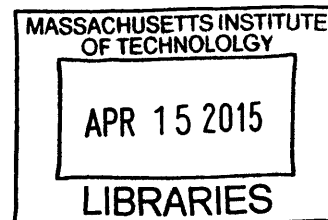


Representing Liquid-Vapor Equilibria of Ternary **ARCHIVES**
Systems Using Neural Networks

by

Mathew M. Swisher



Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Masters of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

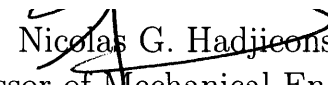
© Massachusetts Institute of Technology 2015. All rights reserved.

Author **Signature redacted**

Department of Mechanical Engineering
January 20, 2015

Signature redacted

Certified by.....


Nicolas G. Hadjiconstantinou
Professor of Mechanical Engineering
Thesis Supervisor

Signature redacted

Accepted by

David E. Hardt
Chairman, Department Committee on Graduate Students

Representing Liquid-Vapor Equilibria of Ternary Systems Using Neural Networks

by

Mathew M. Swisher

Submitted to the Department of Mechanical Engineering
on January 20, 2015, in partial fulfillment of the
requirements for the degree of
Masters of Science in Mechanical Engineering

Abstract

We develop a method based on neural networks for efficiently interpolating equations of state (EOS) for liquid-vapor equilibria of ternary mixtures. We investigate the performance of neural networks both when experimental data are available and when only simulation data are available. Simulation data are obtained from Gibbs Ensemble Monte Carlo simulations, using the TraPPE-EH molecular model. Our investigation uses the mixture of carbon dioxide, methane, and ethane as a validation example, for which experimental data exist. Analysis of the error in a neural-network-generated liquid-vapor coexistence curve shows that the resulting interpolation is robust and accurate, even in the case where the network is trained on a few data points. We use this observation to construct a methodology for accurately locating liquid-vapor equilibria of ternary mixtures without using any experimental data.

Thesis Supervisor: Nicolas G. Hadjiconstantinou
Title: Professor of Mechanical Engineering

Contents

1	Introduction	13
2	Background	17
3	Gibbs Ensemble Monte Carlo	23
3.1	Overview	23
3.2	Implementation	24
3.3	Acceptance Criteria	25
3.3.1	Translation/Rotation	26
3.3.2	Molecule Transfer	26
3.3.3	Volume Transfer for NVT Simulations	27
3.3.4	Volume Change for NPT Simulations	27
3.4	Determining Critical Points	28
3.4.1	Single Component Systems	28
4	Potential and Model	31
4.1	Lennard Jones Potential - Non Bonded Interactions	32
4.2	Combination Rules	33
4.3	Bonded Interactions	33
4.4	Molecular Models	35
4.4.1	TraPPE-UA	35
4.4.2	TraPPE-Explicit Hydrogen	37
4.4.3	TraPPE-SM	39

4.5	Anisotropic United Atom - 4	40
5	GEMC simulation	43
5.1	Pure Component Results	43
5.2	Binary Component Comparison	48
6	Artificial Neural Networks	53
6.1	Problem Statement	54
6.2	Bayesian Regularized Artificial Neural Networks	54
7	Application of Machine Learning to Ternary LVE	57
7.1	Experimental Data	57
7.2	Simulation Data	62
7.3	Training With Few Data Points	65
7.4	Predicting LVE Coexistence Curves Without Experimental Data . . .	69
8	Conclusion	75

List of Figures

2-1	The composition of a ternary mixture can be represented on a ternary plot. In these plots the concentration of each component at a point is given by the three axes. The grid lines corresponding to each axis are the ones departing at 120° measured from the origin. For example point A has a composition of 20% substance 1, 40% substance 2, and 40% substance 3. example of a set of isothermal-isobaric LVE lines is also shown. In this example, the blue line is the isothermal-isobaric coexistence for the vapor phase and the red line is the isothermal-isobaric coexistence line for the liquid phase. The dotted lines are representations of typical experimental data. The solid lines correspond to an approximation of the isothermal-isobaric line from linearly interpolating the binary LVE data points (□).	18
2-2	Two examples of binary LVE isotherms, where blue is the ideal mixing case and red has an azeotrope. Solid lines indicate the fluid phase and dotted lines indicate the vapor phase.	21
4-1	The dihedral angle, or torsion angle, is the angle between normal vectors of the two planes (1,2,3) and (2,3,4). For the example molecule shown, the dihedral angle is -180 degrees.	34

5-1	<p>GEMC simulation results for liquid-vapor equilibrium in pure methane. The liquid-vapor equilibrium curve obtained from experimental results is indicated by the green line. Results from each of TraPPE-UA(\blacktriangleleft), TraPPE-EH(\triangle), AUA-4(\blacktriangleright) are plotted along with the estimated critical point (shown with a filled marker) for comparison. Error bars are smaller than symbol size unless otherwise indicated.</p>	45
5-2	<p>GEMC simulation results for pure ethane. Results from each of TraPPE-UA(\blacktriangleleft), TraPPE-EH(\triangle), AUA-4(\blacktriangleright) are plotted as well as the experimental results (-). The estimated critical point (shown with a filled marker) is shown for comparison. Error bars are smaller than symbol size unless otherwise indicated.</p>	46
5-3	<p>GEMC simulation results for pure propane. Results from each of TraPPE-UA(\blacktriangleleft), TraPPE-EH(\triangle), AUA-4(\blacktriangleright) are plotted as well as the experimental results (-). The estimated critical point (shown with a filled marker) is shown for comparison. Error bars are smaller than symbol size unless otherwise indicated.</p>	47
5-4	<p>GEMC simulation results for liquid-vapor equilibrium for a mixture of CH_4 and C_2H_6 at 180K. Results from each of TraPPE-UA(\blacktriangleleft), TraPPE-EH(\triangle), AUA-4(\blacktriangleright) are plotted along with the experimental results (-). Empty markers and dashed line indicate the vapor phase while the solid markers and solid line indicate the liquid phase. Error bars are smaller than symbol size ($\approx 1\%$).</p>	48
5-5	<p>GEMC simulation results for LVE for a mixture of CO_2 and C_2H_6 at 207K (top) and 213K (bottom). Results from TraPPE-UA(\blacktriangleleft) and TraPPE-EH(\triangle) are plotted along with the experimental results (-). Empty markers and dashed line indicate the vapor phase while the solid markers and solid line indicate the liquid phase. Error bars are smaller than symbol size ($\approx 1\%$).</p>	51

7-1	Ternary liquid-vapor equilibrium results for CO ₂ , CH ₄ , and C ₂ H ₆ at 230K and 1.52MPa (blue) / 3.55MPa (green) / 5.57MPa (red). Experimental results (X) are from Wei et al [16]. Neural network results are denoted by □.	58
7-2	Error Histogram for the error between the experimental data and the interpolated data points, shown in Figure 7-1, as defined by equation (7.1). Blue columns indicate the error from the liquid compositions and the red column indicates the error from the vapor composition. . .	59
7-3	Ternary liquid-vapor equilibrium results for CO ₂ , CH ₄ , and C ₂ H ₆ at 230K and 1.52MPa (blue) / 3.55MPa (green) / 5.57MPa (red). Experimental results (X) are from Wei et al [16]. Neural network results are denoted by □.	62
7-4	Error Histogram for the error between GEMC simulation results and the interpolated data points, shown in Figure 7-3, as defined by equation (7.1). Blue columns indicate the error from the liquid compositions and the red column indicates the error from the vapor composition. . .	63
7-5	An example of the three approximations of the isothermal-isobaric LVE curves being evaluated as reduced initial data sets. The linear approximation is shown in red, the three point approximation in blue, and the four point approximation in fuchsia. For comparison the true isothermal-isobaric curves are shown in black.	65
7-6	Error histogram for the liquid phase of a system composed of CO ₂ , CH ₄ , and C ₂ H ₆ at 230K. The histogram shows the interpolation error distributions resulting from a linear approximation (blue), a three point approximation (green), a four point approximation (yellow), and the error associated with using the full data set (red).	66

7-7	Error histogram for the vapor phase of a system composed of CO ₂ , CH ₄ , and C ₂ H ₆ at 230K. The histogram shows the interpolation error distributions resulting from a linear approximation (blue), a three point approximation (green), a four point approximation (yellow), and the error associated with using the full data set (red).	67
7-8	Diagram of the steps used to calculate LVE using neural networks . .	70
7-9	Ternary liquid-vapor equilibrium results for CO ₂ , CH ₄ , and C ₂ H ₆ at 230K and 1.52MPa (blue) / 2.53MPa (red). Neural network results are indicated by the line with open symbols, while solid symbols indicate simulation results obtained from starting simulations initialized at experimental composition results from Wei et al [16]. Dark shades correspond to the liquid phase and light shades correspond to the vapor phase.	71
7-10	Ternary liquid-vapor equilibrium results for CO ₂ , CH ₄ , and C ₂ H ₆ at 230K. The red and blue markers are the same as in Figure 7-9 and shown for reference. The green line and empty symbols indicate the estimated composition at a temperature of 230K and a pressure of 2.03 MPa. The solid symbols were obtained from GEMC simulations initialized at the interpolated data points. Dark shades correspond to the liquid phase and light shades correspond to the vapor phase. . . .	72

List of Tables

4.1	List of the associated non-bonded interaction parameters for TraPPE-UA	36
4.2	List of parameters used for bonded interactions in TraPPE-UA	36
4.3	List of the associated non-bonded interaction parameters for TraPPE-EH	38
4.4	List of parameters necessary for calculating the bonded potential energies associated with molecules using TraPPE-EH. Note that the interaction site for the C-H bond pseudo atom is located at the center of the bond resulting in an actual distance of .55 Å. Additionally, the X-C-C-H torsion is calculated in a special way and only considers hydrogen pseudo-atoms that are part of a methylene group. Hydrogen atoms that are part of a methyl group are excluded.	38
4.5	List of the associated non-bonded interaction parameters for TraPPE-SM	39
4.6	List of parameters used for bonded interactions associated with TraPPE-SM	39
4.7	List of the associated non-bonded interaction parameters for AUA-4 .	40
4.8	List of parameters used for bonded interactions in AUA-4	41

Chapter 1

Introduction

Reservoir simulations have become increasingly common in the oil-extraction industry [23]. Typically, these calculations solve the compressible Navier-Stokes equations over length scales ranging from hundreds of kilometers to the length scale of porous rock (meters) using massively parallel computer algorithms. One limitation associated with the compressible formulation is the need for an equation of state (EOS) for the reservoir fluid. Unfortunately, in real life situations it is not always possible to find closed form equations of state. For example, this becomes difficult when the reservoir fluid has many constituents. In fact, analytical EOSs are not available for mixtures with three or more components. Reservoir fluids are usually composed of a mixture of various n-alkanes, O₂, CO₂, SO₄, CO₃, HCO₃, H₂O, as well as other compounds, making them particularly challenging to model analytically. An alternative approach is to use experimental data to develop an EOS.

Variations in the compressibility are particularly important near the critical point and in the two phase region. As such, we will be focusing on modeling the two-phase region of liquid-vapor equilibrium for mixtures. For a system in liquid-vapor equilibrium (LVE) the number of degrees of freedom is given by the Gibbs phase rule:

$$F = C - P + 2 \tag{1.1}$$

where F denotes the degrees of freedom available to the system, C is the number of

components in the mixture, and P is the number of phases coexisting. During LVE two phases coexist, so the number of degrees of freedom is the same as the number of components. In a binary mixture (two components), the system is completely defined by two parameters. For example we can use any two of the temperature, the pressure, and the Gibbs free energy. We need to keep in mind that the volume of the state space grows exponentially with the number of degrees of freedom. Completely defining the state of a ternary system (three components) would, as an example, require a temperature, pressure, and Gibbs free energy. It quickly becomes apparent that, for mixtures containing three or more components, it is impractical to completely explore the two phase region using experimental results [1]. As a result, ternary experimental data is often limited to just the few temperatures and pressures of particular interest for any given mixture.

In most cases with complex reservoir fluids, it is difficult to generate a differentiable EOS using experimental data. Fortunately, LVE properties can be calculated using molecular simulation techniques, such as the Gibbs Ensemble Monte Carlo (GEMC) method which will be discussed in Chapter 3. However due to the high computational cost associated with these simulations, some of the drawbacks associated with physical experiments are also present in simulation based approaches, albeit to a lesser degree. Therefore, the need to develop methods which can create an accurate and differentiable EOS based on a minimal number of data points is compelling.

One promising approach is to use machine learning to interpolate the thermodynamic properties of the data. Specifically, we will be using neural networks, though other methods such as Gaussian processes can produce equivalent results [2]. Machine learning is well suited to this task because it allows for the EOS to be defined or redefined for any valid set of independent variables that fully defines the system and is well suited to the high dimensionality of the problem of interest. Additionally, it is updatable, meaning that new data can be incorporated to improve the accuracy of the interpolated EOS.

In this thesis, we develop a method for using artificial neural networks to calculate an approximate representation of the LVE of ternary mixtures using a small set of data

for which the properties of the coexisting phases are known. In Chapter 2 we provide some background information about the LVE of ternary mixtures, the methods used to simulate such systems, such as Gibbs Ensemble Monte Carlo (GEMC), and the challenges associated with these methods. In Chapter 3 we review the theory behind GEMC. In Chapter 4 we discuss the Lennard Jones potential, the Lorentz-Berthelot combining rules for unlike particles, bonded interactions, and the available molecular models that can be used to model the potential energy of alkanes and other molecules of interest. In Chapter 5 we compare pure component and binary mixture LVEs obtained using GEMC simulations to the experimental results to determine which molecular model is the most accurate for our future calculations. In Chapter 6 we discuss the theory Bayesian regularized artificial neural networks for interpolating nonlinear multidimensional data. In Chapter 7 we apply artificial neural networks to available ternary LVE data sets to estimate the number of data points needed for a neural network to produce a reliable representation of LVE coexistence curves. In Chapter 8 we calculate LVE coexistence properties starting from binary coexistence data, thus demonstrating that this approach can be reliably used to obtain reliable LVE representations in the complete absence of experimental data (provided GEMC data can be obtained).

