QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP GENERALIZED ACTIVITY COEFFICIENT MODELS

By

SOLOMON GEBREYOHANNES

Bachelor of Science
Bahir Dar University
Bahir Dar, Ethiopia
2007

Master of Science
Oklahoma State University
Stillwater, Oklahoma
2010

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 2014

QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP GENERALIZED ACTIVITY COEFFICIENT MODELS

Dissertation Approved:

Dr. Khaled A. M. Gasem

Dissertation Adviser

Dr. Robert L. Robinson, Jr.

Dr. Josh D. Ramsey

Dr. Martin Hagan

ACKNOWLEDGEMENTS

First of all, I would like to express my deepest appreciation to my advisor, Professor Khaled A. M. Gasem, for his continuous guidance, support and encouragement throughout my PhD study. I am very grateful and fortunate to have Dr. Gasem as my advisor and teacher. I deeply appreciate his valuable advice on research as well as on my career.

I owe my heartfelt gratitude to Dr. Robert L. Robinson for his guidance and interest in my research work. I am very thankful for his insightful comments and suggestions. I would also like to thank my committee members, Dr. Josh D. Ramsey and Dr. Martin Hagan, for their supervision and constructive comments.

I would like to thank Dr. Brian Neely for his invaluable advice, support and significant contribution. I would also like to thank Dr. Sayeed Mohammad for his advice and helpful discussions. A special thanks to all my colleagues, Dr. Krishna Yerramsetty, Dr. Eric Whitebay, Dr. Younas Dadmohammadi, Christian Odafin, Agelia Abudour, Pongtorn Charoensuppanimit, Vidhya Venugopal and Menelik Negash for their support and valuable inputs. I would also like to thank all my friends for their friendship making my time at OSU full of fun and good memories.

Finally, I would like to thank all my families for their constant love, support and encouragement throughout my graduate study.

Acknowledgements reflect the views of the author and are not endorsed by committee members or Oklahoma State University.

Name: SOLOMON GEBREYOHANNES

Date of Degree: MAY, 2014

Title of Study: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP GENERALIZED ACTIVITY COEFFICIENT MODELS

Major Field: CHEMICAL ENGINEERING

Abstract:

Phase behavior properties of chemical species and their mixtures are essential to design chemical processes involving multiple phases. Thermodynamic models are used in phase equilibria calculations to determine properties, such as phase compositions and partition coefficients at specific temperatures and pressures. In the absence of experimental data, generalized models are employed to predict phase equilibria properties.

The two main objectives of this study are to (1) develop improved generalized models for vapor-liquid equilibria (VLE) and liquid-liquid equilibria (LLE) property predictions using a theory-framed quantitative structure-property relationship (QSPR) modeling approach and (2) implement a new modification to the widely used nonrandom two-liquid (NRTL) activity coefficient model to reduce parameters correlation, which is a limitation of the original model.

In this work, we assembled two databases consisting of 916 binary VLE and 342 binary low-temperature LLE data. Data regression analyses were performed to determine the interaction parameters of various activity coefficient models. Structural descriptors of the molecules were generated and used in developing QSPR models to estimate the regressed interaction parameters. The developed QSPR models for VLE systems provided phase equilibria property predictions within twice the errors obtained through the data regression analyses for VLE systems. For LLE systems, the QSPR models resulted in approximately three to four times the errors found from the regression analyses. Further, our methodology provides *a priori* and easily implementable QSPR models with a wider applicability range than that of the group-contribution model, UNIFAC.

The newly modified model proposed in this work reduced the NRTL model to a one-parameter model and eliminated the parameter correlation. The original and modified NRTL models yield comparable accuracies in representing experimental equilibrium properties. The benefits of our modification include easy generalizability of the parameters, ability to classify VLE behaviors based on a single model parameter and fewer convergence problems in parameter regressions.

TABLE OF CONTENTS

CHAPTER PAGE

CHAPTER PAGE

LIST OF TABLES

TABLE PAGE

TABLE PAGE

TABLE PAGE

LIST OF FIGURES

FIGURE PAGE

FIGURE PAGE

CHAPTER I

INTRODUCTION

1.1. Rationale

Phase equilibrium properties, such as pressure, temperature, compositions, partition coefficients, etc., are required for designing and optimizing separation processes and numerous other unit operations encountered in the chemical industry. These properties are typically determined from experimental measurements; however, conducting experiments requires a substantial investment of money and time. Predicting phase equilibrium properties using reliable generalized models offers an attractive alternative to costly and time consuming experimental measurements.

In phase equilibria calculations, activity coefficient (γ) models are used to account for liquid mixture deviations from ideal behavior. A number of activity coefficient models for predicting vapor-liquid equilibria (VLE) and liquid-liquid equilibria (LLE) have been proposed by various researchers [1-5]. These models provide frameworks that relate activity coefficients with composition and temperature properties. In general, the literature models can be classified as historical and semi-empirical activity coefficient models (Margules [6], Redlich-Kister [6] and Van Laar [6]), theory-based models, which includes local composition and two-liquid models (Wilson [7], NRTL [1] and UNIQUAC [3]) and group-contribution models (UNIFAC [2], ASOG [8]). Activity coefficient models, such as NRTL, UNIQUAC and Wilson require two or three adjustable interaction parameters that are determined through regression of experimental data for a specific system.

Thus, they cannot be applied to predict properties of vapor-liquid equilibrium (VLE) systems for which experimental data are not available.

Although a number of activity coefficient models are reported in the literature, their use is limited by the availability of experimental data. Efforts to minimize the need for experimental data through the development of *a priori* predictive models are on-going [2, 8-10]. Traditionally, to facilitate *a priori* predictions, group-contribution models such as UNIQUAC functional-group activity coefficients (UNIFAC) and analytical-solution-of-groups (ASOG) [2, 8] have been employed to generalize the UNIQUAC and Wilson models, respectively. Models based on quantum chemical calculations such as the conductor-like screening model for real solvents (COSMO-RS) [9, 11] are also used for *a priori* predictions purposes. The UNIFAC parameter matrix published in 2006 [12] has over 4,000 parameters including surface area (q), volume contribution (r) values of 115 sub group and main group interaction (a_{ij}, b_{ij} and c_{ij}) values of 659 interactions.

Despite their potential benefits, group-contribution models suffer from limitations such as the inability to define effectively the functional groups of some chemical species and a lack of model interaction parameters for functional groups that are not represented in the UNIFAC data matrix. In contrast, the COSMO-RS model is more universal compared to the UNIFAC model since COSMO-RS relies on individual chemical elements as opposed to functional groups. For some polar systems, however, the COSMO-RS model results in worse predictions than the UNIFAC model [13] and moreover, sometimes fails to describe the VLE of even nearly-ideal organic systems [13, 14]. Therefore, a need exists for developing accurate and less computationally demanding models capable of *a priori* prediction of equilibria properties.

This work is focused on developing improved generalized models for VLE and LLE property predictions using a theory-framed quantitative structure-property relationship (QSPR) modeling approach. In this approach, theoretical frameworks are used to develop the behavior models, and

QSPR techniques to generalize the substance-specific parameters of the models. Our analysis shows, using a theory-framed QSPR modeling approach, interaction model parameters of various activity coefficient can be generalized for VLE and LLE mixtures. Further, our findings show theory-framed QSPR modeling provides comparable or better accuracy than the available *a priori* models, such as the UNIFAC model.

1.2. Objectives

The goal of this work is to generalize the widely used activity coefficient models using a theory-framed QSPR modeling approach for VLE and LLE binary systems. The following are the four specific objectives that were undertaken to accomplish this goal.

1. Database development

- Assemble VLE and LLE databases with a wide representation of various functional groups and categorize the binary systems based on chemical classes and phase equilibrium behaviors.

2. Behavior representation assessment

- Evaluate the abilities of various activity coefficient models to represent different types of fluid phase behavior using the assembled VLE and LLE databases.
- Assess the behavior representation qualities of the various models for systems encountered in refining and in bi-phasic reactors.

3. QSPR model development:

- Develop improved QSPR generalizations for the interaction parameters of various activity coefficient models applicable to VLE systems.
- Extend the modeling approach to generalize activity coefficient models for LLE systems.

- Perform a rigorous validation of the models using an external test set.

4. Theoretical advancement

- Advance the theory of the current activity coefficient models, such as NRTL, to reduce or avoid the effect of correlation between the model parameters.

This research work provided generalized models for the estimation of interaction parameters of widely used activity coefficient models for VLE and LLE systems. The generalized model predictions are beneficial in reducing the experimentation costs needed for determining phase equilibria properties. In addition, the model resulted in improved generalized property predictions for designing, optimizing and simulating various chemical processes encountered in oil and gas industry. Further application includes providing phase behavior properties of candidate molecules in computer-aided molecular design (CAMD) processes.

1.3. Thesis organization

This work is organized in “manuscript style” and is divided into five stand-alone chapters. Chapter 1 presents the rational and the objectives of this work. Chapter 2 focuses on a QSPR generalized NRTL model for 578 VLE systems (case studies on refining and bi-phasic catalytic systems). Chapter 3 deals with comparison of QSPR generalized UNIQUAC, NRTL, Wilson and UNIFAC models for VLE property predictions. Chapter 4 centers on QSPR generalized NRTL and UNIFAC models for LLE property predictions. Chapter 5 presents modified version of two and one parameter NRTL models for prediction of VLE and LLE properties. Chapter 6 concerns QSPR generalization of the modified one-parameter NRTL model. The final chapter presents a summary of key conclusions drawn from each chapter and potential recommendations for future research. The manuscripts were developed in chronological sequence over a period of three years.

REFERENCES

1. Renon, H. and J.M. Prausnitz, *Local compositions in thermodynamic excess functions for liquid mixtures.* AIChE Journal, 1968. **14**(1): p. 135-144.

2. Gmehling, J., J. Li, and M. Schiller, *A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties.* Industrial & Engineering Chemistry Research, 1993. **32**(1): p. 178-193.

3. Abrams, D.S. and J.M. Prausnitz, *Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems.* AIChE Journal, 1975. **21**(1): p. 116-128.

4. Skjold-Jorgensen, S., B. Kolbe, J. Gmehling, and P. Rasmussen, *Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(4): p. 714-722.

5. Fischer, K. and J. Gmehling, *Further Development, status and results of the PSRK method for the prediction of vapor-liquid equilibria and gas solubilities.* Fluid Phase Equilibria, 1996. **121**(1-2): p. 185-206.

6. Prausnitz, J.M., R.N. Lichtenthaler, and E.G.d. Azevedo, *Molecular thermodynamics of fluid-phase equilibria*. 3rd ed. ed. 1998: Prentice-Hall.

7. Wilson, G.M., *Vapor-liquid equilibrium. XI. A new expression for the excess free energy of mixing.* Journal of the American Chemical Society, 1964. **86**(2): p. 127-130.

8. Gmehling, J., D. Tiegs, and U. Knipp, *A comparison of the predictive capability of different group contribution methods.* Fluid Phase Equilibria, 1990. **54**: p. 147-165.

9. Klamt, A. and F. Eckert, *COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids.* Fluid Phase Equilibria, 2000. **172**(1): p. 43-72.

10. Holderbaum, T. and J. Gmehling, *PSRK: A group contribution equation of state based on UNIFAC.* Fluid Phase Equilibria, 1991. **70**(2–3): p. 251-265.

11. Klamt, A., V. Jonas, T. Bürger, and J.C.W. Lohrenz, *Refinement and parametrization of COSMO-RS.* The Journal of Physical Chemistry A, 1998. **102**(26): p. 5074-5085.

12. Jakob, A., H. Grensemann, J. Lohmann, and J. Gmehling, *Further Development of modified UNIFAC (Dortmund): Revision and extension 5.* Industrial & Engineering Chemistry Research, 2006. **45**(23): p. 7924-7933.

13. Lei, Z., B. Chen, C. Li, and H. Liu, *Predictive molecular thermodynamic models for liquid solvents, solid salts, polymers, and ionic liquids.* Chemical Reviews, 2008. **108**(4): p. 1419-1455.

14. Ravindranath, D., B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, *QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

CHAPTER II

IMPROVED QSPR GENERALIZED INTERACTION PARAMETERS FOR THE NRTL ACTIVITY COEFFICIENT MODEL

2.1. Introduction

Accurate prediction of the phase behavior properties of chemical species and their mixtures is essential for designing and optimizing separation processes and numerous other unit operations encountered in the chemical industry. Predicting phase equilibrium properties, such as phase compositions and partition coefficients at temperatures and pressures of interest, using reliable models offers a more attractive alternative to costly and time consuming experimental measurements.

Within the Gibbsian framework, vapor-liquid equilibrium (VLE) properties are determined using two widely used approaches. The first is the (ϕ/ϕ) approach, where fugacity coefficients (ϕ) for each component in the vapor and liquid phases are calculated using an equation-of-state (EOS) model. The second technique is the split approach (ϕ/γ), where different models are used to predict deviations from ideal behavior. Here, fugacity coefficients and activity coefficients (γ) are used to account for non-ideal behavior in the vapor and liquid phases, respectively. Fugacity coefficients are determined using various EOS models and activity coefficients are calculated using excess Gibbs energy ($\overline{G^E}$) models.

An extensive list of EOS and $\overline{G^E}$ models has been developed over the years originating from different theories to address the needs in various applications [1]. Multiple researchers have suggested various forms of $\overline{G^E}$ based mixing rules to improve the predictions of the EOS models [2-5]. Accurate descriptions of phase behavior, however, remain largely reliant on the availability of VLE experimental data of the targeted systems. Efforts to minimize the need for experimental data through the development of *a priori* predictive models are on-going [6-8]. However, to date, both the EOS models and $\overline{G^E}$ models have limited capabilities for accurate *a priori* predictions.

A better approach for *a priori* predictions of activity coefficients was demonstrated by group-contribution models such as UNIQUAC functional activity coefficient (UNIFAC) and analytical solution of groups (ASOG) [6, 7]. These models are based on functional-group interactions. Since the number of functional groups is much smaller than the number of compounds, a large number of mixtures can be generalized using a smaller number of functional-group interactions [6].

Despite their potential benefits, group-contribution models suffer limitations including an inability to account for the effects of neighboring molecules [9]. Further, the models are only applicable for mixtures consisting of compounds for which functional groups are contained in the UNIFAC data matrix. If the functional groups of interest are not present in the data matrix of UNIFAC, experimental data are required to determine the interaction parameters. Another limitation is the inability to define effectively the functional groups of some chemical species. Detailed reviews for some of the other available generalized activity coefficient models were presented in our previous works [10-12].

The current success of the group-contribution models notwithstanding, a need exists for developing models capable of *a priori* prediction of VLE properties. The current research is an improvement on our previous work [10], where we generalized the system-specific parameters of the nonrandom two-liquid (NRTL) activity coefficient model using a quantitative structure-property relationship

(QSPR) modeling approach. In that initial study [10], 332 binary systems commonly encountered in refinery processes were used to develop two independent QSPR models to predict the two adjustable parameters (a_{12} and a_{21}) in the NRTL model. The non-randomness factor (α_{12}) was set as 0.2 in the previous study [10]. The QSPR-NRTL model parameter generalizations, on-average, have yielded predictions within three times the experimental uncertainties, which represented an improvement over the UNIFAC [7] group-contribution predictions. These good results aside, two issues remained of concern. First, having two separate models for the two NTRL parameters (a_{12} and a_{21}) could result in different parameter values for a specific binary system, depending on the order of components. Second, the database used in the model generalizations was not sufficiently diverse to be representative of the wide array of systems encountered in the chemical industry.

As such, we have a two-fold motivation to undertake the current work. First, we sought to eliminate the potential inconsistency resulting from two separate models for the two NRTL parameters. Second, we wanted to use a more representative database in our model generalization. To address these concerns, a computational strategy was implemented to develop a single QPSR model for the two NRTL model parameters, and a more diverse database encompassing a wide range of functional groups was assembled for the task.

Two case studies were conducted to investigate the predictive capabilities of the proposed QSPR model. In the first case, we examined the predictive capabilities of the generalized model as it applies to the binary systems from the previous database [10], which was focused on systems encountered in refining. The second case study was concerned with mixtures formed in the refining process of pyrolysis oil using bi-phasic reaction processes. Bi-phasic reaction processes use nanoparticle catalysts to selectively catalyze target reactions in organic and aqueous phases [13]. The latter case study was of particular interest because of the growing focus in bi-phasic reaction processes to upgrade pyrolysis oil as well as the diversity of the molecular species encountered in these processes.

2.2. NRTL activity coefficient model

The NRTL equation was developed by Renon and Prausnitz [14] based on the local composition theory of Wilson and the two-liquid solution theory of Scott. Unlike Wilson's equation, the NRTL equation is applicable to partially miscible as well as completely miscible systems [15]. The NRTL activity coefficients of a binary system are given as:

$$ln\,\gamma_1 = x_2^2\left[\tau_{21}\left(\frac{G_{21}}{x_1 + x_2 G_{21}}\right)^2 + \frac{\tau_{12}G_{12}}{(x_2 + x_1 G_{12})^2}\right] \tag{2.1}$$

$$ln\,\gamma_2 = x_1^2\left[\tau_{12}\left(\frac{G_{12}}{x_2 + x_1 G_{12}}\right)^2 + \frac{\tau_{21}G_{21}}{(x_1 + x_2 G_{21})^2}\right] \tag{2.2}$$

where τ_{ij} and G_{ij} are defined as:

$$G_{12} = \exp(-\alpha_{12}\tau_{12}) \qquad G_{21} = \exp(-\alpha_{12}\tau_{21}) \tag{2.3}$$

$$\tau_{12} = \frac{g_{12}-g_{22}}{RT} = \frac{a_{12}}{T} \qquad \tau_{21} = \frac{g_{21}-g_{11}}{RT} = \frac{a_{21}}{T} \tag{2.4}$$

where g_{ij} is an energy parameter characterizing interactions between i and j molecules, α_{12} is the non-randomness factor in the mixture, x_i is the mole fraction of component i, R is the universal gas constant and T is the mixture temperature.

The NRTL equation contains three parameters (defining $\alpha_{ij} = \alpha_{ji}$) that are specific for each binary system. These adjustable parameters are a_{12} or ($g_{12} - g_{22}$), a_{21} or ($g_{21} - g_{11}$), and α_{12}. To be consistent with the DECHEMA LLE database [16] in accommodating liquid-liquid equilibrium systems, the non-randomness factor (α_{12}) was kept constant as 0.2 for all binary systems in this work. We have also investigated the effect of variation of α_{12} on VLE property predictions. Our findings show that variation in α_{12} has little influence in reducing overall prediction errors; moreover, fixing its value has an obvious benefit in reducing parameter correlation. Therefore, we decided to retain a value of 0.2.

Experimental data are usually required to regress the values of the two energy interaction parameters. Therefore, the model cannot be applied directly for systems with no experimental data, and hence, there is a need for a generalized model to estimate the interaction parameters of binary systems *a priori*.

2.3. QSPR methodology

The main elements of the QSPR model development include: (a) database development and regression analysis, (b) structure generation and optimization, (c) molecular descriptor generation, (d) descriptor reduction, and (e) QSPR model development using neural networks. The modeling process starts by compiling a reliable database from credible sources. Next, the structures of components of each system are generated and optimized to find the 3-dimensional (3-D) conformation with the least energy. The optimized molecules are then used to generate 2-D and 3-D descriptors using software such as Dragon [17]. The large number of generated molecular descriptors must now be reduced to find the most significant descriptors for accurate property predictions. Simultaneously, neural network models are developed using the best descriptors. Finally, model interpretation is employed to understand the relationships between the inputs and the outputs of the network. These different elements are described in greater detail below.

2.3.1. Database development

The predictive capability of a QSPR model strongly depends on the accuracy of the experimental data used in the model development process. The VLE data used in this work were collected from several sources. Binary systems with sufficient representation of different functional groups have been included in the database. The general database and the two specialized databases are described in greater detail below.

General database (All binary systems): A low-pressure binary VLE database (Oklahoma State University, OSU database) consisting of 188 binary VLE systems totaling 4716 data points was assembled. This database is comprised of systems of aliphatic and aromatic hydrocarbons, water, alcohols, ethers, sulphides and nitrile compounds. A second database, comprised of 390 binary VLE systems totaling 12,010 data points, was taken from the DECHEMA VLE database [18]. In total, the database compiled in this work consists of a total of 578 binary systems formed from various combinations of 139 different compounds. A total of over 16,500 vapor-liquid equilibrium data points were assembled in the final database (OSU Database II). The data covered a temperature range from 215.15 to 554 K and pressures to 58 bar. However, over 99% of the data were at pressure of less than 10 bar. The pure-component vapor pressure data were taken from DIPPR [19] and DECHEMA [18] databases.

The compounds present in the OSU Database II were classified in a similar manner as the UNIFAC functional-group classification approach [7]. The database is composed of compounds belonging to 31 chemical classes. Figure 1.1 illustrates the data distribution of the binary systems in the OSU Database II based on chemical classes.

Refining systems database: This sub-set database which was adopted from the previous study by Ravindranath *et al.* [10], consists of binary systems that are commonly encountered in refining processes. In this database, 332 binary systems comprising various combinations of 92 compounds are considered. These compounds contain 28 of the 31 chemical classes that are represented in the database. Over 9700 VLE data points at different temperatures were assembled in this database, and a detailed database assessment can be found in a previously published article [10].

Bi-phasic database (compounds formed in bi-phasic reactions): This sub-set database consists of eight compounds that are formed in bi-phasic catalytic reactions. These compounds represent 6 of the 31 chemical classes in the current database. The chemical classes include alcohols, aldehydes, alkanes, furfural, ketones and water. The database is composed of 127 binary systems formed by different combinations of these compounds, and approximately 2800 data points have been assembled in the database. In Figure 2.1, the data shaded in grey are systems consisting of the compounds that are formed in bi-phasic reactions. The figure also shows the number of available binary systems of this sub-set.

2.3.2. NRTL parameter regression methodology

To determine the optimum values of the two adjustable parameters in the NRTL model, a regression analysis using an equal-fugacity equilibrium framework was performed. Specifically, the following equilibrium criteria were applied for the coexisting liquid and vapor phases, subject to mass balance constraints:

$$P^v = P^l$$

$$T^v = T^l \tag{2.5}$$

$$\hat{f}_i^v = \hat{f}_i^l \qquad i = 1,\ldots,N$$

where $\hat{f}_i$ is the fugacity of component i in the mixture, T is the temperature, P is the pressure, and the superscripts, v and l, indicate vapor and liquid, respectively. In the regression analyses, the traditional split approach was employed for VLE system modeling:

$$\hat{\phi}_i^V P y_i = \gamma_i P_i^\circ \phi_i^V x_i \lambda_i \tag{2.6}$$

where for any component i, x_i is the liquid mole fraction, y_i is the vapor mole fraction, $\hat{\phi}_i^V$ is the component fugacity coefficient in the vapor phase, γ_i is the component activity coefficient in the liquid phase, P_i° is the pure-component vapor pressure, ϕ_i^V is the pure-component fugacity coefficient in the vapor phase and λ_i is the Poynting factor. Practically, all the VLE systems considered in this study were at low pressure; hence, the vapor-phase fugacity coefficients were assumed to be 1. We have also investigated the quality of property representation when equation-of-state (EOS) models are used to calculate the vapor-phase fugacity coefficients (results not shown). Our findings show there is no improvement on the overall representation error. This confirms that our assumption is reasonable.

The parameter regression analyses were performed using an objective function, OF_{NRTL}, which is expressed for a binary system by the sum of squares of relative errors in pressure and the activity coefficients of the two components, as shown in Equation 2.7.

$$OF_{NRTL} = \sum_{i=1}^{n}\left(\frac{P^{Exp} - P^{Calc}}{P^{Exp}}\right)_i^2 + \sum_{i=1}^{n}\left(\frac{\gamma_1^{Exp} - \gamma_1^{Calc}}{\gamma_1^{Exp}}\right)_i^2 + \sum_{i=1}^{n}\left(\frac{\gamma_2^{Exp} - \gamma_2^{Calc}}{\gamma_2^{Exp}}\right)_i^2 \tag{2.7}$$

where n is the number of data points, the superscripts *Exp* and *Calc* refer to experimental and calculated values, respectively, and the subscripts 1 and 2 refer to the binary components.

In addition to pressure and activity coefficients, the qualities of predictions are assessed for equilibrium properties such as temperature and component equilibrium K values (*K-values*) of each binary system. The equilibrium *K-value* for component i is the ratio of vapor to liquid mole fraction, which can be recast as follows implementing the equilibrium criteria of Equation 2.6:

$$K_i = \frac{y_i}{x_i} = \frac{\gamma_i P_i^{\circ} \phi_i^V \lambda_i}{\hat{\phi}_i^V P} \tag{2.8}$$

2.3.3. Structure generation, optimization and descriptor calculation

Molecular descriptor calculation requires a series of steps common to all QSPR models. In the current work, ChemBioDraw Ultra 11.0 [20] was used to generate two-dimensional (2D) structures for the molecules in the data set and stored as cdx files. The 3D conformers with the least energy were found by implementing the OpenBabel [21, 22] genetic algorithm (GA) based conformer search, which employs the MMFF94 force field [23]. Dragon [17] software was then used to calculate over 3000 molecular descriptor values for each molecule. Molecular descriptors with at least one missing or undefined values were excluded in the descriptor generation process.

The descriptor set for each binary system is prepared by combining all the descriptors of the individual compounds in the system. Therefore, the first half of the descriptor set belongs to the first component and the second half of the descriptor set belongs to the second component in a binary system.

2.3.4. Descriptor reduction and model development

The current approach in descriptor reduction involves a hybrid strategy, which results in a non-linear wrapper-based model, where descriptor reduction and model development are performed simultaneously. Specifically, a hybrid niche algorithm that combines evolutionary programming (EP) and differential evolution (DE) was used as a wrapper around artificial neural networks (ANNs) to search for the best descriptor subsets from a large number of molecular descriptors. The subsequent discussion will be a brief introduction to ANNs followed by details on the actual descriptor reduction algorithm employed in the current study.

2.3.4.1. Artificial neural networks (ANNs)

Artificial neural networks are inspired by the brain and the interconnections among neurons. Different types of ANNs exist based on architecture, but in the current work, only feed-forward ANNs are relevant and any future reference to ANNs in the current work refers to feed-forward ANNs.

An important aspect of ANNs is the architecture or design, which consists of number of inputs, number of hidden layers and the number of neurons in each hidden layer. In the current work, the number of inputs to an ANN is chosen such that the ratio of data points to the number of inputs is at least ten. The number of hidden layers is fixed at one and the minimum number of hidden neurons is two. In addition, for each ANN, the ratio of the number of training data to the number of adjustable weights and biases was ensured to always be greater than two [24]. This was done as a precaution against over-fitting to the training data.

The current work uses the back-propagation algorithm proposed by Rumelhart et al. [25] to train the ANNs. In the modeling process, over-fitting is avoided by application of a training set (T) and an internal validation set (V) with an early-stopping method [26, 27]. In addition to the T and V sets, an internal test (IT) set was used in selecting the best ANNs during the descriptor search algorithm. Ideally, the training set should be representative of the entire data set, and each data point in the validation and internal test sets should correspond to at least one training data point. In the current work, self-organizing-maps (SOMs) are used to divide the data sets optimally subsequent to the ANN training. The number of map-units (which are analogous to neurons in feed-forward ANNs) in SOM training was adjusted to ensure that the number of training set data points is in the range 65-70% of the entire data set (excluding the external set). The Nguyen-Widrow algorithm was used to initialize weights and biases, which are updated using the Levenberg-Marquardt optimization technique.

2.3.4.2. Genetic representation

A good genetic representation of the solution domain is an important step in developing an efficient evolutionary algorithm. In the current work, the solution space is comprised of single hidden layer ANNs with all possible molecular descriptor subsets of a fixed size of a desired number of descriptors (ND) as inputs, which are determined by the user at the start of the program. The number of hidden neurons in these ANNs lies between a minimum of two and a maximum usually fixed at three times the desired number of descriptors in the model. An individual chromosome in the solution space is represented as a string of real numbers (genes) where each number (gene) corresponds to a particular descriptor.

2.3.4.3. The objective function

Another major aspect of an evolutionary algorithm is the choice of a suitable objective function. In the current work, the objective function used for an individual ANN is the minimization of the root-mean-squared error (RMSE) of the predicted property for the training set data. The minimization of RMSE on the training set is achieved by adjusting the weights using the back-propagation algorithm and the minimization is stopped once the error on the internal validation set increases for six successive iterations of the back-propagation algorithm. In addition, because of the wrapper-type approach of the current work, there is a second tier of optimization associated with the evolutionary algorithm for selecting the best ANN (that has already been optimized) from a large number of possible ANNs. The RMSE values of the predicted parameters relative to the target values were calculated for each of the subsets, T, V and IT. The following objective function, OF_{ANN}, was then computed based on these RMSE values:

$$OF_{ANN} = RMSE_T + RMSE_V + RMSE_{IT} \tag{2.9}$$

2.3.4.4. The algorithm

The algorithm has several parameters that need to be specified by the user, such as: (a) ND, (b) Population size, which is usually set at 400, (c) Number of niches, which is usually set to 1% of the population size to ensure that each niche has 100 individuals, (d) Percentage of population that undergoes MDE operations, which is usually set at 0.1, (e) Percentage of population that undergoes retraining, which is usually set at 0.3 and (f) Percentage of population that undergoes change in the number of hidden neurons, which is usually set at 0.5. At the start of the calculations, the algorithm undergoes an initialization process, where the individual ANNs in a parent population denoted as D are initialized with random descriptor subsets of size ND. The number of hidden neurons for each ANN is initialized to a value of 2. After initialization, the ANNs are trained using a back-propagation (Levenberg-Marquardt) algorithm resulting in network weights that minimize the $RMSE_T$ value. To avoid over-fitting the ANNs to the training data, early-stopping on the internal validation set is used. Specifically, training is stopped when $RMSE_V$ increases for six successive training iterations. Population D then undergoes the following five operations in a single iteration of the algorithm.

a) *Single-point mutation:* A randomly selected gene in each individual's chromosome is mutated/changed to a random descriptor number. The random descriptor number is chosen so that no two genes (descriptor numbers) in a chromosome are the same. The mutated individuals make up a new child population denoted as E.

b) *Modified differential evolution:* 10% of the individuals are randomly selected from population D. Modified differential evolution (MDE) operations are carried out on these individual chromosomes to result in a new mutated population M. The ANNs in M undergo training and the values of the objective function, OF_{ANN}, values are calculated for all individuals. The objective function values of the new ANNs are compared with the objective function values of the corresponding ANNs in population D. This is denoted as individual competition.

c) *Retraining:* 30% of the individuals are selected randomly from population D for retraining using different initial weights. If the new ANN has a lower OF_{ANN} value, then the old ANN is replaced with the new ANN.

d) *Architectural change:* Half the number of individuals are selected randomly from population D. The number of hidden neurons in half of these individuals is increased by 1 and for the rest of the individuals the value is decreased by 1. If the number of hidden neurons for any individual falls below the specified minimum value of 2, then the value is adjusted to the minimum value *o*f 2 for that particular ANN. The resulting new population after the architectural changes is denoted as A. The ANNs in A undergo training and the OF_{ANN} values are calculated for all individuals. Again, corresponding individuals in populations A and D enter individual competition, and population D is updated with fitter individuals.

e) *Rank based selection:* At the end of these four operations, the individual ANNs in the populations ‘D’ and ‘E’ are pooled together and subjected to rank-based selection [28]. In rank-based selection, each individual is ranked based on the number of individuals in the population that ‘dominate’ (an individual with lower objective function value dominates an individual with higher objective function value). The best ranked individuals make up the new population D, which again undergoes the previous four operations in the next iteration. The algorithm is stopped when the change in the mean of the internal test set error, i.e. *mean* ($\mathrm{RMSE}_{\mathrm{IT}}$), for each niche is less than 1% for 100 iterations of the algorithm. This is the stopping criterion for the algorithm.

2.3.4.5. Creating ensembles for final predictions

ANNs are known to be unstable and their predictive performance is dependent heavily on the training data and the training parameters. Therefore, a single outlier in the training data might have disastrous implications on the generalization ability of the model. To prevent this, aggregation or ensembling of ANNs is used, where the predictions of different ANNs are averaged to result in the final predictions [29, 30].

2.3.4.6. External validation

In a recent article, Tropsha *et al.* [24] have emphasized the need to validate QSPR models using external data sets. In the current work, some data were set aside as an external validation set. The performance of the current model on this dataset would indicate the generalization capability of the final model. To create this external data set, three different approaches were implemented:

1. A SOM clustering technique as described in S**ection 2.3.4.1** is used to divide the data (1,156 parameters for 578 systems) into 4 different sets (training, validation and internal test sets and external test set). Using this approach, estimating system specific predictions is not possible. This is due to the fact that the parameters a_{12} and a_{21} of a specific system might lie in different data sets.

2. The entire data set was divided into four sub-sets (training, validation, internal test, and external test sets) based on the functional groups of the components present in the binary systems. The data were divided such that all four data sets have adequate representation from the 31 functional groups shown in Figure 2.1. The proportion of data used for the different data sets was: 50% for the training set, 15% for the internal validation set, 10% for the internal test set and the remaining 25% for the external test set. For instance, there are 24 systems with Alcohol/Alkane interactions in the database. The data division for this type of interactions will be 12, 4, 2 and 6 of the systems assigned to the training, validation, internal test set and external test set, respectively. For interactions with small number of systems, we gave data allocation priority to the training followed by validation and internal tests.

3. In this approach, the training, validation and internal test sets were chosen using the SOM clustering technique. The external test set, however, was selected based on the functional groups of the components present in the binary systems. The external test set was used to evaluate the generalization ability of the model.

2.3.4.7. Modeling scenarios: To meet the objectives of this work, four case studies were constructed to investigate QSPR model parameterization of NRTL parameters. In all case studies, the ideal gas (IG) model was used to describe the gas phase behavior because practically all systems considered in this work are at low pressures. The four case studies are outlined as follows:

Ideal Solution: The ideal solution model was used to predict the phase-equilibrium behavior.

Regressed-NRTL: The NRTL model was used to predict the activity coefficients. The NRTL model parameters were regressed directly from the experimental data by minimizing the objective function OF_{NRTL}.

NRTL-QSPR: The generalized QSPR model was used to provide the NRTL model parameters and then the NRTL model was used to predict the activity coefficients.

UNIFAC-93: The UNIFAC model was used to predict the activity coefficients of each component. The UNIFAC interaction parameters reported by Gmehling *et al.* [7] were used in this case study.

The Regressed-NRTL study was conducted to evaluate the correlative capabilities of the NRTL model; whereas, Ideal Solution, NRTL-QSPR and UNIFAC-93 analyses are focused on assessing the *a priori* predictive capabilities of the ideal solution, the generalized model and the UNIFAC model, respectively.

For the first study, the ideal solution model was used to predict *T*, *P* and *K-values* for the entire database of 578 binary systems. In the Regressed-NRTL study, the two NRTL model parameters, a_{12} and a_{21}, shown in Equation 2.10, were regressed and used directly to predict (a) *P*, K_1 and K_2 for known *T* and x_1 and (b) *T* for known *P* and x_1.

$$a_{12} = \frac{g_{12}-g_{22}}{R} \qquad a_{21} = \frac{g_{21}-g_{11}}{R} \tag{2.10}$$

As expected, property predictions using the regressed NRTL parameters resulted in the most precise representations (lowest prediction error) for the data considered using the current framework. Therefore, the model parameters found in the regression analysis were used as target values in the development of the NRTL-QSPR model. The property prediction errors using the regressed parameters were taken as a benchmark to judge the performance of the NRTL-QSPR model.

2.4. Results and discussion

Four VLE properties (*T*, *P* and *K-values*) were used to analyze the predictive capability of the various models used in the Ideal Solution, Regressed-NRTL, NRTL-QSPR and UNIFAC-93 scenarios. The models used in each case were evaluated by comparing the property prediction errors, as described by RMSE, bias and percentage absolute average deviation (%AAD).

Table 2.1 provides the property prediction errors for the Ideal Solution and the Regressed-NRTL studies. As shown, the Ideal Solution model results in poor predictions compared to the Regressed-NRTL model. The Ideal Solution model has overall %AADs of 12.4, 1.3 and 17.4 for *P, T* and *K-values* predictions, respectively. The NRTL model with regressed parameters has lower overall %AADs of 2.6, 0.2 and 4.9 for *P, T* and *K-values* predictions, respectively. Compared to the Ideal Solution model, the Regressed-NRTL model resulted in error reductions in the property predictions of up to a factor of four.

The Regressed-NRTL study established the best achievable level of prediction errors using the NRTL model. The model parameters (a_{12} and a_{21}) that were obtained by regression in this study were then used as targets in the QSPR model development for the NRTL-QSPR study. QSPR models were developed by applying the three data division approaches discussed in Section 2.3.4.6. The models that were developed using these approaches had similar prediction capabilities. Since there were no significant prediction improvements, we have presented only the results found using the second approach in which the data were divided into four sets based on the functional groups of the components.

The QSPR model development process was initiated by dividing the 578 binary systems into four sets; with 285 in the training set, 89 in the validation set, 65 in the internal test set and 139 in the external test set. Next, a sequential regression process was performed in an effort to reduce the effect of parameter correlation on the predictive accuracy of the NRTL-QSPR model. In this approach, one parameter was fixed at the QSPR generalized value while the other parameter was regressed. This procedure was performed multiple times until the effect of the parameter correlation on the model development was minimized. In each iteration, the parameters a_{12} and a_{21} were regressed alternatively until no significant improvement in the property predictions was observed. The ensemble model was chosen after six iterations of the sequential regression process and consisted of twenty different networks, each having the same descriptors as inputs, but with different network architecture and weights. The best single model of the twenty ANNs was a one-layer neural network with an architecture of 29-5-1.

Figure 2.2a shows the correlation between the two regressed NRTL parameters in the first iteration regression analysis. The figure indicates that there is some level of correlation between the parameters. Figure 2.2b shows the correlation of the regressed parameter values that are used as target values in the final QSPR model (6th iteration model). The plot reveals that the correlation between the two parameters was reduced in the final regression analysis, which demonstrates the efficacy of this method in reducing the observed correlation of the model parameters. The RMSE of the predicted a_{12} and a_{21} from QSPR modeling were 347 and 364, respectively. After six iterations of sequential regression and ANN training, the RMSE values for the two parameters were decreased to 165 and 334. As expected, the reduction in the correlation of the regressed parameters was accompanied by a reduction in the RMSE values of the predicted parameters from the NRTL-QSPR models.

The 29 descriptors that are used as inputs for the ANNs are listed in Table 2.2. The results reveal that structural descriptors of both molecules are equally significant in predicting the NRTL model parameters. Four of the significant descriptors are GETAWAY (GEometry, Topology, and Atom-Weights AssemblY) descriptor types (2 from each component). These descriptors are 3-

dimensional (3-D) descriptors that encode information on the effective position of fragments and substituents in the molecular space [17]. The result also includes three functional group counts such as number of total tertiary C (sp3), non-aromatic conjugated C (sp2) and acceptor atoms for H-bonds (N,O,F). Some of the other significant descriptor classes which appeared more than once are constitutional indices, 2D atom pairs and RDF descriptors.

Table 2.3 summarizes the key improvements of the current study compared to previous work [10]. The current study employed a VLE database with a wide range of functional groups and a modeling technique that provides an internally consistent model.

Figures 2.3 and 2.4 show comparisons of the regressed NRTL model parameters, a_{12} and a_{21}, with the predicted model parameters from the NRTL-QSPR model, respectively. The plots indicate that the NRTL-QSPR predictions are in a good agreement with the Regressed-NRTL model parameters.

Table 2.4 provides the property prediction errors obtained using the QSPR predicted parameters (NRTL-QSPR study) for the training, validation, internal test and external test sets. The %AAD for the VLE predictions in all data sets was about twice the %AAD values calculated in the regression analysis (Regressed-NRTL study). The QSPR predicted parameters resulted in training set %AADs of 5.6, 0.5 and 8.4 for *P, T*, and *K-values* property predictions, respectively. The validation and training set prediction errors were comparable, demonstrating that the network was trained without over fitting. As expected, the generalized model results in slightly higher prediction errors for systems in the internal and external test sets. The %AAD values for the external test set were 7.3, 0.7 and 9.8 for *P, T* and *K-values* predictions, respectively. The errors for the external test set are about 1.5 times the corresponding errors in the training set.

Figures 2.5-2.7 show the distribution of the overall %AAD values for the predictions of pressure, temperature and K values using the NRTL-QSPR predicted model parameters, respectively. The %AAD values for pressure and temperature predictions are less than 6 and 0.6 for nearly 65% and

71% of the data, respectively. Similarly, the %AAD for K-value predictions is less than 8 for approximately 70% of the data. The NRTL-QSPR model yielded higher errors for systems consisting of sulfide, thiol and amide functional groups. These higher prediction errors can be attributed to the lack of similar structures in the training and test sets.

Figures 2.8a-2.8c show experimental and predicted VLE results for 2-methylbutane-hexane, hexane-1-propanol and acetonitrile-ethanol binary systems, respectively. The figures illustrate the capabilities of the newly developed generalized NRTL activity coefficient model in predicting the phase behavior of nearly ideal, highly non-ideal and azeotropic systems. Further, the *a priori* predictions from the generalized NRTL-QSPR model were compared with predictions from the modified UNIFAC model [7] (UNIFAC-93 study). As shown in Table 2.5, the overall prediction errors using the generalized parameters (NRTL-QSPR study) are lower than those produced by the group-contribution method, UNIFAC. %AADs of 9.1, 0.9 and 12.5 are obtained for *P, T* and *K-values*, respectively. These errors are 50% higher than the QSPR predictions. It is also noteworthy that the UNIFAC group interaction parameters were originally determined based on a database [18] that would have included a large proportion of the data sets used in this study. As a result, the UNIFAC model performs better on these systems than what might be expected for newer systems. In addition to higher errors, the UNIFAC model also lacks at least one group interaction parameter for 168 binary systems, which shows the deficiency of the model for generalized property predictions when the group interaction parameters are unavailable. Our results showed that when the missing interaction parameters in the UNIFAC model are set to 0 for prediction purposes, the %AADs increase to 14.5, 1.9 and 14.7 for *P, T* and *K-values*, respectively. In contrast, the NRTL-QSPR model generalization presented herein allows predictions without reliance on any additional phase equilibrium data for the constituent binaries and/or functional groups. Thus, these results indicate that the QSPR modeling approach is effective in generalizing NRTL model parameters for *a priori* property predictions.

