel to the axis means that the parameters are completely uncorrelated. It is interesting to notice that the correlation between $\varepsilon$ and $\sigma$ is different for the CH\textsubscript{3} parameters than for the CH\textsubscript{2} parameters. We provide a preliminary investigation of this phenomenon in Section 5.7.2.
Figure 5.5: Comparison of different uncertainty quantification methods. Panels (a) and (b) correspond to the CH$_3$ and CH$_2$ LJ parameters, respectively. Each uncertainty analysis region is presented at the 95% confidence level. The extrema parameter sets are those from the Type B analysis when both $\rho_l$ and $T_c$ uncertainties are considered. The TraPPE, NERD, and optimal (Mess-UP) parameters are included as a reference.

Clearly the Type A analysis results in a significantly smaller uncertainty region. The Type B confidence region for $\rho_l$ encompasses the Type $A_B$ analysis but extends much further. Even with $T_c$ included, the Type B analysis results in relatively large uncertainty regions. However, certain parameter sets that are accepted with a Type $A_B$ analysis lie outside the multi-property Type B acceptable region. These parameter sets (the Type $A_B$ parameters that are not encompassed by the Type B acceptance region) result in a poor prediction of $T_c$. This demonstrates one advantage of using a multi-property Type B analysis. It is also significant that only a Type B analysis suggests that the TraPPE and NERD CH$_2$ parameters are acceptable. This is important because as the ratio of CH$_2$ to CH$_3$ groups increases the contributions from the CH$_3$ sites will become negligible and only the CH$_2$ parameter uncertainties should contribute. We can thus conclude, without even performing a Type B PoE analysis (i.e. simulating at the extrema), that the TraPPE and NERD models will predict indistinguishable results for larger $n$-alkanes. Therefore, the critical constant
values that we reported in Section 4.11 for large $n$-alkanes using the TraPPE and NERD model are both acceptable predictions.

### 5.5.2 MCS for CH$_3$ and CH$_2$

In this section we demonstrate the MCS results for $\varepsilon$ and $\sigma$ parameter sets. Figure 5.6 presents the Type A and Type A$_B$ single parameter histograms obtained from millions of random numbers and the respective $PDF$ values. Panels (a)-(d) correspond to $\varepsilon_{CH_3}$, $\sigma_{CH_3}$, $\varepsilon_{CH_2}$, and $\sigma_{CH_2}$, respectively. From Figure 5.6 it is clear that the Type A$_B$ single parameter uncertainties are significantly larger than the Type A uncertainties (consistent with Figure 5.5). Furthermore, the histograms for both uncertainty types and each LJ parameter result in a normal distribution.

Since $\varepsilon$ and $\sigma$ are strongly correlated, it is common to create two-dimensional plots of the MCS parameter sets. When the $PDF$ contours are also included, these plots can verify that the MCS parameter sets represent a proper sampling from the $PDF$. For example, Figure 5.7 demonstrates the Type A UQ and MCS results for $\varepsilon_{CH_3}$-$\sigma_{CH_3}$. The parameter uncertainty (or $PDF$) is presented as contour plots with multiple regions representing different confidence levels. Notice that the contour regions are the 81% (90$^2$%), 90% ($\approx$ 95$^2$%), and 95% confidence levels. Therefore, approximately 5% of the MCS parameter sets should be found beyond the outer confidence region. Although only 4% of the 2000 MCS parameter sets presented in Figure 5.7 are outside the 95% confidence region, this is an acceptable result considering the small sample size. Thus, the results from Figure 5.7 suggest that we have correctly implemented the MCS methodology.

Figure 5.8 shows the UQ and MCS results for the $\varepsilon_{CH_3}$-$\sigma_{CH_3}$ and $\varepsilon_{CH_2}$-$\sigma_{CH_2}$ parameter sets. Panels (a)-(b) and (c)-(d) contain the CH$_2$ and CH$_3$ parameters, respectively. Panels (a) and (c) were obtained with a Type A analysis while Panels (b) and (d) represent a Type A$_B$ analysis. Note the difference in the range as the uncertainty region is much larger for the Type A$_B$ analysis. Included in Figure 5.8 are the 100 MCS parameter sets for each compound included in the PoE procedure. Also, notice that the CH$_3$ parameters for $n$-octane (C$_8$) were set as the optimal parameter set ($\varepsilon_{CH_3} = 98.4966$ and $\sigma_{CH_3} = 3.7491$). By only sampling from the CH$_2$ parameter space we can determine how much the uncertainty in just the CH$_2$ parameters affects our results for $n$-octane. Since $n$-octane was used in the parameterization process, the PoE results of $\rho_l$ and $T_c$ for $n$-octane can also validate our results for the CH$_2$ parameters.
Figure 5.6: Histograms of CH$_3$ and CH$_2$ LJ parameters from MCS Type A and A$_B$ analysis. Panels (a)-(d) correspond to $\varepsilon_{\text{CH}_3}$, $\sigma_{\text{CH}_3}$, $\varepsilon_{\text{CH}_2}$, and $\sigma_{\text{CH}_2}$, respectively. The normal distribution fits are also included. Probability density is defined as the number of counts in a single bin divided by both the total number of counts and the bin width.

The main conclusion from Figure 5.8 is that the CH$_3$ and CH$_2$ MCS parameter sets are an accurate representation of the PDF values for both the Type A and Type A$_B$ analysis. Specifically,
the number of parameter sets agrees reasonably well with the confidence level for a given contour. Furthermore, the spatial distribution does not suggest an improper bias towards a specific region of parameter space.

In the case of the Type A analysis of ethane we simulated 2000 MCS sets of $\varepsilon_{CH_3}$ and $\sigma_{CH_3}$. The amount of $\varepsilon$ and $\sigma$ values randomly sampled is limited by the computational cost of the system. Fortunately, we did not observe a significant difference between the Type A results obtained with 100 and 2000 parameters sets for ethane (see Section 5.5.3). Therefore, we only used 100 MCS parameter sets for every other system studied with the Type A and AB analysis in order to significantly reduce the computational cost.

Recall that for the Type B analysis we only simulate the eight extrema points presented in Figure 5.5. The extrema parameter sets are listed in Table 5.3. Although sampling the eight extrema points for both CH$_3$ and CH$_2$ parameter sets would necessitate simulating sixty-four parameter sets, the CH$_3$ uncertainty was found to have a negligible effect on the larger $n$-alkanes. This was verified by simulating C$_{16}$ at the CH$_3$ extrema points while maintaining the CH$_2$ parameters at their optimal values (see Figure F.19). Therefore, we only used the optimal CH$_3$ parameters for C$_{24}$, C$_{36}$ and C$_{48}$ to reduce the number of simulations in the Type B analysis. Furthermore, we did not vary the CH$_3$ parameters for $n$-octane since we wanted to verify that the deviations in $\rho_l$
caused by CH₂ were the same as the DIPPR uncertainties. Similarly, we have provided the CH₃ extrema parameter sets in Table 5.3 to validate the ρ_l uncertainties for ethane.

Rather than simply reporting a single optimal set of force field parameters, we recommend that future force field development studies report several parameter sets. This will enable future researchers to approximate the uncertainties in a property of interest as a result of the propagation
Table 5.3: Extrema parameter sets used in Type B analysis.

<table>
<thead>
<tr>
<th>$P_i$</th>
<th>$\varepsilon_{\text{CH}_3}$ (K)</th>
<th>$\sigma_{\text{CH}_3}$ (Å)</th>
<th>$\varepsilon_{\text{CH}_2}$ (K)</th>
<th>$\sigma_{\text{CH}_2}$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>98.085</td>
<td>3.7582</td>
<td>44.645</td>
<td>3.9920</td>
</tr>
<tr>
<td>$P_2$</td>
<td>98.500</td>
<td>3.7340</td>
<td>44.740</td>
<td>4.0091</td>
</tr>
<tr>
<td>$P_3$</td>
<td>98.780</td>
<td>3.7680</td>
<td>45.240</td>
<td>3.9570</td>
</tr>
<tr>
<td>$P_4$</td>
<td>99.050</td>
<td>3.7510</td>
<td>45.660</td>
<td>3.9770</td>
</tr>
<tr>
<td>$P_5$</td>
<td>98.270</td>
<td>3.7452</td>
<td>45.850</td>
<td>3.9303</td>
</tr>
<tr>
<td>$P_6$</td>
<td>98.950</td>
<td>3.7580</td>
<td>46.052</td>
<td>3.9472</td>
</tr>
<tr>
<td>$P_7$</td>
<td>98.380</td>
<td>3.7645</td>
<td>44.685</td>
<td>4.0025</td>
</tr>
<tr>
<td>$P_8$</td>
<td>98.830</td>
<td>3.7430</td>
<td>45.950</td>
<td>3.9385</td>
</tr>
</tbody>
</table>

of errors in the LJ parameters. For the Type A and Type A$_B$ approaches we recommend that at least 100 parameter sets be reported that are obtained from Monte Carlo Sampling. For this reason, we have provided lists of MCS parameter sets in Table E.1. For the Type B analysis we believe it is sufficient to report the 8 parameter sets from the extrema as reported in Table 5.3. We consider the collective $\varepsilon$ and $\sigma$ parameter sets as the Mess-UP force field. In the subsequent sections, we demonstrate how the Mess-UP parameter sets can be used with the PoE approach to determine the uncertainty in the critical constants of large $n$-alkanes.

5.5.3 PoE for CH$_3$ and CH$_2$

The uncertainty in the LJ parameters leads to uncertainties in the predicted vapor-liquid coexistence curve and critical point constants. In Subsection 5.5.3 we demonstrate the Type A, Type A$_B$, and Type B uncertainties in $\rho_l$ and $T_c$ for ethane and $n$-octane. These uncertainties are obtained from the propagation of errors method detailed in Section 5.4. We provide an example of the PoE methodology for the 2000 parameter sets sampled with the Type A analysis of ethane. We then validate that our force field model and parameter uncertainties do in fact match the experimental data values within the stated uncertainties. This discussion focuses on ethane and $n$-octane as these are the compounds used for optimizing the force field parameters. Finally, we implement the propagation of errors approach for larger compounds. Specifically, we present the uncertainties in $\rho_l$, $T_c$, $\rho_c$, $P_c$, and $Z_c$ for C$_{16}$, C$_{24}$, C$_{36}$ and C$_{48}$ obtained from the Type A, Type A$_B$, and Type
B analysis. This is done to conclude whether or not the uncertainties in large $n$-alkanes are small enough to provide quantitative insight.

**Type A Ethane**

In Figure 5.9 we present the Type A uncertainties in $T_c$ for ethane. Panel (a) contains a histogram of the $T_c$ estimates obtained from the MCS parameter sets. It is important to mention that these estimates have numerical uncertainty since they were obtained by regressing simulation results to the law of rectilinear diameters and the density scaling law. In Panel (b) we compare the numerical, parameter, and overall uncertainties. The numerical uncertainty represents the average numerical uncertainty from regressing the simulation results. The Type A parameter uncertainty is obtained by predicting $T_c$ with the surrogate model for millions of MCS parameter sets. We present two different approximations for the overall uncertainty. The MC Sampling overall uncertainty is obtained by fitting a normal distribution to the histogram presented in Panel (a) for 2000 MCS parameter sets. The other overall uncertainty is obtained using Equation 5.4 and the numerical and parameter uncertainties.

In Figure 5.9 we see that the surrogate model has a slight bias in the estimate for $T_c$. However, this is of lesser importance for our purposes since we are primarily concerned with the uncertainty size. The surrogate model results provide an estimate for the parameter uncertainty without any numerical uncertainty. Notice that the overall uncertainty obtained from the MCS PoE approach is very similar to that obtained from combining the numerical and parameter uncertainties with Equation 5.4. This demonstrates that the MCS PoE approach properly accounts for both numerical and parameter uncertainties.

Another significant observation from Figure 5.9 Panel (a) is that our results with 2000 samples were not much different from our results with only 100 samples. In other words, the histogram and the normal distribution fit were nearly identical. In Figure 5.10 Panels (a)-(c) we see that this is also true for $\rho_c$, $P_c$, and $Z_c$, respectively. For this reason, we only sampled 100 parameter sets for the other systems studied.

As $\rho_c$, $P_c$, and $Z_c$ are not used in the force field parameterization, our surrogate model was not devised to predict these properties accurately. Therefore, we do not attempt to elucidate the contributions from numerical and parameter uncertainties. However, since the numerical uncer-
Figure 5.9: Comparison of the uncertainties in predicting $T_c$ for ethane. Panel (a) compares the histograms obtained using 100 and 2000 MCS parameter sets. The normal distribution fits are also included. Probability density is defined as the number of counts in a single bin divided by both the total number of counts and the bin width. Panel (b) depicts the $T_c$ probability density functions for the numerical, parameter, and overall uncertainties. The numerical uncertainty represents the average numerical uncertainty obtained from the simulations. The surrogate model is used to estimate the parameter uncertainties. The overall uncertainties are obtained in two different ways, as explained in the text.

Ethane and $n$-Octane

In Figure 5.11 we present the experimental $\rho_l$ data (along with the DIPPR correlation), the values from the TraPPE model, and the results from this work. Notice that the Type A/A$_B$ results from this work are represented as a single point at each temperature simulated. These points are the average value obtained from 100 MCS parameter sets. The error bars for each data set are reported at the 95% confidence level. The uncertainties for the correlation to the experimental data are those from the DIPPR 801 database (which implements a Type B analysis). The 95% confidence intervals for the TraPPE results are approximately one symbol size. However, the TraPPE uncertainties for ethane were obtained from the standard error reported by Eggimann et al.
Figure 5.10: Comparison of the histograms obtained using 100 and 2000 MCS parameter sets for ethane. Panels (a)-(c) correspond to $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are also included. The probabilities are defined as the number of counts in a single bin divided by the total number of counts.

There are several key conclusions from Figure 5.11. First, the Mess-UP model predicts $\rho_l$ very accurately over the entire temperature range. (The average value of 100 MCS parameter sets adequately approximates the value that would be obtained with the optimal parameter set.) Notice that the TraPPE model has a clear bias (consistently lower) when compared with the experimental data and correlation (although it is within the correlation uncertainties). Second, the uncertainty regions obtained from the Type A, Type A$_B$, and Type B analysis methods vary greatly in magnitude. For example, even the Type A$_B$ uncertainty analysis results in error bars that are still quite small relative to the correlation uncertainty. This demonstrates that uncertainties obtained from strictly statistical means (Type A and A$_B$) are often smaller than justifiable. By contrast, the Type
Figure 5.11: Comparison of $\rho_l$ experimental data (and a correlation fit to these data) with the TraPPE model and the results from this work. Panels (a) and (b) present results for ethane and $n$-octane, respectively. The correlation uncertainties are those from the DIPPR 801 database. The error bars represent the 95% confidence interval. The numerical uncertainties for the TraPPE model and the Type B results are approximately one symbol size (except for $n$-octane at 540 K). Only the uncertainties for the Type A analysis are included for the MCS results. The error bars for the Type A analysis are approximately half the size of the Type A error bars shown.

B analysis results in $\rho_l$ values that follow the DIPPR correlation uncertainties for the entire temperature range. This validates the assumption that the minimum and maximum values obtained by simulating at the extrema parameter sets adequately reproduces the DIPPR uncertainties. Therefore, we believe that the Type B analysis produces the most meaningful assessment of parameter uncertainty. For this reason, we recommend that future simulation studies utilize the DIPPR 801 uncertainties when performing a UQ+PoE analysis of the force field parameters.

Having demonstrated that our model accurately predicts $\rho_l$ for ethane and $n$-octane, we shift the focus to $T_c$. In Figure 5.12 we present the PoE results of $T_c$ for ethane and $n$-octane. The uncertainty regions are presented as probability density functions. In Figure 5.12 Panel (a) we have included the experimental, TraPPE, Type A, Type $A_B$, and Type B uncertainties. The experimental uncertainties are obtained by assuming the DIPPR uncertainties are the 95% confidence interval for a normal distribution of error [17]. The TraPPE results only account for numerical uncertainties. The TraPPE validation data were used for ethane [72] while the D-optimal design results from
Chapter 4 were used for \textit{n}-octane. The TraPPE uncertainties and the numerical uncertainties in our results are obtained from the rigorous nonlinear analysis methods outlined in Sections 3.3.4 and 4.3. The Type A and Type A$_B$ uncertainties are obtained from the MCS simulations. The distributions displayed in Panel (b) represent the numerical uncertainty for a given extrema parameter set ($P_i$) (see Table 5.3). A conservative estimate for the overall uncertainty with a Type B analysis is taken as the minimum of the lower 95\% confidence intervals and the maximum of the upper 95\% confidence intervals. The Type B results in Panel (a) are presented as a region of constant probability ranging from the lowest to the highest values predicted by the extrema parameter sets.

Figure 5.12: Comparison of the $T_c$ probability densities for ethane and \textit{n}-octane. Panel (a) presents the uncertainties in the experimental data, TraPPE model, and this work. The experimental uncertainties are obtained by assuming the DIPPR uncertainties are the 95\% confidence interval for a normal distribution of error [17]. The TraPPE uncertainty regions for ethane and \textit{n}-octane were obtained in Chapters 3-4, respectively. The Type A and Type A$_B$ results are obtained by fitting a normal distribution to a histogram produced from the 100 MCS parameter sets (2000 in the case of the Type A analysis of ethane). The Type B uncertainty includes the numerical uncertainties of the extrema parameter sets. These numerical uncertainties are presented in Panel (b) for each of the extrema parameter sets ($P_i$).

There are several key observations from Figure 5.12. First, the best estimates (the $T_c$ that corresponds to the peak maximum) for the Mess-UP model are closer to the experimental values
than the best estimates for the TraPPE model. In fact, both the Type A and A\(_B\) peaks overlap considerably with the experimental peaks. Furthermore, the Type B uncertainty spans the entire experimental uncertainty. This demonstrates that it is sufficient to take the minimum and maximum values from the extrema to represent the Type B uncertainty. The Type B range is slightly wider than the experimental uncertainty region because the Type B range includes the numerical uncertainty (at the 95% confidence level) and the 1% uncertainty attributed to the density scaling law parameter (\(\beta\)) [27]. Collectively, these results demonstrate that the UA, LJ 12-6 model is adequate to predict both \(\rho_l\) and \(T_c\) to within the given uncertainty for ethane and \(n\)-octane.

It is also important that the uncertainty in \(T_c\) increases only marginally between the Type A and Type A\(_B\) analysis for \(n\)-octane. This supports the notion that numerical uncertainties are a significant contributor to the overall uncertainty in \(T_c\) for larger compounds. This is significant because the numerical uncertainty for \(T_c\) is the smallest of the four critical constants. Therefore, as previously proposed, it can be expected that the numerical uncertainties in \(\rho_c\), \(P_c\), and \(Z_c\) are larger than their respective parameter uncertainties for the larger \(n\)-alkanes.

**\(C_{16}, C_{24}, C_{36}, \text{ and } C_{48}\)**

In Figure 5.13 we present the \(\rho_l\) results for \(C_{16}, C_{24}, C_{36}, \text{ and } C_{48}\). Notice that experimental data are scarce (and have large uncertainties) for these larger compounds in the temperature range considered. For this reason, the DIPPR correlations are provided to help extrapolate the low temperature \(\rho_l\) data to which the correlations were fitted (although \(C_{48}\) is not included in the DIPPR 801 database). For clarity, we have assigned the DIPPR 801 uncertainties to the experimental data rather than to the DIPPR correlations. We have also included “Simulation Correlations” to guide the eye. The “Simulation Correlations” were obtained for the Mess-UP model with the traditional approach of regressing the simulation results to Equations 3.1-3.2. The TraPPE and NERD results presented in Figure 5.13 were obtained in Chapter 4 and are considered the most accurate simulations with these models for the compounds in question. The error bars are the same as those in Figure 5.11 where the 95% confidence intervals for the TraPPE and Type A analysis are approximately one symbol size. Once again, the error bars for the TraPPE model are limited to numerical uncertainties.
Figure 5.13: Comparison of $\rho_l$ experimental data (and a correlation fit to experimental data), the TraPPE and NERD results reported in Chapter 4 (Messerly et al.), and the PoE approach in this work for $C_{16}$, $C_{24}$, $C_{36}$, and $C_{48}$. The error bars for this work represent the 95% confidence interval for the Type $A_B$ analysis. The DIPPR 801 uncertainties are presented for the experimental data. The error bars for the Type A analysis and the numerical uncertainties for the Type B, TraPPE, and NERD results are approximately one symbol size.

We again observe a systematic improvement in the $\rho_l$ values towards the DIPPR 801 correlations when comparing the results for the TraPPE and Mess-UP parameters. Although both the Mess-UP and the TraPPE model agree with the experimental data to within the DIPPR uncertainties, there does appear to be a slight bias in the TraPPE results. As proposed in Section 5.5.1, the Type B lower and upper bounds agree almost identically with the NERD and TraPPE models, respectively. It is significant that the Type B uncertainties are substantially smaller than the DIPPR uncertainties. This suggests that the DIPPR correlations for $\rho_l$ of these larger $n$-alkanes are actually more reliable than DIPPR 801 staff have presumed.

In Figure 5.14 we present the trends for the critical constants $T_c$, $\rho_c$, $P_c$, and $Z_c$ with respect to carbon number. We compare multiple sources of experimental data [12, 15, 20], several prediction models [9, 10, 12–16, 29], and the Mess-UP simulation results. The uncertainties are reported at the 95% confidence level. Notice that both authors that reported experimental data employed a Type B uncertainty analysis scheme which results in large error bars for certain properties and compounds. Recall that the Type A and $A_B$ uncertainties were obtained by performing
GEMC simulations at 100 MCS parameter sets while only the eight CH$_2$ extrema parameter sets are simulated for the Type B uncertainties. In Figure 5.14 we have only included the largest overall uncertainty obtained either with the Type A, A$_B$, and B analysis methods (which is generally that from the Type B analysis). In Figures F.7-F.18 we provide histograms and PDFs of the Type A, Type A$_B$, and Type B uncertainties for each compound studied. The key conclusions from Figure 5.14 are that the simulation results are reliable enough to elucidate the correct long-chain-length trends even when both the numerical and parameter uncertainties are included. In fact, we arrive at the same conclusions as those in Chapter 4. Specifically, the experimental data reported in 2014 by Nikitin et al. for $T_c$ and $P_c$ are more reliable than the values originally reported [15, 20], $\rho_c$ continues to decrease with increasing carbon number, and $Z_c$ appears to be approaching an asymptote around $\frac{1}{5}$. However, we will see in Chapter 6 that $Z_c$ approaches a minimum of $\frac{1}{5}$ before increasing to an asymptote of $\frac{1}{3}$.

