QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELING & COMPUTER-AIDED MOLECULAR DESIGN: IMPROVEMENTS & APPLICATIONS

By

KRISHNA M. YERRAMSETTY

Bachelor of Engineering (Honors)
Birla Institute of Technology and Science
Pilani, Rajasthan (India)
2005

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, 2012

QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELING & COMPUTER-AIDED MOLECULAR DESIGN: IMPROVEMENTS & APPLICATIONS

Dissertation Approved:

Dr. Khaled A. M. Gasem

Dissertation Adviser

Dr. Robert L. Robinson, Jr.

Dr. Josh D. Ramsey

Dr. Martin Hagan

Outside Committee Member

Dr. Sheryl A. Tucker

Dean of the Graduate College

.

PREFACE

The objective of this work was to develop an integrated capability to design molecules with desired properties. An automated robust genetic algorithm (GA) module has been developed to facilitate the rapid design of new molecules. The generated molecules were scored for the relevant thermophysical properties using non-linear quantitative structure-property relationship (QSPR) models. The descriptor reduction and model development for the QSPR models were implemented using evolutionary algorithms (EA) and artificial neural networks (ANNs). QSPR models for octanol-water partition coefficients (K_{ow}), melting points (MP), normal boiling points (NBP), Gibbs energy of formation, universal quasi-chemical (UNIQUAC) model parameters, and infinite-dilution activity coefficients of cyclohexane and benzene in various organic solvents were developed in this work. To validate the current design methodology, new chemical penetration enhancers (CPEs) for transdermal insulin delivery and new solvents for extractive distillation of the cyclohexane + benzene system were designed.

In general, the use of non-linear QSPR models developed in this work provided predictions better than or as good as existing literature models. In particular, the current models for NBP, Gibbs energy of formation, UNIQUAC model parameters, and infinite-dilution activity coefficients have lower errors on external test sets than the literature models. The current models for MP and K_{ow} are comparable with the best models in the literature. The GA-based design framework implemented in this work successfully identified new CPEs for transdermal delivery of insulin, with permeability values comparable to the best CPEs in the literature. Also, new solvents for extractive distillation of cyclohexane/benzene with selectivities two to four times that of the existing solvents were identified. These two case studies validate the ability of the current design framework to identify new molecules with desired target properties.

This work was made possible by the personal support and encouragement by some talented people at the Oklahoma State University. Firstly, I would like to express my gratitude to my adviser, Professor Khaled A. M. Gasem, for giving me the opportunity to work on this project. My sincere appreciation goes to Dr. Gasem for his endless support, patience, and guidance throughout my graduate studies at Oklahoma State University.

My gratitude is also extended to Dr. Robert L. Robinson, Jr., for his expertise, motivation, and valuable advice during my Ph. D. program. I would also like to thank my graduate advisory committee members, Dr. Martin Hagan, Dr. Josh D. Ramsey, for their valuable input and guidance during the course of this study.

My special thanks go to Dr. Brian J. Neely for his constant guidance, enthusiasm and friendly advice. Also, the friendship and support of my colleagues, Sharath K. Golla, Vijay K. Rachakonda, Dr. Sayeed Mohammad, Eric Whitebay, and Solomon B. Gebreyohannes was most instrumental in making my graduate program a fun and productive experience. Also, I would like to thank the National Institute of Health (Grant#: 1R21EB005749) and the Department of Energy (EPSCoR grant# 0814361) for their financial support without which this work would not have been possible.

Finally, I would like to extend my sincere thanks to my family, for their patience, understanding, and support.

TABLE OF CONTENTS

Chapter Page

Chapter Page

LIST OF TABLES

Table Page

Table Page

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

CHAPTER 1

INTRODUCTION

1.1. Rationale

The demand for newly designed molecules that enhance existing processes and satisfy more stringent operating requirements in technology has been increasing. However, the rational design of molecules with desired properties challenges engineers attempting to meet the needs of various industries, including pharmaceuticals, polymers, petrochemicals, and construction [1-4]. The traditional approach of identifying molecules with desired properties involves testing thousands of molecules for their chemical and physical properties, which is an expensive and laborious undertaking. Hence, rational design techniques, such as computer-aided molecular design (CAMD), have found wide application in recent years [4, 5]. CAMD methods have been employed successfully in a wide range of applications, including solvent design/selection [6], design of chloro-fluro-carbon (CFC) substitutes, alternative process fluids design, polymer design [1], drug design [7], and design for novel molecules with superior properties [3]. A typical CAMD algorithm utilizes two key components, (a) a search method for generating candidate molecules, and (b) models to predict the pertinent physiochemical properties of the generated candidate molecules. Search methods involve mathematical programming, heuristic search approaches or evolutionary approaches. Evolutionary approaches are fast becoming the preferred search algorithms because of their ease of application. However, in most studies, the search space is limited to a certain family of molecular functional groups. This leads to a reduction in computational time at the cost of failing to discover better molecules that may be present outside the search space. Therefore, there is a need for developing generalized molecular search algorithms for CAMD.

Property predictions for the generated molecules are usually done using group-contribution methods, equation-of-state approaches, and quantitative structure-property relationship (QSPR) models. The present state of CAMD is heavily reliant on fragment-based QSPR models for property predictions. This leads to inaccurate predictions when the generated structures have fragments that are not included in the training phase of the models. Models based on molecular descriptors that provide complete 3-dimensional (3D) information of molecules do not suffer from this disadvantage and can be used to predict properties for structures with fragments that have not been included in the training phase. In addition, majority of the QSPR efforts in the literature are based on linear models, which can fail when a strong nonlinear relationship exists between the target property and molecular structure. However, techniques for building reliable nonlinear QSPR models using only relevant molecular descriptors are not well established in the literature and require further development. Specifically, our analyses indicate that: (a) nonlinear QSPR models based on 3D molecular information will outperform linear fragment-based models, and (b) generalized evolutionary search techniques for CAMD that employ nonlinear 3D QSPR models for property prediction lead to better design of molecules.

Accurate QSPR models are important not only for property predictions in CAMD but also for any process design in general, where reliable *a priori* property predictions are sought to avoid experimentation. As such, the present work places equal emphasis on building accurate non-linear QSPR models and developing a generalized CAMD framework, which incorporates non-linear QSPR models based on 3D molecular descriptors as the prediction platform. Therefore, the focus of the present work is to: (a) improve the existing QSPR methodology by developing accurate non-linear models based on 3D molecular information, and (b) develop a generalized CAMD methodology for designing molecules with desired properties. To exemplify the efficacy of the proposed methodology, relevant properties such as octanol-water partition coefficient (K_{ow}), boiling point, melting point, infinite-dilution activity coefficients, and solvent selectivities are modeled using 3D non-linear QSPR, after which new chemical penetration enhancers (CPEs) for improved transdermal insulin permeability and new solvents for extractive distillation are designed using the CAMD framework.

1.2. Goals and Objectives

The two primary goals of this work are to develop robust non-linear 3D QSPR property models and generalized CAMD methodologies for designing new molecules targeted for specific applications. Figure 1.1 describes the overall strategy of the present work, which is carried out in four stages. In the first stage, QSPR models are built using evolutionary algorithms and artificial neural networks (ANNs) to address the major limitations of the existing methods. The second stage involves the application of these improved QSPR methods for predicting K_{ow}, infinite-dilution activity coefficients, boiling points, melting points and solvent selectivities. The third stage focuses on improving the computer-aided molecular design (CAMD) methodology. In the fourth stage, the improved CAMD that incorporates predictions from the third stage are applied for the design of new chemical penetration enhancers (CPEs) for enhancement of insulin permeation through skin and also for designing new solvents for extractive distillation.

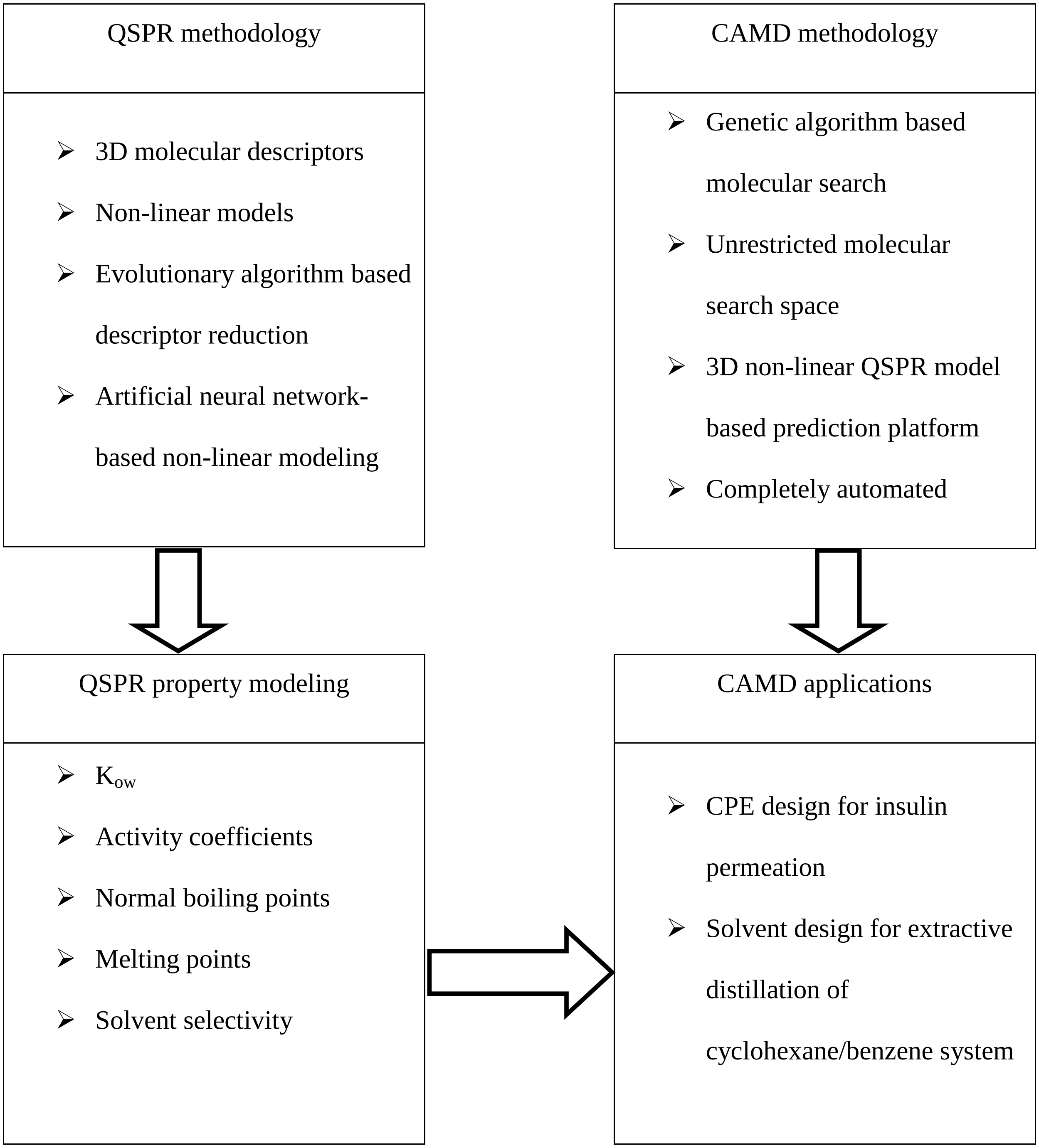
Following are the specific objectives undertaken to achieve the goals of this research:

1. Improve our existing QSPR methodology by developing evolutionary algorithms for selecting the best descriptors for non-linear modeling from a large set of initial descriptors.
2. Apply the improved QSPR methodology to develop *a priori* predictive thermophysical property models, including K_{ow}, infinite-dilution activity coefficients, boiling points, melting points and solvent selectivities.
3. Improve our existing CAMD methodology by (a) generalizing the genetic algorithms for creating new molecules, and (b) automating the different steps involved in CAMD to minimize user supervision.
4. Incorporate the relevant non-linear QSPR models and apply the improved CAMD methodology to discover new CPEs for insulin and new extractive distillation solvents of interest in the energy sector.

The methods advanced in this dissertation have produced a robust general framework for designing new molecules and an improved framework for building accurate models for thermophysical properties. In addition, applications of these improved frameworks have facilitated the design of improved CPEs for insulin, which could contribute to major advancements toward developing transdermal patches for insulin delivery. Similarly, molecular design of new solvents for extractive distillation will be greatly beneficial in reducing the separation cost of difficult-to-separate mixtures.

1.3. Organization of the Dissertation

This dissertation is organized in the “manuscript style,” and divided into eight stand-alone chapters. Chapter 2 describes in detail the QSPR methodology employed in this work to develop the various models for the molecular properties. Chapters 3 to 8 are concerned with the specific details of the development of QSPR models for various thermophysical properties significant for designing new CPEs and extractive distillation solvents. Since the basic modeling methodology is the same for the various models, some sections are repetitive in these chapters. The final chapter describes the CAMD algorithm used in the current work, and it also exemplifies the algorithm for designing new CPEs and new solvents for extractive distillation. For reasons of intellectual property, the names of potential candidate molecules are not disclosed in this dissertation.



QSPR methodology

- 3D molecular descriptors
- Non-linear models
- Evolutionary algorithm based descriptor reduction
- Artificial neural network-based non-linear modeling

CAMD methodology

- Genetic algorithm based molecular search
- Unrestricted molecular search space
- 3D non-linear QSPR model based prediction platform
- Completely automated

QSPR property modeling

- K_{ow}
- Activity coefficients
- Normal boiling points
- Melting points
- Solvent selectivity

CAMD applications

- CPE design for insulin permeation
- Solvent design for extractive distillation of cyclohexane/benzene system

Figure 1.1: **The overall strategy for this dissertation**

REFERENCES

1. Sundaram, A. and V. Venkatasubramanian, *Parametric Sensitivity and Search-Space Characterization Studies of Genetic Algorithms for Computer-Aided Polymer Design.* Journal of Chemical Information and Computer Science, 1998. **38**(6): p. 1177-1191.

2. Venkatasubramanian, V., K. Chan, and J. Caruthers, *Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm.* J Chem Inf Comput Sci, 1995. **35**: p. 188 - 195.

3. Devillers, J., *Genetic Algorithms in Molecular Modeling*. 1996: Academic press.

4. Venkatasubramanian, V., K. Chan, and J.M. Caruthers, *Computer-Aided Molecular Design Using Genetic Algorithms.* Computers and Chemical Engineering, 1994. **18**(9): p. 833–844.

5. Harper, P.M., et al., *Computer-Aided Molecular Design with Combined Molecular Modeling and Group Contribution.* Fluid Phase Equilibria, 1999. **158**: p. 337-347.

6. Godavarthy, S.S., *Design of Improved Solvents for Extractive Distillation.* Ph.D. Dissertation, School of Chemical Engineering. 2004, Oklahoma State University: Stillwater, Oklahoma.

7. Li, J., *CAMD in Modern Drug Discovery.* Drug Discovery Today, 1996. **1**(8): p. 311-312.

CHAPTER 2

QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP (QSPR) MODELING METHODOLOGY

2.1. Introduction

Recent advances in computational technology have created new opportunities for virtual synthesis and evaluation of compounds, which reduce the burden of time and resources associated with traditional experimentation. Computer-aided molecular design (CAMD) is the general term used to describe the process of virtual design of new molecules possessing specific, desired molecular properties. A successful CAMD process needs an accurate prediction platform to compute the relevant thermophysical properties of the generated candidate molecules. Although theory-based models would be preferred, currently, theoretical models are not available for most properties and investigators are forced to rely on empirical or semi-empirical models. A well-known semi-empirical approach for predicting molecular properties is quantitative structure-property relationship (QSPR) modeling, which asserts that quantifiable relationships exist between the thermophysical properties and molecular structure of a substance. When the same techniques are used in predicting activities of biological compounds, the models are usually referred to as quantitative structure-activity relationship (QSAR) models. This is not a strict naming convention, and a QSPR model in the current work refers to any model relating a property to the molecular structure.

The molecular structure of any compound is characterized in terms of certain variables called molecular descriptors, which are usually calculated using quantum-mechanical methods based in theory. In mathematical terms, a QSPR model for any property P is of the following form:

$$\mathrm{P} = f(\text{molecular descriptors}) \tag{2.1}$$

where, f denotes a linear or non-linear mathematical function (model) used to express the property in terms of molecular descriptors. Initially, molecules with known property values are used to optimize the QSPR models, and then these optimized models are used to predict the properties of unknown molecules.

Before outlining the details of QSPR modeling, a brief historical background of various QSPR methodologies will be presented. QSPR techniques have appeared in the literature for over a century. They have facilitated the prediction of thermophysical properties of a molecule based solely on information from its chemical structure [1-3]. Although successful structure-property relationships do not completely eliminate chemical synthesis or experimental validation, a significant reduction in the number of molecules requiring synthesis and validation can be realized. The early major advancements in QSPR-related research were pioneered by Hansch and Fujita [4, 5], who correlated biological activities with hydrophobic, steric and electronic properties of molecular structure, and by Free and Wilson [6], who developed the group-contribution approach to property prediction. After 1980, the availability of inexpensive computational power led to an explosion in the number of QSPR studies, and numerous models have been proposed in the literature to predict varied and often complex thermophysical properties of molecules including normal boiling points, solvent polarity scales, melting points, and refractive indices [4, 7-13]. A thorough review of the history, major areas of applications and software related to QSPR is provided by Katritzky et al. [14].

2.2. Overview of QSPR Methodology

As shown in Figure 2.1, a typical QSPR model development has the following basic steps:

1. ***Database development*** involves collecting representative experimental data of assessed quality and assembling a relevant database. The quality of the data is assessed to establish the experimental uncertainties associated with the data considered.
2. ***Structure generation*** involves the development of the 2-dimensional and 3-dimensional representations of the molecular structures. If 3-dimensional descriptors are needed, then optimization for the lowest conformational energy of the molecules is performed.
3. ***Molecular descriptor calculation*** is undertaken for the molecules in the database using relevant software like CODESSA [15] or DRAGON [16].
4. ***Descriptor reduction*** is the step where the most significant descriptors from the large set of available molecular descriptors are identified.
5. ***Model development*** is the step where the most significant descriptors are correlated with a molecular property using linear or non-linear modeling tools.
6. ***Model validation*** entails evaluation of the predictive performance of the final model.

Typically, the descriptor reduction and model development steps are carried out simultaneously, since information feedback from the model development step is provided to the descriptor reduction step. In fact, these two steps constitute an iterative process that is terminated when certain stopping criteria are met. The model validation and model development steps can also behave in a similar fashion.

Although all QSPR development steps are important, the two critical steps have a major influence on the performance of a QSPR model: descriptor reduction (DR), where the significant structural descriptors are determined, and model development (MD), where the modeling approach is selected. Several approaches have been proposed in the literature for QSPR model development. While the basic steps in QSPR development remain the same, differing techniques are applied for the described steps, (1) – (6). In following sections, a brief overview of these approaches will be provided, along with a description of the techniques used in the current work.

2.3. Database Development

The performance of empirical techniques such as QSPR modeling is heavily dependent on the quality and characterization of data available for use in the training stages. Ideally, the data should include molecules that are similar to the molecules for which the model is intended to be used. For example, to develop a model to predict the octanol-water partition coefficients of drug-like molecules, the training data should ideally include a wide range of drug-like molecules. However, all the models in the current work are developed to be generally applicable to all types of molecules, and so care is taken to ensure that the employed training databases are as diverse as possible. In addition, to ensure accurate QSPR models, only the highest possible quality experimental data with low uncertainties were used for model building in this work, and the sources of the data and uncertainties in the data are provided in future chapters of this dissertation, where ever applicable.

2.4. Structure Generation and Optimization

QSPR models utilize molecular representations ranging from the simplest 1-dimensional (1D) descriptors, which account for gross molecular properties like molecular weight, number of atoms and meting point, to complex 4-dimensional (4D) representations [17, 18], where multiple conformers of a single molecule are considered. The most common molecular representations in QSPR modeling are the 2- and 3-dimenisonal (2D and 3D, respectively) representations. A 2D representation of a molecule encodes the topology and connectivity information and has been used successfully in a wide variety of QSPR models [4-6]. Since a single molecule always has a unique 2D representation, developing 3D QSPR models is inherently more difficult due to the large number of 3D representations based on the number and type of constituent atoms (e.g., Figures 2.2 and 2.3 represent the 2D and 3D structures of salicylic acid). Therefore, finding the “actual” 3D representation, as defined by the lowest conformational energy of a molecule, is essential before inclusion in the QSPR model.

Finding the minimum energy conformation from a large number of possible conformations is a combinatorial optimization problem. Semi-empirical methods such as AM1 [19, 20], PM3 [21, 22] and PM6 [23, 24] have been used widely to calculate the minimum energy of a given conformation. These semi-empirical methods are based on the Hartree-Fock formalism (used for determining the ground-state wave function and ground-state energy of a molecule); however, they involve several approximations and some of these parameters are obtained from empirical data. The semi-empirical methods are employed in computational chemistry to determine the wave-functions and energies of large molecules, for which the full Hartree-Fock treatment would be computationally impractical. In a recent article, Rinnan et al. [25] compared different methods of energy minimization and concluded that the final QSPR models are not influenced significantly by the choice of the energy minimization method, provided the lowest energy conformer has been found *a priori*. However, the majority of QSPR articles in the literature only apply the energy minimization techniques to a randomly chosen 3D conformer of a molecule. This can potentially lead to inaccurate or sub-optimal models. Therefore, in the current work, molecular-structure optimization was performed, which resulted in a global search for the minimum-energy 3D conformation.

While different software packages may be used, structure generation requires a series of steps common to all QSPR models, where initially the 2D structure is drawn based on either names or the simplified molecular input line entry specification (SMILES) and subsequently an optimal 3D structure is identified. In the current work, ChemBioDraw Ultra 11.0 [26] was used to generate 2D structures for the molecules in the data set and stored as cdx files. The conformers with the least energy were found by implementing OpenBabel's [27, 28] genetic algorithm (GA) based conformer search which uses the MMFF94 forcefield [29]. The GA for conformer search can be tailored for accuracy versus computational time by varying four different options that include number of structural conformers or parents in each generation, number of child conformers generated per each parent, mutability parameter for determining the frequency of mutation operations and the number of unchanged generations after which the algorithm is stopped. For further information, the readers are referred to the OpenBabel documentation on conformer searching [30]. In the current work, 30 parent conformers and 5 child conformers were chosen, the mutability parameter was set to 5 and the number of constant generations was set at 25. The optimized molecules were saved in mdl format for subsequent generation of descriptors.

2.5. Descriptor Calculation

The variables used to describe the molecules present in the QSPR database are called molecular descriptors. The accuracy of the final QSPR model depends partly on the accuracy with which these descriptors are calculated. Several types of descriptors can be calculated depending on the representation used for the molecule. As described in the previous section, QSPR models employ generally the following two types of molecular descriptors: (a) 2D descriptors that provide connectivity information concerning the atoms in the molecule, and (b) 3D descriptors that are calculated from the 3-dimensional spatial positioning of atoms of the molecule.

ADAPT software [31, 32] was an early version of an automated program for QSAR/QSPR modeling. ADAPT calculates the following types of descriptors: fragment type, sub-structure type, environment type (providing interconnection information between sub-structures), molecular connectivity type (providing information about the amount of branching in the molecule), and geometric type (describing the shape of the molecule). Some of the earliest commonly used descriptors were linear free energy constants such as the Hammett σ constant (measure of the electronic effects of the aromatic substituent), Taft polar constants (σ^*) (measure of the electronic effects of the polar substituent), Hansch π (measure of the hydrophobicity of the substituent) and Taft steric constant E_s (measure of substituent steric effects) [33]. Usage of these descriptors for developing QSAR models is referred to as the Hansch approach. However, the above models are based on simple linear and additive models and are applicable only to co-generic series of molecules where only the substituents are altered [34]. Also, the above constants are not available for every substituent and therefore are not applicable for a wide range of molecules. An alternative approach to Hansch methodology was proposed by Free and Wilson, who assume that for molecules in a co-generic series, the activity is determined by mutually independent contributions from the substituents in the molecule [6]. There are also hybrid approaches that combine the above two methodologies.

Other common descriptors include topological indices (TIs) that provide molecular connectivity information. The advantage of these descriptors is their relatively short computational time since 3D molecular information is not required. Many TIs with varying advantages and disadvantages have been proposed in the literature. Wiener's index [35, 36] is the earliest of these indices. Balaban [37] presented a basic review on the most widely used TIs prior to 1988 and establish six criteria for a good TI. However, TIs are based only on 2D information of the molecule and therefore cannot be used to represent the spatial conformation of the atoms. In recent years, information content indices based on Shannon information theory have been developed, which can also be considered as TIs [38]. Molecular volume has also been used widely in the early years of QSPR history [39].

Cartier and Rivail [39] were among the earliest researchers to include theoretical quantum chemical descriptors calculated using semi-empirical methods in their QSPR models. Quantum chemistry facilitates a more accurate calculation of the electronic effects than the empirical methods [39]. These effects can be calculated theoretically from the geometry-optimized 3D structure of the molecules. Some of the common quantum chemical descriptors are energies of HOMO and LUMO (ε_{HOMO}, ε_{LUMO}), net atomic charge of atom A (Q_A), molecular polarizability (α) and molecular dipole moment (µ); however, these calculations are based on wave-function theories and involve approximations that limit their applicability to structurally related molecules [40]. Some of the most commonly used routines for calculating the quantum chemistry descriptors are *ab initio* models like the Hamiltonian and the Hatree-Fock-method, semi-empirical methods like the extended Hϋckel theory, complete neglect of differential overlap (CNDO), intermediate neglect of differential overlap (INDO), Austin model 1 (AM1) and parametric model 3 (PM3). Typically, for *ab initio* calculations, the calculation time required is proportional to a high power of the number of electrons in the molecule [40] and therefore, these calculations are computationally expensive. The semi-empirical methods, however, are based on molecular orbital (MO) calculations coupled with experimental data on atoms, which allows for faster calculations than the *ab initio* methods.

One of the widely used software for developing a QSPR model is Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) [15], which was developed by Katritzky et al. [14, 38] as a non-empirical tool for calculating various descriptors such as constitutional, topological, geometrical, electrostatic, thermodynamic, quantum-chemical, molecular orbital (MO)-related and charged partial surface areas (CPSA) descriptors. When implemented, CODESSA does not require experimental data and the descriptors are calculated based entirely on the chemical structure of the molecules. This program has been applied successfully for correlating a large number of physical properties such as boiling points, melting points, and solubility of gases in liquids [40]. In the current work, all QSPR models have been built using descriptors generated by DRAGON 6 [16] software developed by Talete SRL. DRAGON 6 is capable of generating over 4800 descriptors categorized into 0D, 1D, 2D and 3D descriptors. For a detailed list of descriptors calculated by DRAGON, the reader is referred to the DRAGON website [16]. Several successful QSPR models based on DRAGON descriptors have been published in the literature [41-43]. Table 2.1 lists some examples of 2D and 3D descriptors calculated using DRAGON.

2.6. Descriptor Reduction and Model Development

In the current work, around 4800 molecular descriptors may be generated for a given molecule using DRAGON; however, most of these descriptors have negligible influence on a desired property of the molecule and, thus, they must be eliminated systematically to arrive at a tractable set of the most significant descriptors. Reduction or pruning of the descriptor set is a key step in QSPR model development. Various methods exist for descriptor reduction, which include the following: linear orthogonalization [1], principal component analysis (PCA) [44-50], partial least squares (PLS) [51, 52], genetic algorithms (GAs) [53], forward propagating neural networks [54], back propagation neural networks [55], self-organizing maps [53, 56, 57], fuzzy ARTMAP neural networks [58], decision trees [59], logistic regression [60], support vector machines [61, 62], simulated annealing [63], particle swarms [64], ant colony algorithms [65], and various hybrid combinations of the above methods. While each method has its own advantages and limitations and most are efficient methods for pruning a large dataset, they have not been applied widely in conjunction with non-linear QSPR modeling.

Recently, Golla et al. [66, 67] expanded the descriptor set for each molecule by introducing non-linear transformations to all the descriptors. The descriptors were evaluated for significance, and the most significant descriptor is retained and removed from the descriptor pool. The evaluation is repeated and a set of significant descriptors is identified in a sequential fashion. This process of sequential analysis (SA) allows the determination of correlation of the transformed (non-linear) descriptors with the property of interest. An additional benefit of SA is the provision of a rudimentary cause-and-effect type analysis of the descriptor set. The extended dataset can then be used for initial pruning. In this way, the chances of discarding any descriptors that show a non-linear relationship with the property being considered is reduced, if not totally eliminated.

As shown in Table 2.2, the resulting permutations of DR and MD lead to four general modeling types. To date, the DR methods in the literature are largely linear, and the majority of QSPR models reported are also linear (Type I); however, more recent work has employed non-linear QSPR models (Type II). Several QSPR model development efforts in the literature [1, 68, 69] have shown that the relationship between molecular structure and thermo-physical properties is often non-linear. Therefore, use of linear algorithms for descriptor reduction or model development fails to capture the subtle (and even not-so-subtle) relationships between the chemical structure and thermo-physical properties. Further, the inclusion of SA in the DR strategy still results in Type III models that often lead to sub-optimal solutions.

The approach in this work for descriptor reduction involves a hybrid strategy, which results in a Type IV model. 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 (Desc_Sz). The method begins with an initial population of single hidden-layered ANNs (individuals) that have been divided into four different niches. Niches, in the context of this work, are mutually exclusive sub-populations in the original population, which are not allowed to exchange genetic material. Niches are helpful in maintaining genetic diversity in the population [70, 71]. The ANNs in the initial population are assigned random descriptor subsets as inputs. These ANNs then undergo (a) single-point mutation on the descriptor subsets, (b) modified differential evolution (MDE) operations on the descriptor subsets, (c) retraining with different initial weights, and (d) change in the number of hidden neurons, over successive generations. The ANNs that can predict accurately the target property are favored over inaccurate ANNs to remain in the population. Therefore, ANNs in the later generations are, on average, closer to the global minimum of the objective function. The subsequent discussion will be a brief introduction to EP, DE and ANNs, followed by details on the actual descriptor reduction algorithm employed in the current study.

2.6.1. Evolutionary Programming (EP): Evolutionary programming is a stochastic optimization algorithm first developed by Lawrence J. Fogel in 1960 [72]. Similar to other stochastic algorithms, EP is well suited for combinatorial optimization problems where the fitness surface can have multiple local minima. Further, EP can be coded efficiently using real-valued genetic representation of the problem space and, therefore, has an advantage over GAs, which can be coded only using binary genetic representation. The basic EP algorithm has the following three steps that are repeated in each generation until some convergence criterion is met:

1. Randomly generate an initial population of a fixed size, N. Usually, the population size is heuristically determined based on the number of independent variables that describe the fitness surface.
2. Generate children from the parent population using a mutation operation that is chosen from a distribution of possible mutations that range from the most to least severe. Severity of a mutation operation is measured by the amount of functional change between the parent and the offspring.
3. Evaluate the fitness of the child population and select the best individuals from both the parent and child population. The selection is usually done by stochastic tournament, where N individuals are retained for the next generation.

EP has been applied successfully for a diverse range of optimization problems like power system optimization [73], prediction of the effects of genetic modifications [74] and prediction of protein-ligand structures [75]. One of the first applications of EP to QSPR modeling was by Luke in 1994 [76], who compared his methodology with existing QSPR techniques for several commonly used QSPR data sets. Another EP based algorithm is the Mutation and Selection Uncover Model (MUSEUM) [77], which uses only mutation to generate offspring from parents and was shown to be much faster than other regression models. To the author’s knowledge, apart from the two references cited above, no other application of EP to QSPR/QSAR modeling appears in the literature despite its advantages over other evolutionary algorithms.

2.6.2. Differential Evolution (DE): DE is another simple stochastic optimization algorithm similar to GA and was proposed by Price and Storn [78] in 1994. The major difference between GA and DE is that the former uses probability distribution for selection of parents; while in the latter trial vectors are generated. This makes the DE algorithm self-organizing by reducing the number of parameters that need to be pre-set by the user. The basic DE algorithm for minimization has the following steps:

1. Randomly generate an initial population of a fixed size, N. Usually, the population size is heuristically determined based on the dimensionality of the fitness surface denoted as n.
2. Perform the following for successive generations until some stopping criterion is met:

   For each vector **x** in the population, the following steps are conducted:

   a. Choose three different individuals **a**, **b** and **c** that are different from **x**.
   b. Pick a random integer R between 1 to n.
   c. Generate a trial vector **y** ={y_1, y_2, …y_i, …y_n} by iterating over each i from 1 to n
      - Generate a uniformly distributed random number r between 0 and 1.
      - If i=R or if r < CR (cross-over number), then $y_i = a_i + F*(b_i - c_i)$, else $y_i = x_i$, where F is the mutation factor.
   d. If the trial vector **y** has lower objective function than the original vector **x**, then replace **x** with **y.**

DE has been successfully applied to various optimization problems such as heat exchanger network synthesis [79], reservoir system optimization [80], design of temperature profiles for fermentation processes [81] and image pixel clustering [82]. Despite its popularity in other optimization fields, DE has been applied to few QSPR studies to date, which include prediction of atomic charges by Ouyang et al. [83] and predicting skin permeability of insulin in the presence of chemical penetration enhancers by the Oklahoma State University group [84].

2.6.3. Artificial Neural Networks (ANNs): Artificial neural networks are inspired by the brain and the interconnections among neurons, which form a complex network where electrical and other types of signals are exchanged to facilitate functioning of the brain. Although, much of neuronal function in the brain is still unclear, researchers have been able to develop ANNs as limited and simplified models for recreating intelligence artificially. Although the concept of ANNs is well established, the development of the back-propagation learning algorithm by Rumelhart et al. [85] in 1986 led to an explosion in the number of applications of ANNs. In the literature, neural networks have been employed as a non-linear modeling tool for function approximation/regression analysis, time-series forecasting, robotics and data processing. Different types of ANNs exist based on architecture, but in view of the current work, only feed-forward ANNs are relevant and any future reference to ANNs in the current work refers to feed-forward ANNs. Figure 2.4 is a neural interpretation diagram (NID) of a sample feed-forward ANN with 6 inputs, 2 hidden neurons in a single hidden layer and one output (Insulin permeability K_p) [84]. A NID is a diagram representing the neural network structure along with the weights between the different neurons and can be used to interpret the relationships between the output variable and the various input variables to the network. Using this approach, the connections between the neurons will be represented by lines whose thickness depends upon the magnitude of the weight between the corresponding neurons. The thickness of the lines connecting two neurons is proportional to the magnitude of the connection weight between them. Also, to differentiate between the direction of contribution of input variables to the output of a neuron, blue lines and black lines will be used for negatively-contributing and positively-contributing inputs, respectively. A NID therefore, provides qualitative information about the magnitude and the direction of the effect of each input on the output. For a detailed discussion on NIDs and their interpretation, the readers are referred to Olden and Jackson [86].

In a feed-forward ANN, information travels only in the forward direction from the input nodes to the output nodes. The different layers are connected using weights and biases which represent the strength of the signal between the different nodes, and these are updated during the learning phase of the algorithm to minimize the error between the network outputs and the network targets.

An important aspect of ANNs is the architecture or design, which consists of the selection 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 the number of data points to the number of inputs is at least ten. For most applications, using only one hidden layer is adequate; however, choosing the right number of neurons in this hidden layer may not be straight-forward. Choosing too few hidden neurons might lead to an ANN that lacks flexibility to encapsulate the complexity of the data and choosing too many may lead to over-fitting and poor generalization. No theoretical basis exists for choosing the number of hidden neurons and most researchers use trial and error for selection of the architecture leading to the best performance. In the current work, the number of hidden layers is fixed at one, and the minimum number of hidden neurons is two. This number is increased by one for randomly selected ANNs in the population of ANNs, and the better performing ANNs are retained for the next generation/iteration. In addition, for each ANN, the ratio of the number of training data to the number of adjustable weights and biases was ensured to be always greater than two [87]. This was done as a precaution against over-fitting to the training data.

Once, the network architecture has been specified, an ANN is trained on known data before its use as a predictive tool. The most popular learning algorithm for feed-forward ANNs is the back-propagation algorithm proposed by Rumelhart et al. [85], which consists of a forward propagation step and a weight-update step that is repeated until the network performance is satisfactory. The network weights and biases can be updated using several algorithms, but the most popular are the gradient-descent and Levenberg-Marquardt algorithms [88]. However, these algorithms do not guarantee attainment of the global minimum; thus, multiple initializations of the program are often necessary. Iyer and Rhinehart [89] have proposed a multiple initialization method to increase the probability of locating the global minimum. This method is built into the descriptor reduction algorithm used in the current work.

Another important issue associated with ANN training is over-fitting, which results in poor predictive capability. Although several methods for avoiding over-fitting an ANN exist, over-fitting is avoided in the current work by using an internal validation set (V), with an early-stopping method [90, 91]. The validation error normally decreases during the initial phase of training, as does the training set error; however, when the network begins to over-fit the data, the error in the validation set typically begins to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum validation error are retained. Although the early-stopping algorithm is easy to understand and implement, choosing the right training and internal validation sets is not straightforward. In addition to the training (T) and V sets, an internal test (IT) set is generally used in selecting the best ANNs during the descriptor search algorithm. The error on the IT set was used an additional indication of the generalization ability of the individual ANNs. Since, all three data sets (T, V, and IT) are involved in the ANN selection process, the predictive performance of the final ANN model can only be estimated using an external test set that contains data not present in any of the aforementioned three data sets.

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. Several methods exist in the literature for allocation of the data such as random division [92, 93], self-organizing maps (SOMs) [94, 95], Kennard-Stone design [96] and the sphere-exclusion algorithm [97]. In the current work, SOMs are used to divide the data sets optimally prior to the ANN training. The SOM MATLAB toolbox from the Laboratory of Information and Computer Science in the Helsinki University of Technology [98] is used for training the SOMs. SOMs are used to identify clusters of data in the input space, and from each cluster at least one data point is added to the training set. If a cluster has more than one data point, then random selection is used to divide the data in each cluster into the various subsets of T, V and IT. If the data in each cluster cannot be equally divided among the three subsets, preference is given for addition of data points to the training, validation or internal test sets in that order. This process ensures that the training set has the largest number of data points, followed by the validation and internal test sets, respectively. 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 of 65-70% of the entire data set (excluding the external set). Ideally, the training of each ANN is preceded by SOM training using the same inputs for both ANN and SOM; however, SOM training is computationally expensive and therefore, in this work, SOMs were trained once for every iteration of the algorithm for each niche. Here, a niche is a group of individuals that are allowed to exchange genetic material among themselves through DE operations. Individuals belonging to different niches are never allowed to take part in the same DE operation. In each iteration of the evolutionary algorithm, the most commonly-occurring inputs in a niche are used as inputs for the SOMs. Although only one SOM is trained for each niche, the random selection of data from each cluster is carried out separately for each individual in the niche. This ensures slightly different data sets for each individual.

During training of the ANNs in the current work, the inputs and targets (the experimental values of the property that need to be modeled) are normalized to have zero mean and unity standard deviation, which ensures that exceptionally large-valued descriptors or targets do not bias the network. The Nguyen-Widrow algorithm is used to initialize weights and biases, which are updated using the Levenberg-Marquardt optimization technique.

2.6.4. Genetic Representation: A good genetic representation of the solution domain is an important step in developing an efficient evolutionary algorithm. Binary representation is most widely used due to the direct encoding technique for most problems and the applicability for crossover dependent evolutionary algorithms like GA and DE [99]. Real-valued representations on the other hand are better suited for algorithms like EP that are dependent upon mutation as the major evolutionary operator.

In the current work, the solution space is comprised of single hidden layer ANNs with all possible molecular descriptor subsets of a fixed model size (ND) as inputs, which are determined by the user at the start of the program. The number of hidden neurons (NH) in these ANNs lies between a minimum of two and a maximum that is usually fixed at three times the value of ND. Therefore, 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. An example of three sample chromosomes with 100 original descriptors (Desc_sz) and a model size (ND) of five is shown in Table 2.3. Each chromosome is made up of five genes, where each gene represents a descriptor that is used as an input variable to an associated ANN, which is subsequently trained to predict the target property.

Binary representation of the chromosomes entails large memory requirements, and also the algorithm takes longer to converge to a global minimum when compared with real-valued representations. The above considerations are the reason for using real-valued chromosomes in the current work.

2.6.5. The Objective Function: Another major aspect of an evolutionary algorithm is the choice of a suitable objective function. In the current work, a wrapper-based Type IV (see Table 2.1) modeling approach is used for simultaneous descriptor reduction and non-linear model development using ANNs. 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. In general, the objective function for the second tier of optimization in a wrapper-based descriptor reduction approach must be selected such that it is a good estimate of the performance of the underlying linear or non-linear models. The objective functions chosen for linear QSPR models typically maximize statistical measures such as the correlation coefficient (R^2) [100], adjusted R^2, q^2 [101] and Akaike information content [102]. For non-linear models, the root-mean- squared error (RMSE) [103, 104] and absolute average deviation (AAD) [105] are used. In the current work, the entire data set excluding the external test set data was split into training (T), internal validation (V) and internal test sets (IT). The RMSE values between the predicted and target values were calculated for each of these subsets. The following objective function (*F*) was then computed based on these RMSE values:

$$\mathrm{F} = \mathrm{RMSE_T} + \mathrm{RMSE_V} + \mathrm{RMSE_{IT}} \tag{2.2}$$

With proper selection of an objective function, one can apply an algorithm to search for the set of descriptors resulting in an ANN that produces a minimum objective function value.

2.6.6. The Algorithm: The flow chart for the algorithm is given in Figure 2.5. Before execution of the algorithm, the following parameters are set by the user: (a) Desired number of descriptors in the model (ND), (b) Population size (Pop_sz),which is usually set at 400, (c) Number of niches (N_Niche), which is usually set equal to the ratio of

Pop_sz and 100 to ensure that each niche has 100 individuals, (d) Percentage of population that undergoes MDE operations (MDE_p), which is usually set at 0.1, (e) Percentage of population that undergoes retraining (Ret_p), which is usually set at 0.3, and (f) Percentage of population that undergoes change in the number of hidden neurons (Arc_p), which is usually set at 0.5.

The algorithm has an initialization process that executes once. The individual ANNs in a parent population denoted as 'D' are initialized with random descriptor subsets of size ND. The jth gene in the ith individual is represented as D(i,j). The number of hidden neurons for each ANN is initialized to a value of 2. The ANNs are then 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. The objective function *F* for the ith individual in population 'D' is denoted as *F*{D(i)}. Population 'D' then undergoes the following five operations in a single iteration of the algorithm.

1. *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'.
2. *Modified differential evolution:* N (=Pop_size*MDE_p) number of 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'. First, a mutated population, defined as TM, is generated by

combining the genes from three different individuals in population ‘D’. This operation is similar to the mutation operation in the traditional DE algorithm. Next, the mutated population ‘TM’ and the parent population ‘D’ are recombined using the recombination operation of the traditional DE algorithm. The recombined population is denoted as population ‘M’. The ANNs in ‘M’ undergo training and the f{M(i)} 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’. If *f*{M(i)} is lower than *f*{D(i)}, then M(i) is considered fitter than D(i), and therefore, M(i) replaces D(i) in population ’D’. This is denoted as ‘individual competition.’ The pseudo-code for the MDE operations and selection is shown in Figure 2.6.

3. *Retraining:* N (=Pop_size*Ret_p) number of ANNs are selected randomly from population ‘D’ for retraining using different initial weights. The retrained ANNs make up a new population denoted as ‘R’. The corresponding individuals in populations ‘D’ and ‘R’ undergo individual competition and population ‘D’ is updated using the fitter individuals.
4. *Architectural change:* N (=Pop_size*Ret_p) number of individuals are selected randomly from population ‘D’. The number of hidden neurons (NH) in half of these individuals is increased by 1 and for the rest of the individuals the NH value is decreased by 1. If NH for any individual falls below the specified minimum value of 2, then the NH value is adjusted to the minimum value of 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 *f*{A(i)} 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.

5. *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 [106]. 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 N (=Pop_sz) number of 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.

2.6.7. Creating Ensembles: ANNs are known to be highly 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 ensemble formation of ANNs is used, where the predictions of different ANNs are averaged to result in the final predictions [107, 108]. The ANNs in the ensemble can differ with respect to (a) the training data, (b) weights between the different nodes, (c) the number of hidden layers and neurons, and finally (d) the input descriptors. For the current work, specific details concerning ensemble construction are presented below.

Once the algorithm has met the stopping criteria, the descriptors that occur at least 10 times in each niche are identified. These descriptors are termed elite descriptors. Three individuals in each niche that have the most number of elite descriptors are identified.

Non-elite descriptors in these selected individuals are deleted. Each such modified individual from every niche is retrained using a different number of hidden neurons (NH) varied from 2 to 2*ND. The ‘best 100’ ANNs in terms of their objective function value are identified and recorded. This process of retraining using a different number of hidden neurons is carried out for several iterations using different initial weights. If the ANNs identified during an iteration have lower *f* values than any of the ANNs in the ‘best 100’ list, then these fitter ANNs replace the unfit individuals in the ‘best 100’ list. Following each iteration, the ‘best 100’ list is updated. The algorithm is stopped if the ‘best 100’ list stays the same for 100 successive iterations. Of these 100 best ANNs, the 20 networks that have the lowest ‘sum of weights and biases’ values are combined using a simple averaging technique to create an ensemble. Three such ensembles are created in every niche. The predictions from the ensembles from all niches are then averaged to result in the final predictions.

2.7. Conclusions

Since their inception, QSPR modeling techniques have improved significantly and have now become one of the important tools in the virtual design paradigm. The purpose of the current chapter was to introduce the various steps involved in developing a QSPR model, specifically, database development, molecular descriptor calculation, descriptor reduction, and model development. However, selecting the most relevant feature subset from the large set of all possible molecular descriptors is still a difficult task, particularly in the case of wrapper-based techniques where, descriptor reduction and modeling of the target property are carried out simultaneously. In the current chapter, a novel hybrid algorithm that combines evolutionary programming (EP) and differential evolution DE techniques is proposed as a solution for the feature selection problem. The current algorithm employs ANNs as the mapping tool between the molecular descriptors and the target property. To further improve the generalization capability of the model, ensembles of ANNs are created where the final predictions are the simple averages of the predictions by the individual networks.