Table 2.6 shows the property prediction errors for systems that are commonly encountered in refining processes. The table summarizes the VLE prediction errors for the 332 binary systems using the regressed parameters in the Regressed-NRTL study and the generalized parameters in the NRTL-QSPR study. The property predictions using generalized parameters were approximately twice the regression results. Comparable overall prediction errors were found from the previously reported results by Ravindranath *et al.* [10]. Many of the descriptors used in our newly developed model were reported as significant descriptors in the previous work [10] as well. These include descriptors such as number of benzene rings, number of triple bonds, number of acceptor atoms to H-bonds and various polarity related descriptors.

Table 2.7 shows the property prediction errors for systems with compounds that are typically formed in bi-phasic reactions. The table lists VLE prediction errors found using the Regressed-NRTL parameters and the generalized parameters in the NRTL-QSPR study for eight chemicals. The property predictions using generalized parameters were approximately two times that of the regression results. Lower prediction errors were observed for systems with propionaldehyde and 2-propanol in both the Regressed-NRTL and NRTL-QSPR studies. On the other hand, systems consisting of water yielded higher errors in Regressed-NRTL case. This can be attributed to the higher experimental uncertainties associated with water systems and the inability of the model in representing such systems precisely. Further, the mole fractions of water systems tend to be very small which results in larger *percentage* errors.

2.5. Conclusions

This study demonstrates the efficacy of an improved QSPR modeling approach that can be employed to successfully generalize the NRTL model parameters. An internally consistent QSPR model was developed using 578 binary VLE systems consisting of a wide range of functional groups. The QSPR generalized model parameters resulted in reasonable predictions for vapor-liquid phase equilibrium properties. The prediction errors were approximately two times the error of the data regression errors. In general, this QSPR model provided lower errors for *a priori* predictions than the current predictive models such as UNIFAC. Therefore, structural information of compounds can be used with a QSPR model to provide reliable estimates for NRTL model parameters of VLE systems. Further, this work implemented an effective approach for reducing the correlation of model parameters using sequential regression. The newly developed generalized NRTL-QSPR activity coefficient model presented in this work is only applicable to VLE systems. A study is underway to extend this modeling approach to liquid-liquid equilibrium.

Table 2.1: Predictions results for Ideal Solution and Regressed-NRTL case studies

Study	Model (Vapor /Liquid)	Parameters	No. of sys.	Property	No. of pts.	RMSE	Bias	%AAD
Ideal Solution	IG / Ideal Solution	None	578	P (bar)	16667	0.60	-0.10	12.4
				T (K)	16726	8.60	3.80	1.3
				K-values	9952	3.10	-0.45	17.4
Regressed-NRTL	IG / NRTL	a_{12} Regressed a_{21} Regressed	578	P (bar)	16563	0.20	0.00	2.6
				T (K)	16726	2.20	0.30	0.2
				K-values	9937	2.00	-0.15	4.9

Table 2.2. The descriptors used as inputs for the ANNs in the final ensemble for estimating the NRTL model parameters

No	Descriptor description	Component no	Type of descriptor
1	Second Zagreb index by valence vertex degrees	1	Topological indices
2	Number of total tertiary C(sp3)	2	Functional group counts
3	R autocorrelation of lag 5 / weighted by mass	1	GETAWAY descriptors
4	Eigenvalue sum from polarizability weighted distance matrix	1	Eigenvalue-based indices
5	Lowest eigenvalue n. 4 of Burden matrix/weighted by atomic masses	1	Burden eigen values
6	Number of non-aromatic conjugated C(sp2)	1	Functional group counts
7	Shape profile no. 7	1	Randic molecular profiles
8	3D-MoRSE - signal 15 / weighted by atomic van der Waals volumes	1	3D-MoRSE descriptors
9	Squared Moriguchi octanol-water partition coeff. (logP^2)	2	Molecular properties
10	Mean information index on atomic composition	2	Information indices
11	Presence/absence of C - Cl at topological distance 2	1	2D Atom Pairs
12	Leverage-weighted autocorrelation of lag 4 / weighted by van der Waals volume	1	GETAWAY descriptors
13	Presence/absence of C - Br at topological distance 4	1	2D Atom Pairs
14	Radial Distribution Function - 100 / unweighted	2	RDF descriptors
15	Number of triple bonds	2	Constitutional indices
16	Sum of atomic Sanderson electronegativities (scaled on Carbon atom)	2	Constitutional indices
17	Molecular walk count of order 4	1	Walk and path counts
18	H total index / weighted by polarizability	2	GETAWAY descriptors
19	Eigenvalue 02 from edge adj. matrix weighted by resonance integrals	2	Edge adjacency indices
20	1st component shape directional WHIM index / weighted by I-state	2	WHIM descriptors
21	H autocorrelation of lag 0 / weighted by van der Waals volume	2	GETAWAY descriptors
22	Highest eigenvalue n. 3 of Burden matrix / weighted by atomic masses	1	Burden eigen values
23	Number of acceptor atoms for H-bonds (N,O,F)	1	Functional group counts
24	Randic-type eigenvector-based index from polarizability weighted distance matrix	2	Eigenvalue-based indices
25	Spectral moment 08 from edge adj. matrix weighted by dipole moments	2	Edge adjacency indices
26	Number of benzene-like rings	2	Ring descriptors
27	Radial Distribution Function - 025 / unweighted	2	RDF descriptors
28	Moran autocorrelation of lag 1 weighted by Sanderson electronegativity	1	2D autocorrelations
29	Alcohol	2	Atom-centred fragments

Table 2.3. Comparison of our previous and current modeling efforts

		Previous study	This study
Database	Number of systems	332	578
	Number of data points	Over 9,700	Over 16,500
	Number of compounds	92	139
	Characterization	Using Danner's approach	Using functional group characterization ensuring greater degree of representation
Modeling technique	Number of models	Two models for the two (a_{12} and a_{21}) NRTL parameters	A single model for both a_{12} and a_{21} NRTL parameters that provides internally consistent predictions
	Descriptor reduction	Linear and Non-linear technique	Improved Non-linear technique (Genetic Algorithm)
	Artificial Neural Network (ANN) Architecture	a_{12} Model - (29 - 6 - 1); a_{21} Model - (29 - 6 - 1)	Best single Network (29 - 5 - 1); Architectures for the other models in the supplemental material

	Data Split	Training set (221 systems);
		Validation set (111 systems)
			Training set (285 systems);
			Validation set (89 systems);
			Internal Test Set (65 systems);
			External Test Set (139 systems)

Table 2.4. Predictions from the NRTL-QSPR case study

Study	Data split	Model (Vapor/Liquid)	No. of sys.	Property	No. of pts.	RMSE	Bias	%AAD
NRTL-QSPR	Training Set	IG / Generalized NRTL	285	P (bar)	8467	0.34	0.02	5.6
				T (K)	8480	3.73	0.36	0.5
				K-values	5017	0.83	-0.05	8.4
	Validation Set	IG / Generalized NRTL	89	P (bar)	2977	0.10	0.00	6.0
				T (K)	2995	3.68	-0.04	0.5
				K-values	1865	0.66	-0.04	8.2
	Internal Test Set	IG / Generalized NRTL	65	P (bar)	1701	0.20	0.03	7.2
				T (K)	1701	3.58	-0.47	0.6
				K-values	897	4.75	-0.62	9.6
	External Test Set	IG / Generalized NRTL	139	P (bar)	3551	0.32	-0.02	7.3
				T (K)	3551	4.12	0.35	0.7
				K-values	2174	2.99	-0.31	9.8

Table 2.5. Comparison of *a priori* predictions of NRTL-QSPR and UNIFAC-93 case studies

Study	Model (Vapor/Liquid)	No. of sys.	Property	No. of pts.	RMSE	Bias	%AAD
NRTL-QSPR	IG / Generalized NRTL	578	P (bar)	16696	0.28	0.01	6.2
			T (K)	16727	3.79	0.20	0.6
			K-values	9953	1.62	-0.15	8.8
UNIFAC-93	IG / UNIFAC	410*	P (bar)	10572	0.37	-0.04	9.1
			T (K)	10663	7.39	2.02	0.9
			K-values	6126	1.13	-0.10	12.5

*Due to lack of group interaction parameters, 168 systems of the 578 systems were not considered

Table 2.6. Property predictions for systems encountered in refining processes

Study	Model (Vapor/Liquid)	Parameters	No. of sys.	Property	No. of pts.	RMSE	Bias	%AAD
Regressed-NRTL	IG / NRTL	a_{12} Regressed a_{21} Regressed	332	P (bar)	9679	0.19	0.00	2.6
				T (K)	9767	1.94	0.26	0.2
				K-values	6532	0.63	-0.02	5.1
NRTL-QSPR	IG / Generalized NRTL	a_{12} QSPR a_{21} QSPR	332	P (bar)	9767	0.42	0.02	5.6
				T (K)	9767	3.75	0.12	0.5
				K-values	6483	0.81	-0.04	8.7

Table 2.7. Property predictions for systems encountered in bi-phasic reactions

Compound	No. of sys	No. of pts	%AAD					
			Regressed-NRTL			NRTL-QSPR		
			P (bar)	T (K)	K-values	P (bar)	T (K)	K-values
Octane	14	313	1.9	0.1	2.0	7.7	0.6	7.9
1-Propanol	16	315	2.1	0.2	2.8	4.3	0.3	6.8
2-Propanol	5	105	1.0	0.1	2.6	2.5	0.2	5.0
Acetone	36	977	2.1	0.2	5.1	4.7	0.4	7.4
Benzaldehyde	3	70	3.1	0.2	7.7	4.2	0.3	8.2
Propionaldehyde	9	177	0.7	0.1	2.6	1.2	0.1	3.4
Furfural	16	262	3.8	0.4	6.7	8.3	1.0	10.6
Water	28	629	4.1	0.3	6.3	11.3	0.9	13.2
Total	127	2848	2.6	0.2	4.8	6.5	0.6	8.3

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	13																														
2	Aldehyde																															
3	Alkane	24	5	15																												
4	Alkene	10	1	11	3																											
5	Alkyne	6	3	2	6	1																										
6	Amide			6	3	1																										
7	Amine	5		4				4																								
8	Aromatic Bromo			1																												
9	Aromatic Floro	2		2				1																								
10	Benzene Derivative	6	4	14			1	5	1	3	4																					
11	Bromoalkane										1																					
12	Carboxylate	2	6																													
13	Chloroalkane	6		6				7			8	4		2																		
14	Chloroalkene	1									1			8	1																	
15	Chlorobenzene			3				5	1		2	1		2																		
16	Epoxide		2																													
17	Ester	1	1																													
18	Ether	13	2	18	6	4		2		3	5		1	9				3	3													
19	Furfural	1		3	1						2			4	1			1														
20	H2S	1																														
21	Iodoalkane			1							2	1		4					1													
22	Ketone	3	4	20	4	1					7		6	9			1	3	2	2		1	4									
23	Nitrile	5		4	2	2					4			6	3	2			1				1	1								
24	Nitrite														1																	
25	Nitro Compound			3			1				5	1		5	2	2						2		2		2						
26	Pyridine Derivative			4							1	1		2									1	1			2					
27	Sulfide	4		4			1				1	2		5	2				1		1	1	1	1		1						
28	Thiol	1					2	1			1										1			1				4				
29	Thiophene			1			1				1			1												1	1					
30	Toluene Derivative	3	5	4	1			3	1	1	2			3		1			5	1			5	1		2	2	2				
31	Water	8	1	1				8			1								2					2	1		3		1			

	X
Y	■ #

■ # Number of available binary systems consisting of chemicals with functional groups of X and Y

▨ # Number of available binary systems consisting of chemicals with functional groups formed in bi-phasic reactions

□ No VLE data used

Figure 2.1: Database matrix of the compounds in the OSU database II along with the 31 functional groups represented

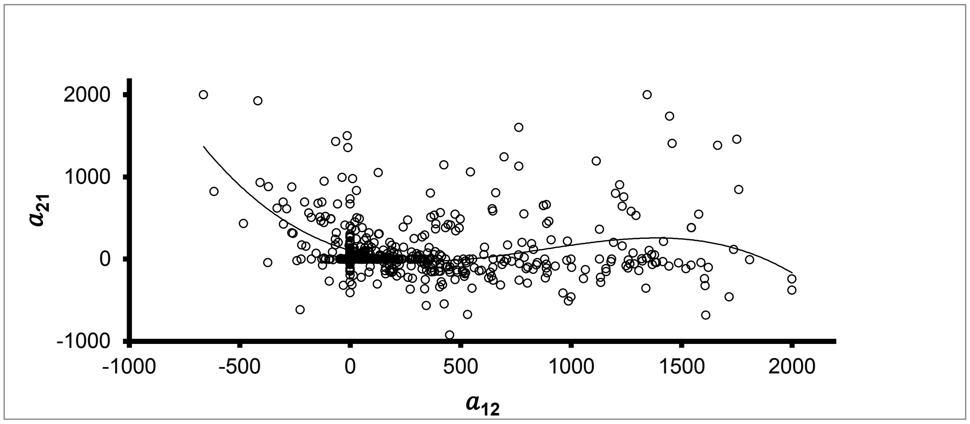


Figure 2.2a. Correlation between the regressed NRTL model parameters in the first iteration regression analysis

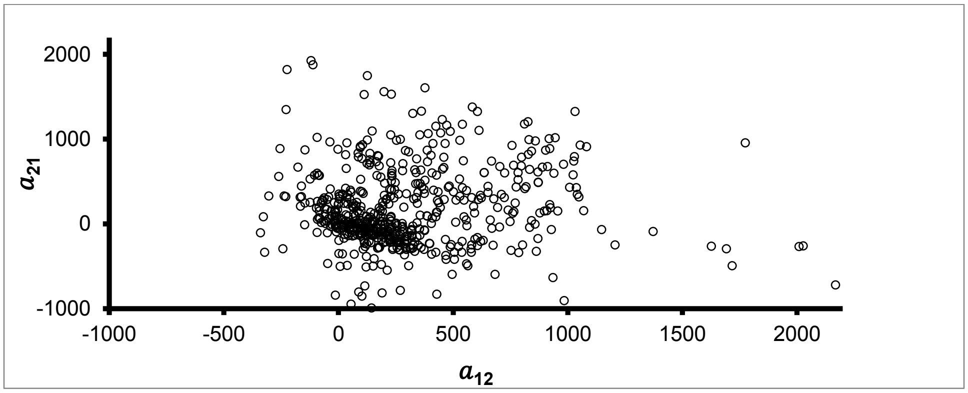


Figure 2.2b. Correlation of regressed NRTL model parameters after six iterations of sequential regression

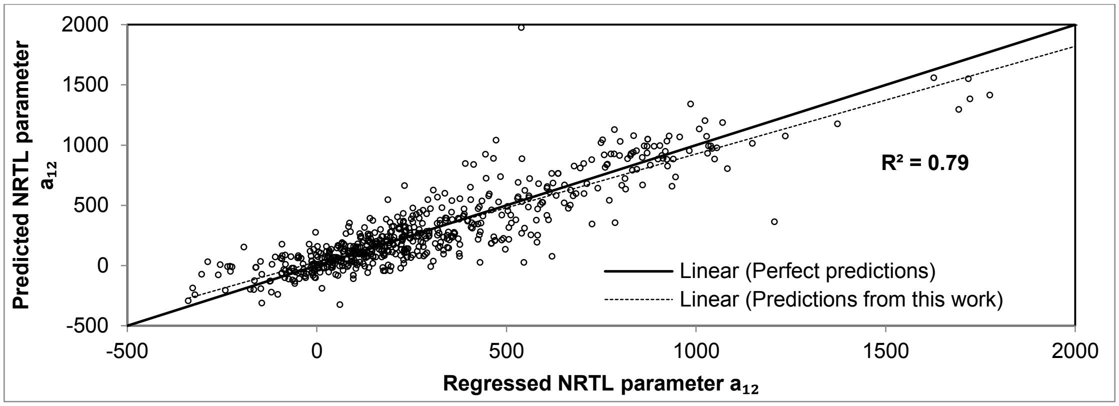


Figure 2.3. Comparison of the regressed NRTL (Regressed-NRTL study) and QSPR (NRTL-QSPR study) predicted a_{12} values for all data

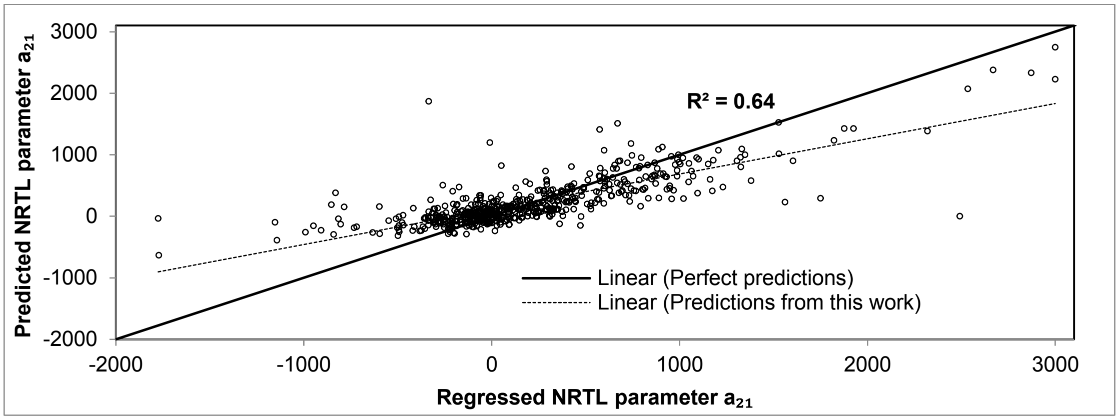


Figure 2.4. Comparison of the regressed NRTL (Regressed-NRTL study) and QSPR (NRTL-QSPR study) predicted a_{21} values for all data

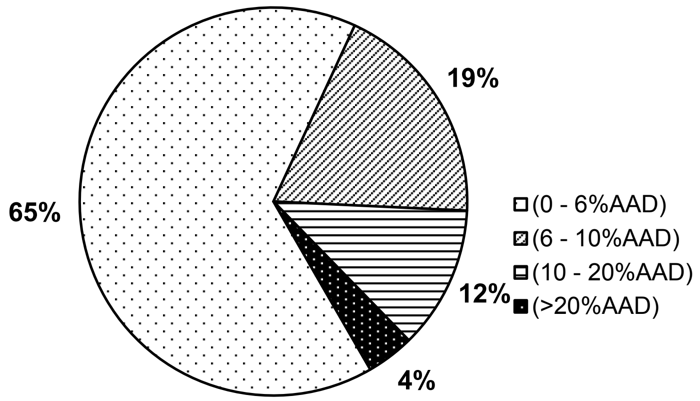


Figure 2.5. %AAD distribution of pressure predictions

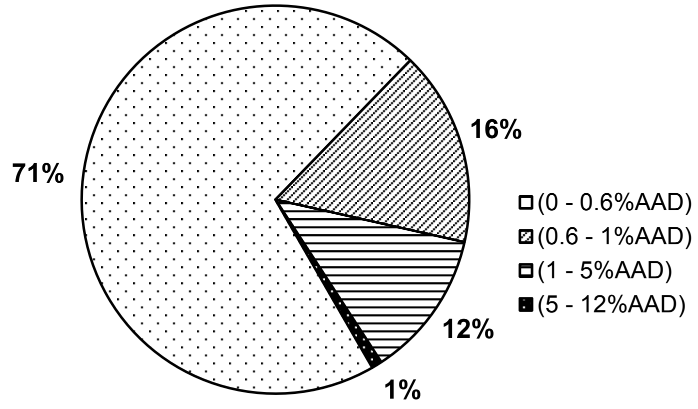


Figure 2.6. %AAD distribution of temperature predictions

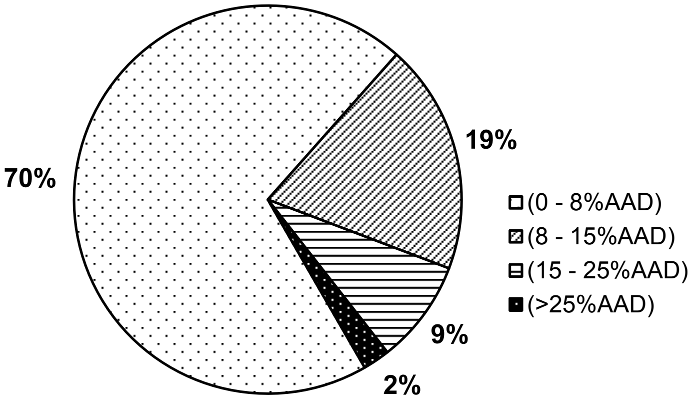


Figure 2.7. %AAD distribution of K-values predictions

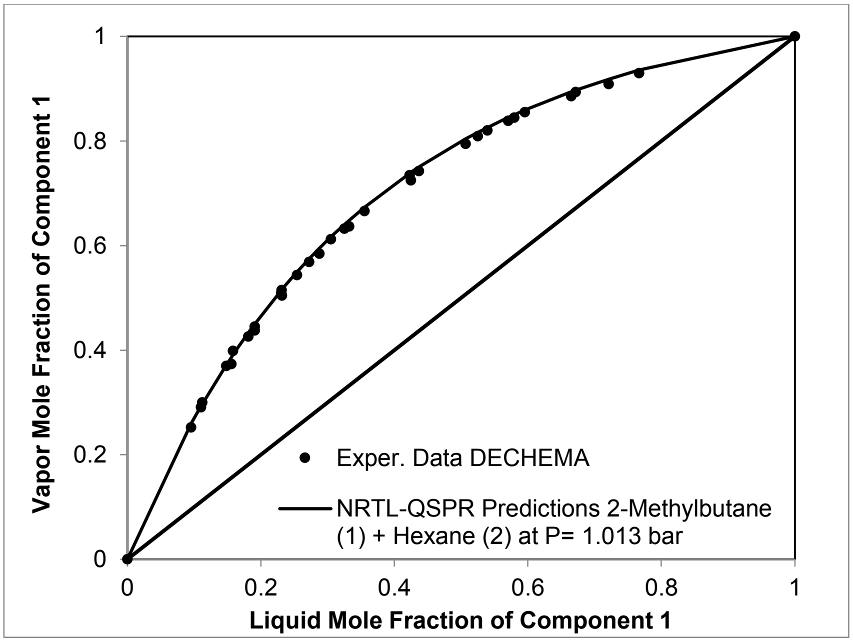


Figure 2.8a. Equilibrium phase compositions for 2-Methylbutane (1) + Hexane (2) system

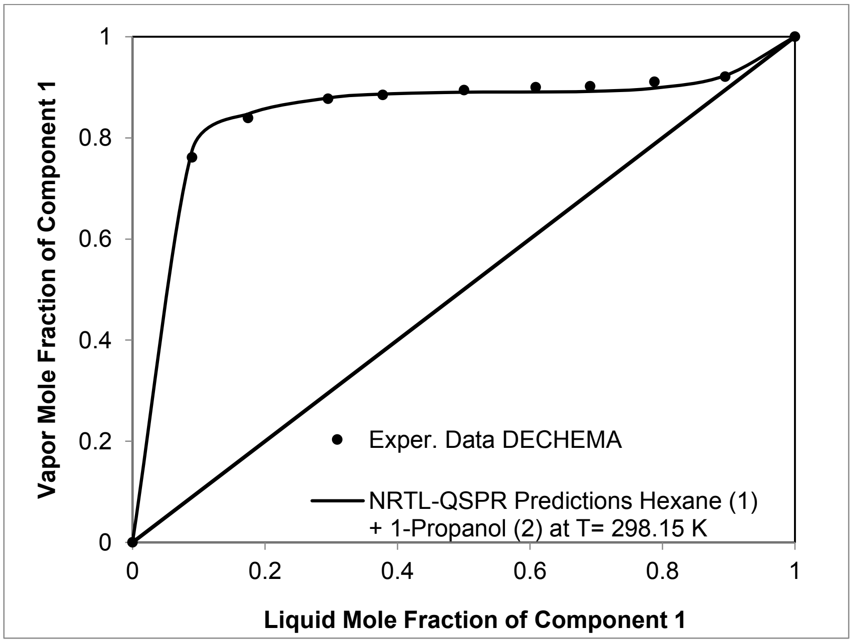


Figure 2.8b. Equilibrium phase compositions for Hexane (1) + 1-Propanol (2) system

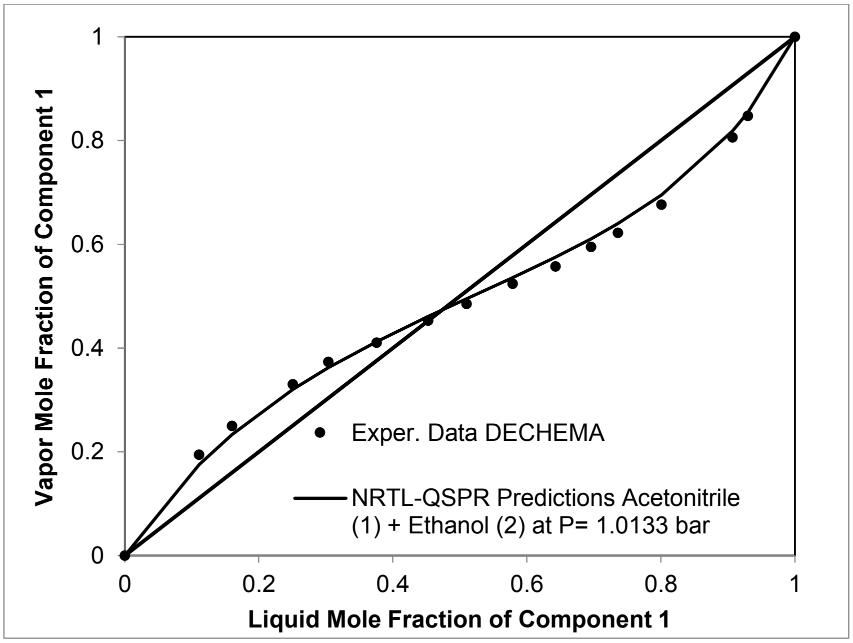


Figure 2.8c. Equilibrium phase compositions for Acetonitrile (1) + Ethanol (2) system

REFERENCES

1. S.I. Sandler, Chemical and engineering thermodynamics. third ed, John Wiley & Sons, Inc., New York, 1999.
2. J. Vidal, Mixing rules and excess properties in cubic equations of state. Chemical Engineering Science. 33(6) (1978). 787-791.
3. M.-J. Huron and J. Vidal, New mixing rules in simple equations of state for representing vapour-liquid equilibria of strongly non-ideal mixtures. Fluid Phase Equilibria. 3(4) (1979). 255-271.
4. M.L. Michelsen, A method for incorporating excess Gibbs energy models in equations of state. Fluid Phase Equilibria. 60(1–2) (1990). 47-58.
5. C. Boukouvalas, N. Spiliotis, P. Coutsikos, N. Tzouvaras, and D. Tassios, Prediction of vapor-liquid equilibrium with the LCVM model: a linear combination of the Vidal and Michelsen mixing rules coupled with the original UNIF. Fluid Phase Equilibria. 92(0) (1994). 75-106.
6. J. Gmehling, D. Tiegs, and U. Knipp, A comparison of the predictive capability of different group contribution methods. Fluid Phase Equilibria. 54 (1990). 147-165.
7. J. Gmehling, J. Li, and M. Schiller, A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties. Industrial & Engineering Chemistry Research. 32(1) (1993). 178-193.
8. T. Holderbaum and J. Gmehling, PSRK: A Group Contribution Equation of State Based on UNIFAC. Fluid Phase Equilibria. 70(2–3) (1991). 251-265.

9. J.M. Prausnitz and F.W. Tavares, Thermodynamics of fluid-phase equilibria for standard chemical engineering operations. AIChE Journal. 50(4) (2004). 739-761.

10. D. Ravindranath, B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior. Fluid Phase Equilibria. 257(1) (2007). 53-62.

11. D. Ravindranath, Structure-based generalized models for pure-fluid saturation properties and activity coefficients, School of Chemical Engineering, M.S. Thesis, Oklahoma State University: Stillwater. (2005).

12. B.J. Neely, Aqueous Hydrocarbon Systems: Experimental Measurements and Quantitative Structure-Property Relationship Modeling, School of Chemical Engineering, Ph.D. Dissertation, Oklahoma State University: Stillwater, Oklahoma. (2007).

13. J. Faria, M.P. Ruiz, and D.E. Resasco, Phase-Selective Catalysis in Emulsions Stabilized by Janus Silica-Nanoparticles. Advanced Synthesis & Catalysis. 352(14-15) (2010). 2359-2364.

14. H. Renon and J.M. Prausnitz, Local compositions in thermodynamic excess functions for liquid mixtures. AIChE Journal. 14(1) (1968). 135-144.

15. J.M. Prausnitz, R.N. Lichtenthaler, and E.G.d. Azevedo, Molecular Thermodynamics of Fluid-Phase Equilibria. 3rd ed. ed, Prentice-Hall, 1998.

16. W. Arlt, M.E.A. Macedo, P. Rasmussen, and J.M. Sorensen, Liquid-Liquid Equilibrium Data Collection. Chemistry Data Series. Vol. V, Parts 1-4, DECHEMA, Frankfurt, Germany, 1979 - 1987.

17. Dragon Professional 5.5, Talete SRL. (2010).

18. J. Gmehling, U. Onken, and W. Arlt, Vapor-Liquid Equilibrium Data Collection, Chemistry Data Series. Chemistry Data Series. Vol. I, Parts 1-8, DECHEMA, Frankfurt, Germany, 1977 - 2001.

19. DIPPR Project 801, Physical and Thermodynamic Properties of Pure Chemicals. (2011).

20. ChemBioOffice 11.0, CambridgeSoft. (2008).

21. R. Guha, M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen, The Blue ObeliskInteroperability in Chemical Informatics. Journal of Chemical Information and Modeling. 46(3) (2006). 991-998.

22. The Open Babel Package 2.3. (2011).

23. T.A. Halgren, Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry. 17(5-6) (1996). 490-519.

24. A. Tropsha, P. Gramatica, and V.K. Gombar, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR & Combinatorial Science. 22(1) (2003). 69-77.

25. D.E. Rumelhart, G.E. Hinton, and R.J. Williams, Learning internal representations by error propagation, in Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. 1986, MIT Press. p. 318-362.

26. L. Prechelt, Automatic early stopping using cross validation: quantifying the criteria. Neural Networks. 11(4) (1998). 761-767.

27. C. Rich, L. Steve, and G. Lee, Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. (2000).

28. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II. Evolutionary Computation, IEEE Transactions on. 6(2) (2002). 182-197.

29. D.K. Agrafiotis, W. Cedeño, and V.S. Lobanov, On the Use of Neural Network Ensembles in QSAR and QSPR. Journal of Chemical Information and Computer Sciences. 42(4) (2002). 903-911.

30. C. Merkwirth, H. Mauser, T. Schulz-Gasch, O. Roche, M. Stahl, and T. Lengauer, Ensemble Methods for Classification in Cheminformatics. Journal of Chemical Information and Computer Sciences. 44(6) (2004). 1971-1978.

CHAPTER III

A COMPARATIVE STUDY OF QSPR GENERALIZED ACTIVITY COEFFICIENT MODEL PARAMETERS FOR VLE MIXTURES

3.1. Introduction

Phase behavior properties of chemical species and their mixtures are required to design chemical processes involving multiple phases. In the absence of experimental data, generalized thermodynamic models are used to predict phase equilibria properties such as pressure, temperature, composition and partition coefficients.

The activity coefficient is a basic phase equilibria property that accounts for liquid mixture deviations from ideal behavior. Although a number of activity coefficient models exist in the literature [1-8], their use is limited by the availability of experimental data. Among these models, nonrandom two-liquid (NRTL) [1], universal quasi-chemical (UNIQUAC) [3] and Wilson [6] are used widely to correlate fluid phase equilibrium data. These models require two or three adjustable interaction parameters that are determined through regression of experimental data for a specific system. Thus, they cannot be applied to predict properties of vapor-liquid equilibrium (VLE) systems for which experimental data are not available.

Typically, to facilitate *a priori* predictions, group-contribution methods such as UNIQUAC functional-group activity coefficients (UNIFAC) and analytical-solution-of-groups (ASOG) [2, 8], have been employed to generalize the UNIQUAC and Wilson models. The premise for group-contribution methods is that an estimation of the property value is possible from the additive sum of contributions of basic molecular and atomic fragments (functional groups). The UNIFAC parameter matrix published in 2006 [9] has over 4,000 parameters including surface area (q), volume contribution (r) values of 115 sub groups, and 659 main group interaction (a_{ij}, b_{ij} and c_{ij}) values.

Although group-contribution methods provide *a priori* VLE property predictions, they suffer from several limitations including (a) the inability to account for the effects of neighboring molecules [10], (b) the lack of UNIFAC interaction parameters for functional groups that are not represented in the UNIFAC data matrix and (c) the inability to define effectively the functional groups of some chemical species. Therefore, a need exists for developing accurate and less computationally demanding models capable of *a priori* prediction of equilibria properties.

The current research has focused on developing an alternative method for generalizing the interaction parameters of the NRTL, UNIQUAC and Wilson models. A theory-framed quantitative structure-property relationship (QSPR) modeling approach was applied to generalize the interaction parameters. In this modeling approach, theoretical frameworks, such as the NRTL, are used to develop the behavior models, and QSPR methodology is used to generalize the substance-specific parameters of the models.

The QSPR modeling technique has been employed to generalize successfully various theoretical frameworks featuring thermophysical property models for predicting pure-component and mixture properties [11-14]. In a recent article [14], we implemented a theory-framed QSPR modeling approach to generalize the NRTL model parameters for VLE binary systems. In that study [14], we developed an internally consistent QSPR model using 578 binary VLE systems. The prediction errors from the QSPR model were approximately two times the error of the data regression analyses [14]. This study demonstrated the potential advantages of theory-framed QSPR modeling over group-contribution methods such as the UNIFAC model, including:

- **Molecule-molecule interactions:** Using theory-frame QSPR activity coefficient modeling, the phase behavior is described as a manifestation of molecular interactions as compared to functional-group interactions.
- **Range of applicability:** The UNIFAC model lacked interaction parameters for about 20% of the systems considered in an earlier study [14]. In contrast, the QSPR model was able to predict properties of all the systems considered, which shows a wider range of applicability.
- **Missing interaction parameters:** When the UNIFAC model was used for systems with at least one missing interaction parameter, the property prediction errors increased about two fold compared to the overall prediction [14]. This shows some limitation in the UNIFAC when the group interaction parameters are unavailable.
- **Model inputs:** One of the disadvantages of the UNIFAC model is its inability to define effectively functional groups of some molecules. In contrast, the QSPR model relies on molecular descriptors which are fixed values.
- **Simplicit**y: Typically, the QSPR model has fewer model parameters (about 300 parameters) compared to the UNIFAC model, which has about 4,000 parameters. This reflects the simplicity of the theory-framed QSPR modeling approach.
- **Ease of modeling**: A theory-framed approach offers the convenience to generalize various theoretical frameworks. Our approach doesn't require an extensive effort to develop generalized models compared to the group-contribution methods which require regression of interaction parameters of each functional-group interaction.

In this work, we further refined the QSPR generalization for the NRTL model parameters by (a) expanding the database to 916 VLE systems to include structural variations of the molecules considered and (b) ensuring the model handles pure and infinite-dilution limits accurately. These improvements lead to QSPR models that are capable of predicting VLE properties at pure and infinite-dilution limits for a wide range of chemical classes. The theory-framed QSPR modeling methodology was also applied to generalize the interaction parameters of the UNIQUAC and Wilson activity coefficient models.

Further, we evaluated the behavior representation capability of various activity coefficient models. The newly assembled VLE database which encompasses a wide range of chemical classes was used to evaluate the representation capabilities of the NRTL, UNIQUAC, Wilson and UNIFAC models. This head-to-head comparison helps to identify the weaknesses and strengths of the activity coefficient models when they are employed in representation of various functional-group interactions.

3.2. Activity coefficient models

A number of activity coefficient models for predicting VLE and LLE properties have been proposed by various researchers [1-5]. In general, the literature models can be classified as historical and semi-empirical activity coefficient models (Margules [7], Redlich-Kister [7] and Van Laar [7]), theory-based models, which includes local composition and two-liquid models (Wilson [6], NRTL [1] and UNIQUAC [3]) and group-contribution models (UNIFAC [2], ASOG [8]). Brief descriptions of those models pertinent to this work are provided below.

3.2.1. Wilson activity coefficient model

Wilson [6] first proposed an equation for excess Gibbs energy ($\overline{G^E}$) by adopting the Flory-Huggins expression [15] for athermal mixtures and introducing a local volume fraction in the equation. Although Wilson's model performs better than other empirical models, the equation is not applicable to LLE property predictions. The Wilson activity coefficient (γ) expression for a binary system is given as:

$$\ln \gamma_i = -\ln(x_i + \Lambda_{ij} x_j) - x_j \left(\frac{\Lambda_{ij}}{x_i + \Lambda_{ij} x_j} - \frac{\Lambda_{ji}}{x_j + \Lambda_{ji} x_i} \right) \tag{3.1}$$

where Λ_{ij} is defined as:

$$\Lambda_{ij} = \frac{v_j}{v_i} \exp\left(-\frac{\lambda_{ij} - \lambda_{ii}}{RT} \right) = \frac{v_j}{v_i} \exp\left(-\frac{a_{ij}}{RT} \right) \tag{3.2}$$

where λ_{ij} is the energy interaction between the i and j molecules, v is the pure component molar volume, x is mole fraction, R is the universal gas constant in $cal\,K^{-1}\,mol^{-1}$ and T is the mixture temperature in K.

The model contains two parameters that are specific for each binary system. These adjustable parameters are a_{12} or $(\lambda_{12} - \lambda_{11})$ and a_{21} or $(\lambda_{21} - \lambda_{22})$. The two parameters account for the differences in mixed (λ_{12} and λ_{21}) and pure (λ_{11} and λ_{22}) component characteristic energy interactions.

3.2.2. NRTL activity coefficient model

Renon and Prausnitz [1] developed the NRTL activity coefficient model based on the local composition theory of Wilson [6] and the two-liquid solution theory of Scott [16]. The model provides precise representation of highly non-ideal VLE and LLE systems [7]. For a binary system, the NRTL activity coefficient is expressed as follows:

$$\ln \gamma_i = x_j^2 \left[\tau_{ji} \left(\frac{G_{ji}}{x_i + x_j G_{ji}} \right)^2 + \frac{\tau_{ij} G_{ij}}{\left(x_j + x_i G_{ij} \right)^2} \right] \tag{3.3}$$

where τ_{ij} and G_{ij} are defined as:

$$G_{ij} = \exp(-\alpha_{ij}\tau_{ij}) \qquad \tau_{ij} = \frac{g_{ij} - g_{jj}}{RT} = \frac{a_{ij}}{RT} \tag{3.4}$$

where g_{ij} is the energy interaction between the i and j molecules and α is the non-randomness factor in the mixture.

The NRTL model has three adjustable parameters (defining $\alpha_{ij} = \alpha_{ji}$) that are unique for a binary system. These parameters are a_{12} or ($g_{12} - g_{22}$), a_{21} or ($g_{21} - g_{11}$), and α_{12}. The parameters account simultaneously for pure-component liquid interactions (g_{11} and g_{22}) and mixed-liquid interactions (g_{12} and g_{21}). The non-randomness factor (α_{12}) varies from 0.2 to 0.47 [7] and can often be set *a priori*. To be consistent with the DECHEMA database [17], the non-randomness factor was kept constant as 0.2 for all binary systems in this work.

3.2.3. UNIQUAC activity coefficient model

Abrams [3] derived the UNIQUAC equation for nonrandom mixtures containing molecules of different sizes [7]. The basis of the UNIQUAC model is that the excess Gibbs energy is the sum of the combinatorial and residual effects. The combinatorial portion attempts to describe the dominant entropic effects, and the residual portion accounts for the intermolecular forces of the system. The combinatorial portion is determined using the composition, size and shape of the components. The residual portion requires two adjustable binary parameters to account for inter-molecular forces. The UNIQUAC model is applicable to a wide range of liquid mixtures that contain both polar and nonpolar fluids. The UNIQUAC model for a binary system is given as:

$$g^E = g^E_{combinatoiral} + g^E_{residual} \tag{3.5}$$

where, g^E is the excess Gibbs energy, $g^E_{combinatorial}$ and $g^E_{residual}$ are the combinatorial and residual terms of the excess Gibbs energy, respectively, and can be expressed as follows:

$$\frac{g^E_{combinatorial}}{RT} = x_i \ln\left(\frac{\phi_i}{x_i}\right) + x_j \ln\left(\frac{\phi_j}{x_j}\right) + \frac{z}{2}\left(q_i x_i \ln\frac{\theta_i}{\phi_i} + q_j x_j \ln\frac{\theta_j}{\phi_j}\right) \tag{3.6}$$

$$\frac{g^E_{residual}}{RT} = -q_i x_i \ln\left[\theta_i + \theta_j \tau_{ji}\right] - q_j x_j \ln\left[\theta_j + \theta_i \tau_{ij}\right] \tag{3.7}$$

where the coordination number *z* is set equal to 10. Segment fraction, *ϕ*, and area fractions, *θ,* are defined as:

$$\phi_i = \frac{x_i r_i}{x_i r_i + x_j r_j} \qquad \theta_i = \frac{x_i q_i}{x_i q_i + x_j q_j} \tag{3.8}$$

where *q* and *r* denote the van der Waals surface area and volume of a component, respectively. The two adjustable parameters, τ_{12} and τ_{21}, are given in terms of characteristic energies, $u_{12} - u_{22}$ and $u_{21} - u_{11}$, by:

$$\tau_{ij} = \exp -\left(\frac{u_{ij} - u_{jj}}{RT}\right) = \exp -\left(\frac{a_{ij}}{RT}\right) \tag{3.9}$$

Activity coefficients are given as:

$$\begin{aligned} \ln\gamma_i &= -\ln\frac{\phi_1}{x_1} + \frac{z}{2} q_1 \ln\frac{\theta_1}{\phi_1} + \phi_2\left(l_1 - \frac{r_1}{r_2} l_2\right) \\ &\quad - q_1 \ln(\theta_1 + \theta_2 \tau_{21}) + \theta_2 q_1 \left(\frac{\tau_{21}}{\theta_1 + \theta_2 \tau_{21}} - \frac{\tau_{12}}{\theta_2 + \theta_1 \tau_{12}}\right) \end{aligned} \tag{3.10}$$

where

$$l_i = \frac{z}{2}(r_i - q_i) - (r_i - 1) \tag{3.11}$$

For a specific binary mixture, the two adjustable parameters are a_{12} or $(u_{12} - u_{11})$ and a_{21} or $(u_{21} - u_{22})$. These parameters account for the differences in mixed (u_{12} and u_{21}) and pure (u_{11} and u_{22}) component characteristic energy interactions. The values of the van der Waals surface area and volume are obtained from the Bondi group-contribution method [11].