5.6 Limitations

In this study we have observed some of the deficiencies in the TraPPE-UA parameterization approach. Specifically, this model over predicts $T_c$ and $\rho_l$ at higher temperatures for larger $n$-alkanes. This is true even for the more refined LJ 12-6 optimal parameters we obtained. The deficiency can be traced back to several possible causes. First, the experimental $\rho_l$ data for ethane and $n$-octane used in the parameterization process may be inaccurate, although the Type B data analysis should account for this possibility. Second, the assumption that the CH$_3$ parameters transfer from ethane to $n$-octane may cause the CH$_2$ parameters to compensate and deviate from their “true” values. A related issue is that the Lorentz-Berthelot (LB) combining rules can place a non-physical constraint on the CH$_2$ parameter results. This study does not consider whether or not the LB combining rules lead to poor extrapolation for larger $n$-alkanes as the ratio of CH$_3$ to CH$_2$ groups changes. In general, model uncertainties may be a concern when predicting properties for compounds and temperatures not included in the parameterization process. For example, the assumption that the CH$_3$ and CH$_2$ parameters transfer to larger $n$-alkanes may not be valid.

The largest potential limitation to consider is the correlation between the CH$_3$ and CH$_2$ parameter values. The contours in Panels (a) and (b) of Figure 5.8 were obtained using the optimal set of CH$_3$ parameters. In essence, we have assumed that the uncertainty in the CH$_3$ optimal
Figure 5.14: Comparison of critical constant trends with respect to carbon number for conflicting experimental data, prediction models, and the simulation results from this work. Panels (a)-(d) correspond to $T_c$, $P_c$, $\rho_c$, and $Z_c$, respectively. Panel (a) excludes smaller $n$-alkanes because the $T_c$ results for ethane and $n$-octane were presented previously in Figure 5.12. The error bars represent the 95% confidence interval. The uncertainties reported from the respective authors were used for the experimental data. The uncertainties for Mess-UP are the largest uncertainty obtained with the Type A, A_B, or B analysis.

parameters from ethane are small enough to have a negligible effect on the CH$_2$ parameters. This assumption reduces the four dimensional parameter space to a pair of independent two dimensional parameter spaces. Therefore, the parameter sets shown in Figure 5.8 were obtained by
independently sampling from the PDFs for the CH₃ and CH₂ parameter spaces. In Figure F.19 we demonstrate that the uncertainty in the CH₃ parameters have a negligible effect on the critical constants for C₁₆. However, the impact of the CH₃ parameter uncertainty on ρ_l for n-octane (which was used in the parameterization process) is likely much larger.

In this section and Section 2.3.2 we discussed some of the possible limitations due to not accounting for the model uncertainties. Since we are not concerned with vapor phase and/or transport properties, we believe that the model uncertainties resulting from the choice of a UA LJ 12-6 model are negligible. However, this study does not account for deeper layers of molecular simulation uncertainty such as the intramolecular model, i.e. rigid or flexible bonds, equilibrium bond lengths, angles, etc. Since other studies have shown that these properties tend to have a small effect on the vapor-liquid coexistence curve we have assumed that the intramolecular model used by the TraPPE force field is sufficiently reliable [66]. In addition, the use of a surrogate model introduces some inexactness into the approach. For this reason, in the following section we provide details regarding the development and validation of our surrogate model.

5.7 Surrogate Model

To date, no exact method exists to predict \( \hat{\rho}_l(T_i; \varepsilon, \sigma) \) for ethane and n-octane. Instead, molecular simulation (in this study, GEMC) is the chosen means of estimating \( \hat{\rho}_l(T_i; \varepsilon, \sigma) \). This presents a challenge that is unique to molecular simulation parameterization. Specifically, in most optimization problems the evaluation of \( \hat{y} \) (and RMS, PDF, etc.) is both cheap and exact. By contrast, molecular simulations are expensive and subject to random error.

In order to overcome both of these obstacles, we implement a surrogate model in this study to predict \( \hat{\rho}_l(T_i; \varepsilon, \sigma) \) and, thereby, evaluate \( RMS(\theta) \) and \( PDF(\theta) \). There are two key advantages of using a surrogate model. First, a surrogate model effectively eliminates the numerical uncertainty, although it may introduce some bias. Second, a surrogate model can be used in a propagation of errors analysis to sample the parameter space. In this section, we discuss the development and validation of the surrogate model used in this chapter.
5.7.1 Development

There are at least three different types of surrogate models used in the simulation literature. The first type is a model that can predict a physical property for a given set of force field parameters [92]. This type of surrogate model is useful because it can reduce the need for performing molecular simulations. For example, the PC-SAFT model utilizes molecular simulations in conjunction with an EOS to optimize the intermolecular potential parameters [95]. The second type of surrogate model simply provides a means of interpolating the $RMS$ for the parameter space (in this case, $\varepsilon$ and $\sigma$) [94]. With this type of model it is no longer necessary to evaluate $RMS$ at every possible $\varepsilon$ and $\sigma$. This is beneficial because evaluating $RMS$ requires performing several simulations for a given set of $\varepsilon$ and $\sigma$.

In this work we utilize a third type of surrogate model referred to by Hulsmann et al. as “Lipra” (linear property approximation) [94]. Rather than interpolate the $RMS$ itself, the Lipra approach interpolates/smooths the $\rho_l$ values at a given temperature with respect to $\varepsilon$ and $\sigma$. By smoothing the $\rho_l$ results as a function of $\varepsilon$ and $\sigma$ we eliminate any outliers and obtain a more continuous contour for $\rho_l$. This produces a more internally consistent set of $\rho_l$ values and thereby reduces the effect of numerical uncertainty in the GEMC results.

The second aspect of our surrogate model is designed to further reduce the numerical uncertainty and to allow for estimating $\rho_l$ at any temperature. This is done by regressing the smoothed $\rho_l$ values for every set of $\varepsilon$ and $\sigma$ to the function previously proposed in Chapter 4

$$\rho_l = \rho_0 + A(T_0 - T) + B(T_0 - T)^\beta$$ (5.5)

where $\rho_0$, $T_0$, and $\beta$ are simply considered fitting parameters. In this case, Equation 5.5 is merely a way of interpolating the simulation results for $\rho_l$ and, thus, $\rho_0$ and $T_0$ are not given any physical interpretation. By using Equation 5.5 we are able to evaluate $\rho_l$ at any temperature (and, thereby, $RMS$ and $PDF$ for any set of $\varepsilon$ and $\sigma$).

In summary, the key steps in the development of our surrogate model are:

1. Create a grid of $\varepsilon$ and $\sigma$ values.
2. Obtain $\rho_i$ at the same temperatures reported by Martin et al. using GEMC simulations for every set of $\varepsilon$ and $\sigma$ values chosen in Step 1.

3. Perform a double interpolation to obtain $\rho_i$ for any possible set of $\varepsilon$ and $\sigma$ but strictly at the temperatures simulated in Step 2.

4. Regress the smoothed $\rho_i$ values from Step 3 to Equation 5.5.

In Figure 5.15 we see the improvement in the $RMS$ contours obtained by using the surrogate model for ethane and $n$-octane. Panels (a) and (c) are used with the Type $A_B$ and B analysis methods while Panels (b) and (d) are utilized with the Type A approach. Notice the difference in the axis ranges in Panels (a)-(b) and (c)-(d) due to the order of magnitude difference in the parameter space required for a Type $A$ analysis compared to the Type $A_B$ and B analysis. In each case the $\varepsilon$ by $\sigma$ grid used in simulation was 10 by 14, where the values were evenly spaced between their respective minimum and maximum values as depicted. It is significant that Panels (a) and (c) were obtained by performing a single simulation at each temperature for a given parameter set. By contrast, we performed 12 and 9 replicate simulations at each temperature to generate the results in Panels (b) and (d), respectively. A large number of replicates were used in an attempt to mitigate the numerical uncertainty that becomes significant when generating contours in such a narrow region of parameter space close to the minimum.

5.7.2 Validation

The aforementioned methodology is admittedly risky due to multiple layers of smoothing and correlating simulation results. For this reason, we believe a validation of our surrogate model is essential. We have implemented several different methods to accomplish this task. First, we compare the contours obtained from the raw simulation data and the surrogate model obtained from the algorithm described in Section 5.7.1. In Figures 5.16-5.19 we include each of the temperatures simulated for ethane and $n$-octane. Figures 5.16-5.17 represent ethane $\rho_i$ while Figures 5.18-5.19 are for $n$-octane $\rho_i$. Figures 5.17 and 5.19 used a very refined grid near the optimum for the Type $A$ analysis while Figures 5.16 and 5.18 scan the much wider range of $\varepsilon$ and $\sigma$ required for the Type $A_B$ and B analysis methods. Included in these figures are each of the temperatures simulated for

132
Figure 5.15: Contours of RMS with respect to $\varepsilon$ and $\sigma$. Panels (a)-(b) correspond to the CH$_3$ contours for ethane while Panels (c)-(d) correspond to the CH$_2$ contours for n-octane. The green and black lines represent the contours for the simulation results and the surrogate model, respectively. Panels (a) and (c) span the much wider parameter space required for a Type $A_B$ and $B$ analysis. Panels (b) and (d) span the parameter space required for a Type $A$ analysis.

developing the surrogate model (i.e. the same temperatures as those used by Martin et al.). The main purpose of these figures is to see that all of the anomalies appear to have been eliminated. We reiterate that Figures 5.16 and 5.17 were obtained by performing a single simulation at each temperature while Figures 5.17 and 5.19 used 12 and 9 replicate simulations, respectively. The similarity between the surrogate models obtained with the coarse and refined grids suggests that the significant computational cost of performing numerous replicate simulations was not merited for developing our surrogate model.
Figure 5.16: Contours of $\rho_l$ for ethane with respect to $\epsilon_{CH_3}$ and $\sigma_{CH_3}$ (coarse grid). The green and black lines represent the contours for the simulation results and the surrogate model, respectively. Panels (a)-(f) correspond to 178, 197, 217, 236, 256, and 275 K.

We believe that Figures 5.16-5.19 are sufficient justification that our surrogate model is a reliable representation of the simulation results. However, we will briefly discuss some of the other methods we employed for validating our model. The second approach, as discussed in Section 5.5, is to verify that the RMS from our surrogate model for the optimal TraPPE parameters is very close to the RMS reported by Martin et al. Third, we verified that the 95% joint confidence region predicted by the surrogate model accurately represents that obtained by the simulation data. In other words, we used Equation 2.16 to analyze the parameter sets that were simulated and determine if they would be accepted at the 95% confidence level. Then, we compared these acceptable parameter sets with those that the surrogate model predicted would be acceptable. We observed that the same parameter sets were acceptable for both methods. Finally, we compared the propagation of error results for both the surrogate model and the simulation results. Specifically, we evaluated the uncertainties in $\rho_l$ and $T_c$. We observed that the uncertainty in $\rho_l$ obtained from the simulations agreed with the predicted uncertainty. For example, notice that the Type B results in Figures 5.11-
Figure 5.17: Contours of $\rho_l$ for ethane with respect to $\varepsilon_{CH_3}$ and $\sigma_{CH_3}$ (refined grid). The green and black lines represent the contours for the simulation results and the surrogate model, respectively. Panels (a)-(f) correspond to 178, 197, 217, 236, 256, and 275 K. Notice the difference in axis ranges compared with Figure 5.16.

5.12 follow the DIPPR uncertainties. Therefore, the surrogate model accurately determines the extrema parameter sets. Furthermore, recall that in Figure 5.9 the overall $T_c$ uncertainty obtained from propagating the numerical and parameter (obtained from the surrogate model) uncertainties was very similar to that from the actual simulations (using MCS).

Apart from validating our surrogate model, a great deal of molecular insight can be teased out from Figures 5.16-5.19. Specifically, notice the contour directionality, i.e. the dependence of $\sigma$ upon $\varepsilon$ for a constant $\rho_l$. The trend for ethane is uniform for all temperatures. That is, an increase in $\varepsilon$ necessitates an increase in $\sigma$ to maintain constant $\rho_l$. This makes sense because increasing the size of the UA sites ($\sigma$) is required to maintain a constant density if the attraction is increased (caused by increasing $\varepsilon$). However, notice that for n-octane (used to obtain the $CH_2$ parameters), the correlation between $\varepsilon$ and $\sigma$ is temperature dependent. To be specific, at low temperatures we observe the same trend as for ethane while at high temperatures the trend is exactly opposite. At high temperatures an increase in attraction ($\varepsilon$) requires a decrease in size ($\sigma$) to maintain
Figure 5.18: Contours of $\rho_l$ for $n$-octane with respect to $\varepsilon_{CH_2}$ and $\sigma_{CH_2}$ (coarse grid). The CH$_3$ parameters are set to the optimal values reported in text. The green and black lines represent the contours for the simulation results and the surrogate model, respectively. Panels (a)-(e) correspond to 390, 440, 490, 515, and 540 K.

Constant density. In addition, due to this change in correlation, at an intermediate temperature (440 K) there is no correlation between $\varepsilon$ and $\sigma$. Surprisingly, at this temperature the liquid density is insensitive to $\sigma$! In other words, $\rho_l$ is solely dependent upon $\varepsilon$. The lack of dependence upon $\sigma$ at intermediate temperatures and the inverse relationship between $\varepsilon$ and $\sigma$ at higher temperatures are completely counter-intuitive results. We believe that in order to understand these conclusions it is important to remember that these are saturated $\rho_l$ contours. In other words, the dependence of vapor chemical potential upon $\varepsilon$ and $\sigma$ needs to be considered. The only explanation that we can provide is that the importance of the vapor chemical potential dependence upon $\varepsilon$ and $\sigma$ increases with temperature for larger molecules.

5.8 Data Evaluation

Since obtaining accurate and precise $\varepsilon$ and $\sigma$ values necessitates reliable and evaluated experimental data, in this section we detail the data that are used in calculating $RMS$ and $PDF$. 

136
Our initial objective was to exactly replicate the work originally done by Martin et al. and, thus, use the same set of data. However, for the Type A and $A_B$ analysis, we found that it was necessary to utilize a slightly different set of data. There are two key differences between the data used by Martin et al. and those used in the Type A and $A_B$ analysis. First, we used solely experimental data. Second, we utilized a larger set of data. In the subsequent paragraphs we explain why these modifications were necessary.

For Equation 2.17 (used in the Type A and $A_B$ analysis) to be valid it is essential that the data used in the objective function ($RMS$) have normally distributed error. This poses a problem for the TraPPE methodology because the TraPPE parameters were optimized to a TRC (Thermodynamics Research Center) correlation for $\rho_l$ rather than truly experimental data. Fortunately, in 2002 highly accurate $\rho_l$ data were measured for ethane that agree strongly with the TRC correlation [98]. Therefore, we utilized these experimental data in our parameter optimization and subsequent uncertainty quantification. Similarly, we have utilized evaluated experimental $\rho_l$ data for $n$-octane in lieu of the TRC correlations [17].
Furthermore, in order to perform any statistical analysis a key factor is the number of data used in the objective function. The original TraPPE methodology used a relatively small data set, namely six \( \rho_l \) values for ethane and five for \( n \)-octane. That being said, a considerable amount of experimental \( \rho_l \) data exist for both ethane and \( n \)-octane. Therefore, the limitation lies not in the availability of experimental data but rather it is found in the computational cost of performing molecular simulations at the temperature corresponding to each experimental data point. With advances in computational power the cost of performing simulations at additional temperatures is less significant. However, we found that a surrogate model was a cheap and reliable way of predicting \( \rho_l \) at temperatures that were not simulated. That being said, we have only utilized experimental \( \rho_l \) data in the temperature range originally used by Martin et al. We have done this for two reasons. First, to be consistent with the work done by Martin et al. Second, because we suspect that our surrogate model may be less accurate for extrapolation than it is for interpolation. The specific \( \rho_l \) data used to calculate RMS for ethane and \( n \)-octane are found in Appendix D.

By contrast with the Type A and A\(_B\) approaches, the Type B analysis does not utilize Equation 2.17. Instead, the Type B approach assumes that the DIPPR correlation uncertainties account for the scatter and inaccuracies in the data. As this approach utilizes the DIPPR correlation uncertainty rather than the experimental data, the Type B approach is ideal for force field developers that are less familiar with analyzing experimental data. Furthermore, by utilizing a correlation uncertainty it is not necessary to predict \( \rho_l \) at a large number of temperature values corresponding to experimental data points. For these reasons, the DIPPR uncertainties are quite useful for quantifying force field parameter uncertainties. We should mention that for the Type B analysis the DIPPR uncertainties were assumed to be constant relative error for ethane. By contrast, for \( n \)-octane we felt it was more appropriate to use a constant absolute error (of 1% the average \( \rho_l \) value) from 380-530 K and from 530-545 K we used the DIPPR \( \rho_c \) uncertainty of 3% (notice the correlation uncertainties in Figure 5.11).

5.9 Conclusions

In this chapter we applied the classical (frequentist) statistical approach of UQ+PoE for the intermolecular force field parameters. We also proposed a non-statistical method that relies upon expert evaluation of data to provide a more physically meaningful assessment of parameter
uncertainty. The DIPPR 801 database has been particularly useful in this regard. Specifically, we demonstrated how a Type A, Type A\textsubscript{B}, and Type B uncertainty analysis can be performed for estimating the uncertainty in the CH\textsubscript{4}, CH\textsubscript{3}, and CH\textsubscript{2} LJ 12-6 parameters. Then, we showed how these uncertainties propagate when predicting \(\rho_l\) and the critical constants for large \(n\)-alkanes. Somewhat surprisingly, our results demonstrate that the numerical uncertainties (from GEMC simulation results and data analysis methods) are typically larger or of a comparable magnitude to the Type A and Type A\textsubscript{B} parameter uncertainties for \(T_c\), \(\rho_c\), \(P_c\), and \(Z_c\). The Type B uncertainties were much larger but provided a more conservative and feasible estimate of uncertainty. Despite these large uncertainties, our results are still conclusive as to which experimental data and prediction models are most reliable for the critical constants of large \(n\)-alkanes.
CHAPTER 6. NEW PREDICTION MODELS

6.1 Introduction

In the previous chapters we compared our simulation results to experimental data and existing prediction models. In this chapter, we take the insight gained from our simulation results to develop new prediction models for $P_c$, $T_c$, and $\rho_c$ (or, more precisely, to regress new coefficients to existing model forms). Essentially, the simulation results helped discern between conflicting and dubious experimental data sets in $P_c$, $T_c$, and $\rho_c$. Therefore, although the simulation results will not be used directly in the development of new prediction models, they will serve as a validation of our models for $P_c$, $T_c$, and $\rho_c$. In addition, we implement our simulation uncertainties to quantify the uncertainty in our prediction models for the critical constants. Also, since internal consistency between the prediction models is an essential aspect of the DIPPR 801 project, we develop a $T_b$ model that agrees with the $T_c$ and $P_c$ models. Unfortunately, it is very difficult to obtain reliable estimates for $T_b$ with molecular simulations so we do not benchmark our $T_b$ model with simulation values. In brief, we aim to develop an internally consistent set of prediction models that agrees with the accepted experimental data, simulation results, and the long-chain-length polymer trends.

6.2 Theoretical Basis

The theoretical basis for the models presented in this chapter is known as Wertheim’s Statistical Associating Fluids Theory (SAFT) [99, 100]. The SAFT framework is a well established molecular based equation-of-state [101]. This specific construct of the SAFT equation is founded in polymer theory for polyethylene, i.e. the infinite chain-length $n$-alkane.
6.2.1 Infinite Chain-Length Behavior

Polymer theory suggests that $T_c$ for the infinite chain-length ($T_c^\infty$) is a constant (non-infinite) value. Most prediction models agree with this conclusion, although the reported value for $T_c^\infty$ ranges between 900-1900 K [19, 102]. Flory-Huggins polymer theory suggests that $\frac{1}{T_c} \propto CN \frac{1}{2} + (2CN)^{-1}$ for large values of $CN$ [19, 103, 104]. In addition, Nikitin developed a Van der Waals Self-Consistent Fluid approximation to determine that $T_c \rightarrow T_c^\infty \propto CN^{-\frac{1}{2}}$ [105]. Each of the theories presented by Flory-Huggins, Wertheim, and Nikitin predict $\rho_c \rightarrow 0 \propto CN^{-\frac{1}{2}}$. Although Flory-Huggins theory does not have a simple expression for $P_c$, both the models from Wertheim and Nikitin suggest that $P_c$ approaches zero for the infinite chain-length at a rate proportional to $CN^{-\frac{3}{2}}$. Although some models have been reported where $P_c$ and/or $\rho_c$ approach a non-zero asymptote [10, 13, 14], several studies have reported $P_c$ and $\rho_c$ models that do approach zero. However, typically the rate at which they approach zero is disregarded.