Chapter 2

Background

One of the difficulties associated with finding the ternary phase envelope is the large amount of data necessary to describe a system with so many dimensions. In many cases, data for binary systems is readily available, but data for ternary systems is limited to a specific temperature or pressure of interest to the previous investigators. It is now relatively common to use molecular simulation to generate data on the liquid-vapor equilibrium surface. Molecular models have been generated by various groups that allow for accurate simulation of various thermodynamic properties, including phase equilibria. The gold standard method for performing simulations using these potentials is the GEMC method developed by Panagiotopoulos et al [4]. The method works by performing simulations of the two coexisting phases in two separate simulation boxes.

While GEMC can obtain results with impressive efficiency, its convergence to a correct equilibrium can be greatly impeded by poor initial choices for coexisting compositions. Poor choices can result in slow convergence, simulation boxes containing the same phase, or the simulation never reaching equilibrium. Working around these problems without reliable information can be extremely computationally expensive. One way of improving the accuracy of the initial compositions is to develop an estimate of the equation of state based on already available data. This estimate can then be used to find an initial guess for a GEMC simulation.

A simple method for modeling the ternary liquid-vapor equilibrium surface in

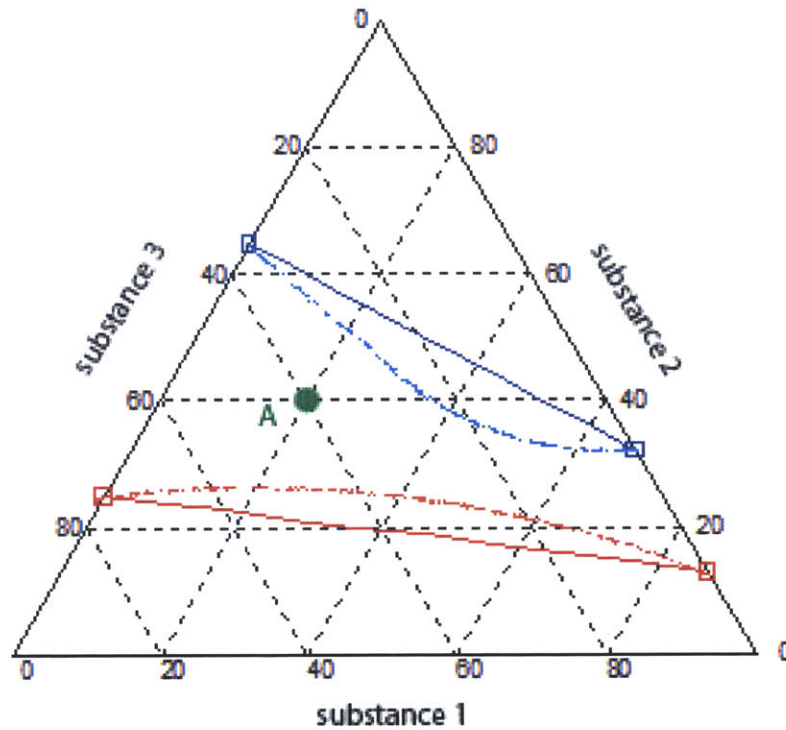


Figure 2-1: The composition of a ternary mixture can be represented on a ternary plot. In these plots the concentration of each component at a point is given by the three axes. The grid lines corresponding to each axis are the ones departing at 120° measured from the origin. For example point A has a composition of 20% substance 1, 40% substance 2, and 40% substance 3.

An example of a set of isothermal-isobaric LVE lines is also shown. In this example, the blue line is the isothermal-isobaric coexistence for the vapor phase and the red line is the isothermal-isobaric coexistence line for the liquid phase. The dotted lines are representations of typical experimental data. The solid lines correspond to an approximation of the isothermal-isobaric line from linearly interpolating the binary LVE data points (\square).

cases where there is limited data, amounts to assuming that the isothermal-isobaric lines connecting the binary LVE data points are linear. We will call this the linear approximation method. This effectively assumes a simplified model in which the chemical potential is constant for a given temperature and pressure.

An example of these lines is shown in Figure 2-1 for the case of a ternary mixture; the substances in this particular mixture (nitrogen, methane, and carbon dioxide) have chemical potentials with very weak dependence on the mole fraction of each component, and for this case the linear assumption is a relatively good approximation.

Here we remind the reader that in general, the chemical potential of substance i in a mixture is given by

$$\mu_i(x_i) = \mu_i^\theta + RT\ln(\gamma_i x_i) \quad (2.1)$$

where R is the gas constant, T is the temperature, μ_i^θ is the chemical potential at a reference state (typically ambient conditions), x_i is the mole fraction of component i , and γ_i is the activity coefficient [20]. The linear approximation neglects the logarithmic term in equation (2.2). As a result using a linear approximation will not always be reasonable, since it is ignoring potentially significant contributions from the change in chemical potential as the composition of the mixture changes. Unfortunately there is little we can do to address this from a thermodynamics standpoint.

Furthermore, there are additional problems that arise from using a linear approximation for the coexistence in a ternary mixture. Strong interactions between the components are a possibility, especially when considering components that are polar and/or ionic. In these cases the activity coefficient, which in ideal cases has a value of one, can be significantly different from one or even depend on the composition of the mixture.

These dependencies can result in large nonlinearities in the liquid-vapor equilibrium surface that are impossible to capture using a linear approximation. One of the most difficult ternary mixtures to model is the combination of methane, ethane, and carbon dioxide [3]. This is due to the strong deviation from Raoult's Law and corresponding positive azeotrope for a system composed of ethane and carbon diox-

ide. Raoult's Law states that the pressure of a solution is a function solely of the pure solvent at the same temperature scaled by the mole fraction of the solvent present.

Azeotropes, also referred to as constant boiling mixtures, can occur when two components have similar vapor pressures and strong intermolecular interactions, resulting in a maximum or minimum in the vapor pressure as a function of the composition of the mixture. An example of an azeotrope is shown in binary example of Figure 2-2. Finding ternary LVE for mixtures with azeotropes is effectively a worst case scenario where the estimates made by a linear approximation of the mole fraction can have an absolute error as high as 15% [16].

As previously mentioned, GEMC is sensitive to initial conditions. One potential solution to this problem is to use a more advanced method to model the liquid-vapor equilibrium surface for the purpose of providing initial conditions to GEMC calculations. We propose the use of machine learning, more specifically neural networks, to interpolate the available data. There are a number of benefits to using neural networks to generate this model. First, neural networks are specifically intended to handle systems with many inputs and many outputs. This is particularly important, because the liquid-vapor coexistence curves for a ternary system need to be fit in a six dimensional space (temperature, pressure, liquid mole fraction of component 1, liquid mole fraction of component 2, vapor mole fraction of component 1, and vapor mole fraction of component 2). Additionally, a neural network can always be retrained with an expanded set of data to improve accuracy. By combining GEMC simulations and neural networks it should be possible to estimate new compositions and refine them to eventually generate a neural network that can accurately generate estimates throughout the two phase region.

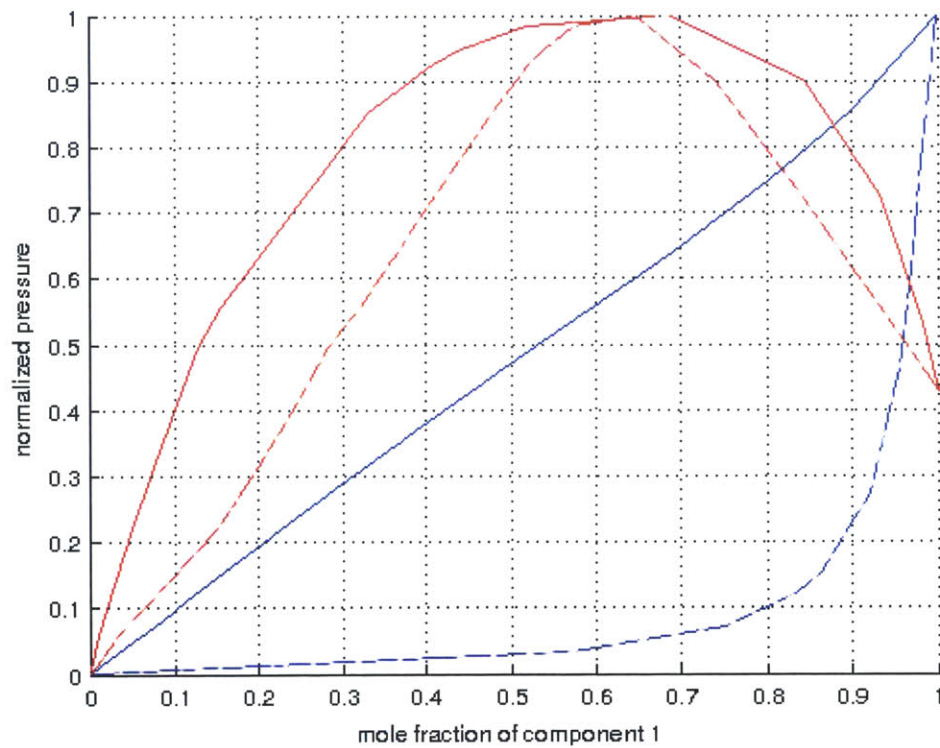


Figure 2-2: Two examples of binary LVE isotherms, where blue is the ideal mixing case and red has an azeotrope. Solid lines indicate the fluid phase and dotted lines indicate the vapor phase.

Chapter 3

Gibbs Ensemble Monte Carlo

Gibbs Ensemble Monte Carlo (GEMC) [4] is the prevalent method of simulating phase equilibria in fluids. The method's popularity is due to its intuitive formulation as well as its straightforward implementation. Additionally, GEMC requires significantly less a priori information about the phase diagram than alternative simulation methodologies.

3.1 Overview

GEMC is a Metropolis Monte Carlo (MC) algorithm [5]. Metropolis MC are methods for sampling statistical mechanical equilibrium distributions using importance sampling. By setting up a Markov Chain [24], these methods generate samples of the desired distribution by only considering the relative probability of consecutive states, thus never calculating the distribution function normalization, which is a high dimensional integral (similar to the one of interest).

In the case of using Metropolis MC for molecular simulation, every Monte Carlo move amounts to a perturbation of the system, such as displacing a molecule. The energy of the new state is compared to that of the old state. If the move results in a lower system energy, the move is accepted. If the move increases the system energy then the move is accepted with probability

$$p = \frac{\exp(-\beta E_{new})}{\exp(-\beta E_{old})} \quad (3.1)$$

This generates states with probability proportional to $\exp(-\beta E)$ where E is the equilibrium state energy. In the case where a move is rejected, the system stays at the old state for another iteration.

3.2 Implementation

The greatest difficulty associated with performing an accurate simulation of a two phase system comes from the presence of interfacial effects which are never absent unless the system is truly infinite. GEMC simulations avoid this difficulty by using two communicating but not spacially adjacent (see below) simulation boxes, each containing one of the phases of interest. These boxes can be brought to equilibrium with each other by exchange of molecules, but are each able to remain in one phase because of the energy barrier associated with creating an interface within a simulation box [4][5]. Particles in the two boxes do not interact directly, and as a result, no interface exists between the two boxes.

In addition to being in internal equilibrium, each of the phases needs to be in equilibrium with each other by having equal temperature, pressure, and chemical potentials for each component. Thermal equilibrium is established by allowing translation and rotation of molecules within their starting box. Pressure is equilibrated by transferring volume between the two simulations (NVT GEMC) or expanding/contracting the boxes independently (NPT GEMC). The chemical potential is equilibrated by using particle transfer moves between the two boxes.

The full statistical mechanics of the Gibbs Ensemble was developed by Smit et al [6] and Smit and Frenkel [7]. For a one component system in the canonical ensemble (NVT) divided into two subregions, they give the partition function as

$$Q_{NVT} = \frac{1}{\Delta^{3N} N!} \sum_{N_I=0}^N \binom{N}{N_I} \int_0^V dV_I V_I^{N_I} V_{II}^{N_{II}} \int d\xi_I^{N_I} \exp[-\beta U_I(N_I)] \int d\xi_{II}^{N_{II}} \exp[-\beta U_{II}(N_{II})] \quad (3.2)$$

where Δ is the de Broglie wavelength, $\beta = \frac{1}{K_b T}$, K_b is the Boltzmann constant, ξ_I and ξ_{II} are the scaled coordinates of the particles, with the roman numeral subscript indicating the corresponding simulation box and $U(N)$ is the intermolecular potential energy associated with N molecules. Since the total volume and number of molecules are conserved, $V_{II} = V - V_I$ and $N_{II} = N - N_I$.

Smit et al. [6] showed that the partition function given in equation (3.2) and a free energy minimization procedure will, for a system with a first-order phase transition, result in the two subregions reaching the correct equilibrium density.

3.3 Acceptance Criteria

To minimize the free energy, GEMC uses the Metropolis Algorithm, which utilizes the relative probability of two states without requiring knowledge of the partition function for the system [4]. Specifically, the relative probability between states is required for acceptance rejection purposes. For canonical (NVT) conditions, the probability distribution obeys

$$\varrho_{NVT}(N_I, V_I; N, V, T) \propto \frac{N!}{N_I! N_{II}!} \exp[N_I \ln V_I + N_{II} \ln V_{II} - \beta U_I(N_I) - \beta U_{II}(N_{II})] \quad (3.3)$$

The probability density function for a NPT system obeys [4]

$$\varrho_{NPT}(N_I, V_I; N, P, T) \propto \frac{N!}{N_I! N_{II}!} \exp[N_I \ln V_I + N_{II} \ln V_{II} - \beta U_I(N_I) - \beta U_{II}(N_{II}) - \beta P(V_I + V_{II})] \quad (3.4)$$

3.3.1 Translation/Rotation

Translation and rotation moves are fairly simple. In the case of a translation move, a molecule is chosen at random from a box and is moved some small distance. In the case of a rotation move, a molecule is rotated by a small angle in a random direction, but this only applies if the molecular model consists of more than one bead. The probability of acceptance is straightforward to calculate in this case, because only the configuration is affected. The acceptance probability is:

$$\min [1, \exp(-\beta\Delta U)] \quad (3.5)$$

Here $\Delta U = U'_1 - U_1 + U'_2 - U_2$, where U_1 and U'_1 are the energies in box one before and after the trial move and U_2 and U'_2 are the energies in box two before and after the trail move. This equation is valid for both NVT and NPT GEMC simulations.