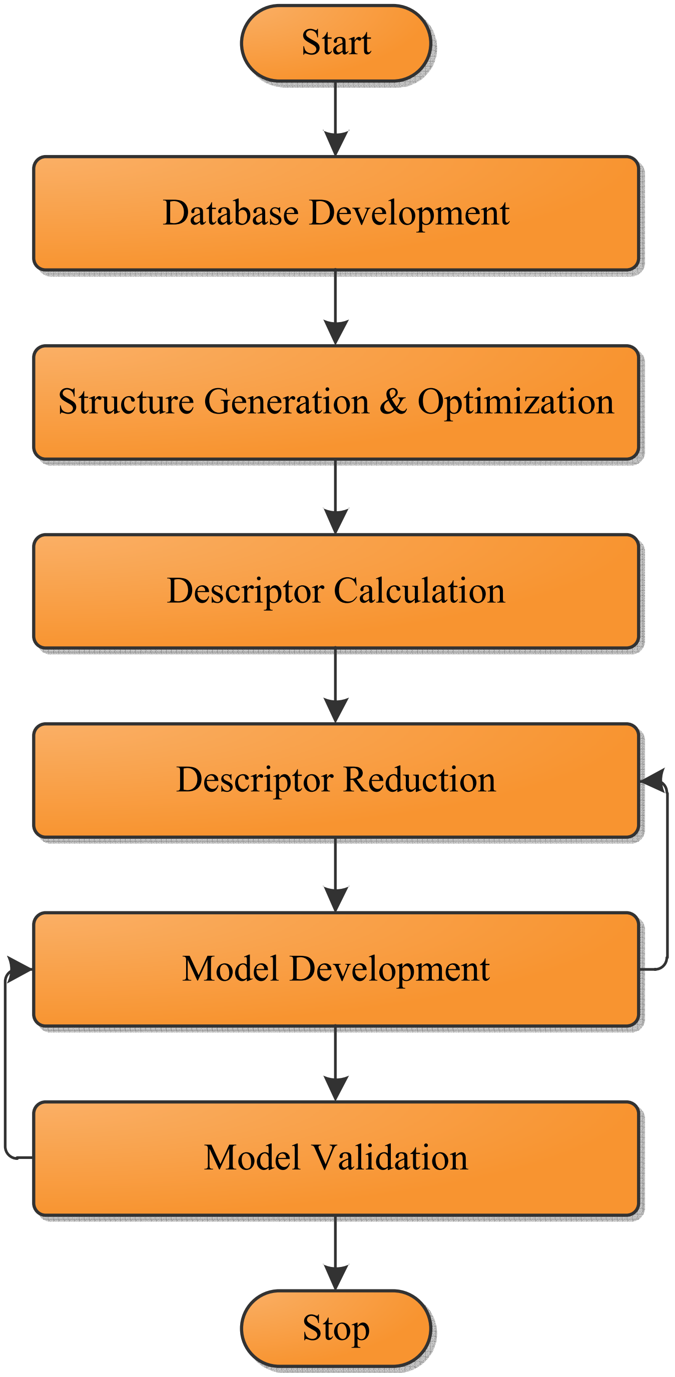
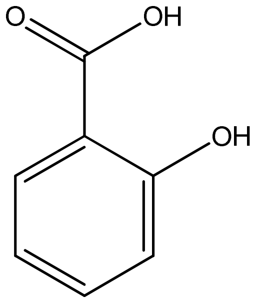


Figure 2.1: QSPR methodology



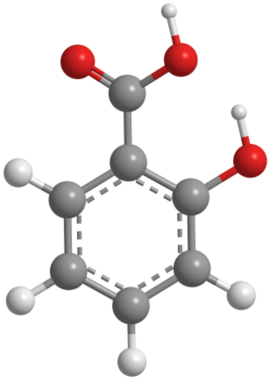


Figure 2.2: 2D representation of salicylic acid

Figure 2.3: 3D representation of salicylic acid

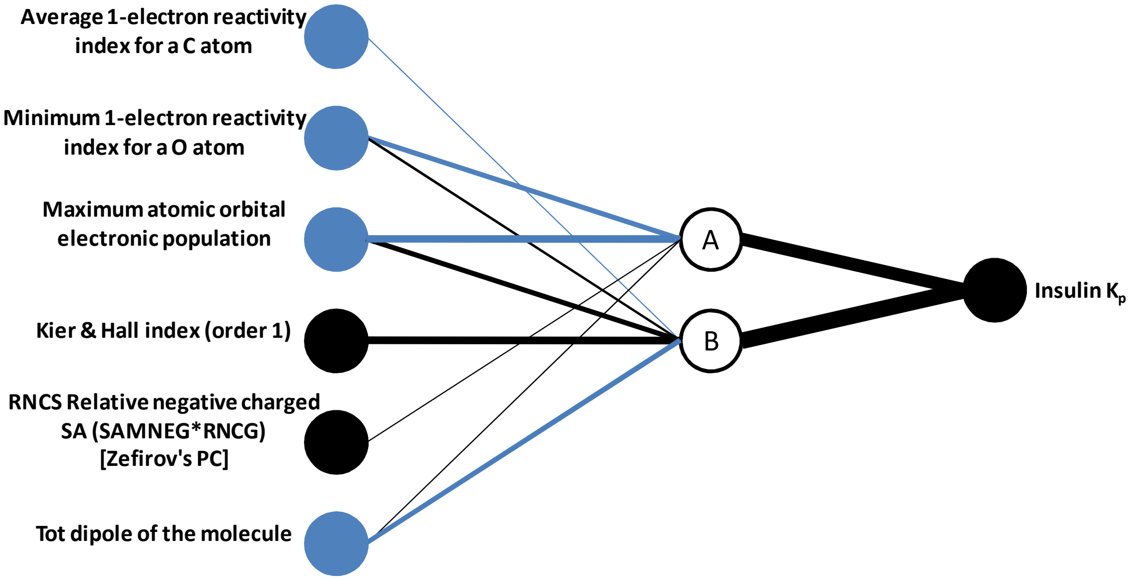

Average 1-electron reactivity index for a C atom
Minimum 1-electron reactivity index for a O atom
Maximum atomic orbital electronic population
Kier & Hall index (order 1)
RNCS Relative negative charged SA (SAMNEG*RNCG) [Zefirov's PC]
Tot dipole of the molecule
A
B
Insulin K_p

Figure 2.4: Neural interpretation diagram (NID) of a sample ANN

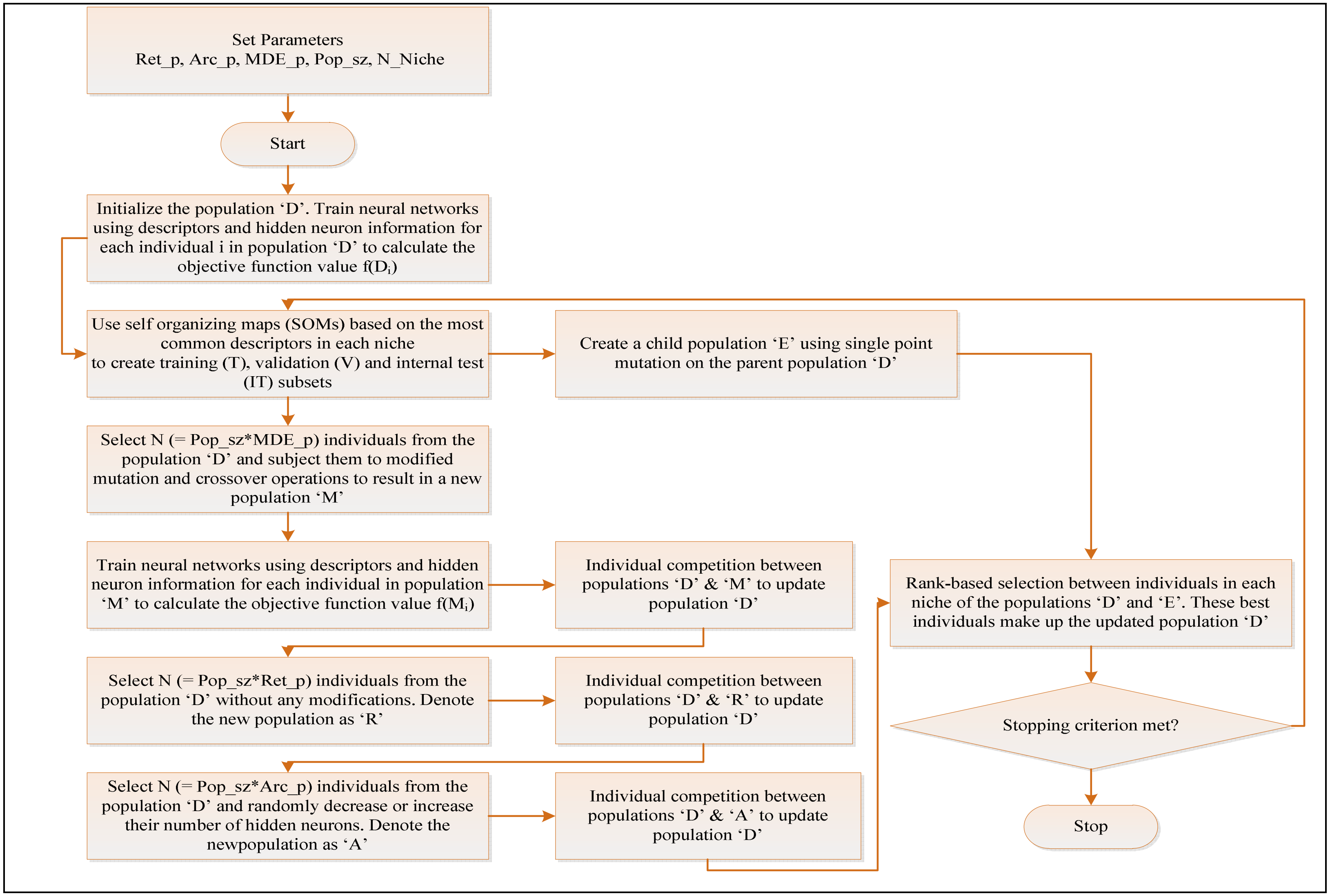

Set Parameters
Ret_p, Arc_p, MDE_p, Pop_sz, N_Niche
Start
Initialize the population 'D'. Train neural networks using descriptors and hidden neuron information for each individual i in population 'D' to calculate the objective function value f(Di)
Use self organizing maps (SOMs) based on the most common descriptors in each niche to create training (T), validation (V) and internal test (IT) subsets
Create a child population 'E' using single point mutation on the parent population 'D'
Select N (= Pop_sz*MDE_p) individuals from the population 'D' and subject them to modified mutation and crossover operations to result in a new population 'M'
Train neural networks using descriptors and hidden neuron information for each individual in population 'M' to calculate the objective function value f(Mi)
Individual competition between populations 'D' & 'M' to update population 'D'
Rank-based selection between individuals in each niche of the populations 'D' and 'E'. These best individuals make up the updated population 'D'
Select N (= Pop_sz*Ret_p) individuals from the population 'D' without any modifications. Denote the new population as 'R'
Individual competition between populations 'D' & 'R' to update population 'D'
Stopping criterion met?
Select N (= Pop_sz*Arc_p) individuals from the population 'D' and randomly decrease or increase their number of hidden neurons. Denote the newpopulation as 'A'
Individual competition between populations 'D' & 'A' to update population 'D'
Stop

Figure 2.5: Flowchart for the EP + DE algorithm used in the current work

```
For $i=1$ to $N$
    Select $a, b, c \in \{1, 2,.., Pop\_sz\}$ such that $a \neq b \neq c$
    For $j=1$ to $ND$ (Mutation operation)
        Generate rand, a random number between 0 & 1
        If ($rand \leq 0.25$), then $TM(i,j) = D(a,j)$
        If ($0.25 < rand \leq 0.5$), then $TM(i,j) = D(b,j)$
        If ($0.5 < rand \leq 0.75$), then $TM(i,j) = D(c,j)$
        If ($0.75 < rand \leq 1$), then $TM(i,j)$ is randomly selected from
        $\{0,1,2,\ldots,Desc\_sz)$
    For $j=1$ to $ND$ (Recombination operation)
        Generate rand, a random number between 0 & 1
        If ($rand \leq CR$), then $M(i,j) = TM(i,j)$
        If ($rand > CR$), then $M(i,j) = D(i,j)$
    If $f\{M(i)\} < f\{D(i)\}$
        $D(i) = M(i)$
```

Figure 2.6: Pseudo code for modified differential evolution (MDE) operations

Table 2.1: Examples of 2D and 3D descriptors calculated by DRAGON

2D Descriptors	Kier flexibility index	Molecular walk count of order 1	Randic ID number	Balaban X index
3D Descriptors	Radial distribution functions	3D Morse descriptors	Randic-type R matrix connectivity	Total symmetry index

Table 2.2: Types of QSPR models based on the linearity or non-linearity of the underlying descriptor reduction and model development methods

	Descriptor Reduction	**QSPR Model**
Type I	Linear	Linear
Type II	Linear	Non-linear
Type III	Modified non-linear	Non-linear
Type IV	Non-linear	Non-linear

Table 2.3: Three sample chromosomes of size five, chosen from a set of 100 descriptors

Chromosome #	**Descriptor 1**	**Descriptor 2**	**Descriptor 3**	**Descriptor 4**	**Descriptor 5**
Chromosome 1	23	45	54	3	98
Chromosome 2	23	49	22	9	67
Chromosome 3	34	44	1	7	100

REFERENCES

1. Godavarthy, S.S., R.L. Robinson Jr., and K.A.M. Gasem, *An Improved Structure-Property Model for Predicting Melting-Point Temperatures.* Industrial and Engineering Chemistry Research, 2006. **45**(14): p. 5117-5126.

2. 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.

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

4. Hansch, C. and T. Fujita, *r-s-p Analysis. A Method for the Correlation of Biological Activity and Chemical Structure.* Journal of the American Chemical Society, 1964. **86**: p. 1616-1626.

5. Hansch, C. and T. Fujita, *p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure.* Journal of the American Chemical Society, 1964. **86**(8): p. 1616-1626.

6. Free, S.M. and J.W. Wilson, *A Mathematical Contribution to Structure-Activity Studies.* Journal of Medicinal Chemistry, 1964. **7**(4): p. 395-399.

7. Katritzky, A.R., V.S. Lobanov, and M. Karelson, *Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure-Property Relationship.* Journal of Chemical Information and Computer Sciences, 1998. **38**: p. 28-41.

8. Katritzky, A.R., L. Mu, and M. Karelson, *QSPR Treatment of the Unified Nonspecific Solvent Polarity Scale.* Journal of Chemical Information and Computer Sciences, 1997. **37**: p. 756-761.

9. Stanton, D.T., et al., *Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles.* Journal of Chemical Information and Computer Sciences, 1992. **32**: p. 306-316.

10. Wessel, M.D. and P.C. Jurs, *Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure.* Journal of Chemical Information and Computer Sciences, 1995. **35**(5): p. 841-850.

11. Karelson, M., *Molecular Descriptors in QSAR/QSPR*. 1st ed. 2000, New York: John Wiley & Sons.

12. Katritzky, A.R., et al., *Prediction of Melting Points for the Substituted Benzenes:□ A QSPR Approach.* Journal of Chemical Information and Computer Sciences, 1997. **37**(5): p. 913-919.

13. Katritzky, A.R., S. Sild, and M. Karelson, *Correlation and Prediction of the Refractive Indices of Polymers by QSPR.* Journal of Chemical Information and Computer Sciences, 1998. **38**(6): p. 1171-1176.

14. Katritzky, A.R., et al., *The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors.* ChemInform, 2003. **34**(18): p. no-no.

15. Semichem Inc., A.R. Katritzky, and M. Karelson, *Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA)*. 1998: Shawnee, KS.

16. *Dragon Professional 6*. 2010, Talete SRL.

17. Vedani, A., et al., *Multiple-Conformation and Protonation-State Representation in 4D-QSAR:□ The Neurokinin-1 Receptor System.* Journal of Medicinal Chemistry, 2000. **43**(23): p. 4416-4427.

18. Streich, D., M. Neuburger-Zehnder, and A. Vedani, *Induced Fit—the Key for Understanding LSD Activity? A 4D-QSAR Study on the 5-HT2A Receptor System.* Quantitative Structure-Activity Relationships, 2000. **19**(6): p. 565-573.

19. Hashimoto, H., et al., *Molecular Structures of Carotenoids as Predicted by MNDO-AM1 Molecular Orbital Calculations.* Journal of Molecular Structure, 2002. **604**(2-3): p. 125-146.

20. Yatsenko, A.V. and K.A. Paseshnichenko, *On the Suitability of AM1 for the Modeling of Molecules Containing Amino Groups.* Journal of Molecular Structure: THEOCHEM, 1999. **492**(1-3): p. 277-283.

21. Dos Santos, H.F. and W.B. De Almeida, *MNDO/AM1/PM3 Quantum Mechanical Semiempirical and Molecular Mechanics Barriers to Internal Rotation: A Comparative Study.* Journal of Molecular Structure: THEOCHEM, 1995. **335**(1-3): p. 129-139.

22. Estrada, E., I. Perdomo-López, and J.J. Torres-Labandeira, *Molecular Modeling (Mm2 and PM3) and Experimental (Nmr and Thermal Analysis) Studies on the Inclusion Complex of Salbutamol and Β-Cyclodextrin.* The Journal of Organic Chemistry, 2000. **65**(25): p. 8510-8517.

23. Alparone, A., V. Librando, and Z. Minniti, *Validation of Semiempirical PM6 Method for the Prediction of Molecular Properties of Polycyclic Aromatic Hydrocarbons and Fullerenes.* Chemical Physics Letters, 2008. **460**(1-3): p. 151-154.

24. Correa, A., et al., *A Comparison of the Performance of the Semiempirical PM6 Method Versus DFT Methods in Ru-Catalyzed Olefin Metathesis*, in *Green Metathesis Chemistry*, V. Dragutan, et al., Editors. 2010, Springer Netherlands. p. 281-292.

25. Rinnan, Å., N. Christensen, and S. Engelsen, *How the Energy Evaluation Method Used in the Geometry Optimization Step Affect the Quality of the Subsequent QSAR/QSPR Models.* Journal of Computer-Aided Molecular Design, 2010. **24**(1): p. 17-22.

26. *Chembiooffice 11.0*. 2008, CambridgeSoft.

27. Guha, R., et al., *The Blue Obelisk-Interoperability in Chemical Informatics.* Journal of Chemical Information and Modeling, 2006. **46**(3): p. 991-998.

28. *The Open Babel Package, Version 2.3*. 2011; Available from: http://openbabel.sourceforge.net/.

29. 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.

30. The Open Babel Developers. *Obconformersearch Class Reference*. 2007 [cited 2011 July 26]; Available from: http://openbabel.org/dev-api/classOpenBabel_1_1OBConformerSearch.shtml.

31. Stuper, A.J. and P.C. Jurs, *Adapt: A Computer System for Automated Data Analysis Using Pattern Recognition Techniques.* Journal of Chemical Information and Computer Sciences, 1976. **16**(2): p. 99-105.

32. Rose, S.L. and P.C. Jurs, *Computer-Assisted Studies of Structure-Activity Relationships of N-Nitroso Compounds Using Pattern Recognition.* Journal of Medicinal Chemistry, 1982. **25**(7): p. 769-776.

33. Greenberg, M.J., *Dependence of Odor Intensity on the Hydrophobic Properties of Molecules. A Quantitative Structure Odor Intensity Relationship.* Journal of Agricultural and Food Chemistry, 1979. **27**(2): p. 347-352.

34. Franke, R., S. Huebel, and W.J. Streich, *Substructural QSAR Approaches and Topological Pharmacophores.* Environ Health Perspect, 1985. **61**: p. 239-55.

35. Wiener, H., *Structural Determination of Paraffin Boiling Points.* Journal of the American Chemical Society, 1947. **69**(1): p. 17-20.

36. Wiener, H., *Correlation of Heats of Isomerization, and Differences in Heats of Vaporization of Isomers, among the Paraffin Hydrocarbons.* Journal of the American Chemical Society, 1947. **69**(11): p. 2636-2638.

37. Balaban, A., *Topological Indices and Their Uses: A New Approach for the Coding of Alkanes.* Journal of Molecular Structure: THEOCHEM, 1988. **165(**3-4): p. 243-253.

38. Katritzky, A.R. and E.V. Gordeeva, *Traditional Topological Indexes Vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research.* Journal of Chemical Information and Computer Sciences, 1993. **33**(6): p. 835-857.

39. Cartier, A. and J.L. Rivail, *Electronic Descriptors in Quantitative Structure--Activity Relationships.* Chemometrics and Intelligent Laboratory Systems, 1987. **1**(4): p. 335-347.

40. Karelson, M., V.S. Lobanov, and A.R. Katritzky, *Quantum-Chemical Descriptors in QSAR/QSPR Studies.* Chemical Reviews, 1996. **96**(3): p. 1027-1044.

41. Gharagheizi, F., *QSPR Studies for Solubility Parameter by Means of Genetic Algorithm-Based Multivariate Linear Regression and Generalized Regression Neural Network.* QSAR & Combinatorial Science, 2008. **27**(2): p. 165-170.

42. Goodarzi, M., T. Chen, and M.P. Freitas, *QSPR Predictions of Heat of Fusion of Organic Compounds Using Bayesian Regularized Artificial Neural Networks.* Chemometrics and Intelligent Laboratory Systems, 2010. **104**(2): p. 260-264.

43. Papa, E., S. Kovarich, and P. Gramatica, *Development, Validation and Inspection of the Applicability Domain of QSPR Models for Physicochemical Properties of Polybrominated Diphenyl Ethers.* QSAR & Combinatorial Science, 2009. **28**(8): p. 790-796.

44. Brown, S.D., et al., *Chemometrics.* Analytical Chemistry, 1988. **60**(12): p. 252-273.

45. Brown, S.D. and R.S. Bear Jr., *Chemometrics.* Analytical Chemistry, 1990. **62**: p. 84R.

46. Brown, S.D., R.S. Bear Jr., and T.B. Blank, *Chemometrics.* Analytical Chemistry, 1992. **64**(12): p. 22-49.

47. Brown, S.D., et al., *Chemometrics.* Analytical Chemistry, 1994. **66**(12): p. 315-359.

48. Brown, S.D., et al., *Chemometrics.* Analytical Chemistry, 1996. **68**(12): p. 21-62.

49. Jolliffe, I.T., *Principal Component Analysis*. 1986, Berlin: Springer.

50. Malinowski, E.R. and D.G. Howery, *Factor Analysis in Chemistry*. 1st ed. 1980, New York: Wiley-Interscience.

51. Lindgren, F., et al., *Interactive Variable Selection (IVS) for PLS. Part 1: Theory and Algorithms*. Journal of Chemometrics, 1994. **8**(5): p. 349-363.

52. Wold, S., et al., *Multi-Way Principal Components and PLS Analysis.* Journal of Chemometrics, 1987. **1**(1): p. 41-56.

53. Bayram, E., et al., *Genetic Algorithms and Self-Organizing Maps: A Powerful Combination for Modeling Complex QSAR and QSPR Problems.* Journal of Computer-Aided Molecular Design, 2004. **18**(7): p. 483-493.

54. Bishop, C.M., *Neural Networks for Pattern Recognition*. 1998, Oxford: Clarendon Press.

55. Cibas, T., et al., *Variable Selection with Neural Networks.* Neural Computation, 1996. **12**: p. 223-248.

56. Kohonen, T., *Self-Organising Maps*. 1995, Heidelberg: Springer-Verlag.

57. Kohonen, T., *Comparison of SOM Point Densities Based on Different Criteria.* Neural Computation, 1999. **11**: p. 2081-2095.

58. Espinosa, G., et al., *A Fuzzy ARTMAP-Based Quantitative Structure-Property Relationship (QSPR) for Predicting Physical Properties of Organic Compounds.* Industrial and Engineering Chemistry Research, 2001. **40**(12): p. 2757-2766.

59. DeLisle, R.K. and S.L. Dixon, *Induction of Decision Trees Via Evolutionary Programming.* Journal of Chemical Information and Computer Sciences, 2004. **44**(3): p. 862-870.

60. Worth, A.P. and M.T.D. Cronin, *The Use of Discriminant Analysis, Logistic Regression and Classification Tree Analysis in the Development of Classification Models for Human Health Effects.* Journal of Molecular Structure: THEOCHEM, 2003. **622**(1): p. 97-111.

61. Asar, A. and E. Bastos. *A Comparative Estimation of Machine Learning Methods on QSAR Data Sets*. in *SAS Conference Proceedings: SAS Users Group International*. April 10-13, 2005 Philadelphia, Pennsylvania.

62. Burbidge, R., et al., *Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis.* Computers & Chemistry, 2001. **26**(1): p. 5-14.

63. de Vicente, J., J. Lanchares, and R. Hermida, *Placement by Thermodynamic Simulated Annealing.* Physics Letters A, 2003. **317**(5-6): p. 415-423.

64. Clerc, M. and J. Kennedy, *The Particle Swarm-Explosion, Stability, and Convergence in a Multidimensional Complex Space.* IEEE Transactions on Evolutionary Computation, 2002. **6**(1): p. 58-73.

65. Dorigo, M., V. Maniezzo, and A. Colorni, *Ant System: Optimization by a Colony of Cooperating Agents.* IEEE Transactions on Systems, Man and Cybernetics, Part B, 1996. **26**(1): p. 29-41.

66. Golla, S., et al., *Quantitative Structure-Property Relationship Modeling of Skin Sensitization: A Quantitative Prediction.* Toxicology in Vitro, 2009. **23**(3): p. 454-465.

67. Golla, S., et al., *Quantitative Structure-Property Relationships Modeling of Skin Irritation.* Toxicology in Vitro, 2009. **23**(1): p. 176-184.

68. Neely, B.J., et al. *Improved Quantitative Structure Property Relationship Models of Infinite-Dilution Activity Coefficients of Aqueous Systems*. in *Proceedings of the Sixth International Petroleum Environmental Conference*. 2004. Albuquerque, NM.

69. Ravindranath, D., et al., *QSPR Generalization of Activity Coefficient Models for Predicting Vapor-Liquid Equilibrium Behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

70. Wei, L. and M. Zhao, *A Niche Hybrid Genetic Algorithm for Global Optimization of Continuous Multimodal Functions.* Applied Mathematics and Computation, 2005. **160**(3): p. 649-661.

71. Horn, J., N. Nafpliotis, and D.E. Goldberg. *A Niched Pareto Genetic Algorithm for Multiobjective Optimization*. in *Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on*. 1994.

72. Fogel, L.J., *Intelligence through Simulated Evolution : Forty Years of Evolutionary Programming*. 1999, New York ; Chichester: Wiley. 162-174.

73. Sinha, N., R. Chakrabarti, and P.K. Chattopadhyay, *Evolutionary Programming Techniques for Economic Load Dispatch.* Evolutionary Computation, IEEE Transactions on, 2003. **7**(1): p. 83-94.

74. Patil, K., et al., *Evolutionary Programming as a Platform for in Silico Metabolic Engineering.* BMC Bioinformatics, 2005. **6**(1): p. 308.

75. Gehlhaar, D.K., et al., *Molecular Recognition of the Inhibitor AG-1343 by HIV-1 Protease: Conformationally Flexible Docking by Evolutionary Programming.* Chemistry & Biology, 1995. **2**(5): p. 317-324.

76. Luke, B.T., *Evolutionary Programming Applied to the Development of Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships.* Journal of Chemical Information and Computer Sciences, 1994. **34**(6): p. 1279-1287.

77. Kubinyi, H., *Variable Selection in QSAR Studies. II. A Highly Efficient Combination of Systematic Search and Evolution.* Quantitative Structure-Activity Relationships, 1994. **13**(4): p. 393-401.

78. Storn, R. *Differential Evolution (DE) for Continuous Function Optimization (an Algorithm by Kenneth Price and Rainer Storn).* [cited 2008 December]; Available from: http://www.icsi.berkeley.edu/~storn/code.html.

79. Yerramsetty, K.M. and C.V.S. Murty, *Synthesis of Cost-Optimal Heat Exchanger Networks Using Differential Evolution.* Computers & Chemical Engineering, 2008. **32**(8): p. 1861-1876.

80. Reddy, M.J. and D.N. Kumar, *Multiobjective Differential Evolution with Application to Reservoir System Optimization.* Journal of Computing in Civil Engineering, 2007. **21**(2): p. 136-146.

81. Oonsivilai, R. and A. Oonsivilai, *Differential Evolution Application in Temperature Profile of Fermenting Process.* WTOS, 2010. **9**(6): p. 618-628.

82. Das, S. and A. Konar, *Automatic Image Pixel Clustering with an Improved Differential Evolution.* Applied Soft Computing, 2009. **9**(1): p. 226-236.

83. Ouyang, Y., F. Ye, and Y. Liang, *A Modified Electronegativity Equalization Method for Fast and Accurate Calculation of Atomic Charges in Large Biological Molecules.* Physical Chemistry Chemical Physics, 2009. **11**(29): p. 6082-6089.

84. Yerramsetty, K.M., et al., *A Skin Permeability Model of Insulin in the Presence of Chemical Penetration Enhancer*. International Journal of Pharmaceutics, 2010. **388**(1-2): p. 13-23.

85. Rumelhart, D.E., 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.

86. Olden, J.D., D.A. Jackson, and P.R. Peres-Neto, *Predictive Models of Fish Species Distributions: A Note on Proper Validation and Chance Predictions.* Transactions of the American Fisheries Society, 2002. **131**(2): p. 329 - 336.

87. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

88. Hagan, M.T., H.B. Demuth, and M.H. Beale, *Neural Network Design*. 1996, Boston ; London: PWS Pub. 1 v. (various pagings).

89. Iyer, M.S. and R.R. Rhinehart, *A Method to Determine the Required Number of Neural-Network Training Repetitions.* IEEE transactions on neural networks, 1999. **10**(2): p. 427-432.

90. Prechelt, L., *Automatic Early Stopping Using Cross Validation: Quantifying the Criteria.* Neural Networks, 1998. **11**(4): p. 761-767.

91. Rich, C., L. Steve, and G. Lee, *Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping*. 2000.

92. Wu, W., et al., *Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set.* Chemometrics and Intelligent Laboratory Systems, 1996. **33**(1): p. 35-46.

93. Yasri, A. and D. Hartsough, *Toward an Optimal Procedure for Variable Selection and QSAR Model Building.* Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1218-1227.

94. Gramatica, P., P. Pilutti, and E. Papa, *Validated QSAR Prediction of OH Tropospheric Degradation of VOCs: Splitting into Training−Test Sets and Consensus Modeling.* Journal of Chemical Information and Computer Sciences, 2004. **44**(5): p. 1794-1802.

95. Gramatica, P., E. Giani, and E. Papa, *Statistical External Validation and Consensus Modeling: A QSPR Case Study for Koc Prediction.* Journal of Molecular Graphics and Modelling, 2007. **25**(6): p. 755-766.

96. Kennard, R.W. and L.A. Stone, *Computer Aided Design of Experiments.* Technometrics, 1969. **11**(1): p. 137-148.

97. Gunturi, S.B. and R. Narayanan, *In Silico ADME Modeling 3: Computational Models to Predict Human Intestinal Absorption Using Sphere Exclusion and kNN QSAR Methods.* QSAR & Combinatorial Science, 2007. **26**(5): p. 653-668.

98. *Laboratory of Computer and Information Science*. 2005 [cited 2011; Available from: http://www.cis.hut.fi/somtoolbox/.

99. Rothlauf, F. and D.E. Goldberg, *Representantions for Genetic and Evolutionary Algorithms*. 2002, Heidelberg, Alemania : Physica.

100. Toropov, A.A. and A.P. Toropova, *Modeling of Acyclic Carbonyl Compounds Normal Boiling Points by Correlation Weighting of Nearest Neighboring Codes.* Journal of Molecular Structure: THEOCHEM, 2002. **581**(1-3): p. 11-15.

101. Shen, M., et al., *Development and Validation of K-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates.* Journal of Medicinal Chemistry, 2003. **46**(14): p. 3013-3020.

102. Duchowicz, P.R., et al., *Application of the Replacement Method as Novel Variable Selection in QSPR. 2. Soil Sorption Coefficients.* Chemometrics and Intelligent Laboratory Systems, 2007. **88**(2): p. 197-203.

103. Karelson, M., et al., *Neural Networks Convergence Using Physicochemical Data.* Journal of Chemical Information and Modeling, 2006. **46**(5): p. 1891-1897.

104. Soto, A., et al., *A Wrapper-Based Feature Selection Method for ADMET Prediction Using Evolutionary Computing*, in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, E. Marchiori and J. Moore, Editors. 2008, Springer Berlin / Heidelberg. p. 188-199.

105. Ravindranath, D., et al., *QSPR Generalization of Activity Coefficient Models for Predicting Vapor–Liquid Equilibrium Behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

106. Deb, K., et al., *A Fast and Elitist Multiobjective Genetic Algorithm: Nsga-II.* Evolutionary Computation, IEEE Transactions on, 2002. **6**(2): p. 182-197.

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

108. Merkwirth, C., et al., *Ensemble Methods for Classification in Cheminformatics.* Journal of Chemical Information and Computer Sciences, 2004. **44**(6): p. 1971-1978.

CHAPTER 3

A NON-LINEAR QSPR MODEL FOR OCTANOL-WATER PARTITION COEFFICIENT

3.1. Introduction

The octanol-water partition coefficient (K_{ow} or log K_{ow}) is used to denote the lipophilicity of a molecule. It is a thermophysical property that expresses the ratio of concentrations of a compound in coexisting phases comprised of octanol and water. Leo et al. [1] were the first authors to review comprehensively the octanol-water partition system and its applications. Although lipophilicity has many uses, its importance in the field of drug delivery is paramount. Several researchers have studied the effects of lipophilicity on the biological activity of drugs [2-6] and on their transport properties [7-9]. Their findings and many other studies indicate the importance of lipophilicity in the evaluation of new drugs or prodrugs [10].

The experimental procedures for measuring the K_{ow} values are detailed by Sangster [11]; however, only a relatively small percentage of existing commercial chemicals have been tested experimentally for their K_{ow} values [10]. This is particularly so for highly hydrophobic compounds with K_{ow} values > 10^6. The low solubility of these compounds in the aqueous-rich phase renders K_{ow} measurements difficult to undertake, and therefore, few values exist at this range. Further, K_{ow} experiments are, in general, time and labor intensive, and they are impractical to carry out for the large number of potential drugs identified in the developmental stages of drug discovery. As such, a need exists for reliable predictive models to determine accurately, K_{ow} values without the need for experimentation. Therefore, this work focuses on the following objectives:

1. Develop an accurate non-linear QSPR model to predict the K_{ow} values using a database made up of diverse set of compounds.
2. Compare the current modeling approach with existing modeling approaches in the literature, on common external set data. This would further establish the efficacy of the modeling approach used in the work.

3.2. State of the Art in Predicting K_{ow} Values

Although rudimentary predictive models for K_{ow} were established nearly half a century ago, advances in computational capabilities has led to the more recent development of a diverse variety of models. An article by Mannhold et al. [12] lists and compares the state-of-the-art models available currently for K_{ow}. These models for K_{ow} can be broadly classified as:

1. Fragment-based methods that divide the molecule into various fragments (either at the molecular or atomic level), and then sum the contributions of these individual fragments to provide the final value. Examples include KowWIN based on the algorithms developed by Meylan and Howard [10], CLOGP [13], and Ghose-Crippen models [14-16].
2. Molecular-property based methods that utilize characteristics of the entire molecule to predict for K_{ow}. These characteristics are usually referred to as molecular descriptors and are normally calculated from the three-dimensional structure of the molecule or from the topology of the molecule. Common examples of this model classification are

QLOGP based on the algorithm by Bodor and Buchwald [17], GBLOGP by Totrov [18], and ALOGPS [12, 19-21].

Fragment-based methods have been highly successful in developing accurate models for K_{ow}. After comparing the performance of currently available methods on three different datasets, Mannhold et al. [12] ranked the best methods, which included fragment-based methods like AB/LogP [22], CLOGP [13] and KowWIN [10]. The fragment-based methods rely solely on the two-dimensional structure of the molecules; as such, thousands of molecules typically generated in a virtual molecule design system can be processed in a relatively short period of time. A disadvantage of the fragment-based methods is the lack of parameter values when a structure cannot be decomposed to sub-structures for which the fragment values are available [23]. Additionally, these methods cannot be used to attach any physical significance to the structural factors affecting the value of the partition coefficient.

The molecular descriptor based methods do not need additional correction factors and provide better physical insight into the factors affecting the partition coefficient. However, finding the optimal three-dimensional structure of the molecule is a time consuming task, which limits the ability of these methods to handle large numbers of molecules in a reasonable amount of time. Most molecular-property based methods do not use all available molecular descriptors, but limit themselves instead to a small subset of descriptors, which have already been proven to be effective like E-state indices [24, 25] by ALOGPS [19, 20] and VLOGP [26], topological descriptors by TLOGP [27] and molecular size and H-bonding descriptors by QLOGP [17]. As a result, these models provide little insight in the identification of specific molecular properties that affect the partition coefficient.

In this work, we propose a non-linear quantitative structure-property relationship (QSPR) model for predicting the octanol-water partition coefficient. The basic premise of a QSPR methodology asserts that a thermophysical property to some degree is a function of its structural attributes [28]. QSPR models have been able to predict successfully a number of thermophysical properties such as normal boiling point [29-31], melting point [32-34], refractive index [35, 36] and glass transition temperature [37]. All the molecular-property based models for K_{ow} discussed previously represent different types of QSPR models, where a variety of structural descriptors are employed, including constitutional, topological, geometrical, electrostatic, quantum-chemical and/or thermodynamic descriptors.

Our model utilizes all the descriptors (including three-dimensional descriptors) of the molecule generated using CODESSA PRO [38]. Nearly 800 descriptors belonging to various classes like constitutional, topological, geometrical, electrostatic, quantum-chemical and thermodynamic were generated for each molecule. Using a wrapper-based algorithm, we determined that 50 descriptors resulted in an accurate model for the partition coefficient; nevertheless, the pruning of descriptors from 800 to 50 is not a trivial task, particularly in non-linear QSPR modeling. A recent review article by Dudek et al. [39] summarizes the different types of descriptor pruning techniques in use. In general, these methods fall into two categories:

1. Filter-based methods: These methods are implemented before the mapping of the structural attributes to the property of interest. Some examples include methods based on mutual information [40].
2. Wrapper-based methods: In these methods, the selection of best descriptors is undertaken along with the mapping or the learning step. The error in the mapped model is used as the judging criterion for the selection process. Some common examples include sequential forward selection, sequential backward elimination, genetic algorithms (GA), simulated annealing (SA) and ant colony optimization.

The filter-based methods are quite fast but may not result in the selection of the best subset of descriptors. A more reliable descriptor subset selection procedure is observed with wrapper-based methods, but these methods are slow in the final stages of the algorithm. This is particularly true of the stochastic methods such as GA, SA and ant colony optimization. In this work, we propose a novel wrapper-based algorithm for the selection of the best subset of descriptors using an evolutionary algorithm called differential evolution (DE) [41], which uses artificial neural networks as the non-linear mapping functions. DE has been proven to be as effective as or better than GA and also is easier to implement [42, 43].

3.3. QSPR Methodology

The development of a QSPR model involves the following series of steps: (a) data set generation, (b) descriptor calculation, (c) descriptor reduction and model development, and (d) model validation. These elements are described below.

3.3.1. Data Set Generation: Experimental octanol-water partition coefficient (log K_{ow}) values were taken from the PhysProp database [44] by Syracuse Research. This database had experimental log K_{ow} values listed for 13,553 compounds. Of these, 350 were not included in the modeling effort because they are either metal containing compounds, inorganics (compounds without a carbon atom), or isomers of another molecule. Also, for 33 compounds, ChemBioDraw Ultra 11.0 (ChemBioOffice 2008 suite) [45] was unable to generate two-dimensional (2D) structures from the molecule names in the PhysProp database [44], thus they were excluded. As such, a total of 13,170 molecules were selected for further analysis. However, only 11,308 molecules could be optimized for their most favorable (lowest energy) three-dimensional conformation using our automated procedure. While characterization of this large database is beyond the scope of this work, Hansch et al. [46] have stated that the Log K_{ow} values can be experimentally determined to an average deviation of ±0.05 for most solutes. For solutes that have a Log K_{ow} value lower than -3 and greater than 6, as well as solutes that are relatively insensitive to gas chromatography, the average deviation expected is ±0.1.

3.3.2. Descriptor Calculation: Descriptor calculation requires a series of steps common to all QSPR models. In the current work, ChemBioDraw Ultra 11.0 [45] was used to generate two-dimensional (2D) structures for the molecules in the data set and stored as .cdx files. These 2D structures were then used to generate three-dimensional (3D) structures. Each 2D structure can be translated into a large number of 3D conformations; however, only the conformation with the lowest conformational energy is considered representative of the natural state of the molecule. When considering the multiple minima of the total-energy curve, finding this 3D conformation is not a trivial task. Chem3D Pro 11.0 (CambridgeSoft 2008 suite) [45] is a commercial software used commonly to minimize the total-energy of a 3D conformation; however, the software is not guaranteed to find the global minimum energy conformation. Therefore, an optimization using several initial 3D conformations will have an improved chance of locating the global minimum. This process is not integrated into Chem3D Pro 11.0 and requires the 3D structure to be manually reinitialized to a different starting conformation each time before optimization. This operation not only places an increased time and effort burden on the user, but it is not a reliable method of locating the global minimum. To alleviate this problem, we have used an automated strategy for identifying the 3D conformation with the least total energy. Chem3D Pro 11.0 was used as the optimizing engine, but it was controlled using its Component Object Model (COM) interface with Microsoft Visual Studio 8 (2005) as the back-end. The 3D structures were further optimized using AMPAC 6.0 [47], and the final optimized structures were provided to CODESSA PRO [38] for descriptor calculation. CODESSA PRO has the capability to generate over 800 descriptors; however, due to structural complexity, this number may be lower for a particular structure and for such structures the missing descriptors were assigned a zero value.

3.3.3. 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 happen simultaneously. Specifically, an evolutionary algorithm called 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 whose size is denoted as Desc_Sz. The method begins with an initial population of single or doubles hidden layered ANNs (individuals). The ANNs in the initial population are assigned random descriptor subsets as inputs. These ANNs then undergo mutation and cross-over operations over successive generations. In each generation, the ANNs that can accurately predict the target property are favored over inaccurate ANNs to remain in the population. Therefore, ANNs in the later generations are, on average, closer to the global minimum of the objective function. The subsequent discussion will be a brief introduction to DE and ANNs followed by details on the actual descriptor reduction algorithm employed in the current study

Differential Evolution (DE): DE is another simple stochastic optimization algorithm similar to GA and was proposed by Price and Storn [41] in 1994. The major difference between GA and DE is that the former uses probability distribution for selection of parents, while in the latter trail vectors are generated. This makes the DE algorithm self-organizing by reducing the number of parameters that need to be pre-set by the user. The basic DE algorithm for minimization has the following steps:

1. Randomly generate an initial population of a fixed size, N. Usually the population size is heuristically determined based on the dimensionality of the fitness surface denoted as n.
2. Do the following for successive generations until some stopping criterion is met:

   For each vector **x** in the population the following steps are conducted:

   a. Choose three different individuals **a**, **b** and **c** that are different from **x**.
   b. Pick a random integer R between 1 to n.
   c. Generate a trial vector **y** $=\{y_1, y_2, \ldots y_i, \ldots y_n\}$ by iterating over each i from 1 to n

      - Generate a uniformly distributed random number r between 0 and 1.

- If i=R or if r < CR (cross-over number), then y_i = a_i + F*(b_i - c_i), else y_i = x_i, where F is the mutation factor.

d. If the trail vector **y** has lower objective function than the original vector **x**, then replace **x** with **y.**

DE has been successfully applied to various optimization problems such as heat exchanger network synthesis [48], reservoir system optimization [49], design of temperature profiles for fermentation processes [50] and image pixel clustering [51]. Despite its popularity in other optimization fields, DE has been applied to few QSPR studies to date, which include prediction of atomic charges by Ouyang et al. [52] and insulin skin permeability in the presence of chemical penetration enhancers by the Oklahoma State University group [53].

Artificial Neural Networks (ANNs): Artificial neural networks are inspired by the brain and the interconnections between neurons, which form a complex network where electrical and other types of signals are exchanged to facilitate functioning of the brain. Although, much of neuronal function in the brain is still unclear, researchers have been able to develop ANNs as limited and simplified models for recreating intelligence artificially. Different types of ANNs exist based on architecture, but in view of the current work, only feed-forward ANNs are relevant and any future reference to ANNs in the current work refers to feed-forward ANNs.

In a feed-forward ANN, information travels only in the forward direction from the input nodes to the output nodes. The different layers are connected using weights and biases which represent the strength of the signal between the different nodes, and these are updated during the learning phase of the algorithm to minimize the error between the network outputs and the network targets.

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. For most applications, using just one hidden layer is adequate; however, choosing the right number of neurons in this hidden layer may not be straightforward. Choosing too few hidden neurons might lead to an ANN that is not flexible enough to encapsulate the complexity of the data and choosing too many may lead to over-fitting and poor generalization. There is no theoretical basis for choosing the number of hidden neurons and hidden layers, and most researchers use trial and error for selection of the architecture leading to the best performance. Since a two hidden-layer network is capable of reasonable approximation of any non-linear function, the maximum number of hidden layers in the current work was limited to two [54]. In addition, for each ANN, the ratio of the number of training data to the number of adjustable weights and biases was ensured to be always greater than two [55]. This was done as a precaution against over-fitting to the training data.

Once, the network architecture has been specified, an ANN is trained on known data before use as a predictive tool. The most popular learning algorithm for feed-forward ANNs is the back-propagation algorithm proposed by Rumelhart et al. [56], which consists of a forward propagation step and a weight update step that are repeated until the network performance is satisfactory. The network weights and biases can be updated using several algorithms, but the most popular are the gradient-descent and Levenberg-

Marquardt algorithms [57]. However, these algorithms do not guarantee location of the global minimum; thus, multiple initializations of the program are often necessary. Iyer and Rhinehart [58] have proposed a multiple initialization method to increase the probability of locating the global minimum. This method is built into the descriptor reduction algorithm used in the current work.