The interaction parameters of the Wilson, NRTL and UNIQUAC models are usually determined using experimental equilibrium data. Therefore, the models cannot be applied for systems lacking experimental data, and hence, a generalized model is required to predict the interaction parameters in the absence of experimental data.

3.3. QSPR methodology

The following steps are employed in the development of QSPR models for generalizing the Wilson, NRTL and UNIQUAC model parameters: (1) database development, (2) parameter regression analyses for VLE systems using the Wilson/NRTL/UNIQUAC models, (3) molecular structure generation and optimization, (4) descriptor generation and (5) descriptor reduction and QSPR model development using neural networks.

Figure 3.1 shows a schematic representation of the steps in developing the QSPR model. The initial step consists of compiling a reliable database of binary VLE data. Next is the regression analyses of the interaction parameters of the Wilson/NRTL/UNIQUAC models for the VLE systems in the database. Then, 2-dimensional (2D) structures of components in each binary system are generated. The 2D structures are then optimized to find a 3-dimensional (3D) representation of the molecules with the minimum conformation energy. The optimized 3D molecular structures are used to generate molecular descriptors using software such as DRAGON [18] and CODESSA [19].

The next step is descriptor reduction where the large number of generated molecular descriptors are reduced to find the most significant descriptors for accurate property predictions. Simultaneously, these significant descriptors are used to develop a neural network model. Finally, the relationships between the descriptors and the model parameters (Wilson/NRTL/UNIQUAC) are investigated. The main stages of the model development process are described in greater detail below.

3.3.1. Database development

A comprehensive VLE database was assembled from available sources by insuring sufficient representation of different functional groups in the database. A low-pressure binary VLE database (Oklahoma State University, OSU database I) [11] consisting of 188 binary VLE systems totaling 4716 data points was assembled. This database is comprised of systems of aliphatic and aromatic hydrocarbons, water, alcohols, ethers, sulphides and nitrile compounds. A second database, comprised of 388 binary VLE systems totaling 12,010 data points, was taken from DECHEMA [20]. A third database consisting of 384 binary systems totaling over 19,000 data points was taken from NIST-TDE [21]. In total, the database compiled in this work consists of 916 binary systems formed from various combinations of 140 different compounds. In addition to pressure, temperature and mole fraction (PTXY) data, we have collected over 500 data points of infinite-dilution activity coefficient values (γ^{∞}) for 137 of the 916 VLE systems in the database [20]. Further, pure-component vapor pressure data were collected from DIPPR [22] and DECHEMA [20]. A total of over 35,000 vapor-liquid equilibrium data points were assembled in the final database (Oklahoma State University, OSU-VLE Database III). The data covered a temperature range from 128 to 554 K and pressures to 58 bar; however, over 99% of the data were at pressure of less than 10 bar.

To illustrate the distribution of data by functional groups, the compounds present in the OSU-VLE Database III were classified in a similar manner as the UNIFAC functional group classification approach [2]. The database is composed of compounds belonging to 31 chemical classes.

Figure 3.2 illustrates the data distribution of the binary systems in the OSU database III based on chemical classes. The number of systems represented for each type of functional-group interaction is shown in the figure. Systems containing alcohol or alkane components are represented extensively in the database due to their abundant data.

3.3.2. Interaction parameter regression

Regression of the interaction parameters of the NRTL, Wilson and UNIQUAC models were performed to evaluate their respective representation capabilities. The regression analyses were performed by applying the Gibbs equilibrium criteria of a closed system containing coexisting liquid and vapor phases, subject to mass balance constraints. The split approach, as shown in Equation 3.12, was employed in the phase equilibria calculations.

$$\hat{\phi}_i^V P y_i = \gamma_i P_i^{\circ} \phi_i^V x_i \lambda_i ; \qquad i = 1, n \tag{3.12}$$

where n is the number of components, and for any component i, $\hat{\emptyset}^V$ is the component fugacity coefficient in the vapor phase, y is the vapor mole fraction, γ is the component activity coefficient in the liquid phase, P is the mixture pressure, P° is the pure-component vapor pressure, $\emptyset^V$ is the pure-component fugacity coefficient in the vapor phase, x is the liquid mole fraction and λ is the Poynting factor. Since most of the VLE systems considered in this study were at low pressure, the vapor-phase fugacity coefficients were assumed to be 1. We have also investigated the quality of representation when equation-of-state (EOS) models are used to calculate the vapor-phase fugacity coefficients (results not shown). Our findings show there is no improvement on the overall representation error. This confirms that our assumption is reasonable.

The Poynting factor is expressed as follows:

$$\lambda_i = \exp\left(\frac{v_i^L (P - P_i^\circ)}{RT}\right) \tag{3.13}$$

where v^L is the liquid molar volume and is determined using the Rackett equation [23].

The objective function, *OF*, used in the parameter regression analyses, was the weighted sum of squares of the relative errors in pressure, K-value, infinite-dilution activity coefficients and weighted absolute sum of model parameters, is shown as follows.

$$\begin{aligned} OF = \sum_{i=1}^{n} w_1 \left(\frac{P^{Exp} - P^{Calc}}{P^{Exp}}\right)_i^2 + w_2 \sum_{i=1}^{n} \left(\frac{K_{values}^{Exp} - K_{values}^{Calc}}{K_{values}^{Exp}}\right)_i^2 \\ + w_3 \sum_{i=1}^{n} \left(\frac{\gamma_{values}^{\infty\ Exp} - \gamma_{values}^{\infty\ Calc}}{\gamma_{values}^{\infty\ Exp}}\right)_i^2 + w_4 (Par) \end{aligned} \tag{3.14}$$

where the weights were: $w_1 = 1$; $w_2 = 1/15$; $w_3 = 1/10$; $w_4 = 2E - 6$; *n* is the number of data points, *Par* is $|a_{12}| + |a_{21}|$ and the superscripts *Exp* and *Calc* refer to experimental and calculated values, respectively. This objective function and associated weights were developed after evaluating the VLE property representations employing various objective function formulations. Equation 3.14 was selected due to the balance it provided in the model representation errors for temperature, pressure, equilibrium constants, activity coefficient and vapor mole fraction, and reduction in correlation of the model parameters (a_{12} and a_{21}) [24].

3.3.3. Descriptor calculation

In this study, ChemBioDraw Ultra 11.0 [25] software was used to generate 2D and 3D structures of the molecules. Then, Open Babel software was used to optimize the 3D structures by minimizing the conformational energy of the molecules using a genetic algorithm (GA) based conformer search [26, 27], which employs the MMFF94 force field [28]. The optimized molecules are then used to generate 2344 DRAGON [29] and 598 CODESSA [19] 0D, 1D, 2D, and 3D descriptors.

3.3.4. Descriptor input

In this study, 2942 structural descriptors are calculated for each compound in the database. The input descriptor set for each binary system is prepared by calculating the differences of all the individual descriptors of the compounds in the binary system. The use of difference of descriptors as inputs is a novel approach, which enables us to develop QSPR models that satisfy the pure-limit behavior of activity coefficient properties. For a hypothetical mixture of X and Y, where X and Y are the same molecule, the values of the activity coefficients for both components are ones; i.e., the interaction parameter values are zeros which requires that the QSPR input values (descriptor differences) to be zeros. Hence, the QSPR model is able to identify such systems and provide prediction values that satisfy the limiting behavior or zero interaction parameters.

3.3.5. Descriptor reduction and model development

In this work, the descriptor reduction involves a hybrid strategy where descriptor reduction and model development happen simultaneously. This approach employs evolutionary programming (EP) and differential evolution (DE) as a wrapper around artificial neural networks (ANNs) to search for the best descriptor subsets from total number of molecular descriptors. A detailed discussion on this step can be found in our previous works [14, 30, 31].

In the model development process, the entire data set was divided into four sub-sets (training, validation, internal test, and external test sets). The data was divided while insuring adequate representation of all functional-group interactions in each of the data sets. The proportion of data for the different data sets was: 50% for the training set, 15% for the internal validation set, 10% for the internal test set and the remaining 25% for the external test set. For example, there are 21 systems with ketone/alkane interactions in the database. The data division for this type of interaction will be 11, 3, 2 and 5 of the systems assigned to the training, validation, internal test set and external test sets, respectively. For interactions with a small number of systems, data allocation priority was given to the training followed by validation and internal test sets.

The training, validation and internal test set data were used in the descriptor reduction and model development process. The validation data set is used to avoid over-fitting by employing an early-stopping method [29, 32]. In addition, the internal test data was used to select the best ANNs during the descriptor reduction algorithm. The external test set data was set aside in the model development process and used to assess the generalization (*a priori* prediction) capability of the developed model.

3.3.6. Modeling scenarios

Eight case studies were performed to investigate the representation and prediction capability of the various models. In all case studies, the ideal gas (IG) model was used to describe the gas phase behavior. The eight case studies are outlined as follows:

Ideal Solution: The ideal solution model was used to predict the phase-equilibrium behavior.

Regressed-NRTL: The NRTL model with regressed parameters was used to represent VLE properties.

Regressed-Wilson: The Wilson model with regressed parameters was used to represent VLE properties.

Regressed-UNIQUAC: The UNIQUAC model with regressed parameters was used to represent VLE properties.

NRTL-QSPR: The generalized QSPR model was used to provide the NRTL model parameters, and then the NRTL model was used to predict the activity coefficients.

Wilson-QSPR: The generalized QSPR model was used to provide the Wilson model parameters, and then the Wilson model was used to predict the activity coefficients

UNIQUAC-QSPR: The generalized QSPR model was used to provide the UNIQUAC model parameters, and then the UNIQUAC model was used to predict the activity coefficients

UNIFAC-2006: The UNIFAC model was used to predict the activity coefficients of each component. The UNIFAC interaction parameters reported by Gmehling et al. [9] were used in this case study.

The case studies with regressed parameters from experimental data were conducted to evaluate the correlative capabilities of the activity coefficient models. In contrast, the Ideal Solution, NRTL-QSPR, Wilson-QSPR, UNIQUAC-QSPR and UNIFAC-2006 case studies were focused on assessing the *a priori* predictive capabilities of each of the listed models.

The representation and prediction capabilities of the models were assessed for equilibrium properties such as pressure (P), activity coefficients (γ^{∞}), temperature (T), vapor mole fraction (y_1) and equilibrium K-value (average of K_1 and K_2). In the first case study, the ideal solution model was used to predict *T*, *P,* y_1 and *K-value* for the entire database. In the Regressed-NRTL, Regressed-Wilson and Regressed-UNIQUAC studies, the two NRTL, Wilson and UNIQUAC model parameters, a_{12} and a_{21}, shown in Equations 3.2, 3.4 and 3.9, were regressed. The regression was done by preforming bubble-point pressure calculations. The regressed or QSPR predicted parameters are directly used directly to calculate (a) *P,* y_1*,* γ^{∞} and *K-value* for known *T* and x_1 and (b) *T* for known *P* and x_1 .

3.4. Results and discussion

The results of this study are focused on (a) assessment of model representation of equilibrium properties, (b) QSPR generalized predictions and (c) limiting-behavior property prediction assessments. The results for each of these studies are discussed in the following sections.

3.4.1 Representation assessment

The NRTL, Wilson and UNIQUAC models were used to correlate experimental *P*, *T*, *x* and y data of 916 binary systems. The representation capabilities of the models were analyzed by calculating the root-mean-squared error (RMSE), bias and percentage absolute average deviation (%AAD).

Table 3.1 provides the property prediction errors for the ideal solution and representations of the Regressed-NRTL, Regressed-Wilson and Regressed-UNIQUAC case studies. As expected, the ideal solution model resulted in poor predictions compared to the NRTL, Wilson and UNIQUAC activity coefficient models. The overall %AADs for the ideal solution model were 13.5, 1.5, 15.3 and 19.2 for *P*, *T, y_1* and *K-value* predictions, respectively. The activity coefficient models reduced the errors by about four fold compared to the ideal solution model. The NRTL model with regressed parameters provided overall representation %AADs of 2.1, 0.2, 4.3 and 5.5 for *P*, *T, y_1* and *K-value*, respectively. The UNIQUAC model with regressed parameters provided overall representation %AADs of 1.9, 0.2, 4.1 and 5.3 for *P*, *T, y_1* and *K-value* properties, respectively. The Wilson model with regressed parameters provided overall representation %AADs of 1.9, 0.2, 4.0 and 5.2 for *P*, *T, y_1* and *K-value*, respectively. The three activity coefficient models resulted in comparable overall representation capabilities for correlating *P*, *T*, *y_1* and *K-value* experimental data.

Figure 3.3 shows the distribution of pressure regression errors for the NRTL, Wilson and UNIQUAC models by functional-group interactions. Results of each functional-group interaction is shaded in variations of grey based on the %AAD ranges given in the figure key.

As shown in the error matrix, all three models have comparable representation capabilities for all type of interactions with the exception of the water systems. As expected, all models provided accurate representation when the components of the systems have the same functional groups (diagonal elements of the triangular matrix). This is due to the fact that components with the same functional groups are structurally similar and produce nearly-ideal behavior (interaction), thus easier property correlation. All models resulted in relatively high errors for most of the systems containing water. In particular, the errors were above 8% (about 4 times higher than the overall results) for systems containing water and aldehyde, amide, benzene derivatives, epoxide, ether or furfural.

Table 3.2 shows the property predictions of the ideal solution and representations of the Regressed-NRTL, Regressed-Wilson and Regressed-UNIQUAC models for binary VLE systems containing water. As shown, the ideal solution model resulted in higher errors compared to the activity coefficient models. The property representation errors of the Regressed-NRTL, Regressed-Wilson and Regressed-UNIQUAC models for water systems were about twice higher than the results found for the overall data. These higher representation errors could be attributed to a combination of factors, including (a) the higher experimental uncertainties associated with water systems, and (b) the inability of the models in representing such systems precisely. Further, the mole fraction of aqueous systems tend to be very small which results in greater *percentage* errors since the denominators are small values.

Table 3.3 presents comparison of pressure representations of the NRTL, UNIQUAC and Wilson models for 13 functional-group interactions where the UNIQUAC model provided more than 20% lower %AADs than the NRTL and Wilson models. Most of these systems tend to exhibit highly-non ideal behavior due to the presence of polar chemicals such as water, alcohols, aldehydes, ketones, sulfides, etc.

Table 3.4 provides comparison of pressure representations of the NRTL, UNIQUAC and Wilson models for 11 functional-group interactions where the Wilson model resulted in more than 20% higher %AADs than the NRTL and UNIQUAC models. Six of the listed interactions were water systems. Although the results provided insights into the model performance for water systems, more data are needed to represent adequately each interaction and provide conclusive comparison.

3.4.2. QSPR generalized predictions

The Regressed-NRTL, Regressed-Wilson and Regressed-UNIQUAC studies established the best achievable level of prediction errors that can be attained by QSPR generalized models. As such, the regressed model parameters (a_{12} and a_{21}) were used as targets when developing the QSPR models.

In the QSPR model development process, a sequential regression approach was performed in order to reduce the effect of correlation of the parameters on accuracy of the generalized model. In this method, a QSPR model is developed by using the initial regressed parameters as targets. Next, the regression analysis is repeated by regressing only one of the parameters while fixing the other as the generalized value from the QSPR model. The parameters found in this step are then used to develop a new QSPR model. These alternative regression and QSPR modeling steps are repeated multiple times until the effect of correlation is reduced and no significant improvement in predictive capability is observed. The final ensemble QSPR models were chosen after five iterations of the sequential regression process and consist of twenty different networks, each having the same descriptors as inputs, but with different network architecture and weights.

Tables 3.5, 3.6 and 3.7 provide the list of the 30 molecular descriptors used as inputs in developing the QSPR models for predicting the NRTL, UNIQUAC and Wilson model parameters, respectively. In the tables, DR and CO represent molecular descriptors calculated using DRAGON [18] and CODESSA [19], respectively. The lists show that functional-group counts, electrostatic, quantum chemical and molecular properties are significant in predicting the interaction parameters of the NRTL, UNIQUAC and Wilson models. Specific descriptors that are related to polarity and LogP (octanol-water partition coefficient) were selected as important descriptors. Polarity signifies the distribution of the electrons (charge) which plays a significant role on how molecules interact with each other. LogP represents the distribution of molecules in aqueous and organic phases and it provides an insight on hydrophilic and hydrophobic interactions of molecules of various types interacting in the presence of organic and aqueous phases at equilibrium.

Figures 3.4a, 3.4b and 3.4c show comparisons of the regressed NRTL, UNIQUAC and Wilson model parameters with the predicted model parameters from the NRTL-QSPR, UNIQUAC-QSPR and Wilson-QSPR models, respectively. The figures show there is a good agreement between the regressed and QSPR predicted parameters, which is signified by squared correlation coefficient (R^2) values close to 1.

Tables 3.8, 3.9 and 3.10 provide the property prediction errors obtained using the QSPR predicted parameters from the NRTL-QSPR, UNIQUAC-QSPR and Wilson-QSPR studies. The results are classified into training, validation, internal test and external test sets. In addition to providing results for all systems, the table also provides results categorized by water containing and highly non-ideal systems. The ratio of %AAD values from QSPR model predictions and regression results are shown as the %AAD multiplier in the tables. All three QSPR models provided VLE predictions about twice higher than the regression analysis %AAD values for all categories including water containing and highly non-ideal systems. Further, the results show the errors for the training and validation data sets were comparable. This indicates that the models were developed without over fitting the training set. In addition, the predictions for the external and internal test sets were comparable to the overall prediction quality, which demonstrates the capability of the model for generalized (*a priori*) predictions.

Figure 3.5 shows the distribution of pressure regression errors for the NRTL-QSPR, Wilson-QSPR and UNIQUAC-QSPR models by functional-group interactions. The figures indicate all three models provide pressure predictions within 5 %AAD for most of the functional-group interactions present in the database. The exceptions are water containing systems where on average the pressure results were approximately10%, which is twice higher than the overall results.

Figures 3.6a, 3.6b, 3.6c and 3.6d show the QSPR predicted equilibrium phase compositions of n-heptane-ethylbenzene, propionic aldehyde-acetone, benzene-tert-butyl alcohol and furfural-ethanol, respectively. The figures indicate all three QSPR models were able to match the experimental composition data accurately. This demonstrates the capabilities of the QSPR models for predicting VLE properties of nearly-ideal and highly-non ideal systems.

The generalization capability of the QSPR models were compared with the predictions from the UNIFAC-2006 model [9]. Table 3.11 shows the results of the NRTL-QSPR, Wilson-QSPR, UNIQUAC-QSPR and UNIFAC-2006 case studies. The results show the QSPR models resulted in comparable predictions to that of the UNIFAC model for 853 systems. When the UNIFAC model is used for systems with at least one missing interaction parameter, the prediction errors increased more than two fold. This shows the limitation of the model for generalized predictions in the absence of interaction parameters. Further, the UNIFAC interaction parameter matrix [9] used in the UNIFAC-2006 study has over 4,000 parameters. In contrast, the QSPR model has about 300 model parameters (neural network weights and biases). Thus, our methodology provides *a priori* and easily implementable QSPR models with wider applicability range than that of the UNIFAC model.

3.4.3. Limiting-behavior prediction assessment

Table 3.12 shows the representation and prediction of infinite-dilution activity coefficients for 137 binary systems using regressed and QSPR predicted model parameters of the NRTL, UNIQUAC and Wilson models and UNIFAC-2006. The Regressed-NRTL, Regressed-Wilson and Regressed-UNIQUAC models provided overall representation %AADs of 8.7, 8.2 and 8.7 for $\gamma\infty$, respectively. The generalized NRTL-QSPR, Wilson-QSPR and UNIQUAC-QSPR models provided $\gamma\infty$ predictions with approximately twice the error found in the regression analyses. The UNIFAC-2006 model resulted in relatively lower error compared to the QSPR models. It is noteworthy that the UNIFAC interaction parameters are regressed using the DECHEMA database [17] which included a large proportion of the PTXY and all $\gamma\infty$ data used in this study. Consequently, the UNIFAC model performs better for these systems than newer systems that are not used in the interaction regression step.

Table 3.13 shows infinite-dilution activity coefficient representation of the NRTL, UNIQUAC and Wilson models for 14 systems where the UNIQUAC model resulted in more than 50% lower %AADs compared to the NRTL and Wilson models. The systems listed contain polar compounds including water, ethanol, methanol, etc. This indicates the UNIQUAC model handles infinite-dilution activity coefficient property better for systems with polar components than the NRTL and Wilson models.

3.5. Conclusion

In this study, we assessed the representation capability of the NRTL, UNIQUAC and Wilson models and generalized the model parameters of the three activity coefficient models using a QSPR modeling approach. A database of 916 binary VLE data consisting of 140 compounds which belong to 31 chemical classes were collected in this study. Our assessment revealed all three models have comparable representation capability for correlating experimental phase equilibria properties. Further, all three models resulted in relatively higher errors for water containing systems. Although further investigation is needed, our study shows the UNIQUAC model tends to provide slightly lower VLE properties errors for systems containing polar compounds.

QSPR models were developed to predict the model parameters of the NRTL, UNIQUAC and Wilson models by ensuring the limiting behavior of mixtures are obeyed. The predictive capabilities of the QSPR generalized models were assessed for phase equilibria properties including pressure, temperature, vapor mole fractions, equilibrium constants and infinite- dilution activity coefficients. Overall, the QSPR generalized models provided predictions within twice the regression results. In addition, we found comparable property predictions between the newly developed QSPR model and the UNIFAC model. The UNIFAC model, however, had a limited range of applicability due to lack of interaction parameters. Thus, our methodology provides a potential alternative approach for generalizing activity coefficient models.

Table 3.1. VLE property predictions of the Ideal Solution model and representation capability of the NRTL, UNIQUAC and Wilson models

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
Ideal Solution	None	P (bar)	916	33283	0.68	-0.13	13.5	97
		T (K)	916	33283	9.29	4.15	1.5	28
		y_1	677	18210	0.10	-0.01	15.3	100
		K-value	676	18205	6.79	-0.82	19.2	100
Regressed-NRTL	a_{12} & a_{21}	P (bar)	916	33841	0.15	0.00	2.1	14
		T (K)	916	33841	1.35	0.10	0.2	1
		y_1	675	18199	0.03	0.00	4.3	48
		K-value	675	18199	5.09	-0.31	5.5	54
Regressed-UNIQUAC	a_{12} & a_{21}	P (bar)	916	33845	0.14	0.00	1.9	14
		T (K)	916	33845	1.29	0.08	0.2	2
		y_1	675	18199	0.03	0.00	4.1	50
		K-value	675	18199	4.69	-0.24	5.3	50
Regressed-Wilson	a_{12} & a_{21}	P (bar)	916	33841	0.19	-0.01	1.9	17
		T (K)	916	33841	1.35	0.13	0.2	2
		y_1	675	18199	0.03	0.00	4.0	49
		K-value	675	18199	3.92	-0.15	5.2	56

Table 3.2. VLE property predictions of the Ideal Solution model and representation capability of the NRTL, UNIQUAC and Wilson models for systems containing aqueous systems

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
Ideal Solution	None	P (bar)	55	4303	1.91	-0.40	27.6	71
		T (K)	55	4303	15.54	9.52	3.0	13
		y_1	47	2313	0.22	-0.03	40.4	100
		K-values	47	2313	23.15	-7.78	46.5	100
Regressed-NRTL	a_{12} & a_{21}	P (bar)	55	4344	0.40	-0.02	4.8	12
		T (K)	55	4344	2.47	0.41	0.4	1
		y_1	47	2313	0.06	-0.01	10.6	48
		K-values	47	2313	17.40	-3.98	11.6	48
Regressed-UNIQUAC	a_{12} & a_{21}	P (bar)	55	4344	0.33	-0.01	4.2	13
		T (K)	55	4344	2.24	0.30	0.4	1
		y_1	47	2313	0.06	-0.01	9.4	50
		K-values	47	2313	15.76	-3.12	10.5	50
Regressed-Wilson	a_{12} & a_{21}	P (bar)	55	4344	0.59	-0.07	5.0	17
		T (K)	55	4344	2.74	0.68	0.4	2
		y_1	47	2313	0.06	0.00	9.3	49
		K-values	47	2313	14.72	-2.17	10.8	56

Table 3.3. Comparison of pressure representations of the NRTL, UNIQUAC and Wilson models for 13 functional-group interactions where the UNIQUAC model provided more than 20% lower %AADs compared to the NRTL and Wilson models

No	Chemical Class 1	Chemical Class 2	No. of sys.	No. of pts.	%AAD on pressure		
					Regressed-UNIQUAC	Regressed-NRTL	Regressed-Wilson
1	Alcohol	Water	9	305	1.8	2.2	2.4
2	Aldehyde	Water	1	7	0.4	0.5	9.6
3	Alkene	Nitro Compound	2	20	1.5	1.9	1.9
4	Amide	Chloroalkene	1	13	1.1	1.7	1.9
5	Amine	Sulfide	1	9	2.8	5.0	4.9
6	Aromatic Bromo	Aromatic Floro	1	9	0.3	0.7	0.6
7	Aromatic Bromo	Sulfide	1	10	0.8	1.7	1.3
8	Aromatic Floro	Toluene Derivative	1	12	0.8	1.3	1.1
9	Carboxylate	Water	3	217	5.3	6.4	6.6
10	Ether	Thiophene	1	35	0.4	0.8	0.5
11	Ketone	Water	5	1604	4.2	5.6	5.1
12	Nitrile	Sulfide	1	9	1.1	1.5	2.6
13	Nitro Compound	Sulfide	2	18	3.3	5.1	4.6
				Average	1.8	2.7	3.3

Table 3.4. Comparison of pressure representations of the NRTL, UNIQUAC and Wilson models for 11 functional-group interactions where the Wilson model provided more than 20% higher %AADs compared to the NRTL and UNIQUAC models

No	Chemical Class 1	Chemical Class 2	No. of sys.	No. of pts.	%AAD on pressure		
					Regressed-UNIQUAC	Regressed-NRTL	Regressed-Wilson
1	Aldehyde	Water	1	7	0.4	0.5	9.6
2	Alkane	Amide	6	175	3.6	3.5	4.9
3	Alkane	Furfural	3	45	6.4	6.1	9.0
4	Alkene	Furfural	1	18	3.8	3.8	6.1
5	Bromoalkane	Water	1	9	1.2	1.1	1.8
6	Nitrile	Pyridine Derivative	1	21	0.7	0.5	0.9
7	Nitrile	Sulfide	1	9	1.1	1.5	2.6
8	Nitrite	Water	1	37	3.0	3.1	4.9
9	Nitro Compound	Water	2	63	4.7	4.6	7.3
10	Sulfide	Water	1	478	4.3	3.7	7.3
11	Thiol	Water	1	10	5.6	3.1	9.4
				Average	3.2	2.9	5.8

Table 3.5. The descriptors used as inputs for the ANNs in the final ensemble for estimating the NRTL model parameters

Descriptor name	Descriptor description	Source	Type of descriptor
SM3_G/D	spectral moment of order 3 from distance/distance matrix	DR	3D matrix-based descriptors
nROH	number of hydroxyl groups	DR	Functional group counts
BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)	DR	Molecular properties
HACA-2/SQRT(TMSA) [Zefirov's PC]	HACA-2/SQRT(TMSA) [Zefirov's PC]	CO	Electrostatic
SM4_X	spectral moment of order 4 from chi matrix	DR	2D matrix-based descriptors
Max 1-electron react. index for a O atom	Max 1-electron react. index for a O atom	CO	Quantum Chemical
GATS1e	Geary autocorrelation of lag 1 weighted by Sanderson electronegativity	DR	2D autocorrelations
TDB02e	3D Topological distance based descriptors - lag 2 weighted by Sanderson electronegativity	DR	3D autocorrelations
Max partial charge for a N atom [Zefirov's PC]	Max partial charge for a N atom [Zefirov's PC]	CO	Electrostatic
Min (>0.1) bond order of a O atom	Min (>0.1) bond order of a O atom	CO	Quantum Chemical
HATS0e	leverage-weighted autocorrelation of lag 0 / weighted by Sanderson electronegativity	DR	GETAWAY descriptors
Mor03i	signal 03 / weighted by ionization potential	DR	3D-MoRSE descriptors
Min e-e repulsion for a C atom	Min e-e repulsion for a C atom	CO	Quantum Chemical
P_VSA_LogP_5	P_VSA-like on LogP, bin 5	DR	P_VSA-like descriptors
HOMO - LUMO energy gap	HOMO - LUMO energy gap	CO	Quantum Chemical
Ui	unsaturation index	DR	Molecular properties
DLS_03	modified drug-like score from Walters et al. (6 rules)	DR	Drug-like indices
HBCA H-bonding charged surface area [Quantum-Chemical PC]	HBCA H-bonding charged surface area [Quantum-Chemical PC]	CO	Quantum Chemical
HACA-2/SQRT(TMSA) [Quantum-Chemical PC]	HACA-2/SQRT(TMSA) [Quantum-Chemical PC]	CO	Quantum Chemical
HACA-1 [Quantum-Chemical PC]	HACA-1 [Quantum-Chemical PC]	CO	Quantum Chemical
AAC	mean information index on atomic composition	DR	Information indices
Max atomic orbital electronic population	Max atomic orbital electronic population	CO	Quantum Chemical
F02[C-C]	Frequency of C - C at topological distance 2	DR	2D Atom Pairs
ITH	total information content on the leverage equality	DR	GETAWAY descriptors
C-028	R--CR--X	DR	Atom-centred fragments
TDB05e	3D Topological distance based descriptors - lag 5 weighted by Sanderson electronegativity	DR	3D autocorrelations
SpMaxA_B(m)	normalized leading eigenvalue from Burden matrix weighted by mass	DR	2D matrix-based descriptors
PNSA-3 Atomic charge weighted PNSA [Quantum-Chemical PC]	PNSA-3 Atomic charge weighted PNSA [Quantum-Chemical PC]	CO	Quantum Chemical
H-050	H attached to heteroatom	DR	Atom-centred fragments
NssO	Number of atoms of type ssO	DR	Atom-type E-state indices

Table 3.6. The descriptors used as inputs for the ANNs in the final ensemble for estimating the UNIQUAC model parameters

Descriptor name	Descriptor description	Source	Type of descriptor
HA dependent HDCA-1 [Zefirov's PC]	HA dependent HDCA-1 [Zefirov's PC]	CO	Electrostatic
BLTA96	Verhaar Algae base-line toxicity from MLOGP (mmol/l)	DR	Molecular properties
Psi_e_1s	electrotopological state pseudoconnectivity index - type 1s	DR	Topological indices
SpMaxA_X	normalized leading eigenvalue from chi matrix	DR	2D matrix-based descriptors
MLOGP2	squared Moriguchi octanol-water partition coeff. (logP^2)	DR	Molecular properties
P_VSA_LogP_5	P_VSA-like on LogP, bin 5	DR	P_VSA-like descriptors
Min n-n repulsion for a H-O bond	Min n-n repulsion for a H-O bond	CO	Quantum Chemical
GATS1e	Geary autocorrelation of lag 1 weighted by Sanderson electronegativity	DR	2D autocorrelations
WiA_B(p)	average Wiener-like index from Burden matrix weighted by polarizability	DR	2D matrix-based descriptors
P_VSA_MR_6	P_VSA-like on Molar Refractivity, bin 6	DR	P_VSA-like descriptors
MATS4i	Moran autocorrelation of lag 4 weighted by ionization potential	DR	2D autocorrelations
Mor04m	signal 04 / weighted by mass	DR	3D-MoRSE descriptors
Max net atomic charge for a Cl atom	Max net atomic charge for a Cl atom	CO	Quantum Chemical
HOMO - LUMO energy gap	HOMO - LUMO energy gap	CO	Quantum Chemical
SsNH2	Sum of sNH2 E-states	DR	Atom-type E-state indices
CSI	eccentric connectivity index	DR	Topological indices
HACA-2/SQRT(TMSA) [Zefirov's PC]	HACA-2/SQRT(TMSA) [Zefirov's PC]	CO	Electrostatic
O%	percentage of O atoms	DR	Constitutional indices
H-049	H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	DR	Atom-centred fragments
HBCA H-bonding charged surface area [Quantum-Chemical PC]	HBCA H-bonding charged surface area [Quantum-Chemical PC]	CO	Quantum Chemical
nBM	number of multiple bonds	DR	Constitutional indices
ALOGP2	squared Ghose-Crippen octanol-water partition coeff. (logP^2)	DR	Molecular properties
REIG	first eigenvalue of the R matrix	DR	GETAWAY descriptors
R4u+	R maximal autocorrelation of lag 4 / unweighted	DR	GETAWAY descriptors
Kier&Hall index (order 1)	Kier&Hall index (order 1)	CO	Topological
X1v	valence connectivity index of order 1	DR	Connectivity indices
FNSA-3 Fractional PNSA (PNSA-3/TMSA) [Quantum-Chemical PC]	FNSA-3 Fractional PNSA (PNSA-3/TMSA) [Quantum-Chemical PC]	CO	Quantum Chemical
RNCS Relative negative charged SA (SAMNEG*RNCG) [Quantum-Chemical PC]	RNCS Relative negative charged SA (SAMNEG*RNCG) [Quantum-Chemical PC]	CO	Quantum Chemical
RTp	R total index / weighted by polarizability	DR	GETAWAY descriptors
Max 1-electron react. index for a Cl atom	Max 1-electron react. index for a Cl atom	CO	Quantum Chemical

Table 3.7. The descriptors used as inputs for the ANNs in the final ensemble for estimating the Wilson model parameters

Descriptor name	Descriptor description	Source	Type of descriptor
MLOGP	Moriguchi octanol-water partition coeff. (logP)	DR	Molecular properties
HACA-2/TMSA [Zefirov's PC]	HACA-2/TMSA [Zefirov's PC]	CO	Electrostatic
nROH	number of hydroxyl groups	DR	Functional group counts
AAC	mean information index on atomic composition	DR	Information indices
MLOGP2	squared Moriguchi octanol-water partition coeff. (logP^2)	DR	Molecular properties
R1v+	R maximal autocorrelation of lag 1 / weighted by van der Waals volume	DR	GETAWAY descriptors
RTi+	R maximal index / weighted by ionization potential	DR	GETAWAY descriptors
HATS1p	leverage-weighted autocorrelation of lag 1 / weighted by polarizability	DR	GETAWAY descriptors
Polarity parameter / square distance	Polarity parameter / square distance	CO	Electrostatic
H-046	H attached to C0(sp3) no X attached to next C	DR	Atom-centred fragments
Max total interaction for a C-C bond	Max total interaction for a C-C bond	CO	Quantum Chemical
HOMO - LUMO energy gap	HOMO - LUMO energy gap	CO	Quantum Chemical
Min (>0.1) bond order of a N atom	Min (>0.1) bond order of a N atom	CO	Quantum Chemical
CATS2D_01_LL	CATS2D Lipophilic-Lipophilic at lag 01	DR	CATS 2D
HOMO energy	HOMO energy	CO	Quantum Chemical
FNSA-3 Fractional PNSA (PNSA-3/TMSA) [Quantum-Chemical PC]	FNSA-3 Fractional PNSA (PNSA-3/TMSA) [Quantum-Chemical PC]	CO	Quantum Chemical
HDCA H-donors charged surface area [Quantum-Chemical PC]	HDCA H-donors charged surface area [Quantum-Chemical PC]	CO	Quantum Chemical
HA dependent HDCA-1/TMSA [Quantum-Chemical PC]	HA dependent HDCA-1/TMSA [Quantum-Chemical PC]	CO	Quantum Chemical
H-050	H attached to heteroatom	DR	Atom-centred fragments
X5A	average connectivity index of order 5	DR	Connectivity indices
HATS2u	leverage-weighted autocorrelation of lag 2 / unweighted	DR	GETAWAY descriptors
FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]	CO	Electrostatic
SsOH	Sum of sOH E-states	DR	Atom-type E-state indices
Max electroph. react. index for a O atom	Max electroph. react. index for a O atom	CO	Quantum Chemical
Min e-e repulsion for a H atom	Min e-e repulsion for a H atom	CO	Quantum Chemical
F01[N-O]	Frequency of N - O at topological distance 1	DR	2D Atom Pairs
Polarity parameter (Qmax-Qmin)	Polarity parameter (Qmax-Qmin)	CO	Electrostatic
nCp	number of terminal primary C(sp3)	DR	Functional group counts
NssO	Number of atoms of type ssO	DR	Atom-type E-state indices
SM5_G	spectral moment of order 5 from geometrical matrix	DR	3D matrix-based descriptors

Table 3.8. Predictions from the NRTL-QSPR case study

Data set	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD	%AAD multiplier
Training Set	P (bar)	460	20301	0.13	0.00	3.6	45	1.8
	T (K)	460	20301	2.08	0.11	0.3	5	1.8
	y_1	339	10187	0.03	0.00	5.2	49	1.2
	K-values	339	10187	5.36	-0.23	6.7	56	1.2
Validation Set	P (bar)	167	5101	0.24	0.01	5.0	23	2.4
	T (K)	167	5101	2.33	0.17	0.4	2	2.4
	y_1	117	2910	0.04	0.00	6.4	50	1.5
	K-values	117	2910	7.21	-0.98	7.8	50	1.5
Internal Test Set	P (bar)	101	2702	0.17	-0.01	5.4	25	2.5
	T (K)	101	2702	3.40	0.54	0.5	2	2.4
	y_1	77	1475	0.05	0.00	7.1	45	1.6
	K-values	77	1475	5.95	-0.67	8.6	100	1.5
External Test Set	P (bar)	188	5741	0.45	0.02	4.6	24	2.3
	T (K)	188	5741	2.60	0.19	0.4	3	2.3
	y_1	142	3627	0.04	0.00	6.5	57	1.7
	K-values	142	3627	1.99	-0.12	8.1	64	1.6
Highly non-ideal	P (bar)	348	14929	0.39	0.00	5.6	45	2.1
	T (K)	348	14929	3.10	0.48	0.5	5	2.1
	y_1	262	8203	0.05	0.00	7.4	57	1.4
	K-values	262	8203	8.69	-0.98	9.1	100	1.4
Water systems	P (bar)	55	4344	0.41	-0.03	9.0	29	1.9
	T (K)	55	4344	4.71	1.35	0.8	3	2.0
	y_1	47	2313	0.09	-0.01	16.0	57	1.5
	K-values	47	2313	18.72	-4.89	17.4	100	1.5
All data	P (bar)	916	33845	0.25	0.00	4.3	45	2.1
	T (K)	916	33845	2.41	0.19	0.4	5	2.1
	y_1	675	18199	0.04	0.00	5.9	57	1.4
	K-values	675	18199	5.44	-0.39	7.4	100	1.3

Table 3.9. Predictions from the UNIQUAC-QSPR case study

Data set	**Property**	**No. of sys.**	**No. of pts.**	**RMSE**	**Bias**	**%AAD**	**Max %AAD**	**%AAD multiplier**
Training Set	P (bar)	460	20301	0.16	0.01	4.4	67	2.3
	T (K)	460	20301	2.61	0.08	0.4	9	2.3
	y_1	339	10187	0.03	0.00	5.2	49	1.2
	K-values	339	10187	5.38	-0.15	7.3	53	1.3
Validation Set	P (bar)	167	5101	0.21	0.00	6.4	41	3.3
	T (K)	167	5101	2.83	0.08	0.5	3	3.3
	y_1	117	2910	0.04	0.00	6.4	50	1.6
	K-values	117	2910	7.12	-0.92	9.3	53	1.8
Internal Test Set	P (bar)	101	2702	0.13	-0.01	5.2	22	2.6
	T (K)	101	2702	3.21	0.52	0.5	3	2.5
	y_1	77	1475	0.05	0.00	7.1	45	1.6
	K-values	77	1475	5.93	-0.67	8.9	85	1.6
External Test Set	P (bar)	188	5741	0.31	0.00	5.8	32	3.2
	T (K)	188	5741	3.24	0.15	0.5	4	3.2
	y_1	142	3627	0.04	0.00	6.5	57	1.7
	K-values	142	3627	1.89	-0.11	8.8	66	1.9
Highly non-ideal	P (bar)	348	14929	0.30	-0.01	6.7	67	2.7
	T (K)	348	14929	3.72	0.58	0.6	9	2.8
	y_1	262	8203	0.05	0.00	7.4	57	1.5
	K-values	262	8203	8.70	-0.86	9.8	85	1.6
Water systems	P (bar)	55	4344	0.37	-0.03	10.2	33	2.4
	T (K)	55	4344	5.18	0.99	0.9	5	2.5
	y_1	47	2313	0.09	-0.01	16.0	57	1.7
	K-values	47	2313	18.78	-4.29	17.5	85	1.7
All data	P (bar)	916	33845	0.21	0.01	5.1	67	2.7
	T (K)	916	33845	2.86	0.14	0.4	9	2.7
	y_1	675	18199	0.04	0.00	5.9	57	1.4
	K-values	675	18199	5.45	-0.33	8.1	85	1.5

Table 3.10. Predictions from the Wilson-QSPR case study

Data set	**Property**	**No. of sys.**	**No. of pts.**	**RMSE**	**Bias**	**%AAD**	**Max %AAD**	**%AAD multiplier**
Training Set	P (bar)	460	20300	0.16	0.01	3.4	41	1.7
	T (K)	460	20300	1.85	0.04	0.3	4	1.7
	y_1	339	10187	0.03	0.00	5.1	39	1.2
	K-values	339	10187	4.91	-0.11	6.5	53	1.2
Validation Set	P (bar)	167	5101	0.48	-0.02	5.4	55	2.9
	T (K)	167	5101	2.93	0.12	0.5	4	2.9
	y_1	117	2910	0.04	0.00	6.3	44	1.7
	K-values	117	2910	7.02	-0.80	7.7	46	1.6
Internal Test Set	P (bar)	101	2702	0.18	0.00	5.1	21	2.6
	T (K)	101	2702	3.05	0.46	0.4	2	2.5
	y_1	77	1475	0.05	0.00	7.1	39	1.7
	K-values	77	1475	5.90	-0.66	8.6	81	1.6
External Test Set	P (bar)	188	5741	0.42	-0.01	5.0	35	2.7
	T (K)	188	5741	2.80	0.11	0.4	4	2.7
	y_1	142	3627	0.04	0.00	6.6	89	1.8
	K-values	142	3627	1.92	-0.02	8.3	89	1.8
Highly non-ideal	P (bar)	348	14928	0.47	-0.02	5.5	41	2.2
	T (K)	348	14928	2.94	0.50	0.5	4	2.2
	y_1	262	8203	0.05	0.00	7.3	89	1.5
	K-values	262	8203	8.29	-0.73	8.7	89	1.5
Water systems	P (bar)	55	4344	0.80	-0.10	8.3	55	1.7
	T (K)	55	4344	4.88	0.72	0.7	4	1.7
	y_1	47	2313	0.08	0.00	15.2	89	1.6
	K-values	47	2313	17.74	-3.68	15.4	89	1.4
All data	P (bar)	916	33844	0.31	0.00	4.3	55	2.2
	T (K)	916	33844	2.43	0.12	0.4	4	2.2
	y_1	675	18199	0.04	0.00	5.8	89	1.5
	K-values	675	18199	5.19	-0.27	7.3	89	1.4

Table 3.11. Comparison of *a priori* predictions of the NRTL-QSPR, UNIQUAC-QSPR, Wilson-QSPR and UNIFAC-2006 case studies

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
NRTL-QSPR	Generalized a_{12} & a_{21}	P (bar)	916	33845	0.25	0.00	4.3	45
		T (K)	916	33845	2.41	0.19	0.4	5
		y_1	675	18199	0.04	0.00	5.9	57
		K-value	675	18199	5.44	-0.39	7.4	100
UNIQUAC-QSPR	Generalized a_{12} & a_{21}	P (bar)	916	33845	0.21	0.01	5.1	67
		T (K)	916	33845	2.86	0.14	0.4	9
		y_1	675	18199	0.04	0.00	5.9	57
		K-value	675	18199	5.45	-0.33	8.1	85
Wilson-QSPR	Generalized a_{12} & a_{21}	P (bar)	916	33844	0.31	0.00	4.3	55
		T (K)	916	33844	2.43	0.12	0.4	4
		y_1	675	18199	0.04	0.00	5.8	89
		K-value	675	18199	5.19	-0.27	7.3	89
UNIFAC-2006	UNIFAC - All interactions present	P (bar)	853[a]	31609	0.51	0.00	5.1	100
		T (K)	853	31609	4.74	-0.06	0.4	25
		y_1	634	17056	0.04	0.00	5.6	100
		K-value	633	17045	6.03	0.11	6.9	100
UNIFAC-2006	UNIFAC - One or more missing interactions	P (bar)	45	1226	0.36	-0.05	11.2	71
		T (K)	45	1226	8.64	1.46	1.1	13
		y_1	30	893	0.07	0.00	13.3	50
		K-value	30	893	1.33	-0.09	15.1	100

[a] Due to a lack of group interaction parameters, 63 systems of the 916 systems were not considered.