3.3.2 Molecule Transfer

In a molecule transfer move, a random molecule type is selected with uniform probability. Then a random molecule of that type is moved to the other box. Unlike the previous move, the acceptance probability needs to account for the fact that there has been a change in the multiplicity of the systems due to the change in the number of molecules in each box. It is given by:

$$\min \left[1, \left(\frac{N_{II,j} \times V_I}{(N_{I,j} + 1) \times V_{II}} \right) \exp(-\beta\Delta U_I - \beta\Delta U_{II}) \right] \quad (3.6)$$

In this equation, as before, the roman numeral subscripts indicate the two simulation boxes. As written, equation (3.6) is for the transfer of a molecule from simulation box II to simulation box I . If the system has multiple components, j indicates the species of the molecule being transferred. This move applies to both types of simulation; however, the species type will only come into play for NPT simulations. This move also satisfies the condition that probability of transferring a molecule from an empty box is zero.

3.3.3 Volume Transfer for NVT Simulations

Volume transfer is one of the moves that can be used to equilibrate the pressure between the two simulation boxes. In this move, a volume change of size ΔV is applied to the two boxes (with opposite signs). All molecule positions are rescaled to the new size of the simulation boxes. The resulting acceptance probability is given by:

$$\rho_{VolumeTransfer} = \min \left[1, \exp \left(-\beta\Delta U_I - \beta\Delta U_{II} + N_I \times \ln \frac{V_I + \Delta V}{V_I} + N_{II} \times \ln \frac{V_{II} - \Delta V}{V_{II}} \right) \right] \quad (3.7)$$

where the last two terms in the exponential account for the entropy associated with volume change. It is important to note that ΔV must be sampled uniformly around a value that adjusts during the simulation run to keep the acceptance ratio at a desirable level. This move is only applicable to constant NVT simulations, since it conserves total volume while equilibrating the pressure between the two boxes.

3.3.4 Volume Change for NPT Simulations

Much like the volume transfer move, the volume change move allows the two simulation boxes to approach an equilibrated pressure; however, in the case of this move, both boxes are adjusted independently towards a set pressure P_{const} . The acceptance probability for each box is given by:

$$\min \left[1, \exp \left(-\beta\Delta U_I + N_I \times \ln \frac{V_I + \Delta V}{V_I} - \beta P_{const} \Delta V \right) \right] \quad (3.8)$$

As with the volume transfer move, ΔV should be sampled uniformly around a value that adjusts during the simulation run to keep the acceptance ratio at a reasonable level. This move is only applicable to constant NPT simulations since it does not conserve the total volume, but does fix the average pressure at the specified value.

3.4 Determining Critical Points

Determining the critical points of the phase diagram can be very challenging using molecular simulation. Simulations that are performed at conditions near the critical point usually produce results which deviate significantly from the experimental results. This is due to the decreasing energy needed to form an interface and results in both the liquid and vapor phases being present inside one or both of the simulation boxes. Unfortunately, it is difficult to separate the error in the simulation results from the true change in the macroscopic thermodynamic properties as one approaches the critical point. It is possible to obtain some improvement in accuracy, using the histogram reweighting method developed by Ferrenberg and Swendsen [8], by creating histograms of the energy and the density for each of the states that were accepted. The histograms can then be analyzed to determine the state or states with the highest probabilities. However, this method can be very sensitive to the size of the simulated system. Increasing the number of molecules reduces the noise in the histogram, making it easier to distinguish multiple peaks in the data set. Unfortunately, the accuracy close to the critical point is still limited by the finite box size. This is because at the critical point, the characteristic correlation length of the system goes to infinity. The finite size of the simulation box means that the macroscopic thermodynamic properties will not be reliable once the correlation length approaches or exceeds the size of the simulation box.

3.4.1 Single Component Systems

In the case of a single component system, it is relatively easy to obtain an accurate estimate of the location of the critical point by fitting simulation results to the expected theoretical results for a non-classical system near a critical point. These equations take into account the density fluctuations that occur near the critical point resulting in a flatter peak. This can be done using the rectilinear diameter rule [9]

$$(\rho_l + \rho_g)/2 = p_c + C(T_c - T) \tag{3.9}$$

as well as the scaling relationship for the width of the coexistence curve [9]

$$\rho_l - \rho_g = A(T_c - T)^b \quad (3.10)$$

Here, A and C are both fitting parameters, ρ_l and ρ_g correspond to the density of the liquid and gas respectively, T_c is the critical temperature, ρ_c is the critical density, and $b = .325$ for a non-classical three dimensional system. Both equations (3.9) and (3.10) were derived empirically using experimental results and then later derived for a hard sphere potential model [18]. Fitting simulation results requires a careful balance, since it is important to obtain data as close to the critical point as possible for the fit to be as accurate as possible; at the same time, it is also important that the data is sufficiently far from the critical point to prevent finite size effects from resulting in a poor fit.

Chapter 4

Potential and Model

In the GEMC method, the probability of accepting a Monte Carlo move is dependent on the change in potential energy from the old state to the new, candidate, state. In this Chapter we discuss the potential energy models used in our work. Molecular models are required because ab-initio calculation of intermolecular forces and interaction energies is too expensive. Instead, models which capture the main features of the former are built from basic building blocks, such as electrostatic interactions and van der Waals interactions between one or more interaction sites (or beads) and can be as simple as a single site for an entire molecule. The force field parameters are usually determined by fitting to various known properties, such as the enthalpy of vaporization and/or the critical point.

Typically, to minimize computational cost, it is desirable to use the simplest force field that can accurately predict the relevant physics. More complexity is needed when dealing with molecules that are ionic, polar, or have large aspect ratios. When dealing with mixtures, other important factors come into play, such as the interactions between unlike molecules. In such cases, special attention needs to be paid to consistently choose force fields that are compatible with each other. In many cases, accurately modeling a mixture requires fitting the force field to some properties associated with benchmark mixtures. This usually means using one type of force field for the whole simulation; one exception is the Transferable Potential for Phase Equilibria (TraPPE) group of force fields [11] [12]. These were developed by

the Siepmann group, and focus on the liquid vapor equilibria of alkanes, but latter additions also incorporated models for some ionic and polar molecules [13]. TraPPE potentials are unusual because they are designed to be transferable. This is done by breaking alkanes down into their component chemical groups and designing a force field for the chemical groups made of one or more interaction sites. These force fields are then optimized to recreate the experimental results for several different molecules that contain the group, as well as for mixtures containing those molecules. A similar process can be used to ensure that the several types of TraPPE force fields can be used in the same simulation.

Another type of available model is the Anisotropic United Atom revision 4 (AUA-4) that was developed by Ungerer [14]. In this Chapter, we will be looking at several of the TraPPE force fields as well as AUA-4.

The total potential energy of a molecule modeled using a potential is divided into bonded and non bonded contributions. The non bonded potentials, typically Lennard Jones and electrostatic interactions, are used only for the interactions of pseudo-atoms belonging to different molecules or for pseudo-atoms that are sufficiently far apart that they do not have any bonded contributions to the potential energy.

4.1 Lennard Jones Potential - Non Bonded Interactions

Perhaps in search of simplicity, the force fields used for modeling coexistence are typically based on the well-known Lennard-Jones potential plus electrostatic interactions, which for the interaction between molecules A and B can be written in the form

$$U = \sum_{i \in A} \sum_{j \in B} \left(4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \left[\frac{q_i q_j}{r_{ij}} \right] \right) \quad (4.1)$$

Here, i and j index the interaction sites (locations) in molecules A and B respectively, r_{ij} is the distance between the sites, ϵ_{ij} is the well depth, σ_{ij} is the core diameter, and q_i is the partial charge at interaction site i . Partial charges are the non-integer charge

values (when measured in elementary charge units) created due to the asymmetric distribution of electrons in chemical bonds. The primary difference among the different force fields that use the Lennard Jones Potential is the values of the parameters ϵ_{ij} and σ_{ij} as well as the differences in the number and location of interaction sites.

4.2 Combination Rules

For an interaction between two like particles, equation (4.1) seems fairly straightforward. However, in the case of an interaction between a pair of unlike molecules there is no longer a well defined well depth or interaction distance. The primary way of modeling this is through combination rules. In our case, we will be making use of the popular Lorentz-Berthelot combining rules:

$$\sigma_{ij} = \frac{\sigma_{ii} + \sigma_{jj}}{2} \quad (4.2)$$

$$\epsilon_{ij} = \sqrt{\epsilon_{ii}\epsilon_{jj}} \quad (4.3)$$

These rules are not very accurate, but are usually sufficient (more complex rules do not produce noticeable improvements in simulation result accuracy), and due to their simplicity have become very common.

4.3 Bonded Interactions

Bonded interactions for all molecular models can be broken down into types based on the number of interaction sites involved. Interactions between two beads that are directly bonded (1-2 interaction) are considered as having a fixed length for all of the models being considered. In more complex cases, flexible bond lengths can be used; however, this phenomenon is not typically significant at the temperatures of interest for LVE. The bonded interaction for a triplet is an angle potential (1-3 interaction). In the models being considered, this interaction is modeled as a harmonic potential

given by:

$$u_{bend}(\theta) = \frac{k_\theta}{2} \times (\theta - \theta_{eq})^2 \quad (4.4)$$

where k_θ is the force constant associated with the bending of the bond angle and u_{bend} is the potential energy stored by the deviation of the bond angle θ from the equilibrium angle θ_{eq} .

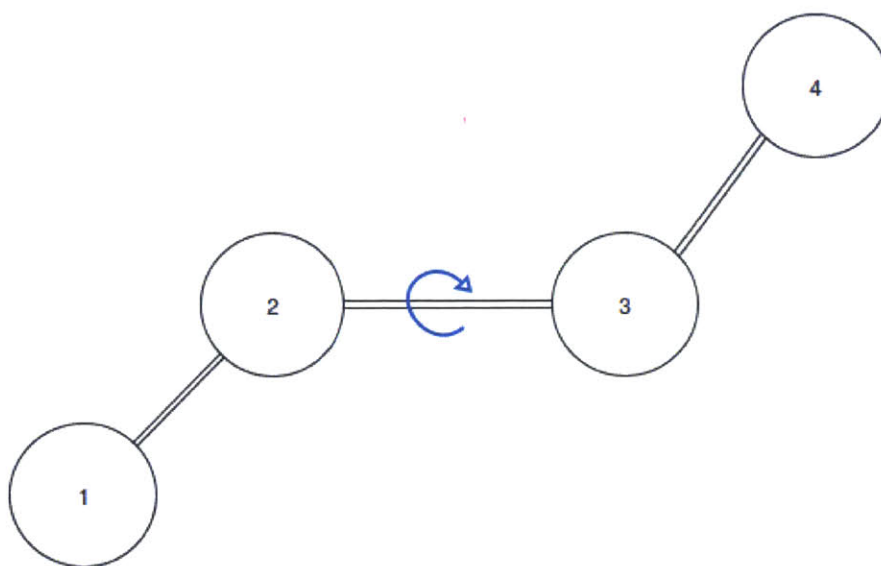


Figure 4-1: The dihedral angle, or torsion angle, is the angle between normal vectors of the two planes (1,2,3) and (2,3,4). For the example molecule shown, the dihedral angle is -180 degrees.

In the case of a large molecule, interactions between sites separated by three bonds become important (1-4 interactions). The potential energy for these interactions is a function of the dihedral angle. An example of the dihedral angle in butane is shown in figure 4-1. The potential energy for this interaction is expressed using trigonometric functions. For example, in TraPPE models the torsion is modeled using:

$$u_{tors}(\phi) = c_0 + c_1[1 + \cos(\phi)] + c_2[1 - \cos(2\phi)] + c_3[1 + \cos(3\phi)] \quad (4.5)$$

where c_i and e_i are constants defined in the molecular model and ϕ is the dihedral angle.

In the case of AUA models, the potential energy stored in torsion is calculated using an eighth order Ryckaert-Bellemans potential to more accurately model the torsional potential.

$$u_{tors}(\theta) = \sum_{j=0}^8 a_j (\cos\phi)^j \quad (4.6)$$

Here $j = 0, 1, \dots, 8$ corresponds to the Fourier mode and a_j scales the response to the j^{th} mode as defined in the molecular model. The Ryckaert-Bellemans potential can be rewritten in the same form as equation (4.5) by expanding the Fourier series.

4.4 Molecular Models

4.4.1 TraPPE-UA

The Transferable Potentials for Phase Equilibria - United Atom model is the most popular and most extensive force field in the TraPPE family [11]. As with any United Atom approach, the total number of interaction sites in a molecule is kept as small as possible. This is by creating pseudo-atoms that lump a number of real atoms together. In the case of TraPPE-UA, a carbon atom together with all of its bonded hydrogen atoms is represented as a single interaction site located at the site of the carbon atom. When modeling n-alkanes, these pseudo atoms will be CH₄, CH₃, CH₂, CH and C. Lennard Jones parameters for each of these groups were determined by fitting to the critical points as well as the saturated liquid densities for n-alkanes from methane to dodecane. The parameters associated with these groups can be found in Table 4.1. Additional pseudo atoms exist that can be useful for modeling other, more complicated molecules. The parameters for the bonded interactions between these pseudo atoms are given in Table 4.2.