Another important issue associated with ANN training is over-fitting, which results in poor predictive capability. Although several methods for avoiding over-fitting exist, in the current work over-fitting is avoided by application of a training set (T) and an internal validation set (V) with an early-stopping method [59, 60]. The validation error normally decreases during the initial phase of training, as does the training set error; however, when the network begins to over-fit the data, the error in the validation set typically begins to rise. When the validation error increases for a specified number of iterations, the training is stopped, and the weights and biases at the minimum validation error are retained. Although the early-stopping algorithm is easy to understand and implement, choosing the right training and internal validation sets is not straightforward. Ideally, the training set should be representative of the entire data set, and each datum in the validation set should correspond to at least one training datum. Several methods exist in the literature for allocation of the data such as random division [61, 62], self-organizing maps (SOMs) [63, 64], Kennard-Stone design [65] and sphere exclusion algorithm [66]. In the current work, the random sphere exclusion algorithm [67] with a specified dissimilarity level of was used to divide the data into training and validation sets. The dissimilarity level was chosen to divide the training and validation sets in the range 70-80% and 20-30% of the data samples, respectively.

During training of the ANNs in the current work, the inputs and targets (the experimental values of the property that need to be modeled), are normalized to have zero mean and unity standard deviation, which ensures that exceptionally large-valued descriptors or targets do not bias the network. The Nguyen-Widrow algorithm is used to initialize weights and biases, which are updated using the Levenberg-Marquardt optimization technique.

Genetic Representation: A good genetic representation of the solution domain is an important step in developing an efficient evolutionary algorithm. Binary representation is most widely used due to the direct encoding technique for most problems and the applicability for crossover dependent evolutionary algorithms like GA and DE [68]. Real-valued representations on the other hand are better suited for algorithms like evolutionary programming (EP) that are dependent upon mutation as the major evolutionary operator. In the current work, the solution space is comprised of single hidden layer ANNs with all possible molecular descriptor subsets of a fixed size, ND, as inputs, which are determined by the user at the start of the program. The number of hidden neurons, NH, in these ANNs lies between a minimum of two and a maximum usually fixed at three times the value of ND. 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. An example of three sample chromosomes with 100 original descriptors (Desc_Sz = 100) and a model size, ND equal to 5, is shown in Table 3.1. Each chromosome is made up of five genes, where each gene represents a descriptor that is used as an input variable to an associated ANN, which is subsequently trained to predict the target property.

Binary representation of the chromosomes entails large memory requirements, and the algorithm requires longer convergence times to reach a global minimum when compared with real-valued representations. The above considerations provide the basis for using real-valued chromosomes in the current work.

The Objective Function: Another major aspect of an evolutionary algorithm is the choice of a suitable objective function. In the current work, a wrapper-based modeling approach is used for simultaneous descriptor reduction and non-linear model development using ANNs. 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. In general, the objective function for the second tier of optimization in a wrapper-based descriptor reduction approach must be selected such that a good estimate of the performance of the underlying linear or non-liner models is achieved. The objective functions chosen for linear QSPR models typically maximize statistical measures such as the correlation coefficient (R^2) [69], adjusted R^2 and q^2 [70], and Akaike information content [71]. For non-linear models, the root-mean- squared error (RMSE) [72, 73] and absolute average deviation (AAD) [74] are used. In the current work, the training set RMSE was used as the objective function, *F*:

$$\mathrm{F} = \mathrm{RMSE_T} \tag{3.1}$$

With proper selection of an objective function, an algorithm can be applied which searches for the set of descriptors resulting in an ANN that results in a minimum value of the objective function.

The Algorithm: The flow chart for the algorithm is given in Figure 3.1. Before execution of the algorithm, the following parameters are set by the user: (a) desired number of descriptors in the model, ND, (b) population size, Pop_Sz, which is usually set at 400, (c) mutation factor, F, is set at 0.75, and (d) crossover factor, CR, is set at 0.8. The algorithm has an initialization process that happens once. The individual ANNs in a parent population denoted as ‘D’ are initialized with random descriptor subsets of size ND. The j^{th} gene in the i^{th} individual is represented as D(i,j). The number of hidden layers and the number of hidden neurons for each ANN is randomly initialized. The ANNs are then trained using a back-propagation, with Levenberg-Marquardt weight updating algorithm, resulting in network weights that minimize the $\mathrm{RMSE_T}$ 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 $\mathrm{RMSE_V}$ decreases for six successive training iterations. The objective function F for the i^{th} individual in population ‘D’ is denoted as $F\{\mathrm{D(i)}\}$. Population ‘D’ then undergoes the DE operations in a single iteration of the algorithm. Specifically, mutated population ‘TM’ is generated by combining the genes from three different individuals in population ‘D’. This operation is similar to the mutation operation in the traditional DE algorithm. Next, the mutated population ‘TM’ and the parent population ‘D’ are recombined using the recombination operation of the traditional DE algorithm. The recombined population is denoted as population ‘M’. The

ANNs in ‘M’ undergo training and the f{M(i)} 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’. If *f*{M(i)} is lower than *f*{D(i)}, then M(i) is considered fitter than D(i), and therefore, M(i) replaces D(i) in population ’D’. This is denoted as ‘individual competition.’

Creating Ensembles for Final Predictions: ANNs are known to be highly 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 [75, 76]. The ANNs in the ensemble can differ with respect to (a) the training data, (b) weights between the different nodes, (c) the number of hidden layers and neurons, and finally (d) based on the input descriptors. In the current work, an ensemble of neural networks was created using networks with the same architecture and inputs as the best network in the final DE population, but differing in the values of the weights between the different layers.

External Validation: In a recent article, Tropsha et al. [55] emphasized the need to validate QSPR models using external data sets. Also, Mannhold et al. [12] have recently compared the various K_{ow} models in the literature using an external public database by Avdeef [77], which consisted of 266 molecules in total. Of these, 214 molecules were similar to the molecules in the PhysProp database [44] and were classified as star set molecules. The remaining 43 molecules were classified as non-star set molecules. Since these molecules were not reported in the PhysProp dataset [44], they can be used as an external validation set for comparing our model with other models in the literature. The performance of the current model on this new dataset would indicate the generalization capability of the final model.

3.4. Results

Thirty, 40, 50 and 60 descriptors were evaluated as inputs to the ANNs. The RMSE values on the training set generally increased with increase in number of descriptors, but no significant difference was observed between the models developed using 50 and 60 descriptors. Therefore, for simplicity, 50 descriptor models were used in the current study. From the DE algorithm, 50-33-35-1 neural network architecture was found to result in the least RMSE for the training set data. Neural networks with the same input data and architecture as the best network identified using the DE + NN algorithm were trained with different random initial weights. Of the networks generated, the five networks resulting in the least RMSE values in the training set were chosen, and their weights were recorded. The final predictions are calculated as a simple average of the individual predictions by these five networks. The addition of further networks did not improve significantly the overall RMSE of the training set.

The RMSE values for the training set and validation set data for the five networks and the resultant average network are shown in Table 3.2. In subsequent discussions, the prediction results are from the ensemble network, which is the average of the predictions by the five best networks. Comparisons of the experimental and predicted K_{ow} values for the training and validation sets are shown in Figure 3.2 and Figure 3.3, respectively. The correlation coefficients (R^2) between the experimental and predicted values for the training and the validation sets are 0.96 and 0.88, respectively. The ensemble RMSE values for training and validation sets are 0.28 and 0.38, respectively, while the ensemble MAE values for the training and validation sets are 0.20 and 0.34, respectively. A histogram of the residuals (no figure shown) was plotted, and the distribution of the residuals around zero was found to be similar to a normal distribution for both the training set and validation set data. The model predictions from this work for the star and the non-star sets from the Avdeef test set are shown in Figures 3.4 and 3.5. The RMSE for the star-set molecules and the non-star set molecules were calculated to be 0.57 and 1.01, respectively. A comparison of the performance of the current model on the Avdeef set with those of the best models in the literature (as reported by Mannhold et al.[12]) is provided in Table 3.3. The best set of descriptors identified in the current work is shown in Table 3.4.

3.5. Discussion

The best network ensemble identified is a combination of two layered networks with fifty input descriptors each. This ensemble was able to account for 96% and 88% of the variation in the training (Figure 3.2) and validation (Figure 3.3) sets, respectively. The statistics presented in Table 3.2 justify the use of a neural network ensemble as compared to individual networks. Network 1 in the ensemble had the lowest RMSE value of 0.329 for the training set, which is nearly 21% higher than the corresponding RMSE value of the entire ensemble (Table 3.2). The different networks in the ensemble had differing weights, which produces varying predictions in different regions of the input space. This further supports the advantage of ensemble networks over an individual network.

The DE + NN approach employed in the current work has been successful in identifying the best descriptors describing the octanol-water partitioning ability of molecules. This is evident from the performance of the model on the external validation set (Table 3.3) where the RMSE of the star-set was 0.57 and that of the non-star set was 1.01. Our model performs better than 28 of the 34 K_{ow} models tested by Mannhold et al. [12] listed in Table 3.3. Of the six models that outperform the current model, four models were developed using fragment-based approaches, which cannot be applied to molecules with unknown fragments. ALOGPS [12, 19, 20] and S+LogP [78] are the only molecular descriptor-based methods that perform better than the current model. The ALOGPS [12, 19, 20] model was developed using the DIPPR database and 75 E-state indices as inputs. The marginally better performance of the ALOGPS model could be attributed to the larger number of input descriptors when compared to the current model, which only employs 50 descriptors. S+LogP [78] was developed using 217 input descriptors and was trained using the same database from which the star-set molecules were extracted. Therefore, the better predictions by the S+LogP model could be due to the inclusion of the same or similar molecules in the training set. Also to note, this external validation data set is limited in its size and therefore cannot truly be used to test the generalization ability of the current model. Mannhold et al. [12] had also tested the 34 models using a Pfizer proprietary dataset of aroud 96,000 compounds. After careful analysis of the performance of the different models, Mannhold et al. [12] reported that, the molecular descriptor-based methods consistently outperform the fragment-based methods, with ALOGPS and S+LogP being the best methods. Testing on such large external data sets could clearly establish the generalization ability of the models built using the approach described herein. Also, the current model was built using only descriptors generated by CODESSA [38]. Other descriptors such as functional-group descriptors and WHIM descriptors (molecular descriptors obtained as statistical indices of the atoms projected onto the 3 principal components obtained from weighted covariance matrices of the atomic coordinates) available in DRAGON [79] could be employed in the future to further improve the model. The model in the current work is purely empirical and is therefore limited in the ability to generalize to systems beyond its applicability domain. To remedy this limitation, a theory-based K_{ow} model could be built that utilizes QSPR generalized activity coefficient models to predict the solubility of solute molecules in each of the two phases of the octanol-water system.

Table 3.4 lists the best set of descriptors for the final ensembles, the majority of which are quantum-chemical descriptors related to the presence of hetero-atoms in the molecule. These descriptors account for the electronegative effects of the hetero-atoms present in the molecule. These descriptors were found to be important in other K_{ow} models in the literature as well. Specifically, the MLOGP model by Moriguchi et al. [80] is developed using 11 descriptors, the majority of which describe the presence or the electronegative effects of hetero-atoms in the molecule. Also, these Moriguchi [80] descriptors, with the addition of descriptors that account for the charge and polarizability of the molecule, are an important part of the S+LogP model [78].

3.6. Conclusions

1. A hybrid algorithm that combines differential evolution algorithms (DE) and artificial neural networks (ANNs) provides an accurate predictive model for octanol-water partition coefficient that compares favorably with viable literature models.
2. An ensemble of neural networks that differ in the values of the weights between layers provides improved predictive generalizations compared to a single network and

yielded predictions with an R^2 of 0.98 and 0.96 for the training and validation sets of a database involving 11,308 molecules.

3. The RMSE on the external test set for the current model was 0.57 on star set molecules and 1.01 on non-star set molecules. These results compare favorably with the other molecular-descriptor-based method such as ALOGPS [12, 19-21]. The current model performs worse than a few fragment-based methods in the literature; however, unlike the fragment-based methods, the current model can be applied to molecules with unknown fragments.
4. The resulting model from this work can be used to predict *a priori* the octanol-water partition coefficient of new molecules.

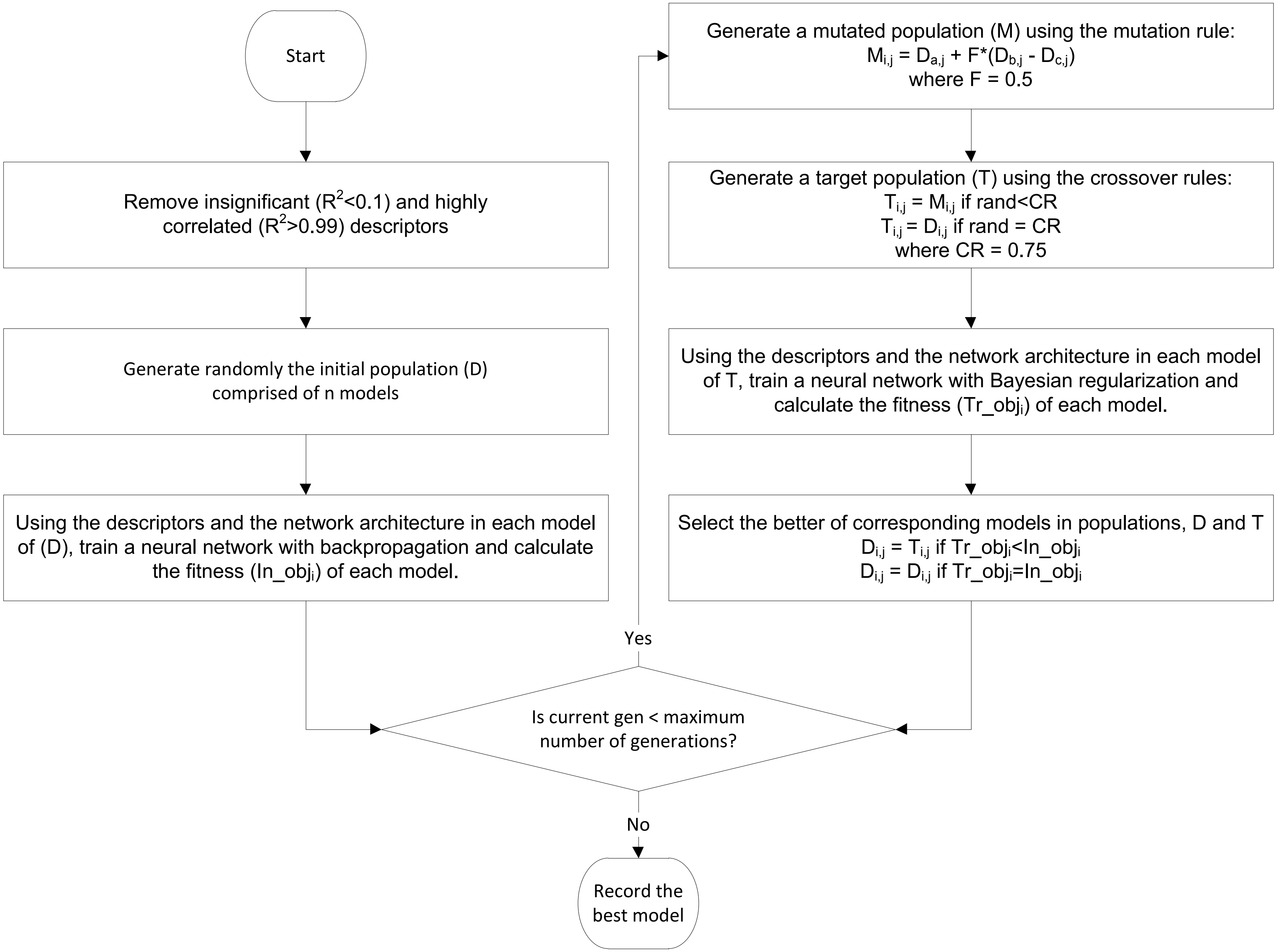

Start
Remove insignificant (R^2<0.1) and highly correlated (R^2>0.99) descriptors
Generate randomly the initial population (D) comprised of n models
Using the descriptors and the network architecture in each model of (D), train a neural network with backpropagation and calculate the fitness (In_obj_i) of each model.
Is current gen < maximum number of generations?
Yes
No
Record the best model
Generate a mutated population (M) using the mutation rule:
$M_{i,j} = D_{a,j} + F*(D_{b,j} - D_{c,j})$
where F = 0.5
Generate a target population (T) using the crossover rules:
$T_{i,j} = M_{i,j}$ if rand<CR
$T_{i,j} = D_{i,j}$ if rand = CR
where CR = 0.75
Using the descriptors and the network architecture in each model of T, train a neural network with Bayesian regularization and calculate the fitness (Tr_obj_i) of each model.
Select the better of corresponding models in populations, D and T
$D_{i,j} = T_{i,j}$ if Tr_obj_i<In_obj_i
$D_{i,j} = D_{i,j}$ if Tr_obj_i=In_obj_i

Figure 3.1: Flowchart for the differential evolution algorithm

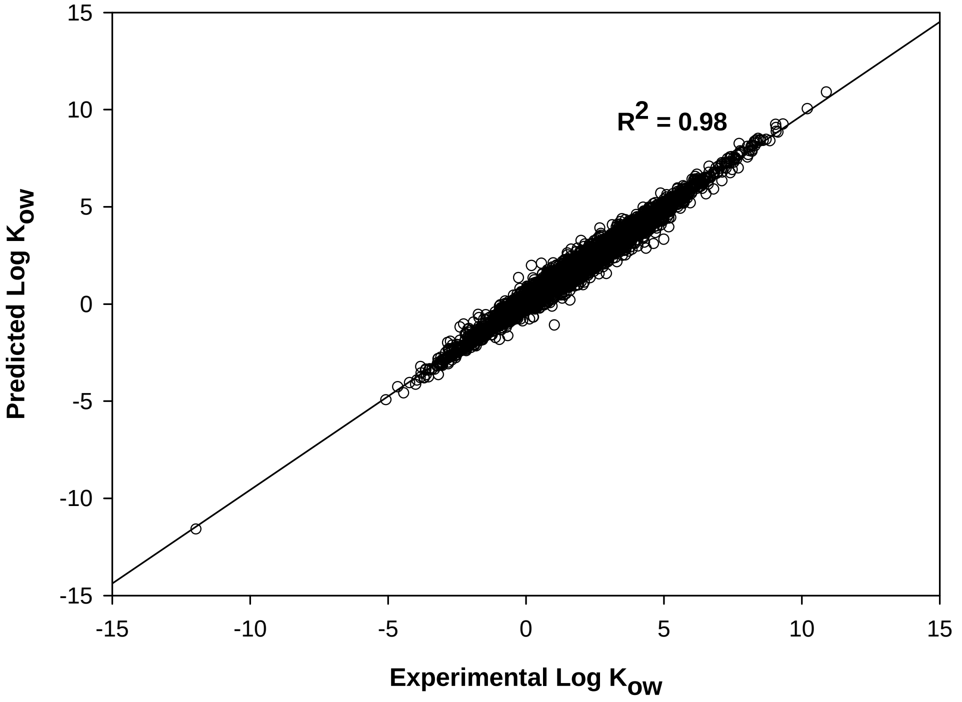


Figure 3.2: Comparison between the experimental and predicted values of Log K_{ow} in the training set

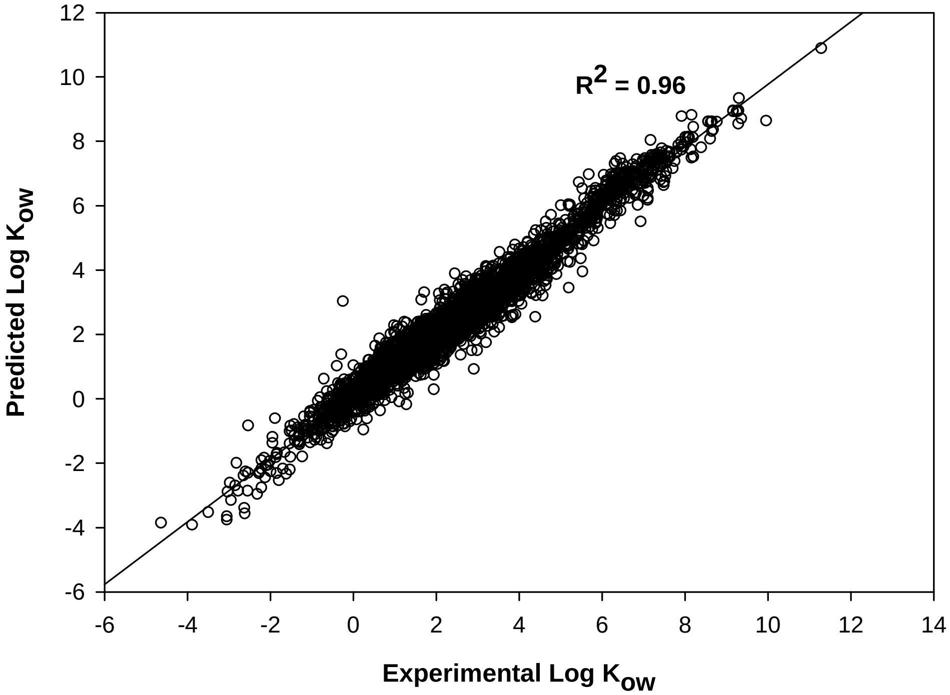


Figure 3.3: Comparison between the experimental and predicted values of Log K_{ow} in the validation set

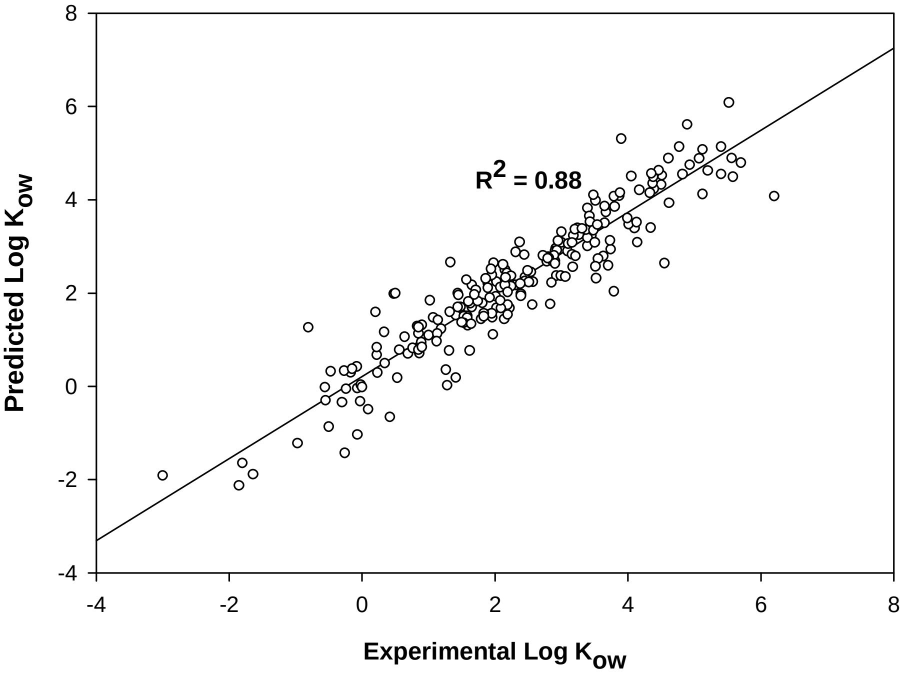


Figure 3.4: Comparison between the experimental and predicted values of Log K_{ow} for the star set compounds in the external validation set

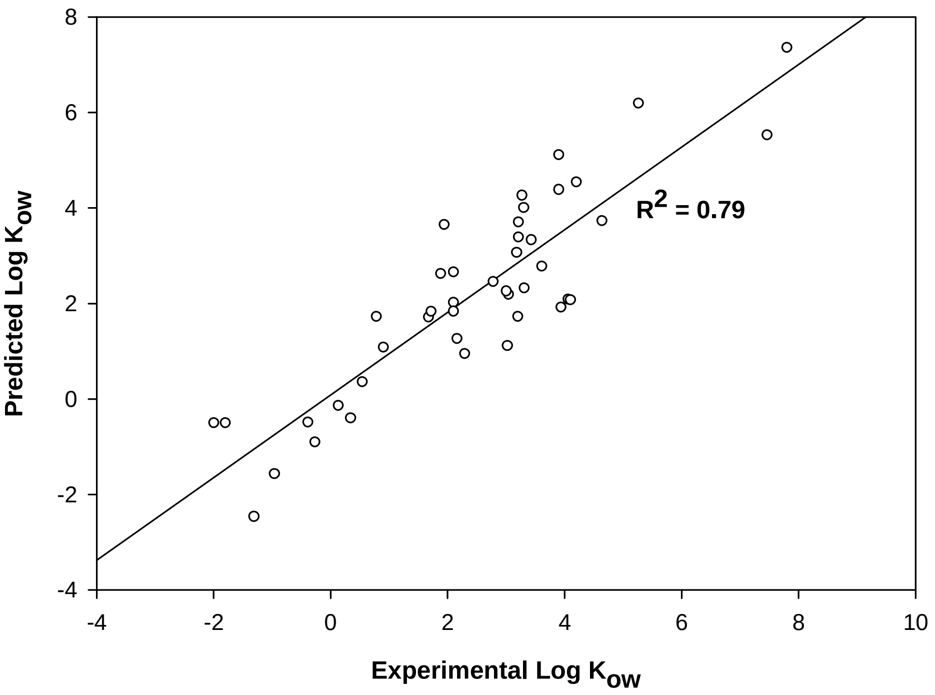


Figure 3.5: Comparison between the experimental and predicted values of Log K_{ow} for the non-star set compounds in the external validation set

Table 3.1: Three sample chromosomes of size five, chosen from a set of 100 descriptors

Chromosome #	Input Descriptor 1	Input Descriptor 2	Input Descriptor 3	Input Descriptor 4	Input Descriptor 5
Chromosome 1	23	45	54	3	98
Chromosome 2	23	49	22	9	67
Chromosome 3	34	44	1	7	100

Table 3.2: Training and validation set root-mean-squared error (RMSE) values for the five best networks and the ensemble

	Network 1	Network 2	Network 3	Network 4	Network 5	Ensemble
Training RMSE	0.329	0.343	0.345	0.348	0.350	0.279
Validation RMSE	0.447	0.450	0.446	0.452	0.462	0.380

Table 3.3: A comparison of the ensemble predictions with available literature models on the Avdeef test set as determined by Mannhold et al. [12]

Method	Star Set (223 molecules)				Non-Star set (43 molecules)			
	RMSE	% of Molecules within Error Range			RMSE	% of Molecules within Error Range		
		< 0.5	0.5-1	> 1		< 0.5	0.5-1	> 1
AB/LogP	0.41	84	12	4	1.00	42	23	35
S+logP	0.45	76	22	3	0.87	40	35	26
ACD/LogP	0.50	75	17	7	1.00	44	33	23
CLOGP	0.52	74	20	6	0.91	47	28	26
VLOGP OPS	0.52	64	21	7	1.07	47	28	26
ALOGPS	0.53	71	23	6	0.82	33	28	26
This work	*0.57*	*71*	*21*	*8*	*1.01*	*37*	*35*	*28*
MiLogP	0.57	69	22	9	0.86	49	30	21
XLOGP3	0.62	60	30	10	0.89	47	23	30
KowWIN	0.64	68	21	11	1.05	40	30	30
CSLogP	0.65	66	22	12	0.93	58	19	23
ALOGP	0.69	60	25	16	0.92	28	40	33
MolLogP	0.69	61	25	14	0.93	40	35	26
ALOGP98	0.70	61	26	13	1.00	30	37	33
OsirisP	0.71	59	26	16	0.94	42	26	33
VLOGP	0.72	65	22	14	1.13	40	28	33
TLOGP	0.74	67	16	13	1.12	30	37	30
ABSOLV	0.75	53	30	17	1.02	49	28	23
QikProp	0.77	53	30	17	1.24	40	26	35
QuantlogP	0.80	47	30	22	1.17	35	26	40

Table 3.3 (cont'd): A comparison of the ensemble predictions with available literature models on the Avdeef test set as determined by Mannhold et al. [12]

Method	Star Set (223 molecules)				Non-Star set (43 molecules)			
	RMSE	% of Molecules within Error Range			RMSE	% of Molecules within Error Range		
		< 0.5	0.5-1	> 1		< 0.5	0.5-1	> 1
SLIPPER-2002	0.80	62	22	15	1.16	35	23	42
COSMOFrag	0.84	48	26	19	1.23	26	40	33
XLOGP2	0.87	57	22	20	1.16	35	23	42
QLOGP	0.96	48	26	25	1.42	21	26	53
VEGA	1.04	47	27	26	1.24	28	30	42
CLIP	1.05	41	25	30	1.54	33	9	49
LSER	1.07	44	26	30	1.26	35	16	49
MLOGP (Sim+)	1.26	38	30	33	1.56	26	28	47
NC+NHET	1.35	29	26	45	1.71	19	16	65
SPARC	1.36	45	22	32	1.70	28	21	49
MLOGP (Dragon)	1.52	39	26	35	2.45	23	30	47
LSER UFZ	1.60	36	23	41	2.79	19	12	67
AAM	1.62	22	24	53	2.10	19	28	53
HINT	1.8	34	22	44	2.72	30	5	65
GBLOGP	1.98	32	26	42	1.75	19	16	65

Table 3.4: The list of the best set of descriptors identified in this work

No	Name of the Descriptor	Type of Descriptor
1	Number of I atoms	Constitutional
2	Randic index (order 3)	Topological
3	Number of N atoms	Constitutional
4	Relative number of F atoms	Constitutional
5	HACA-1 [Zefirov's PC]	Electrostatic
6	Min electroph. react. index for a Br atom	Quantum-chemical
7	Average Complementary Information content (order 1)	Topological
8	Number of N atoms	Constitutional
9	Complementary Information content (order 1)	Topological
10	Relative number of I atoms	Constitutional
11	Number of O atoms	Constitutional
12	Max net atomic charge for a H atom	Electrostatic
13	Number of atoms	Constitutional
14	Max bond order of a C atom	Quantum-chemical
15	Relative number of N atoms	Constitutional
16	Number of O atoms	Constitutional
17	Min resonance energy for a Br-C bond	Quantum-chemical
18	Information content (order 2)	Topological
19	min(#HA, #HD) [Quantum-Chemical PC]	Quantum-chemical
20	Vib heat capacity (300K)	Thermodynamic

Table 3.4 (cont'd): The list of the best set descriptors identified in this work

No	Name of the Descriptor	Type of Descriptor
21	Avg bond order of a O atom	Quantum-chemical
22	Min e-n attraction for a C-S bond	Quantum-chemical
23	Max partial charge for a N atom [Zefirov's PC]	Electrostatic
24	Max nucleoph. react. index for a O atom	Quantum-chemical
25	Max e-n attraction for a H-N bond	Quantum-chemical
26	count of H-donors sites [Zefirov's PC]	Electrostatic
27	Translational entropy (300K) / # of atoms	Thermodynamic
28	(1/6)X GAMMA polarizability (DIP)	Quantum-chemical
29	Min e-n attraction for a Cl-N bond	Quantum-chemical
30	DPSA-3 Difference in CPSAs (PPSA3-PNSA3) [Zefirov's PC]	Electrostatic
31	Min n-n repulsion for a Br-C bond	Quantum-chemical
32	Min electroph. react. index for a N atom	Quantum-chemical
33	Max n-n repulsion for a H-N bond	Quantum-chemical
34	Kier shape index (order 2)	Topological
35	Max e-n attraction for a C-S bond	Quantum-chemical
36	Max total interaction for a H-O bond	Quantum-chemical
37	Exch. eng. + e-e rep. for a C-H bond	Quantum-chemical
38	Min e-e repulsion for a S atom	Quantum-chemical
39	Max atomic state energy for a F atom	Quantum-chemical
40	Max total interaction for a F-P bond	Quantum-chemical

Table 3.4 (cont'd): The list of the best set descriptors identified in this work

No	Name of the Descriptor	Type of Descriptor
41	Min e-n attraction for a C-Cl bond	Quantum-chemical
42	Max resonance energy for a N-O bond	Quantum-chemical
43	Max e-n attraction for a H-P bond	Quantum-chemical
44	Exch. eng. + e-e rep. for a N-N bond	Quantum-chemical
45	Max e-n attraction for a N atom	Quantum-chemical
46	Min e-n attraction for a O-Si bond	Quantum-chemical
47	Max e-n attraction for a O atom	Quantum-chemical
48	Min n-n repulsion for a C-P bond	Quantum-chemical
49	Max total interaction for a F-P bond	Quantum-chemical
50	Min e-n attraction for a H-S bond	Quantum-chemical

REFERENCES

1. Leo, A., C. Hansch, and D. Elkins, *Partition Coefficients and Their Uses.* Chemical Reviews, 1971. **71**(6): p. 525-616.

2. Corwin, H. and M.C. John, *Lipophilic Character and Biological Activity of Drugs II: The Parabolic Case.* Journal of Pharmaceutical Sciences, 1973. **62**(1): p. 1-21.

3. Corwin, H. and J.D. William, III, *Linear Relationships between Lipophilic Character and Biological Activity of Drugs.* Journal of Pharmaceutical Sciences, 1972. **61**(1): p. 1-19.

4. Glave, W.R. and H. Corwin, *Relationship between Lipophilic Character and Anesthetic Activity.* Journal of Pharmaceutical Sciences, 1972. **61**(4): p. 589-591.

5. Renau, T.E., et al., *Effect of Lipophilicity at N-1 on Activity of Fluoroquinolones against Mycobacteria.* Journal of Medicinal Chemistry, 1995. **38**(15): p. 2974-2977.

6. Roland, J., *Parabolic Relationship between Lipophilicity and Biological Activity of Aliphatic Hydrocarbons, Ethers and Ketones after Intravenous Injections of Emulsion Formulations into Mice.* Acta Pharmacologica et Toxicologica, 1975. **37**(1): p. 56-64.

7. Wils, P., et al., *High Lipophilicity Decreases Drug Transport across Intestinal Epithelial Cells.* Journal of Pharmacology and Experimental Therapeutics, 1994. **269**(2): p. 654-658.

8. Waterhouse, R.N., *Determination of Lipophilicity and Its Use as a Predictor of Blood-Brain Barrier Penetration of Molecular Imaging Agents.* Molecular Imaging and Biology. **5**(6): p. 376-389.

9. van Bree, J.B., et al., *Characterization of an "in Vitro" Blood-Brain Barrier: Effects of Molecular Size and Lipophilicity on Cerebrovascular Endothelial Transport Rates of Drugs.* Journal of Pharmacology and Experimental Therapeutics, 1988. **247**(3): p. 1233-1239.

10. Meylan, W.M. and P.H. Howard, *Atom/Fragment Contribution Method for Estimating Octanol-Water Partition Coefficients.* Journal of Pharmaceutical Sciences, 1995. **84**(1): p. 83-92.

11. Sangster, J., *Octanol-Water Partition Coefficients : Fundamentals and Physical Chemistry*. 1997, New York: Wiley.

12. Mannhold, R., et al., *Calculation of Molecular Lipophilicity: State-of-the-Art and Comparison of Log P Methods on More Than 96,000 Compounds.* Journal of Pharmaceutical Sciences, 2009. **98**(3): p. 861-93.

13. Leo, A. and D. Hoekman, *Calculating Log P(Oct) with No Missing Fragments: The Problem of Estimating New Interaction Parameters.* Perspectives in Drug Discovery and Design, 2000. **18**(1): p. 19-38.

14. Arup, K.G., P. Avis, and M.C. Gordon, *Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships III: Modeling Hydrophobic Interactions.* Journal of Computational Chemistry, 1988. **9**(1): p. 80-90.

15. Arup, K.G. and M.C. Gordon, *Atomic Physicochemical Parameters for Three-Dimensional Structure-Directed Quantitative Structure-Activity Relationships I. Partition Coefficients as a Measure of Hydrophobicity.* Journal of Computational Chemistry, 1986. **7**(4): p. 565-577.

16. Ghose, A.K. and G.M. Crippen, *Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions.* Journal of Chemical Information and Computer Sciences, 2002. **27**(1): p. 21-35.

17. Bodor, N. and P. Buchwald, *Molecular Size Based Approach to Estimate Partition Properties for Organic Solutes.* Journal of Physical Chemistry B, 1997. **101**(17): p. 3404-3412.

18. Totrov, M., *Accurate and Efficient Generalized Born Model Based on Solvent Accessibility: Derivation and Application for Log P Octanol/Water Prediction and Flexible Peptide Docking.* Journal of Computational Chemistry, 2004. **25**(4): p. 609-619.

19. Tetko, I.V. and P. Bruneau, *Application of ALOGPS to Predict 1-Octanol/Water Distribution Coefficients, Logp, and Logd, of Astrazeneca in-House Database.* Journal of Pharmaceutical Sciences, 2004. **93**(12): p. 3103-3110.

20. Tetko, I.V. and G.I. Poda, *Application of ALOGPS 2.1 to Predict Log D Distribution Coefficient for Pfizer Proprietary Compounds.* Journal of Medicinal Chemistry, 2004. **47**(23): p. 5601-5604.

21. Tetko, I.V. and V.Y. Tanchuk, *Application of Associative Neural Networks for Prediction of Lipophilicity in ALOGPS 2.1 Program.* Journal of Chemical Information & Computer Science, 2002. **42**: p. 1136-1145.

22. Pranas, J., D. Remigijus, and P. Alanas, *Fragmental Methods in the Design of New Compounds. Applications of the Advanced Algorithm Builder.* Quantitative Structure-Activity Relationships, 2002. **21**(1): p. 23-37.

23. Lemke, T.L., et al., *Foye's Principles of Medicinal Chemistry*. 2008, Philadelphia: Lippincott Williams & Wilkins.

24. Hall, L.H. and L.B. Kier, *The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity.* Journal of Chemical Information and Computer Sciences, 2000. **40**(3): p. 784-791.

25. Hall, L.H. and L.B. Kier, *Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information.* Journal of Chemical Information and Computer Sciences, 2002. **35**(6): p. 1039-1045.

26. Gombar, V.K. and K. Enslein, *Assessment of N-Octanol/Water Partition Coefficient: When Is the Assessment Reliable?* Journal of Chemical Information and Computer Sciences, 1996. **36**(6): p. 1127-1134.

27. Junghans, M. and E. Pretsch, *Estimation of Partition Coefficients of Organic Compounds: Local Database Modeling with Uniform-Length Structure Descriptors.* Fresenius' Journal of Analytical Chemistry, 1997. **359**(1): p. 88-92.

28. Balaban, A.T., *From Chemical Topology to 3D Geometry.* Journal of Chemical Information and Computer Sciences, 1997. **37**: p. 645-650.

29. Wiener, H., *Structural Determination of Paraffin Boiling Points.* Journal of the American Chemical Society, 1947. **69**(1): p. 17-20.

30. Stanton, D.T., et al., *Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles.* Journal of Chemical Information and Computer Sciences, 1992. **32**: p. 306-316.

31. Katritzky, A.R., V.S. Lobanov, and M. Karelson, *Normal Boiling Points for Organic Compounds: Correlation and Prediction by a Quantitative Structure-Property Relationship.* Journal of Chemical Information and Computer Sciences, 1998. **38**: p. 28-41.

32. Katritzky, A.R., et al., *Prediction of Melting Points for the Substituted Benzenes: A QSPR Approach.* Journal of Chemical Information and Computer Sciences, 1997. **37**: p. 913-919.

33. Bergström, C.A.S., et al., *Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs.* Journal of Chemical Information and Computer Sciences, 2003. **43**(4): p. 1177-1185.

34. Karthikeyan, M., R.C. Glen, and A. Bender, *General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks.* Journal of Chemical Information and Modeling, 2005. **45**(3): p. 581-590.

35. Katritzky, A.R., S. Sild, and M. Karelson, *Correlation and Prediction of the Refractive Indices of Polymers by QSPR.* Journal of Chemical Information and Computer Sciences, 1998. **38**(6): p. 1171-1176.

36. Xu, J., et al., *Prediction of Refractive Indices of Linear Polymers by a Four-Descriptor QSPR Model.* Polymer, 2004. **45**(26): p. 8651-8659.

37. Katritzky, A.R., et al., *Prediction of Polymer Glass Transition Temperatures Using a General Quantitative Structure−Property Relationship Treatment.* Journal of Chemical Information and Computer Sciences, 1996. **36**(4): p. 879-884.

38. Semichem Inc., A.R. Katritzky, and M. Karelson, *Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA).* 1998: Shawnee, KS.

39. Dudek, A.Z., T. Arodz, and J. Galvez, *Computational Methods in Developing Quantitative Structure-Activity Relationships (QSAR): A Review.* Comb Chem High Throughput Screen, 2006. **9**(3): p. 213-28.

40. Tarca, L.A., B.P.A. Grandjean, and F. Larachi, *Feature Selection Methods for Multiphase Reactors Data Classification.* Industrial & Engineering Chemistry Research, 2005. **44**(4): p. 1073-1084.

41. Storn, R. *Differential Evolution (DE) for Continuous Function Optimization (an Algorithm by Kenneth Price and Rainer Storn).* [cited 2008 December]; Available from: http://www.icsi.berkeley.edu/~storn/code.html.

42. Aubert, B., et al., *Measurement of the Pseudoscalar Decay Constant fDs Using Charm-Tagged Events in e+e- Collisions at Square Root S=10.58 GeV.* Phys Rev Lett, 2007. **98**(14): p. 141801.

43. Robič, T. and B. Filipič, *Demo: Differential Evolution for Multiobjective Optimization Evolutionary Multi-Criterion Optimization*, C. Coello Coello, A. Hernández Aguirre, and E. Zitzler, Editors. 2005, Springer Berlin / Heidelberg. p. 520-533.

44. Syracuse Research Corporation, *Physical/Chemical Property Database (Physprop)*. 1994: Syracuse, NY.

45. *Chembiooffice 11.0*. 2008, CambridgeSoft.

46. Hansch, C., A. Leo, and D.H. Hoekman, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*. 1995: American Chemical Society.

47. Semichem Inc., *Ampac*. 1998: Shawnee, KS.

48. Yerramsetty, K.M. and C.V.S. Murty, *Synthesis of Cost-Optimal Heat Exchanger Networks Using Differential Evolution.* Computers & Chemical Engineering, 2008. **32**(8): p. 1861-1876.

49. Reddy, M.J. and D.N. Kumar, *Multiobjective Differential Evolution with Application to Reservoir System Optimization.* Journal of Computing in Civil Engineering, 2007. **21**(2): p. 136-146.

50. Oonsivilai, R. and A. Oonsivilai, *Differential Evolution Application in Temperature Profile of Fermenting Process.* WTOS, 2010. **9**(6): p. 618-628.

51. Das, S. and A. Konar, *Automatic Image Pixel Clustering with an Improved Differential Evolution.* Applied Soft Computing, 2009. **9**(1): p. 226-236.

52. Ouyang, Y., F. Ye, and Y. Liang, *A Modified Electronegativity Equalization Method for Fast and Accurate Calculation of Atomic Charges in Large Biological Molecules.* Physical Chemistry Chemical Physics, 2009. **11**(29): p. 6082-6089.

53. Yerramsetty, K.M., et al., *A Skin Permeability Model of Insulin in the Presence of Chemical Penetration Enhancer.* International Journal of Pharmaceutics, 2010. **388**(1-2): p. 13-23.

54. Hornik, K., M. Stinchcombe, and H. White, *Multilayer Feedforward Networks Are Universal Approximators.* Neural Networks, 1989. **2**(5): p. 359-366.

55. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

56. Rumelhart, D.E., 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.

57. Hagan, M.T., H.B. Demuth, and M.H. Beale, *Neural Network Design*. 1996, Boston ; London: PWS Pub. 1 v. (various pagings).

58. Iyer, M.S. and R.R. Rhinehart, *A Method to Determine the Required Number of Neural-Network Training Repetitions.* IEEE transactions on neural networks, 1999. **10**(2): p. 427-432.

59. Prechelt, L., *Automatic Early Stopping Using Cross Validation: Quantifying the Criteria.* Neural Networks, 1998. **11**(4): p. 761-767.

60. Rich, C., L. Steve, and G. Lee, *Overfitting in Neural Nets: Backpropagation, Conjugate Gradient, and Early Stopping*. 2000.

61. Wu, W., et al., *Artificial Neural Networks in Classification of NIR Spectral Data: Design of the Training Set.* Chemometrics and Intelligent Laboratory Systems, 1996. **33**(1): p. 35-46.

62. Yasri, A. and D. Hartsough, *Toward an Optimal Procedure for Variable Selection and QSAR Model Building.* Journal of Chemical Information and Computer Sciences, 2001. **41**(5): p. 1218-1227.

63. Gramatica, P., E. Giani, and E. Papa, *Statistical External Validation and Consensus Modeling: A QSPR Case Study for Koc Prediction.* Journal of Molecular Graphics and Modelling, 2007. **25**(6): p. 755-766.

64. Gramatica, P., P. Pilutti, and E. Papa, *Validated QSAR Prediction of OH Tropospheric Degradation of VOCs:□ Splitting into Training−Test Sets and Consensus Modeling.* Journal of Chemical Information and Computer Sciences, 2004. **44**(5): p. 1794-1802.

65. Kennard, R.W. and L.A. Stone, *Computer Aided Design of Experiments.* Technometrics, 1969. **11**(1): p. 137-148.

66. Gunturi, S.B. and R. Narayanan, *In Silico ADME Modeling 3: Computational Models to Predict Human Intestinal Absorption Using Sphere Exclusion and kNN QSAR Methods.* QSAR & Combinatorial Science, 2007. **26**(5): p. 653-668.

67. Golbraikh, A. and A. Tropsha, *Predictive QSAR Modeling Based on Diversity Sampling of Experimental Datasets for the Training and Test Set Selection.* Journal of Computer-Aided Molecular Design, 2002. **16**(5): p. 357-369.

68. Rothlauf, F. and D.E. Goldberg, *Representantions for Genetic and Evolutionary Algorithms*. 2002, Heidelberg, Alemania : Physica.

69. Toropov, A.A. and A.P. Toropova, *Modeling of Acyclic Carbonyl Compounds Normal Boiling Points by Correlation Weighting of Nearest Neighboring Codes.* Journal of Molecular Structure: THEOCHEM, 2002. **581**(1-3): p. 11-15.

70. Shen, M., et al., *Development and Validation of K-Nearest-Neighbor QSPR Models of Metabolic Stability of Drug Candidates.* Journal of Medicinal Chemistry, 2003. **46**(14): p. 3013-3020.

71. Duchowicz, P.R., et al., *Application of the Replacement Method as Novel Variable Selection in QSPR. 2. Soil Sorption Coefficients.* Chemometrics and Intelligent Laboratory Systems, 2007. **88**(2): p. 197-203.