For example, the $\rho_c$ model reported by Teja et al. [12] approaches zero at a rate around $CN^{-0.165}$ while Nannoolal’s model [9] has a convergence of $CN^{-0.227}$. Also, Nannoolal’s $P_c$ model approaches zero at a rate of $CN^{-2.14}$ while Lemmon’s $P_c$ model form (same as Teja’s) exponentially approaches an asymptote, which could either be set to zero or regressed from the data [12,14]. The significance in the rate at which $P_c$ and $\rho_c$ approach zero can be found in the $Z_c$ curve. Weirtheim’s polymer theory states that $Z_c$ approaches a constant value of $\frac{1}{3}$ [99, 100, 106]. By contrast, the models reported by Teja et al. and Tsonopoulos et al. predict $Z_c^\infty \rightarrow \infty$ while Nannoolal’s model predicts $Z_c^\infty \rightarrow 0$ [9,12,13]. For $Z_c$ to approach a non-zero, non-infinite value requires that $P_c$ and $\rho_c$ approach zero at the correct rates. Only a handful of models have reported non-zero, non-infinite values for $Z_c^\infty$. The original set of Nikitin models [20], the most recent set of Nikitin models [15], and Vega’s models [83] predict $Z_c^\infty$ values of 0.1, 0.15, and 0.2, respectively. The reason these values are considerably less than $\frac{1}{3}$ is because the available $Z_c$ data are less than 0.275 and trending downward. Our goal is to develop a set of prediction models that not only matches the available experimental data but also results in $Z_c^\infty \rightarrow \frac{1}{3}$.
6.2.2 Internal Consistency

We have included two constraints in our optimization to ensure internal consistency between the physical properties. First, as mentioned previously, we force \( Z_c^\infty \rightarrow \frac{1}{3} \). This leads to an internally consistent set of \( T_c \), \( P_c \), and \( \rho_c \) models. The second constraint is that the carbon number at which \( P_c = 1 \) atm is the same as the carbon number at which \( T_c = T_b \). Although this constraint should be obvious to ensure \( T_c \), \( P_c \), and \( T_b \) are internally consistent, it appears to have never been implemented in the literature. In addition, we verify that the curves for \( P_c \), \( T_c \), and \( T_b \) do not have an inflection point. However, since \( \rho_c \) and \( Z_c \) both have a minimum and a maximum they will inevitably have an inflection point, so we used an alternative approach to verify that the \( \rho_c \) trend is well behaved. Specifically, since \( dV_c \) is intimately related to the volume contribution of an additional methylene group, we required that \( d^3V_c \) not have an inflection point, i.e. \( d^3V_c < 0 \).

6.3 Methodology

We implemented a systematic multi-property optimization scheme to develop the new prediction models (i.e. new coefficient values for a pre-existing model form). The general outline for this method is

1. Develop \( P_c \) model
2. Develop \( T_c \) model
3. Develop \( \rho_c \) model subject to constraint that \( Z_c \rightarrow \frac{1}{3} \)
4. Develop \( T_b \) model with constraint that \( T_c = T_b \) at same \( CN \) that \( P_c = 1 \) atm

For each step in this algorithm a similar approach is implemented to “develop” the corresponding model:

1. Evaluate available experimental data
2. Propose different prediction model forms
3. Assess statistical significance of regression parameters
4. Confirm that the models have no inflection points or other abnormalities
5. Compare models with simulation results

6. Verify correct long-chain-length trends

In Step 6 we plot a specific critical constant with the corresponding linearized form. Specifically, we plot $P_c$ with respect to $CN^{-\frac{3}{2}}$, $T_c$ and $\rho_c$ with respect to $CN^{-\frac{1}{2}}$, and $\frac{1}{T_c}$ with respect to $CN^{-\frac{1}{2}} + (2CN)^{-1}$.

6.4 Model Development

In this section we detail the decisions made in developing the proposed prediction models. First, we present the different experimental data sets and which data points are accepted. Second, we compare different model forms and the number of parameters included in the models. We then explain how we imposed the constraints. This is followed by an assessment of the regression parameters. Finally, we present the proposed prediction models.

6.4.1 Experimental Data Evaluation

The primary advantage of our molecular simulation results is that they are reliable enough (accurate with small uncertainties) to determine which experimental data to accept in developing our prediction models. For example, as presented in Chapters 4 and 5, our simulation results suggest that Nikitin’s 2014 experimental $P_c$ data should be accepted [15]. Also, the $P_c$ data for C$_5$-C$_{15}$ from Teja et al. [12] are used while those for C$_{16}$-C$_{18}$ are not used since these data show a strong discontinuity with those from Nikitin et al. We also excluded C$_{19}$ from Nikitin’s data set to improve the internal consistency of the accepted points.

The recommendations are slightly different for $T_c$. Although Nikitin’s 2014 experimental $T_c$ data are in better agreement with our simulation results than his original data, the 2014 data still appear to have a bias compared to the simulation results and Teja’s data. For this reason, we did not include any of Nikitin’s $T_c$ data in our optimization. Instead, we only included Teja’s data for C$_5$-C$_{18}$. This decision can be supported by a Flory-Huggins plot, i.e. by plotting Nikitin’s 2014 $T_c$ values as $\frac{1}{T_c}$ with respect to $CN^{-\frac{1}{2}} + (2CN)^{-1}$. As will be shown in Section 6.5.1, the 2014 Nikitin data follow a slightly different trend than Teja’s data and our simulation results when viewed on a Flory-Huggins plot.
Experimental data for $\rho_c$ are only available for carbon numbers as large as C$_{18}$. In order to extend the range over which estimates for $\rho_c$ are available, we also extrapolated $\rho_l$ data for the largest $n$-alkanes found in the DIPPR 801 database (C$_{10}$-C$_{36}$). Specifically, the critical density was predicted with the Rackett equation, $\rho_l$ data, and the prediction models for $P_c$ and $T_c$. This is one reason why $P_c$ and $T_c$ were obtained prior to $\rho_c$ in our sequential multi-property optimization. In addition, we did not use an exponent value of $\frac{2}{7}$ (0.2857) in the Rackett equation, rather we used an $n$-alkane family specific value of 0.2679 that was obtained from compounds C$_{5}$-C$_{18}$. However, since the Rackett equation is not a thermodynamically rigorous expression, it was important to validate that these results were reliable. As will be shown in Section 6.5.1 our simulation results agree well with these extrapolated estimates for $\rho_c$, so we accepted these Rackett-extrapolated $\rho_c$ values from C$_{16}$-C$_{36}$ as well as Teja’s experimental data from C$_{5}$-C$_{18}$ [12].

The situation is different for $T_b$ as our simulation results do not provide any assistance. In this case, we accepted the same values as those used by DIPPR 801 staff to develop the current DIPPR 801 model. The main data set consists of recommended values by Ambrose et al. from C$_{5}$-C$_{18}$ [107]. However, a $T_b$ value measured at Wiltec Research Co. of 725.1 K for C$_{30}$ is also used in this regression [108]. Furthermore, the constraint that $T_c = T_b$ at the same CN that $P_c = 1$ atm effectively adds another data point at very large CN.

### 6.4.2 Model Form

Several different models for the dependence of the critical constants and normal boiling point on chain-length have been proposed in the literature. Some of these models have a strong theoretical basis while others are mainly empirical. The reliability of extrapolating to larger compounds is intimately connected with the level of theory behind the model. One such set of theoretically based models are those from Nikitin et al. [15, 20, 29, 105] The general mathematical form for the Nikitin models is

$$\hat{y}(CN, b, \delta) = \sum_{i=0}^{d} b_i CN^{\left(\frac{-\delta + 1}{2}\right)}$$

(6.1)

where $\hat{y}$ is the prediction model for either $P_c$, $T_c$, $\rho_c$, or $T_b$, $d$ is the order of the model (number of terms included in summation), $b_i$ is the $i$th regression parameter, and $\delta$ is a constant specific to the property being predicted. Specifically, the value of $\delta$ is 3 for $P_c$, 1 for $\rho_c$, and 0 for $T_c$ and $T_b$. With
increasing chain-length the contribution from the higher order terms becomes negligible and

\[
\lim_{CN \to \infty} \hat{T}_c \to T^\infty_c \propto CN^{\frac{-1}{2}} \tag{6.2}
\]

\[
\lim_{CN \to \infty} \hat{P}_c \to 0 \propto CN^{-\frac{3}{2}} \tag{6.3}
\]

\[
\lim_{CN \to \infty} \hat{\rho}_c \to 0 \propto CN^{\frac{-1}{2}} \tag{6.4}
\]

where \(T^\infty_c\) is the \(b_0\) parameter for \(\hat{T}_c\). Therefore, since these models are consistent with the polymer theories of Wertheim and Flory-Huggins they have the correct long-chain-length functionality. We also wish to match the experimental data available for smaller compounds to a high degree of accuracy. Fortunately, the Nikitin models also result in very low residuals when compared with the experimental data. Therefore, rather than comparing several different model forms (many of which have the incorrect limiting trends) we will focus our attention on the value of \(d\) in Nikitin’s models. In other words, we varied the order of the models and evaluated the impact this had on reproducing the experimental data, the smoothness of the curve, and the shape of the long-chain-length trend.

### 6.4.3 Enforcing Constraints

Since the mathematical form of Nikitin’s models has the correct dependence on \(CN\), these models predict a non-infinite, non-zero value for the infinite chain-length compressibility factor \((Z^\infty_c)\). Furthermore, L’Hôpital’s rule results in a succinct expression relating \(Z^\infty_c\) to the regression parameters in \(P_c\), \(T_c\), and \(\rho_c\)

\[
Z^\infty_c = \frac{P^0_c M_{CH_2}}{R_g T^\infty_c \rho^0_c} \tag{6.5}
\]

where \(M_{CH_2}\) is the mass of a methylene group \((14.0265 \text{ gm/mol})\), and \(P^0_c\) and \(\rho^0_c\) are the \(b_0\) regression parameters for \(P_c\) and \(\rho_c\), respectively. For a given \(P^0_c\) and \(T^\infty_c\), \(\rho^0_c\) can be calculated to satisfy the constraint that \(Z^\infty_c = \frac{1}{3}\). Therefore, we impose this constraint on \(\rho^0_c\) after regressing \(P_c\) and \(T_c\) independently. Essentially, this constraint reduces the number of regression parameters by fixing the value of \(\rho^0_c\). As the \(Z_c\) constraint determines \(\rho^0_c\) it also provides a slope for the \(\rho_c\) curve near the origin when \(\rho_c\) is plotted with respect to \(CN^{\frac{-1}{2}}\). This constraint is helpful because \(\rho_c\) is not in the linear range, even with the extrapolated \(\rho_c\) values for compounds as large as \(C_{36}\). Therefore,
the short chain-length trend is determined by experimental data, the trend for slightly larger chain-lengths is influenced by the values extrapolated from \( \rho \), and the infinite chain-length trend is a result of the constraint for \( Z^\infty_c \).

The constraint that \( T_c = T_b \) at the same \( CN \) that \( P_c = 1 \) atm is imposed by reducing the number of regression parameters for \( T_b \) by 1. Specifically, the first term \((b_0)\) in Equation 6.1 for \( \hat{T}_b \) is not used for optimizing the \( T_b \) model. Instead, \( b_0 \) for \( \hat{T}_b \) is obtained by satisfying this constraint while the remaining parameters are regressed from \( T_b \) data. Similar to \( T^\infty_c \), \( b_0 \) for \( \hat{T}_b \) is the normal boiling point of the infinite chain-length \( (T^\infty_b) \). It is important to recognize that when \( T_c < T_b \) a compound simply ceases to boil at 1 atmosphere pressure. Therefore, the \( T_b \) model is obsolete (i.e. non-physical) for compounds longer than the carbon number where \( T_c = T_b \). This is the reason why we chose to use a sequential multi-property optimization for \( P_c, T_c, \) and \( T_b \). Since \( P_c \) and \( T_c \) are regressed first (and independently), only \( T^\infty_b \) accounts for this consistency check. As \( T^\infty_b \) has no physical meaning (compared to \( T^\infty_c \)) it is used simply to enforce this constraint.

6.4.4 Insight from Regression Parameters

When Nikitin et al. optimized their models it appears that they did not consider the effect that the model order has on the regression parameters. Specifically, the parameters \( P^0_c, T^\infty_c, \) and \( \rho^0_c \) determine \( Z^\infty_c \), which is very influential on the \( Z_c \) trend. For example, the higher the order of \( \hat{P}_c \) the larger the value for \( P^0_c \). In fact, the \( P^0_c \) value obtained with a third order form of \( \hat{P}_c \) can be an order of magnitude less than that for the sixth order form (the exact ratio depends on the data included).

Range of Feasible Parameter

As mentioned previously, in order to verify that \( \rho_c \) is well behaved we investigated \( \frac{d^3 VC}{dCN^3} \). For a specific order of \( \hat{P}_c \) and a given set of data, there is a limited range of \( \rho^0_c \) values that will result in \( \frac{d^3 VC}{dCN^3} < 0 \) for any value of \( CN \). We investigated this range of acceptable \( \rho^0_c \) values for the 4th and 5th order models with the aforementioned \( \rho_c \) data set.

Figure 6.1 presents regions of acceptable parameters for \( P^0_c, T^\infty_c, \) and \( \rho^0_c \). The points represent combinations of \( P^0_c \) and \( T^\infty_c \) that were considered when developing the prediction models. These various \( P^0_c \) and \( T^\infty_c \) values were obtained using different data sets and model orders.
The starred parameter set corresponds to the prediction models presented in Section 6.5.1. We have also included the Nikitin values for $P_c^0$ and $T_c^\infty$ as a reference. The contours represent the ranges of $\rho_c^0$ values that result in a well behaved model (i.e. $\frac{d^3 V_C}{dCN^3} < 0$). The different contours represent the fourth and fifth order $\rho_c$ model. To be specific, only $P_c^0$ and $T_c^\infty$ sets that are within the $2.15 \leq \rho_c^0 \leq 2.85$ contours are acceptable for a fourth order $\hat{\rho}_c$ while those within the $2.85 \leq \rho_c^0 \leq 3.25$ contours are acceptable for a fifth order $\hat{\rho}_c$. Note that Nikitin et al. used a fourth order model for $\rho_c$ [29]. Panels (a) and (b) were obtained using $Z_c^\infty$ values of $\frac{1}{3}$ and $\frac{1}{5}$, respectively. We compared the contours for these $Z_c^\infty$ values since they were both proposed in the literature using the SAFT model [83, 99]. This comparison is useful for reducing the immense number of model combinations that are possible. For example, if attempting to design a prediction model where $Z_c \rightarrow \frac{1}{3}$ it is not necessary to consider models that result in $P_c^0 < 400$ or $P_c^0 > 1000$.

Furthermore, since the $P_c^0$ and $T_c^\infty$ values Nikitin reported [15, 20] are well outside the fourth and fifth order $\rho_c^0$ contours in both Panels (a) and (b), they are incompatible with $Z_c \rightarrow \frac{1}{3}$ or $\frac{1}{5}$.

![Figure 6.1: Contours of acceptable combinations of $P_c^0$, $T_c^\infty$, and $\rho_c^0$. Panels (a) and (b) were obtained using $Z_c^\infty$ values of $\frac{1}{3}$ and $\frac{1}{5}$, respectively. We have provided the acceptable $\rho_c^0$ ranges using both a 4th and 5th order model. The points represent combinations of $P_c^0$ and $T_c^\infty$ that were obtained using different data sets and model orders. The starred parameter set corresponds to the prediction models presented in Section 6.5.1. We have also included the Nikitin values for $P_c^0$ and $T_c^\infty$ as a reference.](image-url)
6.4.5 Proposed Prediction Models

In this section we present the equations selected and the regression coefficients obtained for our prediction models. In brief, we decided upon 6th, 4th, 5th, and 5th order models for $P_c$, $T_c$, $\rho_c$, and $T_b$, respectively. The resulting model parameters are included in Table 6.1 and, to facilitate use of our models, we have provided the corresponding equations below the table. As a reference, we have also included the parameters reported in the various studies by Nikitin et al. Notice that since Nikitin et al. used only a third order $P_c^0$ their $P_c^0$ term is much lower than that from this work. This explains why their $Z_c^\infty$ value was so much lower than the predicted value of $\frac{1}{3}$. Furthermore, this demonstrates how knowledge of the long-chain-length theory can provide insight into selecting the model order. Also, notice the difference in $\rho_c^0$ values. This is a result of using $Z_c^\infty = \frac{1}{3}$ to determine $\rho_c^0$ rather than just the experimental data between $C_5$-$C_{18}$.

Although Nikitin’s group has already reported parameter sets for their models, we believe that the parameters we have obtained with their models improve the prediction of the thermodynamic properties in question for multiple reasons. First, we have meticulously scrutinized the data used in regression where Nikitin et al. accepted all of their own data and those from Teja et al. when developing their models. This evaluation was described in Section 6.4.1. Second, we have considered the effect of including additional parameters in the regression. In other words, we have chosen the prediction model order based upon how well the model matches the experimental data, agrees with the simulation results, and extrapolates to long-chain-lengths. It is significant that for each property we employed a model with at least one additional parameter compared to those from Nikitin’s various studies.

Statistically Significant Parameters

There are two concerns when including a large number of parameters in a regression model. The first is that the model is well behaved, particularly upon extrapolation. We confirmed that our models are well behaved by verifying that there are no inflection points. The second concern is that each regression parameter is statistically significant. In other words, a model is considered over-parameterized if a similar performance can be achieved with fewer parameters. We performed this analysis by comparing $S(\theta)$ of the different model orders using Equation 2.19. Specifically, if $S(\theta)$
Table 6.1: Model parameters for proposed prediction models.
The values from Nikitin’s various models are provided as a reference. [15, 20, 29, 105]

<table>
<thead>
<tr>
<th>Property</th>
<th>Order</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
<th>$b_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_c$, New Model</td>
<td>6</td>
<td>802.44</td>
<td>-10698</td>
<td>67181</td>
<td>-214686</td>
<td>337553</td>
<td>-207762</td>
</tr>
<tr>
<td>$P_c$, Nikitin (2014)</td>
<td>3</td>
<td>215.84</td>
<td>-646.58</td>
<td>555.06</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>$P_c$, Nikitin (1997)</td>
<td>3</td>
<td>138.78</td>
<td>-78.53</td>
<td>-476.45</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>$T_c$, New Model</td>
<td>4</td>
<td>1393</td>
<td>-3788</td>
<td>5121</td>
<td>-2836</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>$T_c$, Nikitin (2014)</td>
<td>3</td>
<td>1314</td>
<td>-2996</td>
<td>2496</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>$T_c$, Nikitin (1997)</td>
<td>3</td>
<td>1259</td>
<td>-2654</td>
<td>1992</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>$\rho_c$, New Model</td>
<td>5</td>
<td>2.915</td>
<td>-19.04</td>
<td>70.28</td>
<td>-130.6</td>
<td>93.56</td>
<td>N/A</td>
</tr>
<tr>
<td>$\rho_c$, Nikitin (1998)</td>
<td>3</td>
<td>1.809</td>
<td>-4.888</td>
<td>4.606</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>$T_b$, New Model</td>
<td>5</td>
<td>1532</td>
<td>-6249</td>
<td>11484</td>
<td>-8254</td>
<td>312.5</td>
<td>N/A</td>
</tr>
<tr>
<td>$T_b$, Nikitin (2000)</td>
<td>4</td>
<td>1338.3</td>
<td>-4510.5</td>
<td>6411.5</td>
<td>-4011.6</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

\[
\hat{P}_c = b_0 CN^{-\frac{3}{2}} + b_1 CN^{-2} + b_2 CN^{-\frac{5}{2}} + b_3 CN^{-3} + b_4 CN^{-\frac{7}{2}} + b_5 CN^{-4}
\]
\[
\hat{T}_c = b_0 + b_1 CN^{-\frac{1}{2}} + b_2 CN^{-1} + b_3 CN^{-\frac{3}{2}}
\]
\[
\hat{\rho}_c = b_0 CN^{-\frac{1}{2}} + b_1 CN^{-1} + b_2 CN^{-\frac{3}{2}} + b_3 CN^{-2} + b_4 CN^{-\frac{5}{2}}
\]
\[
\hat{T}_b = b_0 + b_1 CN^{-\frac{1}{2}} + b_2 CN^{-1} + b_3 CN^{-\frac{3}{2}} + b_4 CN^{-2}
\]

for a lower order model satisfies Equation 2.19 at the $\alpha = 0.95$ level, the final term in the higher order model is not statistically justified. However, the statistical significance is not as important as the extrapolation behavior. For this reason, we considered some model orders that had more parameters than statistically necessary, but we concluded that it was not necessary or desirable to exceed the maximum number of statistically significant parameters.

6.4.6 Uncertainty in Prediction Models

The final task in developing a prediction model is assigning an appropriate uncertainty. Although most models in the literature neglect this important step, some studies have performed elaborate statistical analysis methods to quantify the uncertainty [10, 11]. However, when we employed the Type A analysis discussed in Chapter 2 we obtained uncertainties that were much smaller than we deemed reasonable considering the wide range of conflicting experimental data and existing prediction models. For this reason, we decided to implement an alternative approach.
This section focuses on the methodology used for quantifying the uncertainties while the uncertainties themselves are found in Section 6.6.

Rather than relying upon the random scatter in the data used in the regression, we utilized the Type B simulation uncertainties. Specifically, we varied $b_0$ for each property and obtained a model fit by regressing the remaining $b_i$ parameters to the accepted data. Then, we considered a model fit to be acceptable if it matched our simulation results to within the Type B uncertainties. The uncertainty in the prediction models is determined from the minimum and maximum values of the acceptable model fits. This provides a Type B assessment of uncertainty which is consistent with the uncertainties reported by DIPPR 801. We did not require that a model be well behaved (i.e. have no inflection points) for it to be considered acceptable. For the most part this was not an issue, although it might explain the odd shape in some of the uncertainty bans.

Unfortunately, in the case of $\rho_c$, our simulation results are not consistent over the entire range of $CN$. Therefore, we accepted any model that matched either $C_{36}$ or $C_{48}$, along with the smaller chain-lengths, at the 95% confidence level. When assigning our uncertainties in $\rho_c$ we did not impose any constraints that $Z_c \rightarrow \frac{1}{3}$. For the $Z_c$ uncertainty we assumed that the $P_c$, $T_c$, and $\rho_c$ uncertainties follow a normal distribution and are uncorrelated so that we can employ a standard propagation of errors approach. The assumption of a normal distribution is not valid for $P_c$ with very long-chain-lengths (around $C_{300}$) because the distribution becomes skewed by the lower bound that $P_c$ remain positive. We have not accounted for the skewness since these prediction models are mainly intended for compounds smaller than $C_{300}$. The assumption of uncorrelated errors in $P_c$, $T_c$, and $\rho_c$ is mathematically valid since the constraint that $\rho_c^0$ satisfy $Z_c \rightarrow \frac{1}{3}$ is not applied when assigning uncertainty to $\rho_c$. However, it is likely that the experimental data for $P_c$, $T_c$, and $\rho_c$ are indeed correlated to some extent. Since we did not account for this possible correlation, this may explain the very large uncertainties in $Z_c$ observed in Section 6.5.1.