Pseudo Atom Type	$\epsilon/k_b[K]$	$\sigma[\text{\AA}]$
CH	10	4.68
CH ₂	46	3.75
CH ₃	98	3.75
CH ₄	148	3.73

Table 4.1: List of the associated non-bonded interaction parameters for TraPPE-UA

Parameter	Value	Units
Bond Length	1.54	\AA
Bending		
Bond Angle (θ_0)	112	degrees
k_θ/k_B	62500	K/rad ²
Torsion		
c_0/k_b	0	K
c_1/k_b	355.03	K
c_2/k_b	-68.19	K
c_3/k_b	791.32	K

Table 4.2: List of parameters used for bonded interactions in TraPPE-UA

4.4.2 TraPPE-Explicit Hydrogen

In certain cases, United Atom methods can have difficulty accurately representing certain components or mixtures. In the case of TraPPE-UA, the pseudo atoms result in a relatively coarse model of the atom. While this saves computational time, it can limit the ability of the model to accurately predict the interactions in mixtures with ionic compounds or non polar compounds. TraPPE-EH includes interaction sites for every atom in the molecule [12]. The additional features provided by the extra interaction sites can provide a higher level of accuracy than could be achieved with a simpler model. The downside to this increased accuracy is the significant increase in the number of interaction sites, which greatly increases the computational cost of performing a simulation.

TraPPE-EH molecular models are more complicated to create. It is important to note that the interaction site for hydrogen molecules is not located at the nucleus of the hydrogen atom, but is instead shifted with the electron cloud to the center of the H-C bond. This is necessary because the lone electron of the hydrogen atom is no longer around the nucleus but bonded to a carbon atom. As in the TraPPE-UA case, various types of bonds between the different carbon pseudo-atoms. However, the hydrogen pseudo-atoms are treated as being rigidly attached to their corresponding carbon atom. The flexibility in the hydrogen-carbon bonds is unnecessary because they do not contribute to the bending of the overall chain structure and so would only come into play at extremely high energies. The necessary parameters for this model are shown in Table 4.3 and Table 4.4. The torsional energy for the Hydrogen pseudo-atoms is treated as a special case with:

$$u_{tors}(X - C - C - H) = c_X [1 - \cos(3\phi)] \quad (4.7)$$

This equation only applies to Hydrogen atoms that are part of a methylene group. Methyl group hydrogen are excluded from contributing to the torsional energy. The parameters associated with this force field can be found in Tables 4.3 and 4.4.

Pseudo Atom Type	$\epsilon/k_b[K]$	$\sigma[\text{\AA}]$
C-H bond	15.3	3.31
C(H ₂)	5.00	3.65
C(H ₃)	4.00	3.30
C(H ₄)	0.01	3.31

Table 4.3: List of the associated non-bonded interaction parameters for TraPPE-EH

Parameter	Value	Units
Bond Length (C-C)	1.535	\AA
Bond Length (C-H)	0.55	\AA
Bending		
Bond Length (C-H Bond)	1.10	\AA
Bond Angle (C-C-C)	112.7	degrees
Bond Angle (C-C-H)	110.7	degrees
Bond Angle (H-C-H)	107.8	degrees
k_θ/k_B	58765	K/rad ²
Torsion (C-C-C-C)		
c_0/k_b	0	K
c_1/k_b	355.03	K
c_2/k_b	-68.19	K
c_3/k_b	791.32	K
Torsion for Hydrogen in Methyl Group (X-C-C-H)		
c_C/k_b	854	K
c_H/k_b	717	K

Table 4.4: List of parameters necessary for calculating the bonded potential energies associated with molecules using TraPPE-EH. Note that the interaction site for the C-H bond pseudo atom is located at the center of the bond resulting in an actual distance of .55 \AA . Additionally, the X-C-C-H torsion is calculated in a special way and only considers hydrogen pseudo-atoms that are part of a methylene group. Hydrogen atoms that are part of a methyl group are excluded.

4.4.3 TraPPE-SM

TraPPE models have also been developed for a variety of small molecules that are compatible with the TraPPE potentials for the n-alkane models described above [13]. TraPPE-Small Molecule includes models for carbon dioxide, nitrogen, helium, oxygen, ethylene oxide, and ammonia. For our purposes we will be making use of both carbon dioxide and nitrogen models. TraPPE-SM models have been fit to produce accurate binary mixing results with the other TraPPE models. The parameters associated with these models are listed in Tables 4.5 and 4.6.

Pseudo Atom Type	$\epsilon/k_b[K]$	$\sigma[\text{\AA}]$	$q[e]$
CO ₂			
C	27.0	4.68	+ .70
O	79.0	3.75	- .35
N ₂			
N	148	3.73	- .482
Center Of Mass (COM)	0	0	+ .964

Table 4.5: List of the associated non-bonded interaction parameters for TraPPE-SM

CO ₂		
Parameter	Value	Units
Bond Length	1.160	\AA
Bond Angle (O-C-O)	180	degrees
N ₂		
Bond Length	0.550	\AA
Bond Angle (N-COM-N)	180	degrees

Table 4.6: List of parameters used for bonded interactions associated with TraPPE-SM

4.5 Anisotropic United Atom - 4

The AUA-4 model was created as an alternative to the TraPPE potentials [14]. Originally developed for modeling equilibrium and transport properties, the fourth revision was created to fix deficiencies in the third version that resulted in inaccurate coexistence diagrams. This was done by fitting the potential to vapor pressures, liquid densities, and vaporization enthalpies. AUA-4 is a united atom potential that consolidates groups of atoms into a pseudo atom to allow for reduced computational costs, allowing for larger or longer simulations than models that have interaction sites for every atom. AUA-4 is similar to the TraPPE-UA model in that it is a transferable potential that consolidates molecular groups into pseudo-atoms that are used to build various molecules; however, it is unique in the fact that it does not restrict the location of the pseudo atoms to the location of the carbon atom. The offset, included in Table 4.8, moves the pseudo-atom from the location of the carbon atom in the average direction of the bonded hydrogens. Theoretically this offset distance should allow for the potential to provide improved accuracy over TraPPE-UA while being more computationally efficient than TraPPE-EH. One of the primary drawbacks of this model is that it is less widely used than the TraPPE models and as such, it does not have as extensive a selection of compatible molecular models to choose from for non-alkanes. Additionally, because they are fit to different properties, there is no guarantee of accuracy when molecules modeled using AUA-4 are used in conjunction with molecules modeled with TraPPE.

Pseudo Atom Type	$\epsilon/k_b[K]$	$\sigma[\text{\AA}]$
CH ₂	86.291	3.4612
CH ₃	120.15	3.6072
CH ₄	149.92	3.7327

Table 4.7: List of the associated non-bonded interaction parameters for AUA-4

Parameter	Value	Units
Bond Length	1.535	Å
Offset CH ₂ (δ_{CH_2})	.38405	Å
Offset CH ₃ (δ_{CH_3})	.21584	Å
Bending		
Bond Angle (θ_0)	114	degrees
k_θ/k_B	62500	K/rad ²
Torsion		
a_0/k_b	1001.35	K
a_1/k_b	2129.52	K
a_2/k_b	-303.06	K
a_3/k_b	-3612.27	K
a_4/k_b	2226.71	K
a_5/k_b	1965.93	K
a_6/k_b	-4489.34	K
a_7/k_b	-1736.22	K
a_8/k_b	2817.37	K

Table 4.8: List of parameters used for bonded interactions in AUA-4

Chapter 5

GEMC simulation

Gibbs Ensemble Monte Carlo was performed using the Gibbs program developed by Panagiotopoulos and Errington [4] [5]. For pure substances we used the NVT version of GEMC to determine the liquid-vapor coexistence curves. Simulations were performed using 600 particles in each simulation box and initially equilibrated for at least 2,000,000 Monte Carlo moves before data was produced over a production run of at least 5,000,000 MC moves.

For both binary and ternary mixtures, the NPT version of GEMC was used. Simulations were initially run with inter-box moves disabled to allow for the density in each of the simulation boxes to equilibrate. Equilibration took place over 5,000,000 MC moves. Then, the resulting configuration was restarted with inter-box moves enabled. In this case, the minimum number of moves for equilibration was increased to 15,000,000 and the results were averaged over 20,000,000 MC moves. In the case of mixtures, the initial number of molecules per box was 1000 (2000 total molecules).

5.1 Pure Component Results

Simulations for the pure component systems show that each of the models is implemented correctly. As shown in Figures 5-1, 5-2, and 5-3, each of the models predicts the liquid-vapor equilibrium curve with small deviations. These figures also show the estimated location of the critical point calculated using the method of section 3.4.1.

The good agreement with experiments serves to demonstrate the capability to accurately predict critical points in the case of pure components from GEMC simulation results.

By closely comparing the simulation results to the experimental data obtained by Younglove [15], we can evaluate the relative strengths and weaknesses of each of the molecular models, and connect the small deviations from the experimental data to the basic assumptions inherent in each model.

Comparing the TraPPE-UA models results to the experimental data shows that the model is highly accurate for each of the three components shown (methane, ethane, and propane). Combined with its low computational cost, it is easy to see why this model is so widely used.

The TraPPE-EH model is also shown to have impressive accuracy, though slightly less accurate than the United Atom version. This is attributed to the fact that this model uses fitted potentials for the individual atoms which makes it more difficult to obtain very accurate results for a particular molecule type. The added complexity of the model does not appear to result in any improvement when simulating a pure component system, however it does result in a large increase in computational cost.

The AUA-4 model seems to have a fairly consistent tendency to overestimate the density of the fluid phase. This error is fairly apparent in Figure 5-1 and in figure 5-2. AUA-4 has very good accuracy when simulating the density of the vapor phase. Overall there are only some minor concerns with the accuracy of this model and it is too soon to determine if the slight inaccuracy in the fluid phase will cause significant problems while simulating mixtures.

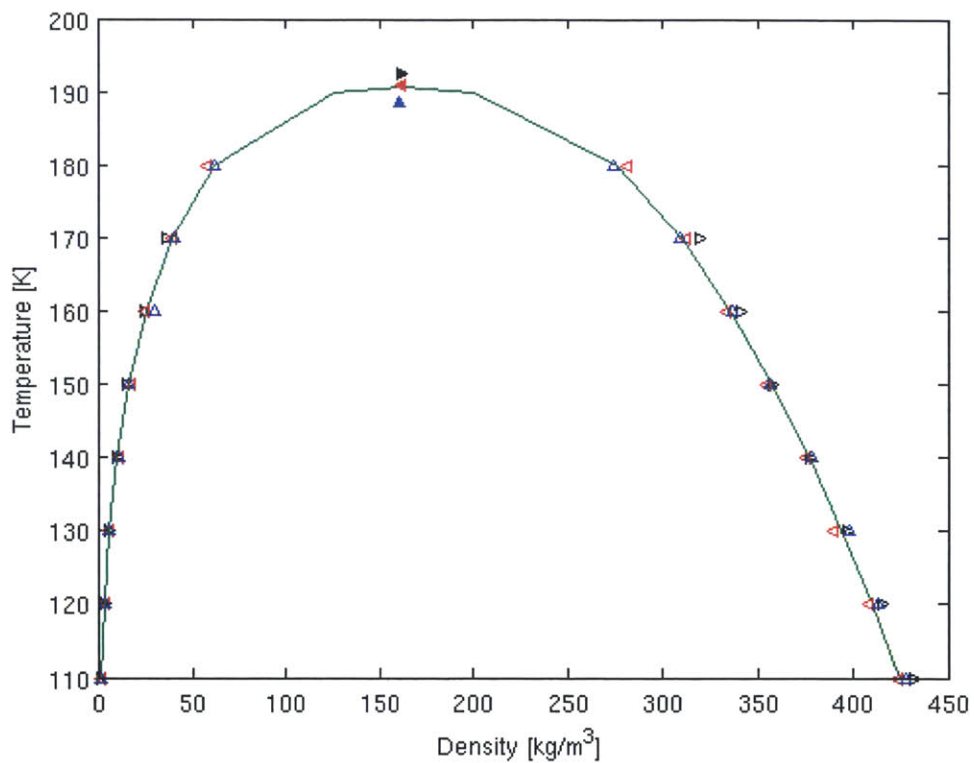


Figure 5-1: GEMC simulation results for liquid-vapor equilibrium in pure methane. The liquid-vapor equilibrium curve obtained from experimental results is indicated by the green line. Results from each of TraPPE-UA(\triangleleft), TraPPE-EH(\triangle), AUA-4(\triangleright) are plotted along with the estimated critical point (shown with a filled marker) for comparison. Error bars are smaller than symbol size unless otherwise indicated.