72. Karelson, M., et al., *Neural Networks Convergence Using Physicochemical Data.* Journal of Chemical Information and Modeling, 2006. **46**(5): p. 1891-1897.

73. Soto, A., et al., *A Wrapper-Based Feature Selection Method for ADMET Prediction Using Evolutionary Computing*, in *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*, E. Marchiori and J. Moore, Editors. 2008, Springer Berlin / Heidelberg. p. 188-199.

74. RAVINDRANATH, D., *Structure-Based Generalized Models for Pure-Fluid Saturation Properties and Activity Coefficients*, School of Chemical Engineering. 2005, Oklahoma State University: Stillwater, Oklahoma.

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

76. Merkwirth, C., et al., *Ensemble Methods for Classification in Cheminformatics.* Journal of Chemical Information and Computer Sciences, 2004. **44**(6): p. 1971-1978.

77. Avdeef, A., *Prediction of Log P with Property-Based Method*, in *Molecular Drug Properties*, R. Mannhold, Editor. 2003. p. 381-406.

78. SPS, *ADMET Predictor(TM)* 2007, Simulations Plus, Inc.

79. *Dragon Professional 6*. 2010, Talete SRL.

80. MORIGUCHI, et al., *Simple Method of Calculating Octanol/Water Partition Coefficient*. Vol. 40. 1992, Tokyo, JAPAN: JAACC.

CHAPTER 4

A NON-LINEAR QSPR MODEL FOR MELTING POINT TEMPERATURE

4.1. Introduction

Melting point (MP) is the temperature at which the solid and liquid phases of a substance co-exist in equilibrium. These temperatures are invariably reported at atmospheric pressure. MP is an important property for identifying compounds and for analyzing purity. In addition, MP is used for predicting aqueous solubilities [1], boiling points [2] and eutectic compositions [3]. Aqueous solubility has enormous practical significance in the pharmaceutical industry for predicting the bioavailability and toxicity potential of drugs.

The solid structural form of any compound is held together by molecular interactions such as ionic, polar, dispersion and hydrogen bonding, which are enthalpic forces, and by positional, expansion, rotational and conformation flexibilities, which are entropic forces [4]. Melting occurs when the thermal agitation inside the solids overcomes these enthalpic and entropic forces. Thermodynamically, at the melting point T_m, the Gibbs free energy of phase transition becomes zero, which is expressed as:

$$\Delta G_m = \Delta H_m - T_m \Delta S_m = 0 \tag{4.1}$$

$$T_m = \frac{\Delta H_m}{\Delta S_m} \tag{4.2}$$

where, ΔH_m reflects the enthalpic forces and ΔS_m reflects the entropic forces. Melting of a substance occurs when the Gibbs free energy of the liquid state of the substance becomes lower than its solid state.

Inorganic compounds generally have high melting points due to the strong electrostatic forces between the constituent ions. The strongest intermolecular force generally exhibited by organic compounds is intermolecular hydrogen bonding, which is relatively weaker than electrostatic forces. Therefore, organic molecules will melt at lower temperatures than inorganic compounds [5]. For large molecules, however, induced dipole interactions become significant and can impact the crystal structure [6]. Further, molecular motion, categorized into oscillations or thermal librations, reorientations, and phase transitions, can also influence the structure of the crystal and affect the melting point [7]. All atoms in a molecule undergo oscillations, which become significant at higher temperatures. Some groups of atoms in a molecule or sometimes the whole molecule can undergo rotations or translations even at temperatures far below their melting points. If these reorientation motions become too easy or too frequent, the crystal structure becomes plastic or pre-liquid-like [7]. Some substances exhibit polymorphism where the compound can crystallize in many different crystal forms and, due to the polymorphism, these substances do not necessarily have only one clearly defined melting point [6]. These phase transformations are usually difficult to observe and the stability of each phase is not understood clearly. Brown and Brown [8] provide a good discussion on the thermodynamic aspects of melting and the effects of structural symmetry and flexibility on melting.

The models available in the literature for predicting the MP values are discussed in the next section. The majority of these models are developed using limited data, and hence, their general applicability is limited. In the current work, efforts have been made to develop QSPR models with much wider applicability using a much larger database of MP values. This work focuses on the following objectives:

1. Develop an accurate non-linear QSPR model to predict the MP values using a database made up a diverse set of compounds.
2. Compare the current modeling approach with existing modeling approaches in the literature on common external set data. This would further establish the efficacy of the modeling approach used in the work.

4.2. State of the Art in Melting Point Prediction

Despite the relative ease of measuring accurately melting point temperatures, modeling of MP has historically been one of the more difficult properties to model. This is due largely to the effect of secondary structural effects such as intermolecular hydrogen bonding and polymorphism, as discussed above. Hughes et al. [9] compared the predictive accuracies of models for octanol-water partition coefficients (Log K_{ow}), MP and aqueous solubility and reported that MP models are significantly less reliable than either Log K_{ow} or solubility models. Bergström et al. [10] and others [6, 11] have suggested that this lack of accuracy is due to the inability of the currently available molecular descriptors to describe the crystal structure of various compounds.

The earliest model for melting point prediction was developed by Mills in 1884 [12], after which the majority of MP models have been based on either group-contribution methods (GCM) or quantitative structure-property relationship (QSPR) methods. Katritzky et al. [5] provide a good review of the different approaches for melting point modeling prior to 2001. In GCM approaches, any molecular property is assumed to be the sum of contributions from predefined groups of atoms in that molecule. Joback and Reid [13] developed one of the earliest GCM approaches for prediction of melting points and boiling points along with other physical properties. Later Constantinou and Gani [14] developed a GCM approach based on UNIversal Functional Activity Coefficients (UNIFAC) groups that lead to better correlations than a simple GCM approach by considering second order group interactions. Wang et al. [15] have improved the GCM approach by taking into account position group contributions along with first and second order group contributions. The average deviation of prediction for their model was 14.5 K as opposed to 29.3 K and 27.8 K for the models by Joback and Reid [13], and Constantinou and Gani [14], respectively. Simamora and Yalkowsky [16] have used group contributions along with rotational symmetry (which is a non-additive and non-constituent molecular property) to develop a model with a standard deviation of 37.5 K for 1690 compounds. Yalkowsky and coworkers [17-20] have estimated melting points from the ratio of enthalpy and entropy of melting. Enthalpy of melting was estimated using GCM approaches [20], while entropy of melting was estimated using two sets of parameters. The first parameter set [17, 20] included molecular symmetry, σ (indicates the number of identical images that can be produce by rigid rotation of the molecule) and molecular flexibility, τ (empirically derived from the number of twist angles present in the molecule). The second parameter set [19] was eccentricity, ε (accounts for entropy of expansion) and spirality, μ (accounts for entropy of configuration). The models developed using only the molecular symmetry and flexibility numbers were able to predict the entropy of melting with an average error of 21% for 376 compounds [17]. When this model was coupled with the GCM model for enthalpy of melting, the resulting absolute average errors in melting point for a test set of 120 compounds was 36 K [18]. Also, the models for entropy of melting based on eccentricity and spirality were able to reduce the average absolute error on the melting points for a test set of 106 compounds by 52% (from 90 K to 43 K) [19]. The GCM approaches suffer from major disadvantages such as their inability to (a) model structures containing undefined functional groups, and (b) account for the interaction between different functional groups and for their spatial arrangement.

In QSPR approaches, the entire molecule is parameterized using molecular descriptors that are calculated using molecular mechanics or quantum mechanical methods. Bergström et al. [10] have built a QSPR model based only on the 2D and 3D molecular descriptors that were able to account for 63% of the variation in melting point data. This same data set was employed by Deeb et al. [21] to develop an improved model using genetic algorithms (GAs) and artificial neural networks (ANNs). This model could account for 70% of the variation in the melting point data and had a prediction root-mean squared error (RMSE) of 36 K. Modarresi et al. [4] have used the Bergström et al. [10] training dataset along with stepwise regression and genetic algorithms for descriptor selection. Their final model was comparable to the Bergström et al. [10] model in its accuracy. Karthikeyan [6] put together a large melting point database of 4,173 compounds, which is, to date, the most diverse database used for melting point prediction. Further, the author employed the 277 drug-like molecules used by Bergström et al. [10] as an external test set to gauge the predictive performance of the models. Principal component analysis (PCA), along with ANNs, was used for descriptor reduction and model development. The resulting model had a mean absolute error (MAE) of 32.6 K. Several other researchers have utilized the Karthikeyan [6] data set with differing techniques for descriptor selection and model development such as k nearest neighbor regression with genetic parameter optimization [22], ensembles of extreme learning machines [23], and artificial ant colony algorithms [24]. The RMSE on the internal test set for these models range from 45-49 K.

Recently, Godavarthy et al. [25] have developed a QSPR model for melting point prediction using symmetry descriptors. They report an RMSE value of 13 K on a training set of over 1200 molecules; however, the calculation of symmetry numbers has been performed manually in this work, as the rules used to evaluate the symmetry numbers cannot be translated into programming languages. This places a severe limitation on the applicability of this model, particularly in automated evaluation of the properties of new molecules. Apart from the general models mentioned above, there are other models in the literature that are applicable only to restricted classes of compounds such as alkanes [26], aldehydes [27], ketones [27], amines [27], substituted benzenes [28], and polychlorinated biphenyls [29]. These models will not be discussed further, since the objective of the current study is to develop a generalized melting point model that can be applied to a diverse structural range of compounds.

4.3. QSPR Methodology

The development of a QSPR model involves the following series of steps: (a) data set generation, (b) descriptor calculation, (c) descriptor reduction and model training, and (d) model validation. These elements are described below.

4.3.1. Data Set Generation: Experimental melting point values of 4,173 organic compounds, ranging from 287-665.5 K were taken from the article by Karthikeyan [6]. To date, this is the most comprehensive open-literature database available for MP values. From this database, we have removed compounds that are salts, as well as compounds that are stereo-isomers of other structures in the database. This pruning of the Karthikeyan database resulted in 3,878 melting point data. Additionally, we have included 952 melting point values from other sources [14, 16, 30-33] to further enrich the diversity of the database. The resulting OSU database is comprised of 4,830 melting point data; however, 43 of these molecules could not be optimized structurally for the most favorable (lowest energy) three-dimensional conformation using our automated procedure. The melting points of the resulting 4,787 molecules range from 74-662.15K. (Figure 4.1 provides the distribution of MP data). The molecular weights of these compounds vary from 16.05 g/mol to 786.04 g/mol, and the octanol-water partition coefficient K_{ow} (calculated by the DRAGON [34] software using the Ghose-Crippen ALOGP model) varies from -8.7 and 15. Details on the database characterization are given in Table 4.1. This principal dataset was used for training the model. Further, the molecules in the database are also characterized based on their drug-likeness as calculated using DRAGON [35] (Table 4.1). A score of 0 implies that the molecule has no characteristics of a drug, while a score of 1 implies the molecule has all the characteristics of being a drug.

In addition to the principal dataset, another set of 277 compounds was used for external validation. This dataset is the same as the external test set used by Karthikeyan [6] and was originally composed by Bergström et al. [10]. The purpose of this dataset is to provide a reliable estimation of the generalization capability and predictive capability of the final model.

4.3.2. Descriptor Calculation: See section 2.5.

4.3.3. Descriptor Reduction and Model Development: See section 2.6.

External Validation: In a recent article, Tropsha et al. [36] emphasized the need to validate QSPR models using external data sets. In the current work, a dataset of 277 molecules as identified by Karthikeyan [6] was used as the external validation set. The performance of the current model on this new dataset would indicate the generalization and predictive capability of the final model.

4.4. Results

10-descriptor, 15-descriptor, and 20-descriptor-models were tested, but no significant difference was observed among the models. Therefore, for the sake of simplicity, 10-descriptor models were used in the final models in the current study. Using less than 10 descriptors resulted in a significant increase in the training RMSE values for databases made up of more than 150 data points. Figure 4.2 is a comparison between the experimental and predicted melting point values for the data in the principal dataset. As can be expected from the distribution of melting point data (Figure 4.1), the deviation between the predictions and the experimental values is the lowest for compounds with melting points in the range 300-600 K, and the deviation becomes progressively higher beyond this interval. The correlation coefficient (R^2) between the experimental and predicted values is 0.86. The prediction residual errors are plotted in Figure 4.3, which demonstrates that the residuals are distributed normally except at temperatures beyond 550 K where the residuals show a distinctive downward bias. Further, the RMSE and the MAE values for the principal dataset predictions are 39.5 K and 30.2 K, respectively. The RMSE values for the individual ensembles range from 40.0 K to 41.6 K. The descriptors used for creating the eight different ensembles are shown in Table 4.2. Note that the neural networks in the ensembles are allowed to have a maximum of 10 elite inputs, but most frequently have a lower number. The descriptors F01[C-N], GATS1e, RBN, and Hy are common to all the ensembles, while the descriptors CIC3, NdssC, nDB, ATSC1i, and IC4 also occur frequently. The types and physical meaning of these commonly occurring descriptors are extracted from the DRAGON [33] help file and provided in Table 4.3.

The comparison between the experimental and the predicted melting points for the external test set is provided in Figure 4.4, where the calculated R^2 value is 0.43, signifying considerable differences between the calculated and the experimental values. The RMSE and MAE values for the external set predictions are 42.5 K and 33.9 K, respectively, which are only slightly higher than the corresponding values for the principal dataset. The corresponding RMSE values for the individual ensembles range from 41.8-45.0 K.

4.5. Discussion

For his model, Karthikeyan [6] reports MAE values of 37.6 K and 39.8 K for the training and internal validation sets, respectively; however, the MAE for the current model over the much larger principal dataset is 30.2 K, which is significantly lower. These error levels are considerably higher than the typical experimental error of 1-2 K reported for organic compounds, which suggest that the MP temperatures are hard to correlate with the existing molecular descriptors. Table 4.4 compares the prediction set error in this work with the errors calculated for other models in the literature that have used the same prediction set. Except for the linear model by Bergström et al. [10], the remaining models, which are all non-linear in nature, have statistically similar RMSE values for the prediction set. The important difference, however, is the number of descriptors used in the final model; this model employed 27 different molecular descriptors across eight ensembles, whereas the model by Karthikeyan [6] is made up of 26 principal components comprising more than 100 different molecular descriptors and, therefore, is more complex. An interesting aspect of the melting point ensembles created in the present work is the narrow difference in RMSE values between the best performing ensemble and the worst performing ensemble. The best performing ensemble had an RMSE of 41.8 K as opposed to 45 K for the worst predicting ensemble. Consequently, the difference between the average ensemble and the best ensemble is minimal. This could be attributed to the lack of molecular descriptors that can encode accurately the intermolecular interactions or the crystal structure of the molecule.

Also of significance is the observation that the current model generally over-predicts for compounds that melt at lower than 425 K and under-predicts for temperatures higher than

425 K. This is the reason for the downward bias observed in the residuals for MP values higher than 550 K in Figure 4.3. This is a trend also observed by Karthikeyan [6], Nigsch et al. [22] and Bhat et al. [23] and could be explained, in part, by the lower numbers of high- and low-melting molecules in the database employed in the current work.

Table 4.2 lists the most common descriptors across the eight different ensembles, which surprisingly are all 2D descriptors or constitutional descriptors independent of the 3D conformation of the molecule. Karthikeyan [6], in his article reports a similar trend in that the 2D descriptors performed better than 3D descriptors in prediction of melting point temperatures. The physical meanings of some descriptors in Table 4.3 are difficult to interpret, such as the 2D autocorrelation descriptors and the topological information indices; however, the common occurrence of these descriptors in the final model implies that the melting point is correlated with the 2D shape of the molecule and presence of electronegative groups in the molecule. Other common descriptors are easier to understand, like F01[C-N] which represents the number of C-N bonds in the molecule, Hy which represents the hydrophilicity and nDB which represents the number of double bonds in the molecule. Hydrophobicity or hydrophilicity, electronegativities, and partial charges have been found to be important molecular descriptors in Karthikeyan's model [6] as well. An interesting descriptor in Table 4.3 is RBN which represents the number of bonds in the molecule that can be freely rotated around them. This descriptor is theoretically similar to the molecular symmetry number (σ) proposed by Yalkowsky and coworkers [17, 19, 20] to model the entropy of melting.

4.6. Conclusions

1. A non-linear QSPR model for melting point temperature was developed using wrapper-based descriptor pruning techniques.
2. The RMSE on the external test set for the current model was 42.5 K, which compares favorably with the value of 41.4 K for Karthikeyan’s model [6]; however, the number of descriptors used in the current work is 27 as compared to more than 100 descriptors used by Karthikeyan.
3. According to the current work, the 2-dimensional shape of the molecule, hydrophilicity of the molecule, and the presence of electronegative charges in the molecule have an effect on the melting point temperature. Further, the number of rotatable bonds in the molecule is important in determining the MP temperature.
4. Like other literature models for predicting MP temperatures, the current model has relatively high prediction errors due to the lack of descriptors that can encode effectively the intermolecular forces and crystal structure information.
5. The resulting models from this work can be used to predict *a priori* the melting point temperatures of new molecules with reasonable accuracy.

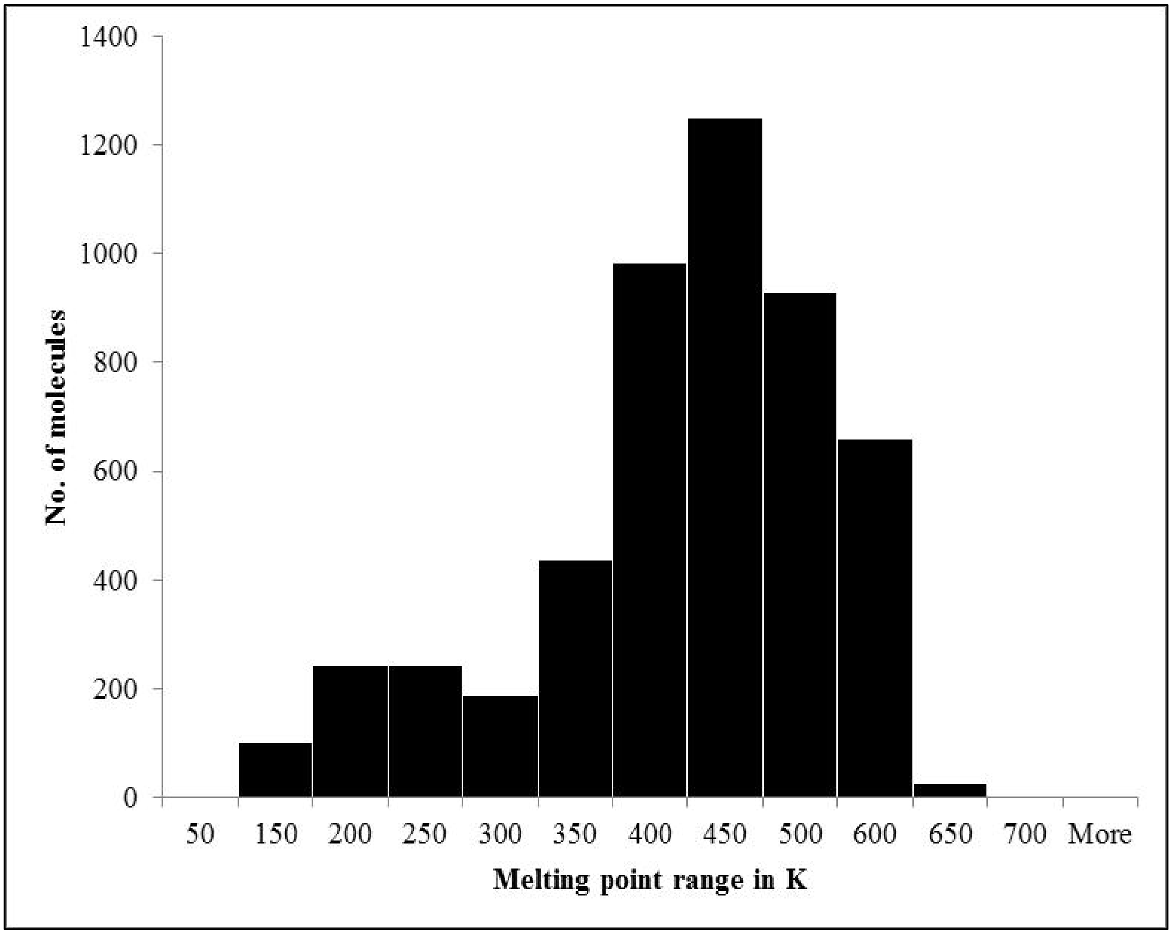


Figure 4.1: Distribution of the melting points in the principal data set

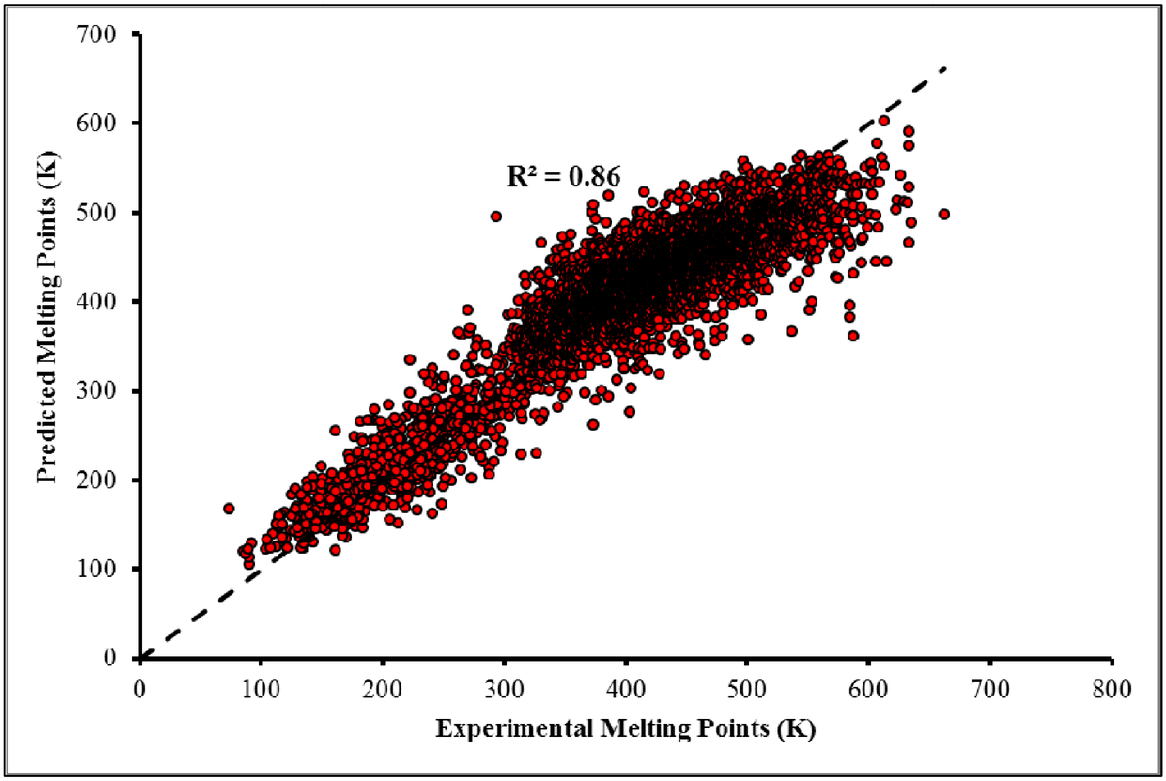


Figure 4.2: Comparison between the experimental and predicted melting point temperatures for the principal dataset. The broken line represents perfect predictions

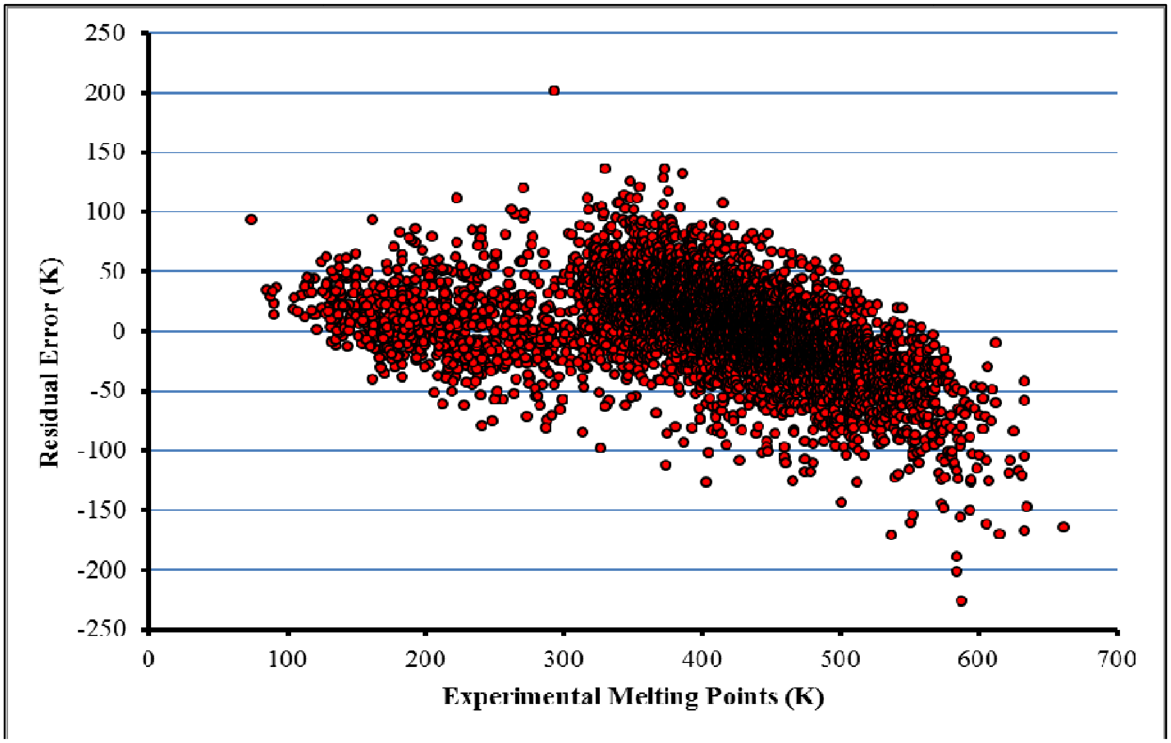


Figure 4.3: Residual error plot of the model predictions on the principal dataset

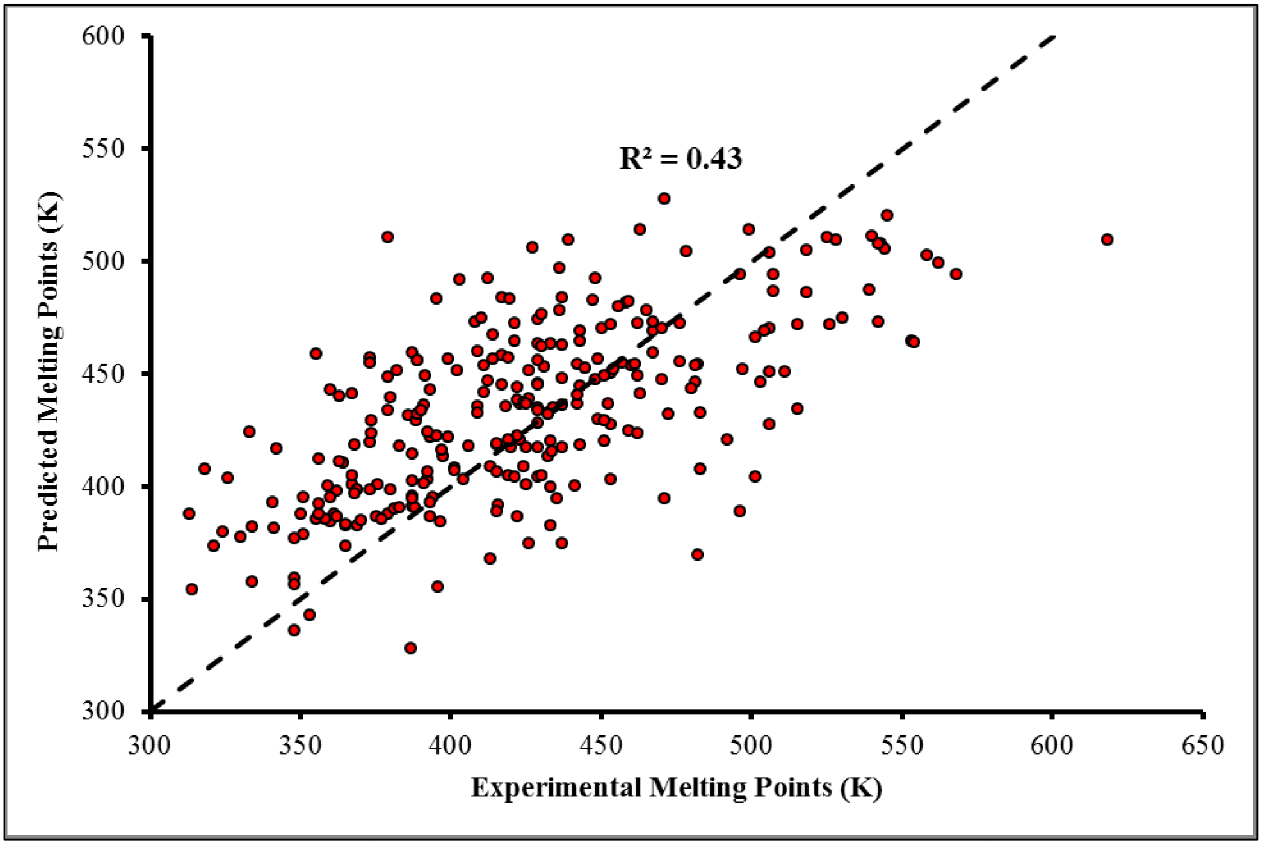


Figure 4.4: Comparison between the experimental and predicted melting point temperatures for the external test dataset. The broken line represents perfect predictions

Table 4.1: Characteristics of the principal dataset made up of 4,787 molecules

Molecular Property	Minimum	Maximum	Mean	Standard Deviation
Melting point (K)	74	662.15	398.8	101.9
Molecular weight (g/mol)	16.05	786.04	279.4	117.7
Octanol-water partition coefficient (Log K_{ow})	-8.7	15.0	2.6	2.3
DRAGON drug like score (0-1)	0.26	1.0	0.82	0.11

Table 4.2: List of the descriptors used in the final eight ensembles

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	F01[C-N]	F01[C-N]	F01[C-N]	F01[C-N]	F01[C-N]	F01[C-N]	F01[C-N]	F01[C-N]
2	GATS1e	GATS1e	GATS1e	GATS1e	GATS1e	GATS1e	GATS1e	GATS1e
3	RBN	RBN	RBN	RBN	RBN	RBN	RBN	RBN
4	Hy	Hy	Hy	Hy	Hy	Hy	Hy	Hy
5	CIC3	NdssC	NdssC	NdssC	CIC3	CIC3	CIC3	CIC3
6	nDB	C-040	SP02	SpPos_B (p)	nDB	nDB	Mor11e	C-040
7	nHAcc	IC4	IC4	IC4	RFD	P_VSA_LogP_4	VR3_Dz (Z)	RDF015m
8	SM3_Dt	ATSC1i	ATSC1i	ZM1Per	SM2_B (p)	SM2_B (p)	ATSC1i	SM3_Dt
9	IAC	SpAbs_B(p)	EE_B(e)	---	WiA_G/D	WiA_G/D	P_VSA_LogP_4	ARR
10	---	---	---	---	P_VSA_LogP_4	---	---	---

Table 4.3: Physical meaning of the commonly occurring descriptors in the ensembles

Descriptor	Descriptor Type	Physical Meaning
F01[C-N]	2D atom pairs	Number of Carbon-Nitrogen (C-N) bonds in the molecule.
GATS1e	2D autocorrelation	Geary coefficient, calculated from molecular graph by summing the products of atomic Sanderson electronegativities of the terminal atoms of all the paths of unit path length. Geary coefficient is a distance-type function varying from zero to infinite. Strong spatial autocorrelation produces small values of this index.
RBN	Constitutional indices	Number of bonds which allow free rotation around themselves. These are defined as any single bond, not in a ring, bound to a nonterminal heavy atom. Excluded from the count are amide C–N bonds because of their high rotational energy barrier
Hy	Molecular property	A hydrophilicity descriptor defined by Todeschini et al. [82] based on the number of hydrophilic groups (-OH, -SH, -NH), the number of carbon atoms and the number of atoms excluding hydrogen.
IC4	Topological information index	A topological information index calculated for an H-included molecular graph and based on neighbor degrees and edge multiplicity. It is calculated by partitioning graph vertices into equivalence classes; the topological equivalence of two vertices is that the corresponding neighborhoods of the 4th order are the same. The vertex neighborhood can be thought of as an open sphere comprising all the vertices in the graph, such that their distance from the considered vertex is less than 4.
CIC3	Topological information index	The Complementary Information Content (CIC3) measures the deviation of the information content IC3 from its maximum value, which corresponds to the vertex partition into equivalence classes containing one element each.
NdssC	Atom-type E-state indices	Number of atoms of type dssC
nDB	Constitutional descriptor	Number of double bonds in the molecule

Table 4.3 (cont'd): Physical meaning of the commonly occurring descriptors in the ensembles

Descriptor	Descriptor Type	Physical Meaning
ATSC1i	2D autocorrelation	Centered Broto-Moreau correlation, calculated from molecular graph by summing the products of ionization potentials of the terminal atoms of all the paths of unit path length.

Table 4.4: Comparison of the current model with literature models on the basis of predictions for 277 drug-like molecules

Researchers	Model Type	No. of Descriptors Used in the Model	RMSE (K)	MAE (K)
This work	Stochastic optimization and ANNs	27 descriptors across eight ensembles	42.5	33.9
Karthikeyan [6]	Principal component analysis (PCA) and ANNs	26 principal components made up of more than 100 descriptors	41.4	32.6
Bergström et al. [10]	Partial least squares (PLS)	121 descriptors	49.8	---
Nigsch et al. [22]	Genetic algorithms and k-nearest neighbor (k-NN)	146 descriptors across 15 nearest neighbors	42.2	---

REFERENCES

1. Jain, P. and S.H. Yalkowsky, *Prediction of Aqueous Solubility from SCRATCH.* International Journal of Pharmaceutics, 2010. **385**(1-2): p. 1-5.

2. Walters, A.E., P.B. Myrdal, and S.H. Yalkowsky, *A Method for Estimating the Boiling Points of Organic Compounds from Their Melting Points.* Chemosphere, 1995. **31**(4): p. 3001-3008.

3. Law, D., et al., *Prediction of Poly(Ethylene) Glycol-Drug Eutectic Compositions Using an Index Based on the Van't Hoff Equation.* Pharmaceutical Research, 2002. **19**(3): p. 315-321.

4. Modarresi, H., J.C. Dearden, and H. Modarress, *QSPR Correlation of Melting Point for Drug Compounds Based on Different Sources of Molecular Descriptors.* Journal of Chemical Information and Modeling, 2006. **46**(2): p. 930-936.

5. Katritzky, A.R., et al., *Perspective on the Relationship between Melting Points and Chemical Structure.* Crystal Growth & Design, 2001. **1**(4): p. 261-265.

6. Karthikeyan, M., R.C. Glen, and A. Bender, *General Melting Point Prediction Based on a Diverse Compound Data Set and Artificial Neural Networks.* Journal of Chemical Information and Modeling, 2005. **45**(3): p. 581-590.

7. Gavezzotti, A. and M. Simonetta, *Crystal Chemistry in Organic Solids.* Chemical Reviews, 1982. **82**(1): p. 1-13.

8. Brown, R.J.C. and R.F.C. Brown, *Melting Point and Molecular Symmetry.* Journal of Chemical Education, 2000. **77**(6): p. 724-null.

9. Hughes, L.D., et al., *Why Are Some Properties More Difficult to Predict Than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P.* Journal of Chemical Information and Modeling, 2008. **48**(1): p. 220-232.

10. Bergström, C.A.S., et al., *Molecular Descriptors Influencing Melting Point and Their Role in Classification of Solid Drugs.* Journal of Chemical Information and Computer Sciences, 2003. **43**(4): p. 1177-1185.

11. Clark, M., *Generalized Fragment-Substructure Based Property Prediction Method.* Journal of Chemical Information and Modeling, 2005. **45**(1): p. 30-38.

12. Mills, E.J., *On Melting Point and Boiling Point as Related to Composition.* Philosophical Magazine (series 5), 1884. **17**(105): p. 173-187.

13. Joback, K.G. and R.C. Reid, *Estimation of Pure-Component Properties from Group-Contributions.* Chemical Engineering Communications, 1987. **57**(1-6): p. 233-243.

14. Constantinou, L. and R. Gani, *New Group Contribution Method for Estimating Properties of Pure Compounds.* AIChE Journal, 1994. **40**(10): p. 1697-1710.

15. Wang, Q., P. Ma, and S. Neng, *Position Group Contribution Method for Estimation of Melting Point of Organic Compounds.* Chinese Journal of Chemical Engineering, 2009. **17**(3): p. 468-472.

16. Simamora, P. and S.H. Yalkowsky, *Group Contribution Methods for Predicting the Melting Points and Boiling Points of Aromatic Compounds.* Industrial & Engineering Chemistry Research, 1994. **33**(5): p. 1405-1409.

17. Dannenfelser, R.-M. and S.H. Yalkowsky, *Predicting the Total Entropy of Melting: Application to Pharmaceuticals and Environmentally Relevant Compounds.* Journal of Pharmaceutical Sciences, 1999. **88**(7): p. 722-724.

18. Jain, A., G. Yang, and S.H. Yalkowsky, *Estimation of Melting Points of Organic Compounds.* Industrial & Engineering Chemistry Research, 2004. **43**(23): p. 7618-7621.

19. Johnson, J.L.H. and S.H. Yalkowsky, *Two New Parameters for Predicting the Entropy of Melting:□ Eccentricity (E) and Spirality (M).* Industrial & Engineering Chemistry Research, 2005. **44**(19): p. 7559-7566.

20. Zhao, L. and S.H. Yalkowsky, *A Combined Group Contribution and Molecular Geometry Approach for Predicting Melting Points of Aliphatic Compounds.* Industrial & Engineering Chemistry Research, 1999. **38**(9): p. 3581-3584.

21. Deeb, O., M. Goodarzi, and S. Alfalah, *Prediction of Melting Point for Drug-Like Compounds Via QSPR Methods.* Molecular Physics, 2011. **109**(4): p. 507-516.

22. Nigsch, F., et al., *Melting Point Prediction Employing K-Nearest Neighbor Algorithms and Genetic Parameter Optimization.* Journal of Chemical Information and Modeling, 2006. **46**(6): p. 2412-2422.

23. Bhat, A.U., S.S. Merchant, and S.S. Bhagwat, *Prediction of Melting Points of Organic Compounds Using Extreme Learning Machines.* Industrial & Engineering Chemistry Research, 2008. **47**(3): p. 920-925.

24. O'Boyle, N., et al., *Simultaneous Feature Selection and Parameter Optimisation Using an Artificial Ant Colony: Case Study of Melting Point Prediction.* Chemistry Central Journal, 2008. **2**(1): p. 1-15.

25. Godavarthy, S.S., R.L. Robinson Jr., and K.A.M. Gasem, *An Improved Structure-Property Model for Predicting Melting-Point Temperatures.* Industrial and Engineering Chemistry Research, 2006. **45**(14): p. 5117-5126.

26. Charton, M. and B. Charton, *Quantitative Description of Structural Effects on Melting Points of Substituted Alkanes.* Journal of Physical Organic Chemistry, 1994. **7**(4): p. 196-206.

27. Katritzky, A.R. and E.V. Gordeeva, *Traditional Topological Indexes Vs Electronic, Geometrical, and Combined Molecular Descriptors in QSAR/QSPR Research.* Journal of Chemical Information and Computer Sciences, 1993. **33**(6): p. 835-857.

28. Katritzky, A.R., et al., *Prediction of Melting Points for the Substituted Benzenes:□ A QSPR Approach.* Journal of Chemical Information and Computer Sciences, 1997. **37**(5): p. 913-919.

29. Gramatica, P., N. Navas, and R. Todeschini, *3D-Modelling and Prediction by WHIM Descriptors. Part 9. Chromatographic Relative Retention Time and Physico-Chemical Properties of Polychlorinated Biphenyls (PCBs).* Chemometrics and Intelligent Laboratory Systems, 1998. **40**(1): p. 53-63.

30. Bondi, A., *Physical Properties of Molecular Crystals, Liquids, and Gases*. 1968, New York: John Wiley and Sons.

31. Lyman, W.J., W.F. Reehl, and D.H. Rosenblatt, *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*. 1990, Washington, D.C. : American Chemical Society.

32. Reid, R.C., J.M. Prausnitz, and T.K. Sherwood, *The Properties of Gases and Liquids*. 1977: McGraw-Hill.

33. Krzyzaniak, J.F., et al., *Boiling Point and Melting Point Prediction for Aliphatic, Non-Hydrogen-Bonding Compounds.* Industrial & Engineering Chemistry Research, 1995. **34**(7): p. 2530-2535.

34. *Dragon Professional 5.5*. 2010, Talete SRL.

35. *Dragon Professional 6*. 2010, Talete SRL.

36. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

CHAPTER 5

A NON-LINEAR QSPR MODEL FOR GIBBS ENERGY OF FORMATION

5.1. Introduction

A frequent problem encountered by chemists is the inability to determine if a particular compound can be synthesized using certain reaction principles. Further, in virtual design paradigms, whether a designed compound can exist at a specified temperature and pressure is often in question. The solution to these problems is based on an understanding of the thermodynamic potentials of the reactants and products involved in the compound synthesis. These thermodynamic potentials are the driving forces for all natural processes to their equilibrium states [1]. Free energy, which is usually expressed as the Helmholtz function, A, or the Gibbs function, G, is a measure of the thermodynamic potential, and consequently, is an important property in thermodynamics [2]. The Helmholtz function is generally applied to a system with constant number of particles, temperature, and volume (constant NVT), whereas the Gibbs function is generally applicable to a system with constant number of particles, temperature and pressure (constant NPT). Since most experiments are carried out at constant temperature and pressure, the

Gibbs function, also known as the free enthalpy, is the commonly used form to represent the free energy [2].

Consider the formation of a compound P from its constituent elements R_1 and R_2.

$$\mathrm{n_1R_1 + n_2R_2 \leftrightarrow P} \tag{5.1}$$

where, n_1 and n_2 are the number of moles of R_1 and R_2, respectively.

At a given temperature, the equilibrium constant of this equation can be written as follows:

$$\mathrm{K_f = \frac{(P)}{(R_1)^{n_1}.\,(R_2)^{n_2}}} \tag{5.2}$$

where, K_f is the equilibrium constant for the formation reaction. Thermodynamically, this equilibrium constant is related to the change in Gibbs free energy (ΔG_f) in the following manner:

$$\mathrm{\Delta G_f = -R * T * ln\,(K_f)} \tag{5.3}$$

where, R is the molar gas constant and T is the temperature.

According to Equation 5.2, the equilibrium for the reaction (Equation 5.1) will be shifted to the right if K_f is greater than 1, and shifted to the left if K_f is less than 1. Considering the relation between K_f and ΔG_f from Equation 5.3, a negative value of ΔG_f implies the reaction is shifted to the right, and a positive value of ΔG_f implies the reaction is shifted to the left. In other words, a compound is stable if the value of ΔG_f at that particular temperature is negative. In addition, for a reaction system involving products and reactants, the Gibbs free energy of the reaction is equal to the sum of the free energies of formation of the reactants subtracted from the free energies of formation of the products. For example, consider a reaction where C and D react to form the products F and G as shown:

$$\mathrm{C} + \mathrm{D} \leftrightarrow \mathrm{F} + \mathrm{G} \tag{5.4}$$

The free energy of the above reaction is written as:

$$\Delta\mathrm{G}_{\mathrm{reaction}} = \Delta\mathrm{G}_{\mathrm{fF}} + \Delta\mathrm{G}_{\mathrm{fG}} - \Delta\mathrm{G}_{\mathrm{fC}} - \Delta\mathrm{G}_{\mathrm{fD}} \tag{5.5}$$

This free energy of the reaction $\Delta G_{reaction}$ can then be used to estimate the equilibrium of the reaction shown in Expression 5.4.

The preceding discussion illustrates the importance of the Gibbs free energy of formation for estimating the stability of a compound relative to its elements, and for estimating the position of equilibrium for a given reaction. However, experimental determination of the free energy is difficult, particularly for systems with multiple minimum energy configurations separated by low-energy barriers [2]. Further, the component properties such as entropy and chemical potential are difficult to measure. Also, other popular computational techniques such as molecular dynamics (MD) and Monte Carlo (MC) simulations are impractically expensive to carryout for pure systems [3].

Therefore, a need exists for models that can reliably predict the Gibbs energy of formation values for a wide range of compounds. The models available in the literature for predicting the Gibbs energy of formation are discussed in the next section. The majority of these models are developed using limited data, and hence, their general applicability is limited. In the current work, efforts have been made to develop QSPR models with much wider applicability using a much larger database of Gibbs energy of formation values. This work focuses on the following objectives:

1. Develop an accurate non-linear QSPR model to predict the Gibbs energy of formation using a database made up of diverse set of compounds.
2. Validate the current modeling approach by employing an external test set of compounds that has not been used to develop the model.
3. Compare the current modeling approach with existing modeling approaches in the literature, on common training and external set data. This would further establish the efficacy of the modeling approach used in the work.

5.2. State of the Art in Predicting the Gibbs Energy of Formation

The earliest work involving Gibbs energy of formation modeling was carried out by van Krevelen and Chermin [1], who used the group-contribution (GCM) approach to estimate the Gibbs energy with a mean average error (MAE) of 3.1 kcal/mol on the entire training data (data that has been used for model development) set. Joback [4] developed an improved GCM approach and reported a MAE of 1.01 kcal/mol on a training data set of 328 compounds. Constantinou and Gani [5] have further improved the GCM approach by including second-order group contributions, and they report a MAE value of 0.78 kcal/mol for their model on the training data. Mavrovouniotis [6] has used an analogous GCM method to model the energy of formation of biochemical compounds in aqueous solutions. More recently, Ivanciuc et al. [7] have employed information-theory along with quantitative structure-property relationship (QSPR) modeling techniques, to develop a model for predicting the free energy of alkanes between C_6 and C_{10}. Wang et al. [8] have used a relatively more diverse set of 180 small to medium sized organic molecules (with less than 10 carbon atoms) and employed the density functional theory (DFT) with neural network corrections to model the Gibbs free energy. Their best neural network had a root-mean-squared error (RMSE) of 3.1 kcal/mol on an external test set of 30 molecules. In another recent work, Yan [9] developed a QSPR model for free energy based only on the 2-dimensional (2D) descriptors of the molecules, and it employed the same set of compounds as Wang et al. [8], except for three compounds that were deemed incompatible with their descriptor generation software. Yan's model [9] was built using Kohonen's self-organizing neural networks and produced a mean absolute error (MAE) of 11.2 kcal/mol for an external test set (data that has not been used for model development) made up of 27 molecules.