Table 3.12. Infinite-dilution activity coefficient representation and prediction of various activity coefficient models

Property	Model	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
Infinite dilution activity coefficient ($\gamma\infty$)	Regressed-NRTL	137	549	3.54	-0.21	8.7	84
	Regressed-UNIQUAC			2.33	0.00	8.2	81
	Regressed-Wilson			2.36	0.38	8.7	73
	NRTL-QSPR	137	549	6.70	-1.36	19.0	104
	UNIQUAC-QSPR			7.70	-1.62	20.9	111
	Wilson-QSPR			6.06	-0.99	23.0	143
	UNIFAC-2006			3.30	0.33	12.2	145

Table 3.13. Infinite-dilution activity coefficient representation of the NRTL, UNIQUAC and Wilson models for systems where the UNIQUAC model resulted in more than 50% lower %AADs compared to the NRTL and Wilson models

No	Compound 1	Compound 2	No. of pts.	%AAD on $\gamma\infty$		
				Regressed-UNIQUAC	Regressed-NRTL	Regressed-Wilson
1	Acetonitrile	Butane	1	0.01	0.7	0.1
2	Benzene	Triethylamine	6	4.1	9.6	6.4
3	Ethanol	Triethylamine	1	0.3	2.6	2.8
4	Benzene	Nitrobenzene	6	0.1	0.4	1.6
5	Acetonitrile	Methyl cyclohexane	2	0.02	0.9	4.4
6	Hexane	Nitrobenzene	6	1.8	15.9	7.1
7	Hexane	Ethanol	16	3.2	9.6	13.5
8	Methanol	Benzene	3	2.2	5.8	8.5
9	Dichloromethane	Triethylamine	1	0.03	8.6	7.0
10	Chloroform	Triethylamine	2	4.8	8.1	7.7
11	Carbondisulfide	Acetonitrile	1	0.4	1.7	4.9
12	p-Xylene	Ethyl acetate	1	3.0	5.4	6.8
13	Acetone	Water	3	0.5	16.6	7.2
14	Ethanol	Chlorobenzene	3	1.4	8.1	2.1
			Average	1.6	6.7	5.7

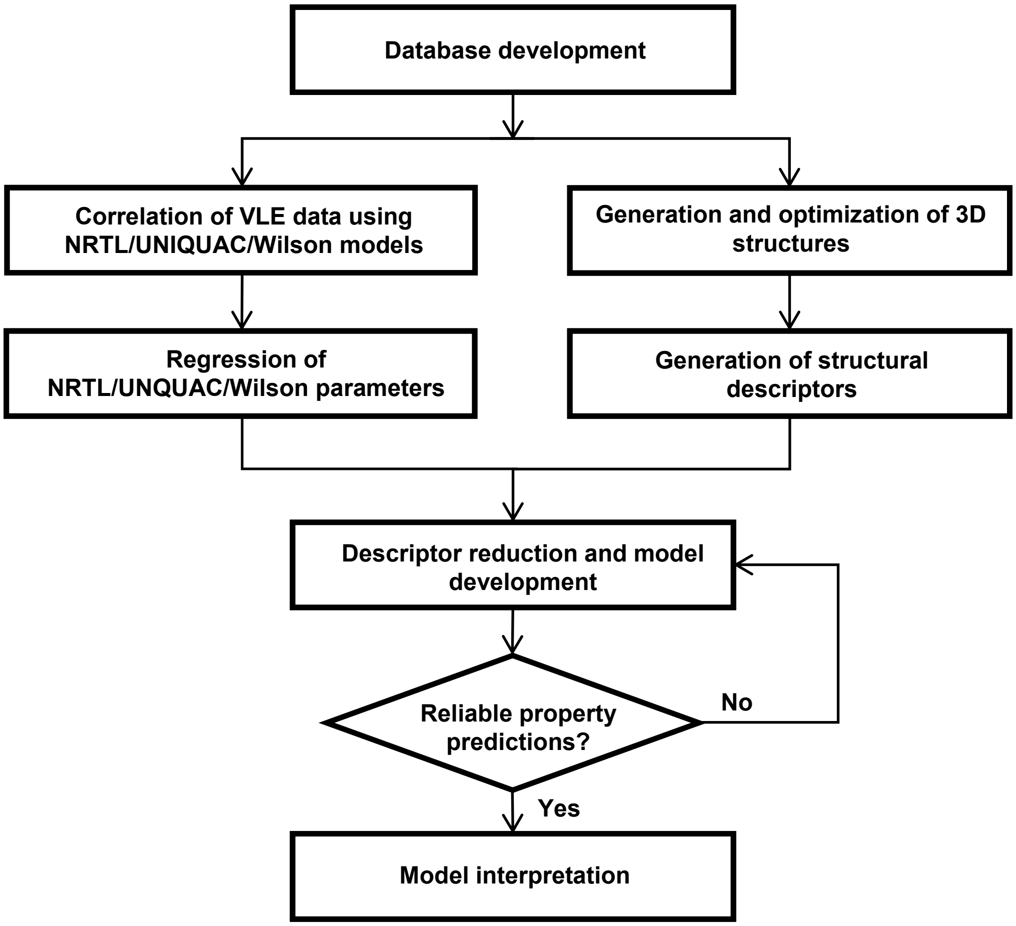

Database development
Correlation of VLE data using NRTL/UNIQUAC/Wilson models
Generation and optimization of 3D structures
Regression of NRTL/UNQUAC/Wilson parameters
Generation of structural descriptors
Descriptor reduction and model development
Reliable property predictions?
No
Yes
Model interpretation

Figure 3.1. Schematic of the QSPR model development process

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	13																														
2	Aldehyde	10	1																													
3	Alkane	24	5	14																												
4	Alkene	9	1	10	3																											
5	Alkyne	5	3	5	6	2																										
6	Amide	6	2	6	2	1																										
7	Amine	5		4			3	4																								
8	Aromatic Bromo	5		3			1																									
9	Aromatic Floro	2		2			1	1	1																							
10	Benzene Derivative	6	3	13	5		1	5	1	3	4																					
11	Bromoalkane	15		5					1	1	8																					
12	Carboxylate	2	5	9	1						6	1	3																			
13	Chloroalkane	5		5	2	2	4	6		2	8	3	4	2																		
14	Chloroalkene	19	1	7		1	1				1		1	8	1																	
15	Chlorobenzene	9		2	2		1	4	1	1	2	1		2	1																	
16	Epoxide	7	3	6							1		2	4																		
17	Ester	1	1	8	1	1	1	1			4	1	1	5	1	1	1															
18	Ether	12	2	21	3	3	2	2		3	5	2	1	9	2	2	1	3	3													
19	Furfural	1		3	1						2			4	1			1														
20	H2S	1		1																												
21	Iodoalkane	3	1	1							2	1	1	4				1	1													
22	Ketone	3	4	21	3	1	2	5	1		8	1	6	8	7	3	1	3	2	2		1	4									
23	Nitrile	4		4	2	2	1	1	1		4		4	6	3	1		1	1				1	1								
24	Nitrite	1													1									1								
25	Nitro Compound	12		3	2	2	1		1		5	1	2	5	1	2		3	3			2	3	2		2						
26	Pyridine Derivative	14		4			1			1	1	1	1	2	1	1		2	1				1	1			1					
27	Sulfide	4		4	3	3	1	1	1		1	2	2	5	2	1		1	1		1	1	1	1		2						
28	Thiol	1		7			2	1			1								1		1		1	1				3				
29	Thiophene	4		1	2		1				1			1					1							1	1					
30	Toluene Derivative	3	6	4	2		1	2	1	1	2	3	2	3	1	1	1		5	1		1	5	1		2	2	2	2	1	1	
31	Water	9	1	2			1	10			3	1	3		1		2	1	4	1			5	3	1	2	3	1	1			

X

Y ■# Number of available binary systems consisting of chemicals with functional groups of X and Y

□ No VLE data used

Figure 3.2. Database matrix of the compounds in the OSU-VLE database III

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	2 2 2																														
2	Aldehyde	3 3 3	1 1 1																													
3	Alkane	4 4 3	1 1 1	1 1 1																												
4	Alkene	4 3 3	2 2 2	1 1 1	1 1 1																											
5	Alkyne	3 3 2	1 1 1	1 1 1	0 0 0	1 1 1																										
6	Amide	2 2 2	3 3 3	4 4 5	2 2 1	1 1 1																										
7	Amine	2 1 1		3 3 3			2 2 2	1 1 1																								
8	Aromatic Bromo	2 2 2		1 1 1			6 3 4																									
9	Aromatic Floro	3 3 2		0 0 0			3 3 3	0 1 0	1 0 1																							
10	Benzene Derivative	2 1 1	2 2 2	1 1 1	2 1 2		3 3 3	2 2 2	2 2 2	2 2 2	1 1 1																					
11	Bromoalkane	2 2 2		2 2 2					0 0 0	1 1 1	3 3 3																					
12	Carboxylate	2 2 2	1 1 1	2 2 2	0 0 0						2 2 2	3 3 2	1 1 1																			
13	Chloroalkane	2 2 2		1 1 1	1 1 0	0 0 0	2 2 2	2 2 2		1 1 1	2 2 2	2 2 2	2 2 2	1 1 1																		
14	Chloroalkene	2 2 2	1 1 1	1 1 1		1 1 1	2 1 2				1 1 1		2 2 2	2 2 2	1 1 1																	
15	Chlorobenzene	3 3 3		1 1 1	2 2 2		3 3 3	1 1 1	0 0 0	0 0 0	0 0 0	1 1 1		1 1 1	2 2 2																	
16	Epoxide	2 1 1	2 2 2	1 1 1							2 2 2		1 1 1	4 3 4																		
17	Ester	6 6 6	1 1 1	1 1 1	0 0 0	0 0 0	5 5 5	2 2 2			2 1 2	1 1 1	3 3 3	1 1 1	1 1 1	1 1 1	5 5 5															
18	Ether	2 2 2	2 1 1	2 2 2	2 2 2	2 1 2	2 2 1	1 1 1		0 0 0	1 1 1	5 5 5	1 1 1	1 1 1	2 2 2	1 1 1	1 1 1	1 1 1	2 2 2													
19	Furfural	3 3 3		6 6 9	4 4 6						3 3 3			3 3 3	1 1 1			1 1 1														
20	H2S	4 4 4		3 2 1																												
21	Iodoalkane	5 4 4	2 2 2	1 1 1							1 1 1	1 1 1	1 1 1	1 1 1				0 0 0	1 0 0													
22	Ketone	1 1 1	1 1 1	2 2 2	1 1 1	1 1 1	2 2 2	1 1 1	1 1 1		2 2 2	1 1 1	1 1 1	1 1 1	2 2 2	2 2 2	1 1 1	1 1 1	0 0 0	1 1 1		2 2 2	1 1 1									
23	Nitrile	0 0 0		3 3 3	2 3 2	4 4 4	2 2 2	1 1 1	1 1 1		2 2 2		1 2 2	2 2 1	2 1 1	2 2 2		0 0 1	1 1 1				1 1 1	2 2 2								
24	Nitrite	2 2 2													3 3 3									4 3 3								
25	Nitro Compound	4 4 4		5 5 5	2 2 2	1 1 1	2 2 2		7 7 7		2 2 2	2 2 2	1 0 0	2 2 2	2 2 1	1 1 1		1 1 1	2 2 2			2 2 2	2 2 2	1 1 1		2 2 2						
26	Pyridine Derivative	2 2 2		1 1 1			0 0 0			1 1 1	2 2 2	8 8 8	0 0 0	4 4 4	1 1 1	0 0 0		2 2 2	2 2 2				1 1 2	0 1 1			0 1 0					
27	Sulfide	2 2 2		1 0 0	1 1 1	3 3 3	1 1 1	5 3 5	2 1 1		2 3 2	2 2 2	1 1 1	3 2 2	1 1 1	2 1 1		2 2 1	2 2 2		1 1 1	1 1 1	2 2 2	2 1 3		5 3 5						
28	Thiol	6 5 5		4 3 3			1 2 2	1 1 1			0 0 0								0 0 0		5 5 5		1 0 0	1 0 0				2 2 2				
29	Thiophene	2 1 1		1 1 1	2 2 2		4 4 4				0 0 0			0 0 0					1 0 0							0 0 0	2 2 2					
30	Toluene Derivative	3 3 3	3 3 3	1 1 1	2 2 2		6 6 6	2 2 2	0 0 0	1 1 1	0 0 0	2 2 2	1 1 1	2 2 2	1 1 1	1 1 1	3 3 3		2 2 2	2 2 2		1 1 1	1 1 1	2 2 1		1 1 1	1 1 1	3 4 3	4 4 4	2 2 2	5 5 5	
31	Water	2 2 2	0 0 10	8 8 10			2 2 2	6 6 5			9 9 8	1 1 2	6 5 7		7 6 4		7 7 8	7 5 3	8 8 9	8 8 6			6 4 5	4 3 4	3 3 5	5 5 7	2 1 1	4 4 7	3 6 9			

Key	
Color	Pressure %AAD Range
#	%AAD<3
#	3<%AAD<6
#	6<%AAD<10
#	10<%AAD<20
#	%AAD>20

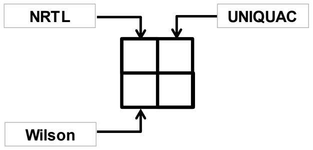


Figure 3.3. Pressure representation of the Regressed-NRTL, Regressed-UNIQUAC and Regressed-Wilson models by type of interactions

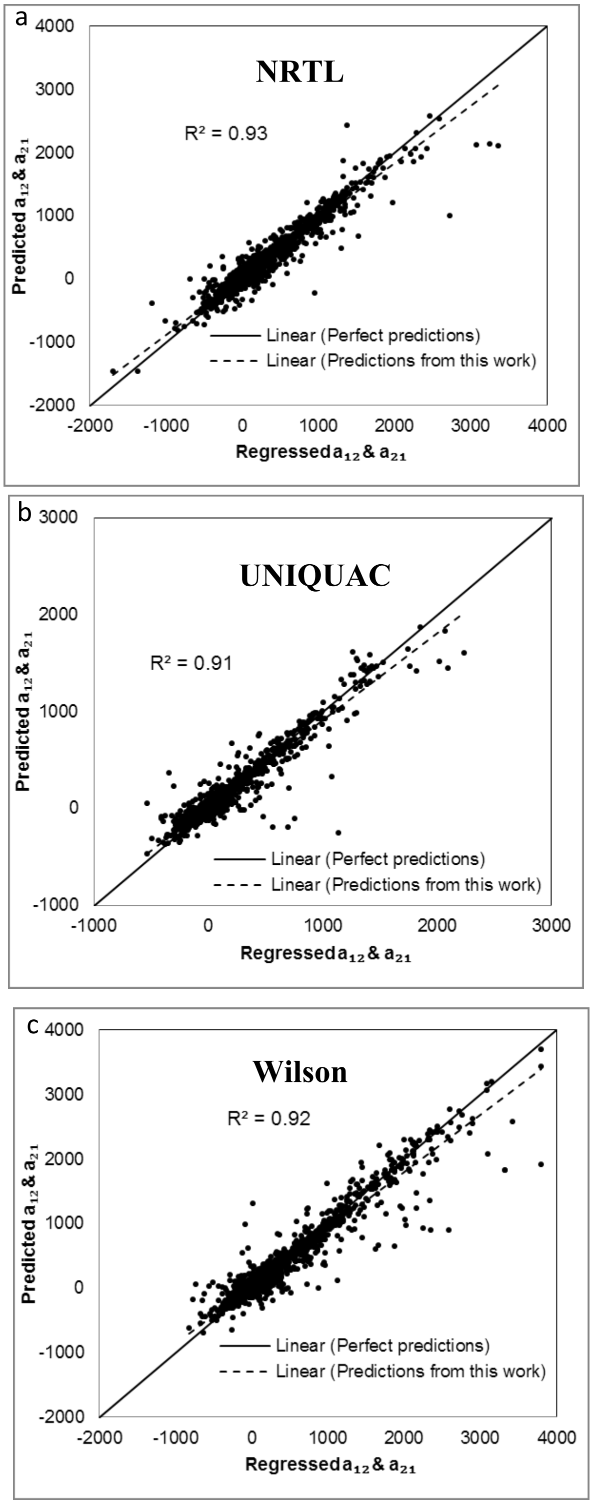


Figure 3.4. Comparison of the regressed and QSPR predicted parameters for (a) NRTL, (b) UNIQUAC and (c) Wilson models

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	3 3 4																														
2	Aldehyde	8 12 8	2 2 2																													
3	Alkane	5 4 5	3 4 4	2 3 3																												
4	Alkene	4 6 4	2 6 2	1 2 1	2 2 2																											
5	Alkyne	4 4 2	3 3 3	5 4 4	1 1 1	3 3 2																										
6	Amide	4 6 6	6 6 4	5 10 6	3 5 5	4 3 4																										
7	Amine	5 12 7		3 10 5			3 6 4	4 3 3																								
8	Aromatic Bromo	4 6 4		3 5 2			6 6 6																									
9	Aromatic Floro	8 10 8		4 4 2			2 6 2	1 3 1	4 4 1																							
10	Benzene Derivative	2 2 3	3 2 2	2 2 2	6 2 4		4 6 6	1 3 3	2 2 2	3 3 4	1 1 2																					
11	Bromoalkane	7 4 5		7 12 9					1 1 1	6 3 1	4 4 5																					
12	Carboxylate	5 4 3	2 3 2	5 4 4	1 2 4						4 4 4	4 16 14	3 3 3																			
13	Chloroalkane	4 4 4		2 7 2	3 1 1	6 8 4	6 10 5	7 8 9		1 3 2	4 4 5	3 6 4	4 4 3	2 2 2																		
14	Chloroalkene	4 4 5	4 2 3	4 4 4		2 2 1	3 5 6				1 1 1		6 8 5	5 3 4	1 3 3																	
15	Chlorobenzene	4 5 4		3 3 3	5 6 3		4 3 5	2 4 4	1 2 0	4 4 2	0 2 1	3 1 2		2 8 2	5 5 5																	
16	Epoxide	4 5 3	2 2 2	2 5 6							12 3 2		2 2 3	5 5 5																		
17	Ester	7 4 1	2 1 3	2 3 2	2 4 1	7 10 3	4 5 3	2 3 4			4 2 3	4 3 3	4 6 6	2 2 2	1 2 1	1 2 1	13 18 15															
18	Ether	4 3 4	2 6 2	3 4 3	4 4 2	2 3 3	5 4 4	6 5 8		3 4 2	3 2 2	10 8 6	2 3 1	2 4 3	3 8 5	4 3 2	3 3 2	2 2 2	3 4 4													
19	Furfural	3 5 4		6 10 10	8 10 8						3 4 3			7 12 12	4 8 3			1 6 2														
20	H2S	4 5 6		4 10 1																												
21	Iodoalkane	7 7 7	3 1 5	1 2 2							4 3 4	1 4 1	2 2 5	3 5 2				0 1 2	1 4 5													
22	Ketone	2 2 2	1 2 1	3 3 3	2 4 4	4 2 1	7 5 5	4 4 4	1 2 2		3 2 3	1 2 1	2 2 2	2 3 2	4 5 4	2 4 3	3 2 4	1 1 2	3 4 3	2 2 5		3 3 7	1 1 2									
23	Nitrile	3 4 3		4 5 4	6 7 5	10 12 8	2 5 4	3 3 2	4 5 5		4 3 5		3 3 3	2 6 3	3 7 4	4 4 6		1 2 1	2 6 1				4 3 3	2 2 3								
24	Nitrite	3 2 3													4 4 3									6 4 4								
25	Nitro Compound	4 6 5		9 9 7	2 5 2	10 7 7	2 5 2		7 7 7		4 4 6	3 3 2	1 2 1	4 6 4	4 2 2	1 5 2		2 4 1	2 3 3			6 9 6	6 6 4	2 1 2		6 3 4						
26	Pyridine Derivative	3 6 3		4 12 9			1 1 5			2 1 2	2 2 2	8 11 9	0 2 1	12 9 7	1 2 1	0 2 2		3 3 3	4 2 3				2 1 1	1 5 2			1 2 1					
27	Sulfide	4 9 5		2 3 3	1 7 14	8 13 5	2 3 3	6 4 5	4 4 6		3 6 2	9 18 18	2 3 2	5 5 5	2 3 1	7 9 4		4 11 7	3 2 2		5 3 5	2 1 1	2 4 2	3 5 5		8 7 8						
28	Thiol	9 11 7		5 5 4			4 10 2	1 4 2			2 4 1								1 1 1		5 5 5		1 1 2	3 2 2				5 5 4				
29	Thiophene	2 8 8		3 1 2	3 4 2		6 4 6				1 1 2			1 3 0					1 5 2							1 3 0	1 1 3					
30	Toluene Derivative	4 3 4	12 11 11	3 2 2	2 3 2		6 7 7	3 3 3	2 1 1	4 5 4	1 0 1	3 4 2	1 3 1	2 3 2	3 3 2	1 5 1	4 7 5		4 4 4	2 3 3		2 3 4	2 2 3	3 1 1		2 2 3	2 1 2	5 6 6	5 7 4	2 3 2	12 5 10	
31	Water	4 4 3	6 3 3	10 10 12			4 8 2	8 11 22			11 14 9	1 1 1	10 9 8		15 9 4		14 9 9	8 4 4	9 10 11	8 8 7			9 9 6	8 10 11	6 7 5	9 12 9	4 8 4	14 21 18	21 33 17			

Key	
Color	Pressure %AAD Range
#	%AAD<3
#	3<%AAD<6
#	6<%AAD<10
#	10<%AAD<20
#	%AAD>20

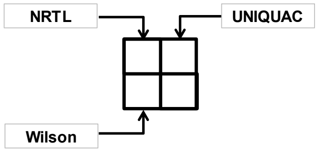


Figure 3.5. Pressure predictions of regressed NRTL-QSPR, UNIQUAC-QSPR and Wilson-QSPR models by type of interactions

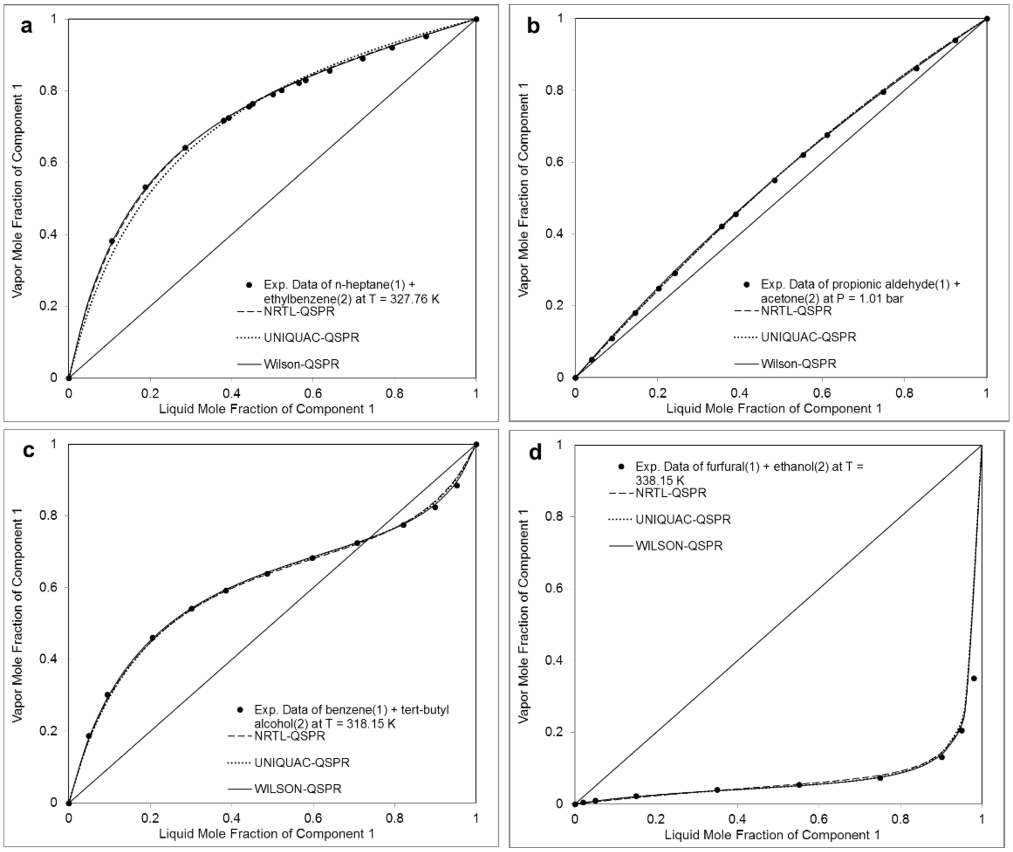


Figure 3.6. QSPR equilibrium phase composition predictions for (a) n-heptane (1) + ethylbenzene (2), (b) propionic aldehyde (1) + acetone (2), (c) benzene (1) + tert-butyl alcohol (2) and (d) furfural (1) + ethanol (2)

REFERENCES

1. Renon, H. and J.M. Prausnitz, *Local compositions in thermodynamic excess functions for liquid mixtures.* AIChE Journal, 1968. **14**(1): p. 135-144.
2. Gmehling, J., J. Li, and M. Schiller, *A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties.* Industrial & Engineering Chemistry Research, 1993. **32**(1): p. 178-193.
3. Abrams, D.S. and J.M. Prausnitz, *Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems.* AIChE Journal, 1975. **21**(1): p. 116-128.
4. Skjold-Jorgensen, S., B. Kolbe, J. Gmehling, and P. Rasmussen, *Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(4): p. 714-722.
5. Fischer, K. and J. Gmehling, *Further development, status and results of the PSRK method for the prediction of vapor-liquid equilibria and gas solubilities.* Fluid Phase Equilibria, 1996. **121**(1-2): p. 185-206.
6. Wilson, G.M., *Vapor-liquid equilibrium. XI. A new expression for the excess free energy of mixing.* Journal of the American Chemical Society, 1964. **86**(2): p. 127-130.
7. Prausnitz, J.M., R.N. Lichtenthaler, and E.G.d. Azevedo, *Molecular thermodynamics of fluid-phase equilibria*. 3rd ed. 1998: Prentice-Hall.
8. Gmehling, J., D. Tiegs, and U. Knipp, *A comparison of the predictive capability of different group contribution methods.* Fluid Phase Equilibria, 1990. **54**: p. 147-165.

9. Jakob, A., H. Grensemann, J. Lohmann, and J. Gmehling, *Further development of modified UNIFAC (Dortmund): Revision and extension 5.* Industrial & Engineering Chemistry Research, 2006. **45**(23): p. 7924-7933.

10. Prausnitz, J.M. and F.W. Tavares, *Thermodynamics of fluid-phase equilibria for standard chemical engineering operations.* AIChE Journal, 2004. **50**(4): p. 739-761.

11. Ravindranath, D., B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, *QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

12. Neely, B.J., *Aqueous hydrocarbon systems: Experimental measurements and quantitative structure-property relationship modeling*, in *School of Chemical Engineering, Ph.D. Dissertation*. 2007, Oklahoma State University: Stillwater, Oklahoma.

13. Godavarthy, S.S., R.L. Robinson Jr., and K.A.M. Gasem, *SVRC-QSPR model for predicting saturated vapor pressures of pure fluids.* Fluid Phase Equilibria, 2006. **246**(1-2): p. 39-51.

14. Gebreyohannes, S., K. Yerramsetty, B.J. Neely, and K.A.M. Gasem, *Improved QSPR generalized interaction parameters for the nonrandom two-liquid activity coefficient model.* Fluid Phase Equilibria, 2013. **339**(0): p. 20-30.

15. Flory, P.J., *Thermodynamics of high polymer solutions.* The Journal of Chemical Physics, 1942. **10**(1): p. 51-61.

16. Scott, R.L., *Corresponding states treatment of nonelectrolyte solutions.* The Journal of Chemical Physics, 1956. **25**(2): p. 193-205.

17. Arlt, W., M.E.A. Macedo, P. Rasmussen, and J.M. Sorensen, *Liquid-liquid equilibrium data collection*. Chemistry Data Series. Vol. V, Parts 1-4. 1979 - 1987: DECHEMA, Frankfurt, Germany.

18. *Dragon Professional 6.0.9*. 2011, Talete SRL.

19. Katritzky, A.R., V.L. Lobanov, and M. Karelson, *Codessa 2.7.8*. 2007.

20. Gmehling, J., U. Onken, and W. Arlt, *Vapor-liquid equilibrium data collection*. Chemistry Data Series. Vol. I, Parts 1-8. 1977 - 2001: DECHEMA, Frankfurt, Germany.

21. *NIST-TDE, NIST Standard Reference Database 103b ThermoData Engine*. 2012.

22. *DIPPR Project 801, Physical and Thermodynamic Properties of Pure Chemicals*. 2011.

23. Rackett, H.G., *Equation of state for saturated liquids.* Journal of Chemical and Engineering Data, 1970. **15**(4): p. 514-517.

24. Tassios, D., *The number of roots in the NRTL and LEMF equations and the effect on their performance.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(1): p. 182-186.

25. *ChemBioOffice 11.0*. 2008, CambridgeSoft.

26. Guha, R., M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen, *The blue obelisk-interoperability in chemical informatics.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 991-998.

27. *The Open Babel Package 2.3*. 2011, Last accessed on: http://openbabel.sourceforge.net/.

28. Halgren, T.A., *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 490-519.

29. Prechelt, L., *Automatic early stopping using cross validation: quantifying the criteria.* Neural Networks, 1998. **11**(4): p. 761-767.

30. Yerramsetty, K.M., B.J. Neely, and K.A.M. Gasem, *A non-linear structure–property model for octanol–water partition coefficient.* Fluid Phase Equilibria, 2012. **332**(0): p. 85-93.

31. Bagheri, M., K. Yerramsetty, K.A.M. Gasem, and B.J. Neely, *Molecular modeling of the standard state heat of formation.* Energy Conversion and Management, 2013. **65**(0): p. 587-596.

32. Caruana, R., S. Lawrence, and L. Giles, *Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping*. 2000, Advances in Neural Information Processing Systems 13, MIT Press: Cambridge, MA. p. 402-408.

CHAPTER IV

GENERALIZED NRTL INTERACTION MODEL PARAMETERS FOR PREDICTING LLE BEHAVIOR

4.1. Introduction

Knowledge of phase behavior properties of chemicals is essential for designing and optimizing processes that involve separation of components from a mixture. Thermodynamic models are used in phase equilibria calculations to predict properties, such as phase compositions and partition coefficients at specific temperatures and pressures. Accuracy of thermodynamic models used to predict equilibrium phase behavior is dependent on the availability of experimental data. Reliable generalized predictions reduce the experimental burden in phase behavior modeling.

Phase equilibria properties are typically determined using equation-of-state (EOS) and activity coefficient (γ) models. A number of activity coefficient models for predicting vapor-liquid equilibria (VLE) and liquid-liquid equilibria (LLE) have been proposed by various researchers [1-5]. The nonrandom two-liquid (NRTL) [1] model is an activity coefficient model that is widely used in phase equilibria calculations. The NRTL model requires three interaction parameters that are determined through regression of experimental data for a specific system.

Many of the activity coefficient models in the literature can only be used to correlate existing data, and as such, they cannot be applied for *a priori* prediction of VLE and LLE behaviors. Traditionally, group-contribution methods (GCM) are used to generalize the interaction parameters of activity coefficient models. Examples of GCM models include UNIQUAC functional-group activity coefficients (UNIFAC) and analytical-solution-of-groups (ASOG) [2, 6]. Despite their potential benefits, group-contribution models suffer from limitations such as the inability to define effectively the functional groups of some chemical species and a lack of model interaction parameters for functional groups that are not represented in the UNIFAC data matrix. Thus, a need exists for an alternative approach to develop generalized models that are capable of *a priori* prediction of VLE and LLE properties.

In this work, we generalize the interaction parameters of the NRTL model for LLE systems using a theory-framed quantitative structure-property relationship (QSPR) modeling approach. In this approach, the NRTL model is used as a theoretical framework to develop the behavior model, and QSPR to generalize the substance-specific parameters of the model.

The QSPR modeling approach has been employed to generalize various theoretical frameworks for property predictions of pure components and mixtures [7-10]. In our previous work [10], we applied this theory-framed QSPR modeling approach to generalize the NRTL model parameters for VLE binary systems. The model provided property prediction errors that were approximately two times the error of the data regression errors [10]. Further, the developed QSPR model provided wider range of applicability and lower prediction errors compared to the UNIFAC model.

In this study, we extended the modeling methodology to predict the properties of LLE binary systems. For this purpose, a representative LLE database was assembled from literature sources. The data were used to develop a QSPR model for the estimation of the interaction parameters of the NRTL model. Further, the predictions from the QSPR model were evaluated using an external test set and also compared with available activity coefficient models from the literature.

4.2. NRTL activity coefficient model

In 1968 Renon and Prausnitz [1] developed the NRTL activity coefficient model based on the local composition theory of Wilson [11] and the two-liquid solution theory of Scott [12]. The model provides precise representation of highly non-ideal VLE and LLE systems [13]. The NRTL activity coefficient (γ) expression for a binary system is shown in Equation 4.1.

$$\ln \gamma_i = x_j^2 \left[\tau_{ji} \left(\frac{G_{ji}}{x_i + x_j G_{ji}} \right)^2 + \frac{\tau_{ij} G_{ij}}{\left(x_j + x_i G_{ij} \right)^2} \right] \tag{4.1}$$

with τ_{ij} and G_{ij} defined as follows:

$$G_{ij} = \exp(-\alpha_{ij}\tau_{ij}) \qquad \tau_{ij} = \frac{g_{ij} - g_{jj}}{RT} = \frac{a_{ij}}{RT} \tag{4.2}$$

where g_{ij} is the energy interaction between i and j molecules, α is the non-randomness factor in the mixture, R is the universal gas constant in $cal\,K^{-1}\,mol^{-1}$ and T is the mixture temperature in K.

The NRTL model has three adjustable parameters that are unique for a system. These parameters are a_{12} or ($g_{12} - g_{22}$), a_{21} or ($g_{21} - g_{11}$), and α_{12}. The parameters account simultaneously for pure-component liquid interactions (g_{11} and g_{22}) and mixed-liquid interactions (g_{12} and g_{21}). The non-randomness factor (α_{12}) varies from 0.2 to 0.47 [13] and is usually set at a constant value of 0.2 for LLE systems [14]. In this study, we used a value of 0.2 for the non-randomness factor.

4.3. QSPR methodology

The QSPR methodology applied to generalize the interaction parameters of the NRTL model involves the following specific steps: (1) database assembly, (2) parameter regression analysis, (3) structure generation and optimization, (3) descriptor reduction, and (5) QSPR model development.

Figure 4.1 illustrates a schematic representation of the steps involved in the development of QSPR models. Initially, a LLE database of experimental binary system data is assembled, and the interaction parameters of the NRTL model are regressed to fit the LLE properties of the database systems. Then, 2-dimensional (2D) structures of components in each binary system are generated and optimized to find a 3-dimensional (3D) representation with the least conformation energy. The optimized structures are then used to generate molecular descriptors using software such as DRAGON [15] and CODESSA [16]. Next, the initial pool of descriptors is analyzed through a reduction process to identify the most significant descriptors for predicting the interaction parameters. Simultaneous with the descriptor reduction, these significant descriptors are used to develop a neural network model. Finally, model interpretation is performed to understand the relationship between the important descriptors and the property of interest. The main elements of the model development process are described in greater detail below.

4.3.1. Database development

In this study, a database of LLE binary systems was collected from the DECHEMA LLE database [14]. The assembled database (OSU-LLE database) consists of 342 low-temperature (10 – 40 °C) binary LLE systems. These low-temperature systems are comprised of different combinations of 257 compounds. Approximately, 1200 low-temperature data points have been assembled.

The compounds present in the OSU-LLE database were classified in a similar manner as the UNIFAC functional-group classification approach [2]. Our LLE database is composed of compounds belonging to 28 chemical classes.

Figure 4.2 shows the data distribution of the binary LLE systems in the OSU-LLE database based on chemical classes. The number of systems represented for each type of functional-group interaction is shown in the figure. The matrix shows systems containing water represent about 70% of the data, which is due to the abundance of LLE experimental data for water containing systems in the literature.

4.3.2. Interaction parameter regression

The interaction parameters of the NRTL model were regressed to correlate the experimental binary LLE data assembled in this study. The regression analyses were performed by applying the Gibbs equilibrium criteria of a closed system containing two coexisting liquid phases, while subject to mass balance constraints. The phase equilibria calculation was performed by equating the component fugacities across the two liquid phases, as shown in Equation 4.3.

$$\begin{aligned} \hat{f}_i^{L_1} &= \hat{f}_i^{L_2} \\ \gamma_i^{L_1} x_i^{L_1} &= \gamma_i^{L_2} x_i^{L_2} \end{aligned} \qquad \text{i=1, 2} \tag{4.3}$$

where for any component i, $\hat{f}$ is the component fugacity in the liquid phases, γ is the component activity coefficient in the liquid phase, x is the liquid mole fraction and the superscripts L_1 and L_2 are liquid phases 1 and 2 in the liquid mixture, respectively.

The objective function, *OF*, used in the parameter regression analyses, was the sum of squares of the relative errors in liquid mole fractions in the two phases, as shown in Equation 4.4.

$$OF = \sum_{i=1}^{n} \left(\frac{x_{1Exp}^{L_1} - x_{1Calc}^{L_1}}{x_{1Exp}^{L_1}} \right)_i^2 + \sum_{i=1}^{n} \left(\frac{x_{1Exp}^{L_2} - x_{1Calc}^{L_2}}{x_{1Exp}^{L_2}} \right)_i^2 \tag{4.4}$$

where *n* is the number of data points and the superscripts *Exp* and *Calc* refer to experimental and calculated values, respectively.