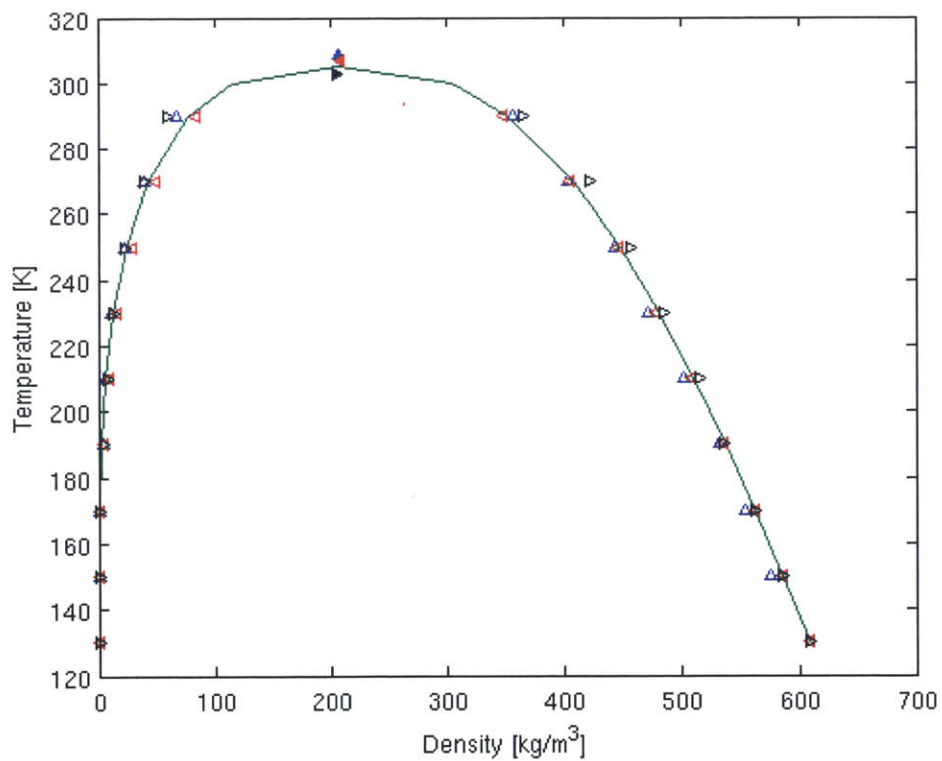


Figure 5-2: GEMC simulation results for pure ethane. Results from each of TraPPE-UA(\square), TraPPE-EH(\triangle), AUA-4(\triangleright) are plotted as well as the experimental results (-). The estimated critical point (shown with a filled marker) is shown for comparison. Error bars are smaller than symbol size unless otherwise indicated.

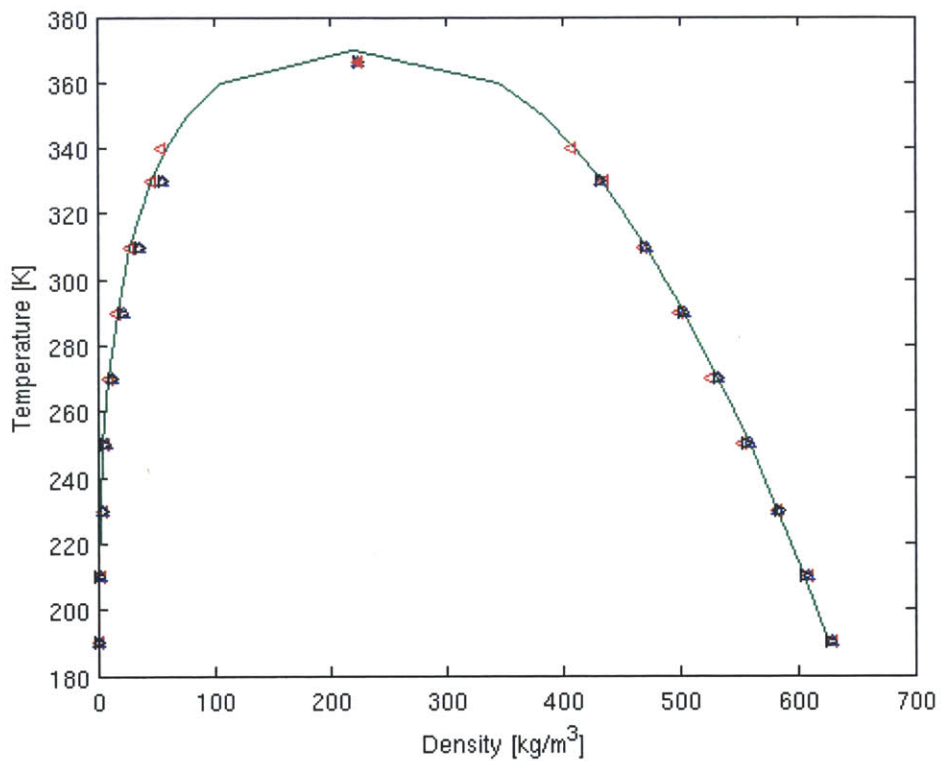


Figure 5-3: GEMC simulation results for pure propane. Results from each of TraPPE-UA(\triangleleft), TraPPE-EH(\triangle), AUA-4(\triangleright) are plotted as well as the experimental results (-). The estimated critical point (shown with a filled marker) is shown for comparison. Error bars are smaller than symbol size unless otherwise indicated.

5.2 Binary Component Comparison

It is important that the molecular model we choose for ternary simulations is able to account for highly non ideal mixtures that do not follow Raoult's Law. In order to test how accurately each force field can model the interaction between unlike molecules, we can perform simulations of the relevant binary mixtures. Since our ternary mixture will be composed of methane, ethane, and carbon dioxide, binary LVE simulations will be performed for a mixture containing methane and carbon dioxide, as well as a mixture containing carbon dioxide and ethane. The latter is an ideal test case, because the intermolecular interactions between ethane and carbon dioxide molecules are strong and result in many nonlinear features, including an azeotrope.

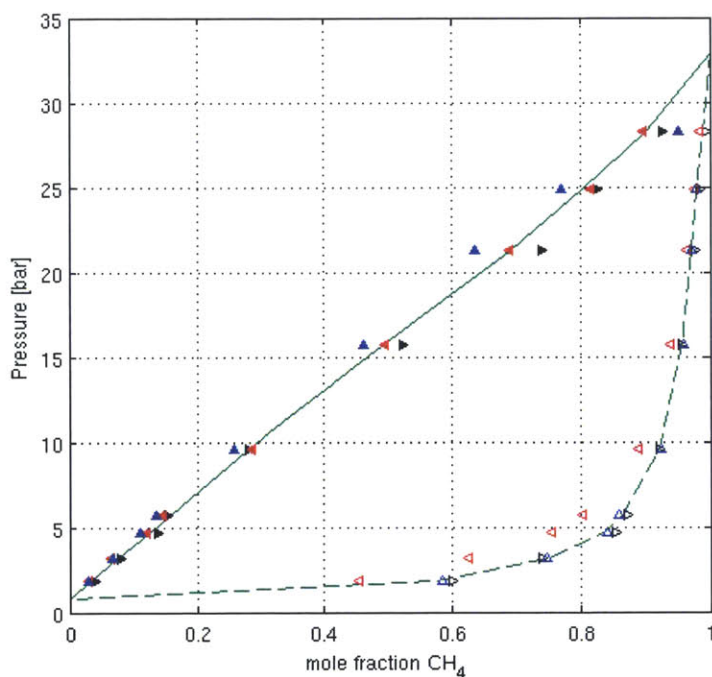


Figure 5-4: GEMC simulation results for liquid-vapor equilibrium for a mixture of CH_4 and C_2H_6 at 180K. Results from each of TraPPE-UA(\triangleleft), TraPPE-EH(\triangle), AUA-4(\triangleright) are plotted along with the experimental results (-). Empty markers and dashed line indicate the vapor phase while the solid markers and solid line indicate the liquid phase. Error bars are smaller than symbol size ($\approx 1\%$).

Figure 5-4 demonstrates that the TraPPE-UA model is too simple to account for the intermolecular interactions that occur in mixtures. Even in the case of a mixture of methane and ethane, which is usually approximated as ideal, this model fails to accurately predict the vapor composition. The vapor composition has consistently less methane by mole fraction at almost every trial pressure compared to the experimental results obtained by Wei et al. [16].

The results in Figure 5-4 also show that the TraPPE-EH model and the AUA-4 model both have inaccuracies in modeling the liquid phase. Fortunately for our purposes, the error in the results obtained by these two potentials appear to be less severe than the error for TraPPE-UA. It is somewhat interesting that the errors associated with these two potentials are of approximately the same magnitude but in opposite directions, with the TraPPE-EH model underestimating the mole fraction of methane in the liquid phase and the AUA-4 model overestimating the mole fraction of methane in the liquid phase.

The much more difficult case of a binary mixture composed of carbon dioxide and ethane, in Figure 5-5, shows that there are limits to what can be simulated with any given potential model of the type (4.1). In this case, the AUA-4 model predicts that there is no two phase equilibrium for any of the pressures tested at the temperatures of 207K and 213K. Additionally the TraPPE-UA and TraPPE-EH models had significant difficulty performing accurate simulation of this mixture, even though both models were able to obtain results indicating LVE. As discussed in Chapter 2, the deviation from ideal mixing is due to the interactions between unlike molecules. Even if the force field is fitted to account for this non ideal behavior, the azeotrope presents additional challenges. As shown in Figure 5-5 for a temperature of 213K, near the critical point there are four different compositions potentially existing at the same temperature and pressure. The implementation of GEMC being used here can only have two coexisting phases. This results in an extended region around the critical point in which GEMC results are unreliable. The coexistence curve for a temperature of 207K does not have this problem because the CO_2 in the mixture solidifies before a critical point is reached.

The effects of having an azeotrope will become less important when performing GEMC simulations for ternary mixtures. The addition of another component helps prevent the properties for an individual simulation box from changing to another phase. Additionally, the effects of the azeotrope will vanish quickly as the concentration of a third component is increased. From the simulation results in Figures 5-4 and 5-5, it is clear that using the TraPPE-EH force field is the best choice going forward.

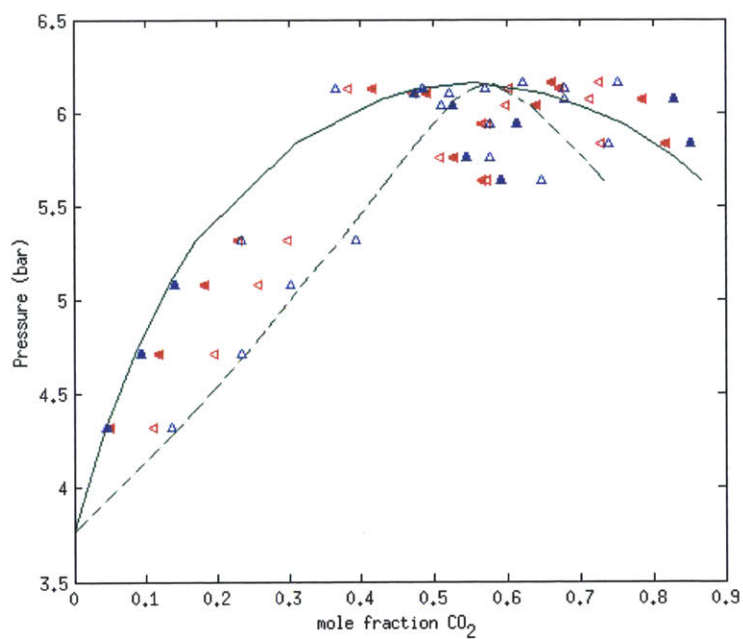
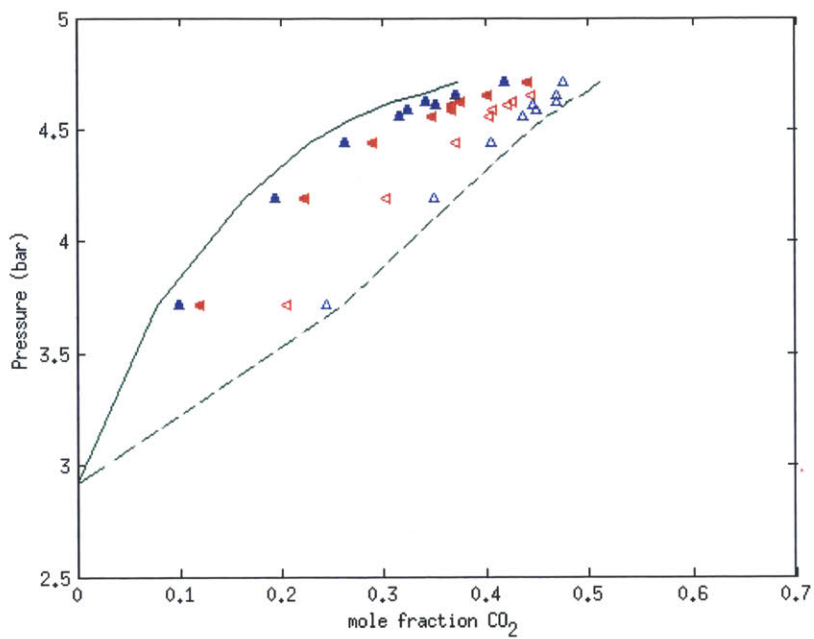


Figure 5-5: GEMC simulation results for LVE for a mixture of CO₂ and C₂H₆ at 207K (top) and 213K (bottom). Results from TraPPE-UA(\triangleleft) and TraPPE-EH(\triangle) are plotted along with the experimental results (-). Empty markers and dashed line indicate the vapor phase while the solid markers and solid line indicate the liquid phase. Error bars are smaller than symbol size ($\approx 1\%$).

Chapter 6

Artificial Neural Networks

As discussed in Chapter 1, analytical equations of state are not available for ternary mixtures. Additionally, increasing the number of degrees of freedom by adding components to the reservoir fluid quickly results in a system that is too large to describe through experimental techniques. Molecular simulation can be used to partly replace experiments; however, this approach also becomes very expensive as the number of components increases and the volume of the phase diagram increases. To address this we require a method for generating a reliable approximation to the EOS that is based on available data; in other words, we require a method which can interpolate data in a large number of dimensions.

Machine learning provides a convenient and popular method for creating interpolation models of nonlinear many-input many-output systems. These methods are especially useful when the class of functions and/or the order of the functions that relate the inputs and outputs is unknown. Rather than guessing a class and order of functions and risking over or under fitting the data, we can instead use machine learning to build a model. In this work, we will be using Bayesian regularized artificial neural networks (BRANNs). The benefit of using Bayesian regularization is that it is very robust in its ability to handle higher order systems with difficult to fit data. It also has the additional benefit of eliminating the computationally expensive step of model validation [17][19]. BRANNs are relatively easy to use because regularization penalizes higher order fits of the data, which protects against over fitting.