5.3. QSPR Methodology

The development of a QSPR model involves the following series of steps: (a) data set generation, (b) descriptor calculation, (c) descriptor reduction and model training, and (d) model validation. These elements are described below.

5.3.1. Data Set Generation: Experimental Gibbs energy of formation, ΔG_f, values at 298K for 1,126 organic compounds were taken from the chemical properties handbook by Yaws [10]. Of these compounds, the structures for four molecules could not be found and the descriptors for 14 other molecules could not be calculated using the DRAGON [11] software. Therefore, the final database used for modeling is composed of 1,108 molecules. To date, this is the most comprehensive database available for ΔG_f values. The ΔG_f values of the molecules in the final OSU database lie in the range -1970 kJ/mol to 665 kJ/mol (Figure 5.1 provides the distribution of ΔG_f data). The molecular weights of these compounds vary from 16.05 g/mol to 446.74 g/mol, and the octanol-water partition coefficient K_{ow} (calculated using the Ghose-Crippen ALOGP model in DRAGON [11]) varies between -2.9 and 9.8. Additionally, the molecules are characterized based on their drug-likeness as calculated using DRAGON [11]. A score of 0 implies that the molecule has no characteristics of a drug, while a score of 1 implies the molecule has all the characteristics of being a drug. Further details on the database characterization are given in Table 5.1.

In addition to the above data sets, additional ΔG_f data of 180 diverse organic compounds were extracted from the article by Yan [9], which were originally taken from the Chemical Properties Handbook [10]. Henceforth in this work, this data will be referred to as the Yan's database to differentiate it from the OSU data set. To validate the current modeling approach, Yan's data were used to develop a QSPR model to predict the ΔG_f values and the resulting model was compared with the prediction results by Yan.[9]. To ensure a fair comparison, the same training and external test data employed by Yan were used in the current work.

5.3.2. Descriptor Calculation: See Section 2.5

5.3.3. Descriptor Reduction and Model Development: See Section 2.6

External Validation: In a recent article, Tropsha et al. [12] emphasized the need to validate QSPR models using external data sets. Therefore, another model was built by separating some data from the original database and allocating it to an external test set. However, the data cannot be randomly separated, as the external set might not be representative of the training set. Therefore, a self-organizing map (SOM) network was created using the best descriptors identified in the first ensemble, which was developed using the entire database. This SOM was used to identify clusters in the data and partition the data into T, IV and IT sets as explained in Section 5.3.3. The number of map units in this SOM was varied until the percentage of data points in the IT set is at least 15% of the size of the entire final data set of 1108 molecules. This IT set was then set aside as an external test set and the remaining data was used for developing another model *de novo*, by repeating the search for the best descriptors, best network architecture and network weights. In the current work, 177 molecules were identified as an external test set using this procedure, and the remaining 931 data points were again divided into T, IV and IT sets and subjected to the descriptor search algorithm as discussed in Section 5.3.3. For clarity, in this work, the model created using all 1108 data points for training will be referred to as Model 1 and the model created using just the 931 data points as Model 2. Model 1 will be used in the computer-aided molecular design (CAMD) algorithms because of its larger training set size, and Model 2 will be used to assess approximately the generalization capability of Model 1, as advocated by Tropsha et al. [12].

5.4. Results

5.4.1 Model 1: Ten-descriptor, 20-descriptor and 30-descriptor models were tested; the 20-descriptor models had lower training set errors than the 10-descriptor models, but no significant difference was observed between the 20-descriptor and 30-descriptor models. Therefore, for the sake of simplicity, 20-descriptor models were used in the final models in the current study. Going lower than 10 descriptors resulted in a significant increase in the training RMSE values for databases made up of more than 150 data points. Therefore, twenty was chosen as the minimum number of input descriptors. Figure 5.2 is a comparison between the experimental and predicted ΔG_f values for Model 1. The correlation coefficient (R^2) between the experimental and predicted values is 0.99. The prediction residual errors in kJ/mol are plotted in Figure 5.3 which clearly demonstrates, that the residuals are almost symmetrically distributed around the horizontal axis, as should be expected from an unbiased model. A histogram of the residuals (no figure shown) was plotted, where the distribution of the residuals around zero was found to be similar to a normal distribution. In addition, the RMSE and the MAE values for the training data set predictions are 17.4 kJ/mol and 9.7 kJ/mol, respectively. The RMSE values for the individual ensembles range from 18.3 kJ/mol to 20.6 kJ/mol. The results from the overall ensemble are slightly better than the results for the individual ensembles, which validates the use of ensembles with different descriptors as inputs.

The different descriptors used for creating the eight different ensembles are shown in Table 5.2. Note that the neural networks in the ensembles are allowed to have a maximum of 20 elite inputs, but most frequently they end up having a slightly lower number of elite descriptors as inputs, after the insignificant descriptors have been removed as described in Section 5.3.3. The descriptors nN, Ho_D/Dt, MAXDN, P_VSA_v_3, P_VSA_p_3, SdO and SM1_Dz (Z) are the most common across the ensembles. The types and physical meanings of these commonly occurring descriptors, as extracted from the DRAGON [11] help file, are provided in Table 5.3.

5.4.2. Model 2: For Model 2, 20-descriptor inputs were chosen for model development. Figure 5.4 is a comparison between the experimental and predicted ΔG_f values for the training data of 931 compounds. The correlation coefficient (R^2) between the experimental and predicted training data is 0.99. The prediction residual errors on this data are near-symmetrically distributed around the horizontal axis (no figure shown**)**. The

RMSE and MAE values for the training set data are 21.4 kJ/mol and 11.4 kJ/mol, respectively. Figure 5.5 compares the experimental and predicted ΔG_f values of the external test set of 177 compounds. The RMSE and MAE values for the external test set are calculated to be 32.4 kJ/mol and 16.4 kJ/mol, respectively.

The descriptors used for creating the eight different ensembles for Model 2 are tabulated in Table 5.4. The descriptors nN, nHet, MAXDN, C-024, ON0V, P_VSA_p_3, and SM15_EA (dm) are the most common across the ensembles. The types and physical meanings of these commonly occurring descriptors, as extracted from the DRAGON [11] help file, are provided in Table 5.5.

5.4.3. Model for Yan's Database: Ten descriptor-models were developed in the current work to correlate the molecules in the Yan's database [9]. One molecule could not be optimized for its 3D structure, and therefore was left out of the modeling process. For the current model, the RMSE and MAE values are calculated to be 21.5 kJ/mol and 16.6 kJ/mol, respectively for the training set comprising 152 compounds. For the external test set comprising 27 compounds, the RMSE and MAE values are calculated to be 29.2 kJ/mol and 21.1 kJ/mol, respectively. A comparison between the experimental and predicted ΔG_f values for the external data of 27 compounds is provided in Figure 5.6. The correlation coefficient (R^2) between the experimental and predicted external test data is calculated to be 0.98.

The errors for all models developed in the current work are tabulated in Table 5.6.

5.5. Discussion

The RMSE value for the Model 1 training set is approximately 25% lower than the corresponding value for Model 2. This difference can be attributed to the larger training set size for Model 1, which allows for better training across the different classes of compounds. Due to the larger training data set, Model 1 would be expected to perform similarly to Model 2 on unseen data (external data set). Therefore, the predictive performance of Model 2 on an external test set can be used as an approximation for determining the generalization capability of Model 1. Few works exist in the open literature relating to the prediction of Gibbs energy of formation, and to the best knowledge of the author, the models by Wang et al. [8] and Yan [9] are the only generalized Gibbs energy of formation models in the literature, where an external test set has been used to assess the predictive capability. These models are compared with the current model in Table 5.7. Wang et al. [8] have reported a RMSE value of 3.1 kcal/mol (13.0 kJ/mol) on an external test set of 30 molecules, using a model that was trained on a data set of 150 molecules. Using 177 compounds from the same database as Wang et al. [8], Yan [9] reported a MAE value of 11.2 kcal/mol (46.9 kJ/mol) for an external test set comprising 27 molecules.

Also, the compounds that exhibit the largest deviations in the various models were further examined manually to identify any correlation between their higher errors and the presence/absence of certain functional groups. The majority of the molecules that had high deviations contain at least one oxygen atom. Table 5.8 lists the 5 compounds in the external set, for which the Model 2 predictions and the experimental values of ΔG_f differ by more than 100 kJ/mol. The first molecule is a geometric isomer and the high error can be attributed to the lack of sufficient number of geometric isomers in the training data. The last molecule is an inorganic compound, and the high prediction error in this case may be due to the lack of sufficient number of inorganics in the training set. The functional groups of the other three molecules, however, are represented sufficiently in the training set, and any reason for the high error is unclear.

Table 5.3 lists the most common descriptors for the eight different ensembles for Model 1. Surprisingly, these descriptors are 2D descriptors or constitutional descriptors independent of the 3D conformation of the molecule. Due to the black-box nature of the artificial neural networks (ANNs), a quantitative assessment of the significance of the different descriptors on the Gibbs energy is not possible. However, approximate qualitative interpretations can be made based on the type of descriptors. For example, the presence of the 2D matrix-based descriptors, Ho_D/Dt and SM1_Dz (Z) indicates a correlation between ΔG_f and the shape of the molecule. In addition, the presence of the descriptors, MAXDN, P_VSA_p3, and SdO indicates that the charge distribution around the molecule also has an effect on the ΔG_f values. The number of nitrogen atoms, denoted by the descriptor nN also has an effect on the Gibbs energy values. As expected, some of the common descriptors for Model 2 (listed in Table 5.5) are identical to the common descriptors for Model 1. For example, the descriptors nN, MAXDN, P_VSA_p3 are common across both Model 1 and Model 2. The other descriptors in Model 2, though not exactly identical to Model 1 descriptors, are drawn from the same sub-category of descriptors as in Model 1, and describe the shape of the molecule. This suggests a strong correlation between the shape and Gibbs energy of formation for a molecule. In addition, the common descriptors observed in the current work are similar to the 2D descriptors such as σ charge and lone-pair electronegativity, employed by Yan [9].

To compare the efficacy of the current modeling approach, the Yan's data set [9] had been employed to develop a QSPR model. The results from this model are provided in Table 5.9, along with the results by Wang et al.[8], using the same training and external test set data. The current model performs significantly better on both the training data and the external set data, which indicates better generalization capability of the current model, when compared with the model by Yan [9]. The same data, but different training and external test set partitions were employed by Wang et al. [8] in their DFT correction approach of modeling the ΔG_f values. The considerably lower error for the model by Wang et al. [8] could be due to fact that all these molecules are small molecules (with lower than 10 carbon atoms) and the Gibbs energy of formation values are calculated using the density-functional theory, and not by the standard molecular-descriptor approach. Although, the systematic deviations in the density-functional theory calculations for small molecules (with lower than 10 carbon atoms) can be accurately corrected using multi-linear regression or neural networks, the deviations from the experimental values for medium to large sized molecules is significantly large and alternative strategies of modeling are usually preferred [8]. The current QSPR model does not suffer from this disadvantage and can be applied to molecules of any size, which makes it ideal for predicting the Gibbs energies of formation values for new molecules identified during the virtual design process.

5.6. Conclusions

1. A non-linear QSPR model for ΔG_f at 298K was developed using wrapper-based descriptor pruning techniques.
2. Two models were developed. Model 1 was built using ΔG_f values for 1,108 compounds, and all this data was used for model development, and Model 2 was built by employing ΔG_f values of 931 compounds from the original database of 1,108 compounds with 177 compounds reserved as an external test set.
3. The RMSE values on the training sets for Model 1 and Model 2 are 17.4 kJ/mol and 21.4 kJ/mol, respectively. The RMSE value for Model 2 on the external test set is 32.4 kJ/mol.
4. According to the current work, the 2-dimensional shape of the molecule and the distribution of electronegative charges in the molecule significantly affect the Gibbs energy of formation values.
5. The current model developed using the Yan's data set performs significantly better than the model by Yan [9] on an external test set of 27 compounds. The MAE value on the external test set for the model by Yan is 47 kJ/mol as compared to a MAE value of 21 kJ/mol from the current model.
6. The resulting models from this work can be used to accurately predict *a priori* the Gibbs energy of formation of new molecules and thereby their stability.

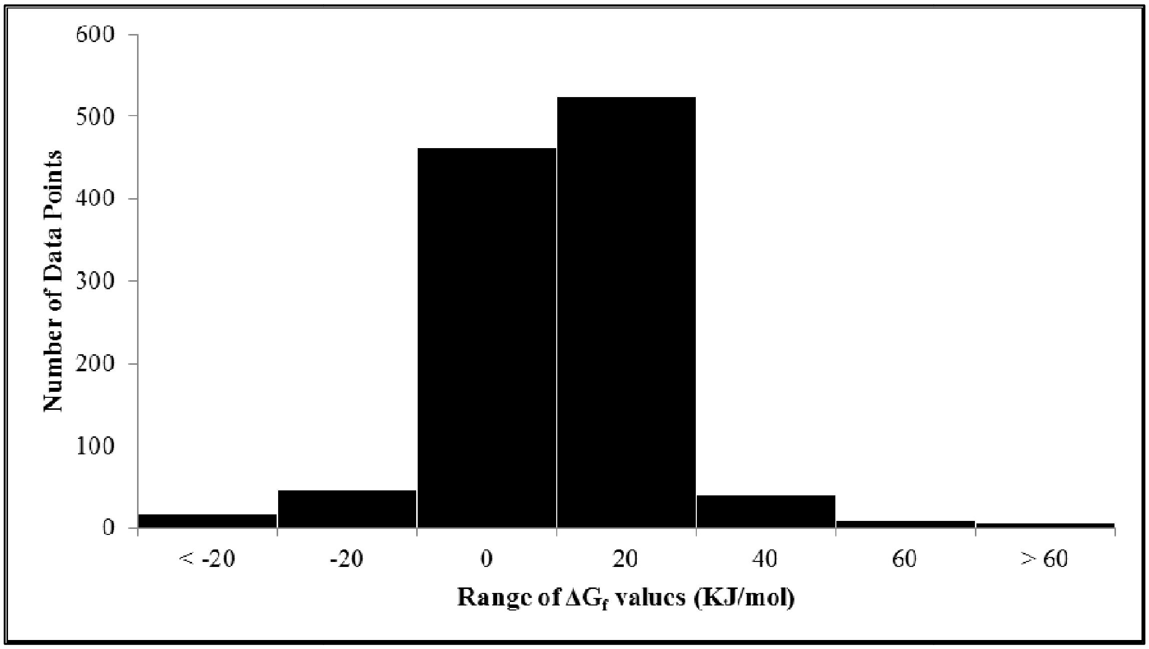


Figure 5.1: Distribution of the ΔG_f values in the final data set

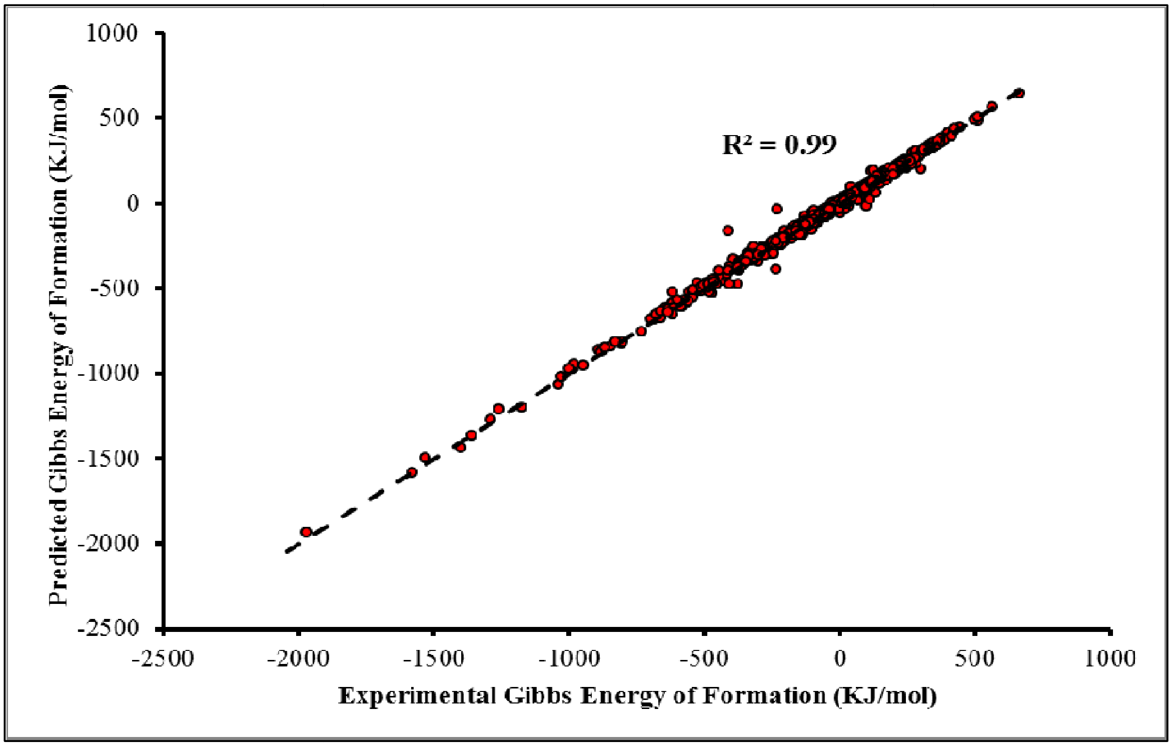


Figure 5.2: Comparison between the experimental and predicted ΔG$_f$ values for Model 1. The broken line represents perfect predictions

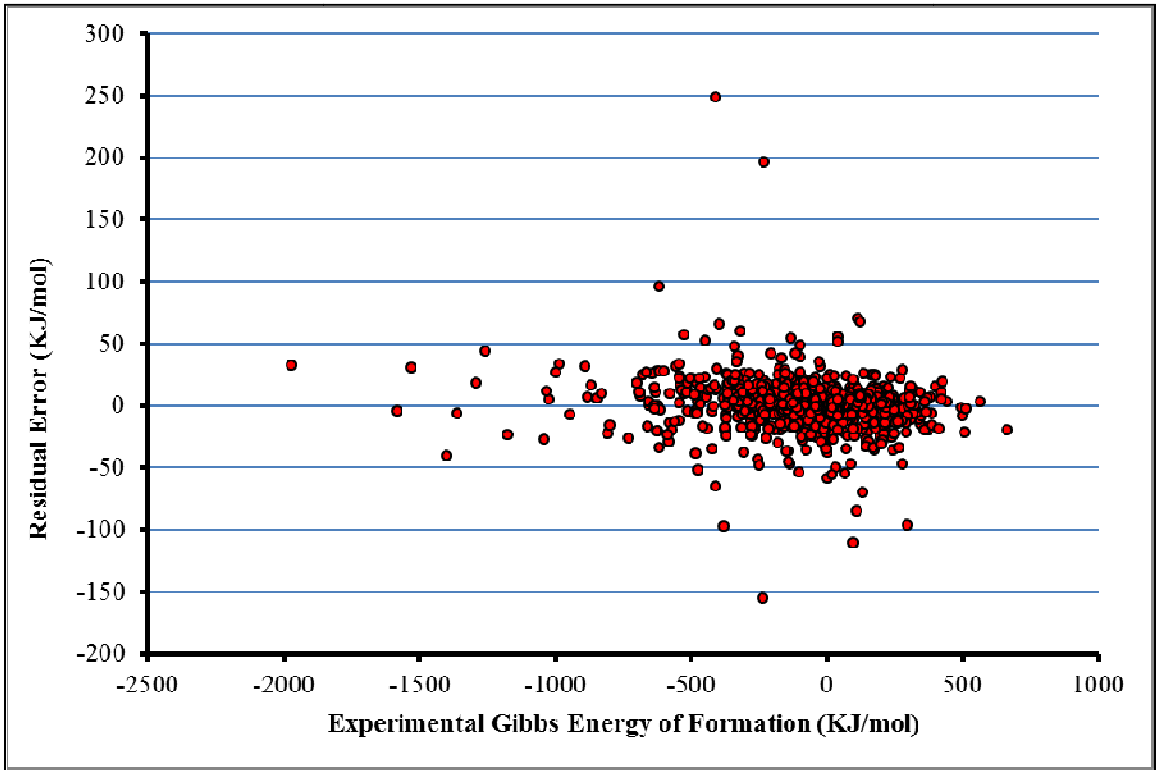


Figure 5.3: Residual error plot of the Model 1 predictions

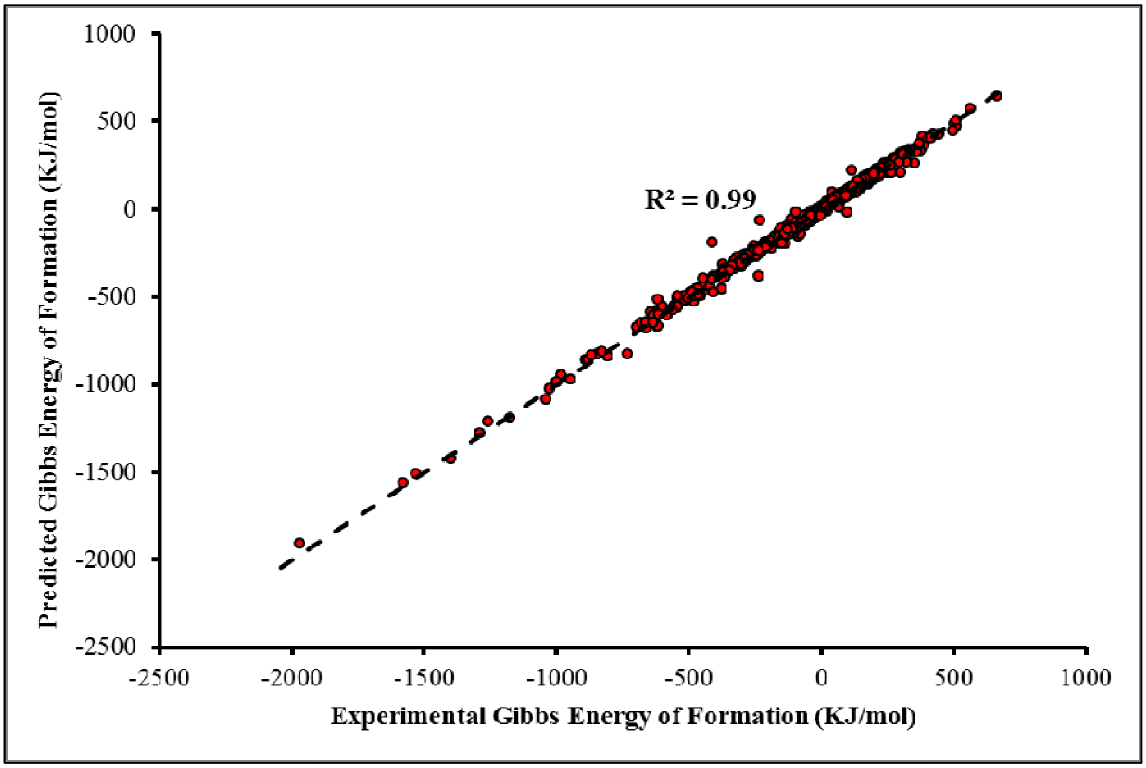


Figure 5.4: Comparison between the experimental and predicted ΔG_f values for the training data in Model 2. The broken line represents perfect predictions

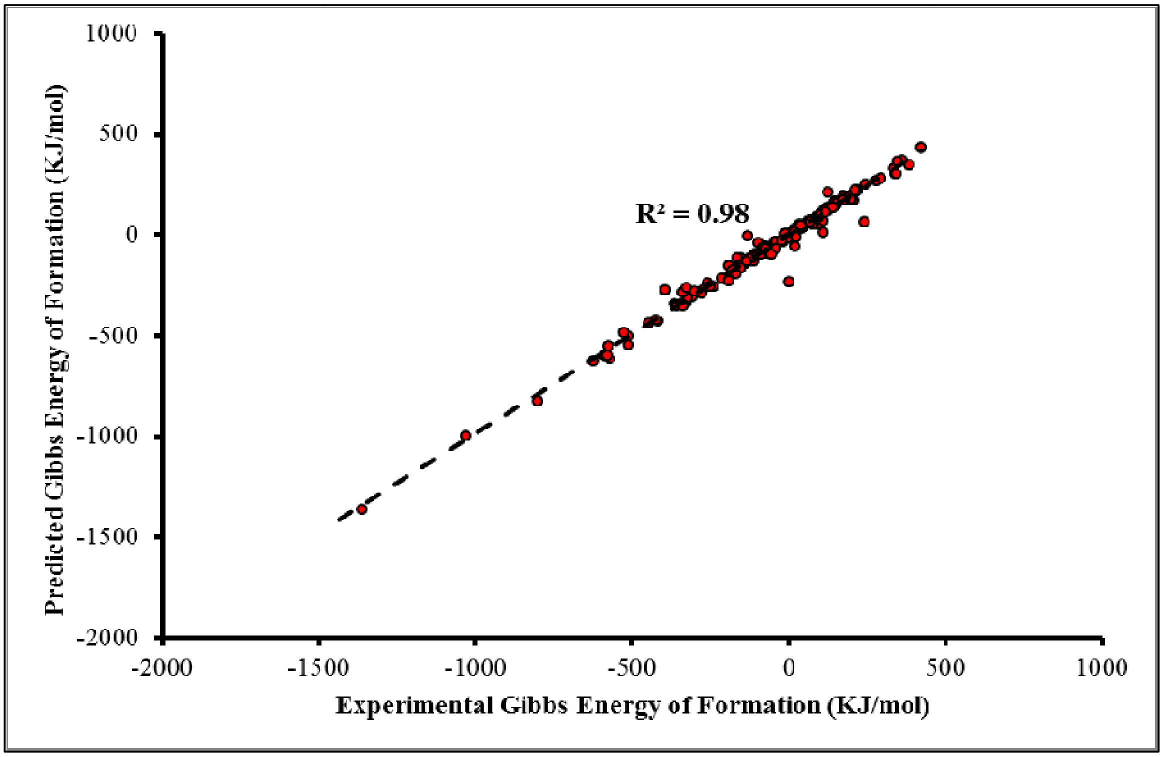


Figure 5.5: Comparison between the experimental and predicted ΔG_f values for the external test set in Model 2. The broken line represents perfect predictions

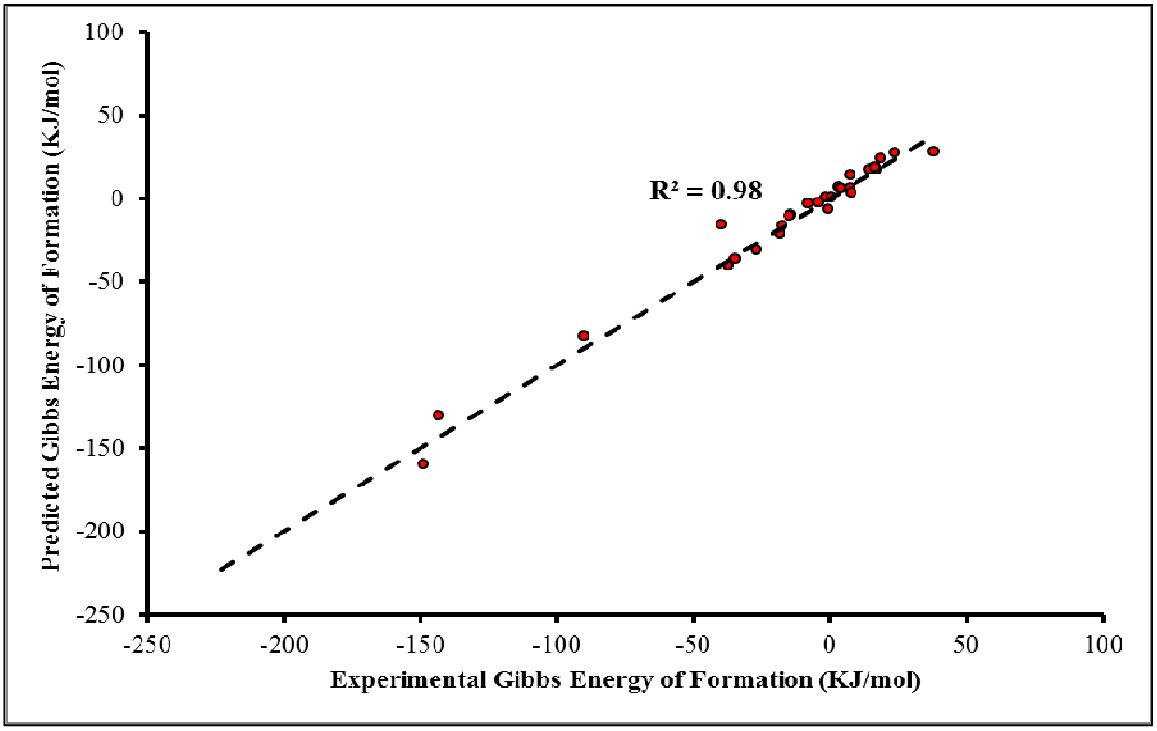


Figure 5.6: Comparison between the experimental and predicted ΔG_f values for the external test set on the Yan's database. The broken line represents perfect predictions

Table 5.1: Characteristics of the final OSU data set

Molecular Property	Minimum	Maximum	Mean	Standard Deviation
ΔG_f (kJ/mol)	-1970.0	665.0	-47.2	268.1
Molecular weight (g/mol)	16.05	446.74	129.3	55.5
Octanol-water partition coefficient(Log K_{ow})	-8.7	15.0	2.6	2.1
DRAGON drug like score (0-1)	0.49	1.0	0.78	0.1

Table 5.2: List of the descriptors used in the final eight ensembles for Model 1

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	nN	nN	nN	nN	nN	nN	nN	nN
2	SAdon	TDB03u	Ho_D/Dt	Ho_D/Dt	Ho_D/Dt	Ho_D/Dt	Ho_D/Dt	Ho_D/Dt
3	ZM1V	ZM1V	Eta_betaS_A	Eta_betaS_A	X%	X%	Eta_beta	Eta_beta
4	MAXDN	MAXDN	EE_H2	EE_H2	MAXDN	MAXDN	nBO	nBO
5	P1m	B01[O-O]	EE_B(s)	EE_B(s)	EE_B(s)	VR2_B(m)	EE_Dz(m)	EE_Dz(m)
6	P_VSA_v_3	P_VSA_v_3	SRW02	SRW02	P_VSA_v_3	P_VSA_v_3	Chi0_EA	Chi0_EA
7	P_VSA_v_2	P_VSA_v_2	P_VSA_p_3	P_VSA_p_3	nCar	Hypertens-80	P_VSA_p_3	P_VSA_p_3
8	SM6_Dz(m)	SM6_Dz(m)	SM1_Dz(Z)	SM1_Dz(Z)	SM3_D	SM3_D	SM1_Dz(Z)	SM1_Dz(Z)
9	B01[O-O]	BBI	SM6_B(p)	SM6_B(p)	SM4_B(e)	SM4_B(e)	L1v	Dz
10	SpMaxA_Dt	SpDiam_Dz(p)	SpDiam_EA(dm)	SpDiam_EA(dm)	ON0	ON0	nCbH	nCbH

Table 5.2 (cont'd): List of the descriptors used in the final eight ensembles for Model 1

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
11	SpMaxA_Dz(i)	Eta_F	SpMAD_Dz(p)	SM1_Dz(p)	nHAcc	nHAcc	SpMaxA_Dz(i)	SpMaxA_Dz(i)
12	SpPosLog_B(p)	SAdon	SpAD_AEA(ed)	SpAD_AEA(ed)	Eta_B	P_VSA_MR_2	SpPosA_Dz(v)	SpPosA_Dz(v)
13	SdO	SdO	SaaCH	SaaCH	SdO	SdO	NdO	NdO
14	VR3_Dz(Z)	VR3_Dz(Z)	ATS2s	ATS2s	ATS1p	ATS1p	ATS1s	ATS1s
15	SCBO	SCBO	TIC4	TIC4	S1K	P_VSA_m_4	X1Mad	X1Mad
16	RDF010p	RDF010p	Mor22e	Mor22e	P_VSA_m_4	nCar	nF	nF
17	MPC04	nRNHR	SM1_Dz(p)	HATS3m	P_VSA_MR_2	nROR	SpAD_RG	SpAD_RG
18	DLS_02	L1m	Mor20e	Mor20e	nROR	S1K	nRNHR	LPRS
19	nBM	nBM	HATS3m	---	nC	nC	LPRS	nRNHR
20	---	---	CATS2D_01_LL	---	EE_A	---	Dz	---

Table 5.3: Physical meaning of the commonly occurring descriptors in the ensembles for Model 1

Descriptor	**Descriptor Type**	**Physical Meaning**
nN	Constitutional descriptor	Number of nitrogen atoms in the molecule.
Ho_D/Dt	2D matrix based descriptor	Hosoya-like index, which is a topological index, calculated by applying a logarithmic function to the distance/detour matrix.
MAXDN	Topological Indices	Maximal electrotopological negative variation [13], which is an E-state index calculated as the maximum negative value of ΔI_i in the molecule. ΔI_i is the intrinsic state of the ith atom.
P_VSA_v_3	P_VSA descriptor	The amount of van der Waals surface area of the molecule that has a value of van der Waals volume between 1 and 1.3 [14].
P_VSA_p_3	P_VSA descriptor	The amount of van der Waals surface area of the molecule that has a value of polarizability between 1 and 2 [14].
SdO	E-state indices	Sum of the electrotopological state values of all '=O' atom types in the molecule [15].
SM1_Dz (Z)	2D matrix based descriptor	The spectral moment of order 1 from Barysz matrix weighted by atomic number [16].

Table 5.4: List of the descriptors used in the final eight ensembles for Model 2

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	nN	nN	nN	nN	nN	nN	nN	nN
2	nHet	nHet	nHet	nHet	nHet	nHet	nHet	nHet
3	MAXDN	MAXDN	MAXDN	MAXDN	MAXDN	MAXDN	MAXDN	MAXDN
4	C-024	C-024	C-024	C-024	C-024	C-024	C-024	C-024
5	ON0V	ON0V	ON0V	ON0V	ON0V	ON0V	ON0V	ON0V
6	P_VSA_v_3	P_VSA_v_3	P_VSA_p_3	P_VSA_p_3	P_VSA_p_3	P_VSA_p_3	P_VSA_p_3	P_VSA_p_3
7	SM15_EA(dm)	SM15_EA(dm)	F01[O-O]	F01[O-O]	SM15_EA(dm)	SM14_EA(dm)	SM15_EA(dm)	SM15_EA(dm)
8	RDF010s	RDF010s	RDF010s	RDF010s	RDF015m	RDF015m	RDF015m	RDF015m
9	SpDiam_B(i)	SpDiam_B(i)	SpDiam_B(i)	SpDiam_B(i)	SpDiam_G/D	SpDiam_G/D	SpDiam_G/D	SpDiam_G/D
10	SpAbs_B (v)	SpAbs_B (v)	SpAbs_B (v)	SpAbs_B (v)	SM5_B (p)	SM5_B (p)	SM5_B (p)	SpPos_H2

Table 5.4 (cont'd): List of the descriptors used in the final eight ensembles for Model 2

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
11	SpMaxA_B(m)	SpMaxA_B(m)	SpMaxA_B(m)	SpMaxA_B(m)	SpPosLog_L	SpPosLog_L	SpPosLog_L	SpPosLog_L
12	SM3_RG	SM3_RG	SM3_Dz (i)	SM3_Dz (i)	HyWi_B (i)	HyWi_B (i)	HyWi_B (i)	HyWi_B (i)
13	SpAD_Dz(i)	SpAD_Dz(i)	P_VSA_m_4	P_VSA_m_4	SM14_EA(dm)	SpAD_Dz(i)	P_VSA_p_1	P_VSA_p_1
14	VR1_H2	Eta_betaP_A	CATS2D_03_LL	CATS2D_03_LL	Eta_betaP	Eta_betaP	Eig14_EA(ed)	Eig14_EA(ed)
15	ATSC2p	ATSC2p	Mor01u	Mor01u	MLOGP	MLOGP	Chi0_AEA(bo)	Chi0_AEA(bo)
16	Gu	Gu	VR2_B(e)	VR2_B(e)	ALOGP	ALOGP	nBM	nBM
17	ATSC4p	VR1_H2	QXXm	QXXm	nX	nX	nX	nX
18	SM2_L	---	RDF020i	RDF020i	N%	---	SpPos_H2	---
19	Mor11s	Mor11s	Mor11s	---	GGI8	GGI8	GGI8	GGI8
20	---	---	---	---	---	---	SpMAD_Dz(p)	SpMAD_Dz(p)

Table 5.5: Physical meaning of the commonly occurring descriptors in the ensembles for Model 2

Descriptor	**Descriptor Type**	**Physical Meaning**
nN	Constitutional descriptor	Number of Nitrogen atoms in the molecule.
nHet	Constitutional descriptor	Number of heteroatoms in the molecule
MAXDN	Topological Indices	Maximal electrotopological negative variation [13], which is an E-state index calculated as the maximum negative value of ΔI_i in the molecule. ΔI_i is the intrinsic state of the ith atom.
C-024	Atom centered fragments	Number of carbon atoms of the type R—CH—R
ON0V	Topological indices	The overall modified Zagreb index of order 0 by valence vertex degrees [17].
P_VSA_p_3	P_VSA descriptor	The amount of van der Waals surface area of the molecule that has a value of polarizability between 1 and 2 [14].
SM15_EA(dm)	Edge adjacency indices	The spectral moment of order 15 from edge adjacency matrix weighted by dipole moment [18].

Table 5.6: The errors for all models developed in this work

Model	Training Set			External Test Set		
	RMSE (kJ/mol)	MAE (kJ/mol)	R^2	RMSE (kJ/mol)	MAE (kJ/mol)	R^2
Model 1	17.4	9.7	0.99	---	---	---
Model 2	21.4	11.4	0.99	32.4	16.4	0.98
Model for Yan data	21.5	16.6	0.99	29.2	21.1	0.98

Table 5.7: Comparison of the current model with literature models on the basis of predictions on external test set molecules

Researchers	Type of model	No. of molecules in the external test set	RMSE (kJ/mol)
This work	Stochastic optimization and ANNs	177	32.4
Wang et al. [8]	Density-functional theory and ANNs	30	13.7
Yan [9]	Pair-wise correlation analysis and ANNs	47	46.9

Table 5.8: List of molecules in the external test set for Model 2 that had an absolute error of more than 100 kJ/mol

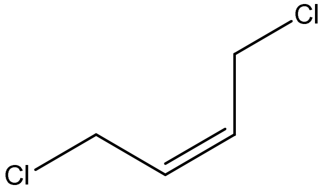
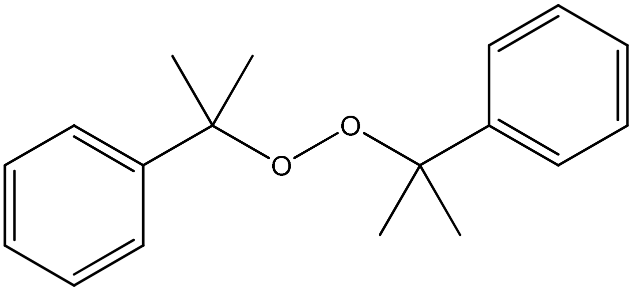
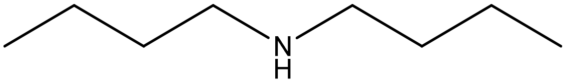
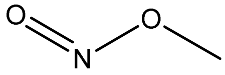
Name	Structure	Experimental ΔG_f (kJ/mol)	Predicted ΔG_f (kJ/mol)
1,4-dichloro-cis-2-butene		108.5	11.3
Dicumyl peroxide		242.0	64.7
Di-n-butylamine		-130.0	-7.6
Methyl nitrite		1.0	-232.5
Carbon di-oxide	O=C=O	-394.4	-274.6

Table 5.9: Comparison of the current model with literature models on the Yan data set

Researchers	Model Type	Training Set MAE	Number of Molecules in External Test Set	External Test Set MAE
This work	Stochastic optimization and ANNs	16.6	45	21.1
Yan [9]	Pair-wise correlation analysis and ANNs	48.1	27	46.9
Wang et al. [8]*	Density-functional theory correction using ANNs	13.4	30	13.0

* The external test set used in the referenced work is different from the one employed by the other models

REFERENCES

1. van Krevelen, D.W. and H.A.G. Chermin, *Estimation of the Free Enthalpy (Gibbs Free Energy) of Formation of Organic Compounds from Group Contributions.* Chemical Engineering Science, 1951. **1**(2): p. 66-80.

2. Castro, E.A., et al., *QSPR Modeling of Gibbs Free Energy of Organic Compounds by Weighting of Nearest Neighboring Codes.* Structural Chemistry, 2005. **16**(3): p. 305-324.

3. Klamt, A., et al., *Prediction of Aqueous Solubility of Drugs and Pesticides with COSMO-RS.* Journal of Computational Chemistry, 2002. **23**(2): p. 275-281.

4. Joback, K.G., *A Unified Approach to Physical Property Estimation Using Multivariate Statistical Techniques,* Department of Chemical Engineering. 1984, Masachussets Institute of Technology: Masachussets.

5. Constantinou, L. and R. Gani, *New Group Contribution Method for Estimating Properties of Pure Compounds.* AIChE Journal, 1994. **40**(10): p. 1697-1710.

6. Mavrovouniotis, M.L., *Group Contributions for Estimating Standard Gibbs Energies of Formation of Biochemical Compounds in Aqueous Solution.* Biotechnology and Bioengineering, 1990. **36**(10): p. 1070-1082.

7. Ivanciuc, O., et al., *Evaluation in Quantitative Structure−Property Relationship Models of Structural Descriptors Derived from Information-Theory Operators.* Journal of Chemical Information and Computer Sciences, 2000. **40**(3): p. 631-643.

8. Wang, X., et al., *Improving the Accuracy of Density-Functional Theory Calculation:□ The Statistical Correction Approach.* The Journal of Physical Chemistry A, 2004. **108**(40): p. 8514-8525.

9. Yan, A., *Modeling of Gibbs Energy of Formation of Organic Compounds by Linear and Nonlinear Methods†.* Journal of Chemical Information and Modeling, 2006. **46**(6): p. 2299-2304.

10. Yaws, C.L., *Chemical Properties Handbook.* 1999, McGraw-Hill.

11. *DRAGON Professional 6*. 2010, Talete SRL.

12. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

13. Gramatica, P., M. Corradi, and V. Consonni, *Modelling and Prediction of Soil Sorption Coefficients of Non-Ionic Organic Pesticides by Molecular Descriptors.* Chemosphere, 2000. **41**(5): p. 763-777.

14. Paul, L., *A Widely Applicable Set of Descriptors.* Journal of Molecular Graphics and Modelling, 2000. **18**(4-5): p. 464-477.

15. Hall, L.H., L.B. Kier, and B.B. Brown, *Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices.* Journal of Chemical Information and Computer Sciences, 1995. **35**(6): p. 1074-1080.

16. Ivanciuc, O., *QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs.* Journal of Chemical Information and Computer Sciences, 2000. **40**(6): p. 1412-1422.

17. Danail, B., *Overall Connectivity — a Next Generation Molecular Connectivity.* Journal of Molecular Graphics and Modelling, 2001. **20**(1): p. 65-75.

18. Estrada, E., *Spectral Moments of the Edge Adjacency Matrix in Molecular Graphs. 1. Definition and Applications to the Prediction of Physical Properties of Alkanes†.* Journal of Chemical Information and Computer Sciences, 1996. **36**(4): p. 844-849.

CHAPTER 6

A NON-LINEAR QSPR MODEL FOR NORMAL BOILING POINT TEMPERATURE

6.1. Introduction

Boiling point is an important thermophysical property that is defined as the temperature at which the liquid and vapor phases of a pure substance co-exist in equilibrium. If measured at atmospheric pressure, the boiling point is referred to as the normal boiling point temperature (henceforth called NBP). Boiling point is one of the properties typically investigated first [1], when identifying new compounds. Apart from estimating the volatility of a compound, NBP information can also be used along with flash point data, to assess the flammability of the compound. Further, the NBP is also used widely to predict other physical properties including critical temperature [2], enthalpies of vaporization [3, 4], flash points [5] and gas chromatographic retention indices [4].

In most extractive distillation process, the solvents used are expensive and are, therefore, recovered and recycled. The NBP of a solvent often determines the process layout. When a solvent with a low-boiling point is used, the solvent is usually recovered along with one of the solutes from the condenser stream; however, the use of a high-boiling point solvent requires recovery from the re-boiler. Therefore, the boiling point is often the first property measured for a new solvent.

Normal boiling points are easy to determine experimentally; however, when a chemical is unavailable, hazardous to handle or yet to be synthesized, a reliable procedure to estimate its boiling point is required. In fact, the rapid growth of combinatorial chemistry provides large numbers of prospective new molecules, which then need to be synthesized and tested; thus, providing the opportunity and impetus for the development of an accurate predictive model for NBP predictions.

The models available in the literature for predicting the NBP values are discussed in the next section. The majority of these models are developed using limited data, and hence, their general applicability is limited. In the current work, efforts have been made to develop QSPR models with much wider applicability using a much larger database of NBP values. This work focuses on the following objectives:

1. Develop an accurate non-linear QSPR model to predict the NBP values using a database made up of diverse set of compounds.
2. Validate the current modeling approach by employing an external test set of compounds that has not been used to develop the model.
3. Compare the current modeling approach with existing approaches in the literature, on common training and external set data. This would further establish the efficacy of the modeling approach used in the work.