Finally, we implemented a slightly different approach for assigning the uncertainties in our $T_b$ prediction model since we did not have any simulation uncertainties for $T_b$. In this case, we accepted any model fit that resulted in an intersection with the minimum and maximum uncertainties in $T_c$ at the respective maximum and minimum $CN$ values where $P_c = 1$ atm ($C_{140}$-$C_{300}$). The $T_b$ uncertainty is determined from the minimum and maximum values for $T_b$ of the acceptable models based upon this criterion.
6.5 Results

In Subsection 6.5.1 we depict the available experimental data (where we distinguish between “accepted” and “not used” according to the description in Section 6.4.1), the simulation results (see Chapters 4-5), and the new prediction models (see Subsection 6.4.5) for the critical constants of the \(n\)-alkanes along with their corresponding uncertainties (see Subsection 6.4.6). Only the recommended values are displayed for the simulation results (see Appendix H). Specifically, the recommended value is a weighted average of the results obtained with the various force fields and system sizes. The simulation uncertainties represent the largest uncertainty reported in Chapter 5 which accounts for both numerical and parameter uncertainties (typically the Type B uncertainty is the largest). We have not included error bars for compounds that were simulated in Chapter 4 but not Chapter 5 since only numerical uncertainties were reported in Chapter 4.

In this section we display \(P_c\), \(T_c\), and \(\rho_c\) with respect to \(CN\) as well as \(CN^{\frac{\delta}{2}}\), where \(\delta\) is the property specific exponent. We also include the \(Z_c\) result as an internal consistency check for the critical constants. Furthermore, in subsection 6.5.2 we demonstrate how well the Rackett equation performs for predicting \(\rho_l\) with the new critical constant models. Then, we present the new \(T_b\) model (and uncertainty) and compare it with the current DIPPR 801 model. We also demonstrate that \(T_c = T_b\) at the same \(CN\) that \(P_c = 1\) atm.

6.5.1 \(P_c\), \(T_c\), \(\rho_c\), and \(Z_c\)

We begin by presenting \(P_c\) in Figure 6.2. In Panels (a) and (b) we plot \(P_c\) with respect to \(CN\) and \(CN^{\frac{3}{2}}\), respectively. Clearly, the new prediction model follows the experimental data to which it was fit. In addition, it also agrees very well with our simulation results, which were not used in the model development. Furthermore, in Panel (a) we see that the prediction uncertainties not only match our simulation uncertainties but also the uncertainty Nikitin et al. assigned for \(C_{60}\). The discontinuity in slope between the data from Teja et al. and Nikitin et al. is apparent in Panel (b). In Panel (b) we can also see that even Nikitin’s data, which include compounds as large as \(C_{60}\), are still not in the linear regime. That is, a straight-line through these data would not intersect the origin. Also, it is important to mention that \(P_c = 1\) atm around \(C_{219}\). However,
since the uncertainty in $P_c$ increases substantially with increasing chain-length, this $CN$ may range anywhere between $C_{140}$-$C_{300}$.

Figure 6.2: Performance of new $P_c$ prediction model. Included are “accepted” and “not used” experimental data sets and the recommended values from our simulation results. In Panels (a) and (b) we plot $P_c$ with respect to $CN$ and $CN^{-3}$, respectively.

In Figure 6.3 Panels (a) and (b) we present $T_c$ with respect to $CN$ and $CN^{-1}$, respectively, while in Panel (c) we plot $\frac{1}{T_c}$ with respect to $CN^{-1} + \frac{2}{CN}$. Again, despite not being used in the model development, our simulation results agree very well with the new prediction model, which also matches the experimental data to which it was regressed. In Panel (a) we see that the prediction uncertainty matches our simulation uncertainties and overlaps with the uncertainties reported by Nikitin et al. for $C_{60}$. Furthermore, in Panel (b) we can see that the uncertainty in $T_c^\infty$ is approximately 75 K. Also, in Panel (c) it is clear that Nikitin’s data demonstrate a small bias compared to Teja’s data. In addition, in Panel (c) it appears that Nikitin’s 2014 data are not internally consistent over the entire $CN$ range, i.e. there are slight discontinuities and changes in slope, particularly between $C_{22}$ and $C_{23}$ (around 0.23 on the horizontal axis).

Figure 6.4 Panels (a) and (b) depict $\rho_c$ with respect to $CN$ and $CN^{-1}$, respectively. Once again, our prediction model performs very well at matching both the experimental data and the simulation results, which were not used for regressing the model parameters. Recall that we have
Figure 6.3: Performance of new $T_c$ prediction model. Included are “accepted” and “not used” experimental data sets and the recommended values from our simulation results. In Panels (a) and (b) we present $T_c$ with respect to $CN$ and $CN^{-\frac{1}{2}}$, respectively, while in Panel (c) we plot $\frac{1}{T_c}$ with respect to $CN^{-\frac{1}{2}} + \frac{2}{CN}$.

included $\rho_c$ estimates obtained from $\rho_l$ in our optimization procedure. The purpose of this was to provide $\rho_c$ values for chain-lengths larger than $C_{18}$ (the largest measured experimental value) and, thereby, improve the predictability of the model. Panel (b) justifies the use of these extrapolated values as they are internally consistent and linear. However, notice that neither the experimental $\rho_c$ values or those extrapolated from $\rho_l$ are in the linear regime (i.e. a straight-line through the data in Panel (b) does not intersect with the origin). This demonstrates why it was useful to determine
\( \rho_c^0 \) from the constraint that \( Z_c^\infty \rightarrow \frac{1}{3} \). It also explains why our model is much different than those obtained from only experimental \( \rho_c \) data. The peculiar shape of the prediction model uncertainty displayed in Panel (a) is caused by the internal inconsistencies between the \( C_{36} \) and \( C_{48} \) simulation uncertainties.

Figure 6.4: Performance of new \( \rho_c \) prediction model. Included are experimental data, extrapolated \( \rho_c \) from \( \rho_l \) and the recommended values from our simulation results. In Panels (a) and (b) we plot \( \rho_c \) with respect to \( CN \) and \( CN^{-\frac{1}{2}} \), respectively.

Finally, Figure 6.5 Panels (a) and (b) contain the results for \( Z_c \) with respect to a linear and log scale of \( CN \), respectively. Since experimental data for \( \rho_c \) have only been measured by Teja et al. we included the current DIPPR 801 values for comparison. Notice that our model appears to reach a minimum around \( C_{45} \) that is slightly higher than our simulation results. This is curious considering that our \( P_c \), \( T_c \), and \( \rho_c \) models all agree well with the simulation results. We attribute this to the propagation effect that small deviations in \( P_c \), \( T_c \), and \( \rho_c \) can lead to large deviations in \( Z_c \). Thus, we did not optimize our prediction models for \( P_c \), \( T_c \), and \( \rho_c \) to match this \( Z_c \) simulation point which comes with relatively large uncertainty. The proposed critical constant prediction models provide a compromise between the infinite chain-length trends reported by Vega et al. (empirical, soft-SAFT) and Wertheim (theoretical, polymer SAFT). Specifically, our \( Z_c \) trend approaches \( \frac{1}{5} \) before subsequently increasing to \( \frac{1}{3} \). In other words, we are able to match what would be expected
based on the available experimental data and simulation results while still reaching the polymer theory limit. Furthermore, since the upper and lower uncertainties in $Z_c$ were obtained without the constraint that $Z_c \to \frac{1}{3}$ these uncertainties account for possible deviations from Wertheim’s theory. For this reason, it is not concerning that the $Z_c$ uncertainties are larger than the simulation uncertainties.

Figure 6.5: Performance of new $Z_c$ prediction model. Included are experimental data, current DIPPR 801 values, and the recommended values from our simulation results. In Panels (a) and (b) we plot $Z_c$ with respect to $CN$ on a linear and log scale, respectively.

### 6.5.2 $\rho_l$

One advantage in developing reliable predictions of the critical constants is the ability to predict several other thermophysical properties. For example, the Rackett equation can be used to predict $\rho_l$ from the critical constants. This is especially important for larger compounds where experimental data are less prevalent, particularly near the critical point. In Figure 6.6 Panel (a) we plot the purely predictive $\rho_l$ curve obtained from our $P_c$, $T_c$, and $\rho_c$ models with the Rackett equation (using an $n$-alkane family specific $\gamma$ value of 0.2679). Also included are the accepted experimental $\rho_l$ data for the largest $n$-alkanes included in the DIPPR 801 database and the simulation results. For clarity, we have only plotted the predicted $\rho_l$ curves in the temperature range where
experimental data or simulation results are available. Although the agreement is not perfect, our predicted $\rho_l$ curves are considerably better than the predicted $\rho_l$ trends using the current DIPPR 801 values for the critical constants. In Figure 6.6 Panel (b) we provide a comparison between the predicted $\rho_l$ curves with our critical constant models and the current DIPPR 801 critical constant values. To facilitate a comparison, we have only included the even numbered compounds from C$_{16}$-C$_{36}$ that have $\rho_l$ data or simulation results above 0.6$T_c$ (C$_{16}$, C$_{18}$, C$_{20}$, C$_{24}$, C$_{28}$, and C$_{36}$). (C$_{48}$ is not included in Panel (b) as it is not currently included in the DIPPR 801 database.) The difference is negligible for compounds with $CN < 16$ since more reliable critical constant data are available for smaller $n$-alkanes. Notice that Panel (a) is on a reduced temperature scale (where $T_c$ is that of our prediction model) while Panel (b) is on an absolute temperature scale.

Figure 6.6: Comparison of experimental $\rho_l$ data with the predicted $\rho_l$ obtained with the Rackett equation. The Rackett exponent ($\gamma$) has been set to a family specific value of 0.2679. Panel (a) plots the predicted $\rho_l$ values using the $P_c$, $T_c$, and $\rho_c$ models developed in this work. Panel (b) compares the predicted $\rho_l$ trends using the new critical constant models and the current DIPPR 801 values. For clarity, the predictions are only plotted in the temperature range where experimental data or simulation results are available. In Panel (a) we have included C$_{10}$-C$_{30}$, C$_{32}$, C$_{36}$, and C$_{48}$. In Panel (b) we have only included C$_{16}$, C$_{18}$, C$_{20}$, C$_{24}$, C$_{28}$, and C$_{36}$ as $\rho_l$ data and/or simulation results exist for these compounds above 0.6$T_c$. In Panels (a) and (b) the top curve represents C$_{10}$ and C$_{16}$, respectively.
6.5.3 $T_b$

In Figure 6.7 we have included the experimental $T_b$ data, the current DIPPR 801 $T_b$ model, the new $T_b$ model, the new $T_c$ model, and the $CN$ at which $P_c = 1$ atm. There are three main conclusions from this figure. First, notice that $T_c = T_b$ at $C_{219}$ which is also where $P_c = 1$ atm, i.e. $T_c$, $P_c$, and $T_b$ are internally consistent, fulfilling one of our objectives. Second, the new $T_b$ model is very similar to the current $T_b$ model. In fact, they do not deviate by more than 1% until around $C_{70}$. Furthermore, the assigned uncertainty for the new prediction model encompasses the current DIPPR 801 model until $C_{125}$. For example, the deviation at $C_{100}$ (the largest compound for which the current $T_b$ model was intended) is only 25 K, or 2%, which is also well within the uncertainties reported for the current DIPPR 801 model. This demonstrates that only a small modification to the current DIPPR $T_b$ model is required so as to be consistent with the new $P_c$ and $T_c$ models. Third, the new $T_b$ model matches the experimental $T_b$ data to the same level of accuracy as the current DIPPR $T_b$ model. In addition, the current $T_b$ model under predicts the $C_{30}$ value by 1 K whereas the new prediction model under predicts the $C_{30}$ value only by 0.5 K.

![Figure 6.7: Performance of new $T_b$ prediction model. Included are experimental data from Ambrose et al. and VonNiederhausern et al., the $T_c$ model presented in Table 6.1, the $CN$ where $P_c = 1$ atm using the $P_c$ model presented in Table 6.1, and the current DIPPR 801 $T_b$ model.](image-url)
6.6 Model Performance

One clear way to quantify the model performance is to assess the deviation between the model and the data to which it was regressed. There are several different approaches for such an analysis. In Figure 6.8 we present a plot of the residuals with respect to \( CN \). The percent deviation is defined as \( \frac{\hat{y}(CN_i) - y_i}{y_i} \times 100 \% \). Due to the order of magnitude difference in the various percent deviations and the ranges in carbon number, we have plotted \( P_c \) and \( \rho_c \) in Panel (a) while \( T_c \) and \( T_b \) are found in Panel (b). For the most part, our residuals have random scatter, as desired.

![Figure 6.8: Residual plots for the different prediction models. The percent deviation is defined as \( \frac{\hat{y}(CN_i) - y_i}{y_i} \times 100 \% \). Panel (a) contains \( P_c \) and \( \rho_c \) while \( T_c \) and \( T_b \) are found in Panel (b).](image)

In Figure 6.9 we also provide the average absolute deviation (AAD) and the maximum absolute deviation (MAD). The AAD and MAD are both low for every property considered. The \( P_c \) deviations are the largest, followed by those for \( \rho_c \), \( T_b \), and \( T_c \). Notice that the largest AAD and MAD values correspond to the properties that were regressed to more data and compounds with longer chain-lengths. Recall that \( P_c \), \( \rho_c \), \( T_b \), and \( T_c \) used data for compounds as large as \( C_{60} \), \( C_{36} \), \( C_{30} \), and \( C_{18} \), respectively. Therefore, the large deviations may simply be attributed to
increasing uncertainties (and scatter) in the experimental data, which is typically the case for larger compounds.

Figure 6.9: Comparison of the average absolute deviation (AAD) and maximum absolute deviation (MAD) for each property. The percent deviation is defined as \( \frac{\hat{y}(CN_i) - y_i}{y_i} \times 100\% \).

It is noteworthy that our models match the new Nikitin data for \( T_c \) and the Teja data for \( P_c \) to within the author’s reported uncertainties. For this reason, these data were listed as “not used” rather than “rejected”, in accordance with DIPPR 801 policy. However, accepting these data would greatly affect our model, even if the uncertainties were used to weigh the data. This is because both data sets appear to have a bias which is difficult to account for in regression. Therefore, rather than include these data in our regression we simply verified that the prediction models matched the “not used” data to within the author reported uncertainties.

Finally, in Figure 6.10 we display the relative uncertainties for each property for C\textsubscript{20}-C\textsubscript{100}. The relative uncertainty is defined as \( \frac{\hat{y}(CN_i)_{\text{max}} - \hat{y}(CN_i)_{\text{min}}}{2\hat{y}(CN_i)} \times 100\% \). In this figure we see that the largest relative uncertainties are for \( P_c \) and, consequently, \( Z_c \). The large relative uncertainty in \( P_c \) is likely attributed to \( P_c \) converging to 0. Surprisingly, the \( T_b \) uncertainty is fairly small considering that we varied the value of \( CN \) for which \( T_b = T_c \) between 140-300.
Figure 6.10: Estimated uncertainties for the different prediction models. The relative uncertainty is defined as

\[
\frac{\hat{y}(CN_{i})^{\text{max}} - \hat{y}(CN_{i})^{\text{min}}}{2\hat{y}(CN_{i})} \times 100 \%
\]

6.7 Limitations

The constraint that \( T_c = T_b \) at the same \( CN \) that \( P_c = 1 \text{ atm} \), while thermodynamically rigorous, is limited in its use because \( P_c \) is very flat with respect to \( CN \) in the region near 1 atm. Therefore, a slight difference in \( P_c \) data or the order used can result in a significantly different \( CN \) that \( P_c = 1 \text{ atm} \). Specifically, we observed that \( CN \) could range anywhere from 140-300 with just a small change in methodology (changing the model order and/or data accepted). Therefore, constraining \( T_b \) to intersect with \( T_c \) at the same \( CN \) could result in a poor prediction of \( T_b \). However, based upon our observations we believe that the intersection at \( C_{219} \) is well behaved. To be specific, \( T_b \) matches the experimental data well, it follows the current DIPPR correlation very closely up until \( C_{60} \), and it does not have an inflection point. We deliberately obtained this good behavior by considering numerous \( P_c, T_c, \) and \( T_b \) combinations and disregarding those that resulted in \( T_b \) curves that deviated strongly from the data or the current DIPPR correlation and/or had an abnormal shape.

It is also important to note that various theories have been proposed in an attempt to predict the infinite chain-length trends for the critical constants and normal boiling point, and some of
these theories disagree with the theoretical construct for our models. The remainder of this section discusses the theories not in agreement with our models.

First, in one of the earlier studies from Nikitin’s group, they derived that \( Z_c \to 0 \propto \frac{1}{3}CN^{-1} \) for the Van der Waals Chain-molecule model [105]. However, this proposed behavior was not internally consistent with the other trends they reported—namely the trends for \( P_c \) and \( \rho_c \) mathematically suggest a non-zero value for \( Z_c^\infty \). More specifically, the model parameters reported in the different Nikitin studies predict a \( Z_c^\infty \) value around 0.1 and 0.15 [15, 20, 29]. Thus, we did not employ the constraint that \( Z_c^\infty = 0 \) since it appeared inconsistent with the rest of their work.

We have already alluded to the study by Vega et al. that reports a \( Z_c^\infty \) of \( \frac{1}{3} \) [83]. This is interesting since this was the same group that reported the value of \( \frac{1}{3} \). They argued that these two values are not inconsistent since they arrived at the value of \( \frac{1}{3} \) by including an empirical correction factor in the original model that predicts \( Z_c \to \frac{1}{3} \). Specifically, they used Wertheim’s SAFT theory to support the value of \( \frac{1}{3} \) while they used the empirical soft-SAFT model for the \( \frac{1}{3} \) asymptote [83]. They arrived at a value of \( \frac{1}{3} \) by using two different values for \( \sigma \) (the LJ parameter, which is also a parameter in the soft-SAFT EoS). Although this provided good agreement with experimental \( Z_c \) data, assuming two different \( \sigma \) values does not appear to have a strong theoretical basis. Furthermore, the \( Z_c \) correlation they presented does not approach \( \frac{1}{3} \) smoothly. We experienced the same difficulty when developing our own model, namely, when we constrained \( Z_c \to \frac{1}{3} \) it did not converge monotonically.

Finally, Chickos et al. proposed that \( T_c \) and \( T_b \) do not intersect, rather they converge and, therefore, \( P_c \to 1 \text{ atm} \) [102]. These conclusions were based primarily upon extrapolating \( T_c \) and \( T_b \) data for several families and concluding that \( T_c \) and \( T_b \) for a specific family approach the same infinite chain-length value. However, the resulting \( T_c^\infty \) and \( T_b^\infty \) values obtained from extrapolation have such large uncertainties that it is difficult to draw a definitive conclusion on the matter. Furthermore, since all of the most recent polymer theories support \( P_c^\infty = 0 \), we have not required \( P_c \) to approach an asymptote at 1 atm or for \( T_c \) and \( T_b \) to converge.

### 6.8 Conclusions

In this chapter we regressed new coefficients to the same general mathematical form as the models proposed by Nikitin et al. for \( P_c, T_c, \rho_c, \) and \( T_b \). The coefficients reported in this study
differ from those reported by Nikitin et al. for three reasons. First, we meticulously scrutinized the data included in the model development. Second, we required that the $P_c$, $T_c$, and $\rho_c$ models be consistent with Wertheim’s polymer theory and that the $P_c$, $T_c$, and $T_b$ models be internally consistent. Third, we used additional terms in the model expansion, i.e. regression parameters.

For example, we included six parameters for $P_c$, four for $T_c$, and five for $\rho_c$ and $T_b$. By contrast, Nikitin used three parameters for $P_c$, $T_c$, and $\rho_c$ and four parameters for $T_b$. For this reason, it was important to ensure that the models were not over-fitting the data. This was accomplished by verifying that each regression parameter was statistically significant and that the models were well behaved, i.e. no inflection points. An additional criterion for selecting the number of regression parameters was that the model agree well with the simulation results. This is an example of how molecular simulation can provide insight into developing new prediction models. In addition, we utilized the Type B simulation uncertainties to provide a quantitative estimate for the prediction model uncertainties. We believe this is a reliable approach since we accounted for numerous layers of uncertainty in our simulation results and used a Type B analysis to determine the uncertainties in the force field parameters.

Furthermore, by reducing the simulation uncertainties in $P_c$ and $T_c$ we were able to discern between conflicting data. The decision of which data to include is pivotal in model development. We used Teja’s experimental data for $T_c$ and $\rho_c$ from C$_5$-C$_{18}$ and $P_c$ data from C$_5$-C$_{15}$. We also used Nikitin’s 2014 $P_c$ data from C$_{20}$-C$_{60}$. In addition, we demonstrated how extrapolating $\rho_l$ to predict $\rho_c$ can be beneficial for obtaining estimates for large $n$-alkanes. By including these points (from C$_{16}$-C$_{36}$) and constraining our $\rho_c$ correlation such that $Z_c \rightarrow \frac{1}{3}$, we obtained a more reliable estimate of $\rho_c$ for long-chain-lengths. Finally, the improved predictions of $P_c$ and $T_c$ had an impact on the new $T_b$ prediction model. Specifically, constraining $T_b = T_c$ at the same CN that $P_c = 1$ atm resulted in a slightly higher $T_b$ model than currently reported by DIPPR 801.
CHAPTER 7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

The importance of this research is founded in the ultimate goal of molecular simulation, namely, that the accuracy of molecular simulation can one day approach that of experiment. Uncertainty quantification is a vital tool to realize this goal. This research consisted of a systematic analysis of the uncertainty quantification and propagation of errors for several layers of the uncertainty hierarchy in molecular simulation. Specifically, we quantified the uncertainty: inherent to GEMC, from regression of data, design conditions, finite-size effects, model inadequacies, and intermolecular parameters. By being able to quantify the reliability of the predicted critical constants from GEMC we were able to elucidate the correct long-chain-length trends for $T_c, \rho_c, P_c,$ and $Z_c$ for large $n$-alkanes. In fact, we utilized our simulation results to assist in the development of new models for $T_c, \rho_c, P_c,$ and $T_b$ for large $n$-alkanes. We also utilized our simulation uncertainties to assign the prediction model uncertainties. Therefore, we successfully demonstrated that molecular simulation can play a significant role in thermophysical property data evaluation and prediction.