6.1 Problem Statement

Our objective is the calculation of the mole fraction y of one of the three substances along the liquid and vapor coexistence lines given the mole fraction of one of the substances x_1 , the temperature, and pressure (note that this fully defines the coexistence point since the mole fraction of the third substance, x_2 , can be calculated from $x_2 = 1 - x_1 - y$). In the present case, we treat the liquid and vapor lines separately; in other words, we are seeking to create a neural network with one output, that is, of the form

$$y = \sum_{i=1}^{N_P} w_i h_i(x_i) \equiv g(\mathbf{X}) \quad (6.1)$$

where N_P is the number of parameters (or weights), \mathbf{X} denotes the vector of input (independent) variables with elements $x_i, i = 1, \dots, N_P$, and $h_i(x_i)$ denotes the basis function associated with the i th independent variable.

6.2 Bayesian Regularized Artificial Neural Networks

The values of w may be found by minimizing the error

$$\sum_{i=1}^{N_D} [y_i - g(\mathbf{X}_i)]^2 \quad (6.2)$$

over N_D data points. Here, \mathbf{W} denotes the vector whose entries are $w_i, i = 1, \dots, N_P$.

This basic least squares approach is prone to overfitting. To mitigate this, a penalty term is added that penalizes large values of w_i . In its simplest form, this leads to the objective function

$$E(\mathbf{W}) = \sum_{i=1}^{N_D} [y_i - g(\mathbf{X}_i)]^2 + \lambda \sum_{j=1}^{N_P} w_j^2 \quad (6.3)$$

with the constraint $0 \leq \lambda \leq 1$. The solution now requires determination of the vector \mathbf{W} and the optimal value of λ . In the present work a solution is obtained via the BRANN procedure implemented in the MATLAB [®] package Neural Network

Toolbox which uses the sigmoid function as a basis, namely

$$h_i(x_i) = \frac{1}{1 + \exp(-x_i)} \quad (6.4)$$

and a slightly modified form of (6.3), namely

$$S(\mathbf{W}) = \beta \sum_{i=1}^{N_D} [y_i - g(\mathbf{X}_i)]^2 + \alpha \sum_{j=1}^{N_P} w_j^2 \quad (6.5)$$

This formulation avoids overfitting by introducing Bayes theorem which leads to explicit expressions for the hyperparameters α and β , as well as the effective number of parameters N_P (see [17] for more details). The choice of the sigmoid function is motivated by efficiency considerations. Creating a neural network that fits 100 data points takes less than 0.3% of the calculation time of a single GEMC simulation. More details can be found in References [17] and [19].

Chapter 7

Application of Machine Learning to Ternary LVE

In this Chapter, we will examine the ability of a Bayesian regularized neural network to create a model of an LVE coexistence curve given a set of data points. We also investigate how many data points are needed to accurately create such a model. For testing the ability of a neural network to reproduce LVE coexistence data, we will train the neural network using both experimental data as well as GEMC simulation results. To quantify the number of points needed, we train the neural network using a series of data sets, based on the experimental data, which contain only a small number of data points. By analyzing the error between the experimental data and the resulting predictions obtained using smaller data sets, we can estimate how many data points are needed to describe LVE coexistence without introducing significant error.

7.1 Experimental Data

Our first objective is to confirm that machine learning can make an accurate prediction about the liquid and vapor compositions of a ternary mixture under "optimal conditions", in which we have an abundance of data, and under conditions for which the LVE surface is smooth. In this case, we will attempt to build a neural network

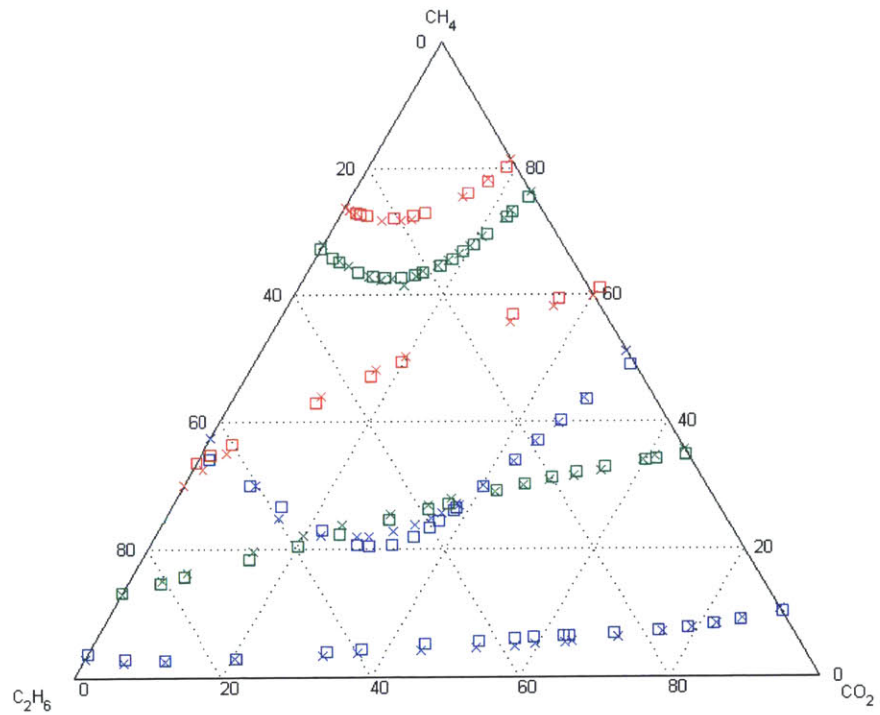


Figure 7-1: Ternary liquid-vapor equilibrium results for CO₂, CH₄, and C₂H₆ at 230K and 1.52MPa (blue) / 3.55MPa (green) / 5.57MPa (red). Experimental results (X) are from Wei et al [16]. Neural network results are denoted by □.

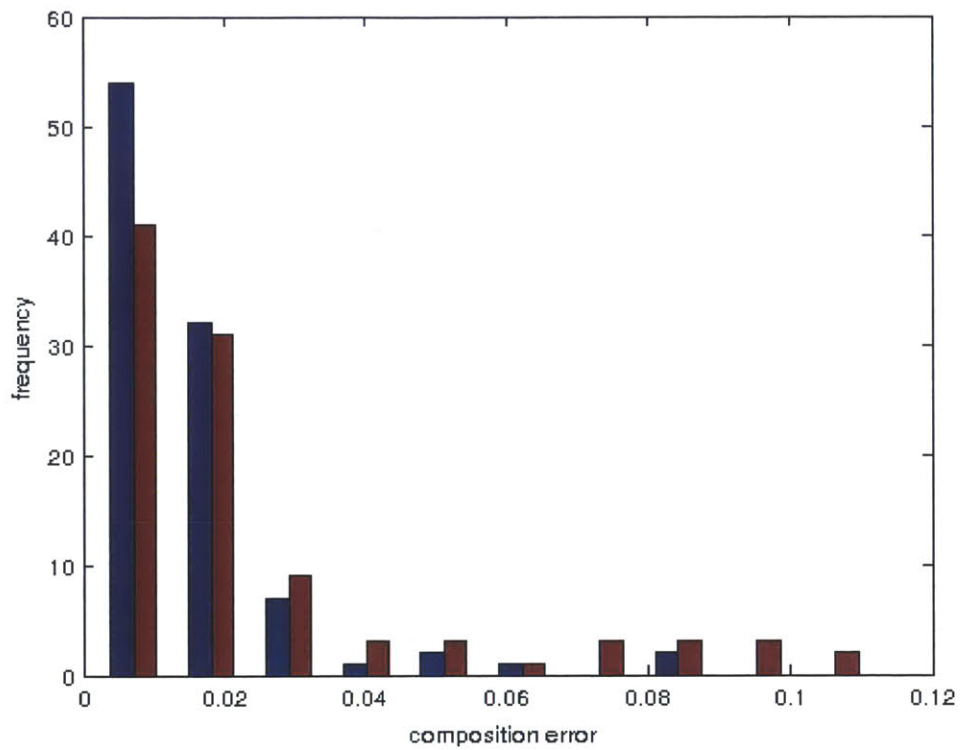


Figure 7-2: Error Histogram for the error between the experimental data and the interpolated data points, shown in Figure 7-1, as defined by equation (7.1). Blue columns indicate the error from the liquid compositions and the red column indicates the error from the vapor composition.

model of the methane, ethane, and carbon dioxide system using the experimental data obtained by Wei et al [16]. The actual neural network will be implemented using the MATLAB Neural Net Toolbox discussed in the previous Chapter. Due to the limitations in the available experimental ternary LVE composition data, we will only be considering a temperature of 230K. Using the experimental data set provides the neural network with the best chance of converging to an accurate equation of state, because the experimental data is, by comparison to the simulation data considered in section 7.2, noise free. Additionally, the experimental data also includes composition data at and near the critical points and other regions where GEMC simulations break down.

One of the benefits of training this neural network is that it will provide information on the maximum number of weights needed to represent the data over the phase space covered by the data. As discussed in Chapter 6, the use of a regularized training method prevents over fitting. Since it is impossible to over fit the neural network, we do not have an upper limit on the number of basis functions that can be used. Establishing the minimum number of weights needed to represent the data gives us the minimum number of basis functions that are needed for the neural network.

Knowing the number of basis functions needed allows us to not include unused basis functions in the neural network, improving the efficiency of future training calculations. The neural network training session that resulted in the model whose results are shown in Figure 7-1, indicated that a network with seven basis functions was sufficient to adequately model the system. Using a larger number of basis functions size does not change any of the results significantly, though some fine tuning in the number of basis functions can result in a slight improvement in the error of a neural network's fit.

Figure 7-1 illustrates how well the neural network reproduces the experimental data. As expected, the experimental results match the estimations made by the neural network almost exactly. The error in the predicted compositions needs to be quantified so that we can make comparisons to the error from the neural network. We can calculate the total error at a given point as the Euclidean distance between

the neural network estimate (\hat{x}) and the experimental result (x), namely

$$Error = \sqrt{(\hat{x}_1 - x_1)^2 + (\hat{x}_2 - x_2)^2 + (\hat{x}_3 - x_3)^2} \quad (7.1)$$

As expected, Figure 7-2 shows that the error created by using a neural network is very small. On average, the mean error in the liquid phase is 1.54% and the mean error in the vapor phase is 2.40%. The histogram also shows that there are a few outliers which result in a maximum error of 8.96% in the fluid and a maximum error of 11.20% in the vapor.

7.2 Simulation Data

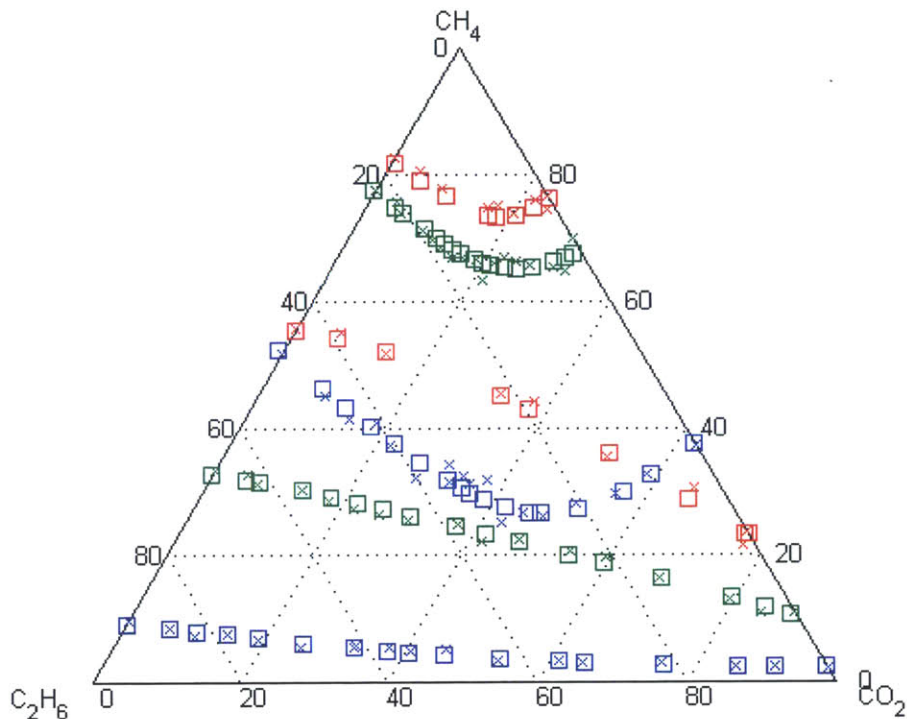


Figure 7-3: Ternary liquid-vapor equilibrium results for CO₂, CH₄, and C₂H₆ at 230K and 1.52MPa (blue) / 3.55MPa (green) / 5.57MPa (red). Experimental results (X) are from Wei et al [16]. Neural network results are denoted by □.

After confirming that neural networks can be used to create a model of the LVE, we will proceed to investigate how the method performs when relying on noisy data. Note that even though neural networks can assign a weight to the importance of each of the training data points, thus reflecting our prior knowledge of the error uncertainty associated with each of them, in our case, where the goal is to use as little a priori knowledge as possible, estimating the accuracy of the data points used to train the network becomes impractical.