6.2. State of the Art in Predicting Normal Boiling Point Temperatures

According to Katritzky [6], the boiling point of a compound is determined by the intermolecular forces in the liquid state, and by the difference in the molecular internal partition function between the vapor and liquid phases. Therefore, the boiling point temperature of a compound should be predictable from its chemical structure. Accordingly, many models have been developed to correlate the NBP values with the molecular structure of the compounds. One of the first reported efforts was by Walker [7], who attempted to correlate the boiling point with the number of carbon atoms and molecular weight. Horvath [8], Nendza [9], Lyman et al. [10], and Katritzky et al. [6] have summarized the early work (until the 1990's) on boiling point prediction. The majority of the early prediction approaches were based on group-contribution methods (GCM), where any molecular property is assumed to be a sum of contributions from predefined groups of atoms in that molecule. Joback and Reid [11] developed one of the earliest GCM approaches for prediction of melting points and boiling points along with other physical properties. They reported a mean absolute error (MAE) of 12.9 K for a database of 438 compounds. Later, Constantinou and Gani [12] developed a GCM approach based on UNIversal Functional Activity Coefficients (UNIFAC) groups that lead to better correlations than a simple GCM approach by considering second-order group interactions. They report a MAE value of 5.4 K on their training data (data that has been used for model development) of 392 compounds, extracted from the Design Institute for Physical Properties Research (DIPPR) database [13]. Stein and Brown [14] have improved the Joback and Reid [11] approach, primarily by increasing the number of functional groups. They employed a training set database of 4426 compounds and an external test (data that has not been used for model development) set of 6584 compounds. Their model produced MAE values of 15.5 K and 20.4 K for the training and test sets, respectively; however, the majority of their experimental data were measured at pressures lower than 1 atm, and these data were then extrapolated to atmospheric pressure using a vapor pressure equation.

Despite their popularity, GCM approaches suffer from major disadvantages such as their inability to model structures containing undefined functional groups and to account for the interaction between different functional groups and for their spatial arrangement. An alternative to the GCM approach is the quantitative structure-property relationship (QSPR) method, where the entire molecule is parameterized using molecular descriptors calculated through molecular mechanics or quantum mechanical methods. Using a data set of about 150 compounds, Sola and coworkers [15] demonstrated that the QSPR approach to modeling the NBP values is more accurate than the best available GCM approach. The pioneering work in predicting boiling points using QSPR techniques was carried out by Wiener [16], who introduced the path number (Wiener index), which is defined as the sum of the distances between any two carbon atoms in the molecule. Using this descriptor, Wiener [16] was able to calculate the boiling points of 94 paraffins within a deviation of one degree Celsius. Other early contributions include the topological indices developed by Randic [17], and Kier and Hall [18] which have been employed successfully to model the boiling points of alkanes and amines. More recently, a plethora of QSPR models for boiling point prediction have appeared in the literature, with the majority of the developed models dealing with a specific class of compounds such as alkanes [19-24]. Dearden [25] provides a detailed review of these methods and also tests them using an external test set of 100 organic molecules. Dearden [25] notes that almost all the models have standard errors in single figures and employ graph theoretical descriptors, also known as topological descriptors; these descriptors are helpful in describing the branching in molecules.

Dearden [25] also reviews some of the generalized QSPR models that are based on a diverse set of compounds, such as the models based on the Comprehensive Descriptors for Structural and Statistical Analysis (CODESSA) [26] and ADAPT [24] software. One of the most accurate generalized models for NBP prediction was developed by Hall and coworkers, who employ E-state indices [27, 28] and report a MAE value of 3.93 K for a training set of 298 compounds, and a MAE value of 3.86 K for an external test set of 30 compounds. Katritzky et al. [29] developed a generalized QSPR model using 584 diverse organic compounds for training and 28 additional compounds (mostly fluorinated and chlorinated compounds) as an external test set. They report a root-mean-squared error (RMSE) value of 14.6 K for the training set and a RMSE value of 9.7 K on the test set, which is comparable to the estimated experimental RMSE of 11.4 K for the entire data set. Chalk and coworkers [30] developed a generalized model for predicting NBP values based on semi-empirical molecular orbital (MO) descriptors, using a large training data set of 6000 compounds and a representative external test set of 629 compounds. However, the experimental errors for this data were not available and therefore, the quality of the resulting model is questionable. They report standard deviations of 16.5 K and 19.0 K for the training and test sets, respectively.

Despite the availability of sufficient experimental NBP data, the majority of generalized QSPR models in the literature are trained using fewer than 300 compounds. The only comprehensive model using a reasonably sized data set was developed by Katritzky et al. [29], but their model was not tested sufficiently using an external test set. In this work, we augmented the data provided by Katritzky et al. [29] with additional data from DIPPR [13] to develop generalized NBP models and tested these models for their predictive ability using an external test set.

6.3. QSPR Methodology

The development of a QSPR model involves the following series of steps: (a) data set generation, (b) descriptor calculation, (c) descriptor reduction and model training, and (d) model validation. These elements are described below.

6.3.1. Data Set Generation: Experimental NBP values were extracted from the DIPPR database. The DIPPR database provides an estimated maximum error for each datum, and only data that have an estimated error of less than 5% were used for training the models in the current work. In total, the DIPPR database has 1,317 NBP values with estimated maximum errors less than 5%. Of these, 101 compounds are either inorganics or salts and were removed from the database. The pruned DIPPR database was combined with the database employed by Katritzky and coworkers [31]. Katritzky's database [29] is made up of data from DIPPR, the CRC handbook of chemistry and physics [32], and the Aldrich catalog of fine chemicals [33]. The experimental uncertainties for the specialty fine chemicals are not given but are expected to be higher than 10%. After removing duplicates, the combined database has values for 1,321 compounds; however, the final OSU-NBP database used for modeling is made up of 1,320 NBP values, after removing phosphoric acid ester, which could not be optimized structurally for the most favorable (lowest energy) three-dimensional conformation using our automated procedure (see Section 6.2.2).

The OSU-NBP database is one of the most comprehensive databases available for NBP values in the open literature. The NBP values of the molecules in this database are in the range of 111.66 K to 716.15 K. Figure 6.1 provides the distribution of NBP values in the OSU-NBP database. The molecular weights of these compounds vary from 16.05 g/mol to 607.44 g/mol, and the octanol-water partition coefficient, K_{ow}, (calculated by the DRAGON [34] software using the Ghose-Crippen ALOGP model) varies between -2.3 and 12.9. In addition, the molecules are characterized based on their drug-likeness as calculated using DRAGON [34], where score of 0 implies that the molecule has no characteristics of a drug, while a score of 1 implies the molecule has all the characteristics of being a drug. Further details on the database characterization are provided in Table 6.1.

In addition to the above data sets, NBP data of 394 diverse organic compounds were extracted from the article by Ghavami et al. [35]. This data contains 52 alcohols, 22 amines, 69 alkanes, 156 mono-alkenes, 9 ethers, 69 alkyl benzenes, and 17 alkyl halides. Henceforth in this work, this data will be referred to as the Ghavami's database to differentiate it from the OSU-NBP data set. To validate the current modeling approach, Ghavami's data were used to develop a QSPR model to predict the NBP values and the resulting model was compared with the prediction results by Ghavami and coworkers

[35]. To ensure a fair comparison, the same training and external test data employed by Ghavami et al. [35] were used in the current work.

6.3.2. Descriptor Calculation: See Section 2.5

6.3.3. Descriptor Reduction and Model Development: See section 2.6

External Validation: In a recent article, Tropsha et al. [36] emphasized the need to validate QSPR models using external data sets. Therefore, another model was built by separating some data from the original OSU-NBP database and allocating it to an external test set. However, the data cannot be randomly separated, as the external set might not be representative of the training set. Therefore, a self-organizing map (SOM) network was created using the best descriptors identified in the first ensemble, which was developed using the entire database. This SOM was used to identify clusters in the data and partition the data into T, IV and IT sets as explained in Section 6.3.3. The number of map units in this SOM was varied until the percentage of data points in the IT set is at least 15% of the size of the entire OSU_NBP data set of 1320 molecules. This IT set was then set aside as an external test set and the remaining data was used for developing another model *de novo*, by repeating the search for the best descriptors, best network architecture and network weights. In the current work, 203 molecules were identified as an external test set using this procedure, and the remaining 1,117 data points were again divided into T, IV and IT sets and subjected to the descriptor search algorithm as discussed in Section 6.3.3. For clarity in this work, the model created using all 1,320 data points in the OSU-NBP database for model development will be referred to as Model 1 and the model created using just the 1,116 data points as Model 2. Model 1 will be used in the computer-aided molecular design (CAMD) algorithms because of its larger training set size, and Model 2 will be used to assess the generalization capability of Model 1, as advocated by Tropsha et al. [36].

6.4. Results

6.4.1 Model 1: 10-descriptor, 15-descriptor, and 20-descriptor-models were tested, but no significant difference was observed between the models. Therefore, for the sake of simplicity, 10-descriptor models were used in the final models in the current study. Going lower than 10 descriptors resulted in a significant increase in the training RMSE values for databases made up of more than 150 data points. Therefore, ten was chosen as the minimum number of input descriptors. Figure 6.2 is a comparison between the experimental and predicted NBP values for Model 1. The correlation coefficient (R^2) between the experimental and predicted values is 0.97. The prediction residual errors in K are plotted in Figure 6.3 for Model 1, which clearly demonstrates that the residuals are almost symmetrically distributed around the horizontal axis, as should be expected from an unbiased model. A histogram of the residuals (not shown) was plotted, where the distribution of the residuals around zero was found to be similar to a normal distribution. The RMSE, MAE, and the average absolute percentage deviation (%AAD) values for the training data set predictions are 14.4 K, 9.3 K, and 2.3%, respectively, and the RMSE values for the individual ensembles range from 14.7 K to 16.8 K. The results from the overall ensemble are slightly better than the results for the individual ensembles, which validates the use of ensembles with different descriptors as inputs.

The different descriptors used for creating the eight different ensembles are shown in Table 6.2. Note that the neural networks in the ensembles are allowed to have a maximum of 10 elite inputs, but most ensembles frequently have a lower number of elite descriptors as inputs, after the insignificant descriptors have been removed as described in Section 6.3.3. The descriptors AMR, P_VSA_p_2, piID, TIC0, and SpPosLog_G are the most common to all the ensembles. The types and physical meanings of these commonly occurring descriptors, as extracted from the DRAGON [34] help file, are provided in Table 6.3.

6.4.2. Model 2: For Model 2, 10-descriptor models were chosen as the final models. Figure 6.4 is a comparison between the experimental and predicted NBP values of the training data of 1,117 compounds. The correlation coefficient (R^2) between the experimental and predicted training data is 0.98. The prediction residual errors on this data are near-symmetrically distributed around the horizontal axis (no Figure shown**)**. The RMSE, MAE, and %AAD values for the training set data are 13.1 K, 8.6 K, and 2.1%, respectively. Figure 6.5 compares the 204 experimental and predicted NBP values of the external test. The RMSE, MAE, and %AAD values for the external test set are calculated to be 17.8 K, 10.2 K, and 2.6%, respectively.

The descriptors used for creating the eight different ensembles for Model 2 are tabulated in Table 6.4. The descriptors AMR, P_VSA_p_2, GATS1s, nHM, and Eta_sh_p are the most common across all the ensembles. The types and physical meanings of these commonly occurring descriptors, extracted from the DRAGON [34] help file are provided in Table 6.5.

6.4.3. Model for Ghavami's Database: Ghavami et al. [35] had 10 topological descriptors to develop their model, and so to ensure a fair comparison, 10 descriptor-models were developed in the current work as well. For the current model, the RMSE and MAE values are calculated to be 1.8 K and 1.3 K, respectively for the training set comprising 354 compounds. For the external test set comprising 40 compounds, the RMSE and MAE values are calculated to be 2.1 K and 1.5 K, respectively. A comparison between the experimental and predicted NBP values for the external data of 40 compounds is provided in Figure 6.6. The correlation coefficient (R^2) between the experimental and predicted external test data is calculated to be nearly 1.0.

The errors for all models developed in the current work are tabulated in Table 6.6.

6.5. Discussion

The RMSE values for the training set of both Model 1 and Model 2 are almost equal. Due to the larger training set, Model 1 would be expected to perform similarly to Model 2 on unseen data (external dataset). Therefore, the predictive performance of Model 2 on an external test set can be used as an approximation for determining the generalization capability of Model 1. The %AAD value for the external test set is within the maximum experimental uncertainty (5%) in the data used for modeling.

Few recent works in the open literature employ an external test set comprised of diverse molecules to test the predictive capability of a developed NBP model. To the best knowledge of the author, the models by Stein and Brown [14], and Chalk et al. [30] are the only models that are developed using a diverse set of organic molecules, and they are also assessed for their predictive capability using an external data set (Table 6.7). Stein and Brown [14] employ experimental boiling point data measured at sub-atmospheric pressures, which are then extrapolated to atmospheric pressure using a vapor pressure equation. They report a RMSE value of 20.4 K on an external test set of about 6,500 molecules, using a GCM approach. Chalk et al. [30] employ quantum mechanics and artificial neural networks (ANNs) to develop their QSPR models and they report RMSE and MAE values of 19 K and 13 K, respectively, on an external test set of 629 molecules, using a model that had been trained on a data set of 6000 molecules. Model 2 from the current work resulted in RMSE and MAE values of 17.8 K and 10.2 K respectively, for an external set of 203 compounds. These results are better than the results reported by Chalk et al. [30] for their external test set. Also, the compounds that exhibit the largest deviations in the various models were further examined manually to identify any correlation between their higher errors and the presence/absence of certain functional groups. However, no particular trends were observed between the functional groups present in the molecule and the prediction error for the molecule. The higher errors for some molecules could be due to the high experimental uncertainty in the data for those molecules.

Most of the descriptors in Table 6.2 are 2D descriptors or constitutional descriptors and are independent of the 3D conformation of the molecule. Table 6.3 lists the most common descriptors for the eight different ensembles for Model 1. Due to the black-box nature of the ANNs, a quantitative assessment of the significance of the different descriptors on the NBP values is not possible; however, approximate qualitative interpretations can be made based on the type of descriptors. For example, AMR, which denotes the molar refractivity calculated according to the Ghose-Crippen model [37], occurs in all eight ensembles and therefore, must be correlated with NBP. Egolf and Jurs [38] have also reported a correlation between NBP and molar refractivity, and they attribute molar refractivity to be a measure of the polarizability of the molecule, which consequently describes the ability of a molecule to form bonds with neighboring molecules in the liquid state. The descriptor P_VSA_p_2, which also describes the polarizability of a molecule, was found to be occurring frequently across the ensemble. Additional commonly occurring descriptors in the ensembles are piID and TIC0, which describe the degree of unsaturation (presence of multiple bonds) present in the molecule and the molecular complexity, respectively. In addition, Table 6.2 contains some 2D-matrix based descriptors that describe the 2-dimensional shape of the molecule.

As expected, some of the common descriptors for Model 2 (listed in Table 6.5) are identical to the common descriptors for Model 1. For example, the descriptors AMR and P_VSA_p2 are common across both Model 1 and Model 2. The other descriptors in Model 2, though not exactly identical to Model 1 descriptors, are drawn from the same sub-category of descriptors as in Model 1. A few descriptors in both Models 1 and 2 are either 2D matrix-based descriptors or other descriptors that describe the shape of the molecule. This suggests some correlation between the shape and NBP values of a molecule.

To compare the efficacy of the current modeling approach, the Ghavami's data set had been employed to develop a QSPR model. The results from this model are provided in

Table 6.8, along with the results by Ghavami et al. [35], using the same training and external test set data. The current model performs significantly better on both the training data and the external set data, which indicates better generalization capability of the current model. The poor performance of the model by Ghavami et al. [35] could be due to the absence of 3D molecular descriptors in their modeling, which proves the efficacy of 3D QSPR modeling when compared to 2D QSPR modeling.

6.6. Conclusions

1. In the current work, a non-linear QSPR model for the normal boiling point prediction was developed using wrapper-based descriptor pruning techniques.
2. Two models were developed in the current work: Model 1 was built using NBP values for 1,320 compounds, where all data was used for model development, and Model 2 was developed using just 1,116 compounds from the OSU-NBP database, while the remaining 204 compounds were employed as an external test set.
3. The RMSE values on the training sets for Model 1 and Model 2 are 14.7 K and 13.1 K, respectively. The RMSE value for Model 2 on the external test set is 17.8 K. The models by Stein and Brown [14], and Chalk et al. [30] are the only works in the literature for predicting the NBP values for a wide range of molecular classes using an unbiased external test set. The predictive accuracy of Model 2 from this work, on an external test set of 204 compounds is better than the accuracy of the models by Stein and Brown [14], and Chalk et al. [30] (Table 6.8).
4. According to the descriptors identified by the current work, the polarizability and the 2-dimensional shape of the molecule significantly affect the NBP values.

5. The current model developed using the Ghavami data set performs significantly better than the model by Ghavami and co-workers [35] on an external test set of 40 compounds. The RMSE value on the external test set for the model by Ghavami et al. [35] is 6.8 K as opposed to RMSE value of 2.1 K from the current model.
6. The resulting models from this work can be used to accurately predict *a priori* the NBP values of organic compounds.

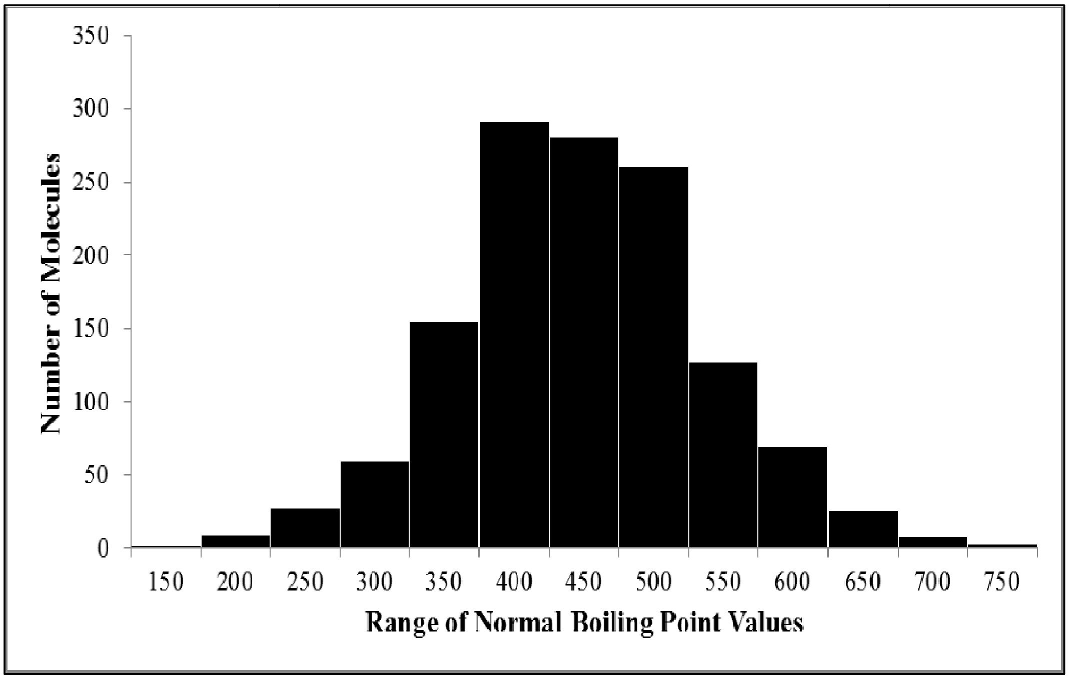


Figure 6.1: Distribution of the normal boiling points in the OSU-NBP data set

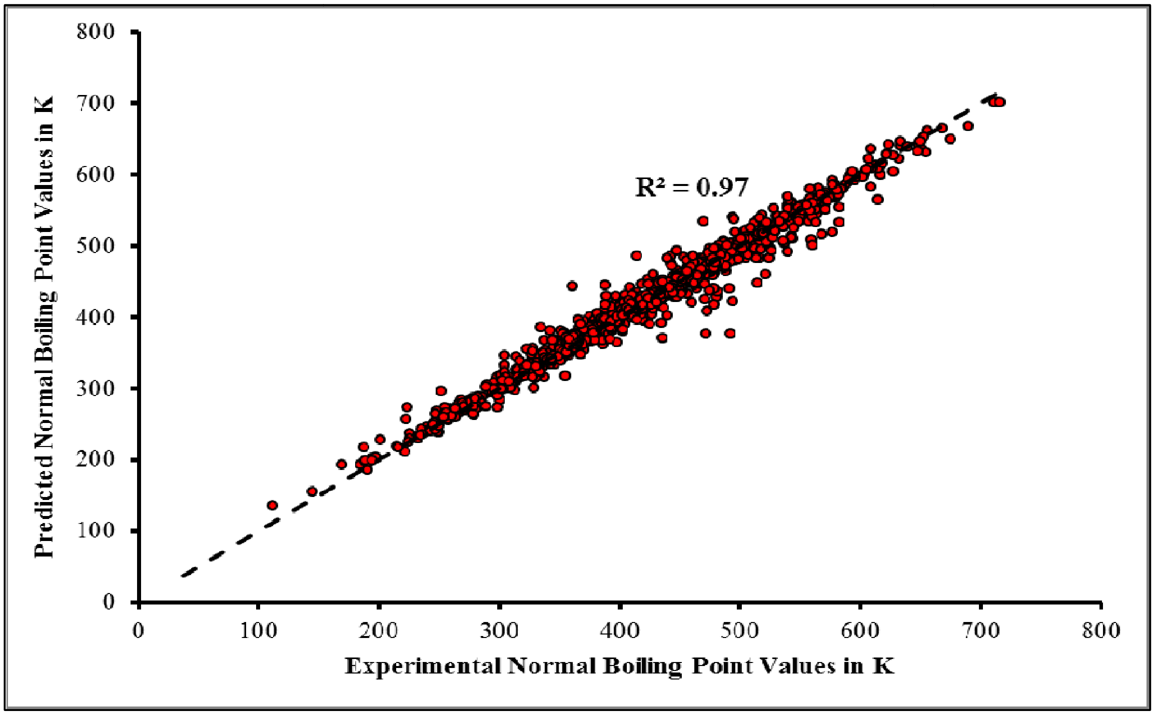


Figure 6.2: Comparison between the experimental and predicted NBP temperatures for Model 1. The broken line represents perfect predictions

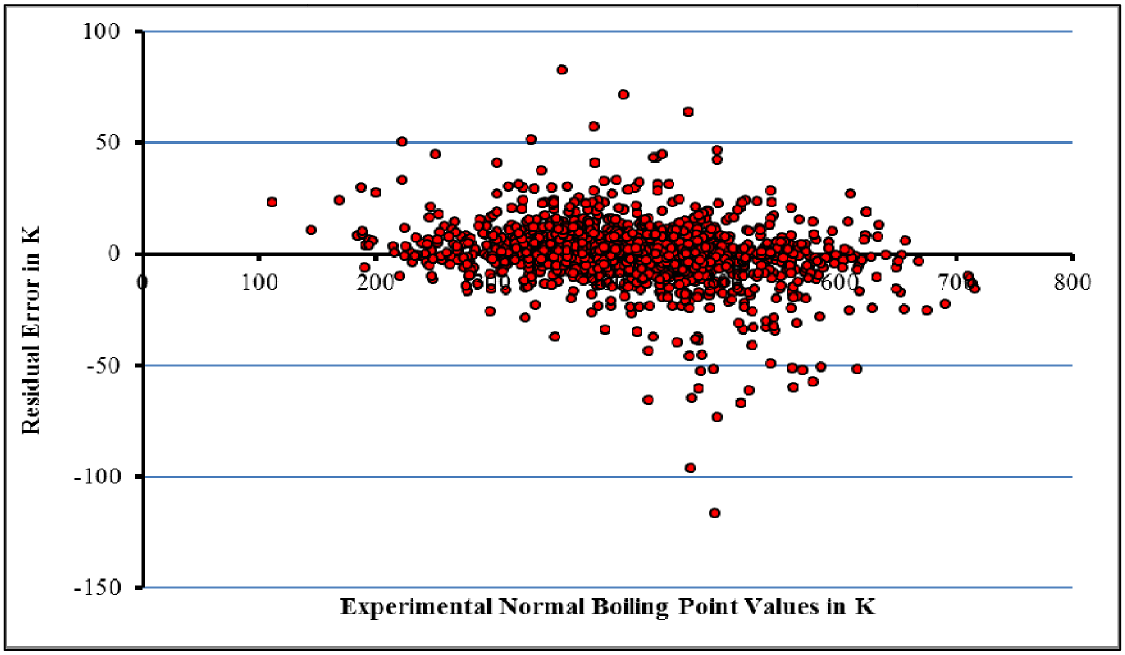


Figure 6.3: Residual error plot for Model 1 predictions

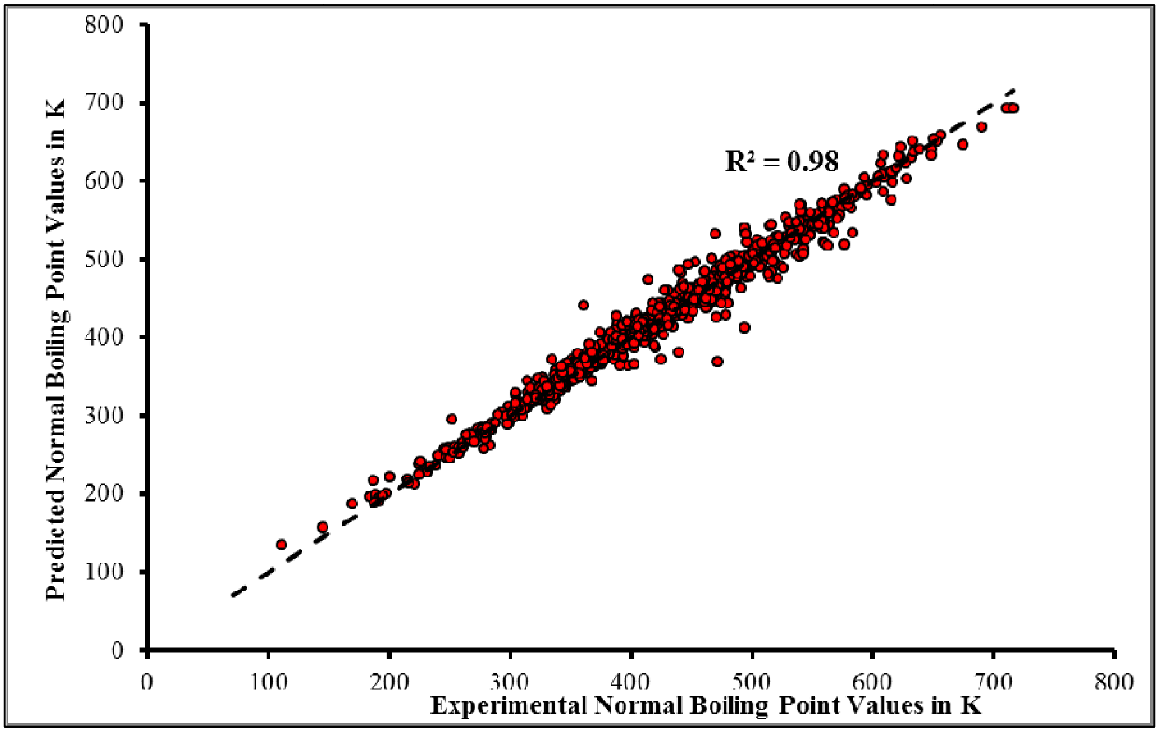


Figure 6.4: Comparison between the experimental and predicted NBP temperatures for the training set in Model 2. The broken line represents perfect predictions

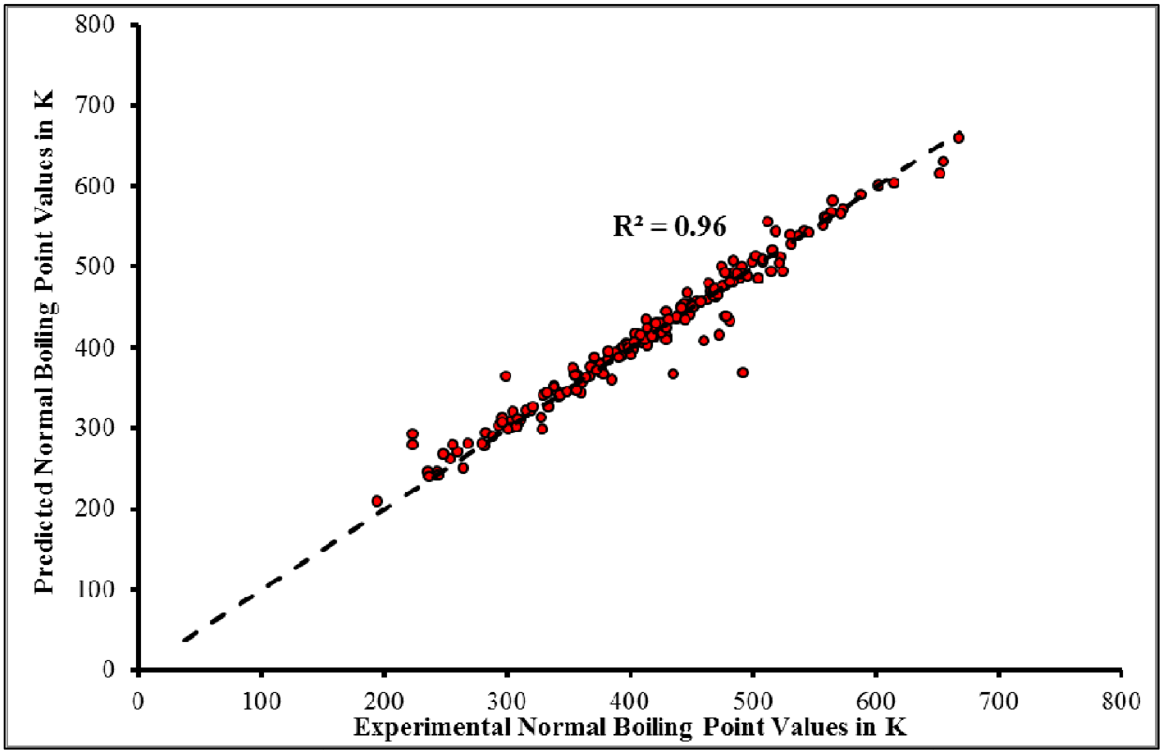


Figure 6.5: Comparison between the experimental and predicted NBP temperatures for the external test set in Model 2. The broken line represents perfect predictions

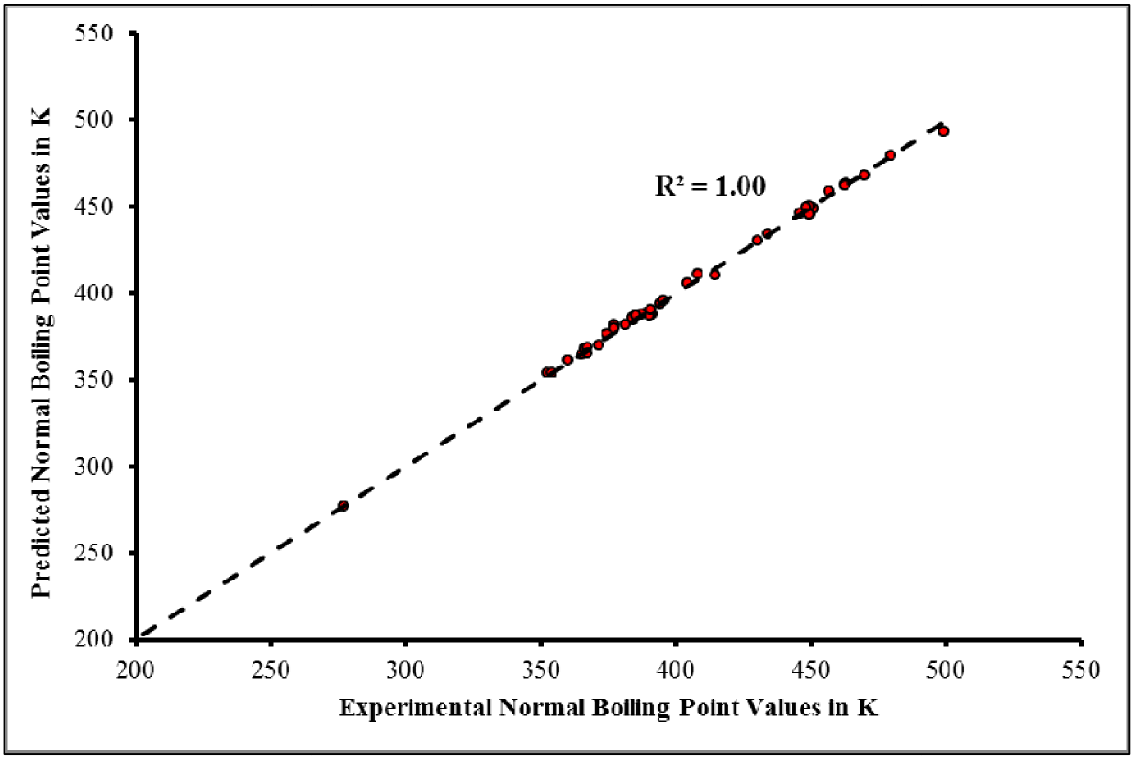


Figure 6.6: Comparison between the experimental and predicted NBP temperatures for the external test set compounds in the Ghavami database [35]. The broken line represents perfect predictions

Table 6.1: Characteristics of the final data set of 1,320 molecules

Molecular Property	Minimum	Maximum	Mean	Standard Deviation
NBP (K)	111.7	716.2	422.5	88.7
Molecular weight (g/mol)	16.0	607.4	127.6	54.9
Octanol-water partition coefficient(Log K_{ow})	-2.3	12.9	2.2	1.7
DRAGON drug like score (0-1)	0.49	1.0	0.80	0.10

Table 6.2: List of the descriptors used in the final eight ensembles for Model 1

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	AMR	AMR	AMR	AMR	AMR	AMR	AMR	AMR
2	P_VSA_p_2	P_VSA_p_2	P_VSA_v_2	P_VSA_v_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2
3	piID	piID	AAC	AAC	X2	X2	piID	piID
4	Ho_B (s)	Ho_B (s)	SM6_Dt	SM6_Dt	TIC0	TIC0	TIC0	TIC0
5	H%	H%	NssO	NssO	SpPosLog_G	---	SpPosLog_G	SpPosLog_G
6	SM08_EA (dm)	SM08_EA (dm)	SpPosA_A	SpPosA_A	N-072	N-072	SpPosLog_D	SpPosLog_D
7	VE2_L	VE2_L	VR3_Dz (m)	VR3_Dz (m)	SpAbs_Dz (p)	SpAbs_Dz (p)	Ho_Dz(p)	Ho_Dz(p)
8	piPC02	SM02_AEA (ed)	SM4_RG	SM4_RG	Mp	Mp	Chi0_EA	Chi0_EA
9	RDF010u	RDF010u	SpMAD_RG	---	SpMAD_Dt	SpMAD_Dt	---	---

Table 6.3: Physical meaning of the commonly occurring descriptors in the ensembles for Model 1

Descriptor	Descriptor Type	Physical Meaning
AMR	Molecular property	Ghose-Crippen molar refractivity
P_VSA_p_2	P_VSA like descriptor	The amount of van der Waals surface area of the molecule that has a value of polarizability between 0.4 and 1 [39]
piID	Walk and path count	The total number of weighted paths obtained by summing the weights of all paths of any length (from 0 to the maximum path length) in the graph. The weight of each path is calculated by multiplying the conventional bond order of all the edges of the path [40]
TIC0	Information index	Calculated as nAT times IC0, nAT being the total number of molecule atoms, and IC0 being the mean information content of order 0 [41]
SpPosLog_G	3D matrix based descriptor	Logarithmic spectral positive sum from geometrical matrix (a square matrix of Euclidian distances for each pair of atoms in the molecule).

Table 6.4: List of the descriptors used in the final eight ensembles for Model 2

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	AMR	AMR	AMR	AMR	AMR	AMR	AMR	AMR
2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2	P_VSA_p_2
3	GATS1s	GATS1s	GATS1s	GATS1s	GATS1s	GATS1s	ChiA_RG	ChiA_RG
4	RBF	RBF	nHM	nHM	F01[C-O]	F01[C-O]	nHM	nHM
5	Eta_sh_p	Eta_sh_p	EE_D/Dt	EE_D/Dt	SM6_D/Dt	SM6_D/Dt	Eta_sh_p	Eta_sh_p
6	SM1_Dz (p)	SM3_RG	nRCONHR	nRCONHR	Ho_Dz (i)	Ho_Dz (i)	VR1_G/D	VR1_G/D
7	SM1_Dz (Z)	SM1_Dz (Z)	Mv	Mv	Eta_betaS_A	Eta_epsi_A	R2u	---
8	TDB01s	TDB01s	BLTA96	---	XMOD	---	IDM	IDM
9	---	---	---	---	---	---	SM02_AEA (bo)	SM02_AEA (bo)

Table 6.5: Physical meaning of the commonly occurring descriptors in the ensembles for Model 2

Descriptor	Descriptor Type	Physical Meaning
AMR	Molecular property	Ghose-Crippen molar refractivity
P_VSA_p_2	P_VSA like descriptor	The amount of van der Waals surface area of the molecule that has a value of polarizability between 0.4 and 1 [39]
GATS1s	2D autocorrelation	Geary coefficient, calculated from molecular graph by summing the products of intrinsic states of the terminal atoms of all the paths of unit path length. Geary coefficient is a distance-type function varying from zero to infinite. Strong spatial autocorrelation produces small values of this index [42]
nHM	Constitutional index	Number of heavy atoms
Eta_sh_p	ETA index	Eta p shape index

Table 6.6: The errors for all models developed in this work

Model	**Training Set**			**External Test Set**		
	RMSE (K)	MAE (K)	R^2	RMSE (K)	MAE (K)	R^2
Model 1	14.4	9.3	0.97	---	---	---
Model 2	13.1	8.6	0.98	17.8	10.2	0.96
Model for Ghavami data	1.8	1.3	1.00	2.1	1.5	~1.00

Table 6.7: Comparison of the current OSU-NBP model with literature models, on the basis of external test set predictions

Researchers	Model Type	Number of Molecules in External Test Set	RMSE (K)
This work (Model 2)	Stochastic optimization and ANNs	204	17.8
Chalk et al. [30]	Quantum mechanics and ANNs	629	19.0
Stein and Brown [14]#	Group-contribution method (GCM)	6584	20.4

# Majority of the boiling point data, were measured at pressures less than 1 atm., and then extrapolated to atmospheric pressure

Table 6.8: Comparison of the current model with the model by Ghavami et al. [35] on the Ghavami database

Researchers	Model Type	Training Set RMSE (K)	External Test Set RMSE (K)
This work	Stochastic optimization and ANNs	1.8	2.1
Ghavami et al. [35]	Principal components and ANNs	6.1	6.8

REFERENCES

1. Shriner, R.L., *The Systematic Identification of Organic Compounds*. 8th ed. / Ralph L. Shriner ... [et al.] ed. 2004, Hoboken, NJ ; [Great Britain]: Wiley. ix, 723 p.

2. Fisher, C.H., *Boiling Point Gives Critical Temperature*. Chemical Engineering, 1989. **96**: p. 157.

3. Lyman, W.J., W.F. Reehl, and D.H. Rosenblatt, *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*. 1982, New York: McGraw-Hill.

4. White, C.M., *Prediction of the Boiling Point, Heat of Vaporization, and Vapor Pressure at Various Temperatures for Polycyclic Aromatic Hydrocarbons.* Journal of Chemical & Engineering Data, 1986. **31**(2): p. 198-203.

5. Satyanarayana, K. and M.C. Kakati, *Note: Correlation of Flash Points.* Fire and Materials, 1991. **15**(2): p. 97-100.

6. Katritzky, A.R., et al., *Correlation of Boiling Points with Molecular Structure. 1. A Training Set of 298 Diverse Organics and a Test Set of 9 Simple Inorganics.* The Journal of Physical Chemistry, 1996. **100**(24): p. 10400-10407.

7. Walker, J., *Xxii.-the Boiling Points of Homologous Compounds. Part I. Simple and Mixed Ethers.* Journal of the Chemical Society, Transactions, 1894. **65**: p. 193-202.

8. Horvath, A.L., *Molecular Design : Chemical Structure Generation from the Properties of Pure Organic Compounds*. 1992, Amsterdam, The Netherlands ; New York: Elsevier.

9. Nendza, M., *Structure-Activity Relationships in Environmental Sciences*. 1998: Chapman & Hall.

10. Lyman, W.J., W.F. Reehl, and D.H. Rosenblatt, *Handbook of Chemical Property Estimation Methods: Environmental Behavior of Organic Compounds*. 1990, Washington, D.C. : American Chemical Society.

11. Joback, K.G. and R.C. Reid, *Estimation of Pure-Component Properties from Group-Contributions.* Chemical Engineering Communications, 1987. **57**(1-6): p. 233-243.

12. Constantinou, L. and R. Gani, *New Group Contribution Method for Estimating Properties of Pure Compounds.* AIChE Journal, 1994. **40**(10): p. 1697-1710.

13. Design Institute for Physical Properties Research (DIPPR), *Project 801*. 2010, American Institute of Chemical Engineers.

14. Stein, S.E. and R.L. Brown, *Estimation of Normal Boiling Points from Group Contributions.* Journal of Chemical Information and Computer Sciences, 1994. **34**(3): p. 581-587.

15. Sola, D., et al., *QSPR Prediction of N-Boiling Point and Critical Properties of Organic Compounds and Comparison with a Group-Contribution Method.* Fluid Phase Equilibria, 2008. **263**(1): p. 33-42.

16. Wiener, H., *Structural Determination of Paraffin Boiling Points.* Journal of the American Chemical Society, 1947. **69**(1): p. 17-20.

17. Randic, M., *Characterization of Molecular Branching.* Journal of the American Chemical Society, 1975. **97**(23): p. 6609-6615.

18. Kier, L.B. and L.H. Hall, *Molecular Connectivity in Chemistry and Drug Research*. 1976: Academic Press.

19. Needham, D.E., I.C. Wei, and P.G. Seybold, *Molecular Modeling of the Physical Properties of Alkanes.* Journal of the American Chemical Society, 1988. **110**(13): p. 4186-4194.

20. Stanton, D.T., et al., *Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles.* Journal of Chemical Information and Computer Sciences, 1992. **32**: p. 306-316.

21. Balaban, A.T., et al., *Correlations between Chemical Structure and Normal Boiling Points of Halogenated Alkanes C1-C4.* Journal of Chemical Information and Computer Sciences, 1992. **32**(3): p. 233-237.

22. Balaban, A.T., L.B. Kier, and N. Joshi, *Correlations between Chemical Structure and Normal Boiling Points of Acyclic Ethers, Peroxides, Acetals, and Their Sulfur Analogs.* Journal of Chemical Information and Computer Sciences, 1992. **32**(3): p. 237-244.

23. Wessel, M.D. and P.C. Jurs, *Prediction of Normal Boiling Points of Hydrocarbons from Molecular Structure.* Journal of Chemical Information and Computer Sciences, 1995. **35**(1): p. 68-76.

24. Wessel, M.D. and P.C. Jurs, *Prediction of Normal Boiling Points for a Diverse Set of Industrially Important Organic Compounds from Molecular Structure.* Journal of Chemical Information and Computer Sciences, 1995. **35**: p. 841-850.

25. Dearden, J.C., *Quantitative Structure-Property Relationships for Prediction of Boiling Point, Vapor Pressure, and Melting Point.* Environmental Toxicology and Chemistry, 2003. **22**(8): p. 1696-1709.

26. Katritzky, A.R., V.S. Lobanov, and M. Karelson, *QSPR: The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure.* Chemical Society Reviews, 1995. **24**(4): p. 279-287.

27. Hall, L.H., L.B. Kier, and B.B. Brown, *Molecular Similarity Based on Novel Atom-Type Electrotopological State Indices.* Journal of Chemical Information and Computer Sciences, 1995. **35**(6): p. 1074-1080.

28. Hall, L.H. and L.B. Kier, *The E-State as the Basis for Molecular Structure Space Definition and Structure Similarity*. Journal of Chemical Information and Computer Sciences, 2000. **40**(3): p. 784-791.

29. Katritzky, A.R., S. Sild, and M. Karelson, *Correlation and Prediction of the Refractive Indices of Polymers by QSPR.* Journal of Chemical Information and Computer Sciences, 1998. **38**(6): p. 1171-1176.

30. Chalk, A.J., B. Beck, and T. Clark, *A Quantum Mechanical/Neural Net Model for Boiling Points with Error Estimation.* Journal of Chemical Information and Computer Sciences, 2001. **41**(2): p. 457-462.

31. Katritzky, A.R., V.S. Lobanov, and M. Karelson, *Normal Boiling Points for Organic Compounds:□ Correlation and Prediction by a Quantitative Structure−Property Relationship.* Journal of Chemical Information and Computer Sciences, 1998. **38**(1): p. 28-41.

32. Weast, R.C., *Handbook of Chemistry and Physics: A Ready-Reference Book of Chemical and Physical Data*. 1976: CRC Press.

33. *Aldrich Catalog Handbook of Fine Chemicals*, Aldrich, Editor. 1996: Milwaukee, WI.

34. *Dragon Professional 6*. 2010, Talete SRL.

35. Ghavami, R., A. Najafi, and B. Hemmateenejad, *QSPR Studies on Normal Boiling Points and Molar Refractivities of Organic Compounds by Correlation-Ranking-Based PCR and PC–ANN Analyses of New Topological Indices.* Canadian Journal of Chemistry, 2009. **87**(11): p. 1593-1604.

36. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

37. Ghose, A.K. and G.M. Crippen, *Atomic Physicochemical Parameters for Three-Dimensional-Structure-Directed Quantitative Structure-Activity Relationships. 2. Modeling Dispersive and Hydrophobic Interactions.* Journal of Chemical Information and Computer Sciences, 1987. **27**(1): p. 21-35.

38. Egolf, L.M. and P.C. Jurs, *Prediction of Boiling Points of Organic Heterocyclic Compounds Using Regression and Neural Network Techniques.* Journal of Chemical Information and Computer Sciences, 1993. **33**(4): p. 616-625.

39. Paul, L., *A Widely Applicable Set of Descriptors.* Journal of Molecular Graphics and Modelling, 2000. **18**(4-5): p. 464-477.

40. Randić, M. and P.C. Jurs, *On a Fragment Approach to Structure-Activity Correlations.* Quantitative Structure-Activity Relationships, 1989. **8**(1): p. 39-48.

41. Magnuson, V.R., D.K. Harriss, and S.C. Basak, *Topological Indices Based on Neighborhood Symmetry: Chemical and Biological Applications*, in *Studies in Physical and Theoretical Chemistry*, R.B. King, Editor. 1983, Elsevier: Amsterdam, The Netherlands. p. 178-191.