In Chapter 3 we investigated the numerical uncertainty that is inherent to GEMC simulations. Specifically, we observed that the uncertainty increases exponentially with respect to temperature. To eliminate the need for arduous data analysis when estimating the simulation uncertainties, we developed an error model for predicting the standard deviation in the liquid, vapor, rectilinear, and scaling densities. This model had several uses. First, it allowed for properly weighing the terms in the objective function used to predict $T_c$ and $\rho_c$. Second, it enabled the development of an experimental design to determine the ideal simulation temperatures to minimize the uncertainties in $T_c, \rho_c, P_c,$ and $Z_c$. The other significant contribution in Chapter 3 is a rigorous nonlinear statistical analysis. This approach is important for accurately determining the single parameter uncertainties in $T_c$ or $\rho_c$. It also provides greater insight by assessing the joint confidence region.
of \( T_c \) and \( \rho_c \). Furthermore, a PDF of the 4-dimensional parameter space allows for a Monte Carlo Sampling propagation of errors analysis. The utility of this approach becomes clear in Chapter 4.

In Chapter 4 we proposed an alternative approach for predicting \( P_c \). Specifically, the Vetere approach enables reliable predictions of \( P_c \) for force fields that do not accurately predict \( P_v \). We employed the MCS PoE method (with the PDF mentioned previously) to quantify the uncertainties in \( P_c \) and \( Z_c \). In addition, we developed an experimental design to minimize the uncertainty in the critical constants. Fortuitously, the experimental design also mitigates finite-size effects by avoiding the near critical temperature regime. With these numerous tools for quantifying and reducing numerical uncertainty we perform simulations for larger \( n \)-alkanes utilizing the TraPPE, NERD, and Exp-6 models. By comparing the results for the three different force fields, we obtained a simplistic assessment of model uncertainty. Specifically, we observed that the TraPPE approach slightly over predicts \( T_c \) while all three models predict nearly identical \( P_c \) and \( Z_c \) values. This is significant because it suggests that a simple (i.e transferable, UA, LJ 12-6) force field is reliable enough to predict \( P_c \) and \( Z_c \) trends. For this reason, we used the TraPPE model as a prototype for developing the Mess-UP force field in Chapter 5.

The deepest layer of uncertainty that we addressed in this research is the intermolecular potential parameter uncertainty. Specifically, we quantified the uncertainty in the LJ 12-6 parameters and propagated these uncertainties for the critical constants of large \( n \)-alkanes. We proposed three different approaches for uncertainty quantification. The most physically meaningful assessment of uncertainty utilized the DIPPR 801 uncertainties that account for possible bias in the experimental data. Although this approach resulted in the largest uncertainties in the critical constants, the results were still reliable enough to discern between conflicting experimental data and elucidate the correct long-chain-length trends.

The culmination of our research was the development of new prediction models for \( T_c, \rho_c, P_c, \) and \( T_b \). In Chapter 6 we demonstrated how our simulation results provided insight when developing our prediction models and assigning uncertainties to the prediction models. Specifically, our simulation results suggested that Nikitin’s most recent data for \( P_c \) are reliable [15]. By judiciously selecting the data and model order we developed an improved \( P_c \) prediction model. Furthermore, our simulation trends for \( T_c \) verified that Nikitin’s most recent data are more reliable than their original set [20]. However, their recent data were not included in the regression process since these
data were not internally consistent and appeared to have a bias. Despite only including data for compounds as large as $C_{18}$, our prediction model is in excellent agreement with our simulation results. Next, we generated a $\rho_c$ model that results in a $Z_c$ trend consistent with polymer theory while matching experimental data for smaller $n$-alkanes and simulation results for larger $n$-alkanes. Finally, although our simulation results are not capable of accurately predicting $T_b$, we were able to develop improved $T_b$ models by requiring internal consistency between $T_c$, $P_c$, and $T_b$.

In conclusion, molecular simulation is capable of complementing, validating, and correcting experimental data and prediction models. Uncertainty quantification is pivotal in this pursuit. Most simulation studies only report numerical uncertainties and these uncertainties are commonly determined using improper statistical practices. We have demonstrated how to rigorously quantify and reduce these numerical uncertainties for GEMC. In addition, we have proposed the Vetere approach for predicting $P_c$ and $Z_c$ to overcome force field model deficiencies (i.e. the inability for simple UA, LJ 12-6 models to predict both $\rho_l$ and $P_v$ accurately). Subsequently, we developed an experimental design for reducing the uncertainties in $T_c$, $\rho_c$, $P_c$, and $Z_c$. Having established a set of tools that enable reliable predictions of the critical constants, we verified that finite-size effects are negligible. Then, we quantified the uncertainty that is due to the imprecision in the force field parameters. Furthermore, we demonstrated that our methodology allows for reliable enough predictions to elucidate the correct long-chain-length trends for $n$-alkanes. Finally, we used the insight gained from simulation to develop new prediction models. Therefore, by addressing different layers of uncertainty we have enabled molecular simulation to be a quantitative tool in thermophysical property analysis. This is significant because molecular simulation results have generally not been included in the DIPPR 801 literature review process. We recommend that future projects focus on implementing molecular simulation more in the DIPPR 801 database.

7.2 Recommendations

7.2.1 Implementing Prediction Models in DIPPR 801

Based upon our findings, we recommend that the DIPPR 801 database implement the new prediction models for $T_c$, $P_c$, $\rho_c$, and $T_b$ of $n$-alkanes found in Table 6.1. This will have the most significant impact on the large $n$-alkanes that are currently being added to the database, namely,
C₄₀, C₄₄, C₅₀, C₆₀, and C₈₀. However, our results will also affect compounds already in the database, primarily those in the range C₁₆-C₃₆. Specifically, our T_c model will result in a 3-15 K shift from the current DIPPR 801 values for C₂₁-C₃₆, respectively. However, the P_c model may have the largest impact on the DIPPR 801 database. The proposed P_c model differs from the current DIPPR 801 values by 6%-15% for C₂₀-C₃₆, respectively. Not only are the values of P_c much different than those found in the database but the CN at which P_c = 1 atm is nearly three times what was originally proposed by DIPPR 801 staff. This has a considerable effect on the T_b model. Specifically, the new T_b model diverges from the current DIPPR 801 model for compounds larger than C₇₀. For compounds smaller than C₇₀ the difference in the T_b model is less than 1%. In fact, there is nearly a constant deviation on an absolute scale for the largest compounds currently in the DIPPR 801 database (C₃₆). To be specific, the new T_b model is 0.5 K higher than the current DIPPR T_b model for C₁₈-C₃₆. Finally, the new model for ρ_c does not result in a non-zero asymptote as previously suggested by DIPPR 801 staff. Surprisingly, this only results in about a 3% difference in ρ_c for C₃₆ compared to the current DIPPR 801 values. However, we believe that the new ρ_c curve matches the experimental data and simulation results better while having the correct long-chain-length trend. The overall improvement in the critical constants is most obvious in Z_c. The new models predict a Z_c trend that diverges dramatically from the current DIPPR 801 values at C₁₆. The percent deviation ranges from 7% for C₂₀ to 14% for C₃₆.

7.2.2 Predicting ρ_l and P_v

The proposed prediction models for T_c, P_c, and ρ_c allow for the purely predictive Rackett equation to be more useful. Despite this, when developing a ρ_l correlation we recommend that experimental ρ_l data be used when available. Although we have not assessed the effect of our new values on prediction methods for other physical properties, we would expect to see a significant improvement. An investigation of the impact that our T_c, P_c, and T_b models have on predicting P_v with the Riedel equation could be a beneficial endeavor [78].
7.2.3 Investigating Other Homologous Series

In addition to modifying the $n$-alkane trends, these new prediction models affect every other family in the DIPPR 801 database. Specifically, in order to ensure consistency between families the $P_c$ and $\rho_c$ trends for other homologous series should converge to zero at the appropriate rate. We recommend that the multi-property optimization approach discussed in Chapter 6 be repeated for $n$-alkenes, $n$-alcohols, etc. Whether or not $T^\infty_c$, $T^\infty_b$, and $Z^\infty_c$ values are the same for every homologous series has been debated [19, 102]. We recommend using the same $P^0_c$, $\rho^0_c$, $T^\infty_c$, and $Z^\infty_c$ values for every family (or at least to within the uncertainty in $T^\infty_c$). Since $T^\infty_b$ has no physical significance, we recommend that this parameter be family specific so $T_c = T_b$ at the same CN that $P_c = 1$ atm for each homologous series.

7.3 Future Work

I believe there are three logical directions that this research could proceed. The first obvious continuation of this research is to simulate larger $n$-alkanes. An alternative path is to investigate deeper layers of the uncertainty hierarchy. The third route is to repeat and expand upon our findings for other families. In addition, there are a few sub-projects that could return profitable results. These involve assessing the correlation between CH$_3$ and CH$_2$ parameters and the temperature dependence of the $\epsilon$ and $\sigma$ contours. We will discuss each of these topics in the remainder of this section.

7.3.1 Simulating Larger $n$-Alkanes

Obtaining simulation results for compounds that are larger than those measured experimentally ($C_{60}$) would help validate our prediction models. However, we were unable to simulate compounds larger than $C_{48}$, despite the fact that Siepmann suggested this should be possible with configurational-bias GEMC. Our attempt to simulate $C_{60}$ was frustrated by the infinitesimal probability of a successful insertion step into the liquid phase. We tried increasing the MC move probability for particle transfer but the number of attempts necessary to consistently yield a successful insertion was computationally crippling. In other words, nearly all of the computational resources were devoted to the particle transfer attempts.
Preliminary Simulation Results

In order to overcome this difficulty we investigated using a coarse-grained (CG) model rather than a UA model. We utilized the TEAM-CG model as it predicted the VLCC more accurately than the TraPPE-CG model [109, 110]. Both CG models combine three UA sites into a single site, reducing the number of interaction sites by a factor of three. Therefore, simulating $C_{60}$ with the CG model is a comparable cost to simulating $C_{20}$ with a UA model. We assumed that the CG approach would allow for prediction of compounds three times larger than the UA approach (approximately $C_{150}$). However, we were only able to obtain reliable results (i.e. GEMC simulations that converged) for compounds as large as $C_{120}$. We attribute this again to the difficulty of a particle insertion in the liquid phase. Whereas the difficult for $C_{60}$ with a UA model was the improbability of successively inserting 60 sites, the difficulty for $C_{150}$ with a CG model is the absolute size of the molecule. That is, although a successful particle transfer necessitates inserting fewer sites (50), since each of the CG sites requires a larger volume than a UA site the probability of a CG site is less than a UA site.

Although the CG approach made it feasible to simulate $C_{60}$-$C_{120}$ the performance of the TEAM-CG model depreciated with increasing chain-length. The reason for the poor extrapolation is because grouping together several heavy atoms (carbon) is a greater simplification than grouping together several non-acidic hydrogen atoms. Due to this high degree of coarse-graining, the TEAM-CG model utilizes a temperature dependent intermolecular potential to increase the “flexibility” of their model by introducing additional parameters. However, this temperature dependence appears to lead to a poor performance when simulating compounds not included in the training set. To be specific, we observed a systematic under prediction of $\rho_l$ for compounds larger than $C_{15}$ (TEAM-CG was optimized to match experimental VLCC data for compounds only as large as $C_{12}$). This resulted in a considerable under prediction of $T_c$ for larger compounds. Due to an apparent cancellation of errors, the $\rho_c$ and $P_c$ trends appear to be quite reliable as they agree with our prediction models. Unfortunately, it is not clear whether the $Z_c$ trend has the correct qualitative or quantitative behavior. Specifically, we do not observe a minimum or an asymptote for $Z_c$ of compounds as large as $C_{120}$ (see Appendix I for details).
Future Simulation Work

Based upon our experience with the available CG models, an alternative approach is necessary to obtain reliable results for larger $n$-alkanes. There are at least three different methods for realizing this goal: first, using MD rather than MC; second, using Cassandra or GOMC instead of Towhee; third, developing an improved CG model.

MD has been implemented for compounds as large as C$_{100}$ [26]. However, as explained in Chapter 2 the primary weakness of MD is the large uncertainties and finite-size effects. Reducing the uncertainties in the orthobaric densities would require improving upon the Voronoi Tessellations approach or that proposed by Patel et al. [37, 38] Furthermore, it would be essential to reduce the degree of finite-size effects by determining the minimum number of molecules and the necessary cutoff length. Due to the increased computational cost incurred by utilizing a larger system and longer cutoff length, it is unclear if MD is an attractive approach for accurately predicting the critical constants of large compounds.

Cassandra and GOMC present extremely promising possibilities for performing GEMC simulations of larger compounds [48, 49]. The developers of Cassandra and GOMC have demonstrated that these codes scale exceptionally well with increasing system size. We participated in a Cassandra workshop and have also performed simulations with Cassandra on the BYU Fulton Supercomputer. From our experience, Cassandra is considerably faster than Towhee. We believe that simulating C$_{60}$, and possibly larger compounds, would be feasible with Cassandra. Although GOMC has been shown to be superior to Cassandra and Towhee in a specific scenario [88], we recommend that a future researcher become familiar with all three of these source codes so that the user can decide which code is best for a given situation.

Regardless of whether MD, Cassandra, or GOMC is utilized, a CG model could be extremely useful as it reduces the number of sites by a factor of three compared to a UA model. We recommend developing a CG model that extrapolates more reliably to larger molecules than the TEAM-CG model. We believe that this could be done by re-parameterizing the TEAM-CG model to match experimental data for compounds larger than C$_{12}$. To verify that this model extrapolates well to larger compounds, the CG results for C$_{48}$ should be compared to the UA results for C$_{48}$. This would be a significant contribution as it would permit simulations of compounds as large as C$_{100}$ that could be helpful in validating the new prediction models. Furthermore, CG models are
widely popular in bio-molecular simulations due to their reduction in computational cost for large species.

7.3.2 Additional Layers of Uncertainty

In addition, we recommend investigating deeper layers of the uncertainty hierarchy. The next layers would be the type of site model chosen, the intramolecular model, and the data used in regression. Comparing the performance of an AA, UA, anisotropic-united-atom (AUA), and CG model would be very beneficial to the simulation community. As previously stated, we recommend that the focus be on developing a CG model that can replicate the UA results for C_{48}. We would also recommend investigating the effect of bond lengths between UA sites and the optimal LJ parameters. In this study we did not quantitatively account for different bond lengths or flexible bond lengths. However, the Exp-6 and TraPPE/Mess-UP models use slightly different bond lengths and the NERD model uses a flexible bond length (with the same equilibrium distance as the TraPPE/Mess-UP models). Therefore, it would be insightful to reoptimize the Mess-UP model using the bond lengths from the Exp-6 model and the flexible bond length from the NERD model. Most likely the effect on the optimal $\varepsilon$ and $\sigma$ values would be negligible, especially since the bond deviations are quite small. Since the dihedral angles may have a significant impact on large compounds, we also recommend quantifying the uncertainty related to the torsional potential.

There are several ways to investigate the effect of the data used in the regression on the optimal parameters and the corresponding uncertainties. The most obvious choice is to use a different type of data, i.e. $P_v$ or $B_2$ rather than $\rho_l$. However, it is a well known fact that this will drastically change the optimal parameters and their uncertainties because of the model inabilities to match both vapor and liquid phase properties (recall Figure 5.1). Perhaps a more interesting investigation is the effect of the temperature range and the compounds included in the parameterization. For example, we would recommend re-optimizing the CH$_2$ parameters using larger $n$-alkanes than $n$-octane, similar to what Nath et al. did when developing the NERD model [25]. By utilizing C$_{12}$, C$_{16}$, etc. the CH$_3$ parameters will be less influential on the optimal CH$_2$ parameters. Furthermore, the predictability of the force field will be improved by including several compounds because it is unlikely that a bias exists for each compound.
7.3.3 Extend Mess-UP Force Field

The third future project that we recommend is to investigate other families of compounds, particularly hydrocarbons such as branched alkanes, \( n \)-alkenes, \( n \)-alkynes, etc. Verifying that these families converge to the \( n \)-alkanes would be a significant conclusion for the DIPPR 801 database. By repeating the uncertainty quantification process for the \( n \)-alkene family, for example, it could be possible to determine if the \( T_c \) trend for \( n \)-alkenes converge to or simply maintain a constant deviation from the \( n \)-alkane family trend for \( T_c \). Since we have already developed the tools for quantifying and reducing the numerical uncertainties, the primary focus of this project would be the uncertainties in the LJ 12-6 parameters. We also recommend that this include an investigation into possible advantages of the Bayesian approach for UQ+PoE in comparison with the Type A and Type B methodologies presented in this research.

Polar compounds present an added level of complexity that might require more sophisticated force field models than the simple UA, LJ 12-6 model. For this reason, we believe it would be best to focus on non-polar hydrocarbons rather than alcohols, acids, and other associating and/or polar compounds. Specifically, we recommend extending the Mess-UP force field by regressing parameters for C, CH, CH\( = \), CH\( \equiv \), etc. The compounds and data included in the parameterization process could either be the same as those used in the TraPPE development or an improved training set, if deemed necessary. Again, by assuming transferability of the CH\( _3 \) and CH\( _2 \) parameter sets, scanning the 2-dimensional parameter space systematically is feasible. However, it may become necessary to account for the correlation between the optimal CH\( _3 \), CH\( _2 \) parameter sets and those for additional UA sites.

Accounting for the correlation between the CH\( _3 \) and CH\( _2 \) parameters is one of the biggest difficulties in the UQ+PoE approach. The Monte Carlo Sampling of the CH\( _2 \) parameter space should account for uncertainties in the CH\( _3 \) values. This analysis would be greatly simplified if approximate but reliable analytic expressions were developed to relate the CH\( _2 \) parameters to the CH\( _3 \) parameters. In other words, if the MCS value for \( \sigma_{CH_3} \) deviates from the optimal value the \( \sigma_{CH_2} \) parameter should likely account for this deviation. In Appendix G we present some preliminary work on this topic.

In Figures 5.18-5.19 we presented some curious trends for the \( \varepsilon_{CH_2} \) and \( \sigma_{CH_2} \) correlation with respect to temperature. Recall that at low temperatures of \( n \)-octane (and for all temperatures
of ethane) we observe the expected result that an increase in $\varepsilon$ necessitates an increase in $\sigma$ for a constant $\rho_l$. However, with increasing temperature for $n$-octane the dependence is exactly opposite. Determining the cause of this inversion in correlation could provide significant molecular insight. For example, when developing a force field model it could be beneficial for determining $\varepsilon$ to perform simulations at 440 K for $n$-octane because $\rho_l$ only depends on $\varepsilon$. However, at this temperature you do not obtain any information about $\sigma$. In Section 5.7.2 we proposed a simple explanation for this phenomenon. Essentially, we believe that this observation is unique to saturated liquid density. We do not believe such a trend would be observed for constant pressure liquid density. To test this theory it would be a simple exercise to scan the $\varepsilon_{CH_2}$-$\sigma_{CH_2}$ parameter space by performing simulations at a constant pressure and various temperatures. If this change in the contour direction is still observed an alternative explanation would be necessary.
REFERENCES


[15] Nikitin, E. D., and Popov, A. P., 2014. “Critical temperatures and pressures of C$_{40}$, C$_{44}$, and C$_{60}$ normal alkanes measured by the pulse-heating technique.” Fluid Phase Equilibria, 379, pp. 191–195. 1, 2, 4, 50, 53, 54, 78, 80, 84, 85, 86, 127, 128, 141, 143, 144, 147, 149, 161, 164


[47] Martin, M. G. MCCCS Towhee. 13, 33

[48] Maginn, E. Cassandra. 13, 169

[49] Potoff, J. GOMC. 13, 169


180


APPENDIX A. VALUES USED IN CHAPTER 3

Tables A.1-A.3 contain the data used to generate Figure 3.5 Panels (a), (b), and Table 3.2, respectively. The terms $S_{El}$, $S_{Ev}$, $S_{Er}$, and $S_{Es}$ refer to the standard error of the mean for the liquid, vapor, rectilinear, and scaling densities, respectively. The population standard deviations ($S_D$) in Tables A.1 and A.3 were estimated using the error model presented in Table 3.1. By contrast, since Ref [72] used several (8) replicate simulations, it was assumed that the sample standard deviation ($s$) is an accurate measure of the population standard deviation $S_D$. Therefore, both the traditional and rigorous methods in Table A.2 utilized the uncertainties reported in the literature together with Equation 3.5. This was done so that the comparison could be done on the same basis. In general, the error model should be used for a completely rigorous analysis. Since the error model was obtained from 20 independent replicates, the assumption that $s \approx S_D$ is more reliable. Refer to Chapter 3 for a clarification between when $S_E$ or $s$ is utilized in the “traditional” analysis. No distinction needs to be made for the “rigorous” analysis because the ratio between $S_E$ and $s$ is a constant value. This factor will not affect Equation 2.19 since it will be included in both $S(\theta)$ and $S(\hat{\theta})$.