For this test, we will train the neural network using GEMC simulation results that were obtained by initializing the simulations from the experimental data. The error is calculated as the Euclidean distance between the neural network estimate

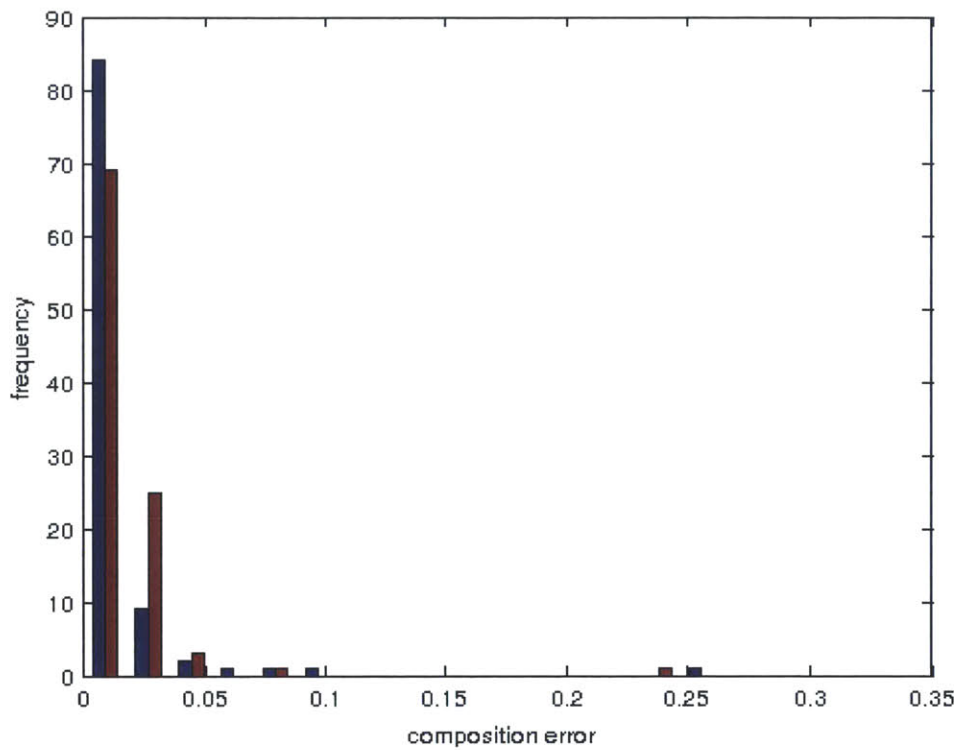


Figure 7-4: Error Histogram for the error between GEMC simulation results and the interpolated data points, shown in Figure 7-3, as defined by equation (7.1). Blue columns indicate the error from the liquid compositions and the red column indicates the error from the vapor composition.

and the corresponding GEMC simulation result. In Figure 7-4 we can see the error histogram comparing the fitting of the neural network to the simulated data it was generated from. Once again, we find that overall the error is still very small, though there are much larger outliers compared to the experimental data case of Section 7.1. This is expected since the neural network is effectively smoothing out the noise in the simulation data, which results in large errors for points if the GEMC simulation does not predict a smooth VLE surface.

On average, the mean error in the liquid phase is 1.34% and the mean error in the vapor phase is 1.78%. The histogram also shows that there are a few outliers which result in a maximum error of 26.64% in the fluid and a maximum error of 24.00% in the vapor.

Now that we have confirmed that a neural network can accurately model a nonlinear ternary LVE surface in favorable conditions with sufficient data, we can investigate the possibility of building a predictive model of the VLE of a system for which there is little or no existing experimental data. For this method to be useful, we will need to determine how many data points are needed to build a robust model.

The level of robustness, or accuracy, will depend on the application. Here, we have two applications in mind. First, we are interested in using the neural network as a source of initial conditions for GEMC simulations; this requires relatively low accuracy, although not when, the results of such simulations are to be used to enrich the original network fit. The second application is a complete LVE coexistence curve description, which requires high accuracy, namely on the order of a few percent.

7.3 Training With Few Data Points

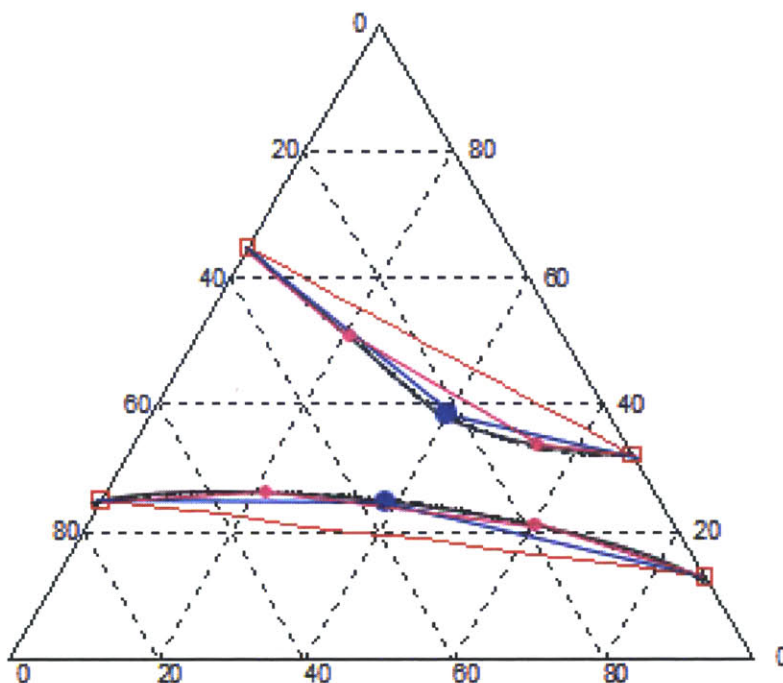


Figure 7-5: An example of the three approximations of the isothermal-isobaric LVE curves being evaluated as reduced initial data sets. The linear approximation is shown in red, the three point approximation in blue, and the four point approximation in fuchsia. For comparison the true isothermal-isobaric curves are shown in black.

To estimate how much error is introduced by using a reduced data set, we will begin with using a limited number of data points from the experimental data. We will be looking at three different cases. The first is a linear approximation of the liquid and vapor equilibrium curves connecting the binary simulation results. The second case assumes the availability of experimental data close to the midpoint of the isothermal-isobaric lines connecting the binary data points, in effect having a three point piece-wise linear approximation for each curve. The third neural network will use a four point approximation constructed from three piece-wise linear segments. For this set of tests, only experimental data is used, because we already have estimated

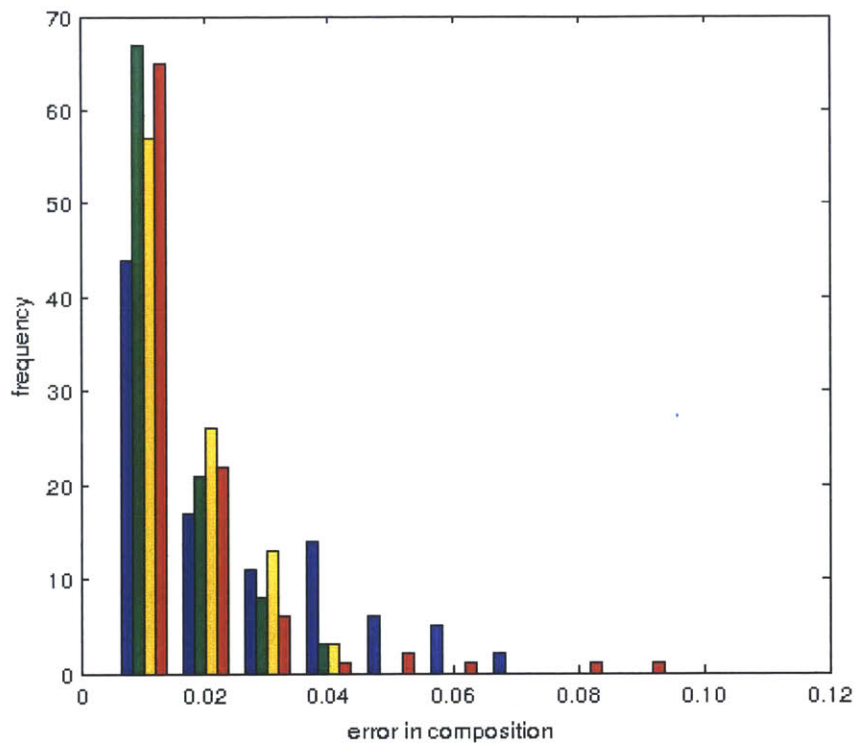


Figure 7-6: Error histogram for the liquid phase of a system composed of CO_2 , CH_4 , and C_2H_6 at 230K. The histogram shows the interpolation error distributions resulting from a linear approximation (blue), a three point approximation (green), a four point approximation (yellow), and the error associated with using the full data set (red).

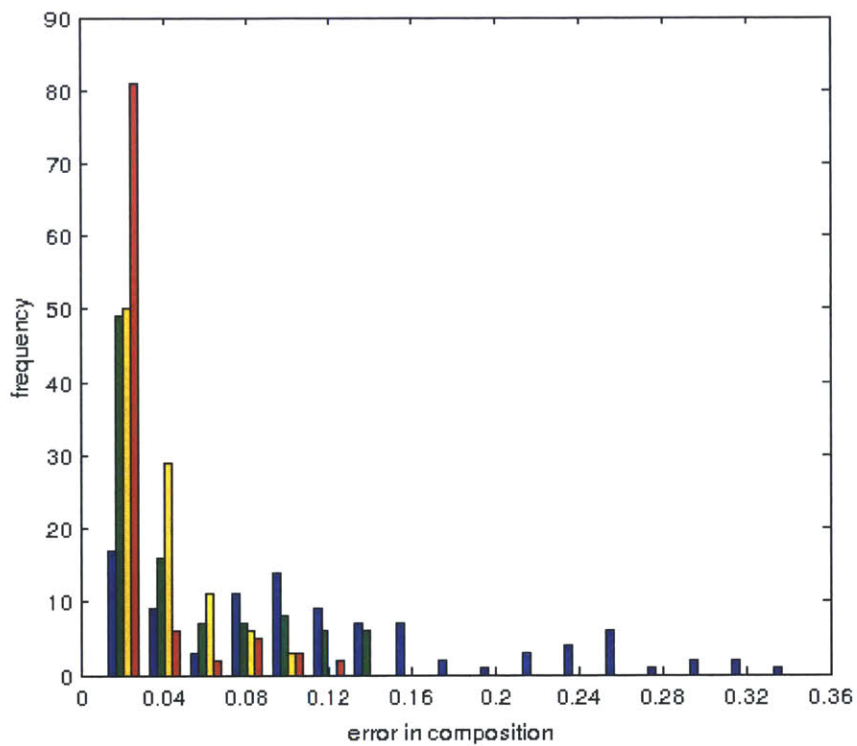


Figure 7-7: Error histogram for the vapor phase of a system composed of CO₂, CH₄, and C₂H₆ at 230K. The histogram shows the interpolation error distributions resulting from a linear approximation (blue), a three point approximation (green), a four point approximation (yellow), and the error associated with using the full data set (red).

the amount of error resulting from using GEMC simulation results. We will compare the error of the resulting neural networks to each other as well as to the error resulting from using every data point from the experimental LVE data set at 230K [16] to train a neural network. The results of this analysis are shown in Figure 7-6 for the liquid phase and in Figure 7-7 for the vapor phase.

Figure 7-6 shows that, for the liquid phase, there is little difference between the different neural networks tested. This is because the liquid curve is nearly linear. It is also notable that using the full data set to fit the neural network resulted in the worst outliers of any of the neural networks. This is a result of the neural network having difficulty fitting the critical points. In the smaller data sets the critical points are not included in the provided data.

The more definitive test, however, is the histogram in Figure 7-7 due to the wider distribution in errors. Both the three and four data point sets result in few outliers. While using a four point approximation of the isothermal-isobaric lines results in a slight reduction in the mean error, small improvements in accuracy are unlikely to provide any real benefit if the purpose of the neural network is to generate new estimates of the initial conditions for future GEMC simulations.

From this section we conclude that the three point approximation produces a sufficiently accurate representation of the LVE coexistence curves. While using additional data is unlikely to hurt the accuracy of the neural network, calculating the additional data will be too computationally expensive to justify the small improvement in accuracy.

7.4 Predicting LVE Coexistence Curves Without Experimental Data

We now have the basics for using neural networks to efficiently calculate LVE coexistence curves in a ternary mixture. We assume that the binary coexistence data for our chosen force field is either available or can be easily calculated using GEMC. This provides the necessary information to implement a linear approximation of the isobaric-isothermal lines connecting the binary LVE compositions. Here the goal is to build up from this simplistic model to one that can accurately model mixtures with complicated interactions between unlike molecules. This means not using experimental ternary data, since, in general, those cannot be assumed to be available.

In the previous section, it was determined that a three point approximation of the isobaric-isothermal connecting lines provided the best balance between accuracy and computational cost. Since our objective is to construct a neural network without using experimental ternary data, we propose to use GEMC results for obtaining the third (intermediate) point in this approximation. We further propose obtaining this data point by using a linear-interpolation-based neural network to provide an estimate for initializing the GEMC simulation.

The primary challenge associated with this scheme is determining if a potentially "poorly" initialized GEMC simulation (the one whose initial condition is given by the neural network based on the linear interpolation of the binary LVE data) has converged to a sufficiently precise solution. Because GEMC uses a fixed number of molecules of each species, making it is possible for the system to be unable to reach the correct equilibrium composition if started sufficiently far from it. The most reliable way to check for this error is to look at the distribution of the total number of particles at the end of the simulation. If the total number of molecules in each box is no longer close to its starting value, this is a strong indication that convergence was not achieved. A reasonable estimate for this threshold is a shift of 20% in the number of molecules in each box. If convergence is not reached, a new simulation can be started at the new compositions with equal numbers of molecules in each box. This can be performed

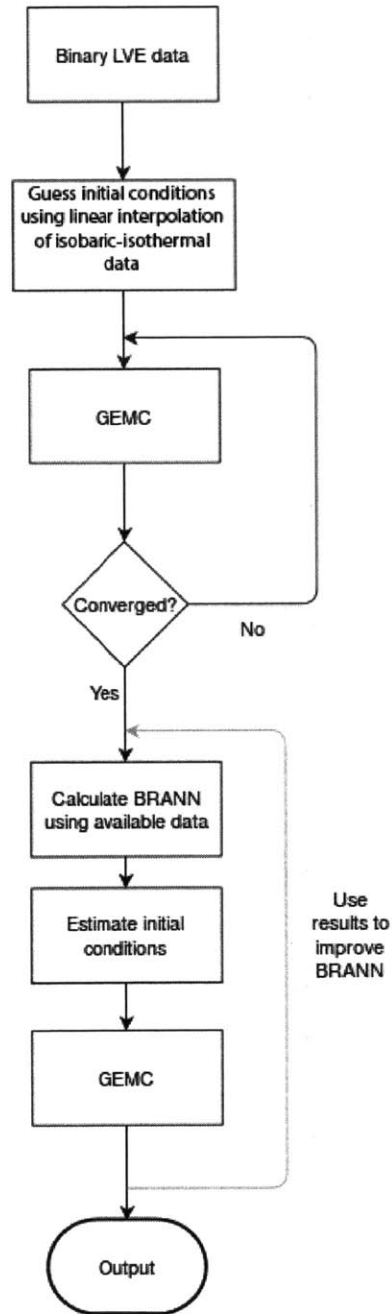


Figure 7-8: Diagram of the steps used to calculate LVE using neural networks

iteratively until the shift in the number of molecules becomes small, suggesting that the resulting compositions of the liquid and vapor phases will have converged to the correct equilibrium state. This process is time consuming but this cost is offset somewhat by the fact that the neural network can be used to provide estimates for additional initial points in the future. Using this procedure, we can calculate real coexistence compositions near the midpoint by initializing from the midpoint of the linear approximation. These calculated coexistence points and the binary LVE data can then be used to train a neural network to interpolate. This process is summarized in the flow chart in Figure 7-8.