42. Geary, R.C., *The Contiguity Ratio and Statistical Mapping.* The Incorporated Statistician, 1954. **5**(3): p. 115-146.

CHAPTER 7

GENERALIZING THE UNIVERSAL QUASI-CHEMICAL (UNIQUAC) MODEL PARAMETERS USING A NON-LINEAR QSPR MODEL

7.1. Introduction

A thorough understanding of chemical phase behavior properties is essential for designing and optimizing chemical and separation processes. Phase equilibria properties such as compositions and partition coefficients are typically measured in laboratory experiments, which require a substantial investment of money and time. The alternative is to predict phase equilibria properties using generalized thermodynamic models.

Vapor-liquid phase equilibria properties are typically determined within one of two computational frameworks. The first is the (ϕ/ϕ) approach, where fugacity coefficients (ϕ) for the vapor and liquid phases are calculated using equation-of-state (EOS) models. The second framework involves the split approach (ϕ/γ), where different models are used to predict the deviation functions, ϕ and γ of each component in each phase. Fugacity coefficients and activity coefficients (γ) are used as non-ideal behavior correction factors to the component ideal fugacities in the vapor phase and liquid phase, respectively. Fugacity coefficients are determined using various EOS models, and activity coefficients are calculated using excess Gibbs energy ($\overline{\mathrm{G}^{\mathrm{E}}}$) models; however, both EOS models and $\overline{\mathrm{G}^{\mathrm{E}}}$ models have limited capabilities for *a priori* predictions.

7.2. State of the Art in Activity Coefficient Modeling

A number of activity coefficient models for predicting vapor-liquid equilibria (VLE) have been proposed by various researchers [1-5], and these models can be classified as follows: (a) empirical and theory-based activity coefficient models such as Margules, Redlich-Kister and van Laar, regular solution, Wilson, non-random two liquid (NRTL) model, and the universal quasi-chemical (UNIQUAC) model [3]; and (b) predictive group-contribution models, such as universal functional activity coefficient (UNIFAC) and analytical solution of groups (ASOG) [2, 6]. Wilson first proposed an equation for excess Gibbs energy ($\overline{G^E}$) using the “local composition” concept that is based on the hypothesis that the local concentration around a molecule is different from the bulk concentration. Although the Wilson model performed better than other empirical models, the equation cannot be used to predict liquid-liquid equilibria (LLE) properties. Renon and Prausnitz [1] proposed the NRTL model based on Wilson’s local composition concept [7] and Scott’s two-liquid solution theory [8]. The NRTL model has three adjustable parameters that can be generalized to multicomponent mixtures using only the binary mixture parameters. One of the model parameters can be set *a priori*, which creates effectively a two parameter model.

Another popular activity coefficient model is the universal quasi-chemical (UNIQUAC) model. Abrams and Prausnitz [3] derived the UNIQUAC equation for nonrandom mixtures containing molecules of different sizes [9]. 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 can be determined using composition, size and shape of the components. The residual portion requires two adjustable binary parameters to account for intermolecular forces. The UNIQUAC model is applicable to a wide range of liquid mixtures that contain polar and nonpolar fluids. Although the UNIQUAC model requires only two adjustable parameters, this equation is more complex than the NRTL model. In addition, the UNIQUAC model is not always as precise for some systems where more than two adjustable binary parameters are needed [9]. Similar to the NRTL model, attempts to generalize the UNIQUAC model parameters have been limited [10].

Many of the activity coefficient models in literature are not generalized models and, as such, they cannot be applied for *a priori* prediction of VLE behaviors. Until recently, the preferred approach for *a priori* predictions of activity coefficients was the use of group-contribution models. These models are based on functional group interactions, such as UNIFAC and ASOG [2, 6]. 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]. The ASOG model estimates activity coefficients by summing the effects of molecular weight and functional group interactions. In the UNIFAC model, activity coefficients are determined based on the UNIQUAC model formulation, where combinatorial and residual effects are summed to determine the activity coefficients. The combinatorial portion from the UNIQUAC model is used directly, whereas the residual portion is calculated by considering the interaction of the functional groups present in the molecules. While successful for many systems, the

UNIFAC model suffers from some limitations, including an inability to account for the effects of neighboring molecules [11]. Further, the models are applicable only 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. A detailed review of other available generalized activity coefficient models can be found elsewhere [12, 13].

Recently, we sought an alternative methodology for providing generalized activity coefficient models that is more effective than group contributions [12]. Specifically, we used the quantitative structure-property relationship (QSPR) modeling approach to generalize the model parameters of the NRTL and UNIQUAC activity coefficient models and provide *a priori* VLE property predictions. The current research is an improvement on the previously reported work. In the earlier study, 332 binary systems that are encountered commonly in refinery processes were used to develop two independent QSPR models to predict the two adjustable parameters in the NRTL equation [11]. However, having two separate models could result in different model parameter values for a specific binary system, depending on the order of components involved. To make the model internally consistent, a single QSPR model for both parameters is required. Moreover, employing a more representative VLE database that goes beyond the needs of refinery processes would produce a more applicable generalized model. Therefore, the objectives of the current work are twofold: (a) to expand the existing database to include compounds comprised of a wider range of functional groups, and (b) to develop a single QPSR model for the two UNIQUAC model parameters.

Two case studies were conducted to investigate the predictive capabilities of the proposed QSPR-UNIQUAC activity coefficient model using (a) binary systems from the previous database [13] where systems in refining processes were the focus, and (b) compounds that are formed in the refining process of pyrolysis oil using bi-phasic reaction processes. The latter was of particular interest because of the growing interest in bi-phasic reaction processes to upgrade pyrolysis oil as well as the diversity of the molecular species encountered in these processes.

7.3. UNIQUAC Activity Coefficient Model Theory

The basis of UNIQUAC is that the excess Gibbs energy is a sum of the combinatorial and residual terms:

$$g^{E} = g^{E}_{comb} + g^{E}_{resid} \tag{7.1}$$

This can be extended to multi-component systems; however, for illustrative purposes the pertinent equations are given for a binary system:

$$\frac{g^{E}_{comb}}{RT} = x_1 \ln\left(\frac{\phi_1}{x_1}\right) + x_2 \ln\left(\frac{\phi_2}{x_2}\right) + \frac{Z}{2}\left(q_1 x_1 \ln\frac{\theta_1}{\phi_1} + q_2 x_2 \ln\frac{\theta_2}{\phi_2}\right) \tag{7.2}$$

and

$$\frac{g^{E}_{resid}}{RT} = -q_1 x_1 \ln\left[\theta_1 + \theta_2\tau_{21}\right] - q_2 x_2 \ln\left[\theta_2 + \theta_1\tau_{12}\right] \tag{7.3}$$

where, g^{E} is the excess Gibbs energy, $g^{E}{}_{comb}$ and $g^{E}{}_{resid}$ are the combinatorial and residual terms of the excess Gibbs energy, respectively, τ is an empirical binary interaction parameter that is experimentally determined, R is the universal gas constant, T is the mixture temperature, x denotes the mole fraction of a component, ϕ is the component area fraction, θ denotes the component volume fraction, and q and r denote the van der Waals surface area and volume of a component, respectively. A single numerical subscript indicates that the property is calculated for either component 1 or 2 of the binary mixture, while τ_{12} and τ_{21} are the interaction parameters between the two molecules, which are determined using Equation 7.5 [3] below.

The area fraction (ϕ_1) and volume fraction (θ_1) in Equation 7.3 are defined as:

$$\phi_1 = \frac{x_1 r_1}{x_1 r_1 + x_2 r_2} \qquad \theta_1 = \frac{x_1 q_1}{x_1 q_1 + x_2 q_2} \tag{7.4}$$

where, x, q, and r, are as defined previously.

The values of the van der Waals surface area and volume are obtained from the Bondi group contribution method [13].

$$\tau_{12} = \exp-\left(\frac{u_{12}-u_{22}}{RT}\right) = \exp-\left(\frac{a_{12}}{T}\right) \qquad \tau_{21} = \exp-\left(\frac{u_{21}-u_{11}}{RT}\right) = \exp-\left(\frac{a_{21}}{T}\right) \tag{7.5}$$

The parameters a_{12} and a_{21} are regressed from the available experimental data, and if experimental data do not exist, modeling of phase equilibria for those systems using UNIQUAC is not possible. Hence, the need exists for developing reliable predictive models to estimate UNIQUAC parameters.

7.4. QSPR Methodology

The development of a QSPR model for activity coefficients in the current work involves the following series of steps: (a) data set generation, (b) data regression to evaluate the best set of coefficients, (c) molecular descriptor calculation, (d) descriptor reduction and model training on the best coefficients, and (e) model validation. These different elements are described in greater detail below.

7.4.1. Database Development: The predictive capability of a QSPR model depends strongly 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 experimental VLE data points in each system were distributed evenly over the entire concentration range of 0 (pure component 2) to 1 (pure component 1). The general database and the two specialized databases are described in greater detail below.

General Database (Binary Systems): A low-pressure binary VLE database consisting of 186 systems totaling 4,716 data points was extracted from the Oklahoma State University (OSU) database. The 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. In total, the database compiled in this work consists of a total of 578 binary systems formed from various combinations of 145 different compounds. As such, a total of over 16,500 vapor-liquid equilibrium data points were assembled in the final database (OSU database II).

The compounds present in the OSU database II 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 7.1 illustrates the data distribution of the binary systems in the OSU database II based on chemical class.

Refining Systems Database: This sub-set database, which was adopted from the previous study by Ravindranath et al. [13], 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 9,700 VLE data at different temperatures were assembled in this database, and a detailed database assessment can be found in the previously published article [11].

Bi-phasic Database (Compounds Formed in Bi-phasic Reactions): Bi-phasic catalytic reaction is a promising technique that can be applied to the pyrolysis oil refining process. This methodology employs nanoparticle catalysts to selectively catalyze the target reactions in the oil and aqueous phases either individually or simultaneously [14]. Pyrolysis oil is an amalgam of different organic compounds such as acids, esters, alcohols, aldehydes, oxygenates, sugars, furans, phenols, guaiacols and syringols [15]. To be used as a transportation biofuel, pyrolysis oil needs to be upgraded, which includes increasing the caloric value of the refining process product by reducing the oxygen content and improving storage stability by reducing the levels of reactive compounds such as aldehydes [16]. To characterize these target reactions, knowledge of the phase behavior or the activity coefficients of the compounds in the pyrolysis oil is important.

The bi-phasic database consists of eight compounds that are formed in bi-phasic catalytic reactions. These compounds are comprised of 6 of the 31 chemical classes that are represented in the OSU database II. These chemical classes include alcohols, aldehydes, alkanes, furfural, ketones and water. The bi-phasic database is composed of 127 binary systems formed by different combinations of these compounds and approximately 2800 data points. In Figure 7.1, 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.

7.4.2. Model Parameter Regressions: To determine the optimum values of the two adjustable parameters a_{12} and a_{21} in the UNIQUAC equation, a regression analysis using an equal-fugacity equilibrium framework with mass balance constraints was performed to estimate the interaction parameters in Equation 7.5.

The vapor-liquid phase equilibrium criteria of a multicomponent closed system at given temperature and pressure are:

$$\hat{f}_i^v = \hat{f}_i^l \qquad i = 1,\ldots,N \tag{7.6}$$

$$T^v = T^l$$

$$P^v = P^l$$

where, $\hat{f}_i$ is fugacity of component i in the mixture, T is the temperature, P is the pressure, and the superscripts, *v* and *l*, indicate the vapor and liquid phases, respectively. In the regression analysis, a split approach was employed to express the component fugacities:

$$\phi_i P y_i = \gamma_i f_i^0 \lambda_i x_i \tag{7.7}$$

where, x_i is the liquid mole fraction, y_i is the vapor mole fraction, ϕ_i is the vapor fugacity, γ_i is the liquid activity coefficient, f_i° is the liquid fugacity at saturation, and λ_i is the Poynting factor. In this study, the bubble-point iteration function was employed:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} K_i x_i = 1 \tag{7.8}$$

where, n is the number of components, *K* is the equilibrium constant for component i, *x* is the component mole fractions in the liquid phase and *y* is the component mole fractions in the liquid vapor phase.

The parameter regression analysis was performed using an objective function, *F*, which is expressed for a binary system as the sum of squares of the relative errors in pressure and the activity coefficients of the two components, as follows:

$$\min F = \sum_{i=1}^{n} \left(\frac{P^{Exp} - P^{Calc}}{P^{Exp}} \right)_i + \sum_{i=1}^{n} \left(\frac{\gamma_1^{Exp} - \gamma_1^{Calc}}{\gamma_1^{Exp}} \right)_i + \sum_{i=1}^{n} \left(\frac{\gamma_2^{Exp} - \gamma_2^{Calc}}{\gamma_2^{Exp}} \right)_i \tag{7.9}$$

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 quality of the predictions is assessed using temperature and equilibrium constants of each binary system. The equilibrium constant, *K*, for component i is defined as the ratio of vapor to liquid mole fraction, or:

$$K_i = \frac{y_i}{x_i} = \frac{\gamma_i f_i^0}{\phi_i P} \tag{7.10}$$

7.4.3. Descriptor Calculation: The descriptors were calculated for each compound in the database using the method described in Section 2.5. 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 solute (component 1) and the second half of the descriptor set belongs to the solvent (component 2) in a binary system.

7.4.4. Descriptor Reduction and Model Development: See Section 2.6.

External Validation: In a recent article, Tropsha et al. [17] 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 data set would indicate the generalization capability of the final model. To create this external data set, three different approaches were implemented:

1. A self-organizing map (SOM) clustering technique as described in Section 2.6 was used to divide the data (1,156 parameters for 578 systems) into 4 different sets (training, validation, internal test, and external test sets). Using this approach, performing system-specific predictions is not possible because the parameters a_{12} and a_{21} of a specific system might lie in different data sets.
2. The entire data set was also 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 the four data sets have adequate representation from the 31 functional groups shown in Figure 7.1. The proportion of data used for the different data sets as follows: 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.
3. In the final 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 predictive capability of the model.

7.4.5. Case Studies: To meet the objectives of this work, four case studies were constructed to investigate QSPR model parameterization of the UNIQUAC parameters. In all these case studies, the ideal gas (IG) model was used to predict fugacity coefficients in the vapor phase, since all systems considered in this work are at low pressures. The four case studies are outlined as follows:

Case 1: *Ideal Solution model:* The ideal solution model was used to predict the phase-equilibria properties.

Case 2: *UNIQUAC model:* The UNIQUAC model was used to predict the activity coefficients. The UNIQUAC model parameters were regressed directly from the experimental data.

Case 2Q: UNIQUAC-QSPR model: The UNIQUAC model was used to predict the activity coefficients based on interaction parameters provided by the newly developed generalized QSPR parameter model.

Case 3U: *UNIFAC model:* The UNIFAC model [2] was used to predict the activity coefficients of each component. The UNIFAC interaction parameters reported by Gmehling et al. [2] were used in this case study. This case study allows a direct comparison between the current modeling approach and the best models reported in the literature.

Cases 1 was conducted to evaluate the correlative capabilities of the UNIQUAC model, whereas Cases 1, 2Q and 3U are focused on assessing the *a priori* predictive capabilities of the ideal solution, the generalized model (UNIQUAC /QSPR) and the UNIFAC model, respectively.

For the first case, the ideal solution model was used to predict *T*, *P*, K_1 and K_2 for the entire database of 578 binary systems. In Case 2, the two UNIQUAC model parameters, a_{12} and a_{21}, shown in Equation 7.11 were regressed and used directly to predict *T*, *P*, K_1 and K_2.

$$a_{12} = \frac{g_{12} - g_{22}}{R} \qquad a_{21} = \frac{g_{21} - g_{11}}{R} \tag{7.11}$$

Property predictions using the regressed UNIQUAC parameters resulted in the minimum error possible for the considered systems in Case 2. Therefore, the model parameters found in the regression analysis were used as target values in the development of the QSPR models. The property prediction errors using the regressed parameters were taken as a benchmark to judge the performance of the QSPR model.

Figure 7.2 shows the correlation between the two regressed UNIQUAC parameters in Case 2. The figure indicates that there is some level of correlation between the two parameters. The parameter correlation may hinder the accuracy of the QSPR models and the capability to provide reliable predictions from the structure of the components. A sequential parameter regression approach was applied in an effort to reduce the correlation of the model parameters. In this approach, one parameter was fixed at the generalized value while the other parameter was regressed. This procedure was performed multiple times until the effect of the correlation on the model development was minimized. A flowchart for the model development process employed in the current work is provided in Figure 7.3.

7.5. Results

Four VLE properties (*T, P, K_1* and *K_2*) were used to analyze the predictions or the representations (Case 2) of the various models used in Cases 1, 2, 2Q and 3U. The models used in each case were evaluated by comparing the property prediction errors, as described by root-mean-squared error (RMSE), bias and percentage absolute average deviation (AAD).

Table 7.1 presents the property prediction errors for the ideal solution (Case 1) and UNIQUAC (Case 2) models. The ideal solution model has overall AAD values of 12.4%, 1.3%, 13.2% and 21.6% for *T, P, K_1,* and *K_2* predictions, respectively. The UNIQUAC (Case 2) model with regressed parameters shows lower overall AAD values of 2.5%, 0.2%, 3.5% and 6.2% for *T, P, K_1* and *K_2* predictions, respectively. Case 2 establishes the best achievable level of prediction errors using the UNQUAC model. The model parameters (a_{12} and a_{21}) that were obtained by regression in Case 2 were then used as targets in the QSPR model development for Case 2Q. Our goal was to develop a QSPR model which would be capable of predicting *T, P, K_1* and *K_2* within twice the AAD value of the data regression in Case 2.

QSPR models were developed by applying the three data division approaches discussed in Section 7.4.5. 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 with each containing binaries with comparable functional groups. The QSPR model development process was initiated by dividing the 578 binary systems into four sets; 285 for training, 89 for validation, 65 for internal testing, and 139 for external testing. Regressed parameters from Case 2 were used as targets for developing the QSPR models. Models with 10, 20, 30 and 40 input descriptors were developed. The models with 30 and 40 descriptors had lower training RMSE values than the 10- and 20-descriptor models. For sake of simplicity, 30-descriptor models were further examined in the current work since the 40-desciptor models did not provide a statistically significant reduction in the training RMSE values when compared with the 30-descriptor models. The final model was chosen after nine iterations of sequential regression process, where the parameters a_{12} and a_{21} were regressed alternatively. The final ensemble model consisted of 20 different networks, each having the same descriptors as inputs but with different network architecture and weights.

Figures 7.4 and 7.5 compare the regressed UNIQUAC model parameters from Case 2, with the predicted model parameters from the UNIQUAC-QSPR (Case 2Q) model, for all data excluding the external test set. The plots indicate that the QSPR predictions are in good agreement with the regressed model parameters. Similarly, Figures 7.6 and 7.7 compare the regressed UNIQUAC model parameters from Case 2, with the predicted model parameters from the UNIQUAC-QSPR (Case 2Q) model, for the external test set. The generalization capability of the QSPR model was further analyzed by predicting *T, P*, K_1 and K_2 properties using the predicted model parameters. Table 7.2 shows the property prediction errors obtained using the UNIQUAC-QSPR predicted parameters (Case 2Q) for the training, validation, internal test and external test sets. The AAD values for the VLE predictions in all data sets were approximately twice the AAD values calculated in the UNIQUAC regression analysis (Case 2). The QSPR predicted parameters resulted in training set AAD values of 6.4%, 0.6%, 7.2% and 11.8% for *T, P,*

K_1, and K_2 property predictions, respectively. The validation and training set prediction errors were comparable, which demonstrates sufficient network training 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 8.6%, 0.7%, 8.2% and 14.2% for T, P, K_1, and K_2 predictions, respectively. The 29 elite descriptors (discussed in Section 2.6) that are used as inputs for the ANNs in the final ensemble model are listed in Table 7.3. Component numbers are used to denote whether the particular descriptor belongs to the first component or the second component in the binary system. Also, sample VLE plots for three systems are provided in Figures 7.8-7.10. For each of these systems, the experimental mole fractions in the liquid and vapor phases are compared with the model predictions from this study.

Further, the results from the Case 2Q predictions were compared to the predictions by the modified UNIFAC model [2] (Case 3U). The UNIFAC model could not be applied to 28 systems, due to the lack of parameters. Table 7.4 shows the overall prediction errors found using the generalized parameters (Case 2Q) and UNIFAC (Case 3U). Table 7.5 shows the property prediction errors for systems with compounds that are typically encountered in refining, and Table 7.6 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 parameters in Case 1 and generalized parameters in Case 2Q for eight chemicals.

7.6. Discussion

As expected, accounting for the non-ideal behavior through the UNIQUAC model (Case 2) resulted in significant error reductions (up to 4 times in the property predictions) when compared to the ideal solution model (Table 7.1). For the generalized UNIQUAC-QSPR (Case 2Q) model, the errors in property prediction for the external test set are about 1.1 - 1.3 times the corresponding errors in the training set, which is satisfactory (Table 7.2). Also to note, the UNIQUAC-QSPR model typically had higher errors for systems consisting of sulfide, chloro-alkane and amine functional groups. These higher prediction errors can be attributed to the lack of adequate representative structures in the training set.

A closer examination of Figures 7.4-7.7 suggests that the model leads to inaccurate predictions for parameter a_{21}, when compared to predictions for parameter a_{12}, as evident from the flat prediction curve for values close to zero in Figure 7.5. This could be due to the order of regression employed in this work, where parameter a_{12} was regressed initially, followed by the regression of parameter a_{21}. To prove this, another iteration of regression was performed on parameter a_{21}, while fixing the values of parameter a_{12} at the QSPR prediction values from the previous iteration. This led to better predictions for parameter a_{21}, but decreased the accuracy of the predictions for parameter a_{12} (no figure shown). The predictions for *T*, *P*, K_1, and K_2, however, did not significantly change during this additional iteration, which highlights the effects of parameter correlation, where each parameter (a_{12} or a_{21}) can have a range of optimum values of the other parameter with similar prediction results. To illustrate this assertion, five different binary systems from the flat prediction region in Figure 7.5 were selected, and a sensitivity analysis was performed by varying the a_{12} value systematically, and optimizing for the a_{21} parameters, while simultaneously recording the AAD values for pressure predictions. For the systems studied, a wide range of parameter values was identified that led to only a

25% increase in the AAD value on the pressure predictions. Figure 7.11 illustrates this for one binary system, where a_{12} values in the range -250 to 550 and a_{21} values in the range -320 to 380 lead to statistically similar AAD values in the pressure predictions. This suggests that for some systems, optimizing just one parameter, either a_{12} or a_{21} is sufficient to result in good predictions, as long as the other parameter is within a certain range. This also explains the poor agreement between the regressed and predicted a_{21} values for the systems in the external test set (Figure 7.7)

Table 7.3 lists the 29 elite descriptors that are used as inputs to the ANNs in the final ensemble model. Due to the nature of the ANNs, a quantitative assessment of the significance of these descriptors is not possible. However, the number of descriptors associated with the solute and the solvent molecules in Table 7.3 are almost the same. Of the 29 best input descriptors, four are molecular representation of structures based on electron diffraction (3D-MoRSE) descriptors [18]. These descriptors are used to describe the 3-dimensional (3D) structure of any molecule using a fixed number of variables. Also common are the GEometry, Topology, and Atom-Weights AssemblY (GETAWAY) descriptors, which according to Consonni and coworkers encode both the geometrical information given by the influence molecular matrix and the topological information provided by the molecular graph, weighted by the chemical information encoded in selected atomic weightings [19]. These descriptors contain information concerning the 3D structure of the molecule. In addition, the best descriptor list also has three binary fingerprint descriptors that describe the presence of carbon-carbon and oxygen-oxygen bonds at certain topological distances in the molecule.

Figure 7.12 shows the correlation of the regressed parameter values that are used as target values in the final QSPR model (9th iteration model). The plot reveals that the correlation between the two parameters was significantly reduced in the final regression analysis. This shows that the sequential regression technique was successful in reducing the correlation of the model parameters. The RMSE values between the regressed and the predicted parameter values from QSPR were 218 and 219 for the a_{12} and a_{21} model parameters respectively. After nine iterations of sequential regression analysis, the RMSE values decreased to 62 and 133 for a_{12} and a_{21} model parameters respectively. As expected, the reduction in the correlation of the regressed parameters was accompanied by reduction in the RMSE values between the regressed and the predicted parameter values from the QSPR models.

Table 7.4 shows the comparison between predictions from the generalized QSPR model with predictions from the modified UNIFAC model (Case 3U). As can be seen from Table 7.4, the overall results of the QSPR model are better compared to the UNIFAC group-contribution method. The AAD values of UNIFAC are 19.7%, 1.7%, 20.4% and 28.4% for T, P, K_1 and K_2 predictions, respectively. The current QSPR model resulted in approximately three times lower errors than that of UNIFAC predictions, which indicates that a QSPR modeling approach is effective in generalizing UNIQUAC model parameters for *a prior* property prediction. This could be attributed partially to the ability of the descriptors in the QSPR model to describe the 3D structures of the solute and the solvent; whereas, the UNIFAC model is based only on the 2D structural information and may be deficient in describing completely the solute-solvent interactions.

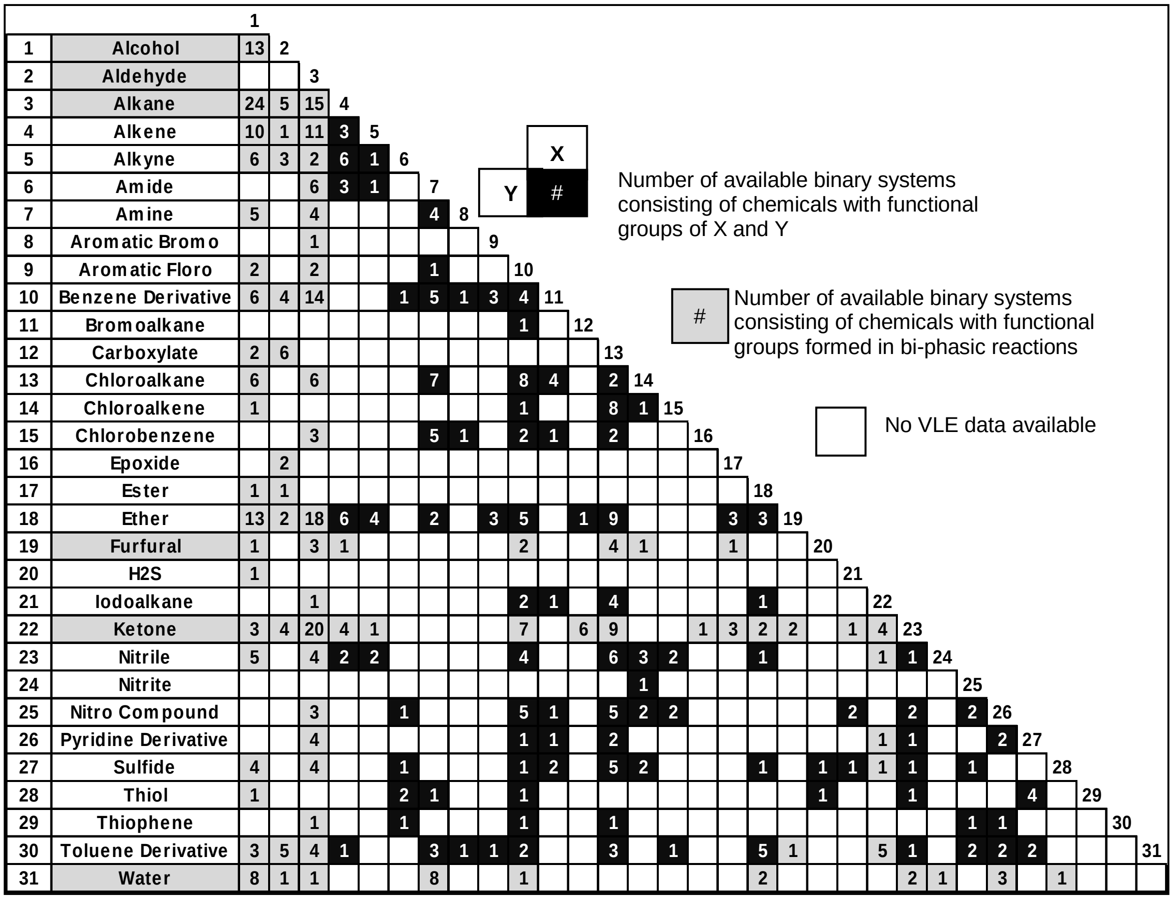
Table 7.5 shows the property prediction errors for systems that are commonly encountered in refining processes. The table provides VLE prediction errors using the regressed parameters in Case 2, and generalized parameters in Case 2Q, for the 332 binary systems. The property predictions using generalized parameters were approximately twice the regression results. Comparable overall prediction errors were found from the previously reported results by Ravindranath et al. [11], who employed two different QSPR models to predict for the two UNIQUAC model parameters. Some of the descriptors used in our newly developed model were reported as significant descriptors in the previous work [11] as well. These include descriptors such as atomic charge for N, O, C atoms, electro negativity and C - C bond related descriptors.

Finally, Table 7.6 shows the property prediction errors for systems with compounds that are typically formed in bi-phasic reactions. 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 Case 2 and Case 2Q; however, systems consisting of water and furfural gave higher errors in both Case 2 and 2Q. This can be attributed to a higher degree of non-ideality of the components and/or the lower quality of the data for systems with similar compounds.

7.7. Conclusions

1. In the current work, a non-linear QSPR model was developed to generalize successfully the UNIQUAC model parameters using an extensive database of 578 binary systems. As compared to previous works, where two different QSPR models were employed to predict for the two UNIQUAC parameters, the current work successfully employed just one QSPR model.

2. This work demonstrated an effective approach for the reduction of the correlation of model parameters using a sequential regression technique.
3. The prediction AAD values on an external test set of 139 binary systems were 8.6%, 0.7%, 8.2% and 14.2% for *T, P,* K_1, and K_2 predictions, respectively. Our QSPR model resulted in *a priori* predictions with errors approximately twice the errors obtained regressing experimental data.
4. According to the current work, 3D descriptors of the species involved have a significant effect on the UNIQUAC parameter values. This could be the reason for the higher accuracy of the current QSPR model compared to the existing UNIFAC group-contribution method.
5. The generalized UNIQUAC model was used to predict the equilibrium properties for 127 binary systems comprised of compounds typically formed in bi-phasic catalytic reactions. The AAD values for these systems were calculated to be 9.2%, 0.8%, 8.1%, and 15.9% for *T, P,* K_1, and K_2 predictions, respectively. This case study illustrates that the QSPR-generalized UNIQUAC model can be employed to predict the activity coefficients for binary systems with reasonable accuracy, even when no experimental data are available.



		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			

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 available

Figure 7.1: Database matrix of the compounds in the OSU database II along with the 31 functional groups represented

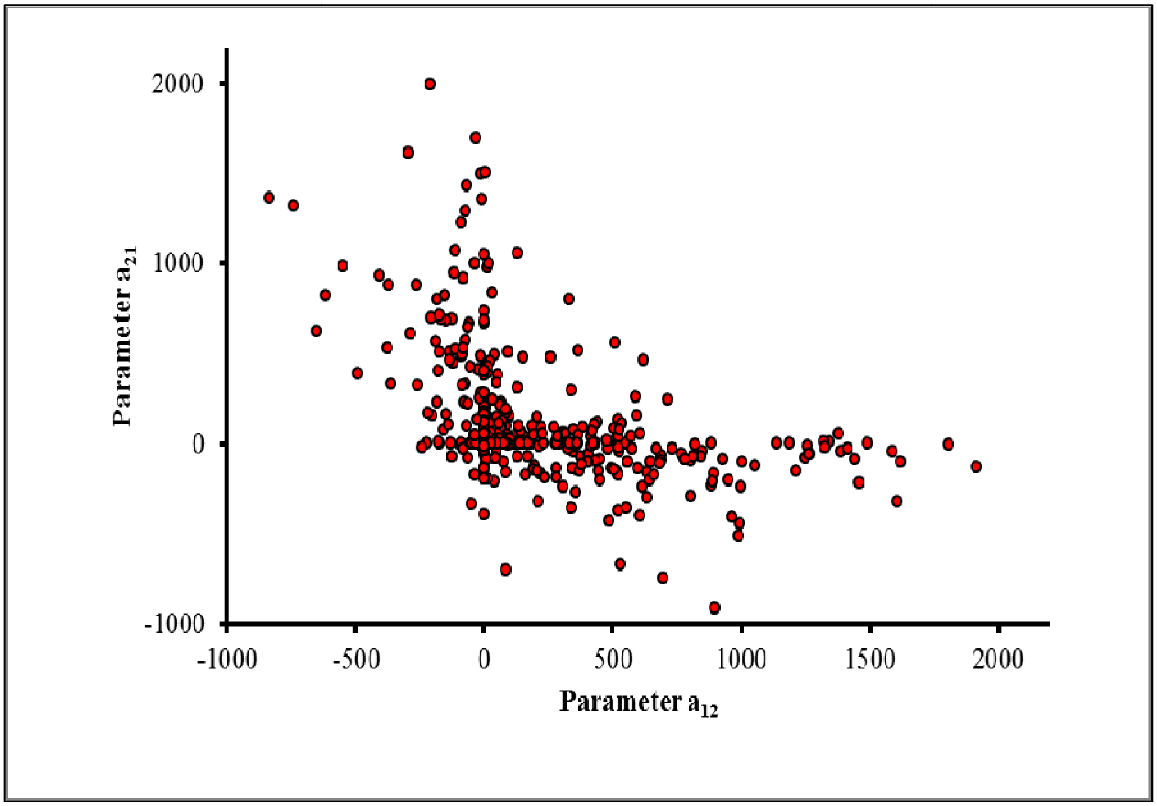


Figure 7.2: Correlation between the regressed UNIQUAC (Case 2) model parameters

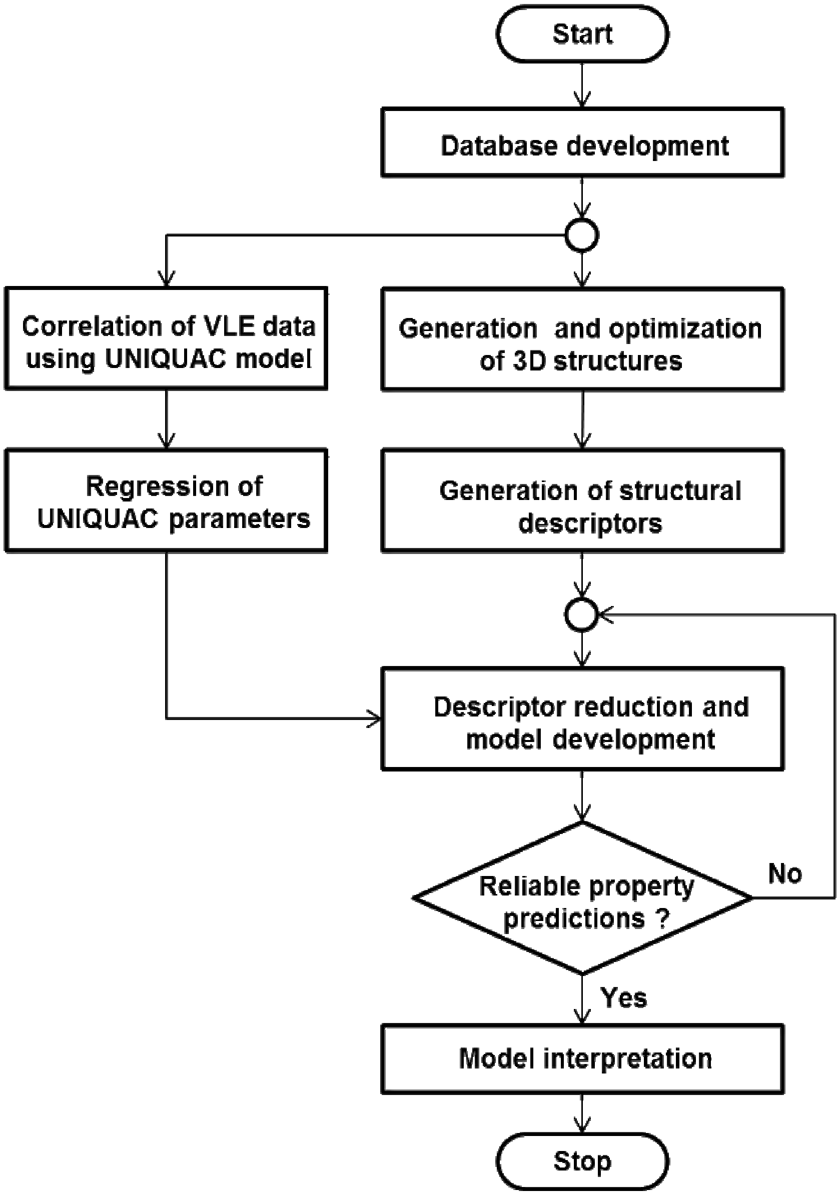

Start
Database development
Correlation of VLE data using UNIQUAC model
Generation and optimization of 3D structures
Regression of UNIQUAC parameters
Generation of structural descriptors
Descriptor reduction and model development
Reliable property predictions ?
No
Yes
Model interpretation
Stop

Figure 7.3: Schematic of the model development process employed in this work

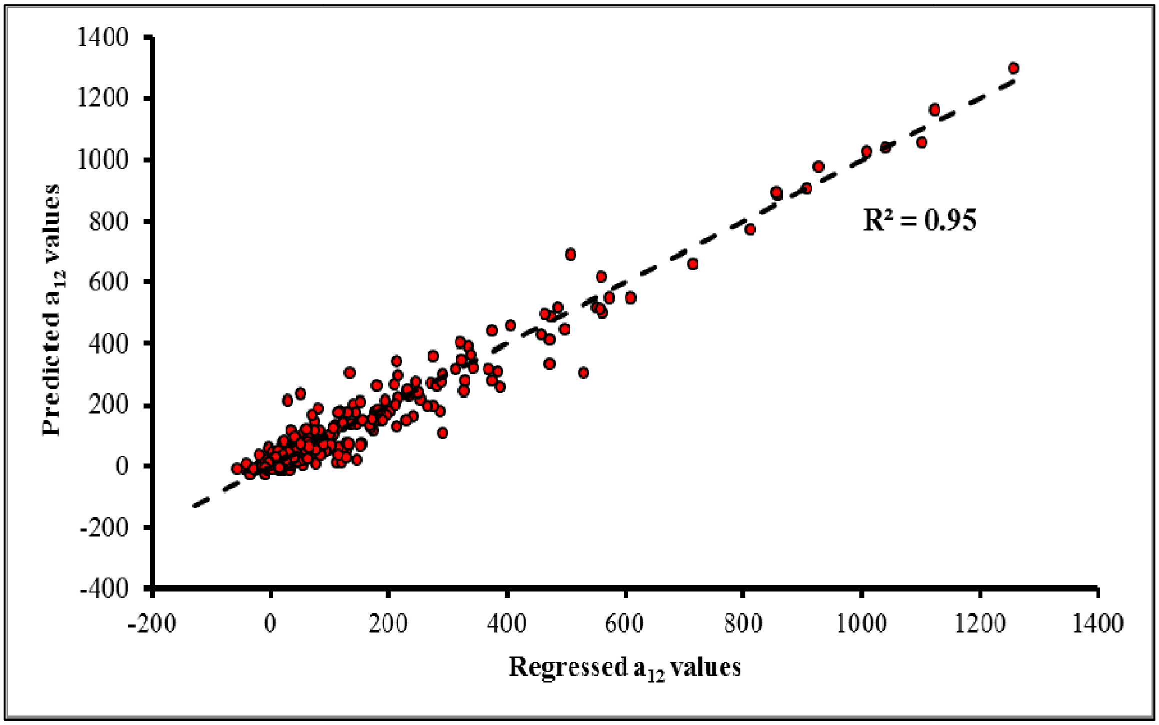


Figure 7.4: Comparison of the UNIQUAC-regressed (Case 2) and UNIQUAC-QSPR (Case 2Q) predicted a_{12} values for all data excluding the external test set. The broken line represents perfect predictions

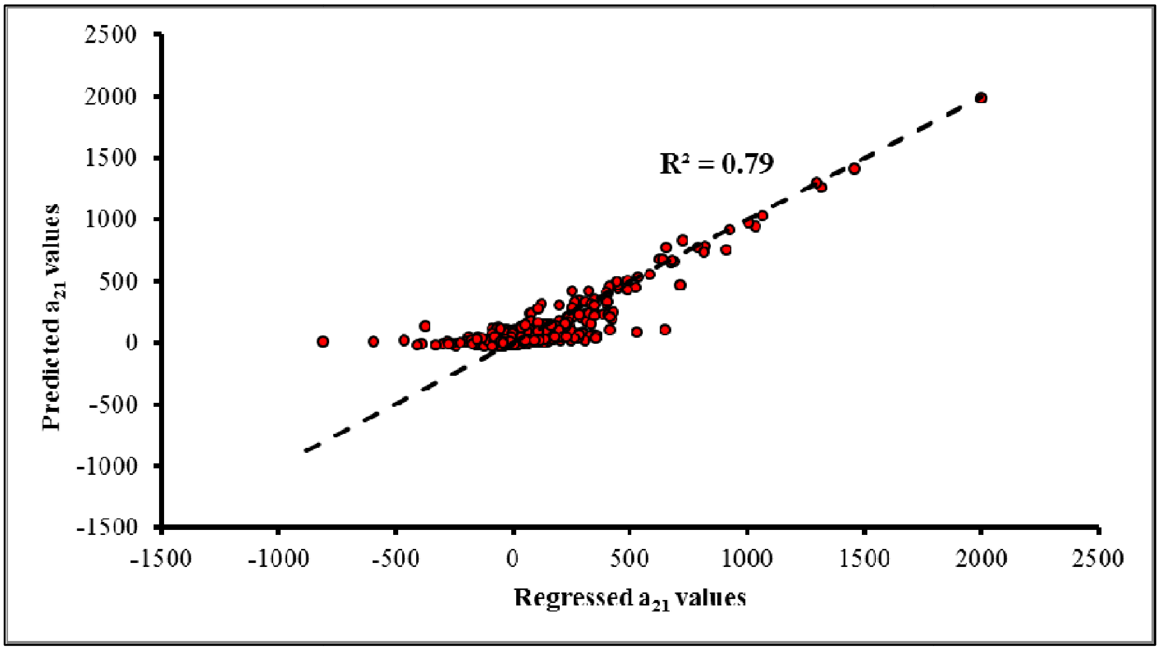


Figure 7.5: Comparison of the UNIQUAC-regressed (Case 2) and UNIQUAC-QSPR (Case 2Q) predicted a_{21} values for all data excluding the external test set. The broken line represents perfect predictions

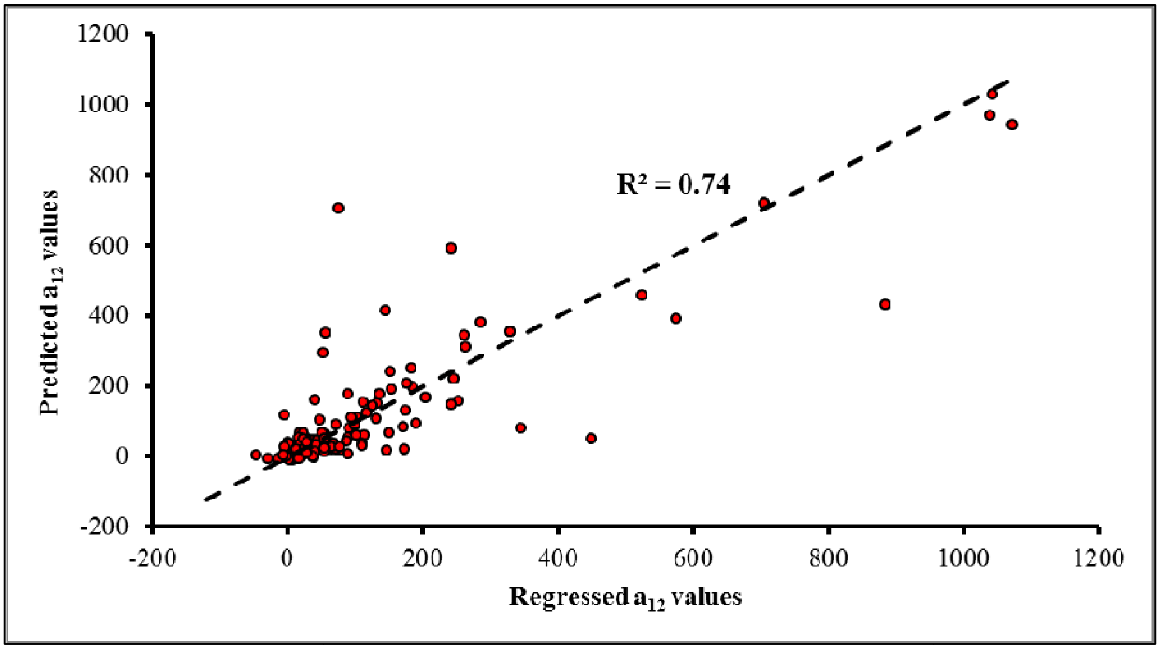


Figure 7.6: Comparison of the UNIQUAC-regressed (Case 2) and UNIQUAC-QSPR (Case 2Q) predicted a_{12} values for the external test set. The broken line represents perfect predictions