In addition to liquid mole fractions, the quality of predictions are assessed for equilibrium properties such as equilibrium K-values of each binary system. The K-value for component *i* is the ratio of the liquid mole fraction in the two phases, which is shown in Equation 4.5.

$$K_i = \frac{x_i^{L_1}}{x_i^{L_2}} \qquad (4.5)$$

4.3.3. Descriptor calculation

Descriptor calculation was performed using various computational chemistry software. First, ChemBioDraw Ultra 11.0 [17] software was used to generate 2D and 3D structures of the molecules. Open Babel software [18] was then used to optimize the 3D structures by minimizing conformation energy of the molecules. Genetic algorithm (GA) based conformer search [18, 19], which employs the MMFF94 force field [20], was used in the structure optimization. The optimized molecules are then used to generate 2461 DRAGON [21] and 604 CODESSA [16] 0D, 1D, 2D, and 3D descriptors. Examples of these descriptors and their associated class are listed below.

Constitutional Descriptors: These descriptors, which include MW, number of atoms, etc., reflect the chemical composition of a compound without any information about its molecular geometry and atomic connectivity.

Topological: Topological descriptors are determined using graphical representation of the molecule. Mean-square distance index, polarity number, eccentric connectivity index, etc. are included in this category.

Geometrical: These descriptors, which include gravitational indices, radius of gyration, sphericity, asphericity, etc., are computed based on size indices.

Charge Descriptors: These descriptors are used to describe electronic nature of the molecule and are defined in terms of atomic charges. Some of descriptors in this category include for example maximum positive charge, total positive charge and total absolute charge.

Quantum Chemical: These descriptors are computed from molecular wave functions, characteristics of molecular orbitals and solvation energies. Some of descriptors in this category include ionization potentials, electron affinities and the HOMO/LUMO energy gap.

Molecular Properties: These descriptors describe physico-chemical and biological properties obtained from literature models. These descriptors include the octanol-water partition coefficient, hydrophilic factor, partition coefficient, dipole moment and similar characteristics.

4.3.4. Descriptor input

The calculated descriptors are used in the development of the QSPR model. For each binary system, the input descriptor set is prepared by calculating the differences of the individual descriptors of the compounds in each binary system. This novel approach ensures that the QSPR model results satisfy the pure limit behavior of activity coefficients. For a hypothetical mixture of X and Y where X and Y are the same molecule, the value of activity coefficients are ones; thus, the value of the interaction parameters are zeros, which requires that the QSPR input values (descriptor differences) to be zeros. Using this approach, the descriptor set up forces the model to obey the pure component behavior limits (γ=1, a_{12}=a_{21}=0) in the final QSPR prediction.

4.3.5. Descriptor reduction and model development

In this step, the large number of descriptor inputs is reduced to find the most significant descriptors for accurate property predictions. The model development employed in this study is a hybrid strategy where descriptor reduction and model development happen simultaneously. The hybrid algorithm uses evolutionary programming (EP) and differential evolution (DE) as a wrapper around artificial neural networks (ANNs) to search for the best descriptor subsets from a large number of molecular descriptors. A detailed discussion on our descriptor reduction and model development methodology can be found in our previous works [10, 22, 23].

In the model development process, the entire data set was divided into four sub-sets (training, validation, internal test and external test). The proportion of data for the different data sets was: 50% for the training set, 15% for the internal validation set, 10% for the internal test set and the remaining 25% for the external test set. The data division was performed while insuring adequate representation of all the functional-group interactions within all the data sets. For example, there are 42 LLE systems with alcohol/water interactions in the database. The data division for this type of interactions will be 21, 6, 4 and 11 of the systems assigned to the training, validation, internal test and external test sets, respectively. For interactions with a small number of systems, data allocation priority was extended to the training followed by validation and internal test sets.

The descriptor reduction and model development process was performed using all data excluding the external test set. The validation data set is used to avoid over-fitting by applying an early-stopping method [21, 24]. The internal test set data was used to select the best ANNs during the descriptor reduction algorithm. In model development, the external test set data was set aside and was used only to assess the generalization (*a priori* prediction) capability of the developed model.

4.3.6. Modeling scenarios

Three case studies were performed to assess the representation and prediction of three models for LLE property behavior. The three case studies are outlined as follows:

NRTL-Regressed-LLE: The NRTL model with regressed a_{12} and a_{21} parameters was used to represent LLE properties.

NRTL-QSPR-LLE: The generalized QSPR model was used to provide the NRTL model parameters, and then the NRTL model was used to predict the activity coefficients.

UNIFAC-1981-LLE: The UNIFAC model for LLE systems was used to predict the activity coefficients of each component for LLE systems. The UNIFAC interaction parameters reported by Gmehling *et al.* [25] were used in this case study.

The NRTL-Regressed-LLE study was conducted to evaluate the representation capability of the NRTL model. The NRTL-QSPR-LLE and UNIFAC-1981-LLE case studies were focused on assessing the *a priori* predictive capabilities of the QSPR generalized NRTL model and the UNIFAC model, respectively.

The representation and prediction capabilities of the models were assessed for using the equilibrium properties, liquid mole fraction (x_1 and x_2) and equilibrium K-values (K_1 and K_2), of the LLE binaries. In the NRTL-Regressed-LLE study, the two model parameters, a_{12} and a_{21}, shown in Equation 4.2, were regressed. GEOS software [26], developed by our research team for predictions of thermophysical properties using various models, was employed to correlate the VLE data using the NRTL model. Flash calculation was employed in the regression analyses. The regressed or QSPR predicted parameters are used directly to calculate x_1 in Phase 1, x_1 in Phase 2, K_1 and K_2 for known T and component mole fraction (z_1).

4.4. Results and discussion

Experimental T and x data of 342 binary LLE systems were used to evaluate the correlative capabilities of the NRTL model. The results were analyzed by calculating the root-mean-squared error (RMSE), bias and percentage absolute average deviation %AAD.

Table 4.1 provides the property representation errors for the NRTL-Regressed-LLE case study. As shown, the NRTL model has overall %AADs of 12.5, 2.0, 15.9 and 10.8 for x_1 in Phase 1, x_1 in Phase 2, K_1 and K_2, respectively. The NRTL model representation result for LLE systems is significantly higher than the representation results found for VLE systems in our previous study [10]. This is primarily due to the high temperature dependence of the NRTL interaction parameters, as opposed to the insignificant temperature effect associated with the interaction parameters of the VLE systems. This is similar issue that was faced in the DECHEMA LLE database analyses, where the interaction parameters for LLE systems in the DECHEMA LLE database were given on a point-by-point (temperature-by-temperature) basis [14]. In this study, we limited the temperature range between 10 and 40 °C to reduce the effect of temperature on the property predictions.

Figure 4.3 shows the distribution of the x_1 in Phase 1 regression errors for the NRTL model by functional-group interaction. The result of each functional-group interaction is shaded in variations of grey based on the %AAD ranges given in the figure key. As indicated in the matrix, the NRTL model provides representation of mole fractions within AADs of 15% for most of the functional-group interactions with the exception of the water systems. The model resulted in relatively high errors for the systems containing water which could be attributed to the large uncertainties in the experimental measurements and the inability of the model to represent such systems. Further, the mole fractions of aqueous systems tend to be very small, which results in higher *percentage* errors.

The NRTL-Regressed-LLE study established the benchmark for the best achievable level of prediction errors for QSPR generalization. The regressed model parameters (a_{12} and a_{21}) were used as targets in developing the QSPR model.

The list of 30 molecular descriptors that were used as inputs in developing the QSPR model are shown in Table 4.2. DR and CO represent molecular descriptors calculated using DRAGON [15] and CODESSA [16], respectively. The result indicates functional group, electrostatic, quantum chemical and GETAWAY descriptors are significant in predicting the interaction parameters. Some of descriptors identified in this study were similar to the descriptors found in our previous study for VLE interaction parameter predictions [10].

Figures 4.4 and 4.5 show comparisons of the regressed and predicted a_{12} and a_{21} values for the training and validation sets, respectively. The figures indicate good agreement between the regressed and QSPR predicted parameters. Consequently, the QSPR model resulted in comparable predictions for the training and validation sets which suggests that the model was trained without over-fitting. Similarly, Figure 4.6 shows comparison of the regressed and predicted a_{12} and a_{21} values for the external test set. The R^2 value between the regressed and predicted values for the external test set were 0.8 and 0.7 for a_{12} and a_{21}, respectively. Although the level of agreement here is lower than that for the training and validation sets, these results are still indicative of good generalized parameter predictions from the QSPR model.

Table 4.3 provides the LLE property prediction errors obtained using the QSPR predicted parameters from the NRTL-QSPR-LLE study. The results are classified into training, validation, internal test and external test sets. The LLE property predictions for the QSPR model were about three to four times the regression analyses %AAD values. The table also indicates the model resulted in comparable errors in all data sets. Further, a number of systems failed to converge to an appropriate two-phase equilibrium solution. The parameters generated by our newly developed model led to converged two-phase solutions for 305 out of the 342 systems. Convergence failure is due to the fact that unlike VLE systems the LLE interaction parameters are highly temperature dependent and very sensitive to small temperature variations.

Table 4.4 shows the representation and predictions of the NRTL-Regressed-VLE and NRTL-QSPR-VLE case studies for VLE systems from our previous study [10]. The result shows the NRTL model was able to provide precise representation for VLE systems. In addition, the QSPR model was able to generalize successfully the NRTL interaction parameters for VLE systems within twice the regression errors. This result reveals that although the NRTL model is able to handle VLE system properties well, the model lacks robustness when applied to LLE property predictions. Thus, further study is required to improve the NRTL model capability in capturing the temperature dependence of LLE interaction parameters. In this regard, better accounting of temperature dependence for LLE systems may be attained by incorporating equation-of-state interaction concepts within the NRTL model. Further, we need to investigate the capability of the UNIQUAC model for LLE systems. For better accounting of the temperature dependence, the residual part of the UNIQUAC model could be modified by learning from the theoretical formulation of equation-of-state models.

Figure 4.7 shows the distribution of x_1 in Phase 1 QSPR prediction errors for the NRTL model by functional-group interactions. The figure shows the QSPR model resulted in prediction of pressure within 30% for about half of the functional-group interactions present in the database. The matrix also shows that the model provided %AADs between 30 and 50 for most of the water systems.

Table 4.5 shows LLE property prediction comparisons of the NRTL-QSPR-LLE and UNIFAC-1981-LLE case studies. Overall, the QSPR model yielded predictions with 38, 8, 51 and 44 %AAD for x_1 in Phase 1, x_1 in Phase 2, K_1 and K_2, respectively. The predictions are within 3 to 4 times the regression errors. The table also shows that the QSPR model has an approximate 11% failure rate. The UNIFAC model [25] resulted in about 3 to 7 times the regression errors. A failure rate of 35% was observed for UNIFAC-1981-LLE predictions. In comparison to the QSPR model, the UNIFAC-1981-LLE model resulted in larger prediction errors as well as the three times larger failure rate. This demonstrates the efficacy of our modelling approach in providing improved and reliable predictions, as well as an increased range of applicability when compared to the prevalent UNIFAC model.

4.5. Conclusion

In this study, the interaction parameters of the NRTL model were generalized using a QSPR modeling approach for LLE systems. A database consisting of 342 low-temperature binary LLE systems from combinations of 257 compounds was assembled. The structural descriptors of the molecules were calculated and used as inputs in the generalized QSPR model. The newly developed QSPR generalized model provided LLE property predictions within 3 to 4 times the overall errors found in the experimental data regression analyses. In comparison to the UNIFAC-1981-LLE model, the QSPR model provided lower errors as well as a wider range of applicability for LLE property predictions based on the failure to convergence rate. Our findings indicate that the QSPR modeling approach is effective in generalizing the interaction parameters of the NRTL activity coefficient model.

The study revealed that the NRTL model lacks a robustness to handle the temperature dependence of the interaction parameters for LLE systems. Therefore, further studies need to be focused on modifying the NRTL model to better capture the temperature dependence of the parameters. A potential area that may lead to better accounting of temperature dependence is incorporating equation-of-state interaction concepts within the NRTL model. Further, we need to investigate the capability of the UNIQUAC model for LLE systems by modifying the residual part of the model by learning from the theoretical formulation of equation-of-state models.

Table 4.1. LLE experimental data representation of the NRTL model using regressed parameters

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
NRTL-Regressed-LLE	a_{12} & a_{21}	x_1 in Phase 1	342	1183	0.03	0.00	12.5	30
		x_1 in Phase 2	237	840	0.03	0.00	2.0	24
		K_1	237	840	47.30	10.85	15.9	66
		K_2	237	840	0.12	0.01	10.8	100

Table 4.2. The descriptors used as inputs for the ANNs in the final ensemble for estimating the NRTL model parameter

No	Descriptor name	Descriptor description	Source	Type of descriptor
1	ARR	aromatic ratio	DR	Ring descriptors
2	CATS2D_09_DL	CATS2D Donor-Lipophilic at lag 09	DR	CATS 2D
3	H2m	H autocorrelation of lag 2 / weighted by mass	DR	GETAWAY descriptors
4	SAtot	total surface area from P_VSA-like descriptors	DR	Molecular properties
5	ICR	radial centric information index	DR	Topological indices
6	QYYi	quadrupole y-component value / weighted by ionization potential	DR	Geometrical descriptors
7	SpMAD_D	spectral mean absolute deviation from topological distance matrix	DR	2D matrix-based descriptors
8	SpMax7_Bh(i)	largest eigenvalue n. 7 of Burden matrix weighted by ionization potential	DR	Burden eigenvalues
9	Mor17v	signal 17 / weighted by van der Waals volume	DR	3D-MoRSE descriptors
10	SP03	shape profile no. 3	DR	Randic molecular profiles
11	B03[C-O]	Presence/absence of C - O at topological distance 3	DR	2D Atom Pairs
12	MATS7e	Moran autocorrelation of lag 7 weighted by Sanderson electronegativity	DR	2D autocorrelations
13	RDF040i	Radial Distribution Function - 040 / weighted by ionization potential	DR	RDF descriptors
14	RDF100u	Radial Distribution Function - 100 / unweighted	DR	RDF descriptors
15	RDF125p	Radial Distribution Function - 125 / weighted by polarizability	DR	RDF descriptors
16	SM2_L	spectral moment of order 2 from Laplace matrix	DR	2D matrix-based descriptors
17	Mor08e	signal 08 / weighted by Sanderson electronegativity	DR	3D-MoRSE descriptors
18	Mor05u	signal 05 / unweighted	DR	3D-MoRSE descriptors
19	Avg 1-electron react. index for a O atom	Avg 1-electron react. index for a O atom	CO	Quantum Chemical
20	RTe+	R maximal index / weighted by Sanderson electronegativity	DR	GETAWAY descriptors
21	Max e-e repulsion for a F atom	Max e-e repulsion for a F atom	CO	Quantum Chemical
22	Image of the Onsager-Kirkwood solvation energy	Image of the Onsager-Kirkwood solvation energy	CO	Quantum Chemical
23	R2e	R autocorrelation of lag 2 / weighted by Sanderson electronegativity	DR	GETAWAY descriptors
24	F01[N-O]	Frequency of N - O at topological distance 1	DR	2D Atom Pairs
25	FPSA-1 Fractional PPSA (PPSA-1/TMSA) [Zefirov's PC]	FPSA-1 Fractional PPSA (PPSA-1/TMSA) [Zefirov's PC]	CO	Electrostatic
26	nHAcc	number of acceptor atoms for H-bonds (N,O,F)	DR	Functional group counts
27	Min partial charge for a O atom [Zefirov's PC]	Min partial charge for a O atom [Zefirov's PC]	CO	Electrostatic
28	RTu+	R maximal index / unweighted	DR	GETAWAY descriptors
29	Min e-n attraction for a C-N bond	Min e-n attraction for a C-N bond	CO	Quantum Chemical
30	RDF050e	Radial Distribution Function - 050 / weighted by Sanderson electronegativity	DR	RDF descriptors

Table 4.3. Predictions from the NRTL-QSPR-LLE model

Data Set	Parameters	Property	No. of converged systems	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD	%AAD multiplier
Training set	a_{12} & a_{21}	x_1 in Phase 1	167 out of 182	167	602	0.08	-0.01	33.1	85	3
		x_1 in Phase 2		140	515	0.08	0.01	7.6	68	4
		K_1		140	515	56.49	10.37	44.9	100	3
		K_2		140	515	0.19	0.00	42.7	100	4
Validation set	a_{12} & a_{21}	x_1 in Phase 1	44 out of 49	44	157	0.07	-0.02	43.2	91	3
		x_1 in Phase 2		19	64	0.11	0.06	10.2	50	5
		K_1		19	64	62.58	41.56	62.2	100	4
		K_2		19	64	0.17	-0.08	34.3	92	4
Internal test set	a_{12} & a_{21}	x_1 in Phase 1	32 out of 36	32	98	0.04	-0.01	40.9	95	3
		x_1 in Phase 2		8	27	0.06	0.03	7.2	18	8
		K_1		8	27	17.09	7.09	38.2	100	3
		K_2		8	27	0.07	-0.04	31.2	80	4
External test set	a_{12} & a_{21}	x_1 in Phase 1	62 out of 75	62	214	0.12	-0.04	47.7	96	4
		x_1 in Phase 2		42	145	0.10	0.04	9.8	47	4
		K_1		42	145	65.91	44.05	69.4	100	4
		K_2		42	145	0.13	-0.06	52.8	100	5
All data	a_{12} & a_{21}	x_1 in Phase 1	305 out of 342	305	1071	0.08	-0.02	38.3	96	3
		x_1 in Phase 2		209	751	0.08	0.02	8.3	68	4
		K_1		209	751	58.12	19.85	51.2	100	3
		K_2		209	751	0.17	-0.02	43.5	100	4

Table 4.4. Representation and predictions of the NRTL-Regressed and the NRTL-QSPR case studies for VLE systems from our previous study [10]

Study	Model (Vapor/Liquid)	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD
NRTL-Regressed-VLE	IG/NRTL	P (bar)	578	16563	0.20	0.00	2.6
		T (K)		16726	2.20	0.30	0.2
		K-values		9937	2.00	−0.15	4.9
NRTL-QSPR-VLE	IG/Generalized NRTL	P (bar)	578	16696	0.28	0.01	6.2
		T (K)		16727	3.79	0.20	0.6
		K-values		9953	1.62	−0.15	8.8

Table 4.5. Comparison of *a priori* predictions of the NRTL-QSPR and the UNIFAC-1981-LLE case studies

Model	Parameters	Property	No. of converged systems	No. of pts.	RMSE	Bias	%AAD	Failure factor
NRTL-QSPR-LLE	Generalized a_{12} & a_{21}	x_1 in Phase 1	305 out of 342	1071	0.08	-0.02	38.3	11%
		x_1 in Phase 2		751	0.08	0.02	8.3	
		K_1		751	58.12	19.85	51.2	
		K_2		751	0.17	-0.02	43.5	
UNIFAC-1981-LLE using ASPEN PLUS	UNIFAC-1981-LLE	x_1 in Phase 1	152 out of 237*	578	0.03	-0.01	71.8	36%
		x_1 in Phase 2		578	0.05	0.00	7.5	
		K_1		578	35.06	4.08	47.2	
		K_2		578	0.04	0.02	76.8	

*only systems with complete T-x-x (temperature and x_1 in Phase 1 and 2) experimental data are considered

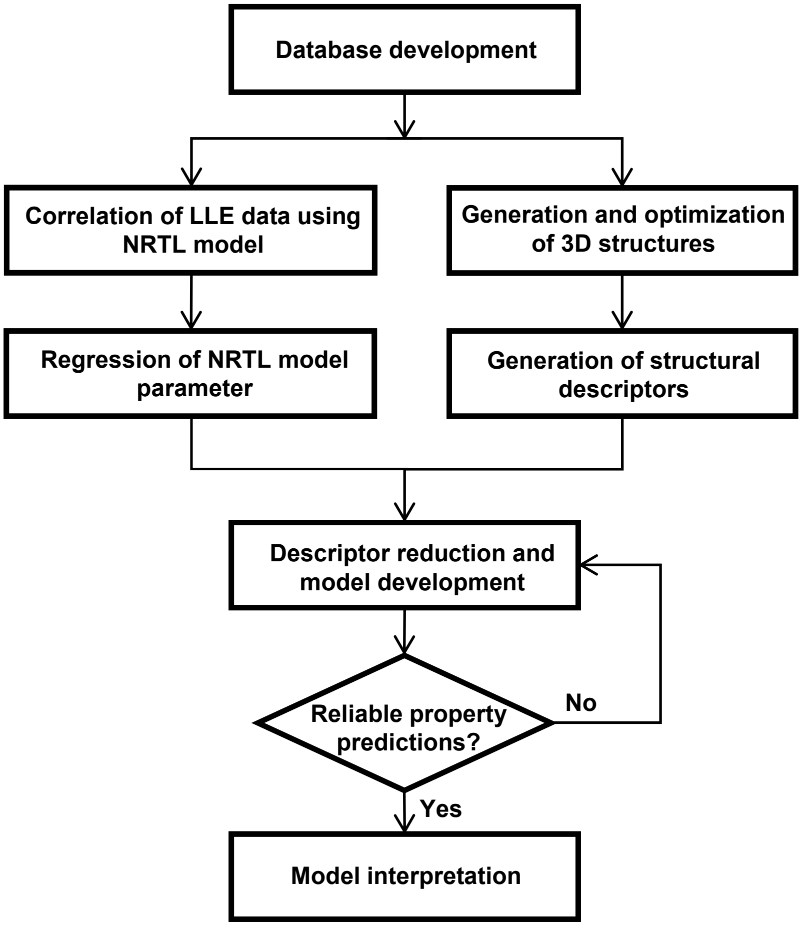

Database development
Correlation of LLE data using NRTL model
Generation and optimization of 3D structures
Regression of NRTL model parameter
Generation of structural descriptors
Descriptor reduction and model development
Reliable property predictions?
No
Yes
Model interpretation

Figure 4.1. Schematic of the QSPR model development process

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Alcohol	4																											
2	Aldehyde																												
3	Alkane	17																											
4	Alkene																												
5	Alkyne	2																											
6	Amine			12	1																								
7	Aromatic Bromo																												
8	Aromatic Floro																												
9	Benzene Derivative	1					2		1																				
10	Bromoalkane																												
11	Carboxylic Acid			10																									
12	Chloroalkane																												
13	Chlorobenzene																												
14	Diol	1		1						5		1																	
15	Ester														1														
16	Ether																												
17	Floroalkane			4																									
18	Furfural Derivative	2		14																									
19	Ketone	1		2																									
20	Nitrile	2		8													1												
21	Nitro Compound	5		7	1										1														
22	Phenol Derivative			6																									
23	Pyridine Derivative			2																									
24	Sulfide	1		1								1										1							
25	Sulfone	1		2																									
26	Thiophene						2								1														
27	Toluene Derivative	1		1											1			1											
28	Water	42	4	33	2	3	4	1		14	6	27	13	1		1	23		4	19	1	4	3	5				3	

X

Y # Number of available binary LLE systems consisting of chemicals with functional groups of X and Y

□ No LLE data used

Figure 4.2. Database matrix of the compounds in the OSU-LLE database

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Alcohol	2.9																											
2	Aldehyde																												
3	Alkane	11.9																											
4	Alkene																												
5	Alkyne	7.7																											
6	Amide			10.2	13																								
7	Amine																												
8	Aromatic Bromo																												
9	Aromatic Floro	16.2					4.1		11.2																				
10	Benzene Derivative																												
11	Bromoalkane			10.4																									
12	Carboxylate																												
13	Chloroalkane																												
14	Chloroalkene	1.1		4.5						5.7		2.3																	
15	Chlorobenzene														2.5														
16	Epoxide																												
17	Ester			13.1																									
18	Ether	11.2		7.7																									
19	Furfural	0.3		16.4																									
20	H_2S	7.7		7.3													3.6												
21	Iodoalkane	13.5		15.6	6.7										9.7														
22	Ketone			12.1																									
23	Nitrile			8.6																									
24	Nitrite	20.5		5								10.7										5							
25	Nitro Compound	8.7		8.7																									
26	Pyridine Derivative						2.5								1.3														
27	Sulfide	19.2		16.5											5.1			8.1											
28	Thiol	13.8	14.2	19.7	11.6	12	17.8	3.6		11.4	12.5	9.7	13.5	8		21.5	17.7		12.5	13.8	16.7	7.1	4.5	16.8				13.8	

Color	%AAD Range
#	%AAD<10
#	10<%AAD<20
#	20<%AAD<50
#	50<%AAD<100

Figure 4.3. Representation of x_1 in Phase 1 using the regressed NRTL model by type of interaction

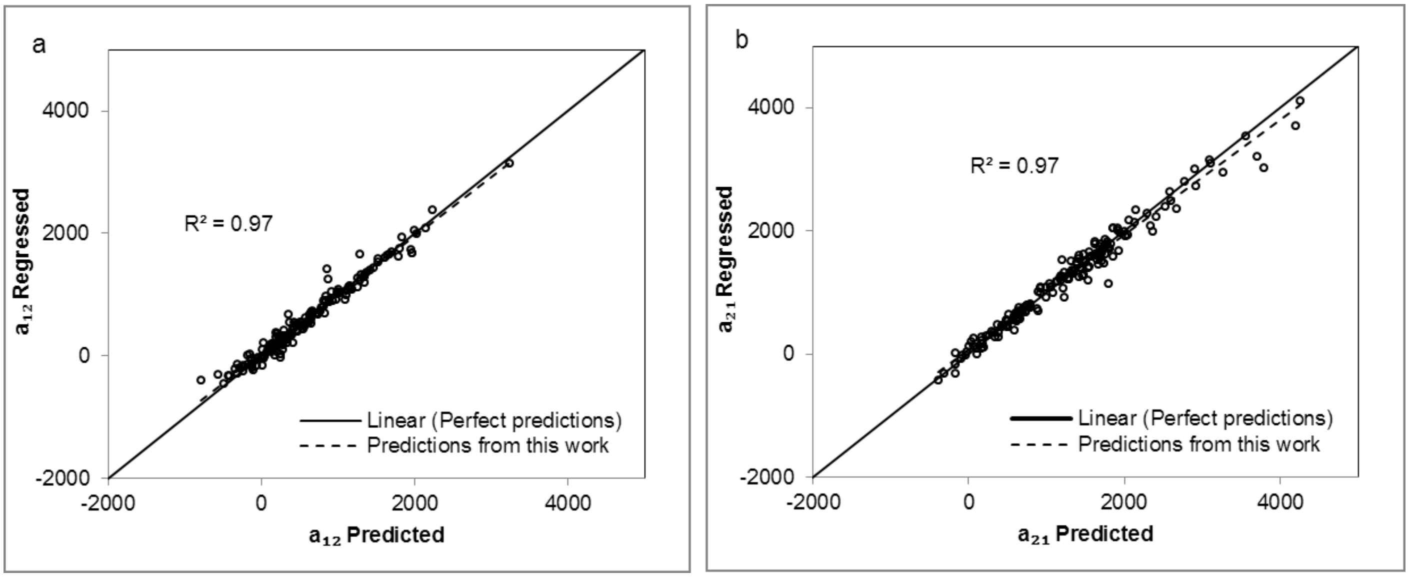


Figure 4.4. Comparison of the regressed and QSPR predicted (a) a_{12} and (b) a_{21} values in the training set

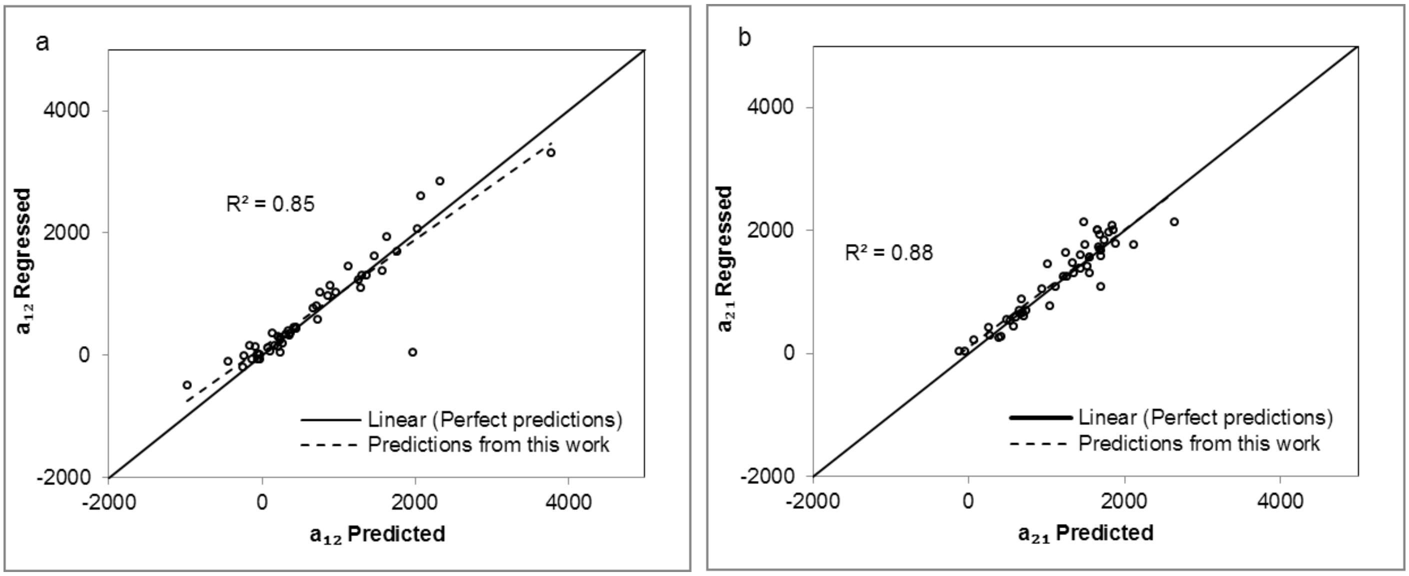


Figure 4.5. Comparison of the regressed and QSPR predicted (a) a_{12} and (b) a_{21} values in the validation set

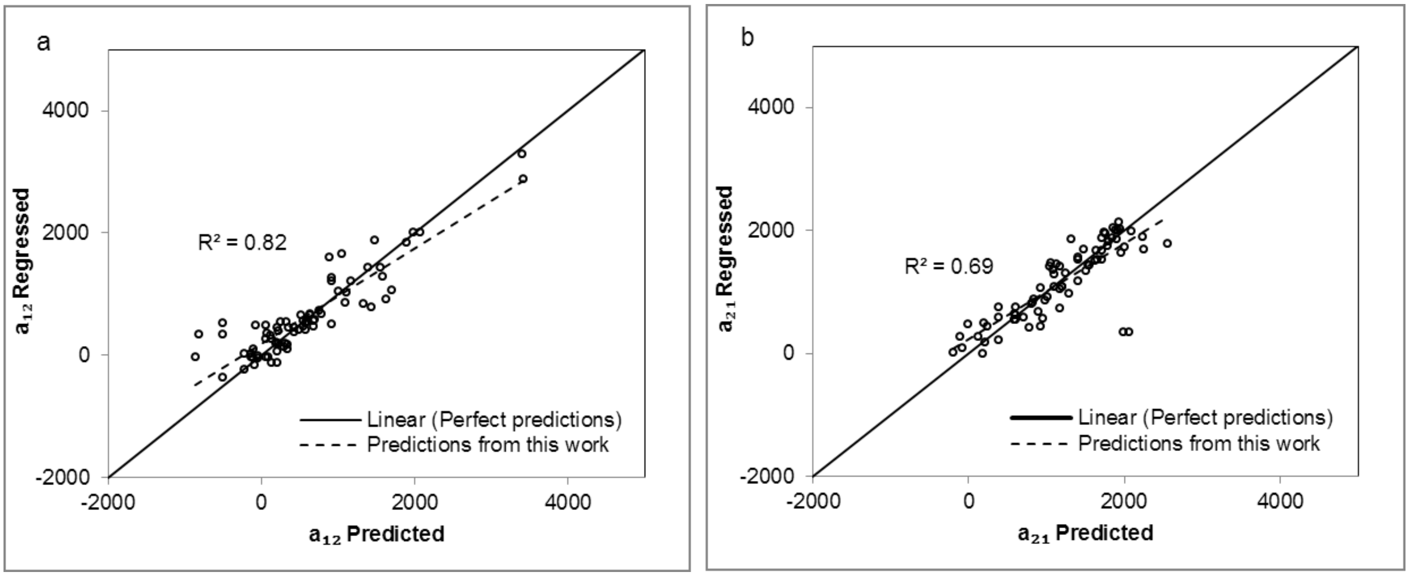


Figure 4.6. Comparison of the regressed and QSPR predicted (a) a_{12} and (b) a_{21} values in the external test set

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Alcohol	31																											
2	Aldehyde																												
3	Alkane	53																											
4	Alkene																												
5	Alkyne	65																											
6	Amide			26	73																								
7	Amine																												
8	Aromatic Bromo																												
9	Aromatic Floro	16					19		14																				
10	Benzene Derivative																												
11	Bromoalkane			46																									
12	Carboxylate																												
13	Chloroalkane																												
14	Chloroalkene	4								48		15																	
15	Chlorobenzene														15														
16	Epoxide																												
17	Ester			40																									
18	Ether	11		22																									
19	Furfural	11		42																									
20	H_2S	61		27													48												
21	Iodoalkane	51		53	22										49														
22	Ketone			31																									
23	Nitrile			31																									
24	Nitrite	20		16								12										29							
25	Nitro Compound	50		71																									
26	Pyridine Derivative						20								41														
27	Sulfide	22		46														15											
28	Thiol	38	31	38	51	47	61	15		34	49	42	40			44	44		36	48	83	34	11	29				35	

Color	%AAD Range
#	%AAD<10
#	10<%AAD<20
#	20<%AAD<50
#	50<%AAD<100

Figure 4.7. Prediction of x_1 in Phase 1 using the regressed NRTL-QSPR-LLE model by type of interaction

REFERENCES

1. Renon, H. and J.M. Prausnitz, *Local compositions in thermodynamic excess functions for liquid mixtures.* AIChE Journal, 1968. **14**(1): p. 135-144.
2. Gmehling, J., J. Li, and M. Schiller, *A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties.* Industrial & Engineering Chemistry Research, 1993. **32**(1): p. 178-193.
3. Abrams, D.S. and J.M. Prausnitz, *Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems.* AIChE Journal, 1975. **21**(1): p. 116-128.
4. Skjold-Jorgensen, S., B. Kolbe, J. Gmehling, and P. Rasmussen, *Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(4): p. 714-722.
5. Fischer, K. and J. Gmehling, *Further development, status and results of the PSRK method for the prediction of vapor-liquid equilibria and gas solubilities.* Fluid Phase Equilibria, 1996. **121**(1-2): p. 185-206.
6. Gmehling, J., D. Tiegs, and U. Knipp, *A comparison of the predictive capability of different group contribution methods.* Fluid Phase Equilibria, 1990. **54**: p. 147-165.
7. Ravindranath, D., B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, *QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

8. Neely, B.J., *Aqueous hydrocarbon systems: Experimental measurements and quantitative structure-property relationship modeling*, in *School of Chemical Engineering, Ph.D. Dissertation*. 2007, Oklahoma State University: Stillwater, Oklahoma.

9. Godavarthy, S.S., R.L. Robinson Jr., and K.A.M. Gasem, *SVRC-QSPR model for predicting saturated vapor pressures of pure fluids.* Fluid Phase Equilibria, 2006. **246**(1-2): p. 39-51.

10. Gebreyohannes, S., K. Yerramsetty, B.J. Neely, and K.A.M. Gasem, *Improved QSPR generalized interaction parameters for the nonrandom two-liquid activity coefficient model.* Fluid Phase Equilibria, 2013. **339**(0): p. 20-30.

11. Wilson, G.M., *Vapor-liquid equilibrium. XI. A new expression for the excess free energy of mixing.* Journal of the American Chemical Society, 1964. **86**(2): p. 127-130.

12. Scott, R.L., *Corresponding states treatment of nonelectrolyte solutions.* The Journal of Chemical Physics, 1956. **25**(2): p. 193-205.

13. Prausnitz, J.M., R.N. Lichtenthaler, and E.G.d. Azevedo, *Molecular thermodynamics of fluid-phase equilibria*. 3rd ed. 1998: Prentice-Hall.

14. Arlt, W., M.E.A. Macedo, P. Rasmussen, and J.M. Sorensen, *Liquid-liquid equilibrium data collection*. Chemistry Data Series. Vol. V, Parts 1-4. 1979 - 1987: DECHEMA, Frankfurt, Germany.

15. *Dragon Professional 6.0.9*. 2011, Talete SRL.

16. Katritzky, A.R., V.L. Lobanov, and M. Karelson, *Codessa 2.7.8*. 2007.

17. *ChemBioOffice 11.0*. 2008, CambridgeSoft.

18. *The Open Babel Package 2.3*. 2011, Last accessed on: http://openbabel.sourceforge.net/.

19. Guha, R., M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen, *The Blue ObeliskInteroperability in Chemical Informatics.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 991-998.

20. Halgren, T.A., *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 490-519.

21. Prechelt, L., *Automatic early stopping using cross validation: quantifying the criteria.* Neural Networks, 1998. **11**(4): p. 761-767.

22. Yerramsetty, K.M., B.J. Neely, and K.A.M. Gasem, *A non-linear structure–property model for octanol–water partition coefficient.* Fluid Phase Equilibria, 2012. **332**(0): p. 85-93.

23. Bagheri, M., K. Yerramsetty, K.A.M. Gasem, and B.J. Neely, *Molecular modeling of the standard state heat of formation.* Energy Conversion and Management, 2013. **65**(0): p. 587-596.

24. Caruana, R., S. Lawrence, and L. Giles, *Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping*. 2000, Advances in Neural Information Processing Systems 13, MIT Press: Cambridge, MA. p. 402-408.

25. Magnussen, T., P. Rasmussen, and A. Fredenslund, *UNIFAC parameter table for prediction of liquid-liquid equilibriums.* Industrial & Engineering Chemistry Process Design and Development, 1981. **20**(2): p. 331-339.

26. Khaled, G., *GEOS (Generalized EOS Predictions)*. 2013.

CHAPTER V

ONE-PARAMETER MODIFIED NRTL ACTIVITY COEFFICIENT MODEL

5.1. Introduction

Phase equilibrium properties, such as pressure, temperature, compositions and partition coefficients are required for the design of chemical separation operations. Generalized thermodynamic models are used widely to describe phase equilibria properties of systems; thus avoiding the need to conduct expensive and time intensive experimental property measurements.

In phase equilibria calculations, activity coefficients (γ) are used to account for component non-ideal liquid behavior in a mixture. A number of activity coefficient models for representing vapor-liquid equilibrium (VLE) and liquid-liquid equilibrium (LLE) systems have been proposed by various researchers [1-5]. These models demonstrate the composition and temperature dependence of activity coefficients. In general, the literature models can be classified as historical semi-empirical activity coefficient models (Margules [6], Redlich-Kister [6] and Van Laar [6]), theory-based models, which include local composition and two-liquid models (Wilson [7], NRTL [1] and UNIQUAC [3]) and group-contribution models (UNIFAC [2], ASOG [8]).

The NRTL model is among the most widely used activity coefficient models in phase equilibria. The model requires three adjustable parameters, which include two energy interaction parameters (a_{12} and a_{21}) and a non-randomness factor (α_{12}). The model provides good representation of experimental equilibrium data for strongly non-ideal mixtures and partially immiscible systems [6].

One of the main disadvantages of the NRTL model is the strong correlation between the two energy interaction parameters (a_{12} and a_{21}). A number of researchers [9-13] have attempted to modify the original NRTL model to eliminate or reduce the correlation. Many of the modified models presented in the literature, however, were not successful in decoupling the energy interaction parameters between like and unlike molecules. Further, they lacked simplicity and were not evaluated for wide range of interactions. Therefore, a need exists for a simple modification of the NRTL model that eliminates the parameter correlation.

In this work, we propose a new modification to the NRTL activity coefficient model addressing the limitation of the original model. The newly modified model recasts the model parameters in such a way that the two new model parameters reflect two different characteristics, namely energy interaction and energy interaction ratio parameters. This new modification enables easier generalization of one of the parameters (energy interaction ratio) in terms of pure-component properties, which essentially reduces the NRTL model to a one-parameter model for a VLE system. As such, our modification eliminates the parameter correlation present in the original model by reducing the number of model parameters. The single model parameter also provides a capability of relating/classifying VLE behaviors based on that parameter value. The ability to identify behaviors of systems with only the parameter value is useful for designing processes involving new systems.

The objectives of this work are (1) to mitigate the limitations of the original NRTL model, namely the parameter correlation and generalizability, (2) to evaluate the representation capability of the original and modified NRTL models for representing mixtures containing various functional groups, (3) to provide a qualitative approach to classifying systems in terms of their behaviors, (4) to assess the applicability of the original and modified NRTL models for multicomponent systems and (5) to evaluate the temperature dependence of model parameters.

To meet the objectives of this work, VLE and LLE databases were assembled from the DECHEMA-VLE [10], DECHEMA-LLE [14], NIST-TDE [15] and DIPPR [16] databases. The VLE systems were classified by chemical class and a qualitative approach was applied to classify the systems by behavior into nearly-ideal, non-ideal and highly non-ideal systems. Further, the data were used to validate the applicability of the models for multiphase and "cross-phase" property predictions, i.e., the applicability of LLE regressed parameters to predict VLE properties and vice versa.