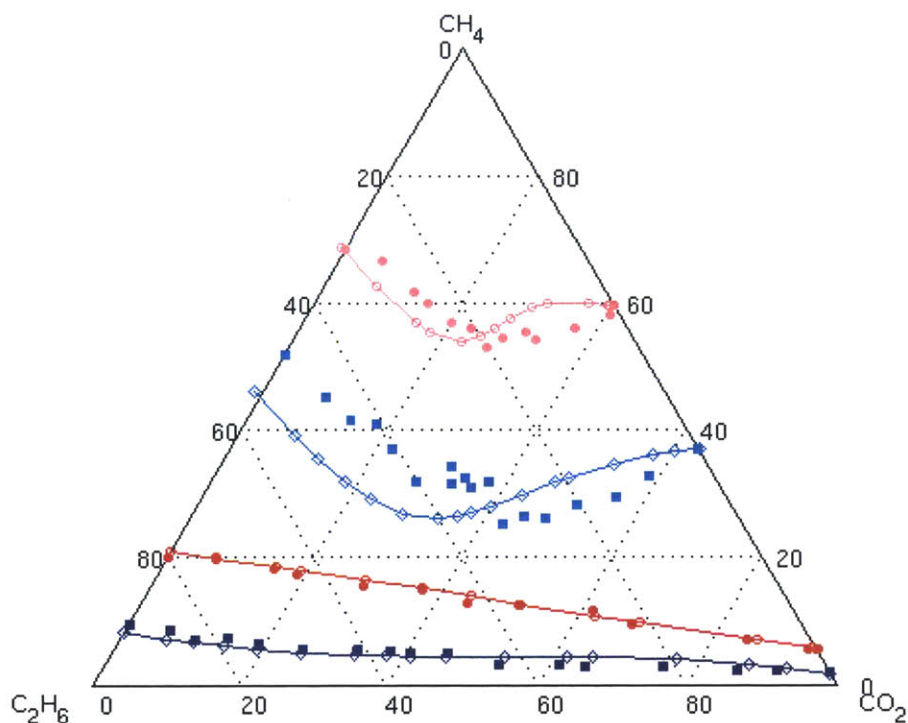


Figure 7-9: Ternary liquid-vapor equilibrium results for CO₂, CH₄, and C₂H₆ at 230K and 1.52MPa (blue) / 2.53MPa (red). Neural network results are indicated by the line with open symbols, while solid symbols indicate simulation results obtained from starting simulations initialized at experimental composition results from Wei et al [16]. Dark shades correspond to the liquid phase and light shades correspond to the vapor phase.

Figure 7-9 shows the results from a neural network that was trained to predict

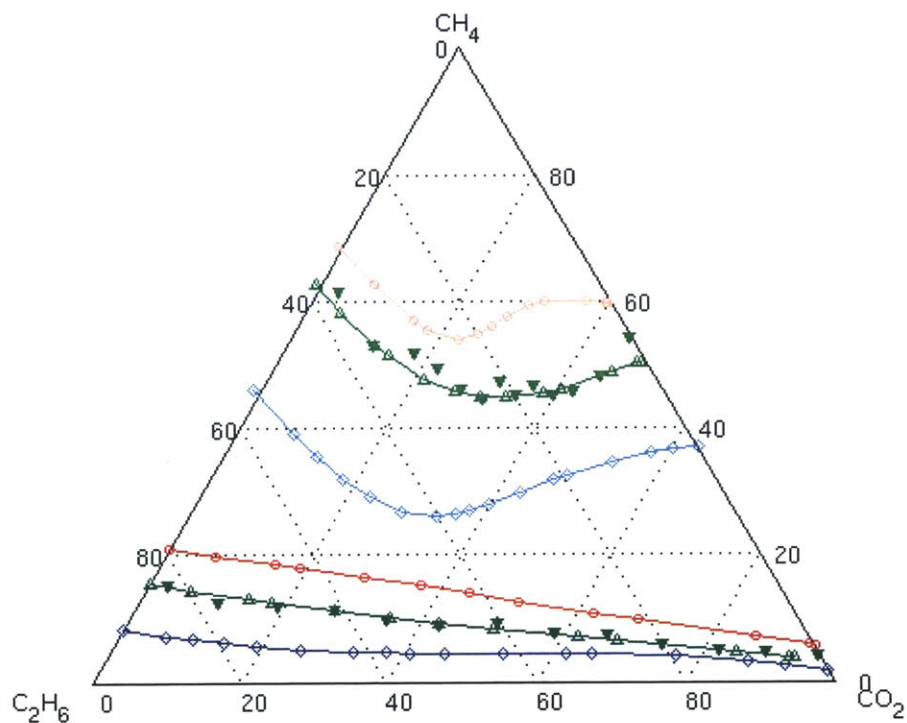


Figure 7-10: Ternary liquid-vapor equilibrium results for CO_2 , CH_4 , and C_2H_6 at 230K. The red and blue markers are the same as in Figure 7-9 and shown for reference. The green line and empty symbols indicate the estimated composition at a temperature of 230K and a pressure of 2.03 MPa. The solid symbols were obtained from GEMC simulations initialized at the interpolated data points. Dark shades correspond to the liquid phase and light shades correspond to the vapor phase.

the LVE based on the given temperature, pressure, and as well as the liquid mole fraction of CO₂ from GEMC simulation results at those same conditions. The neural network used for this figure was trained using a three point approximation of the isothermal-isobaric lines, where the LVE properties were calculated using GEMC simulations, that is, in other words, using the procedure just outlined. These results show that even if we use a very limited data set to train the neural network, it is possible to obtain reasonably accurate predictions.

Additionally as demonstrated in Figure 7-10, using this neural network we can make predictions that are not at the same temperature and pressure as the data that was originally used to create the model. This figure was created using the same neural network as Figure 7-9. The neural network was used to make predictions of the LVE at a pressure of 2.03 MPa, for which no experimental data is available. These predictions were used to initialize GEMC simulations to check the accuracy of the predictions. The GEMC simulation results were all in good agreement with the predicted values. This shows that the neural network can be used to make predictions not just on the same isothermal-isobaric lines, but also make predictions at other points on the LVE surface, given that the points temperature and pressure is bounded by known coexistence data included in the neural network (that is, the neural network is used to interpolate - as opposed to extrapolate - data).

Chapter 8

Conclusion

Machine learning can be very effective at reducing the amount of data necessary to accurately calculate the liquid-vapor equilibrium surface of a ternary mixture. Using a Bayesian regularized neural network allows for the robust and accurate calculation of a model for highly dimensional systems with many inputs and many outputs. While Bayesian regularization is considered an expensive method for training a neural network, its computational cost is roughly two orders of magnitude less than performing a single GEMC simulation. The comparatively low cost of fitting a highly accurate neural network means that it has the potential to significantly reduce the cost of calculating the LVE surface using GEMC simulations alone.

The neural network models calculated in Figures 7-9 and 7-10 demonstrate that from a fairly limited set of coexistence data points it is possible to estimate initial compositions that, from the standpoint of a GEMC simulation, are indistinguishable or are nearly indistinguishable from using the experimental data as the initial guess. In Figures 7-9 and 7-10 we have used just three data points per isobaric isothermal coexistence line. It is clear from our results that while the system does not necessarily predict the LVE surface with complete accuracy, it is close enough that any further GEMC simulations in that region of phase space would converge to a correct result without the need for iteration.

This method does have some limitations. Since the model is interpolated, it is only accurate within the phase space covered by the data used to train the network. As

such, the network cannot be used for extrapolation of properties into regions without any data. This is somewhat alleviated by the fact that the binary coexistence data is relatively simple to calculate, meaning that it should be relatively easy to extend to new temperatures and pressures. It also means that this method will never be able to provide the data at a critical point unless data for nearby points was included in the training data.

The methodology described here can in principle be extended to mixtures involving more components. The largest challenge presented by increasingly complex mixtures is the increasing volume of the state space. This makes it much more difficult to obtain the initial data needed to train the neural network. In the example of a quaternary mixture, one would need to first approximate the LVE coexistence curves for four different ternary mixtures before using the method described to interpolate between the results for the ternary mixtures to obtain the initial data points for the quaternary neural network. While this would be more computationally expensive, we do not foresee any fundamental limitations for performing either the GEMC simulations or training the neural network.

Bibliography

- [1] Bruce, A.D. and Wilding, N.B., "Computational Strategies for Mapping Equilibrium Phase Diagrams", *Adv. Chem. Phys.*, 127, 1-64 (2003).
- [2] Williams, C.K.I., "Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond." *Learning in Graphical Models*. Ed. Michael I. Jordan. Springer Netherlands, 599-621 (1998).
- [3] van't Hof, A., Peters, C.J., and de Leeuw, S.W., "An Advanced Gibbs-Duhem Integration Method: Theory and Applications", *J. Chem. Phys.*, 124, 054906 (2006).
- [4] Panagiotopoulos, A.Z., "Monte Carlo Methods for Phase Equilibria of Fluids", *J. Phys.: Condens. Matter*, 12, R25-R52 (2000).
- [5] Panagiotopoulos, A.Z., "Direct Determination of Phase Coexistence Properties of Fluids by Monte Carlo Simulation in a New Ensemble", *Mol. Phys.*, 61 813-826 (1987).
- [6] Smit, B., de Smedt, Ph., and Frenkel D., "Computer Simulations in the Gibbs Ensemble", *Mol. Phys.*, 68, 931-950 (1989).
- [7] Smit, B. and D. Frenkel., "Calculation of the Chemical Potential in the Gibbs Ensemble", *Mol. Phys.*, 68, 951-958 (1989).
- [8] Ferrenberg A.M. and Swendsen R.H., "New Monte Carlo Technique for Studying Phase Transitions", *Phys. Rev. Lett.*, 63, 1195-1198 (1989).
- [9] Rowlinson, J.S. and Swinton, F. L., *Liquids and Liquid Mixtures* (3rd ed.). Butterworth (1982).
- [10] Potoff, J.J. and Panagiotopoulos, A.Z., "Critical Point and Phase Behavior of the Pure Fluid and a Lennard-Jones Mixture", *J. Chem. Phys.*, 109, 10914-10920 (1998).
- [11] Martin, M.G. and Siepmann, J.I. "Transferable Potentials for Phase Equilibria. 1. United-Atom Description of N-Alkanes, *J. Phys. Chem. B*, 102, 2569-2577 (1998).
- [12] Chen, B. and Siepmann, J.I. "Transferable Potentials for Phase Equilibria. 3. Explicit-Hydrogen Description of Normal Alkanes", *J. Phys. Chem. B*, 103, 5370-5379 (1999).

- [13] Potoff, J.J. and Siepmann, J.I., "Vapor-Liquid Equilibria of Mixtures Containing Alkanes, Carbon Dioxide, and Nitrogen", *AIChE J.*, 47, 1676-1682 (2001).
- [14] Ungerer, P., Beauvais, C., Delhommelle, J., Boutin, A., Rousseau, B. "Optimization of the Anisotropic United Atoms Intermolecular Potential for N-Alkanes", *J. Chem. Phys.*, 112, 5499-5510 (2000).
- [15] Younglove, B.A. and Ely, J.F., "Thermophysical Properties of Fluids. II. Methane, Ethane, Propane, Isobutane, and Normal Butane" *J. Phys. Chem. Ref. Data*, 16, 577-798 (1987).
- [16] Wei, M.S.W., Brown, T.S., and Kidnay, A.J.K. "Vapor + Liquid Equilibria for the Ternary System Methane + Ethane + Carbon Dioxide at 230K and Its Constituent Binaries at Temperatures from 207 to 270K", *J. Chem. Eng. Data*, 40, 726-731 (1995).
- [17] Burden, F. and Winkler, D. "Bayesian Regularization of Neural Networks", *Method. Mol. Bio.*, 458, 23-42 (2009).
- [18] Biswas, A.C., "The Law of Rectilinear Diameter for the Liquid-Gas Phase Transition", *Pramana*, 1, 109-111 (1973).
- [19] Buntine, W.L. and Weigend, A.S., "Bayesian Back-Propagation", *Complex Systems*, 5, 603-643 (1991).
- [20] DeHoff, R. *Thermodynamics in Materials Science*. Taylor and Francis (2006).
- [21] Nabney, I.T. *Netlab: Algorithms for Pattern Recognition*. Springer-Verlag, London (2002).
- [23] Danesh, A. *PVT and Phase Behaviour of Petroleum Reservoir Fluids*. Elsevier (1998).
- [24] Hastings, W.K., "Monte Carlo Sampling Methods Using Markov Chains and Their Applications", *Biometrika*, 57, 97-109 (1970).
- [25] Toxvaerd, S., "Equation of State of Alkanes II", *J. Chem. Phys.*, 107, 5197-5204 (1997).