Table A.1: Summary of the data used in the analysis corresponding to Figure 3.5 Panel (a). The column “Traditional” was used in the “Traditional Linearization” while the “Rigorous” column was used in the “Rigorous Algorithm”. Units for densities and uncertainties are kg/L.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>$\rho_l$</th>
<th>$\rho_v$</th>
<th>$s_l$</th>
<th>$s_v$</th>
<th>$S_{Ev} = S_{Es}$</th>
<th>$S_{Dv}$</th>
<th>$S_{Ds}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>236</td>
<td>0.46898</td>
<td>0.01988</td>
<td>0.00175</td>
<td>0.00168</td>
<td>0.00054</td>
<td>0.00059</td>
<td>0.00042</td>
</tr>
<tr>
<td>256</td>
<td>0.43228</td>
<td>0.03096</td>
<td>0.00524</td>
<td>0.00450</td>
<td>0.00155</td>
<td>0.00090</td>
<td>0.00052</td>
</tr>
<tr>
<td>275</td>
<td>0.39554</td>
<td>0.05556</td>
<td>0.00522</td>
<td>0.00798</td>
<td>0.00213</td>
<td>0.00212</td>
<td>0.00105</td>
</tr>
</tbody>
</table>
Table A.2: Summary of the data used in the analysis corresponding to Figure 3.5 Panel (b). The column “Traditional” was used in the “Traditional Linearization” while the “Rigorous” column was used in the “Rigorous Algorithm”.
Units for densities and uncertainties are kg/L.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>GEMC results according to Ref [72]</th>
<th>Traditional</th>
<th>Rigorous</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$</td>
<td>$\rho_v$</td>
<td>$s_l \approx S_{D_l}$</td>
</tr>
<tr>
<td>275</td>
<td>0.3937</td>
<td>0.0598</td>
<td>0.0003</td>
</tr>
<tr>
<td>279</td>
<td>0.3835</td>
<td>0.0648</td>
<td>0.0004</td>
</tr>
<tr>
<td>283</td>
<td>0.3726</td>
<td>0.0739</td>
<td>0.0003</td>
</tr>
<tr>
<td>288</td>
<td>0.3589</td>
<td>0.090</td>
<td>0.0006</td>
</tr>
</tbody>
</table>
Table A.3: Summary of the data used in the analysis corresponding to Table 3.2. Units for densities and uncertainties are $\frac{kg}{L}$.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>GEMC Results</th>
<th>Blocking Equation 3.5</th>
<th>Blocking Direct</th>
<th>Table 3.1 Equation 3.5</th>
<th>Table 3.1 Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$</td>
<td>$\rho_v$</td>
<td>$S_{E_l}$</td>
<td>$S_{E_v}$</td>
<td>$S_{E_r} = S_{E_s}$</td>
</tr>
<tr>
<td>390</td>
<td>0.6259</td>
<td>0.0042</td>
<td>0.0008</td>
<td>0.0001</td>
<td>0.0004</td>
</tr>
<tr>
<td>440</td>
<td>0.5739</td>
<td>0.0133</td>
<td>0.0009</td>
<td>0.0005</td>
<td>0.0005</td>
</tr>
<tr>
<td>490</td>
<td>0.5119</td>
<td>0.0329</td>
<td>0.0010</td>
<td>0.0012</td>
<td>0.0008</td>
</tr>
<tr>
<td>515</td>
<td>0.4752</td>
<td>0.0507</td>
<td>0.0021</td>
<td>0.0019</td>
<td>0.0014</td>
</tr>
<tr>
<td>543</td>
<td>0.4190</td>
<td>0.0841</td>
<td>0.0041</td>
<td>0.0031</td>
<td>0.0026</td>
</tr>
</tbody>
</table>
APPENDIX B. VAPOR-LIQUID SIMULATION DATA IN CHAPTER 4

Table B.1: Vapor-liquid coexistence data for C₆, C₈, C₁₀, C₁₂, C₁₆, and C₂₄ obtained with the NERD and TraPPE models for a system size of 200 molecules.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>NERD</th>
<th>TraPPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ρ₁ (kg/L)</td>
<td>ρᵥ (kg/L)</td>
</tr>
<tr>
<td>n-Hexane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>360</td>
<td>0.5821</td>
<td>0.00581</td>
</tr>
<tr>
<td>360</td>
<td>0.5820</td>
<td>0.00568</td>
</tr>
<tr>
<td>360</td>
<td>0.5820</td>
<td>0.00576</td>
</tr>
<tr>
<td>440</td>
<td>0.4903</td>
<td>0.03110</td>
</tr>
<tr>
<td>440</td>
<td>0.4921</td>
<td>0.03171</td>
</tr>
<tr>
<td>440</td>
<td>0.4915</td>
<td>0.03034</td>
</tr>
<tr>
<td>n-Octane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>390</td>
<td>0.6113</td>
<td>0.00332</td>
</tr>
<tr>
<td>390</td>
<td>0.6109</td>
<td>0.00320</td>
</tr>
<tr>
<td>390</td>
<td>0.6112</td>
<td>0.00315</td>
</tr>
<tr>
<td>490</td>
<td>0.5044</td>
<td>0.02724</td>
</tr>
<tr>
<td>490</td>
<td>0.5050</td>
<td>0.02774</td>
</tr>
<tr>
<td>490</td>
<td>0.5051</td>
<td>0.02780</td>
</tr>
<tr>
<td>n-Decane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>435</td>
<td>0.6092</td>
<td>0.00339</td>
</tr>
<tr>
<td>435</td>
<td>0.6097</td>
<td>0.00360</td>
</tr>
<tr>
<td>435</td>
<td>0.6094</td>
<td>0.00352</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>NERD</th>
<th>TraPPE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$ (kg/L)</td>
<td>$\rho_v$ (kg/L)</td>
</tr>
<tr>
<td>535</td>
<td>0.5061</td>
<td>0.02697</td>
</tr>
<tr>
<td>535</td>
<td>0.5035</td>
<td>0.02708</td>
</tr>
<tr>
<td>535</td>
<td>0.5076</td>
<td>0.02836</td>
</tr>
</tbody>
</table>

$n$-Dodecane

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>465</td>
<td>0.6118</td>
<td>0.00315</td>
<td>0.6238</td>
<td>0.00355</td>
</tr>
<tr>
<td>465</td>
<td>0.6133</td>
<td>0.00309</td>
<td>0.6229</td>
<td>0.00334</td>
</tr>
<tr>
<td>465</td>
<td>0.6133</td>
<td>0.00317</td>
<td>0.6232</td>
<td>0.00358</td>
</tr>
<tr>
<td>575</td>
<td>0.4984</td>
<td>0.02690</td>
<td>0.5078</td>
<td>0.02937</td>
</tr>
<tr>
<td>575</td>
<td>0.5002</td>
<td>0.02810</td>
<td>0.5069</td>
<td>0.02957</td>
</tr>
<tr>
<td>575</td>
<td>0.4979</td>
<td>0.02687</td>
<td>0.5068</td>
<td>0.02977</td>
</tr>
</tbody>
</table>

$n$-Hexadecane

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>510</td>
<td>0.6138</td>
<td>0.00199</td>
<td>0.6246</td>
<td>0.00226</td>
</tr>
<tr>
<td>510</td>
<td>0.6156</td>
<td>0.00261</td>
<td>0.6239</td>
<td>0.00235</td>
</tr>
<tr>
<td>510</td>
<td>0.6156</td>
<td>0.00222</td>
<td>0.6250</td>
<td>0.00245</td>
</tr>
<tr>
<td>625</td>
<td>0.4996</td>
<td>0.02187</td>
<td>0.5108</td>
<td>0.02259</td>
</tr>
<tr>
<td>625</td>
<td>0.5037</td>
<td>0.02304</td>
<td>0.5078</td>
<td>0.02084</td>
</tr>
<tr>
<td>625</td>
<td>0.5000</td>
<td>0.02063</td>
<td>0.5116</td>
<td>0.02273</td>
</tr>
</tbody>
</table>

$n$-Tetracosane

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>580</td>
<td>0.6058</td>
<td>0.00168</td>
<td>0.6124</td>
<td>0.00133</td>
</tr>
<tr>
<td>580</td>
<td>0.6052</td>
<td>0.00159</td>
<td>0.6124</td>
<td>0.00137</td>
</tr>
<tr>
<td>580</td>
<td>0.6050</td>
<td>0.00169</td>
<td>0.6124</td>
<td>0.00111</td>
</tr>
<tr>
<td>700</td>
<td>0.4954</td>
<td>0.01695</td>
<td>0.5034</td>
<td>0.01657</td>
</tr>
<tr>
<td>700</td>
<td>0.4957</td>
<td>0.01809</td>
<td>0.5026</td>
<td>0.01462</td>
</tr>
<tr>
<td>700</td>
<td>0.4974</td>
<td>0.01724</td>
<td>0.5026</td>
<td>0.01610</td>
</tr>
</tbody>
</table>
Table B.3: Vapor-liquid coexistence data for C_{24}, C_{36}, and C_{48} using the Exp-6, NERD, and TraPPE models with system sizes of 200 and 1600 molecules.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \rho_l ) (kg/L)</td>
<td>( \rho_v ) (kg/L)</td>
</tr>
<tr>
<td>n-Tetracosane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>560</td>
<td>0.6182</td>
<td>0.00049</td>
</tr>
<tr>
<td>560</td>
<td>0.6175</td>
<td>0.00054</td>
</tr>
<tr>
<td>560</td>
<td>0.6189</td>
<td>0.00051</td>
</tr>
<tr>
<td>560</td>
<td>0.6172</td>
<td>0.00046</td>
</tr>
<tr>
<td>560</td>
<td>0.6172</td>
<td>0.00047</td>
</tr>
<tr>
<td>600</td>
<td>0.5848</td>
<td>0.00157</td>
</tr>
<tr>
<td>600</td>
<td>0.5859</td>
<td>0.00165</td>
</tr>
<tr>
<td>600</td>
<td>0.5866</td>
<td>0.00176</td>
</tr>
<tr>
<td>600</td>
<td>0.5849</td>
<td>0.00161</td>
</tr>
<tr>
<td>600</td>
<td>0.5861</td>
<td>0.00183</td>
</tr>
<tr>
<td>680</td>
<td>0.5122</td>
<td>0.00939</td>
</tr>
<tr>
<td>680</td>
<td>0.5114</td>
<td>0.00913</td>
</tr>
<tr>
<td>680</td>
<td>0.5098</td>
<td>0.00919</td>
</tr>
<tr>
<td>680</td>
<td>0.5108</td>
<td>0.00891</td>
</tr>
<tr>
<td>680</td>
<td>0.5137</td>
<td>0.00983</td>
</tr>
<tr>
<td>715</td>
<td>0.4703</td>
<td>0.01834</td>
</tr>
<tr>
<td>715</td>
<td>0.4727</td>
<td>0.01877</td>
</tr>
<tr>
<td>715</td>
<td>0.4722</td>
<td>0.01828</td>
</tr>
<tr>
<td>715</td>
<td>0.4691</td>
<td>0.01791</td>
</tr>
<tr>
<td>715</td>
<td>0.4682</td>
<td>0.01812</td>
</tr>
<tr>
<td>n-Hexatriacontane</td>
<td></td>
<td></td>
</tr>
<tr>
<td>650</td>
<td>0.5885</td>
<td>0.00082</td>
</tr>
<tr>
<td>650</td>
<td>0.5871</td>
<td>0.00099</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$ (kg/L)</td>
<td>$\rho_v$ (kg/L)</td>
</tr>
<tr>
<td>650</td>
<td>0.5867</td>
<td>0.00075</td>
</tr>
<tr>
<td>650</td>
<td>0.5872</td>
<td>0.00092</td>
</tr>
<tr>
<td>650</td>
<td>0.5880</td>
<td>0.00098</td>
</tr>
<tr>
<td>730</td>
<td>0.5181</td>
<td>0.00530</td>
</tr>
<tr>
<td>730</td>
<td>0.5186</td>
<td>0.00578</td>
</tr>
<tr>
<td>730</td>
<td>0.5180</td>
<td>0.00509</td>
</tr>
<tr>
<td>730</td>
<td>0.5197</td>
<td>0.00559</td>
</tr>
<tr>
<td>730</td>
<td>0.5186</td>
<td>0.00580</td>
</tr>
<tr>
<td>765</td>
<td>0.4855</td>
<td>0.01050</td>
</tr>
<tr>
<td>765</td>
<td>0.4845</td>
<td>0.01103</td>
</tr>
<tr>
<td>765</td>
<td>0.4838</td>
<td>0.01020</td>
</tr>
<tr>
<td>765</td>
<td>0.4855</td>
<td>0.01099</td>
</tr>
<tr>
<td>765</td>
<td>0.4866</td>
<td>0.01131</td>
</tr>
<tr>
<td>800</td>
<td>0.4473</td>
<td>0.01914</td>
</tr>
<tr>
<td>800</td>
<td>0.4478</td>
<td>0.01978</td>
</tr>
<tr>
<td>800</td>
<td>0.4484</td>
<td>0.02072</td>
</tr>
<tr>
<td>800</td>
<td>0.4447</td>
<td>0.01875</td>
</tr>
<tr>
<td>800</td>
<td>0.4460</td>
<td>0.01967</td>
</tr>
</tbody>
</table>

$n$-Hexatriacontane

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>$\rho_l$ (kg/L)</th>
<th>$\rho_v$ (kg/L)</th>
<th>$\rho_l$ (kg/L)</th>
<th>$\rho_v$ (kg/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>650</td>
<td>0.5934</td>
<td>0.00073</td>
<td>0.5925</td>
<td>0.00061</td>
</tr>
<tr>
<td>650</td>
<td>0.5935</td>
<td>0.00064</td>
<td>0.5917</td>
<td>0.00057</td>
</tr>
<tr>
<td>650</td>
<td>0.5937</td>
<td>0.00077</td>
<td>0.5923</td>
<td>0.00063</td>
</tr>
<tr>
<td>650</td>
<td>0.5947</td>
<td>0.00071</td>
<td>0.5921</td>
<td>0.00061</td>
</tr>
<tr>
<td>650</td>
<td>0.5937</td>
<td>0.00064</td>
<td>0.5919</td>
<td>0.00060</td>
</tr>
<tr>
<td>730</td>
<td>0.5283</td>
<td>0.00442</td>
<td>0.5306</td>
<td>0.00488</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ρl (kg/L)</td>
<td>ρv (kg/L)</td>
</tr>
<tr>
<td></td>
<td>ρl (kg/L)</td>
<td>ρv (kg/L)</td>
</tr>
<tr>
<td>730</td>
<td>0.5296</td>
<td>0.00498</td>
</tr>
<tr>
<td>730</td>
<td>0.5287</td>
<td>0.00450</td>
</tr>
<tr>
<td>730</td>
<td>0.5277</td>
<td>0.00522</td>
</tr>
<tr>
<td>730</td>
<td>0.5296</td>
<td>0.00496</td>
</tr>
<tr>
<td>775</td>
<td>0.4858</td>
<td>0.01115</td>
</tr>
<tr>
<td>775</td>
<td>0.4865</td>
<td>0.01076</td>
</tr>
<tr>
<td>775</td>
<td>0.4869</td>
<td>0.01034</td>
</tr>
<tr>
<td>775</td>
<td>0.4865</td>
<td>0.01046</td>
</tr>
<tr>
<td>775</td>
<td>0.4864</td>
<td>0.01092</td>
</tr>
<tr>
<td>775</td>
<td>0.4868</td>
<td>0.01935</td>
</tr>
<tr>
<td>775</td>
<td>0.4468</td>
<td>0.01939</td>
</tr>
<tr>
<td>775</td>
<td>0.4498</td>
<td>0.01974</td>
</tr>
<tr>
<td>775</td>
<td>0.4494</td>
<td>0.01952</td>
</tr>
<tr>
<td>775</td>
<td>0.4511</td>
<td>0.01980</td>
</tr>
<tr>
<td>810</td>
<td>0.4468</td>
<td>0.01935</td>
</tr>
<tr>
<td>810</td>
<td>0.4485</td>
<td>0.01939</td>
</tr>
<tr>
<td>810</td>
<td>0.4498</td>
<td>0.01974</td>
</tr>
<tr>
<td>810</td>
<td>0.4494</td>
<td>0.01952</td>
</tr>
<tr>
<td>810</td>
<td>0.4511</td>
<td>0.01980</td>
</tr>
</tbody>
</table>

n-Octatetracontane

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>Exponential-6 Model</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ρl (kg/L)</td>
<td>ρv (kg/L)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ρl (kg/L)</td>
<td>ρv (kg/L)</td>
</tr>
<tr>
<td>700</td>
<td>0.5623</td>
<td>0.00038</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.5614</td>
<td>0.00029</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.5600</td>
<td>0.00029</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.5584</td>
<td>0.00038</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.5595</td>
<td>0.00025</td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.4806</td>
<td>0.00600</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>0.4785</td>
<td>0.00589</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>0.4808</td>
<td>0.00642</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>0.4790</td>
<td>0.00625</td>
<td></td>
</tr>
<tr>
<td>800</td>
<td>0.4806</td>
<td>0.00600</td>
<td></td>
</tr>
</tbody>
</table>

Continued on next page
Table B.4 – continued from previous page

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$ (kg/L)</td>
<td>$\rho_v$ (kg/L)</td>
</tr>
<tr>
<td>835</td>
<td>0.4446</td>
<td>0.01017</td>
</tr>
<tr>
<td>835</td>
<td>0.4441</td>
<td>0.00983</td>
</tr>
<tr>
<td>835</td>
<td>0.4430</td>
<td>0.01106</td>
</tr>
<tr>
<td>835</td>
<td>0.4400</td>
<td>0.01060</td>
</tr>
<tr>
<td>835</td>
<td>0.4445</td>
<td>0.01156</td>
</tr>
<tr>
<td>810</td>
<td>0.4468</td>
<td>0.01935</td>
</tr>
<tr>
<td>810</td>
<td>0.4485</td>
<td>0.01939</td>
</tr>
<tr>
<td>810</td>
<td>0.4498</td>
<td>0.01974</td>
</tr>
<tr>
<td>810</td>
<td>0.4494</td>
<td>0.01952</td>
</tr>
<tr>
<td>810</td>
<td>0.4511</td>
<td>0.01980</td>
</tr>
</tbody>
</table>

$n$-Octatetracontane

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>NERD Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>730</td>
<td>0.5425</td>
</tr>
<tr>
<td>730</td>
<td>0.5431</td>
</tr>
<tr>
<td>730</td>
<td>0.5441</td>
</tr>
<tr>
<td>730</td>
<td>0.5451</td>
</tr>
<tr>
<td>730</td>
<td>0.5446</td>
</tr>
<tr>
<td>770</td>
<td>0.5122</td>
</tr>
<tr>
<td>770</td>
<td>0.5126</td>
</tr>
<tr>
<td>770</td>
<td>0.5131</td>
</tr>
<tr>
<td>770</td>
<td>0.5120</td>
</tr>
<tr>
<td>770</td>
<td>0.5102</td>
</tr>
<tr>
<td>800</td>
<td>0.4829</td>
</tr>
<tr>
<td>800</td>
<td>0.4835</td>
</tr>
<tr>
<td>800</td>
<td>0.4819</td>
</tr>
<tr>
<td>800</td>
<td>0.4830</td>
</tr>
</tbody>
</table>

Continued on next page
## Table B.4 – continued from previous page

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$ (kg/L)</td>
<td>$\rho_v$ (kg/L)</td>
</tr>
<tr>
<td>800</td>
<td>0.4855</td>
<td>0.00656</td>
</tr>
<tr>
<td>830</td>
<td>0.4542</td>
<td>0.01076</td>
</tr>
<tr>
<td>830</td>
<td>0.4549</td>
<td>0.01173</td>
</tr>
<tr>
<td>830</td>
<td>0.4562</td>
<td>0.01147</td>
</tr>
<tr>
<td>830</td>
<td>0.4552</td>
<td>0.01154</td>
</tr>
<tr>
<td>830</td>
<td>0.4546</td>
<td>0.01051</td>
</tr>
<tr>
<td>860</td>
<td>0.4194</td>
<td>0.01665</td>
</tr>
<tr>
<td>860</td>
<td>0.4210</td>
<td>0.01824</td>
</tr>
<tr>
<td>860</td>
<td>0.4190</td>
<td>0.01815</td>
</tr>
<tr>
<td>860</td>
<td>0.4233</td>
<td>0.01933</td>
</tr>
<tr>
<td>860</td>
<td>0.4209</td>
<td>0.01834</td>
</tr>
</tbody>
</table>

**n-Octatetracontane**

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>TrapPE Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>730</td>
<td>0.5528</td>
</tr>
<tr>
<td>730</td>
<td>0.5542</td>
</tr>
<tr>
<td>730</td>
<td>0.5537</td>
</tr>
<tr>
<td>730</td>
<td>0.5525</td>
</tr>
<tr>
<td>730</td>
<td>0.5522</td>
</tr>
<tr>
<td>770</td>
<td>0.5220</td>
</tr>
<tr>
<td>770</td>
<td>0.5210</td>
</tr>
<tr>
<td>770</td>
<td>0.5200</td>
</tr>
<tr>
<td>770</td>
<td>0.5245</td>
</tr>
<tr>
<td>770</td>
<td>0.5223</td>
</tr>
<tr>
<td>800</td>
<td>0.4975</td>
</tr>
<tr>
<td>800</td>
<td>0.4950</td>
</tr>
<tr>
<td>800</td>
<td>0.4961</td>
</tr>
</tbody>
</table>

Continued on next page
Table B.4 – continued from previous page

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>200 Molecules</th>
<th>1600 Molecules</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\rho_l$ (kg/L)</td>
<td>$\rho_v$ (kg/L)</td>
</tr>
<tr>
<td>800</td>
<td>0.4955</td>
<td>0.00512</td>
</tr>
<tr>
<td>800</td>
<td>0.4971</td>
<td>0.00492</td>
</tr>
<tr>
<td>830</td>
<td>0.4681</td>
<td>0.00885</td>
</tr>
<tr>
<td>830</td>
<td>0.4678</td>
<td>0.00890</td>
</tr>
<tr>
<td>830</td>
<td>0.4663</td>
<td>0.00812</td>
</tr>
<tr>
<td>830</td>
<td>0.4674</td>
<td>0.00875</td>
</tr>
<tr>
<td>830</td>
<td>0.4662</td>
<td>0.00802</td>
</tr>
<tr>
<td>860</td>
<td>0.4379</td>
<td>0.01467</td>
</tr>
<tr>
<td>860</td>
<td>0.4343</td>
<td>0.01525</td>
</tr>
<tr>
<td>860</td>
<td>0.4379</td>
<td>0.01516</td>
</tr>
<tr>
<td>860</td>
<td>0.4350</td>
<td>0.01399</td>
</tr>
<tr>
<td>860</td>
<td>0.4369</td>
<td>0.01451</td>
</tr>
</tbody>
</table>
APPENDIX C. SIMULATION RESULTS IN CHAPTER 4

Table C.1 contains the results of the simulations for each condition outlined in Table 4.1. The predicted critical constant values with their corresponding 95% confidence intervals are listed.
Table C.1: The predicted $T_c$, $\rho_c$, $P_c$, and $Z_c$ values for the different compounds, force field models, and system sizes simulated in this study (where $N$ is the total number of molecules). Uncertainties are reported at the 95% confidence level.