5.2. Literature review on one-parameter NRTL activity coefficient models

The nonrandom two-liquid (NRTL) activity coefficient model developed by Renon and Prausnitz in 1968 [6] is based on the local composition theory of Wilson [7] and the two-liquid solution theory of Scott [17]. The model provides precise representation of highly non-ideal VLE and LLE systems. The NRTL activity coefficient expression for components in a binary system is given as:

$$\ln \gamma_i = x_j^2 \left[\tau_{ji} \left(\frac{G_{ji}}{x_i + x_j G_{ji}} \right)^2 + \frac{\tau_{ij} G_{ij}}{\left(x_j + x_i G_{ij} \right)^2} \right] \tag{5.1}$$

where τ_{ij} and G_{ij} are defined as:

$$G_{ij} = \exp(-\alpha_{ij}\tau_{ij}) \qquad \tau_{ij} = \frac{g_{ij} - g_{jj}}{RT} = \frac{a_{ij}}{RT} \tag{5.2}$$

where α_{ij} is the non-randomness factor in the mixture, g_{ij} is energy interaction between i and j component molecules, a_{ij} is energy interaction difference of g_{ij} and g_{jj}, x_i is the mole fraction of component i, R is the universal gas constant and T is the mixture temperature.

The NRTL model contains three parameters (defining $\alpha_{ij} = \alpha_{ji}$) that are specific for each binary system. These adjustable parameters are a_{12} or ($g_{12} - g_{22}$), a_{21} or ($g_{21} - g_{11}$), and α_{12}. The two energy interaction parameters account simultaneously for pure-component liquid interactions (g_{11} and g_{22}) and mixed-liquid interactions (g_{12} and g_{21}). The non-randomness factor (α_{12}) varies from 0.2 to 0.47 [6] and can often be set *a priori*. To be consistent with the DECHEMA database [14], the non-randomness factor was kept constant as 0.2 for all binary systems in this work.

As mentioned previously, the parameters a_{12} and a_{21} in the basic NRTL model are strongly correlated. A number of researchers have tried to modify the NRTL model to eliminate or reduce the correlation between the parameters. Bruin and Prausnitz in 1971 [9] first attempted to reduce the number of parameters in the NRTL model. They presented a new derivation of the model by substituting the local mole fractions with volume fractions. Their newly derived equation is shown, as follows, in Equation 5.3:

$$\begin{aligned}
\ln\gamma_i &= q_{ij}\Phi_{ji}{}^2\tau_{ji} + q_{ji}\frac{x_2}{x_1}\Phi_{ij}(1-\Phi_{ij})\tau_{ij} \\
\Phi_{ij} &= \frac{v_i x_i \exp(-\alpha_{ij}\tau_{ij}/q_{ji})}{v_j x_j + v_i x_i \exp(-\alpha_{ij}\tau_{ij}/q_{ij})} \\
\tau_{ij} &= \frac{g_{ij}-g_{jj}}{RT}
\end{aligned} \qquad (5.3)$$

where Φ_{ij} is the local volume fraction of molecule *i* around molecule *j*; v_j is the molar volume of molecule *i*; q_{ij} is a measure of the number of sites a molecule of type *i* occupies in a pseudo lattice structure.

The newly introduced q_{12} and q_{21} parameters in the modified equation are determined using the following three conditional statements:

$$\begin{array}{lll} v_1 > v_2 & q_{12} = \left(\dfrac{v_1}{v_2}\right)^{1/2} & q_{21} = 1 \\ v_1 \approx v_2 & q_{12} = 1 & q_{21} = 1 \\ v_1 < v_2 & q_{12} = 1 & q_{21} = \left(\dfrac{v_2}{v_1}\right)^{1/2} \end{array} \tag{5.4}$$

For further simplification of the model, Bruin and Prausnitz [9] suggested the pure energy interaction parameters (g_{11} and g_{22}) be estimated from pure component property data specifically using the internal energy of complete vaporization, which is shown as follows in Equation 5.5:

$$g_{ii} = -\beta E_i \tag{5.5}$$

where β is the proportionality constant and E_i is the energy change upon isothermal vaporization from the saturated liquid i to the ideal-gas state. An expression for E_i can be derived from the Clausius-Clapeyron equation [9]. Such modification leaves g_{ij} as the only adjustable parameter in Equation 5.3.

Bruin and Prausnitz [9] tested six variations of the modified NRTL model using 130 binary VLE systems in which about 50 of the systems were aqueous systems. Two of these variations were; (1) NRTL with equal molar volumes (v_1=v_2) and one adjustable parameter, and (2) NRTL with different molar volume, size factor (q) and one adjustable parameter. When the first model was considered, the average %AAD (average absolute percentage deviation) in pressure and AAD (average absolute deviation) in vapor mole fraction were approximately 2 to 8 times the error found using the original NRTL model for aqueous systems. Their second model resulted in comparable errors in pressure and vapor mole fraction to that of the original NRTL model. Although the error reduced significantly in the second case (one parameter with volume ratio), the equation lacks simplicity due to the additional calculation of volume ratio and q_{ij} parameters using the conditional statement shown in Equation 5.4.

Vetere in 1977 [18] followed the work of Bruin and Prausnitz [9] on generalizing the parameters of the NRTL model. He proposed an empirical method to estimate the parameter g_{ij} in the NRTL model. The proposed model employs a modified form of Equation 5.5 to determine the two pure interaction NRTL parameters g_{11} and g_{22}, as shown in Equation 5.6:

$$\begin{aligned} g_{ii} &= -(\Delta H_{vi} - RT) \\ g_{ij} &= f(\delta_i, \delta_j) \end{aligned} \tag{5.6}$$

where H_v is heat of vaporization and δ is the Hildebrand solubility parameter of a pure compound.

The only unknown parameter g_{ij} is determined by using the Hildebrand solubility parameters of the pure compounds. Vetere [11-13] showed the use of various empirical forms of the above concept to estimate the cross-interaction parameter (g_{ij}) of the NRTL model. For aqueous and non-aqueous systems, he presented the following equations to determine g_{ij} [12, 13]:

$$\begin{aligned} (g_{ij} - g_{ii}) + (g_{ij} - g_{jj}) &= A + B(\delta_i - \delta_j) \\ \frac{\tau_{ji}}{\tau_{ij}} &= a + b(g_{ii} - g_{jj}) \end{aligned} \tag{5.7}$$

The values of the new parameters a, b, A and B are generalized for five aqueous and five non-aqueous classes of mixtures. All four parameters were regressed for each chemical class using selected binary systems. The modified NRTL model was evaluated using over 60 binary non-aqueous and a limited number of aqueous systems. The predictive capability of the model was comparable to that of the UNIFAC-1991 [19] model for the selected systems.

Although the generalized model provided an alternative way of estimating g_{ij}, the model suffers various limitations, which includes lack of simplicity because of the additional empirical equations and the four parameters which need to be regressed for each type of chemical class. Another limitation is the inability to define effectively the chemical class of a compound in a binary system.

Further, the parameter generalization was conducted for a small number of binary systems; therefore, the model had limited capability to predict VLE properties of systems with diverse functional-group interactions.

5.3. Modified NRTL models

This section discusses the proposed modified NRTL model for VLE and LLE systems. The first model (mNRTL2) recasts the original NRTL model parameters so that the two new parameters reflect different characteristics. In the second model (mNRTL1), pure-component properties are used to generalize one of the parameters of the proposed model for VLE systems.

5.3.1. Two-parameter modified NRTL model (mNRTL2)

The NRTL model parameters can be written as shown in the following equation:

$$a_{ij} = \left(g_{ij} - g_{jj}\right) = g_{ij}\left(1 - \frac{g_{jj}}{g_{ij}}\right) = g_{ij}\left(1 - r_{ij}\right) = g_{ij}R_{ij} \tag{5.8}$$

Thus, the original two-parameter NRTL model is presented in terms of a binary interaction energy parameter, g_{ij}, and interaction ratio, R_{ij}. Here, various empirical modeling alternatives could be proposed as potential modifications of the NRTL model. Through trial and error, we found the best scenario to be described as follows: employing the following combination rule (half harmonic mean) assumption.

$$g_{ij} = \left[a_{ji}a_{ij}\right]^{1/2} = \left[\left(g_{ij} - g_{ii}\right)\left(g_{ij} - g_{jj}\right)\right]^{1/2} = \left[\frac{g_{ii}g_{jj}}{g_{ii} + g_{jj}}\right] \tag{5.9}$$

and substituting Equation 5.9 into Equation 5.8, the interaction ratio (R_{ij}) becomes,

$$R_{ij} = \frac{1}{R_{ji}} \tag{5.10}$$

Thus, we obtain a modified two-parameter NRTL model (mNRTL2) with two adjustable parameters (g_{ij} and R_{ij}). The modified two-parameter NRTL model works for most of the VLE and LLE systems studied. The effectiveness of the model has been evaluated and compared with the original NRTL model. Results show comparable representations of phase equilibria properties. We also have observed that g_{ij} and R_{ij} are easier to regress than the original NRTL parameters a_{12} and a_{21}. The g_{ij} and R_{ij} values range from -500 to 1500 and from 0 to 4, respectively. Our analysis shows, compared to the original NRTL, the mNRTL2 model is easier to initialize. Values of g_{ij}=200 and R_{ij}=1 tend to be good initial values for most of the VLE systems. Further, the mNRTL2 model has a slightly lower correlation coefficient value of 0.94 compared to the original NRTL model which resulted in a correlation coefficient value of 0.97 for the VLE systems considered in this study.

5.3.2. One-parameter modified NRTL model (mNRTL1)

A generalization for the interaction ratio R_{ij}, in the proposed mNRTL2 model was obtained in terms of pure-fluid properties. After evaluating the ratios of various pure fluid properties, the ratio of acentric factor and critical pressure resulted in the best representation of equilibrium properties of VLE systems, as shown in the following equation:

$$R_{ij} = \left[\left(\frac{\omega_j}{\omega_i} \right) \left(\frac{P_{Ci}}{P_{Cj}} \right) \right] \tag{5.11}$$

where ω is acentric factor, P_c is critical pressure and the *i* and *j* subscripts are molecules type *i* and *j*, respectively. Use of Equation 5.11 permits R_{ij} to be determined from pure substance properties, leaving only one parameter, g_{12}, to be regressed.

Equation 5.11 shows the ratio of pure-fluid properties in the modified one-parameter NRLT model (mNRTL1) that are used to determine the interaction ratio parameter. This modification essentially reduces the NRTL model to a single parameter model. As discussed below, this modified model is capable of describing VLE proprieties, as well as the infinite limits of the equilibrium properties.

5.4. Representation of equilibrium experimental data

The representation capabilities of the proposed models were evaluated using a comprehensive database of VLE experimental data. The database was assembled from available sources by insuring sufficient representation of a variety of functional groups in the database. The experimental VLE data were taken from DECHEMA [10,14] and NIST-TDE [15]. The pure-component vapor pressure data were collected from DIPPR [16] and DECHEMA [10].

5.4.1. VLE database

A low-pressure binary VLE database (Oklahoma State University, OSU database I) consisting of 188 binary VLE systems totaling 4716 data points was assembled [20]. This database is comprised of systems of aliphatic and aromatic hydrocarbons, water, alcohols, ethers, sulphides and nitrile compounds. A second database, comprised of 388 binary VLE systems totaling 12,010 data points, was taken from DECHEMA [10]. A third database consisting of 340 binary systems totaling over 17,000 data points was taken from NIST-TDE [15]. In total, the database compiled in this work consists of a total of 916 binary systems formed from various combinations of 140 different compounds. A total of over 33,000 vapor-liquid equilibrium data points were assembled in the final database (Oklahoma State University, OSU-VLE Database III). In addition to pressure, temperature and mole fraction (PTXY) data, we have collected over 500 data points of infinite-dilution activity coefficient values (γ^{∞}) for 137 of the 916 VLE systems in the database [10]. The data covered a temperature range from 128 to 554 K and pressures to 58 bar; however, over 99% of the data were at pressure of less than 10 bar.

The compounds present in the OSU-VLE Database III were classified in a similar manner as the UNIFAC functional-group classification approach [2]. The database is composed of compounds belonging to 31 chemical classes.

Figure 5.1 illustrates the data distribution of the binary systems in the OSU database III based on chemical classes. The number of systems represented for each type of functional-group interaction is shown in the figure. Systems containing alcohol or alkane components are represented extensively in the database due to their abundant data.

5.4.2. Interaction parameter regression methodology

Regression analyses were conducted to optimize the adjustable parameter or parameters in the original NRTL, mNRTL2 and mNRTL1 models. The regression analyses were performed by applying the Gibbs equilibrium criteria for a closed system to the coexisting liquid and vapor phases, while subject to mass balance constraints. The split approach was employed in the phase equilibria calculations, as follows:

$$\hat{\phi}_i^V P y_i = \gamma_i P_i^{\circ} \phi_i^V x_i \lambda_i ; \qquad i = 1, n \tag{5.12}$$

where n is the number of components, the subscript i represents a particular component, $\hat{\emptyset}^V$ is the component fugacity coefficient in the vapor phase, y is the vapor mole fraction, γ is the component activity coefficient in the liquid phase, P is the mixture pressure, P° is the pure-component vapor pressure, $\emptyset^V$ is the pure-component fugacity coefficient in the vapor phase, x is the liquid mole fraction and λ is the Poynting factor. The VLE systems considered in this study were generally at low pressure; hence, the vapor-phase fugacity coefficients were assumed to be 1. We have also investigated the quality of representation when equation-of-state (EOS) models are used to calculate the vapor-phase fugacity coefficients (results not shown). Our findings show there is no improvement on the overall representation error. This result confirms that our assumption is reasonable.

The Poynting factor is expressed as follows:

$$\lambda_i = \exp\left(\frac{v_i^L (P - P_i^\circ)}{RT} \right) \tag{5.13}$$

where v^L is the liquid molar volume and is determined using the Rackett equation [21].

The objective function, *OF*, used in the parameter regression analyses, was the weighted sums of squares of relative errors in pressure, K-values, infinite-dilution activity coefficients and weighted absolute sum of model parameters, as follows:

$$\begin{aligned} OF = \sum_{i=1}^{n} w_1 \left(\frac{P^{Exp} - P^{Calc}}{P^{Exp}} \right)_i^2 + w_2 \sum_{i=1}^{n} \left(\frac{K_{values}^{Exp} - K_{values}^{Calc}}{K_{values}^{Exp}} \right)_i^2 \\ + w_3 \sum_{i=1}^{n} \left(\frac{\gamma_{values}^{\infty\ Exp} - \gamma_{values}^{\infty\ Calc}}{\gamma_{values}^{\infty\ Exp}} \right)_i^2 + w_4 (Par) \end{aligned} \tag{14}$$

where the weights were $w_1 = 1$; $w_2 = 1/15$; $w_3 = 1/10$; $w_4 = 2E - 6$; *n* is the number of data points, *Par* is $|a_{12}| + |a_{21}|$ for the NRTL model and $|g_{12}|$ for the mNRTL1 model and the superscripts *Exp* and *Calc* refer to experimental and calculated values, respectively.

This objective function and associated weights were developed after evaluating the VLE property predictions employing various objective function formulations. Equation 5.14 was found to be the most suitable since the equation provided a balance of the model prediction errors for temperature, pressure, equilibrium constants, activity coefficient and vapor mole fraction and also reduced the correlation of the two model parameters (a_{12} and a_{21} or g_{12} and R_{12}) [22].

5.4.3. Case studies

Three regression case studies were conducted to investigate representation qualities of the original NRTL, mNRTL2 and mNRTL1 models. In all case studies, the ideal gas (IG) model was used to describe the gas phase behavior. The case studies are described below:

Case original NRTL: The original NRTL model was used to represent the activity coefficients by regressing a_{12} and a_{21}.

Case mNRTL2: The g_{12} and R_{12} parameters in the modified NRTL model were regressed to represent the experimental data.

Case mNRTL1: The one-parameter modified NRTL model was evaluated by regressing g_{12}, with the second parameter set by Equation 5.11.

The representation capabilities of the models were assessed for equilibrium properties such as pressure (*P*), infinite-dilution activity coefficients (γ^{∞}), temperature (*T*), component 1 vapor mole fraction (y_1) and equilibrium K-value (average of K_1 and K_2). The regression was conducted by performing a bubble-point pressure calculation. After the regression analyses, the regressed parameters were used directly to calculate (a) *P*, K_1 and K_2 for known *T* and x_1 and (b) *T* for known *P* and x_1.

5.4.4. Behavior classification

The degree of non-ideality depends on the particular types of molecular interactions encountered by the components of the system considered. Components with similar functional groups, polarity and sizes usually show nearly ideal behavior while components with a high degree of polarity difference exhibit highly non-ideal behavior. Although the types of molecules provide a general idea about mixture behaviors, they do not allow the precise determination of the degree of non-ideality [23]. The alternative is to employ a qualitative approach which relates model parameter values to the behavior of the systems.

The ability to classify behaviors qualitatively of VLE systems is important in process design since this provides an easy method of determining the degree of non-ideality without the need of additional information. Once the type of behavior is identified, an appropriate thermodynamic model can be selected to determine properties of the systems.

Danner [23] presented a behavior classification approach based on the Margules model parameter (A) for 104 VLE systems. In this approach, he presented the relationship of excess Gibbs energy to the A parameter value. Danner classified systems with an A value of between -0.6 and 0.6 to be nearly-ideal while a value greater than 0.6 or less than -0.6 were classified as highly non-ideal.

In this study, we have applied the same approach as Danner [23] to determine the g_{12} values for ideal and non-ideal systems in the mNRTL1 model. Theoretically for ideal systems, the excess Gibbs energy is zero. In order to determine the cutoff point between ideal and non-ideal systems, we plotted G^E/RT as a function x_1 for all the systems in our database. From the plots the maximum $|G^E/RT|$ values were determined and compared with the Margules (A) and mNRTL1 (g_{12}) model parameters.

Figure 5.2 shows the G^E/RT vs. x_1 plots for six VLE binary systems. The systems were selected to demonstrate the change in behavior from nearly ideal to highly non-ideal systems. The degree of non-ideality increases as the maximum $|G^E/RT|$ value increases.

Table 5.1 shows the maximum $|G^E/RT|$ and $\gamma\infty$ values for the selected six binary systems with regressed Margules (A) and mNRTL1 (g_{12}) model parameters. The table indicates for systems 2 and 5 the |A| parameter and maximum $|G^E/RT|$ values are approximately 0.62 and 0.15, respectively. In addition, the $\gamma\infty$ values for systems 2 and 5 are approximately 2 and 0.5, respectively. Based on the Danner [23] classification, these systems are classified as highly-non ideal. After examining all the systems in our database, the boundary for nearly ideal systems was found to occur at a maximum $|G^E/RT|$ value <= 0.15. The relationship of maximum $|G^E/RT|$ and the mNRTL1 (g_{12}) model parameter value is discussed in the Result Section.

5.5. Results and discussion

The results of this study address five main concerns, which are (1) representation of equilibrium properties, (2) behavior classification, (3) cross-phase system predictions (the applicability of LLE regressed parameters to predict VLE properties and vice versa), (4) parameter temperature dependence and (5) multicomponent phase behavior predictions. The results of each focus is presented in the following sub sections.

5.5.1. Regression of equilibrium properties

The representation capabilities of the original NRTL, mNRTL2 and mNRTL1 models were assessed by using experimental $T,\ P,\ y_1,\ K$ and γ^∞ values of 916 binary systems and γ^∞ data of 137 binary systems. The regression errors from each model were analyzed by calculating the root-mean-squared error (RMSE), bias and %AAD.

Table 5.2 provides the property representation errors for the original NRTL, mNRTL2 and mNRTL1 case studies. As shown in the table, the original NRTL model with regressed parameters provided overall %AADs of 2.1, 0.2, 4.3, 5.5 and 8.7 for $P,\ T,\ y_1, K$ and γ^∞, respectively. The mNRTL2 model provided overall %AADs of 2.2, 0.2, 4.4, 5.7 and 10.2 for $P,\ T,\ y_1, K$ and γ^∞, respectively. The results show the mNRTL2 provided comparable results to that of the original NRTL model. This indicates that the modified model performs equally well as the original NRTL model. The one-parameter (mNRTL1) model resulted in overall %AADs of 2.5, 0.2, 4.7, 6.1 and 13.3 for $P,\ T,\ y_1, k$ and γ^∞, respectively. Compared to the mNRTL2 model, the mNRTL1 model provided good VLE property representation with a slight loss of precision. With only one parameter, the mNRTL1 was able to successfully represent VLE properties including infinite-dilution activity coefficients (γ^∞), which has previously been a challenge for one-parameter models.

Table 5.3 shows the property representation errors using the original NRTL, mNRTL2 and mNRTL1 models for binary VLE systems containing water. The property representations errors for water systems were slightly higher than the results found for the overall data. The mNRTL2 model resulted in comparable results with the original NRTL model for the water systems. The mNRTL1 model also provided reasonable precision in representing experimental data for water systems. These higher errors could be due to the high level of experimental uncertainty associated with water systems and the inability of the models in representing such systems precisely. Further, the mole fraction of aqueous systems tend to be small which results in large percentage errors.

Figure 5.3 shows the distribution of pressure representation regression errors for the original NRTL, mNRTL2 and mNRTL1 models by functional-group interactions. As indicated by the error matrix, all three models have comparable representation capabilities for all type of interactions with the exception of the water systems. As expected, all models provided precise representations when the components in the system have the same functional groups (diagonal elements of the triangular matrix). The mNRTL1 model showed slightly higher errors for some of the interactions involving water. The results for water systems are inconclusive since the database lacks good representation of each type of interaction with water for a number of systems.

5.5.2. Behavior classification

Table 5.4 shows the maximum $|G^E/RT|$ range of values that were used to identify the degree of non-ideality in the VLE systems. Systems with a maximum $|G^E/RT|$ value <= 0.15 are classified as nearly-ideal, while $|G^E/RT|$ values > 0.15 are classified as highly non-ideal systems.

Figure 5.4 shows the distribution of g_{12} values based on maximum $|G^E/RT|$ values for 913 systems. The figure shows the nearly-ideal system g_{12} range (approximately between -170 and 220) and the highly non-ideal system ranges. As indicated in the figure, the highly non-ideal system range overlaps the nearly-ideal system range on both the left and right sides. To avoid misclassifying systems in the overlapping region, we considered only correctly classified systems in Figure 5.5, which eliminated the overlap shown in Figure 5.4.

Table 5.5 provides the range of the g_{12} parameter for nearly-ideal and highly non-ideal system classes excluding those systems that are in the overlapping region. The result shows the g_{12} range of the nearly-ideal systems is approximately between -100 and 100 while highly non-ideal systems are >220 and <-180. The systems in the overlapping regions could be considered as non-ideal systems due to the fact that they cannot be classified as nearly-ideal or highly non-ideal. The range of g_{12} for non-ideal systems are between -180 and -100 and 100 and 220. The classification results confirm that when the interaction energy value increases the degree of the non-ideality also increases.

Table 5.6 shows the pressure property representation errors using the original NRTL, mNRTL2 and mNRTL1 models for nearly-ideal, non-ideal and highly non-ideal systems. As expected, the representation quality decreases as degree of non-ideality increases. The result also shows the three models have comparable representation capability for nearly-ideal systems. For non-ideal and highly non-ideal systems, the original NRTL model provided slightly better representation of pressure compared to the modified NRTL models.

5.5.3. Cross-phase property predictions

The representation capabilities of the original NRTL and mNRTL2 models were evaluated using VLE and LLE experimental data. In this study, twenty systems with both binary VLE and LLE experimental data were gathered from the VLE and LLE DECHEMA databases [10, 14]. Regression analyses were carried out for VLE, LLE and VLE-LLE (LLE and VLE data combined) systems. The original NRTL and mNRTL2 model parameters found in the regression analyses were different for the VLE and LLE systems with the same components. We investigated the source effect of model parameters (from VLE or LLE or VLE-LLE regressions) on the property predictions of different phases.

Tables 5.7 and 5.8 show the regression results of the 20 VLE, LLE and VLE-LLE binary systems using the original NRTL and mNRTL2 models. The results show the two models have comparable representation capability for correlating experimental mole fractions. The %AADs on liquid mole fraction were approximately 18 and 20 for LLE and VLE systems, respectively. The error for the combined VLE-LLE data increased slightly to 26% for both models.

The robustness of the two models was investigated by predicting equilibrium properties of VLE, LLE and VLE-LLE systems using NRTL parameters regressed only using LLE, VLE or VLE-LLE data. Tables 5.7 and 5.8 show the prediction of x_1 for VLE and x_1 in Phase 1 for LLE and VLE-LLE data using NRTL parameters regressed only from LLE, VLE or VLE-LLE data. Using both models, the results show parameters from the VLE-LLE regression provided the lowest errors when used for VLE and LLE systems. In both models, the LLE parameters resulted in relatively better VLE and VLE-LLE property prediction compared to the VLE parameters when used for LLE and VLE-LLE systems.

This study revealed the lack of robustness of the NRTL model in handling both VLE and LLE properties with the same regressed parameters. This could be due to the strong temperature dependence of the model parameters, especially for the LLE systems. Improved accounting for the temperature dependence of LLE systems may be attained incorporating equation-of-state interaction concepts within the modified NRTL model.

5.5.4. Temperature dependence of the mNRTL2 model g_{12} parameter

The effect of temperature on the g_{12} parameter of the mNRTL2 model was examined. Six LLE and VLE systems listed in Table 5.9 were collected from the DECHEMA VLE [10] and

DECHEMA LLE databases [14]. Regression analyses were carried out to determine the optimum value of g_{12} and R_{12} parameters for LLE and VLE systems. In the regression analysis, the R_{12} values were fixed at the LLE regressed value while g_{12} is regressed temperature by temperature for both the LLE and VLE systems.

Figures 5.6a and 5.6b show the g_{12} temperature by temperature regression results of the six VLE and LLE systems. The error bars indicate the range of g_{12} values that correspond to a ±25% increase in the property prediction errors. The dotted lines are drawn to clearly indicate g_{12} values that belong to a same component VLE and LLE mixture. Systems 1, 2, 3 and 4 show the value of g_{12} increases as temperature increases, which indicates that VLE systems tend to have higher g_{12} values than LLE systems. The two exceptions are Systems 5 and 6 where the results show an inverse relationship of g_{12} and temperature. In general, the g_{12} parameter has a linear type of relationship with temperature for the LLE and VLE systems.

5.5.5. Multicomponent phase behavior predictions

The prediction capabilities of the original NRTL, mNRTL2 and mNRTL1 models were evaluated for multicomponent systems. The objectives of this study is to assess the representation capability of the three models for ternary VLE property predictions using interaction parameters obtained from regression of binary VLE experimental data. To accomplish this, we assembled a database of 57 ternary VLE systems encompassing a variety of molecular species. Regressed binary model parameters were used to predict the phase equilibrium properties of the ternary systems.

Table 5.10 shows the prediction of ternary properties using the original NRTL, mNRTL2 and mNRTL1 models. The original NRTL equation resulted in %AADs of approximately 3, 0.3 and 9 for pressure, temperature and K-value predictions, respectively. Compared to the original NRTL, the mNRTL2 and mNRTL1 model resulted in slightly higher %AADs. The mNRTL2 and mNRTL1 models provided comparable predictions for ternary systems. In all cases, the averaged errors are within 1.5 times the errors found from binary system regression analyses. The results indicate all three models could be extended to multicomponent phase behavior predictions with only a slight loss of accuracy.

5.6. Conclusion

In this study, we proposed a modification to the widely used NRTL activity coefficient model which addresses the limitation of the original model. The representation capabilities of the models were assessed with 916 VLE and 20 binary LLE systems. The regression results indicate the newly proposed model provides comparable results with the original NRTL model.

The study provided a generalization for the interaction ratio in the newly proposed model using pure-component properties. This reduces the model to only one energy interaction parameter and eliminates the correlation between parameters. Compared to the original NRTL model, the one-parameter model provided VLE equilibrium property representations with a slight loss of accuracy. A study is underway to further generalize the model by relating the energy interaction parameter to the structures of molecules in the binary systems.

Model parameters for VLE and LLE systems are different for both the original and modified models. Further, the model parameters in both models show strong temperature dependence for the LLE systems. This suggests there is room for improving the temperature dependence of activity coefficients in the NRTL model. A potential concept that may lead to improved accounting of temperature dependence is incorporating equation-of-state interaction concepts within the modified NRTL model.

Table 5.1. Maximum $|G^E/RT|$ and $\gamma\infty$ properties for six VLE binary systems with their regressed Margules (A) and mNRTL1 (g_{12}) parameters

Sys	Type	Comp 1	Comp 2	A	g_{12}	Original NRTL		
						Max $\|G^E/RT\|$	$\gamma\infty_1$	$\gamma\infty_2$
1	Positive Excess Gibbs	n-octane	ethylbenzene	0.19	66	0.05	1.2	1.2
2		1,2-dichloroethane	tetrachloroethylene	0.62	236	0.16	1.9	1.8
3		ethyl tertiary butyl ether	ethanol	1.43	547	0.37	3.5	5.9
4	Negative Excess Gibbs	methanol	pyridine	-0.12	-37	-0.04	0.9	0.8
5		hexafluorobenzene	p-xylene	-0.62	-182	-0.16	0.6	0.5
6		butylamine	1-propanol	-1.25	-338	-0.31	0.2	0.3

Table 5.2. VLE property representation capability of the original NRTL, mNRTL2 and mNRTL1 models

Model (L)	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD
Original NRTL	a_{12} & a_{21}	P (bar)	916	33841	0.15	0.00	2.1
		T (K)	916	33841	1.35	0.10	0.2
		y_1	675	18199	0.03	0.00	4.3
		K-value	675	18199	5.09	-0.31	5.5
		$\gamma\infty$	137	549	3.54	-0.21	8.7
mNRTL2	g_{12} & R_{12}	P (bar)	916	33844	0.17	0.00	2.2
		T (K)	916	33844	1.42	0.11	0.2
		y_1	675	18199	0.03	0.00	4.4
		K-value	675	18199	4.84	-0.28	5.7
		$\gamma\infty$	137	549	4.70	-0.34	10.2
mNRTL1	g_{12}	P (bar)	916	33845	0.24	-0.01	2.5
		T (K)	916	33845	1.67	0.16	0.2
		y_1	675	18199	0.03	0.00	4.7
		K-value	675	18199	5.41	-0.21	6.1
		$\gamma\infty$	137	549	6.46	-0.77	13.3

Table 5.3. VLE properties representations capability of original NRTL, mNRTL2 and mNRTL1 models for water systems

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD
Original NRTL	a_{12} & a_{21}	P (bar)	55	4344	0.40	-0.02	4.8
		T (K)	55	4344	2.47	0.41	0.4
		y_1	47	2313	0.06	-0.01	10.6
		K-value	47	2313	17.40	-3.98	11.6
mNRTL2	g_{12} & R_{12}	P (bar)	55	4344	0.35	-0.02	5.7
		T (K)	55	4344	2.72	0.51	0.5
		y_1	47	2313	0.07	-0.01	11.7
		K-value	47	2313	16.29	-3.57	13.2
mNRTL1	g_{12}	P (bar)	55	4344	0.70	-0.10	7.2
		T (K)	55	4344	3.99	1.11	0.6
		y_1	47	2313	0.08	0.00	13.2
		K-value	47	2313	18.44	-2.41	15.6

Table 5.4. Classification of binary systems based on max $|G^E/RT|$

No	Range	Type
1	Max $|G^E/RT| <=0.15$	Nearly-ideal
2	Max $|G^E/RT| > 0.15$	Highly non-ideal

Table 5.5. Range of parameters excluding systems that are in the overlapping region

No	Type	Max G^E/RT Range	g_{12} Range	Average $\gamma\infty$	No. of sys.
1	Nearly-ideal	Max $|G^E/RT| <= 0.15$	$|g_{12}| <= 100$	1.1	401
2	Non-ideal	Max $|G^E/RT| \approx 0.15$	$-180 < g_{12} < -100$ $\& 100 < g_{12} < 220$	2.9	167
3	Highly non-ideal	Max $|G^E/RT| > 0.15$	$g_{12} <= -180$	0.3	21
	Highly non-ideal	Max $|G^E/RT| < -0.15$	$g_{12} >= 220$	12	324

Table 5.6. Regression results of the original NRTL, mNRTL2 and mNRTL1 by type of behavior

Type	**No. of sys.**	**%AAD on pressure**		
		Original NRTL	**mNRTL2**	**mNRTL1**
Nearly-ideal	401	1.6	1.6	1.6
Non-ideal	167	1.9	2.3	2.4
Highly non-ideal	345	2.7	2.9	3.6

Table 5.7. Regression results of 20 VLE and LLE systems using the original NRTL model

Type of data	**No. of sys.**	**No. of pts.**	**%AAD for on x_1 VLE and x_1 in Phase 1 for LLE**			
			Original-NRTL regression	**Parameters from LLE regression**	**Parameters from VLE regression**	**Parameters from VLE-LLE**
LLE	20	108	18.0	18.0	61.8	33.2
VLE	20	1231	19.9	34.0	19.9	25.9
VLE-LLE	20	1340	25.0	31.3	37.4	25.0

Table 5.8. Regression results of 20 VLE and LLE systems using the mNRTL2 model

Type of data	**No. of sys.**	**No. of pts.**	**%AAD for on x_1 VLE and x_1 in Phase 1 for LLE**			
			mNRTL2 regression	**Parameters from LLE regression**	**Parameters from VLE regression**	**Parameters from VLE-LLE**
LLE	20	108	17.8	17.8	64.1	32.0
VLE	20	1231	20.1	34.9	20.1	27.8
VLE-LLE	20	1340	26.6	32.0	38.3	26.6

Table 5.9. LLE and VLE systems used for the temperature dependence study of the g_{12} parameter

No	System	No	System
1	methanol + hexane	4	methanol + cyclohexanol
2	diethyl ether + water	5	acetonitrile + water
3	diisopropyl ether + water	6	nitromethane + cyclohexane

Table 5.10. Prediction results of 57 ternary VLE systems using the original NRTL, mNRTL2 and mNRTL1 models

Study	Parameters	No. of sys.	Property	No. of pts.	RMSE	Bias	%AAD
Original NRTL	a_{12} & a_{21}	57	P (bar)	2212	0.05	0.00	3.0
			T (K)	2212	1.57	0.16	0.3
			y_1	1890	0.04	0.00	8.7
			K-values	1890	0.43	0.01	8.3
mNRTL2	g_{12} & R_{21}	57	P (bar)	2212	0.05	0.01	3.7
			T (K)	2212	1.71	-0.29	0.3
			y_1	1890	0.04	0.00	9.6
			K-values	1890	0.37	0.02	9.1
mNRTL1	g_{12}	57	P (bar)	2212	0.05	0.01	3.8
			T (K)	2212	1.71	-0.31	0.3
			y_1	1890	0.04	0.00	9.9
			K-values	1890	0.38	0.02	9.4

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	13																														
2	Aldehyde	10	1																													
3	Alkane	24	5	14																												
4	Alkene	9	1	10	3																											
5	Alkyne	5	3	5	6	2																										
6	Amide	6	2	6	2	1																										
7	Amine	5		4			3	4																								
8	Aromatic Bromo	5		3			1																									
9	Aromatic Floro	2		2			1	1	1																							
10	Benzene Derivative	6	3	13	5		1	5	1	3	4																					
11	Bromoalkane	15		5					1	1	8																					
12	Carboxylate	2	5	9	1						6	1	3																			
13	Chloroalkane	5		5	2	2	4	6		2	8	3	4	2																		
14	Chloroalkene	19	1	7		1	1				1		1	8	1																	
15	Chlorobenzene	9		2	2		1	4	1	1	2	1		2	1																	
16	Epoxide	7	3	6							1		2	4																		
17	Ester	1	1	8	1	1	1	1			4	1	1	5	1	1	1															
18	Ether	12	2	21	3	3	2	2		3	5	2	1	9	2	2	1	3	3													
19	Furfural	1		3	1						2			4	1			1														
20	H2S	1		1																												
21	Iodoalkane	3	1	1							2	1	1	4				1	1													
22	Ketone	3	4	21	3	1	2	5	1		8	1	6	8	7	3	1	3	2	2		1	4									
23	Nitrile	4		4	2	2	1	1	1		4		4	6	3	1		1	1				1	1								
24	Nitrite	1													1									1								
25	Nitro Compound	12		3	2	2	1		1		5	1	2	5	1	2		3	3			2	3	2		2						
26	Pyridine Derivative	14		4			1			1	1	1	1	2	1	1		2	1				1	1			1					
27	Sulfide	4		4	3	3	1	1	1		1	2	2	5	2	1		1	1		1	1	1	1		2						
28	Thiol	1		7			2	1			1								1		1		1	1				3				
29	Thiophene	4		1	2		1				1			1					1							1	1					
30	Toluene Derivative	3	6	4	2		1	2	1	1	2	3	2	3	1	1	1		5	1		1	5	1		2	2	2	2	1	1	
31	Water	9	1	2			1	10			3	1	3		1		2	1	4	1			5	3	1	2	3	1	1			

X

Y ■# Number of available binary systems consisting of chemicals with functional groups of X and Y

□ No VLE data used

Figure 5.1. Database matrix of the compounds in the OSU-VLE database III

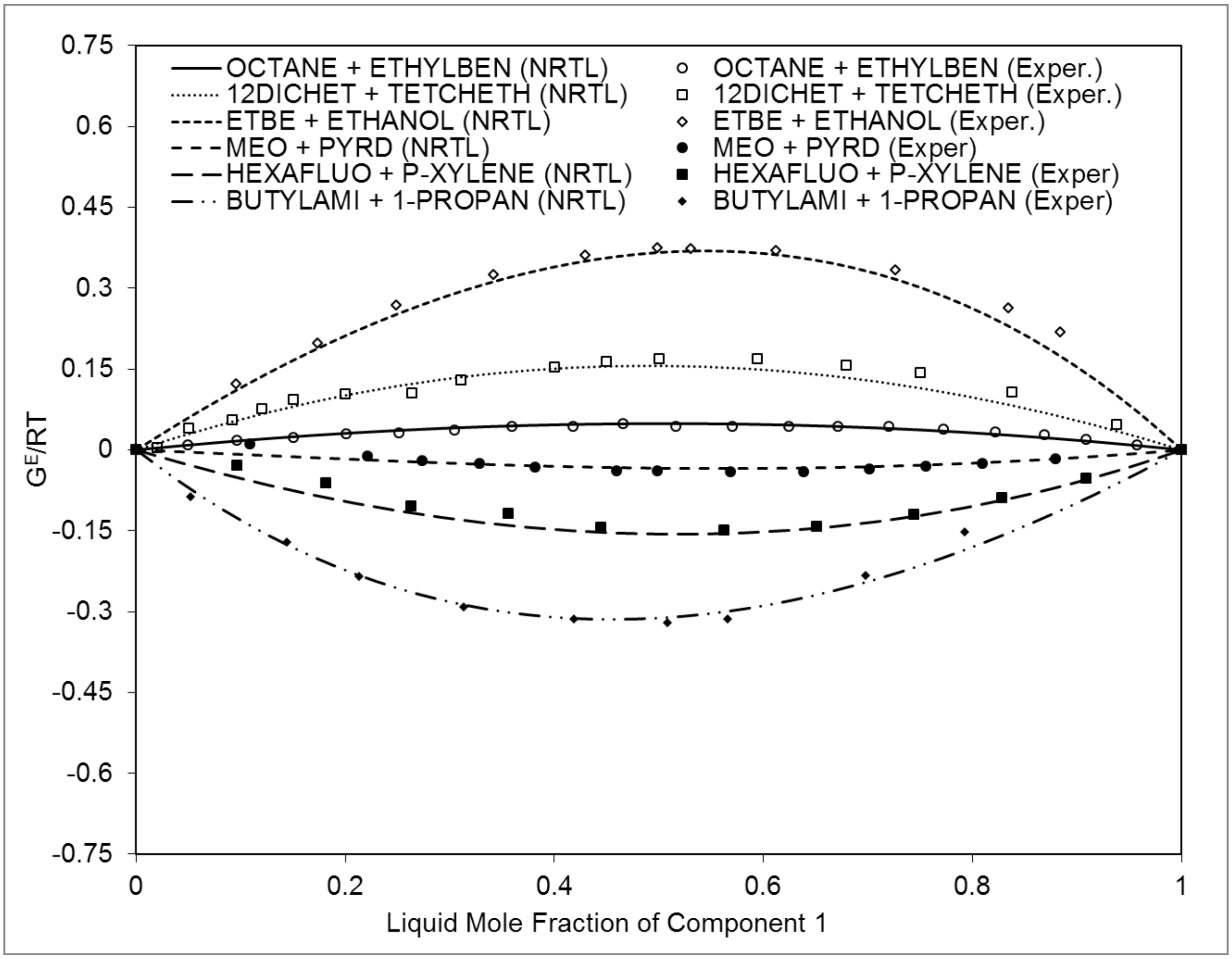


Figure 5.2. Excess Gibbs energy for six VLE binary systems in the OSU-VLE-III database