<table>
<thead>
<tr>
<th>Compound</th>
<th>Model</th>
<th>$N$</th>
<th>$T_c$ (K)</th>
<th>$\rho_c$ (kg/L)</th>
<th>$P_c$ (MPa)</th>
<th>$Z_c$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$-Hexane (C$_6$)</td>
<td>NERD</td>
<td>200</td>
<td>519 ± 1</td>
<td>0.229 ± 0.002</td>
<td>3.10 ± 0.05</td>
<td>0.270 ± 0.003</td>
</tr>
<tr>
<td>$n$-Hexane (C$_6$)</td>
<td>TraPPE</td>
<td>200</td>
<td>505 ± 2</td>
<td>0.240 ± 0.005</td>
<td>3.2 ± 0.1</td>
<td>0.271 ± 0.006</td>
</tr>
<tr>
<td>$n$-Octane (C$_8$)</td>
<td>NERD</td>
<td>200</td>
<td>579 ± 1</td>
<td>0.230 ± 0.001</td>
<td>2.52 ± 0.02</td>
<td>0.260 ± 0.001</td>
</tr>
<tr>
<td>$n$-Octane (C$_8$)</td>
<td>TraPPE</td>
<td>200</td>
<td>571 ± 1</td>
<td>0.238 ± 0.001</td>
<td>2.60 ± 0.03</td>
<td>0.262 ± 0.002</td>
</tr>
<tr>
<td>$n$-Decane (C$_{10}$)</td>
<td>NERD</td>
<td>200</td>
<td>626 ± 2</td>
<td>0.230 ± 0.003</td>
<td>2.15 ± 0.06</td>
<td>0.255 ± 0.004</td>
</tr>
<tr>
<td>$n$-Decane (C$_{10}$)</td>
<td>TraPPE</td>
<td>200</td>
<td>622 ± 1</td>
<td>0.236 ± 0.002</td>
<td>2.18 ± 0.04</td>
<td>0.255 ± 0.003</td>
</tr>
<tr>
<td>$n$-Dodecane (C$_{12}$)</td>
<td>NERD</td>
<td>200</td>
<td>664 ± 2</td>
<td>0.227 ± 0.003</td>
<td>1.81 ± 0.04</td>
<td>0.246 ± 0.004</td>
</tr>
<tr>
<td>$n$-Dodecane (C$_{12}$)</td>
<td>TraPPE</td>
<td>200</td>
<td>662 ± 1</td>
<td>0.233 ± 0.002</td>
<td>1.87 ± 0.03</td>
<td>0.248 ± 0.002</td>
</tr>
<tr>
<td>$n$-Hexadecane (C$_{16}$)</td>
<td>NERD</td>
<td>200</td>
<td>725 ± 3</td>
<td>0.221 ± 0.004</td>
<td>1.38 ± 0.06</td>
<td>0.234 ± 0.006</td>
</tr>
<tr>
<td>$n$-Hexadecane (C$_{16}$)</td>
<td>TraPPE</td>
<td>200</td>
<td>726 ± 2</td>
<td>0.224 ± 0.003</td>
<td>1.41 ± 0.04</td>
<td>0.235 ± 0.004</td>
</tr>
<tr>
<td>$n$-Tetracosane (C$_{24}$)</td>
<td>NERD</td>
<td>200</td>
<td>813 ± 2</td>
<td>0.213 ± 0.003</td>
<td>0.96 ± 0.03</td>
<td>0.225 ± 0.003</td>
</tr>
<tr>
<td>$n$-Tetracosane (C$_{24}$)</td>
<td>TraPPE</td>
<td>200</td>
<td>817 ± 2</td>
<td>0.213 ± 0.002</td>
<td>0.95 ± 0.02</td>
<td>0.222 ± 0.002</td>
</tr>
<tr>
<td>$n$-Tetracosane (C$_{24}$)</td>
<td>Exp-6</td>
<td>200</td>
<td>813 ± 3</td>
<td>0.207 ± 0.003</td>
<td>0.89 ± 0.02</td>
<td>0.217 ± 0.003</td>
</tr>
<tr>
<td>$n$-Tetracosane (C$_{24}$)</td>
<td>Exp-6</td>
<td>1600</td>
<td>815 ± 2</td>
<td>0.206 ± 0.002</td>
<td>0.89 ± 0.02</td>
<td>0.218 ± 0.002</td>
</tr>
<tr>
<td>$n$-Hexatriacontane (C$_{36}$)</td>
<td>NERD</td>
<td>200</td>
<td>887 ± 2</td>
<td>0.199 ± 0.002</td>
<td>0.60 ± 0.01</td>
<td>0.206 ± 0.003</td>
</tr>
<tr>
<td>$n$-Hexatriacontane (C$_{36}$)</td>
<td>NERD</td>
<td>1600</td>
<td>890 ± 2</td>
<td>0.196 ± 0.003</td>
<td>0.59 ± 0.02</td>
<td>0.205 ± 0.004</td>
</tr>
<tr>
<td>$n$-Hexatriacontane (C$_{36}$)</td>
<td>TraPPE</td>
<td>200</td>
<td>901 ± 1</td>
<td>0.200 ± 0.002</td>
<td>0.61 ± 0.01</td>
<td>0.208 ± 0.002</td>
</tr>
<tr>
<td>$n$-Hexatriacontane (C$_{36}$)</td>
<td>TraPPE</td>
<td>1600</td>
<td>902 ± 2</td>
<td>0.197 ± 0.002</td>
<td>0.60 ± 0.02</td>
<td>0.205 ± 0.003</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>Exp-6</td>
<td>200</td>
<td>946 ± 5</td>
<td>0.189 ± 0.004</td>
<td>0.44 ± 0.02</td>
<td>0.202 ± 0.005</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>Exp-6</td>
<td>1600</td>
<td>948 ± 7</td>
<td>0.180 ± 0.006</td>
<td>0.42 ± 0.03</td>
<td>0.198 ± 0.008</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>NERD</td>
<td>200</td>
<td>945 ± 3</td>
<td>0.185 ± 0.005</td>
<td>0.43 ± 0.02</td>
<td>0.197 ± 0.004</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>NERD</td>
<td>1600</td>
<td>947 ± 5</td>
<td>0.186 ± 0.006</td>
<td>0.42 ± 0.05</td>
<td>0.198 ± 0.009</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>TraPPE</td>
<td>200</td>
<td>959 ± 3</td>
<td>0.186 ± 0.004</td>
<td>0.43 ± 0.02</td>
<td>0.196 ± 0.003</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>TraPPE</td>
<td>1600</td>
<td>961 ± 4</td>
<td>0.187 ± 0.005</td>
<td>0.45 ± 0.02</td>
<td>0.200 ± 0.004</td>
</tr>
</tbody>
</table>
APPENDIX D. DATA USED TO PARAMETERIZE MESS-UP FORCE FIELD

Table D.1: The experimental $\rho_l$ data values for ethane and $n$-octane that were used to calculate $RMS$ when developing the Mess-UP force field.

<table>
<thead>
<tr>
<th>Temperature (K)</th>
<th>Ethane $\rho_l$ (kg/L)</th>
<th>Temperature (K)</th>
<th>$n$-Octane $\rho_l$ (kg/L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>185.00</td>
<td>0.5433</td>
<td>393.15</td>
<td>0.6170</td>
</tr>
<tr>
<td>195.00</td>
<td>0.5305</td>
<td>398.15</td>
<td>0.6120</td>
</tr>
<tr>
<td>200.00</td>
<td>0.5240</td>
<td>403.15</td>
<td>0.6080</td>
</tr>
<tr>
<td>210.00</td>
<td>0.5105</td>
<td>413.15</td>
<td>0.5990</td>
</tr>
<tr>
<td>220.00</td>
<td>0.4963</td>
<td>423.15</td>
<td>0.5875</td>
</tr>
<tr>
<td>230.00</td>
<td>0.4813</td>
<td>433.15</td>
<td>0.5770</td>
</tr>
<tr>
<td>240.00</td>
<td>0.4653</td>
<td>448.15</td>
<td>0.5609</td>
</tr>
<tr>
<td>250.00</td>
<td>0.4481</td>
<td>473.15</td>
<td>0.5310</td>
</tr>
<tr>
<td>260.00</td>
<td>0.4291</td>
<td>498.15</td>
<td>0.4978</td>
</tr>
<tr>
<td>265.00</td>
<td>0.4188</td>
<td>513.15</td>
<td>0.4730</td>
</tr>
<tr>
<td>270.00</td>
<td>0.4077</td>
<td>523.15</td>
<td>0.4554</td>
</tr>
<tr>
<td>275.00</td>
<td>0.3958</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## APPENDIX E. MONTE CARLO SAMPLING PARAMETER SETS

Table E.1: MCS parameter sets for $\varepsilon_{CH_3}$, $\sigma_{CH_3}$, $\varepsilon_{CH_2}$, and $\sigma_{CH_2}$ using Type A and $A_B$ analysis methods.

<table>
<thead>
<tr>
<th>Type A</th>
<th>Type $A_B$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{CH_3}$ (K)</td>
<td>$\sigma_{CH_3}$ (Å)</td>
</tr>
<tr>
<td>98.54353</td>
<td>3.749086</td>
</tr>
<tr>
<td>98.51707</td>
<td>3.749428</td>
</tr>
<tr>
<td>98.55345</td>
<td>3.748579</td>
</tr>
<tr>
<td>98.53228</td>
<td>3.749714</td>
</tr>
<tr>
<td>98.50517</td>
<td>3.749329</td>
</tr>
<tr>
<td>98.43242</td>
<td>3.747863</td>
</tr>
<tr>
<td>98.47475</td>
<td>3.748689</td>
</tr>
<tr>
<td>98.44896</td>
<td>3.748436</td>
</tr>
<tr>
<td>98.46549</td>
<td>3.748877</td>
</tr>
<tr>
<td>98.52699</td>
<td>3.749747</td>
</tr>
<tr>
<td>98.50715</td>
<td>3.749858</td>
</tr>
<tr>
<td>98.52038</td>
<td>3.749637</td>
</tr>
<tr>
<td>98.53956</td>
<td>3.749770</td>
</tr>
<tr>
<td>98.51244</td>
<td>3.749296</td>
</tr>
<tr>
<td>98.46946</td>
<td>3.748226</td>
</tr>
<tr>
<td>98.48930</td>
<td>3.749108</td>
</tr>
<tr>
<td>98.46086</td>
<td>3.748800</td>
</tr>
<tr>
<td>98.42118</td>
<td>3.748149</td>
</tr>
<tr>
<td>98.46483</td>
<td>3.748733</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Type A</th>
<th>Type A_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{\text{CH}_3}$ (K)</td>
<td>$\sigma_{\text{CH}_3}$ (Å)</td>
</tr>
<tr>
<td>98.44896</td>
<td>3.748634</td>
</tr>
<tr>
<td>98.50385</td>
<td>3.749461</td>
</tr>
<tr>
<td>98.48731</td>
<td>3.749362</td>
</tr>
<tr>
<td>98.52633</td>
<td>3.749736</td>
</tr>
<tr>
<td>98.39142</td>
<td>3.747819</td>
</tr>
<tr>
<td>98.45557</td>
<td>3.747697</td>
</tr>
<tr>
<td>98.42978</td>
<td>3.747764</td>
</tr>
<tr>
<td>98.51178</td>
<td>3.749505</td>
</tr>
<tr>
<td>98.48930</td>
<td>3.749108</td>
</tr>
<tr>
<td>98.45094</td>
<td>3.748315</td>
</tr>
<tr>
<td>98.52832</td>
<td>3.749714</td>
</tr>
<tr>
<td>98.55808</td>
<td>3.749880</td>
</tr>
<tr>
<td>98.46814</td>
<td>3.748546</td>
</tr>
<tr>
<td>98.51377</td>
<td>3.749351</td>
</tr>
<tr>
<td>98.50914</td>
<td>3.749185</td>
</tr>
<tr>
<td>98.49790</td>
<td>3.749141</td>
</tr>
<tr>
<td>98.53625</td>
<td>3.749395</td>
</tr>
<tr>
<td>98.52766</td>
<td>3.749031</td>
</tr>
<tr>
<td>98.49327</td>
<td>3.749207</td>
</tr>
<tr>
<td>98.50782</td>
<td>3.748921</td>
</tr>
<tr>
<td>98.59114</td>
<td>3.751048</td>
</tr>
<tr>
<td>98.50385</td>
<td>3.749053</td>
</tr>
<tr>
<td>98.47607</td>
<td>3.748315</td>
</tr>
<tr>
<td>98.38349</td>
<td>3.747929</td>
</tr>
<tr>
<td>98.47872</td>
<td>3.748612</td>
</tr>
<tr>
<td>98.38481</td>
<td>3.747091</td>
</tr>
</tbody>
</table>

Continued on next page
Table E.2 – continued from previous page

<table>
<thead>
<tr>
<th>Type A</th>
<th>Type A_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{\text{CH}_3}$ (K)</td>
<td>$\varepsilon_{\text{CH}_2}$ (K)</td>
</tr>
<tr>
<td>98.49657</td>
<td>3.748700</td>
</tr>
<tr>
<td>98.49327</td>
<td>3.748888</td>
</tr>
<tr>
<td>98.54551</td>
<td>3.749560</td>
</tr>
<tr>
<td>98.51840</td>
<td>3.750111</td>
</tr>
<tr>
<td>98.48930</td>
<td>3.749395</td>
</tr>
<tr>
<td>98.49657</td>
<td>3.749494</td>
</tr>
<tr>
<td>98.57725</td>
<td>3.750497</td>
</tr>
<tr>
<td>98.54287</td>
<td>3.749384</td>
</tr>
<tr>
<td>98.54617</td>
<td>3.749736</td>
</tr>
<tr>
<td>98.45557</td>
<td>3.748579</td>
</tr>
<tr>
<td>98.51178</td>
<td>3.749009</td>
</tr>
<tr>
<td>98.41655</td>
<td>3.748293</td>
</tr>
<tr>
<td>98.41259</td>
<td>3.747455</td>
</tr>
<tr>
<td>98.41259</td>
<td>3.747764</td>
</tr>
<tr>
<td>98.53890</td>
<td>3.749968</td>
</tr>
<tr>
<td>98.52832</td>
<td>3.749604</td>
</tr>
<tr>
<td>98.56932</td>
<td>3.750585</td>
</tr>
<tr>
<td>98.52964</td>
<td>3.749803</td>
</tr>
<tr>
<td>98.49790</td>
<td>3.749318</td>
</tr>
<tr>
<td>98.42780</td>
<td>3.748149</td>
</tr>
<tr>
<td>98.46483</td>
<td>3.748502</td>
</tr>
<tr>
<td>98.46880</td>
<td>3.748722</td>
</tr>
<tr>
<td>98.51112</td>
<td>3.749329</td>
</tr>
<tr>
<td>98.54551</td>
<td>3.750574</td>
</tr>
<tr>
<td>98.40267</td>
<td>3.748028</td>
</tr>
<tr>
<td>98.44565</td>
<td>3.748513</td>
</tr>
</tbody>
</table>

Continued on next page
Table E.2 – continued from previous page

<table>
<thead>
<tr>
<th>Type A</th>
<th>Type A_B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{\text{CH}_3}$ (K)</td>
<td>$\sigma_{\text{CH}_3}$ (Å)</td>
</tr>
<tr>
<td>98.54816</td>
<td>3.749615</td>
</tr>
<tr>
<td>98.50253</td>
<td>3.748811</td>
</tr>
<tr>
<td>98.51906</td>
<td>3.749902</td>
</tr>
<tr>
<td>98.56006</td>
<td>3.749770</td>
</tr>
<tr>
<td>98.42317</td>
<td>3.748248</td>
</tr>
<tr>
<td>98.52567</td>
<td>3.748711</td>
</tr>
<tr>
<td>98.44962</td>
<td>3.748160</td>
</tr>
<tr>
<td>98.50253</td>
<td>3.749141</td>
</tr>
<tr>
<td>98.45491</td>
<td>3.747841</td>
</tr>
<tr>
<td>98.53559</td>
<td>3.749207</td>
</tr>
<tr>
<td>98.53824</td>
<td>3.750321</td>
</tr>
<tr>
<td>98.45359</td>
<td>3.748623</td>
</tr>
<tr>
<td>98.54154</td>
<td>3.749637</td>
</tr>
<tr>
<td>98.42780</td>
<td>3.748469</td>
</tr>
<tr>
<td>98.45822</td>
<td>3.748469</td>
</tr>
<tr>
<td>98.49261</td>
<td>3.749130</td>
</tr>
<tr>
<td>98.46814</td>
<td>3.749163</td>
</tr>
<tr>
<td>98.54287</td>
<td>3.750078</td>
</tr>
<tr>
<td>98.44896</td>
<td>3.749097</td>
</tr>
<tr>
<td>98.52236</td>
<td>3.749395</td>
</tr>
<tr>
<td>98.42846</td>
<td>3.748260</td>
</tr>
<tr>
<td>98.52567</td>
<td>3.748822</td>
</tr>
<tr>
<td>98.45689</td>
<td>3.748248</td>
</tr>
<tr>
<td>98.48599</td>
<td>3.749196</td>
</tr>
<tr>
<td>98.53625</td>
<td>3.750232</td>
</tr>
<tr>
<td>98.51972</td>
<td>3.748921</td>
</tr>
</tbody>
</table>

Continued on next page
<table>
<thead>
<tr>
<th>Type A</th>
<th>Type Aβ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\varepsilon_{\text{CH}_3} \ (\text{K})$</td>
<td>$\varepsilon_{\text{CH}_3} \ (\text{K})$</td>
</tr>
<tr>
<td>$\sigma_{\text{CH}_3} \ (\text{Å})$</td>
<td>$\sigma_{\text{CH}_3} \ (\text{Å})$</td>
</tr>
<tr>
<td>$\varepsilon_{\text{CH}_2} \ (\text{K})$</td>
<td>$\varepsilon_{\text{CH}_2} \ (\text{K})$</td>
</tr>
<tr>
<td>$\sigma_{\text{CH}_2} \ (\text{Å})$</td>
<td>$\sigma_{\text{CH}_2} \ (\text{Å})$</td>
</tr>
<tr>
<td>98.45028</td>
<td>3.748050</td>
</tr>
<tr>
<td>98.53427</td>
<td>3.749042</td>
</tr>
<tr>
<td>98.50253</td>
<td>3.748744</td>
</tr>
<tr>
<td>45.39429</td>
<td>3.971887</td>
</tr>
<tr>
<td>45.38877</td>
<td>3.971571</td>
</tr>
<tr>
<td>45.37524</td>
<td>3.968774</td>
</tr>
<tr>
<td>3.748744</td>
<td>3.971887</td>
</tr>
<tr>
<td>3.749042</td>
<td>3.971571</td>
</tr>
<tr>
<td>3.748050</td>
<td>3.971887</td>
</tr>
<tr>
<td>3.757521</td>
<td>3.971703</td>
</tr>
<tr>
<td>3.739528</td>
<td>3.970672</td>
</tr>
<tr>
<td>3.752064</td>
<td>3.965649</td>
</tr>
<tr>
<td>45.32800</td>
<td>3.971703</td>
</tr>
<tr>
<td>45.34229</td>
<td>3.970672</td>
</tr>
<tr>
<td>45.25408</td>
<td>3.965649</td>
</tr>
</tbody>
</table>
APPENDIX F. PROPAGATION OF ERRORS HISTOGRAMS

In this section we present the histograms that were used for determining the uncertainties in \( \rho_l, T_c, \rho_c, P_c, \) and \( Z_c \) in Chapter 5. Recall that these histograms were obtained by simulating 100 MCS parameter sets. The bin count is divided by the total number of counts (100, i.e. the number of MCS parameter sets) to provide a bin probability. The normal distribution fits are only included for visual purposes. To clarify, the vertical axis is not applicable to the normal distribution fits since these should be probability densities (which have a different scale and inverse units).

Figures F.1-F.6 contain the Type A and A\(_B\) \( \rho_l \) histograms for ethane, \( n \)-octane, \( C_{16} \), \( C_{24} \), \( C_{36} \), and \( C_{48} \), respectively. Figures F.7-F.12 provide the Type A and A\(_B\) histograms for \( T_c, \rho_c, P_c, \) and \( Z_c \) of ethane, \( n \)-octane, \( C_{16} \), \( C_{24} \), \( C_{36} \), and \( C_{48} \), respectively. Figures F.13-F.18 depict the Type B results as probability densities representing the numerical uncertainties in \( T_c, \rho_c, P_c, \) and \( Z_c \) of ethane, \( n \)-octane, \( C_{16} \), \( C_{24} \), \( C_{36} \), and \( C_{48} \), respectively for each of the extrema parameter sets listed in Table 5.3. In Figure F.19 we validate that the uncertainty in the CH\(_3 \) parameters has a negligible impact on the results for \( C_{16} \). Finally, in Figure F.20 we present the extrema parameter sets as a reference. We have used the same color scheme in this figure as that found in Figures F.13-F.19 to facilitate comparing the results for the different extrema parameter sets.