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	2 2 2																														
2	Aldehyde	3 3 4	1 1 1																													
3	Alkane	4 4 5	1 1 1	1 1 1																												
4	Alkene	3 3 4	2 2 2	1 1 1	1 1 1																											
5	Alkyne	3 3 4	1 1 1	2 2 2	0 1 1	1 1 2																										
6	Amide	2 2 2	3 3 3	4 4 4	2 2 5	2 4 4																										
7	Amine	2 2 2		3 3 3			2 2 3	1 1 1																								
8	Aromatic Bromo	2 2 3		1 1 1			6 6 6																									
9	Aromatic Floro	3 3 5		0 0 0			3 3 3	0 0 0	1 1 1																							
10	Benzene Derivative	2 2 2	2 2 2	1 1 1	2 2 2		3 3 3	2 1 1	2 2 2	2 2 2	1 1 1																					
11	Bromoalkane	2 2 3		2 3 4					0 0 0	1 1 1	3 3 3																					
12	Carboxylate	2 2 2	1 1 1	2 2 3	0 1 1						2 2 2	3 4 4	1 1 1																			
13	Chloroalkane	1 2 3		1 1 1	1 1 1	0 0 0	2 4 5	2 2 2		1 1 1	2 2 2	2 3 3	2 2 2	1 1 1																		
14	Chloroalkene	2 3 3	1 1 1	1 1 1		1 1 1	2 2 2				1 1 1		2 2 2	2 2 2	1 1 1																	
15	Chlorobenzene	3 3 3		1 1 1	2 2 2		3 3 3	1 1 1	0 0 0	0 0 0	0 1 1	1 1 1		1 1 1	2 2 2																	
16	Epoxide	2 2 2	2 2 2	1 1 1							2 2 2		1 1 1	4 4 4																		
17	Ester	6 6 6	1 1 1	1 1 2	0 0 1	0 0 0	5 5 5	2 2 2			2 2 2	1 1 1	3 3 3	1 1 1	1 1 1	1 1 1	5 5 5															
18	Ether	2 2 2	2 2 2	2 2 2	2 2 2	2 2 2	2 2 4	4 4 4		0 0 0	1 1 1	5 6 6	1 1 1	1 1 1	2 2 2	1 1 1	1 1 1	1 1 1	2 2 2													
19	Furfural	3 3 3		6 6 5	4 4 7						3 3 3			3 3 3	1 1 2			1 1 1														
20	H2S	4 4 4		3 3 7																												
21	Iodoalkane	5 5 5	2 2 2	1 1 1							1 1 1	1 1 1	1 1 1	1 1 1				0 0 0	1 1 1													
22	Ketone	1 1 1	1 1 1	2 2 2	1 1 2	1 1 1	2 2 2	1 1 2	1 1 1		2 2 2	1 1 1	2 2 2	1 2 2	2 2 2	2 2 2	1 1 1	1 1 1	0 1 1	1 1 1		2 2 2	1 1 1									
23	Nitrile	0 0 0		3 4 5	2 2 3	4 4 3	2 2 2	1 2 2	1 3 4		2 3 3		1 1 1	2 2 2	2 2 2	2 2 2		0 0 1	1 1 1				1 1 1	2 2 2								
24	Nitrite	2 2 2													3 3 3									4 4 4								
25	Nitro Compound	4 4 4		5 4 4	2 2 5	1 1 1	2 2 2		7 7 7		2 3 3	2 2 2	1 1 1	2 3 3	2 2 2	1 1 1		1 1 1	2 2 2			2 2 2	2 2 2	1 1 1		2 2 2						
26	Pyridine Derivative	2 2 2		1 1 1			0 0 0			1 2 2	2 2 2	8 8 8	0 0 0	4 4 5	1 1 1	0 0 1		2 2 2	2 2 2				1 1 1	0 1 1			1 1 1					
27	Sulfide	2 2 3		1 1 1	1 1 1	3 3 3	1 1 1	5 5 5	2 4 4		2 2 2	2 2 3	1 1 1	3 3 5	1 1 1	2 3 3		2 2 4	2 2 2		1 3 4	1 1 1	2 2 2	2 2 8		5 5 9						
28	Thiol	6 7 8		4 4 4			1 2 2	1 1 1			0 0 0								0 0 1		5 5 5		1 1 1	1 1 1				2 1 1				
29	Thiophene	2 2 2		1 1 2	2 2 2		4 4 4				0 0 0			0 0 0					1 1 1							0 0 0	2 2 2					
30	Toluene Derivative	3 3 3	3 3 3	1 1 1	1 1 1		6 6 6	2 2 2	0 0 0	1 1 1	0 0 0	2 2 2	1 1 1	2 2 2	1 1 1	1 1 1	3 3 3		2 2 2	2 2 3		1 1 1	1 1 1	2 2 2		1 1 1	1 1 1	3 3 4	4 4 4	2 2 2	5 5 5	
31	Water	2 3 3	0 0 6	8 8 18			2 2 2	5 8 11			9 9 14	1 1 1	6 6 7		7 8 15		7 7 12	7 7 4	8 8 9	8 8 8			6 6 6	4 4 5	3 3 5	5 5 7	2 5 5	4 6 10	3 4 3			

Key	
Color	Pressure %AAD Range
#	%AAD<3
#	3<%AAD<6
#	6<%AAD<10
#	10<%AAD<20

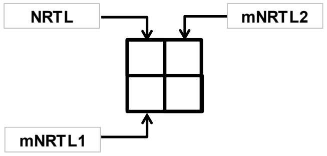


Figure 5.3. Pressure representation of the original NRTL, mNRTL1 and mNRTL2 models by type of interaction

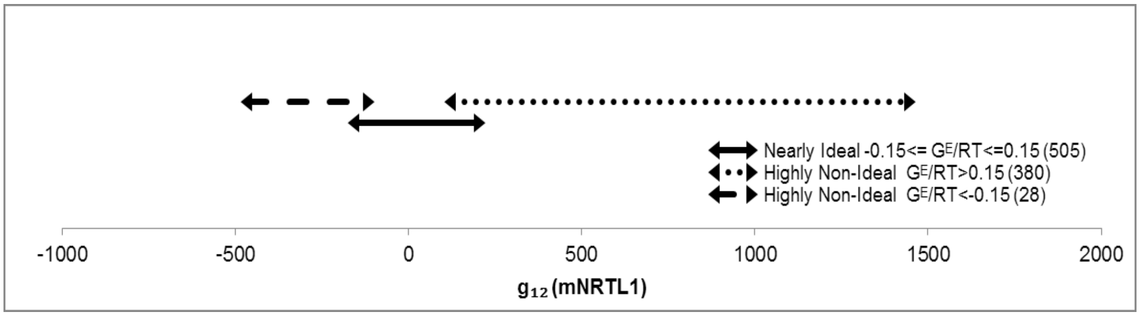


Figure 5.4. Distribution of g_{12} based on maximum $|G^E/RT|$ values in the OSU-VLE-III database

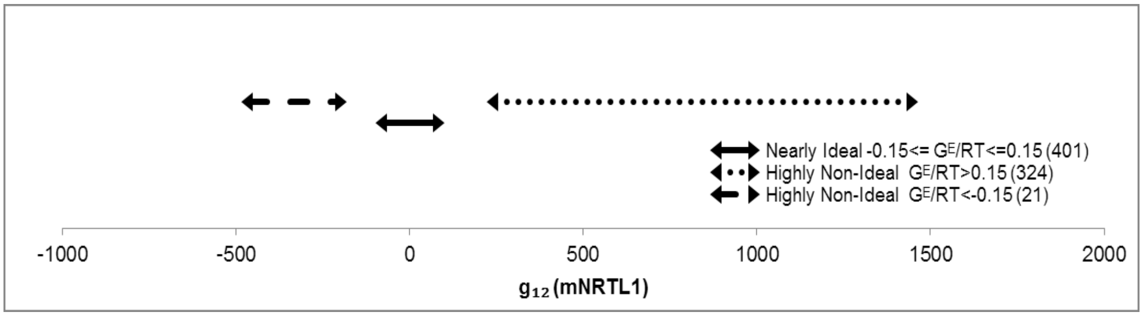


Figure 5.5. Distribution of g_{12} based on maximum $|G^E/RT|$ values excluding overlapping region systems

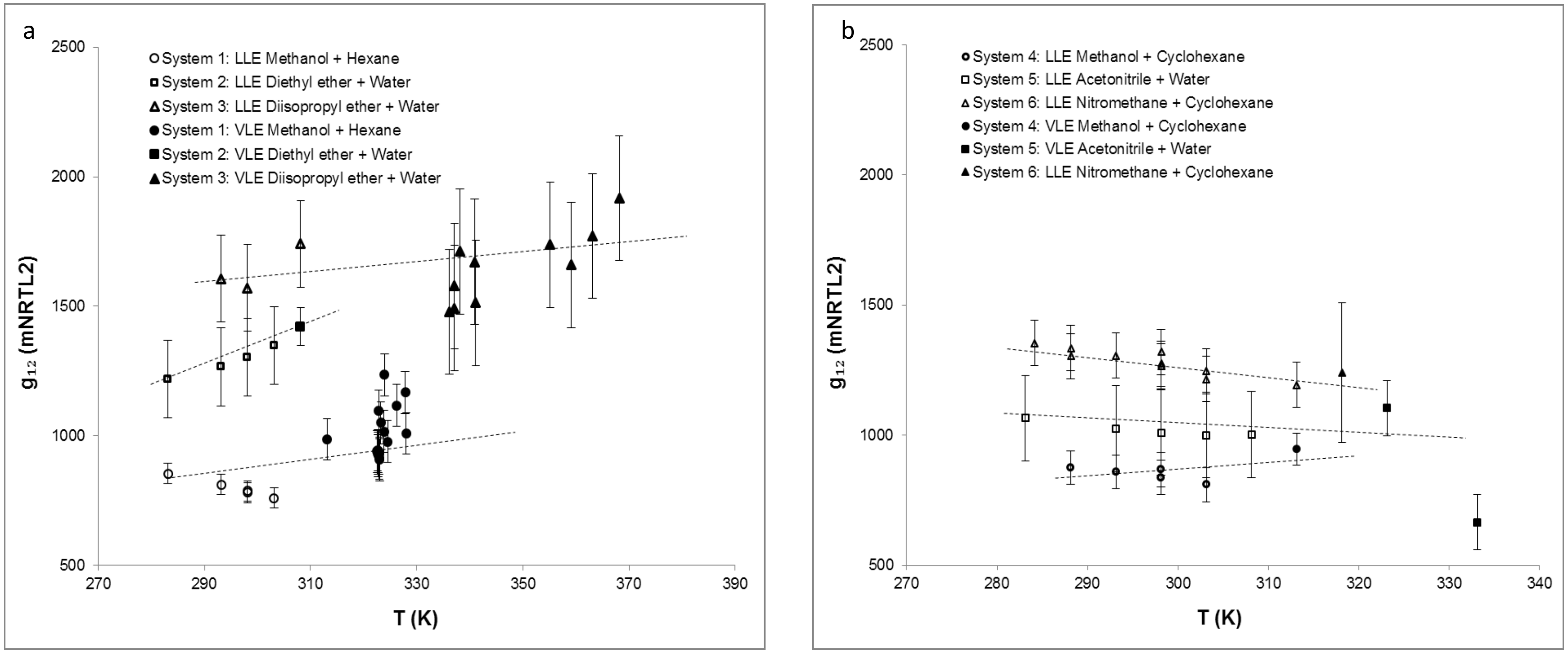


Figure 5.6. Variation of g_{12} with temperature for VLE and LLE systems where R_{12} is fixed as the LLE regressed value (a) systems 1-3 and (b) systems 4-6

REFERENCES

1. Renon, H. and J.M. Prausnitz, *Local compositions in thermodynamic excess functions for liquid mixtures.* AIChE Journal, 1968. **14**(1): p. 135-144.

2. Gmehling, J., J. Li, and M. Schiller, *A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties.* Industrial & Engineering Chemistry Research, 1993. **32**(1): p. 178-193.

3. Abrams, D.S. and J.M. Prausnitz, *Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems.* AIChE Journal, 1975. **21**(1): p. 116-128.

4. Skjold-Jorgensen, S., B. Kolbe, J. Gmehling, and P. Rasmussen, *Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(4): p. 714-722.

5. Fischer, K. and J. Gmehling, *Further development, status and results of the PSRK method for the prediction of vapor-liquid equilibria and gas solubilities.* Fluid Phase Equilibria, 1996. **121**(1-2): p. 185-206.

6. Prausnitz, J.M., R.N. Lichtenthaler, and E.G.d. Azevedo, *Molecular thermodynamics of fluid-phase equilibria*. 3rd ed. 1998: Prentice-Hall.

7. Wilson, G.M., *Vapor-liquid equilibrium. XI. A new expression for the excess free energy of mixing.* Journal of the American Chemical Society, 1964. **86**(2): p. 127-130.

8. Gmehling, J., D. Tiegs, and U. Knipp, *A comparison of the predictive capability of different group contribution methods.* Fluid Phase Equilibria, 1990. **54**: p. 147-165.

9. Bruin, S. and J.M. Prausnitz, *One-Parameter Equation for Excess Gibbs Energy of Strongly Nonideal Liquid Mixtures.* Industrial & Engineering Chemistry Process Design and Development, 1971. **10**(4): p. 562-572.

10. Gmehling, J., U. Onken, and W. Arlt, *Vapor-liquid equilibrium data collection*. Chemistry Data Series. Vol. I, Parts 1-8. 1977 - 2001: DECHEMA, Frankfurt, Germany.

11. Vetere, A., *An improved method to predict VLE equilibria of subcritical mixtures.* Fluid Phase Equilibria, 1996. **124**(1–2): p. 15-29.

12. Vetere, A., *Prediction of vapor-liquid equilibria of aqueous systems in the subcritical range by using the NRTL equation.* Fluid Phase Equilibria, 1994. **99**(0): p. 63-74.

13. Vetere, A., *Prediction of vapor-liquid equilibria of non-aqueous systems in the subcritical range by using the NRTL equation.* Fluid Phase Equilibria, 1993. **91**(2): p. 265-280.

14. Arlt, W., M.E.A. Macedo, P. Rasmussen, and J.M. Sorensen, *Liquid-liquid equilibrium data collection*. Chemistry Data Series. Vol. V, Parts 1-4. 1979 - 1987: DECHEMA, Frankfurt, Germany.

15. *NIST-TDE, NIST Standard Reference Database 103b ThermoData Engine*. 2012.

16. *DIPPR Project 801, Physical and Thermodynamic Properties of Pure Chemicals*. 2011.

17. Scatchard, G., S.E. Wood, and J.M. Mochel, *Vapor-Liquid Equilibrium. VII. Carbon Tetrachloride-Methanol Mixtures1.* Journal of the American Chemical Society, 1946. **68**(10): p. 1960-1963.

18. Vetere, A., *A modified Heil-Prausnitz equation for excess gibbs energy.* The Canadian Journal of Chemical Engineering, 1977. **55**(1): p. 70-77.

19. Hansen, H.K., P. Rasmussen, A. Fredenslund, M. Schiller, and J. Gmehling, *Vapor-liquid equilibria by UNIFAC group contribution. 5. Revision and extension.* Industrial & Engineering Chemistry Research, 1991. **30**(10): p. 2352-2355.

20. Ravindranath, D., B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior. Fluid Phase Equilibria, 2007. 257(1): p. 53-62.

21. Rackett, H.G., *Equation of state for saturated liquids.* Journal of Chemical and Engineering Data, 1970. **15**(4): p. 514-517.

22. Tassios, D., *The number of roots in the NRTL and LEMF equations and the effect on their performance.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(1): p. 182-186.

23. Danner, R.P. and M.A. Gess, *A data base standard for the evaluation of vapor-liquid-equilibrium models.* Fluid Phase Equilibria, 1990. **56**(0): p. 285-301.

CHAPTER VI

GENERALIZED INTERACTION MODEL PARAMETER FOR MODIFIED NRTL ACTIVITY COEFFICIENT MODEL

6.1. Introduction

The activity coefficient is a deviation function that is used to account for non-ideal liquid behavior in a mixture. A number of activity coefficient models have been presented by several researchers in the literature [1-8]. Among the available models, the nonrandom two- liquid model (NRTL) [1] is used widely for designing chemical processes involving highly polar components.

In our previous work [9], we proposed a modification to the NRTL activity coefficient model. The modified model recast the original interaction parameters in such a way that the two new parameters reflect different characteristics, which are the energy interaction and energy interaction ratio parameters. The new formulation enabled us to reduce the number of interaction parameters from two to one by generalizing one of the parameter using pure-component properties. The modified model resulted in comparable representation of experimental phase equilibria properties to those of the original NRTL model.

The interaction parameter (g_{12}) in the modified model is determined by regressing experimental data. In this work, we generalize the interaction parameter of the modified NRTL model using a theory-framed quantitative structure-property relationship (QSPR) modeling approach. In this approach, the modified NRTL model is used as a theoretical framework to develop the behavior model, and QSPR is used to generalize the substance-specific parameter of the model.

The QSPR modeling technique has been employed to generalize successfully various theoretical frameworks for property predictions of pure components and mixtures [10-13]. In recent work [13], we implemented a theory-framed QSPR modeling approach to generalize the original NRTL model parameters for vapor–liquid equilibrium (VLE) binary systems. The model provided property prediction errors that were approximately two times the error of the data regression errors [13].

Our previous study [13] presented the challenge in generalizing highly-correlated parameters, as is the case of the two interaction parameters of the NRTL model. To reduce the effect of parameter correlation on the model reliability, a sequential regression approach was performed in the QSPR model development process. In this approach, one parameter is fixed at the QSPR generalized value while the other parameter was regressed. This procedure is performed multiple times until the effect of the parameter correlation on the model development was minimized.

In the current study, we applied our modified version of the NRTL model which has only one parameter. The advantage of having a single parameter is the avoidance of the sequential regression analysis technique applied previously in the model development process. This improvement leads to an internally consistent model (independent of the order of components) capable of predicting the interaction parameter *a priori*. Further, there is a significant reduction in the computational time required for developing the QSPR model.

The specific objectives of this work are as follows: (1) assemble a representative VLE database; (2) develop a QSPR model that can estimate the interaction parameter of the modified NRTL models *a priori*; (3) perform a rigorous validation of the model using an external test set; and (4) compare the model predictions with available activity coefficient models.

6.2. NRTL activity coefficient model

Renon and Prausnitz [1] developed the NRTL activity coefficient model based on the local composition theory of Wilson [6] and the two-liquid solution theory of Scott [14]. The model provides precise representation of highly non-ideal VLE and liquid–liquid equilibrium (LLE) systems [7]. The NRTL activity coefficient of a binary system is given as follows:

$$\ln \gamma_i = x_j^2 \left[\tau_{ji} \left(\frac{G_{ji}}{x_i + x_j G_{ji}} \right)^2 + \frac{\tau_{ij} G_{ij}}{\left(x_j + x_i G_{ij} \right)^2} \right] \tag{6.1}$$

where τ_{ij} and G_{ij} are defined as:

$$G_{ij} = \exp(-\alpha_{ij}\tau_{ij}) \qquad \tau_{ij} = \frac{g_{ij} - g_{jj}}{RT} = \frac{a_{ij}}{RT} \tag{6.2}$$

where g_{ij} is the energy interaction between i and j molecules, α is the non-randomness factor in the mixture, R is the universal gas constant in *cal K^{-1} mol^{-1}* and T is the mixture temperature in *K*.

The NRTL model has three adjustable parameters that are unique for a system. These parameters are a_{12} or ($g_{12} - g_{22}$), a_{21} or ($g_{21} - g_{11}$), and α_{12}. The parameters account simultaneously pure-component liquid interactions (g_{11} and g_{22}) and mixed-liquid interactions (g_{12} and g_{21}). The non-randomness factor (α_{12}) varies from 0.2 to 0.47 [7] and can often be set a *priori*. To be consistent with the DECHEMA database [15], the non-randomness factor was kept constant as 0.2 for all binary systems in this work.

6.3. One-parameter modified NRTL model (mNRTL1)

In our previous work [9], we proposed a modified version of the NRTL model which reduced the effect of parameter correlation in the original NRTL model. The modified model recasts the original NRTL equation as shown in the following equation:

$$a_{ij}=\left(g_{ij}-g_{jj}\right)=g_{ij}\left(1-\frac{g_{jj}}{g_{ij}}\right)=g_{ij}\left(1-r_{ij}\right)=g_{ij}R_{ij} \tag{6.3}$$

where g_{ij} is the cross interaction energy parameter and R_{ij} is the interaction ratio.

Employing the following combination rule (half harmonic mean) assumption:

$$g_{ij}=\left[a_{ji}a_{ij}\right]^{1/2}=\left[\left(g_{ij}-g_{ii}\right)\left(g_{ij}-g_{jj}\right)\right]^{1/2}=\left[\frac{g_{ii}g_{jj}}{g_{ii}+g_{jj}}\right] \tag{6.4}$$

Substituting Equation 6.4 into Equation 6.3, the interaction ratio (R_{ij}) becomes:

$$R_{ij}=\frac{1}{R_{ji}} \tag{6.5}$$

A generalization for the interaction ratio, R_{ij}, is introduced by using pure-fluid properties, which is shown in the following equation:

$$R_{ij}=\left[\left(\frac{\omega_j}{\omega_i}\right)\left(\frac{P_{Ci}}{P_{Cj}}\right)\right] \tag{6.6}$$

where ω is acentric factor, P_c is critical pressure and, the *i* and *j* subscripts indicate molecule *i* and *j*, respectively.

Equation 6.6 shows the ratio of pure-fluid properties in the modified one-parameter NRTL model (mNRTL1) that are used to determine the interaction ratio parameter (R_{ij}). This modification essentially reduces the NRTL model to a single parameter model (g_{ij}). In this work, we have generalized the g_{ij} parameter of the mNRTL1 model using a QSPR approach.

6.4. QSPR methodology

The QSPR methodology applied to generalize the interaction parameter of the mNRTL1 model involves several steps which includes the following: (1) database development, (2) parameter regression analyses for VLE systems using mNRTL1 model, (3) structure generation and optimization, (4) molecular descriptor generation and (5) descriptor reduction and QSPR model development using non-linear neural network models.

A schematic representation of the steps involved in developing the QSPR model is shown in Figure 6.1. Initially, a representative binary VLE database is assembled. Using the assembled data the interaction parameter of the mNRTL1 model is regressed to fit the VLE properties of the systems in the database. The following step is to generate the 2-dimensional (2D) structures of components in each binary system. The 2D structures are then optimized to find a 3-dimensional (3D) representation with the minimum conformation energy. The optimized 3D molecular structures are used to generate molecular descriptors using software such as DRAGON [16] and CODESSA [17]. The current DRAGON [16] software is capable of generating about 4,800 structural descriptors for each component. Next, the initial sets of descriptors are reduced through a process where the most significant descriptors for predicting the interaction parameter are identified. Simultaneously, these significant descriptors are used to develop a neural network model. The main steps of the model development process are described in greater detail below.

6.4.1. Database development

We have assembled a comprehensive VLE database from available sources by insuring sufficient representation of various functional groups in the database. The experimental VLE data were taken from DECHEMA [18] and NIST-TDE [19]. The pure-component vapor pressure data were collected from DIPPR [20] and DECHEMA [18].

A low-pressure binary VLE database (Oklahoma State University, OSU database I) [10] consisting of 188 binary VLE systems totaling 4716 data points was assembled. The second source of data was the DECHEMA [18] database, from which we collected 388 binary VLE systems totaling 12,010 data points. A third database consisting of 340 binary systems totaling over 17,000 data points was taken from NIST-TDE [19]. The data from the above three sources was compiled and a final database was created (Oklahoma State University, OSU-VLE Database III). The compiled data consists of 916 binary systems formed from various combinations of 140 different compounds totaling over 33,000 vapor-liquid equilibrium data points. The data covered a temperature range from 128 to 554 K and pressures to 58 bar; however, over 99% of the data were at pressure of less than 10 bar. In addition to pressure, temperature and mole fraction (PTXY) data, we have collected over 500 data points of infinite-dilution activity coefficient values (γ^{∞}) for 137 of the 916 VLE systems in the database [18].

The compounds present in the OSU-VLE Database III were classified in a similar manner as the UNIFAC functional-group classification approach [2]. The database is composed of compounds belonging to 31 chemical classes.

Figure 6.2 illustrates the data distribution of the binary systems in the OSU database III based on chemical classes. The figure provides the number of systems represented for each type of functional-group interaction. Due to the abundant data availability, systems containing alcohol or alkane components are highly represented in the database.

6.4.2. Interaction parameter regression

The interaction parameters of the NRTL and mNRTL1 models were regressed to correlate experimental binary VLE data. The regression analyses were performed by applying Gibbs equilibrium criteria for a closed system involving coexisting liquid and vapor phases, subject to mass balance constraints. We applied the split approach, as shown in Equation 6.7, in the phase equilibria calculations.

$$\hat{\phi}_i^V P y_i = \gamma_i P_i^\circ \phi_i^V x_i \lambda_i ; \qquad n = 1, n \tag{6.7}$$

where n is the number of components, the subscript i represents a particular component, $\hat{\emptyset}^V$ is the component fugacity coefficient in the vapor phase, P is the mixture pressure, y is the vapor mole fraction, γ is the component activity coefficient in the liquid phase, P° is the pure-component vapor pressure, $\emptyset^V$ is the pure-component fugacity coefficient in the vapor phase, x is the liquid mole fraction and λ is the Poynting factor. The VLE systems considered in this study were generally at low pressure; hence, the vapor-phase fugacity coefficients were assumed to be 1. We have also investigated the quality of representation when equation-of-state (EOS) models are used to calculate the vapor-phase fugacity coefficients. Our findings show there is no improvement on the overall representation error, which substantiates our assumption (data not shown).

The Poynting factor is expressed given as:

$$\lambda_i = \exp\left(\frac{v_i^L (P - P_i^\circ)}{RT} \right) \tag{6.8}$$

where v^L is the liquid molar volume and is determined using the Rackett equation [21].

The parameter regression analyses was performed by employing the objective function, *OF*, which is the weighted sum of squares of the relative errors in pressure, k-values, infinite-dilution activity coefficients and the weighted absolute sum of model parameters, as shown in Equation 6.9.

$$\begin{aligned} OF = \sum_{i=1}^{n} w_1 \left(\frac{P^{Exp} - P^{Calc}}{P^{Exp}} \right)_i^2 + w_2 \sum_{i=1}^{n} \left(\frac{K_{values}^{Exp} - K_{values}^{Calc}}{K_{values}^{Exp}} \right)_i^2 \\ + w_3 \sum_{i=1}^{n} \left(\frac{\gamma_{values}^{\infty\ Exp} - \gamma_{values}^{\infty\ Calc}}{\gamma_{values}^{\infty\ Exp}} \right)_i^2 + w_4 (Par) \end{aligned} \tag{6.9}$$

where the weights were: $w_1 = 1;\ w_2 = 1/15$; $w_3 = 1/10;\ w_4 = 2E - 6$; *n* is the number of data points, *Par* is $|a_{12}|+|a_{21}|$ for the NRTL model and $|g_{12}|$ for the mNRTL1 model, the superscripts *Exp* and *Calc* refer to experimental and calculated values, respectively.

Various objective function formulations were tested to determine the most suitable objective functions. Equation 6.9 was selected since the equation provided a balance of the model prediction errors for temperature, pressure, equilibrium constants, activity coefficient and vapor mole fraction and reduced correlation of the model parameters (a_{12} and a_{21}) [22].

6.4.3. Descriptor calculation

ChemBioDraw Ultra 11.0 [23] software was used to generate 2D and 3D structures of the molecules. Open Babel software [24] was then used to optimize the 3D structures by minimizing the conformation energy of the molecules. The structure optimization was performed using a genetic algorithm (GA) based conformer search [24, 25], which employs the MMFF94 force field [26]. The optimized molecules are then used to generate 2344 DRAGON [27] and 598 CODESSA [17] 0D, 1D, 2D, and 3D descriptors.

6.4.4. Descriptor input

The structural descriptors from DRAGON [27] and CODESSA [17] were used as input values in the development of the QSPR model. The input descriptor set for each binary system is prepared by calculating the absolute differences of all the individual descriptors of the compounds in the binary system. This novel approach forces the QSPR model to satisfy the pure limit behavior of activity coefficient properties. For a hypothetical mixture of X and Y where X and Y are the same molecule, the activity coefficient values are ones; i.e., the interaction parameter are zeros. For such hypothetical systems, the QSPR input values (descriptor differences) are zeros. Hence, the QSPR model will provide prediction values that satisfy the limiting behavior or zero interaction parameters.

6.4.5. Descriptor reduction and model development

The descriptor reduction applied in this study is a hybrid approach where descriptor reduction and model development happen in parallel. The hybrid approach uses evolutionary programming (EP) and differential evolution (DE) as a wrapper around artificial neural networks (ANNs) to identify the best descriptor subsets from the initial molecular descriptors pool. A detailed discussion on this approach can be found in our previous works [13, 28, 29].

The initial step in the model development process is to divide the entire data set into four sub-sets (training, validation, internal test and external test sets) with a proportion of 50% for the training set, 15% for the internal validation set, 10% for the internal test set and the remaining 25% for the external test set. The data division was performed by ensuring that there is adequate representation of all the functional-group interactions in all the data sets. For example, there are 19 systems with chloroalkene/alcohol interactions in the database. The data division for this type of interactions will be 10, 3, 2 and 4 of the systems assigned to the training, validation, internal test and external test sets, respectively. For interactions with a small number of systems, data allocation is prioritized to the training followed by validation and internal test sets.

All data excluding the external test set were used in the descriptor reduction and model development process. To avoid over-fitting, the validation set data was used by applying an early-stopping method [27, 30]. In addition, the internal test data was used to identify the best ANNs during the descriptor reduction algorithm. In the model development, the external test set data was set aside and was only used to evaluate the generalization (*a priori* prediction) capability of the developed model.

6.4.6. Modeling scenarios

In this study, six case studies were performed to assess the generalization capability of the QSPR model and compare the results with literature models. In all case studies, the ideal gas (IG) model was applied to describe the gas phase behavior. The six case studies are outlined as follows:

Ideal Solution: The ideal solution model was used to predict the phase-equilibrium behavior.

NRTL-Regressed: The NRTL model with regressed a_{12} and a_{21} parameters was used to represent VLE properties.

mNRTL1-Regressed: The mNRTL1 model with a regressed g_{12} parameter was used to represent VLE properties.

mNRTL1-QSPR: The generalized QSPR model was used to provide the mNRTL1 model parameter, and then the mNRTL1 model was used to predict the activity coefficients

UNIFAC-2006: The UNIFAC model was used to predict the activity coefficients of each component. The UNIFAC interaction parameters reported by Gmehling et al. [31] were used in this case study.

The NRTL-Regressed and mNRTL1-Regressed studies were conducted to evaluate the correlative capabilities of the NRTL and mNRTL1 models, respectively. The Ideal Solution, mNRTL1-QSPR and UNIFAC-2006 case studies were focused on assessing the *a priori* predictive capabilities of the ideal solution, the generalized modified NRTL model and the UNIFAC model, respectively.

The representation and prediction capabilities of the models were assessed for equilibrium properties such as pressure (P), activity coefficients (γ^{∞}), temperature (T), vapor mole fraction (y_1)

and equilibrium K-values (K_1 and K_2). In the NRTL-Regressed study, the two model parameters, a_{12} and a_{21}, shown in Equation 6.2, were regressed. In the mNRTL1-Regressed study, the g_{12} model parameter, shown in Equation 6.3, was regressed. Bubble-point pressure calculations were performed in the regression analyses. The regressed or QSPR predicted parameters are used directly to calculate (a) *P*, y_1, γ^{∞} and *K-values* for known T and x_1 and (b) T for known P and x_1 .

6.5. Results and discussion

This work focused on assessing (a) model representation of equilibrium properties, (b) QSPR generalized predictions, (c) limiting-behavior property predictions and (d) multicomponent phase behavior predictions. The results for each of these objectives are discussed in the following sections.

6.5.1. Representation assessment

Experimental *P*, *T*, *x* and *y* data of 916 binary systems were used to evaluate the correlative capabilities of the NRTL and mNRTL1 models. The representation capabilities of the models were analyzed by calculating the root-mean-squared error (RMSE), bias and percentage absolute average deviation %AAD.

Table 6.1 provides the property prediction errors for the ideal solution, NRTL-Regressed and mNRTL1-Regressed case studies. As expected, the ideal solution model resulted in poor predictions compared to the activity coefficient models. When activity coefficient models are used, the error was reduced by about four fold compared to the ideal solution model. The NRTL model with regressed parameters provided overall representation %AADs of 2.1, 0.2, 4.3 and 5.5 for *P, T*, y_1 and *K-values*, respectively. The mNRTL1 model with a regressed parameter resulted in slightly higher overall %AADs of 2.5, 0.2, 4.7, 6.1 and 13.3 for *P, T*, y_1 and *K-values*, respectively.

Figure 6.3 shows the distribution of pressure errors for the NRTL and mNRTL1 models by functional-group interactions. Results of each functional-group interaction is shaded in variations of grey colors based on the %AAD ranges given in the figure key. In general, the two models showed comparable representation capabilities. As indicated by the matrix, both models provided accurate representation when the components of the systems have the same functional groups (diagonal elements of the triangular matrix). The interaction between molecules from the same functional groups is nearly-ideal. Thus, the activity coefficient models represent such systems without difficulty. Both models resulted in relatively high errors for most of the aqueous systems. These higher representation errors could be attributed to the high level of experimental uncertainties associated with water systems, and the inability of the models in representing such systems precisely. Further, the mole fraction of aqueous systems tend to be very small which results in large percentage error.

6.5.2. QSPR generalized predictions

The mNRTL1-Regressed study established the benchmark for the best achievable level of prediction errors for QSPR generalization. The regressed model parameter (g_{12}) from this case study was used as a target when developing the QSPR model.

One of the key tasks in QSPR modeling is determining the number of descriptors that can provide the functional flexibility required to predict accurately target values. For this purpose, we performed a sensitivity analysis to analyze the effect of variation in the number of descriptors on the property predictions. Figure 6.4 shows the quality of pressure and γ^{∞} predictions from QSPR models developed using 10, 15, 20 and 30 descriptors. The result shows the prediction of pressure improved modestly from 5.8 to 4.5 %AAD when the number of descriptors was increased from 10 to 30. The γ^{∞} predictions improved from 26 to 23 %AAD when using 10 and 15 descriptors, respectively. The improvements, however, were not significant when the descriptors are increased from 15 to 30. In general, the QSPR model with 15 descriptors provided comparable property predictions compared to the results found using a model with 30 descriptors; hence, the QSPR model with 15 descriptors was selected in this work due to simplicity and accurate prediction of the properties.

Table 6.2 provides the list of the 15 molecular descriptors used as inputs in developing the QSPR model for predicting the mNRTL1 model parameter. DR and CO represent molecular descriptors calculated using DRAGON [16] and CODESSA [17], respectively. The list reveals constitutional indices, electrostatic, quantum chemical and molecular properties are significant in predicting the interaction parameter. Based on the individual R^2 value, the most important specific descriptors were related to polarity and LogP (octanol-water partition coefficient). Polarity signifies the distribution of the electrons (charge) which plays a significant role on how molecules interact with each other. LogP represents the distribution of molecules in aqueous and organic phases and it provides insight on hydrophilic and hydrophobic interactions of molecules of various types interacting in the presence of organic and aqueous phases at equilibrium. Similar significant descriptors were found in our previously developed QSPR generalized models for the NRTL, UNIQUAC and Wilson models [13].

Figures 6.5 and 6.6 show comparisons of the regressed and predicted g_{12} values for the training and validation sets, respectively. The correlation coefficients (R^2) between the regressed and predicted values for the training and the validation sets are 0.96 and 0.91, respectively. The figures indicate good agreement between the regressed and QSPR predicted parameters. As such, the QSPR model resulted in comparable predictions for the training and validation sets, which indicates the model was trained without over-fitting. Similarly, Figure 6.7 shows the comparison of the regressed and predicted g_{12} values for the external test set. The R^2 value between the regressed and predicted values for the external test set is 0.85, indicating good generalized predictions by the QSPR model.

Table 6.3 provides the VLE property prediction errors obtained using the QSPR predicted parameters from the mNRTL1-QSPR study. The results are classified into training, validation, internal test and external test sets. In addition to providing results for all systems, the table also provides the results for water containing and highly non-ideal systems. The VLE property predictions for the QSPR model were approximately twice the regression analyses %AAD values for all categories including water containing and highly non-ideal systems. Further, the external and internal test set predictions were comparable to the overall prediction, which demonstrates the capability of the model for generalized *a priori* predictions.

Figure 6.8 shows the distribution of pressure errors for the mNRTL1-QSPR model by functional-group interaction. As shown in the figure, the QSPR model resulted in prediction of pressure within 6 %AAD for most of the functional-group interactions present in the database. The matrix also indicates the model provided %AADs between 10% and 20% for most of the water systems.

Figures 6.9a, 6.9b, 6.9c and 9d illustrate the QSPR predicted equilibrium phase compositions of n-heptane-ethylbenzene, propionic aldehyde-acetone, benzene-ethanol and tetrachloromethane-furfural systems, respectively. The figures indicate the prediction of mNRTL1-QSPR and representation of the NRTL and mNRTL1 models. For all the examples, the predictions from the QSPR model are in a good agreement with the experimental composition values. This demonstrates the capabilities of the QSPR model for predicting various type of phase behaviors, including nearly-ideal and highly-non ideal systems.

The modified UNIFAC model [31] was used to compare the generalization capability of the QSPR model. Table 6.4 shows the results of the mNRTL1-QSPR and UNIFAC-2006 case studies. The QSPR model provided comparable predictions to that of the UNIFAC model for 853 VLE systems. When the UNIFAC model is used for systems with at least one missing interaction parameter, the prediction errors increased significantly. This indicates the limitations of the UNIFAC model for generalized predictions when the interaction parameters are missing. In contrast, the QSPR model was able to provide predictions for a wider range of functional groups. Further, the QSPR model uses about 300 model parameters (neural network weights and biases) which is significantly lower than the UNIFAC model [31] which has over 4,000 parameters.

6.5.3. Limiting-behavior prediction assessment

Table 6.5 shows the representation and prediction of infinite-dilution activity coefficients for 137 binary systems using NRTL-Regressed, mNRTL1-Regressed, mNRTL1-QSPR and UNIFAC-2006. The NRTL-regressed and mNRTL1-Regressed models provided overall representation %AADs of 8.7 and 13.3 for the γ^{∞} property, respectively. The generalized mNRTL1-QSPR model provided γ^{∞} predictions within twice the error found in the regression analyses. The UNIFAC-2006 model resulted in relatively lower error compared to the QSPR model.

6.5.4. Multicomponent phase behavior predictions

The prediction capabilities of the NRTL-Regressed, mNRTL1-Regressed, mNRTL1-QSPR and UNIFAC-2006 models was evaluated for multicomponent systems. For the NRTL-Regressed and mNRTL1-Regressed models, the interaction parameters obtained from regression of binary VLE experimental data were used for ternary VLE property predictions.

Table 6.6 shows the prediction of ternary properties using the various models for 57 ternary systems. The UNIFAC-2006 model resulted in relatively lower errors compared to the three models. This is due to the fact that the ternary data used in this study were used to regress the interaction parameters of the UNIFAC model. As a result, the UNIFAC model performs better on these systems. The NRTL-Regressed model resulted in slightly lower errors compared to the mNRTL1-Regressed results. Further, the mNRTL1-QSPR provided comparable predictions to the mNRTL1-Regressed results which indicates that the QSPR model can be extended to multiphase property predictions without a great loss of accuracy.

6.6. Conclusion

In this study, a QSPR modeling approach was applied to generalize the interaction parameter of a modified one-parameter NRTL model. A VLE database consisting of 916 binary VLE system from combinations of 140 compounds was assembled. Structural descriptors of molecules were used as inputs in the QSPR model. The limiting behavior of mixtures were taken into consideration while developing the QSPR model. The predictive capabilities of the generalized model were assessed for phase equilibria properties including pressure, temperature, vapor mole fractions, equilibrium constants and infinite-dilution activity coefficients. The QSPR generalized model provided property predictions within twice the overall errors found in the experimental data regression analyses. The results using the QSPR model were comparable to that of the UNIFAC group-contribution model. Thus, our methodology provides a potential alternative approach for generalization of activity coefficient models. Future studies should focus on extending the methodology applied in this study to LLE systems.