In Figure F.1 we see that the Type A\(_B\) uncertainties in \( \rho_l \) are always larger than those from the Type A analysis. However, the difference appears to be largest at lower temperatures. This is because numerical uncertainty is more significant at higher temperatures. Therefore, since these histograms include both parameter and numerical uncertainties, the Type A and A\(_B\) results become similar at high temperatures. By contrast, at low temperatures the difference between the Type A and A\(_B\) histograms is almost entirely due to the parameter uncertainties since numerical uncertainty is negligible. This phenomenon is more readily observed in Figure F.2. Notice that in Panel (d) the Type A and A\(_B\) histograms are nearly identical at 515 K while in Panel (a) they are much different at 390 K. This is because we have used fewer particles for \( n \)-octane than ethane so the numerical...
uncertainty is even more significant. We do not observe this behavior in Figures F.3-F.6 because the highest temperature simulated was at a reduced temperature of 0.85, which is not close enough to the critical point to result in large numerical uncertainties.

In Figures F.7-F.12, notice that for $T_c$ and $\rho_c$ the Type A and Type $A_B$ regions have a much different range. For this reason, we have used different bin sizes to cover the whole region with a feasible number of bins (between 20-25). Therefore, when comparing probabilities you must consider that a single bin with the Type $A_B$ analysis encompasses multiple bins for the Type A analysis. The situation is different for $P_c$ and $Z_c$, where the ranges are similar. For these properties we have used the same bin sizes. It is interesting that these properties are nearly irrespective of the uncertainty approach. In some cases, due to the relatively small sample size, the $P_c$ or $Z_c$ Type $A_B$ uncertainty is actually smaller than the Type A uncertainty. Most likely this is because the numerical uncertainty is the primary contributor and, therefore, the overall uncertainties will be practically identical. In Figures F.14-F.18 we see that indeed the numerical uncertainty can be quite large for $\rho_c$, $P_c$, and $Z_c$ of larger $n$-alkanes.

Figures F.13-F.18 are provided to demonstrate how the Type B uncertainties were determined for ethane, $n$-octane, $C_{16}$, $C_{24}$, $C_{36}$, and $C_{48}$, respectively. The distributions in these figures represent the numerical uncertainty for each of the eight extrema parameter sets simulated. Panels (a)-(d) contain the uncertainties in $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. As seen in Panel (c) the numerical uncertainty in $P_c$ is large enough that there is considerable overlap between the different extrema. In addition, in Panel (d) the numerical uncertainty in $Z_c$ is more than the Type B parameter uncertainty.

Some useful insight is obtained by comparing Figures F.13-F.18 with Figure F.20. For example, notice that $P_1$, $P_2$, and $P_5$ (the left side of the acceptance region in Figure F.20 Panel (a)) correspond to the lower $T_c$ values in Figure F.13 Panel (a). By contrast, $P_3$, $P_4$, and $P_6$ (the right side of the acceptance region in Figure F.20 Panel (a)) correspond to the higher $T_c$ values in Figure F.13 Panel (a). Similar observations are found for the other critical constants for both the $CH_3$ and $CH_2$ parameter extrema.
Figure F.1: Comparison of the $\rho_l$ histograms obtained using 100 MCS parameter sets from a Type A and $A_B$ analysis for ethane. Panels (a)-(f) correspond to different temperatures. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.2: Comparison of the $\rho_l$ histograms obtained using 100 MCS parameter sets from a Type A and $A_B$ analysis for $n$-octane. Panels (a)-(e) correspond to 390, 440, 490, 515, and 540 K, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.3: Comparison of the $\rho_l$ histograms obtained using 100 MCS parameter sets from a Type A and A$B$ analysis for C$_{16}$. Panels (a)-(b) correspond to 510 and 625 K, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.

Figure F.4: Comparison of the $\rho_l$ histograms obtained using 100 MCS parameter sets from a Type A and A$B$ analysis for C$_{24}$. Panels (a)-(b) correspond to 600 and 700 K, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.5: Comparison of the $\rho_l$ histograms obtained using 100 MCS parameter sets from a Type A and A$_B$ analysis for C$_{36}$. Panels (a)-(b) correspond to 650 and 775 K, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.

Figure F.6: Comparison of the $\rho_l$ histograms obtained using 100 MCS parameter sets from a Type A and A$_B$ analysis for C$_{48}$. Panels (a)-(b) correspond to 730 and 830 K, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.7: Comparison of the histograms obtained using 100 MCS parameter sets from a Type A and $A_B$ analysis for ethane. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.8: Comparison of the histograms obtained using 100 MCS parameter sets from a Type A and A$_B$ analysis for $n$-octane. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.9: Comparison of the histograms obtained using 100 MCS parameter sets from a Type A and A\textsubscript{B} analysis for C\textsubscript{16}. Panels (a)-(d) correspond to $T_c$, $p_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.10: Comparison of the histograms obtained using 100 MCS parameter sets from a Type A and A_B analysis for C_{24}. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.11: Comparison of the histograms obtained using 100 MCS parameter sets from a Type A and A_B analysis for C_{36}. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.12: Comparison of the histograms obtained using 100 MCS parameter sets from a Type A and A\textsubscript{B} analysis for C\textsubscript{48}. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distribution fits are only included for visual purposes. The bin count is divided by the total number of counts (100) to provide a bin probability.
Figure F.13: Comparison of the eight extrema parameter sets from a Type B analysis of ethane. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.14: Comparison of the eight extrema parameter sets from a Type B analysis of $n$-octane. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.15: Comparison of the eight extrema parameter sets from a Type B analysis of $C_{16}$. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.16: Comparison of the eight extrema parameter sets from a Type B analysis of $C_{24}$. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.17: Comparison of the eight extrema parameter sets from a Type B analysis of C\textsubscript{36}. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.18: Comparison of the eight extrema parameter sets from a Type B analysis of C$_{48}$. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.19: Comparison of the eight CH$_3$ extrema parameter sets from a Type B analysis of C$_{16}$. The CH$_2$ parameters are the optimal set reported in Table 5.2. Panels (a)-(d) correspond to $T_c$, $\rho_c$, $P_c$, and $Z_c$, respectively. The normal distributions represent the numerical uncertainties for a given parameter set.
Figure F.20: Extrema parameter sets used in Type B analysis when both \( \rho_I \) and \( T_c \) uncertainties are considered. Panels (a) and (b) correspond to the CH\(_3\) and CH\(_2\) LJ parameters, respectively.
Several studies have demonstrated that the CH$_3$ and CH$_2$ parameters are highly correlated, particularly when they are regressed simultaneously. In the case of a sequential optimization, any uncertainty in the CH$_3$ parameters is inherited by the CH$_2$ parameters. This correlation can be accounted for by developing expressions that relate the deviations in CH$_3$ parameters to the CH$_2$ parameters. For example, we can assume that the Lorentz-Berthelot (LB) cross interactions are constant. Therefore, any deviation in the CH$_3$ parameters will shift the CH$_2$ contours so that the new optimal CH$_2$ parameters fulfill the following criteria

\[
\hat{\sigma}_{CH_3} + \hat{\sigma}_{CH_2} = \sigma_{CH_3}^{MCS} + \sigma_{CH_2}^0 \quad (G.1)
\]

\[
\hat{\varepsilon}_{CH_3} \hat{\varepsilon}_{CH_2} = \varepsilon_{CH_3}^{MCS} \varepsilon_{CH_2}^0 \quad (G.2)
\]

where \(\hat{\sigma}_{CH_3}, \hat{\sigma}_{CH_2}, \hat{\varepsilon}_{CH_3}, \) and \(\hat{\varepsilon}_{CH_2}\) are the optimal LJ parameters, \(\sigma_{CH_3}^{MCS}\) and \(\varepsilon_{CH_3}^{MCS}\) are the CH$_3$ parameters obtained from the Monte Carlo Sampling algorithm and \(\sigma_{CH_2}^0\) and \(\varepsilon_{CH_2}^0\) are the optimal CH$_2$ parameters for the MCS CH$_3$ parameters. Finally, the MCS CH$_2$ parameters are obtained from

\[
\sigma_{CH_2}^* = \sigma_{CH_2}^{MCS} \frac{\sigma_{CH_2}^0}{\hat{\sigma}_{CH_2}} \quad (G.3)
\]

\[
\varepsilon_{CH_2}^* = \varepsilon_{CH_2}^{MCS} \frac{\varepsilon_{CH_2}^0}{\hat{\varepsilon}_{CH_2}} \quad (G.4)
\]

where \(\sigma_{CH_2}^{MCS}\) and \(\varepsilon_{CH_2}^{MCS}\) are the CH$_2$ parameters obtained from the Monte Carlo Sampling algorithm and \(\sigma_{CH_2}^*\) and \(\varepsilon_{CH_2}^*\) are the MCS values after correcting for the correlation between CH$_3$ and CH$_2$ parameters. These equations assume that the ratio between the optimal and MCS parameter sets should be equal regardless of correlation. In Figures G.1-G.2 we demonstrate how MCS CH$_2$ parameter sets are affected by this correlation. Specifically, we sampled 50000 CH$_3$ parameter sets and implemented Equations G.1-G.4 to modify the 50000 sampled CH$_2$ parameter sets. In
Figure G.1 we have plotted the uncorrelated and correlated CH₂ parameter sets. Panels (a) and (b) utilize a Type A and A₇ approach, respectively. Only 1000 parameter sets are included in Figure G.1 for clarity. In Figure G.2 we present histograms of the 50000 MCS parameter sets with and without CH₃ correlation. Panels (a)-(b) represent the Type A analysis while Panels (c)-(d) are for the Type A₇ analysis. Notice that accounting for correlation has a much larger affect on the Type A₇ analysis. This is expected as the Type A₇ analysis results in much larger uncertainties in the CH₃ parameters. Also, clearly using the LB combining rules causes the ε uncertainty to widen significantly while having a small effect on the σ uncertainty.

Figure G.1: Comparison between the uncorrelated and correlated MCS parameter sets for εCH₂ and σCH₂. Panels (a)-(b) correspond to the Type A and Type A₇ analysis, respectively. Only 1000 parameter sets are presented for clarity.
Figure G.2: Comparison between the uncorrelated and correlated histograms for $\varepsilon_{\text{CH}_2}$ and $\sigma_{\text{CH}_2}$. Panels (a)-(b) correspond to the Type A analysis for both the CH$_3$ and CH$_2$ parameters while Panels (c)-(d) use the Type A$_B$ analysis. The normal distribution fits to the histograms are also included. Probability density is defined as the number of counts in a single bin divided by both the total number of counts and the bin width.
APPENDIX H.  RECOMMENDED SIMULATION VALUES

Table H.1 contains the recommended simulation values used in Chapter 6. The best estimate values are a weighted average of the results reported in Chapters 4-5 (weighted by assumed reliability, i.e. TraPPE is weighted more for smaller compounds, NERD is weighted more for larger compounds, Exp-6 is weighted more for \( \rho_c \), and 1600 molecule simulations are weighted more than 200 molecular simulations). Uncertainties are only provided for the compounds simulated in Chapter 5 where the largest uncertainty (Type A or Type B) is reported at the 95% confidence level.

Table H.1: The recommended \( T_c \), \( \rho_c \), \( P_c \), and \( Z_c \) values with their corresponding uncertainties. Uncertainties are only reported for compounds where the parameter uncertainty was included (i.e. those simulated in Chapter 5).

<table>
<thead>
<tr>
<th>Compound</th>
<th>( T_c ) (K)</th>
<th>( \rho_c ) (kg/L)</th>
<th>( P_c ) (MPa)</th>
<th>( Z_c )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n )-Hexane (C(_6))</td>
<td>505 ± N/A</td>
<td>0.229 ± N/A</td>
<td>3.10 ± N/A</td>
<td>0.271 ± N/A</td>
</tr>
<tr>
<td>( n )-Octane (C(_8))</td>
<td>571 ± 5</td>
<td>0.230 ± 0.004</td>
<td>2.52 ± 0.06</td>
<td>0.262 ± 0.003</td>
</tr>
<tr>
<td>( n )-Decane (C(_{10}))</td>
<td>624 ± N/A</td>
<td>0.230 ± N/A</td>
<td>2.15 ± N/A</td>
<td>0.255 ± N/A</td>
</tr>
<tr>
<td>( n )-Dodecane (C(_{12}))</td>
<td>663 ± N/A</td>
<td>0.227 ± N/A</td>
<td>1.81 ± N/A</td>
<td>0.247 ± N/A</td>
</tr>
<tr>
<td>( n )-Hexadecane (C(_{16}))</td>
<td>725 ± 7</td>
<td>0.221 ± 0.004</td>
<td>1.38 ± 0.04</td>
<td>0.234 ± 0.003</td>
</tr>
<tr>
<td>( n )-Tetracosane (C(_{24}))</td>
<td>813 ± 9</td>
<td>0.207 ± 0.006</td>
<td>0.89 ± 0.04</td>
<td>0.217 ± 0.005</td>
</tr>
<tr>
<td>( n )-Hexatriacontane (C(_{36}))</td>
<td>889 ± 10</td>
<td>0.198 ± 0.006</td>
<td>0.60 ± 0.03</td>
<td>0.206 ± 0.005</td>
</tr>
<tr>
<td>( n )-Octatetracontane (C(_{48}))</td>
<td>947 ± 13</td>
<td>0.186 ± 0.007</td>
<td>0.44 ± 0.03</td>
<td>0.200 ± 0.008</td>
</tr>
</tbody>
</table>
APPENDIX I. COARSE-GRAINED MODEL RESULTS

In this section we present the simulation results we obtained using a coarse-grained (CG) model, namely, the TEAM-CG force field [109]. In Table I.1 we provide the simulation details for each system simulated with TEAM-CG. Specifically, we include the compound, temperatures, number of replicates, and number of Monte Carlo moves used in both equilibration and production. In each case we used a system size of 200 molecules and assumed that finite-size effects are negligible (see Chapter 4). As recommended by Cao et al., we employed a larger cutoff distance of 20 Å rather than 14 Å due to the greater volume of a CG site compared to a UA site.

Table I.1: Summary of compounds, temperatures ($T_i$), number of replicates at each temperature ($n$), and the number of Monte Carlo cycles in the equilibration ($MC_E$) and production ($MC_P$) periods.

<table>
<thead>
<tr>
<th>Compound</th>
<th>$T_1$ (K)</th>
<th>$T_2$ (K)</th>
<th>$T_3$ (K)</th>
<th>$T_4$ (K)</th>
<th>$T_5$ (K)</th>
<th>$n$</th>
<th>$MC_E$</th>
<th>$MC_P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$-Pentadecane (C$_{15}$)</td>
<td>500</td>
<td>540</td>
<td>575</td>
<td>610</td>
<td>645</td>
<td>1</td>
<td>50,000</td>
<td>50,000</td>
</tr>
<tr>
<td>$n$-Octadecane (C$_{18}$)</td>
<td>525</td>
<td>560</td>
<td>600</td>
<td>635</td>
<td>670</td>
<td>1</td>
<td>50,000</td>
<td>50,000</td>
</tr>
<tr>
<td>$n$-Tetracosane (C$_{24}$)</td>
<td>570</td>
<td>610</td>
<td>650</td>
<td>690</td>
<td>730</td>
<td>1</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>$n$-Hexatriacontane (C$_{36}$)</td>
<td>620</td>
<td>665</td>
<td>710</td>
<td>755</td>
<td>800</td>
<td>1</td>
<td>100,000</td>
<td>100,000</td>
</tr>
<tr>
<td>$n$-Octatetracontane (C$_{48}$)</td>
<td>665</td>
<td>710</td>
<td>755</td>
<td>800</td>
<td>850</td>
<td>5</td>
<td>100,000</td>
<td>300,000</td>
</tr>
<tr>
<td>$n$-Hexacontane (C$_{60}$)</td>
<td>665</td>
<td>710</td>
<td>755</td>
<td>800</td>
<td>850</td>
<td>5</td>
<td>100,000</td>
<td>700,000</td>
</tr>
<tr>
<td>$n$-Pentaheptacontane (C$_{75}$)</td>
<td>690</td>
<td>740</td>
<td>785</td>
<td>835</td>
<td>885</td>
<td>5</td>
<td>100,000</td>
<td>500,000</td>
</tr>
<tr>
<td>$n$-Nonacontane (C$_{90}$)</td>
<td>705</td>
<td>755</td>
<td>805</td>
<td>855</td>
<td>905</td>
<td>5</td>
<td>100,000</td>
<td>300,000</td>
</tr>
<tr>
<td>$n$-Pentahectane (C$_{105}$)</td>
<td>N/A</td>
<td>755</td>
<td>805</td>
<td>855</td>
<td>905</td>
<td>5</td>
<td>100,000</td>
<td>200,000</td>
</tr>
<tr>
<td>$n$-Icosahectane (C$_{120}$)</td>
<td>N/A</td>
<td>765</td>
<td>815</td>
<td>865</td>
<td>915</td>
<td>5</td>
<td>100,000</td>
<td>150,000</td>
</tr>
</tbody>
</table>
As mentioned in Chapter 7, the probability of a successful insertion MC move into the liquid phase becomes fleetingly small for large compounds at high densities (i.e., lower temperatures). For this reason, we were not able to successfully equilibrate \( \text{C}_{105} \) and \( \text{C}_{120} \) at \( T_r \approx 0.7 \). Also, note that we did not use the D-optimal design since it was unknown \textit{a priori} if the TEAM-CG model would perform well over the entire temperature range. Specifically, according to the values reported by Cao et al., the TEAM-CG model demonstrates some abnormal fluctuations in \( \rho_l \) with respect to temperature. We attribute this to the temperature dependent LJ parameters used in the TEAM-CG model.

In Figure I.1 we present the vapor-liquid coexistence curves for each of the compounds we simulated with the TEAM-CG force field. We have also included a fit to the simulation results and the predicted critical point \((T_c, \rho_c)\). Although not included in Figure I.1, it is worth mentioning that we verified that our simulation results agreed with those of Cao et al. for \( \text{C}_{12} \).

In Figure I.1 we compare the experimental data and the proposed prediction models for the critical constants with the simulation results for the TEAM-CG force field. The error bars included
in Figure I.2 are the 95% confidence intervals (i.e. only the numerical uncertainties). The values and uncertainties presented for C$_5$-C$_{12}$ were obtained by analyzing the simulation results reported by Cao et al. with the rigorous non-linear analysis explained in Chapters 2-4 (as Cao et al. did not report uncertainties).

Figure I.2: Comparison of critical constant trends with respect to carbon number for experimental data, new prediction models, and the simulation results using the TEAM-CG model. Panels (a)-(d) correspond to $T_c$, $P_c$, $\rho_c$, and $Z_c$, respectively. Carbon number is plotted on a log base 10 scale. The error bars represent the 95% confidence interval.
As seen in Figure I.2, the TEAM-CG model agrees well with the $\rho_c$ and $P_c$ prediction models. The deviation from $Z_c$ at large $CN$ is encompassed by the uncertainty in the prediction model. However, there exists a strong deviation from the $T_c$ prediction model for larger $CN$. The prediction model uncertainty and the numerical uncertainty in the simulation results are not large enough to justify this discrepancy. Instead, we believe that this is a result of inadequacies in the force field itself. Specifically, the TEAM-CG model uses a temperature dependent Lennard-Jones potential which was parameterized for $n$-alkanes up to $C_{12}$. It appears that this model performs poorly when extrapolating to the much higher temperatures required for simulating larger compounds. Specifically, as seen in Figure I.3 the TEAM-CG force field suffers from a systematic deviation of $\rho_l$ for compounds as small as $C_{24}$.

Figure I.3: Comparison of $\rho_l$ DIPPR correlations, Mess-UP (both the optimal and the Type B extrema), and TEAM-CG results for $C_{24}$, $C_{36}$, and $C_{48}$. The 95% confidence intervals are smaller than one symbol size (only included for TEAM-CG $C_{48}$).

In Figure I.3 we have included the simulation results for $C_{24}$, $C_{36}$, and $C_{48}$ using the TEAM-CG and Mess-UP force fields. The uncertainties assigned to the Mess-UP force field are those from the Type B analysis. Specifically, the error bars represent the minimum and maximum values obtained from the Type B extrema parameter sets. Only the Type B extrema results (rather
than the Type A or Type A\textsubscript{B} uncertainties) are presented since these are nearly identical to the values from the TraPPE and NERD force fields. In addition, we have provided the DIPPR correlation for C\textsubscript{24} and C\textsubscript{36}. To help guide the eye we have also included a correlation fit to the TEAM-CG simulation results. Clearly, the TEAM-CG force field disagrees with the DIPPR correlations and the UA based force field (Type B extrema). For example, the TEAM-CG C\textsubscript{48} results are such that they overlap considerably with the C\textsubscript{36} DIPPR correlation and Type B extrema.

The systematic deviation in \( \rho_\text{l} \) causes the vapor-liquid coexistence curve (VLCC) to be more narrow and, therefore, results in a lower \( T_\text{c} \) (recall Figure I.2 Panel (a)). The impact on \( \rho_\text{c} \) is not as dramatic, although this could explain the consistent over-prediction of \( \rho_\text{c} \) since a lower \( T_\text{c} \) results in a higher \( \rho_\text{c} \) (notice the slope of \( \rho_\text{r} \) included for C\textsubscript{15} in Figure I.1). Similarly, the \( P_\text{c} \) estimates do not depreciate considerably since \( P_\text{c} \) is related to the slope of \( \rho_\text{l} \) when obtained with the Vetere approach. However, the propagation of these different effects appears to have a significant impact on the \( Z_\text{c} \) trend. Despite the poor prediction of \( T_\text{c} \), the TEAM-CG results can provide some qualitative insight when compared with our new prediction models. As mentioned in Chapter 7, we recommend improving upon the TEAM-CG model so that it extrapolates more reliably to longer chain-lengths and higher temperatures than those used in the parameterization process.