Table 6.1. VLE property predictions of the Ideal Solution model and representations of the NRTL and mNRTL1 models

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
Ideal Solution	None	P (bar)	916	33283	0.68	-0.13	13.5	97
		T (K)	916	33283	9.29	4.15	1.5	28
		y_1	675	18199	0.10	-0.01	15.3	100
		K-values	675	18199	6.79	-0.82	19.2	100
NRTL-Regressed	a_{12} & a_{21}	P (bar)	916	33841	0.15	0.00	2.1	14
		T (K)	916	33841	1.35	0.10	0.2	1
		y_1	675	18199	0.03	0.00	4.3	48
		K-values	675	18199	5.09	-0.31	5.5	54
mNRTL1-Regressed	g_{12}	P (bar)	916	33845	0.24	-0.01	2.5	26
		T (K)	916	33845	1.67	0.16	0.2	2
		y_1	675	18199	0.03	0.00	4.7	51
		K-values	675	18199	5.41	-0.21	6.1	64

Table 6.2. The descriptors used as inputs for the ANNs in the final ensemble for estimating the mNRTL1 model parameter

No	Descriptor name	Descriptor description	Source	Type of descriptor	R²
1	BLTF96	Verhaar Fish base-line toxicity from MLOGP (mmol/l)	DR	Molecular properties	0.35
2	HASA-2/SQRT(TMSA) [Zefirov's PC]	HASA-2/SQRT(TMSA) [Zefirov's PC]	CO	Electrostatic	0.23
3	Polarity parameter / square distance	polarity parameter / square distance	CO	Electrostatic	0.22
4	AAC	mean information index on atomic composition	DR	Information indices	0.15
5	SM3_Dz(p)	spectral moment of order 3 from Barysz matrix weighted by polarizability	DR	2D matrix-based descriptors	0.12
6	HOMO - LUMO energy gap	HOMO - LUMO energy gap	CO	Quantum Chemical	0.12
7	GATS1e	Geary autocorrelation of lag 1 weighted by Sanderson electronegativity	DR	2D autocorrelations	0.11
8	MLOGP2	squared Moriguchi octanol-water partition coeff. (logP^2)	DR	Molecular properties	0.09
9	HOMO energy	HOMO energy	CO	Quantum Chemical	0.04
10	Min e-e repulsion for a C atom	Min e-e repulsion for a C atom	CO	Quantum Chemical	0.03
11	Mor11m	signal 11 / weighted by mass	DR	3D-MoRSE descriptors	0.03
12	nCsp2	number of sp2 hybridized Carbon atoms	DR	Constitutional indices	0.02
13	WiA_B(p)	average Wiener-like index from Burden matrix weighted by polarizability	DR	2D matrix-based descriptors	0.02
14	P_VSA_LogP_4	P_VSA-like on LogP, bin 4	DR	P_VSA-like descriptors	0.01
15	IAC	total information index on atomic composition	DR	Information indices	0.00

Table 6.3. Predictions from the mNRTL1-QSPR case study

Data set	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD	%AAD multiplier
Training set	P (bar)	460	20298	0.17	0.01	4.2	41	1.9
	T (K)	460	20298	2.32	0.11	0.4	5	1.8
	y_1	339	10187	0.04	0.00	5.8	58	1.6
	K-values	339	10187	5.98	-0.18	7.4	63	1.3
Validation set	P (bar)	167	5101	0.42	-0.02	5.7	38	2.2
	T (K)	167	5101	2.70	0.19	0.5	2	2.1
	y_1	117	2910	0.04	0.00	7.1	39	1.6
	K-values	117	2910	7.48	-0.86	8.5	44	1.4
Internal test set	P (bar)	101	2702	0.17	-0.01	6.3	33	2.3
	T (K)	101	2702	3.87	0.62	0.6	4	2.1
	y_1	77	1475	0.05	0.00	7.7	34	1.4
	K-values	77	1475	5.96	-0.66	9.9	100	1.3
External test set	P (bar)	188	5741	0.36	-0.01	5.5	43	2.6
	T (K)	188	5741	3.03	0.23	0.5	4	2.4
	y_1	142	3627	0.04	0.00	7.2	55	1.7
	K-values	142	3627	1.83	-0.12	8.7	55	1.6
Highly non-ideal	P (bar)	348	14926	0.41	-0.03	6.4	41	1.9
	T (K)	348	14926	3.37	0.65	0.6	5	1.8
	y_1	262	8203	0.05	0.00	8.4	58	1.6
	K-values	262	8203	9.39	-0.86	10.1	79	1.4
Water systems	P (bar)	55	4344	0.74	-0.11	11.5	30	1.6
	T (K)	55	4344	5.94	1.45	1.0	5	1.4
	y_1	47	2313	0.09	0.00	17.9	58	1.4
	K-values	47	2313	20.35	-4.22	19.1	79	1.1
All data	P (bar)	916	33842	0.28	0.00	5.0	43	2.1
	T (K)	916	33842	2.75	0.20	0.4	5	2.0
	y_1	675	18199	0.04	0.00	6.6	58	1.6
	K-values	675	18199	5.87	-0.34	8.1	100	1.4

Table 6.4. Comparison of a priori predictions of the mNRTL1-QSPR and UNIFAC-2006 case studies

Model	Parameters	Property	No. of sys.	No. of pts.	RMSE	Bias	%AAD	Max %AAD
mNRTL1-QSPR	Generalized g_{12}	P (bar)	916	33842	0.25	0.00	5.0	43
		T (K)	916	33842	2.41	0.19	0.4	5
		y_1	675	18199	0.04	0.00	6.6	58
		K-values	675	18199	5.44	-0.39	8.1	100
UNIFAC-06	UNIFAC - All interactions present	P (bar)	853[a]	31609	0.51	0.00	5.1	100
		T (K)	853	31609	4.74	-0.06	0.4	25
		y_1	634	17056	0.04	0.00	5.6	100
		K-values	634	17056	6.03	0.11	6.9	100
UNIFAC-06	UNIFAC - One or more missing interactions	P (bar)	46	1308	0.35	-0.05	11.1	71
		T (K)	46	1308	8.56	1.49	1.1	13
		y_1	31	940	0.07	0.00	13.2	50
		K-values	31	940	1.31	-0.09	15.0	100

[a] Due to a lack of group interaction parameters, 63 systems of the 916 systems were not considered.

Table 6.5. Infinite-dilution activity coefficient representation and prediction of various activity coefficient models

Property	Model	No. of sys.	No. of pts.	RMSE	Bias	%AAD
Infinite-dilution activity coefficient ($\gamma\infty$)	NRTL-Regressed	137	549	3.54	-0.21	8.7
	mNRTL1-Regressed			6.46	-0.77	13.3
	mNRTL1-QSPR	137	549	7.84	-1.60	22.6
	UNIFAC-06			3.30	0.33	12.2

Table 6.6. Prediction results of 57 ternary VLE systems using NRTL-Regressed, mNRTL1-Regressed, mNRTL1-QSPR and UNIFAC-2006 models

Study	Parameters	No. of sys.	Property	No. of pts.	RMSE	Bias	%AAD
NRTL-Regressed	a_{12} & a_{21}	57	P (bar)	2212	0.05	0.00	3.0
			T (K)	2212	1.57	0.16	0.3
			y_1	1890	0.04	0.00	8.7
			K-values	1890	0.43	0.01	8.3
mNRTL1-Regressed	g_{12}	57	P (bar)	2212	0.05	0.01	3.8
			T (K)	2212	1.71	-0.31	0.3
			y_1	1890	0.04	0.00	9.9
			K-values	1890	0.38	0.02	9.4
mNRTL1-QSPR	Generalized g_{12}	57	P (bar)	2212	0.05	-0.01	4.0
			T (K)	2212	1.81	0.38	0.3
			y_1	1890	0.04	0.00	9.0
			K-values	1890	0.38	-0.01	9.0
UNIFAC-06	UNIFAC-2006	57	P (bar)	2212	0.04	0.00	2.5
			T (K)	2212	1.49	0.19	0.2
			y_1	1890	0.04	0.00	7.8
			K-values	1890	0.36	0.00	7.4

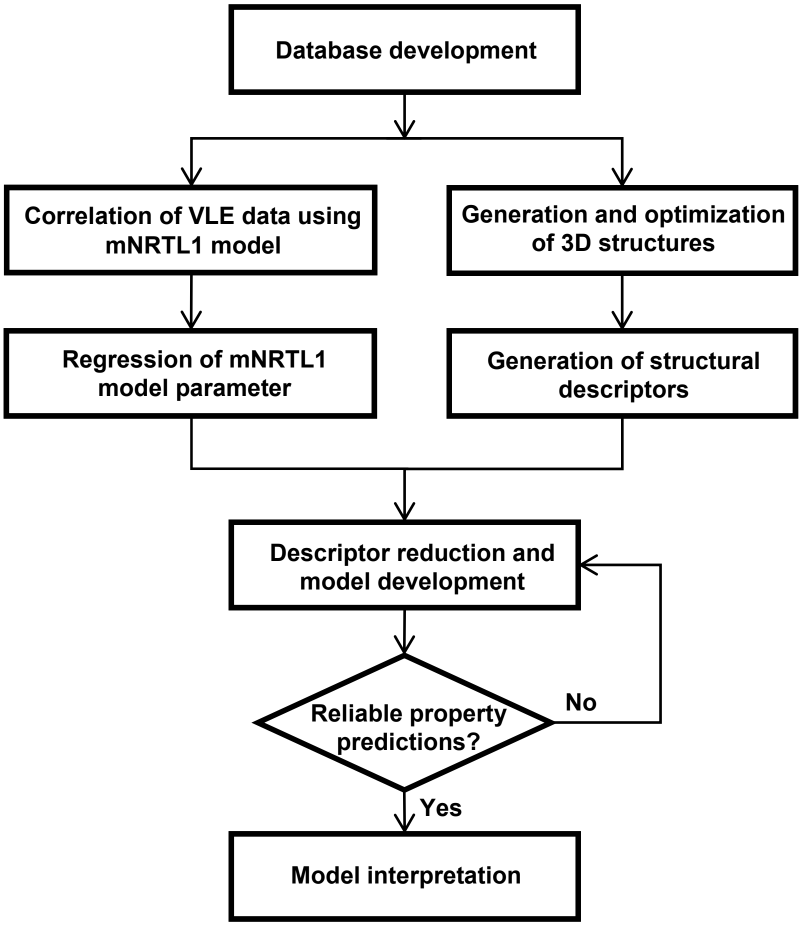

Database development
Correlation of VLE data using mNRTL1 model
Generation and optimization of 3D structures
Regression of mNRTL1 model parameter
Generation of structural descriptors
Descriptor reduction and model development
Reliable property predictions?
No
Yes
Model interpretation

Figure 6.1. Schematic of the QSPR model development process

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	13																														
2	Aldehyde	10	1																													
3	Alkane	24	5	14																												
4	Alkene	9	1	10	3																											
5	Alkyne	5	3	5	6	2																										
6	Amide	6	2	6	2	1																										
7	Amine	5		4			3	4																								
8	Aromatic Bromo	5		3			1																									
9	Aromatic Floro	2		2			1	1	1																							
10	Benzene Derivative	6	3	13	5		1	5	1	3	4																					
11	Bromoalkane	15		5					1	1	8																					
12	Carboxylate	2	5	9	1						6	1	3																			
13	Chloroalkane	5		5	2	2	4	6		2	8	3	4	2																		
14	Chloroalkene	19	1	7		1	1				1		1	8	1																	
15	Chlorobenzene	9		2	2		1	4	1	1	2	1		2	1																	
16	Epoxide	7	3	6							1		2	4																		
17	Ester	1	1	8	1	1	1	1			4	1	1	5	1	1	1															
18	Ether	12	2	21	3	3	2	2		3	5	2	1	9	2	2	1	3	3													
19	Furfural	1		3	1						2			4	1			1														
20	H2S	1		1																												
21	Iodoalkane	3	1	1							2	1	1	4				1	1													
22	Ketone	3	4	21	3	1	2	5	1		8	1	6	8	7	3	1	3	2	2		1	4									
23	Nitrile	4		4	2	2	1	1	1		4		4	6	3	1		1	1				1	1								
24	Nitrite	1													1									1								
25	Nitro Compound	12		3	2	2	1		1		5	1	2	5	1	2		3	3			2	3	2		2						
26	Pyridine Derivative	14		4			1			1	1	1	1	2	1	1		2	1				1	1			1					
27	Sulfide	4		4	3	3	1	1	1		1	2	2	5	2	1		1	1		1	1	1	1		2						
28	Thiol	1		7			2	1			1								1		1		1	1				3				
29	Thiophene	4		1	2		1				1			1					1							1	1					
30	Toluene Derivative	3	6	4	2		1	2	1	1	2	3	2	3	1	1	1		5	1		1	5	1		2	2	2	2	1	1	
31	Water	9	1	2			1	10			3	1	3		1		2	1	4	1			5	3	1	2	3	1	1			

X

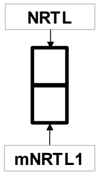
Y #

Number of available binary systems consisting of chemicals with functional groups of X and Y

□ No VLE data used

Figure 6.2. Database matrix of the compounds in the OSU-VLE database III

Key	
Color	Pressure %AAD Range
#	%AAD<3
#	3<%AAD<6
#	6<%AAD<10
#	10<%AAD<20



		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	Alcohol	2 2																														
2	Aldehyde	3 4	1 1																													
3	Alkane	4 5	1 1	1 1																												
4	Alkene	3 4	2 2	1 1	1 1																											
5	Alkyne	3 4	1 1	2 2	0 1	1 2																										
6	Amide	2 2	3 3	4 4	2 5	2 4																										
7	Amine	2 2		3 3			2 3	1 1																								
8	Aromatic Bromo	2 3		1 1			6 6																									
9	Aromatic Floro	3 5		0 0			3 3	0 0	1 1																							
10	Benzene Derivative	2 2	2 2	1 1	2 2		3 3	2 1	2 2	2 2	1 1																					
11	Bromoalkane	2 3		2 4					0 0	1 1	3 3																					
12	Carboxylate	2 2	1 1	2 3	0 1						2 2	3 4	1 1																			
13	Chloroalkane	1 3		1 1	1 1	0 0	2 5	2 2		1 1	2 2	2 3	2 2	1 1																		
14	Chloroalkene	2 3	1 1	1 1		1 1	2 2				1 1		2 2	2 2	1 1																	
15	Chlorobenzene	3 3		1 1	2 2		3 3	1 1	0 0	0 0	0 1	1 1		1 1	2 2																	
16	Epoxide	2 2	2 2	1 1							2 2		1 1	4 4																		
17	Ester	6 6	1 1	1 2	0 1	0 0	5 5	2 2			2 2	1 1	3 3	1 1	1 1	1 1	5 5															
18	Ether	2 2	2 2	2 2	2 2	2 2	2 4	4 4		0 0	1 1	5 6	1 1	1 1	2 2	1 1	1 1	1 1	2 2													
19	Furfural	3 3		6 5	4 7						3 3			3 3	1 2			1 1														
20	H2S	4 4		3 7																												
21	Iodoalkane	5 5	2 2	1 1							1 1	1 1	1 1	1 1				0 0	1 1													
22	Ketone	1 1	1 1	2 2	1 2	1 1	2 2	1 2	1 1		2 2	1 1	2 2	1 2	2 2	2 2	1 1	1 1	0 1	1 1		2 2	1 1									
23	Nitrile	0 0		3 5	2 3	4 3	2 2	1 2	1 4		2 3		1 1	2 2	2 2	2 2		0 1	1 1				1 1	2 2								
24	Nitrite	2 2													3 3									4 4								
25	Nitro Compound	4 4		5 4	2 5	1 1	2 2		7 7		2 3	2 2	1 1	2 3	2 2	1 1		1 1	2 2			2 2	2 2	1 1		2 2						
26	Pyridine Derivative	2 2		1 1			0 0			1 2	2 2	8 8	0 0	4 5	1 1	0 1		2 2	2 2				1 1	0 1			1 1					
27	Sulfide	2 3		1 1	1 1	3 3	1 1	5 5	2 4		2 2	2 3	1 1	3 5	1 1	2 3		2 4	2 2		1 4	1 1	2 2	2 8		5 9						
28	Thiol	6 8		4 4			1 2	1 1			0 0								0 1		5 5		1 1	1 1				2 1				
29	Thiophene	2 2		1 2	2 2		4 4				0 0			0 0					1 1							0 0	2 2					
30	Toluene Derivative	3 3	3 3	1 1	1 1		6 6	2 2	0 0	1 1	0 0	2 2	1 1	2 2	1 1	1 1	3 3		2 2	2 3		1 1	1 1	2 2		1 1	1 1	3 4	4 4	2 2	5 5	
31	Water	2 3	0 6	8 18			2 2	5 11			9 14	1 1	6 7		7 15		7 12	7 4	8 9	8 8			6 6	4 5	3 5	5 7	2 5	4 10	3 3			

Figure 6.3. Pressure representation of the regressed NRTL and mNRTL1 models by type of interaction

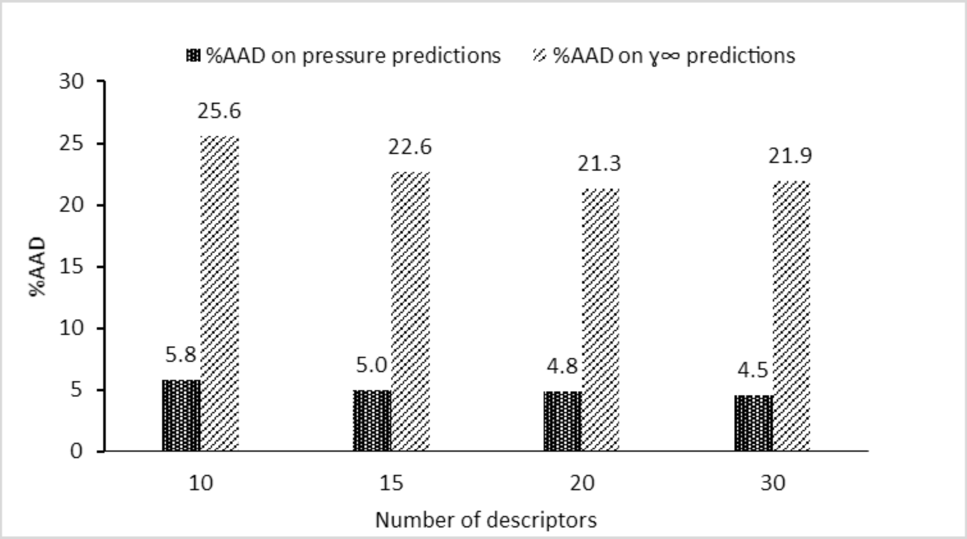


Figure 6.4. Effect of variation in the number descriptors of the prediction of pressure and γ^{∞} values using the mNRTL1-QSPR model

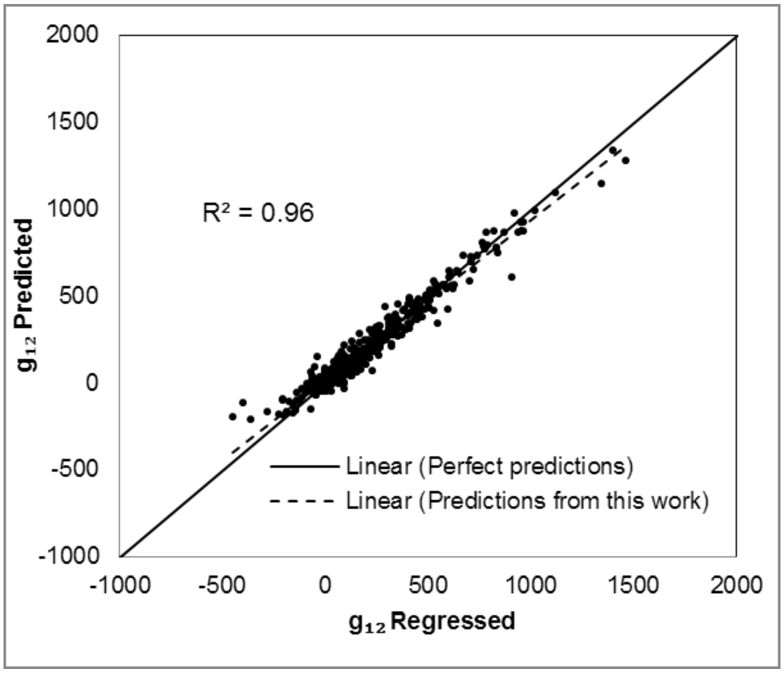


Figure 6.5. Comparison of the regressed and QSPR predicted g_{12} values in the training set.

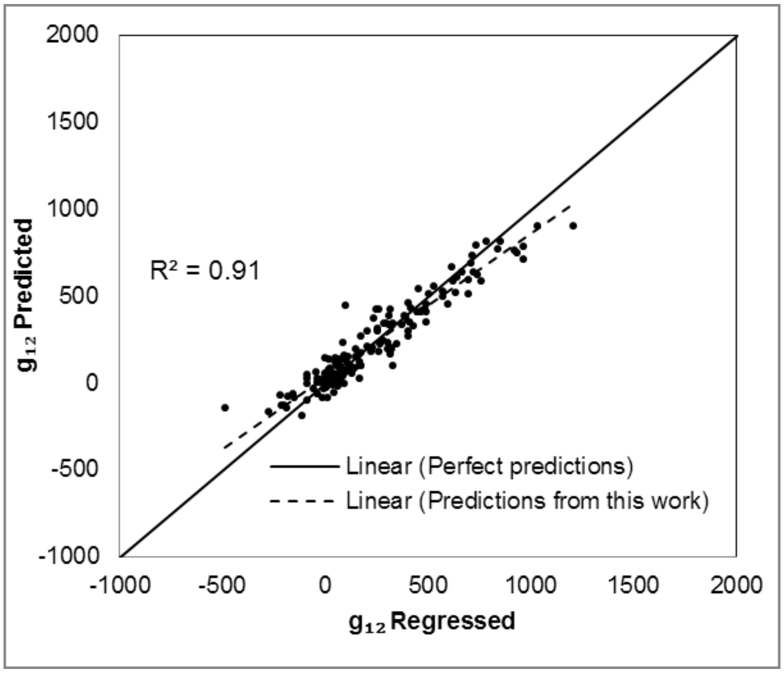


Figure 6.6. Comparison of the regressed and QSPR predicted g_{12} values in the validation set

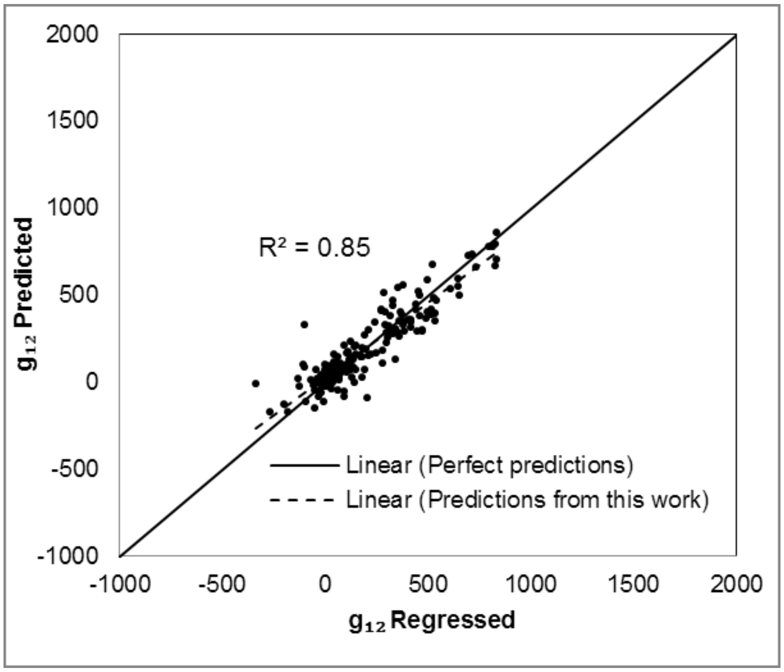


Figure 6.7. Comparison of the regressed and QSPR predicted g_{12} values in the external test set

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
1	**Alcohol**	4																														
2	**Aldehyde**	6.9	2																													
3	**Alkane**	4.6	2	2.6																												
4	**Alkene**	8.1	4	1.1	2																											
5	**Alkyne**	4.1	4	3.9	1	5.9																										
6	**Amide**	5.2	4	4.8	6	8.6																										
7	**Amine**	11		5.7			8	5.9																								
8	**Aromatic Bromo**	3.7		4.3			6																									
9	**Aromatic Floro**	8.7		3.5			6	4.7	1																							
10	**Benzene Derivative**	3.5	2	2.5	5		5	1.9	4	3	0.7																					
11	**Bromoalkane**	7.6		12					0	2	4																					
12	**Carboxylate**	3.3	1	4.2	1						3.6	5.3	2.7																			
13	**Chloroalkane**	3.5		4.5	2	6	9	4.4		2	2.9	7.1	7.4	1																		
14	**Chloroalkene**	5.8	4	8.6		7.5	2				1.7		4.7	4.9	3.1																	
15	**Chlorobenzene**	4.3		2.3	4		4	1.8	0	2	1.7	1.8		2.6	6																	
16	**Epoxide**	7.5	3	1.9							4.4		2.4	6.9																		
17	**Ester**	0.5	2	5.2	3	1.6	5	4.9			1.9	1.7	6	2.8	2.6	2.4	16															
18	**Ether**	4	3	2.9	5	3.9	5	4.1		3	3	10	2.4	3.6	3.3	2.3	6.5	2.1	2.5													
19	**Furfural**	7.9		9.6	7						2.8			6.1	2.4			1														
20	**H2S**	11		18																												
21	**Iodoalkane**	7.4	2	1.5							4.8	0.8	1.2	2.3				1.3	1.7													
22	**Ketone**	1.1	1	2.4	4	1.7	5	7.6	2		2.2	1.1	2	3.2	5.9	3	1.5	2.1	3.9	1.7		3.4	1.4									
23	**Nitrile**	2.9		4.4	6	9.7	4	4.4	5		5.2		3.2	3.2	5.4	6		1	1				1.4	2.4								
24	**Nitrite**	3													6.6									6.7								
25	**Nitro Compound**	4.7		8.9	5	1.3	2		7		3.6	1.3	2.7	7.3	4.4	3.2		2.1	1.6			5.3	4.4	4.2		4.6						
26	**Pyridine Derivative**	10		6.9			2			3	2.2	8.4	1.7	18	3.3	1		2.7	2				1.4	2.5			0.9					
27	**Sulfide**	8.8		7	4	10	1	13	5		1.8	6.4	3.8	5.7	3.6	4.2		9.9	2.1		12	1.6	3.4	6		11						
28	**Thiol**	9		8.4			6	1.9			1.6								0.6		5.3		0.9	0.8				5.4				
29	**Thiophene**	6		1.6	3		4				1.9			1.4					1.7							2.3	5					
30	**Toluene Derivative**	4.2	7	1.3	3		8	3.8	2	4	0.4	2.3	2.3	1.6	1.2	2	7.3		3.4	3.7		1.2	1.4	0.9		1	1.5	5.2	4.1	2.5	7.4	
31	**Water**	6.2	6	20			7	16			15	1.8	6.6		16		17	3.9	14	8			9.7	13	10	13	6	20	30			

Key	
Color	**Pressure %AAD Range**
#	%AAD<3
#	3<%AAD<6
#	6<%AAD<10
#	10<%AAD<20
#	%AAD>20

Figure 6.8. Pressure predictions of the mNRTL1-QSPR by type of interactions

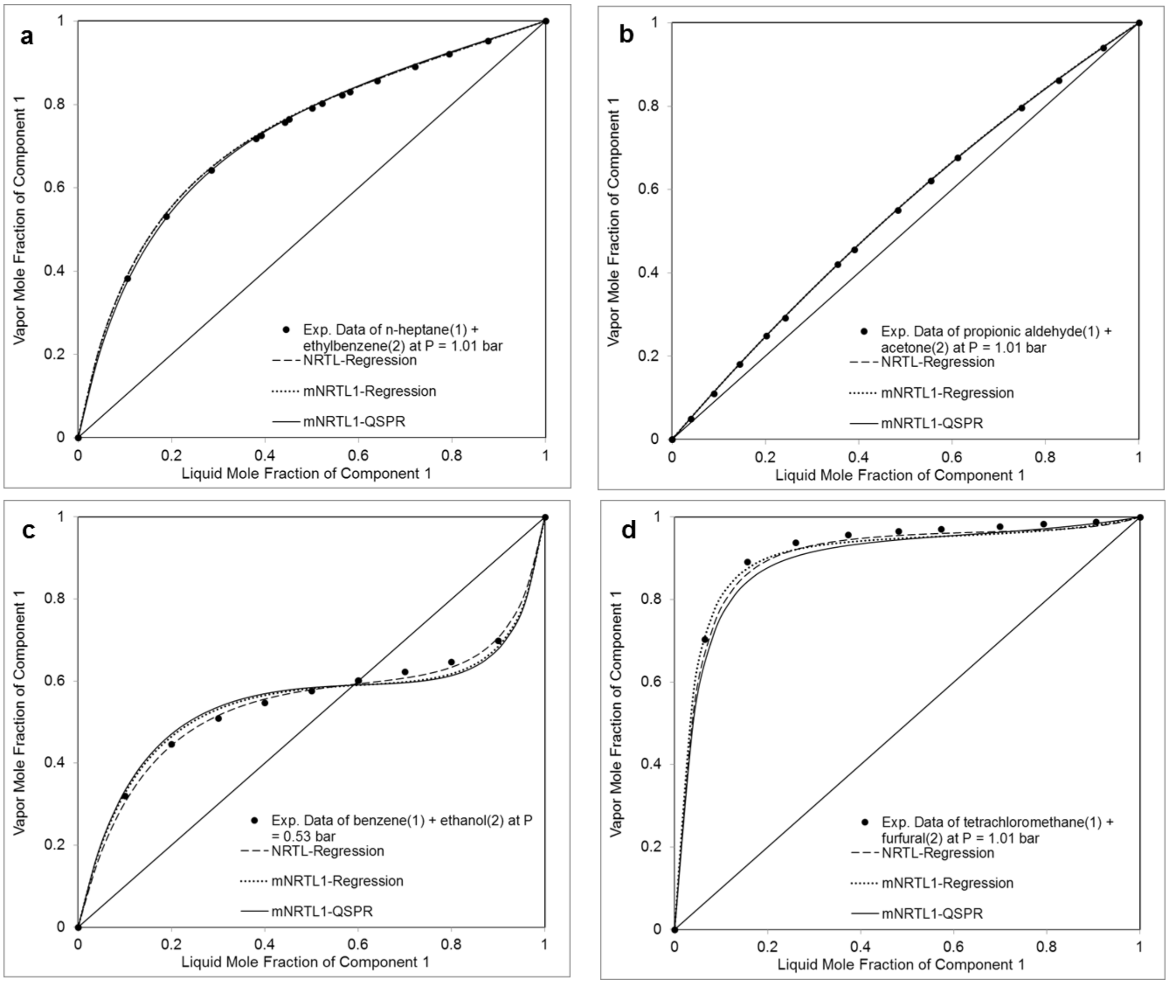


Figure 6.9. Regression and QSPR equilibrium phase composition predictions for (a) n-heptane (1) + ethylbenzene (2), (b) propionic aldehyde (1) + acetone (2), (c) benzene (1) + ethanol (2) and (d) tetrachloromethane (1) + furfural (2) systems.

REFERENCES

1. Renon, H. and J.M. Prausnitz, *Local compositions in thermodynamic excess functions for liquid mixtures.* AIChE Journal, 1968. **14**(1): p. 135-144.

2. Gmehling, J., J. Li, and M. Schiller, *A modified UNIFAC model. 2. Present parameter matrix and results for different thermodynamic properties.* Industrial & Engineering Chemistry Research, 1993. **32**(1): p. 178-193.

3. Abrams, D.S. and J.M. Prausnitz, *Statistical thermodynamics of liquid mixtures: A new expression for the excess Gibbs energy of partly or completely miscible systems.* AIChE Journal, 1975. **21**(1): p. 116-128.

4. Skjold-Jorgensen, S., B. Kolbe, J. Gmehling, and P. Rasmussen, *Vapor-liquid equilibria by UNIFAC group contribution. Revision and extension.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(4): p. 714-722.

5. Fischer, K. and J. Gmehling, *Further development, status and results of the PSRK method for the prediction of vapor-liquid equilibria and gas solubilities.* Fluid Phase Equilibria, 1996. **121**(1-2): p. 185-206.

6. Wilson, G.M., *Vapor-liquid equilibrium. XI. A new expression for the excess free energy of mixing.* Journal of the American Chemical Society, 1964. **86**(2): p. 127-130.

7. Prausnitz, J.M., R.N. Lichtenthaler, and E.G.d. Azevedo, *Molecular thermodynamics of fluid-phase equilibria*. 3rd ed. 1998: Prentice-Hall.

8. Gmehling, J., D. Tiegs, and U. Knipp, *A comparison of the predictive capability of different group contribution methods.* Fluid Phase Equilibria, 1990. **54**: p. 147-165.

9. Gebreyohannes, S., B.J. Neely, and K.A.M. Gasem, *One-parameter modified nonrandom two-liquid (NRTL) activity coefficient model*. 2014, Oklahoma State University: Manuscript submitted for publication.

10. Ravindranath, D., B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, *QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

11. Neely, B.J., *Aqueous hydrocarbon systems: Experimental measurements and quantitative structure-property relationship modeling*, in *School of Chemical Engineering, Ph.D. Dissertation*. 2007, Oklahoma State University: Stillwater, Oklahoma.

12. Godavarthy, S.S., R.L. Robinson Jr., and K.A.M. Gasem, *SVRC-QSPR model for predicting saturated vapor pressures of pure fluids.* Fluid Phase Equilibria, 2006. **246**(1-2): p. 39-51.

13. Gebreyohannes, S., K. Yerramsetty, B.J. Neely, and K.A.M. Gasem, *Improved QSPR generalized interaction parameters for the nonrandom two-liquid activity coefficient model.* Fluid Phase Equilibria, 2013. **339**(0): p. 20-30.

14. Scott, R.L., *Corresponding states treatment of nonelectrolyte solutions.* The Journal of Chemical Physics, 1956. **25**(2): p. 193-205.

15. Arlt, W., M.E.A. Macedo, P. Rasmussen, and J.M. Sorensen, *Liquid-liquid equilibrium data collection*. Chemistry Data Series. Vol. V, Parts 1-4. 1979 - 1987: DECHEMA, Frankfurt, Germany.

16. *Dragon Professional 6.0.9*. 2011, Talete SRL.

17. Katritzky, A.R., V.L. Lobanov, and M. Karelson, *Codessa 2.7.8*. 2007.

18. Gmehling, J., U. Onken, and W. Arlt, *Vapor-liquid equilibrium data collection*. Chemistry Data Series. Vol. I, Parts 1-8. 1977 - 2001: DECHEMA, Frankfurt, Germany.

19. *NIST-TDE, NIST Standard Reference Database 103b ThermoData Engine*. 2012.

20. *DIPPR Project 801, Physical and Thermodynamic Properties of Pure Chemicals*. 2011.

21. Rackett, H.G., *Equation of state for saturated liquids.* Journal of Chemical and Engineering Data, 1970. **15**(4): p. 514-517.

22. Tassios, D., *The number of roots in the NRTL and LEMF equations and the effect on their performance.* Industrial & Engineering Chemistry Process Design and Development, 1979. **18**(1): p. 182-186.

23. *ChemBioOffice 11.0*. 2008, CambridgeSoft.

24. *The Open Babel Package 2.3*. 2011, Last accessed on: http://openbabel.sourceforge.net/.

25. Guha, R., M.T. Howard, G.R. Hutchison, P. Murray-Rust, H. Rzepa, C. Steinbeck, J. Wegner, and E.L. Willighagen, *The blue obelisk-interoperability in chemical informatics.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 991-998.

26. Halgren, T.A., *Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94.* Journal of Computational Chemistry, 1996. **17**(5-6): p. 490-519.

27. Prechelt, L., *Automatic early stopping using cross validation: quantifying the criteria.* Neural Networks, 1998. **11**(4): p. 761-767.

28. Yerramsetty, K.M., B.J. Neely, and K.A.M. Gasem, *A non-linear structure–property model for octanol–water partition coefficient.* Fluid Phase Equilibria, 2012. **332**(0): p. 85-93.

29. Bagheri, M., K. Yerramsetty, K.A.M. Gasem, and B.J. Neely, *Molecular modeling of the standard state heat of formation.* Energy Conversion and Management, 2013. **65**(0): p. 587-596.

30. Caruana, R., S. Lawrence, and L. Giles, *Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping*. 2000, Advances in Neural Information Processing Systems 13, MIT Press: Cambridge, MA. p. 402-408.

31. Jakob, A., H. Grensemann, J. Lohmann, and J. Gmehling, *Further development of modified UNIFAC (Dortmund): Revision and extension 5.* Industrial & Engineering Chemistry Research, 2006. **45**(23): p. 7924-7933.

CHAPTER VII

CONCLUSIONS AND RECOMMENDATIONS

In this chapter, the conclusions and recommendations of the studies presented in Chapters 2-6 are presented.

7.1. Improved QSPR generalized interaction parameters for the NRTL activity coefficient model

The objective of this part of the study was to demonstrate the efficacy of an improved QSPR modeling approach for the generalization of the NRTL model parameters. We developed an internally consistent QSPR model using 578 binary VLE systems consisting of a wide range of functional groups. The conclusions and recommendations from this work are presented below.

Conclusions

- The results showed a QSPR modeling approach is effective in generalizing the NRTL model interaction parameters for a wide range of systems.
- An improved generalization methodology was demonstrated by eliminating the potential inconsistency resulting from the use of separate models for each of the two NRTL parameters. Further, a more representative database compared to a database used previously by our research group [1] was implemented in the generalization process.
- The QSPR generalized model parameters resulted in vapor-liquid phase equilibrium property predictions within twice the error of the data regression errors.

Recommendations

- A larger VLE database should be assembled that consists of chemical classes not currently represented in the database. Such improvement will widen the applicability domain of the QSPR model.
- The generalization methodology should be improved by ensuring the model's capability in reflecting pure and infinite-dilution limits accurately. This improvement will lead to better predictions for activity coefficients at infinite dilution.

7.2. A comparative study of QSPR generalized activity coefficient model parameters for VLE mixtures

The objective of this part of the work was to assess the representation capability of the NRTL, UNIQUAC and Wilson models and generalize the model parameters of the three activity coefficient models using a QSPR modeling approach. The conclusions and recommendations from this work are presented below.

Conclusions

- Our assessment showed the NRTL, UNIQUAC and Wilson models have comparable representation capabilities for all types of interactions. The three models with regressed parameters provided overall representation %AADs of approximately 2, 0.2, 4 and 6 for *T*, *P*, y_1 and *K-value*, respectively. While results for most systems were reasonable, all three models resulted in higher errors (approximately twice errors compared to the overall results) for water containing systems.
- In this study, an improved generalization methodology was implemented. This new improvement resulted in QSPR models that obey the limiting behavior of mixtures.
- An improved generalization methodology was implemented. The new improvement lead QSPR models that obey the limiting behavior of mixtures.

- The developed QSPR models provided phase equilibria property predictions within two times the errors obtained through the data regression analyses.
- Our methodology provides *a priori* and easily implementable QSPR models with wider applicability range than that of the UNIFAC model.

Recommendations

- Additional data for systems containing chemical classes that are not represented in the current database should be assembled. Such improvement will widen the applicability domain of the QSPR model.
- The database should be expanded by adding more water containing systems. This will allow conclusive comparisons to be made of the representation capabilities of the three models for aqueous systems.
- Additional infinite-dilution activity coefficient data should be assembled for all the VLE systems that are represented in the database. The additional data will allow better evaluation of the performance of the QSPR models for predicting limiting behavior for diverse mixture types.

7.3. Generalized NRTL interaction model parameters for predicting LLE behavior

The objective of this part of the work was to generalize the interaction parameters of the NRTL model for LLE systems using a theory-framed quantitative structure-property relationship (QSPR) modeling approach. The conclusions and recommendations from this work are presented below.

Conclusions

- The study demonstrated an effective methodology for generalizing the NRTL model interaction parameters for LLE systems.

- The newly developed QSPR model yielded binary predictions that are approximately 3 to 4 times the errors found from the regression analysis for about 90% of the systems considered.
- In comparison to the UNIFAC-1981-LLE model, the QSPR model provided lower errors, as well as a wider range of applicability for LLE property predictions.
- The NRTL interaction parameters for LLE systems are highly dependent on temperature unlike the interaction parameters of VLE systems.

Recommendations

- A larger LLE database encompassing a wide range of functional-group interactions should be assembled. This improvement will widen to applicability domain of the QSPR model.
- Further research should be focused on modifying the current models to better capture the temperature dependence of the LLE system interaction parameters. A potential area that may lead to better accounting of temperature dependence is incorporating equation-of-state interaction concepts within the NRTL model. Further, we need to investigate the capability of the UNIQUAC model for LLE systems by modifying the residual part of the model by learning from the theoretical formulation of equation-of-state models.

7.4. One-parameter modified NRTL activity coefficient model

The objective of this part of the work was to propose a modified version of the NRTL activity coefficient model which addresses the limitation of the original model, namely strong parameter correlation and generalizability. The modified NRTL model was expressed by recasting the formulation of the model parameters of the original NRTL model. The study presented two versions of a modified NRTL model, namely the two parameter NRTL model (mNRTL2) and the one parameter NRTL model (mNRTL1). The conclusions and recommendations from this work are presented below.

Conclusions

- The study proposed a modification to the original NRTL activity coefficient model which addressed the limitation of the original model.
- The ratio of the interaction energy parameter was generalized by using pure-component properties (mNRTL1), which reduced the model to only one energy interaction parameter and eliminated the parameter correlation. This enabled to qualitatively classify VLE behaviors based on degree of non-ideality.
- The mNRTL1 model provided VLE equilibrium property representations with a slight loss of accuracy compared to the original NRTL model.

Recommendations

- Further study needs to be carried out to generalize the one parameter in the modified model using structural descriptors of molecules. This will result in a generalized activity coefficient model that is capable of *a priori* prediction solely based on structural descriptors.
- Further study should be focused on investigating the temperature dependence of the parameter in the newly modified model using additional VLE and LLE systems. Such investigation will provide an insight on the effect of temperature and the variation of model parameters when going from VLE to LLE and vice versa for a wide range of interactions. Better accounting of the temperature dependence for LLE systems may be attained by incorporating equation-of-state interaction concepts within the modified NRTL model.

7.5. Generalized interaction model parameter for the modified NRTL activity coefficient model

The objective of this part of the study was to generalize the interaction parameter of the modified NRTL model for VLE systems using a theory-framed quantitative structure-property relationship (QSPR) modeling approach. The conclusions and recommendations from this work are presented below.

Conclusions

- The result from this study revealed that a QSPR modeling approach is effective in generalizing the single parameter of the modified NRTL model.
- The study demonstrated the advantage of having a single parameter through the elimination of the sequential regression analysis technique, which was required when generalizing two model parameters.
- The QSPR generalized model provided property predictions within twice the overall errors found in the experimental data regression analysis.

Recommendations

- Additional data for systems consists of chemical classes that are not currently represented in the database should be assembled. Such improvement will widen the applicability domain of the QSPR model.
- The QSPR modeling approach applied in this study should be extended for LLE systems, which will help in evaluating the performance of a generalized modified NRTL model for LLE systems.

REFERENCES

1. Ravindranath, D., B.J. Neely, R.L. Robinson Jr., and K.A.M. Gasem, QSPR generalization of activity coefficient models for predicting vapor-liquid equilibrium behavior. Fluid Phase Equilibria, 2007. 257(1): p. 53-62.

VITA

Solomon Gebreyohannes

Candidate for the Degree of

Doctor of Philosophy

Thesis: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP GENERALIZED ACTIVITY COEFFICIENT MODELS

Major Field: Chemical Engineering

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Chemical Engineering at Oklahoma State University, Stillwater, Oklahoma in May, 2014.

Completed the requirements for the Master of Science in Chemical Engineering at Oklahoma State University, Stillwater, Oklahoma in 2010

Completed the requirements for the Bachelor of Science in Chemical Engineering at Bahir Dar University, Bahir Dar, Ethiopia in Year.

Experience:

- Research and Teaching Assistant, Oklahoma State University, Stillwater, Oklahoma, August 2008 – December 2013
- Graduate Assistant, Bahir Dar University, Bahir Dar, Ethiopia, June 2007 – June 2008

Professional Memberships:

- Chemical Engineering Graduate Student Organization, 2011 – Present
- Honor Society of Omega Chi Epsilon Mu Chapter, 2010 – 2012
- International Society of Automation, OSU chapter, 2010 – 2011
- Golden Key International Honour Society, 2009 – Present
- American Institute of Chemical Engineers, 2008 – Present