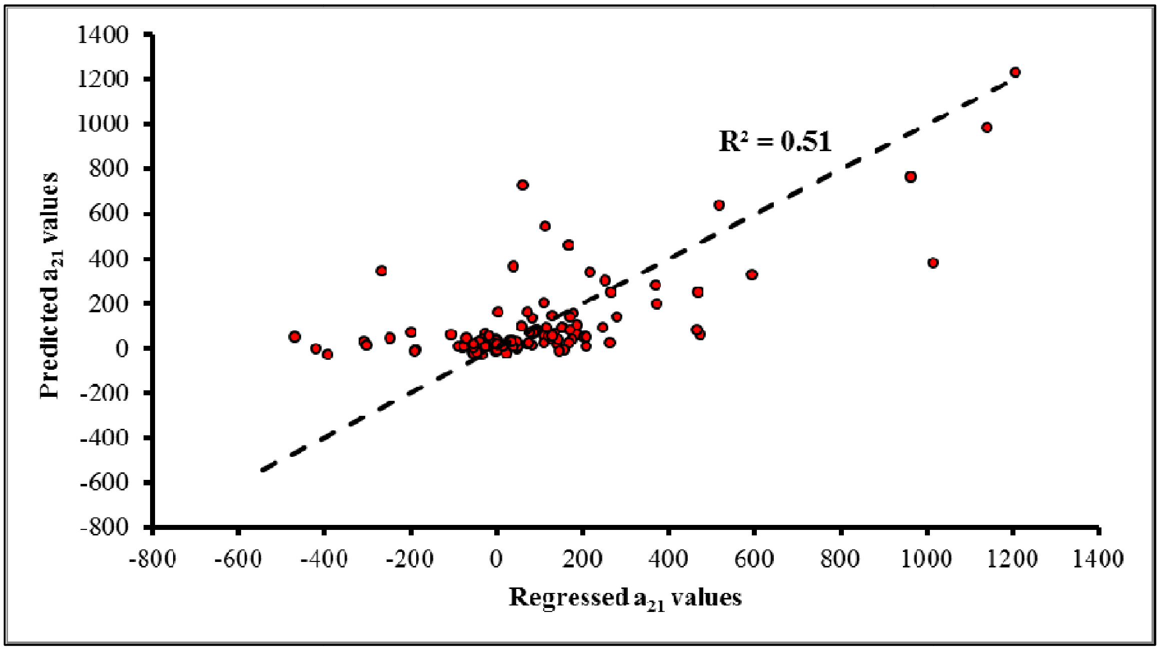


Figure 7.7: Comparison of the UNIQUAC-regressed (Case 2) and UNIQUAC-QSPR (Case 2Q) predicted a_{21} values for the external test set. The broken line represents perfect predictions

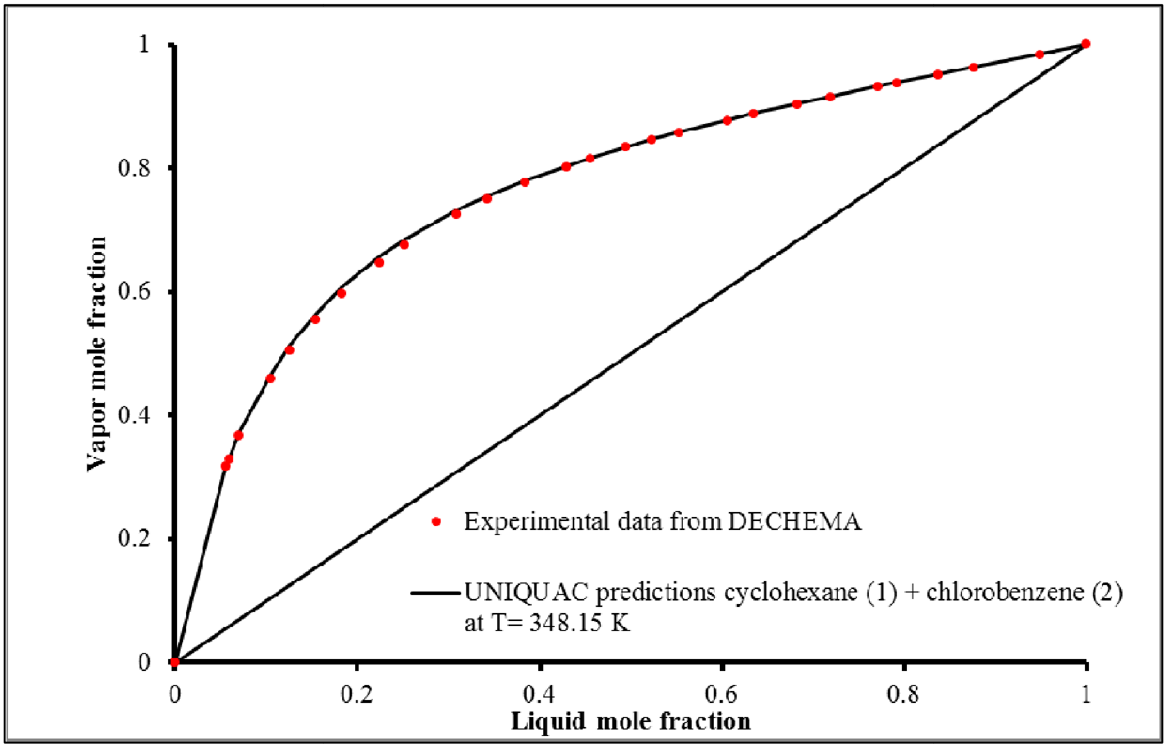


Figure 7.8: Equilibrium phase compositions for cyclohexane (1) + chlorobenzene (2) at T = 348.15 K

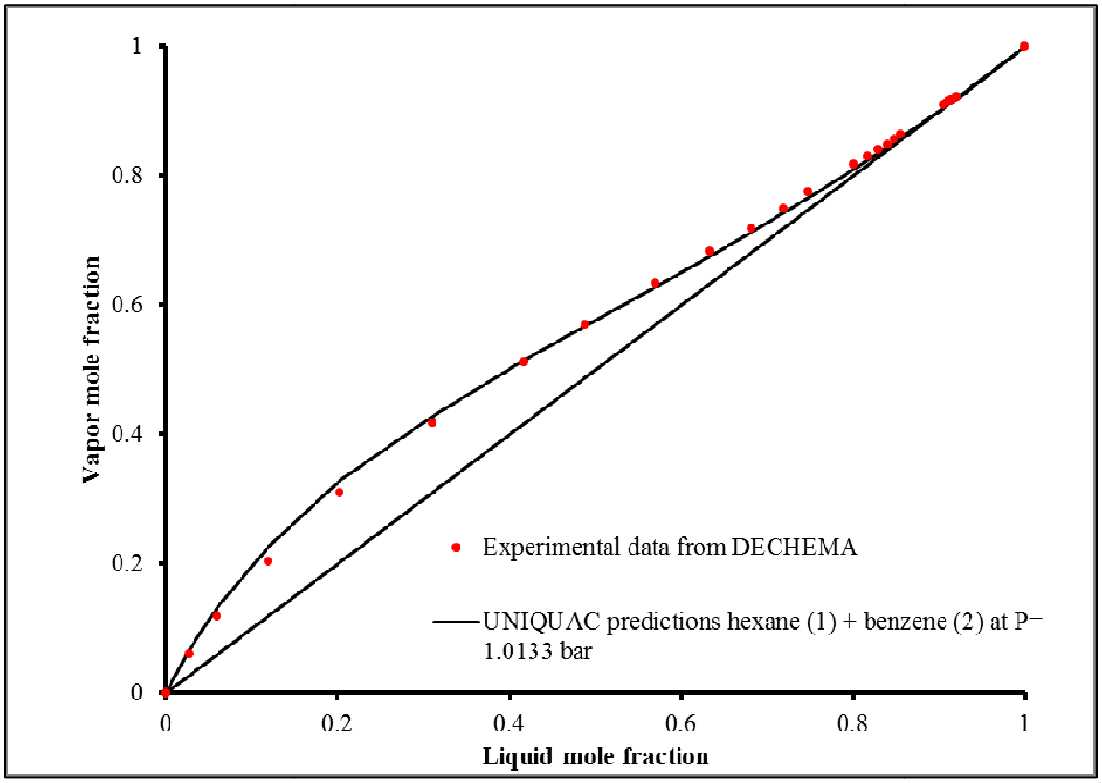


Figure 7.9: Equilibrium phase compositions for hexane (1) + benzene (2) at P = 1.0133 bar

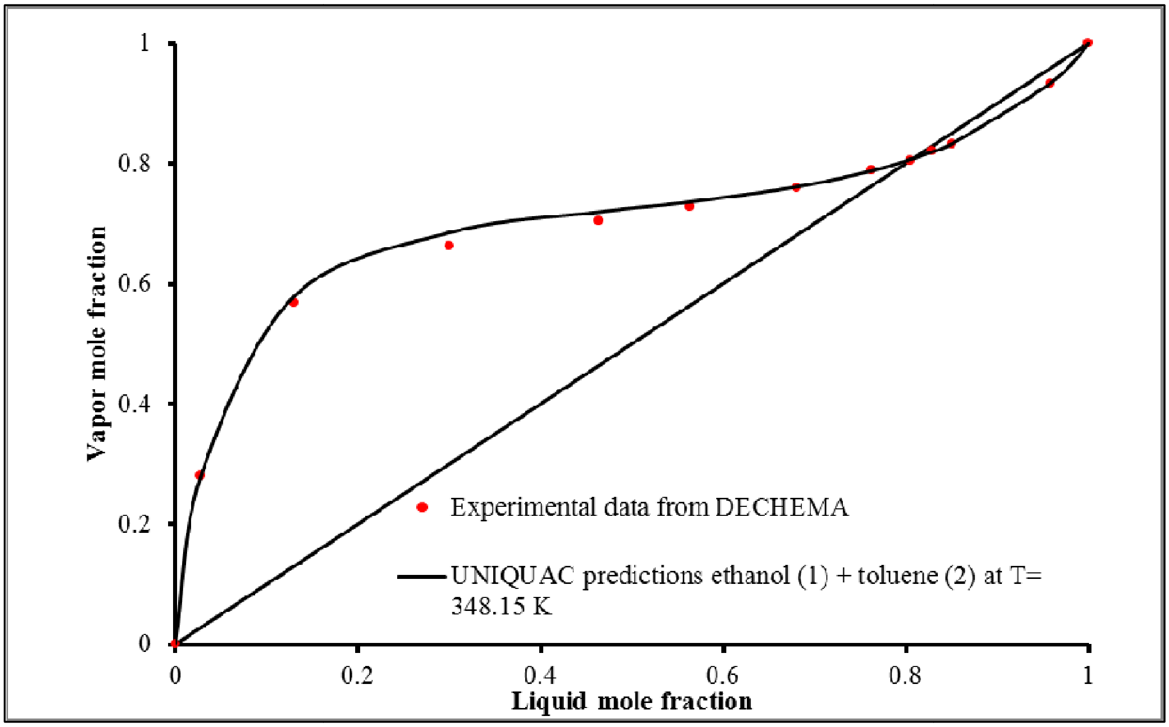


Figure 7.10: Equilibrium phase compositions for ethanol (1) + toluene (2) at T= 348.15 K

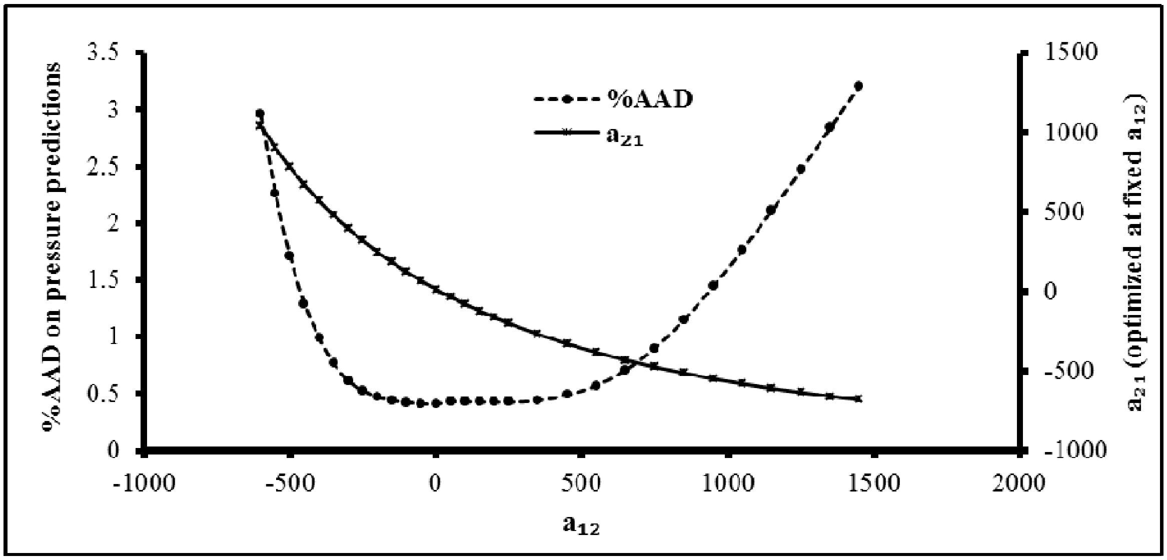


Figure 7.11: The effect of varying a_{12} and a_{21} on the quality of pressure predictions

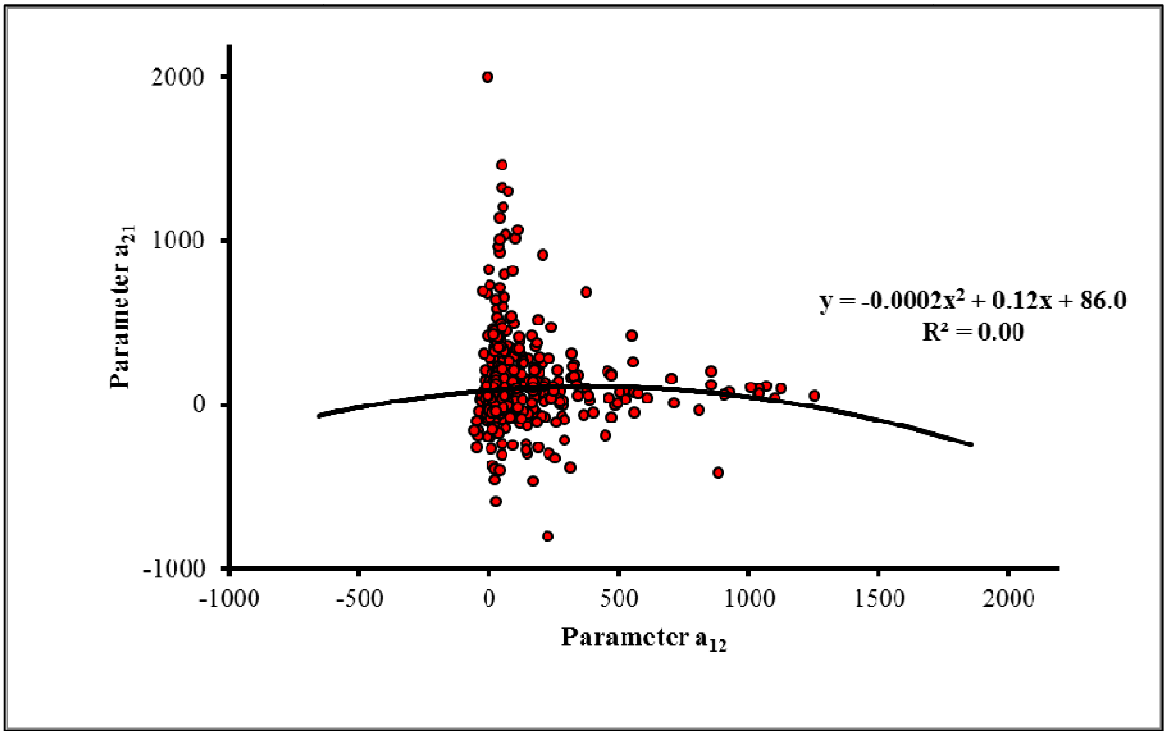


Figure 7.12: Correlation between the regressed UNIQUAC model parameters after nine iterations

Table 7.1: VLE predictions using ideal solution (Case 1) and UNIQUAC (Case 2) models

Case #	Model (V/L)	Parameters	Property	RMSE	Bias	%AAD
1	Ideal Solution	None	*P* (bar)	0.60	-0.10	12.4
			T (K)	8.60	3.80	1.3
			K_1	5.30	-0.70	13.2
			K_2	0.90	-0.20	21.6
2	IG/UNIQUAC	Regressed a_{12} & a_{21}	*P* (bar)	0.17	0.00	2.5
			T (K)	2.13	0.24	0.2
			K_1	3.51	-0.24	3.5
			K_2	0.25	-0.02	6.2

Table 7.2: VLE prediction errors for the UNIQUAC-QSPR (Case 2Q) model

Data Set	# of systems	Property	# of Pts.	RMSE	Bias	%AAD
Training Set	285	*P* (bar)	8451	0.29	0.01	6.4
		T (K)	8479	3.93	0.42	0.6
		K_1	5018	0.97	-0.05	7.2
		K_2	5016	0.53	-0.05	11.8
Validation Set	89	*P* (bar)	2977	0.12	-0.01	6.6
		T (K)	2995	3.98	0.47	0.6
		K_1	1866	0.51	-0.03	6.8
		K_2	1864	1.02	-0.07	10.4
Internal Test Set	65	*P* (bar)	1701	0.14	0.01	8.0
		T (K)	1701	3.76	0.02	0.6
		K_1	897	9.12	-1.25	6.5
		K_2	897	0.47	-0.02	14.3
External Test Set	139	*P* (bar)	3547	0.30	-0.03	8.6
		T (K)	3551	4.27	0.17	0.7
		K_1	2174	5.33	-0.47	8.2
		K_2	2174	0.43	0.00	14.2

Table 7.3: List of the descriptors used in the final ensemble for the UNIQUAC-QSPR (Case 2Q) model

No.	Descriptor	Complete Name of Descriptor	Component	Type of
1	RDF035v	Radial Distribution Function - 3.5 / weighted by atomic van der Waals volumes	1	RDF descriptors
2	Mor31u	3D-MoRSE - signal 31 / un-weighted	1	3D-MoRSE descriptors
3	Mor18v	3D-MoRSE - signal 18 / weighted by atomic van der Waals volumes	1	3D-MoRSE descriptors
4	G3u	3st component symmetry directional WHIM index / unweighted	1	WHIM descriptors
5	WA	Mean Wiener index	1	Topological descriptors
6	R6e	R autocorrelation of lag 6 / weighted by atomic Sanderson electronegativities	1	GETAWAY descriptors
7	F01[C-C]	Frequency of C - C at topological distance 01	1	2D frequency fingerprints
8	P1u	1st component shape directional WHIM index / unweighted	1	WHIM descriptors
9	Mor28v	3D-MoRSE - signal 28 / weighted by atomic van der Waals volumes	1	3D-MoRSE descriptors
10	TPSA(NO)	Topological polar surface area using N,O polar contributions	1	Molecular properties
11	EPS1	Edge connectivity index of order 1	1	Edge adjacency indices
12	nCIR	Number of circuits	1	Constitutional descriptors
13	DISPp	d COMMA2 value / weighted by atomic polarizabilities	1	Geometrical descriptors
14	MWC07	Molecular walk count of order 07	1	Walk and path counts

Table 7.3 (cont'd): List of the descriptors used in the final ensemble for the UNIQUAC-QSPR (Case 2Q) model

No.	Descriptor	Complete Name of Descriptor	Component No.	Type of Descriptor
15	F03[O-O]	Frequency of O - O at topological distance 03	2	2D frequency fingerprints
16	X1v	Valence connectivity index chi-1	2	Connectivity indices
17	nN	Number of Nitrogen atoms	2	Constitutional descriptors
18	B06[C-C]	Presence/absence of C - C at topological distance 06	2	2D binary fingerprints
19	ATS7p	Broto-Moreau autocorrelation of a topological structure - lag 7 / weighted by atomic	2	2D autocorrelations
20	GATS3v	Geary autocorrelation - lag 3 / weighted by atomic van der Waals volumes	2	2D autocorrelations
21	Mor24e	3D-MoRSE - signal 24 / weighted by atomic Sanderson electronegativities	2	3D-MoRSE descriptors
22	HATS0m	Leverage-weighted autocorrelation of lag 0 / weighted by atomic masses	2	GETAWAY descriptors
23	MPC04	Molecular path count of order 04	2	Walk and path counts
24	HATS3e	Leverage-weighted autocorrelation of lag 3 / weighted by atomic Sanderson electronegativities	2	GETAWAY descriptors
25	H-051	H attached to alpha-C	2	Atom-centered fragments
26	BLI	Kier benzene-likeliness index	2	Topological descriptors
27	nROH	Number of hydroxyl groups	2	Functional group counts
28	HATS2e	Leverage-weighted autocorrelation of lag 2 / weighted by atomic Sanderson electronegativities	2	GETAWAY descriptors

Table 7.3 (cont'd): List of the descriptors used in the final ensemble for the UNIQUAC-QSPR (Case 2Q) model

No.	Descriptor	Complete Name of Descriptor	Component No.	Type of Descriptor
29	Jhetv	Balaban-type index from van der Waals weighted distance matrix	2	Topological descriptors

Table 7.4: Cases 2Q and 3U - *a priori* VLE prediction comparison

Case #	Model (V/L)	# of systems	Property	RMSE	Bias	%AAD
2Q	IG/UNIQAUC	578	*P* (bar)	0.24	0.00	7.0
			T (K)	4.00	0.34	0.6
			K_1	1.53	-0.15	7.3
			K_2	0.36	-0.02	12.3
3U	IG/UNIFAC (Due to lack of interaction model parameters 28 systems from 578 systems were not considered)	550	*P* (bar)	4.70	0.20	19.7
			T (K)	12.20	-0.80	1.7
			K_1	4.90	0.10	20.4
			K_2	4.20	0.20	28.4

Table 7.5: Case 2 and 2Q – VLE property predictions for systems that are commonly encountered in refining processes

Case #	Model (V/L)	Parameters	# of systems	Property	RMSE	Bias	%AAD
2	IG/UNIQUAC	a_{12} Regressed a_{21} Regressed	332	*P* (bar)	0.26	0.01	2.4
				T (K)	1.92	0.20	0.2
				K_1	0.85	-0.01	3.4
				K_2	0.24	-0.01	6.1
2Q	IG/UNIQUAC	a_{12} QSPR a_{21} QSPR	332	*P* (bar)	0.31	-0.01	6.4
				T (K)	3.69	0.24	0.6
				K_1	0.50	-0.01	7.5
				K_2	0.30	-0.02	11.9

Table 7.6: Case 2Q – VLE property predictions for systems with compounds that are formed in bi-phasic reactions

Compound	# of sys	# of pts	%AAD							
			Case 2 (Regression)				Case 2Q (UNIQUAC/QSPR)			
			P **(bar)**	***T*** **(K)**	***K_1***	***K_2***	***P*** **(bar)**	***T*** **(K)**	***K_1***	***K_2***
n-octane	14	313	2.5	0.2	2.1	3.5	8.8	0.7	5.9	14.7
1-Propanol	16	315	3.5	0.3	3.6	9.2	6.8	0.5	8.4	13.8
2-propanol	5	105	1.6	0.1	3.7	3.8	5.8	0.4	8.8	8.2
Acetone	36	977	2.7	0.2	4.0	7.6	7.1	0.6	8.0	14.0
Benzaldehyde	3	70	3.5	0.3	5.1	10.6	5.9	0.4	2.4	15.6
Propionaldehyde	9	177	1.0	0.1	2.6	4.2	4.4	0.4	7.0	8.0
Furfural	16	262	5.4	0.6	3.1	14.1	15.2	2.1	4.9	28.0
Water	28	629	4.9	0.4	6.9	12.4	13.7	0.9	13.5	19.8
Total	**127**	**2848**	**3.4**	**0.3**	**4.0**	**8.6**	**9.2**	**0.8**	**8.1**	**15.9**

REFERENCES

1. Renon, H. and J.M. Prausnitz, *Estimation of Parameters for the NRTL Equation for Excess Gibbs Energies of Strongly Nonideal Liquid Mixtures.* Industrial & Engineering Chemistry Process Design and Development, 1969. **8**(3): p. 413-419.

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., et al., *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, 1995. **112**(1): p. 1-22.

6. Gmehling, J., D. Tiegs, and U. Knipp, *A Comparison of the Predictive Capability of Different Group Contribution Methods.* Fluid Phase Equilibria, 1990. **54**(0): p. 147-165.

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. Scott, R.L., *Corresponding States Treatment of Nonelectrolyte Solutions*. Vol. 25. 1956: AIP. 193-205.

9. Prausnitz, J.M., R.N. Lichtenthaler, and E.G. Azevedo, *Molecular Thermodynamics of Fluid-Phase Equilibria*. 1998: Pearson Education.

10. Ravindranath, D., et al., *QSPR Generalization of Activity Coefficient Models for Predicting Vapor–Liquid Equilibrium Behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

11. Prausnitz, J.M. and F.W. Tavares, *Anniversary Article Thermodynamics of Fluid-Phase Equilibria for Standard Chemical Engineering Operations.* AIChE Journal, 2004. **50**(4).

12. Ravindranath, D., *Structure-Based Generalized Models for Pure-Fluid Saturation Properties and Activity Coefficients*, School of Chemical Engineering. 2005, Oklahoma State University: Stillwater, Oklahoma.

13. Ravindranath, D., et al., *QSPR Generalization of Activity Coefficient Models for Predicting Vapor-Liquid Equilibrium Behavior.* Fluid Phase Equilibria, 2007. **257**(1): p. 53-62.

14. Resasco, D.E. and S. Crossley, *Molecular Engineering Approach in the Selection of Catalytic Strategies for Upgrading of Biofuels.* AIChE Journal, 2009. **55**(5): p. 1082-1089.

15. Huber, G.W., S. Iborra, and A. Corma, *Synthesis of Transportation Fuels from Biomass:□ Chemistry, Catalysts, and Engineering.* Chemical Reviews, 2006. **106**(9): p. 4044-4098.

16. Mahfud, F.H., F. Ghijsen, and H.J. Heeres, *Hydrogenation of Fast Pyrolyis Oil and Model Compounds in a Two-Phase Aqueous Organic System Using Homogeneous Ruthenium Catalysts.* Journal of Molecular Catalysis A: Chemical, 2007. **264**(1-2): p. 227-236.

17. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

18. Schuur, J.H., P. Selzer, and J. Gasteiger, *The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity.* Journal of Chemical Information and Computer Sciences, 1996. **36**(2): p. 334-344.

19. Consonni, V., R. Todeschini, and M. Pavan, *Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3D Molecular Descriptors.* Journal of Chemical Information and Computer Sciences, 2002. **42**(3): p. 682-692.

CHAPTER 8

A NON-LINEAR QSPR MODEL FOR THE INFINITE-DILUTION ACTIVITY COEFFICIENTS OF CYCLOHEXANE AND BENZENE IN VARIOUS SOLVENTS

8.1. Introduction

Solvents play an important role in many chemical reactions and separation processes. The design of solvents for specific applications requires an understanding of the interactions between the solute and the solvent at the molecular level. For ideal solutions, the mean strength of the interactions between all the molecules (including solute-solvent interactions) is the same, and therefore, the mixture properties can be described using just the concentrations of the individual species. For non-ideal solutions, the solute-solvent interactions are different from the solute-solute, or the solvent-solvent interactions. These interactions can be described quantitatively using dimensionless quantities called activity coefficients, which are denoted using the Greek symbol gamma (γ). To be specific, the activity coefficient, γ_i, describes the non-ideality for a species 'i' in a mixture. When the solute mole fraction, x_i, approaches zero, the activity coefficient is referred to as the infinite-dilution activity coefficient, and is denoted as γ_i^∞:

$$\gamma_i^\infty = \lim_{x_i \to 0}(\gamma_i) \tag{8.1}$$

Infinite-dilution activity coefficient values are of great importance because they describe only the solute-solvent molecular interactions, without the complication of the solute-solute interactions. The magnitude of the γ^{∞} value provides insight into the molecular forces that exist between the solute and the solvent molecules. From a practical viewpoint, γ^{∞} can be used to design separation equipment, to predict phase equilibria properties and to determine the fate of chemicals in the environment [1].

The experimental determination of γ^{∞} values is time-consuming and expensive. Further, these experimental techniques are difficult for sparingly soluble solutes, and experimental values typically do not exist for novel molecules that are designed *in silico* in the chemical and pharmaceutical industries. Therefore, a need exists for predictive models that can compute γ^{∞} values accurately based on molecular structures of the solute and the solvent molecules.

8.2. State of the Art in Predicting γ^{∞} Values

Since values of γ^{∞} can encompass a range of several orders of magnitude, logarithmic transformations such as log γ^{∞} or ln γ^{∞} are much easier to model, when compared to the original γ^{∞} values. In activity coefficient literature, the most common practice is to model the ln γ^{∞} values and therefore, the same practice is employed in this work. Several predictive models for ln γ^{∞} exist in the literature and are based on group-contribution methods (GCM). The universal functional activity coefficient (UNIFAC) approach by Fredenslund and coworkers [2] is one of the earliest predictive models for ln γ^{∞}, where for typical binary systems a deviation of 20% between experimental and predicted values is reported. Many modifications of the original UNIFAC method have been proposed such as Modified UNIFAC (Dortmund) [3, 4] and Modified UNIFAC (Lyngby) [5].

Voutsas and Tassios [6] compared various methods of calculating the ln γ^{∞} values and reported that the modified UNIFAC methods give better results than the original UNIFAC for athermal alkane/alkane asymmetric mixtures. Another GCM based model is the analytical solutions of groups (ASOG) model by Tochigi and coworkers [7]. Despite their popularity, GCM methods suffer from disadvantages, such as their inability to model structures containing undefined functional groups and to account for the interaction between different functional groups and their spatial arrangement. For polar systems in particular, the UNIFAC approach leads to significantly inaccurate predictions [8]. Alternate strategies such as the linear solvation energy relationship (LSER) [9] models have been found to be more accurate. The LSER method is reported to have an average absolute deviation of 0.294 units for ln γ^{∞} values of 336 organics in water [9]. Molecular simulation methods are the other major class of prediction models for ln γ^{∞}. These methods are usually based on potential energy functions derived from pure-fluid properties such as heats of vaporization, and therefore theoretically, are easier to implement due to the availability of the data, when compared to GCM approaches, which are based on binary mixture thermodynamic data. Lazaridis and Paulaitis [10] developed a free energy perturbation method with Monte Carlo simulations, for predicting ln γ^{∞} values for chlorinated organic compounds in water. However, the deviation between the experimental and predicted values was found to be unacceptable even for simple solutes. Moreover, the method was expensive computationally to employ, even for moderately sized solutes. The conductor-like screening model for real solvents (COSMO-RS) is a relatively new promising simulation methodology for calculating the γ^{∞} values. The COSMO-RS theory describes the interactions in a fluid in terms of local contact interactions of molecular surfaces [11]. Putnam et al. [12] have tested the COSMO-RS model and report that the model predicts reasonably for aqueous binary systems, but predicts poorly for some non-aqueous systems.

An attractive alternative to the previous modeling methods is the quantitative structure-property relationship (QSPR) approach, where the target molecular property is expressed in terms of the structural aspects of the molecule. Mackay and Shiu [13] developed one of the earliest QSPR correlations for ln γ^{∞}, where a correlation between ln γ^{∞} values of polynuclear aromatic hydrocarbons in water with the number of carbon atoms was discovered. This model, developed for a specific class of compounds in a single solvent, resulted in better predictions than the more general universal quasi chemical (UNIQUAC) and UNIFAC models. In a similar study, Medir and Giralt [14] developed a correlation between ln γ^{∞} values and molecular connectivity descriptors for aromatic and aliphatic hydrocarbons in water. While more limited in application, their model provided better predictions than both the UNIQUAC and the UNIFAC models. Neely et al. [15] had developed a neural network based QSPR model for predicting γ^{∞} values of hydrocarbon-water binary systems. Their model had an AAD value of 6% on the training data, but had poor predictive performance on extended temperatures. In a similar work, Mitchell and Jurs [16] developed a QSPR model for a large number of organics in water using the automated data analysis and pattern recognition toolkit (ADAPT). They reported a prediction set (in this case, data that have not been used for model development) error of 0.43 units for ln γ^{∞} values, which is better than corresponding predictions from UNIFAC models. Rani and Dutt [17] performed a similar study consisting of 351 training data (the data used to develop the model) to predict 92 ln γ^{∞}

values for one halocarbon in water, 17 organics in one hydrofluoroparaffin, and one organic in five hydrofluoroparaffins with a reported average absolute deviation (AAD) of 11.8% on the basis of ln γ^{∞} values. In a recent work, Giralt and coworkers [8] used Kohonen self-organizing maps (SOMs), along with fuzzy-ARTMAP neural classifiers to develop QSPR models that had an average absolute error of 0.52 (6.6%) natural log units for a prediction set of 45 organics in water. Schult [18] developed a modified UNIFAC model to predict the ln γ^{∞} values of 20 solutes in n-methyl-2-pyrrolidone and hexadecane. They report AAD values of 8% and 11% for solutes in n-methyl-2-pyrrolidone and hexadecane, respectively.

This brief review of the existing literature highlights the facts that the majority of the models deal with aqueous systems, and the generalized UNIQUAC and UNIFAC models cannot provide reliable predictions when quality experimental data do not exist for the specific solutes and solvents. Further, QSPR techniques have proven to be effective when dealing with a limited class of compounds. This provides the impetus for the current work, where specific QSPR models were built to predict the ln γ^{∞} values for benzene and cyclohexane, separately, in the presence of a varied class of solvent compounds. Specifically, this work focuses on the following objectives:

1. Develop an accurate non-linear QSPR model to predict the ln γ^{∞} values for benzene and cyclohexane using a database made up of diverse set of solvents.
2. Validate the current modeling approach by employing an external test set of compounds that has not been used to develop the model.
3. Compare the current modeling approach with existing approaches in the literature, on common training and external set data. This would further establish the efficacy

of the modeling approach used in the work. Specifically, ln γ^{∞} values for 325 organics in water were extracted from the literature and were used to develop QSPR models.

8.3. QSPR Methodology

The development of a QSPR model involves the following series of steps: (a) data set generation, (b) descriptor calculation, (c) descriptor reduction and model development, and (d) model validation. These elements are described below.

8.3.1. Data Set Generation: Experimental γ^{∞} values at 20°- 40°C were extracted from the DECHEMA chemistry data series [19, 20] for binary systems with cyclohexane and benzene as solutes. The γ^{∞} values have been assumed to be temperature-independent in this narrow temperature range, which is a reasonable assumption considering the experimental uncertainty associated with the data. To support the assumption, plots of ln γ^{∞} versus temperature are provided in Figures 8.1 and 8.2, for benzene and cyclohexane, respectively. These plots suggest that the uncertainties in experimental data are considerable for ln γ^{∞} values close to zero, and for the systems considered, the variation of ln γ^{∞} values with temperature is minimal within the 20°- 40°C range. In all, 175 and 192 unique solvent γ^{∞} values were extracted from the literature, for cyclohexane and benzene, respectively. Approximately 80% of this data was within the 25°- 30°C temperature range. The DECHEMA chemistry data series [19, 20] does not provide estimates for the experimental uncertainties of the data, and therefore the quality of the data used in the current work cannot be assessed.

Cyclohexane: The ln γ^{∞} values of the molecules in the final database for cyclohexane lie in the range of -0.65 to 5.7 natural log units (Figure 8.3 provides the distribution of ln γ^{∞}

data). The molecular weights of these compounds vary from 32.05 g/mol to 426.76 g/mol, and the octanol-water partition coefficient, log K_{ow}, (calculated using the Ghose-Crippen ALOGP model in DRAGON [21]) varies between -1.7 and 11.3. Further details on the database characterization are provided in Table 8.1.

Benzene: The ln γ^{∞} values of the molecules in the final database for benzene lie in the range of -0.76 to 3.6 natural log units (Figure 8.4 provides the distribution of ln γ^{∞} data). The molecular weights of these compounds also vary from 32.05 g/mol to 426.76 g/mol, and the log K_{ow} values (calculated using the Ghose-Crippen ALOGP model in DRAGON [21]) vary between -1.7 and 11.3. Further details on the database characterization are given in Table 8.2.

In addition to the above data sets, additional ln γ^{∞} data of 325 organics in water were extracted from literature [8, 16]. This data was originally compiled by Sherman et al. [9], and has been used to develop models to predict the ln γ^{∞} values of organics in water, by Sherman et al. [9], Mitchell and Jurs [16], and more recently by Giralt et al. [8]. Henceforth in this work, this data will be referred to as the aqueous database to differentiate it from the cyclohexane and benzene data sets. Giralt et al. [8] report that their model performs better than the models by Sherman et al. [9], and Mitchell and Jurs [16] on this aqueous data set. To validate the current modeling approach, the same aqueous data were used to develop a QSPR model to predict the ln γ^{∞} values and the resulting model was compared with the prediction results by Giralt et al. [8]. To ensure a fair comparison, the same training and external test data employed by Giralt et al. [8] were used in the current work.

8.3.2. Descriptor Calculation: See Section 2.5.

8.3.3. Descriptor Reduction and Model Development: See Section 2.6.

External Validation: In a recent article, Tropsha et al. [22] emphasized the need to validate QSPR models using external data sets. Therefore, another model was built by separating some benzene and cyclohexane data from the original database and allocating them to an external test set; however, the data cannot be randomly separated, as the external set might not be representative of the training set. Therefore, a SOM network was created using the best descriptors identified in the first ensemble, which was developed using the entire database. This SOM was used to identify clusters in the data and facilitate the partitioning of data into T, IV and IT sets as explained in Section 8.3.3. The number of map units in this SOM was varied until the percentage of data points in the IT set is 15% of the size of the entire final data set. This IT set was then set aside as an external test set and the remaining data was used for developing another model *de novo*, by repeating the search for the best descriptors, best network architecture and network weights. In the current work, 15% of the molecules were identified as an external test set using this procedure, and the remaining 85% data points were again divided into T, IV and IT sets and subjected to the descriptor search algorithm as discussed in Section 8.3.3. For clarity, the model created using all data points for model development will be referred to as Model 1 and the model developed using just 85% of the data points as Model 2. Model 1 will be used in the computer-aided molecular design (CAMD) algorithms because of its larger training set size, and Model 2 will be used to assess the generalization capability of Model 1, as advocated by Tropsha et al. [22].

8.4. Results

8.4.1. Model 1 for Cyclohexane: Ten-descriptor, 15-descriptor, and 20-descriptor-models were tested, but no significant difference was observed between the models. Therefore, for the sake of simplicity, 10 descriptors were used in the final models. Using less than 10 descriptors resulted in a significant increase in the training RMSE values for databases comprised of more than 150 data points, which provides additional support for the choice of ten input descriptors. Figure 8.5 shows the comparison between the experimental and predicted ln γ^{∞} values for Model 1. The correlation coefficient (R^2) between the experimental and predicted values is 0.94. The prediction residual errors in natural log units are plotted in Figure 8.6**,** which demonstrates clearly that the residuals are almost symmetrically distributed around the horizontal axis, as expected from an unbiased model. A histogram of the residuals (not shown) was plotted, and the distribution of the residuals around zero was found to be similar to a normal distribution. Additionally, the RMSE and the mean average error (MAE) values for the training data set predictions are 0.29 natural log units and 0.22 natural log units, respectively. The RMSE values for the individual ensembles range from 0.30 natural log units to 0.36 natural log units. The results from the overall ensemble are better than the results for the individual ensembles, which validates the use of ensembles with different descriptors as inputs.

The different descriptors used for creating the eight different ensembles are shown in Table 8.3. Note that the neural networks in the ensembles are allowed to have a maximum of 10 elite inputs, but frequently the individual networks will have a slightly lower number of elite descriptors as inputs after the insignificant descriptors have been removed, as described in Section 8.3.3. The descriptors R1s, ALOGP, Chi0_EA (dm), HyWi _B (m) and SpPosLog_Dz (Z) are the most common across the ensembles. The types and physical meanings of these commonly occurring descriptors, as extracted from the DRAGON [21] help file, are provided in Table 8.4.

8.4.2. Model 2 for Cyclohexane: For Model 2, 10-descriptor models were chosen. Figure 8.7 provides a comparison between the experimental and predicted ln γ^{∞} values for the external test data set of 28 compounds. The correlation coefficient (R^2) between the experimental and predicted external test data is 0.83. The prediction residual errors on this data are near-symmetrically distributed around the horizontal axis (no figure shown**)**. The RMSE and MAE values for the training set data of 147 compounds are 0.32 and 0.23 natural log units, respectively. The RMSE and MAE values for the external test set are calculated to be 0.48 and 0.39 natural log units, respectively.

8.4.3. Model 1 for Benzene: Similar to the models for cyclohexane, ten-descriptor, 15-descriptor, and 20-descriptor-models were tested, but no significant difference was observed between the models. Therefore, for the sake of simplicity, 10 descriptor models were used in the final models in the current study. Figure 8.8 shows a comparison between the experimental and predicted ln γ^{∞} values for Model 1 for benzene. The correlation coefficient (R^2) between the experimental and predicted values is 0.93. The prediction residual errors in natural log units are plotted in Figure 8.9**,** which demonstrates clearly that the residuals are almost symmetrically distributed around the horizontal axis, as expected from an unbiased model. A histogram of the residuals (not shown) was plotted, and the distribution of the residuals around zero was found to be similar to a normal distribution. Additionally, the RMSE and the MAE values for the training data set predictions are 0.19 and 0.14 natural log units, respectively. The RMSE values for the individual ensembles range from 0.19-0.24 natural log units.

The different descriptors used for creating the eight different ensembles are shown in Table 8.5. The descriptors MLOGP, SAdon, and Sp_Abs_B (e) are the most common across the ensembles. The types and physical meanings of these commonly occurring descriptors, as extracted from the DRAGON [21] help file, are provided in Table 8.6.

8.4.4. Model 2 for Benzene: For Model 2, 10 descriptor-models were chosen. A comparison between the experimental and predicted ln γ^{∞} values for the external test data of 30 compounds is provided in Figure 8.10. The RMSE and MAE values for the training set data of 162 compounds are 0.19 and 0.15 natural log units, respectively. The RMSE and MAE values for the external test set are calculated to be 0.45 and 0.29 natural log units, respectively. The correlation coefficient (R^2) between the experimental and predicted external test data is 0.66. The prediction residual errors on this data are near-symmetrically distributed around the horizontal axis (no figure shown).

8.4.5. Model for Aqueous Data: Giralt et al. [8] had employed 12 descriptors in their model, and so to ensure a fair comparison, 10 descriptor-models were developed in the current work. For the current model, the RMSE and MAE values are calculated to be 0.38 and 0.28 natural log units, respectively for the training set comprising 280 compounds. For the external test set comprising 45 compounds, the RMSE and MAE values are calculated to be 0.67 and 0.35 natural log units, respectively. A comparison between the experimental and predicted ln γ^{∞} values for the external data of 45 compounds is provided in Figure 8.11. The correlation coefficient (R^2) between the experimental and predicted external test data is calculated to be 0.96.

The errors for all models developed in the current work are tabulated in Table 8.7.

8.5. Discussion

The Model 2 MAE values for the cyclohexane and benzene external test sets are within two times the corresponding training set MAE values, which indicate normal predictive performance, based on comparison with other models in the literature that employ neural networks to model physico-chemical properties [23, 24]. Due to the larger training data set, Model 1 for both cyclohexane and benzene would be expected to perform similar to or better than Model 2 on unseen data (external data set). Therefore, the predictive performance of Model 2 on an external test set can be used as an approximation for determining the generalization capability of Model 1 for both cyclohexane and benzene.

The residual plots (Figures 8.6 and 8.9) of Model 1 for both cyclohexane and benzene exhibit over-prediction for lower values of ln γ^{∞} and under-prediction for higher values of ln γ^{∞}. Similar trends were observed in Model 2 for both cyclohexane and benzene. This could be explained, in part, by the lower numbers of molecules with extreme ln γ^{∞} values in the databases employed in the current work. Further, the compounds that exhibit the largest deviations in the various models were examined manually to identify any correlation between their higher errors and the molecular structure, as typified by the presence/absence of certain functional groups. This examination did not reveal any particular trends between the functional groups present in the molecule and the prediction error for the molecule. The higher errors for some molecules could be due to the higher experimental uncertainty in the data for those molecules.

The prediction results from Model 2 for cyclohexane and benzene are comparable to the existing QSPR models in the literature. In particular, the current model compares favorably to a recent model developed by Giralt and coworkers [8], which had a MAE value of 0.52 natural log units for an external test set of 45 organics in water, and the model by Mitchell and Jurs [16], which had a MAE value of 0.33 for an external test set comprising 25 organics in water.

The largest contributor to the prediction error in the current work could be due to the use of experimental data which is a compilation of all available literature data without consideration of the associated experimental uncertainties. To accumulate sufficient data for a reasonably generalized QSPR model, all data within the temperature range of 20°-40°C have been considered in the current work. The γ^{∞} values have been assumed to be temperature-independent in this narrow temperature range, which is a reasonable assumption considering the experimental uncertainty associated with the data.

Tables 8.4 and 8.6 list the most common descriptors for the eight different ensembles for cyclohexane and benzene, respectively. Due to the black-box nature of the artificial neural networks (ANNs), a quantitative assessment of the relative contribution of the different descriptors to the calculated γ^{∞} values is not possible; however, approximate qualitative interpretations can be made based on the type of descriptors. For example, the presence of the octanol-water partition coefficient in both the cyclohexane and benzene models indicates a strong correlation between the γ^{∞} values and the octanol-water partition coefficient values. This relationship is not surprising considering the theoretical mutual dependence between the two properties [1, 25]. In addition, the presence of the 2D-matrix based descriptors for both cyclohexane and benzene suggests a strong correlation between molecular shape and γ^{∞} values. Also, for the cyclohexane model, topological and 3D geometrical structures of the solvent affect the γ^{∞} values.

To compare the efficacy of the current modeling approach, the aqueous data set had been employed to develop a QSPR model. The results from this model are provided in Table 8.8, along with the results by Giralt et al. [8], using the same training and external test set data. Although, the model by Giralt et al. [8] has a significantly lower training set MAE, the current model performs better on the external set data, which indicates better generalization capability on new molecules unseen by the model. Mitchell and Jurs [16] also employed the same aqueous database to develop their QSPR model, but an external test set of only 25 molecules was used to validate their model. They report an MAE value of 0.33 for their external test set. A direct comparison between the model by Mitchell and Jurs [16] and the model from the current work is not possible due to the differences in the training and test sets employed. However, despite the larger external test set employed in the current work, the difference between the MAE values on the external test set between this model and the model by Mitchell and Jurs [16] is insignificant.

8.6. Conclusions

1. Separate non-linear QSPR models for ln γ^{∞} values in the temperature range of 20°-40°C were developed, using wrapper-based descriptor pruning techniques, for systems containing cyclohexane and benzene as solutes in a variety of solvents.
2. Two models each were developed for cyclohexane and benzene, as follows: Model 1 was created using ln γ^{∞} values for all available data in the model development; Model 2 was developed by employing ln γ^{∞} values of 85% of data from the original database, with 15% of the compounds reserved as an external test set.

3. For cyclohexane, the RMSE values of the training sets for Model 1 and Model 2 are 0.29 and 0.32 natural log units, respectively. The RMSE value for Model 2 of the external test set is 0.48 natural log units.
4. For benzene, the RMSE values of the training sets for Model 1 and Model 2 are 0.19 and 0.19 natural log units, respectively. The RMSE value for Model 2 of the external test set is 0.45 natural log units.
5. According to the descriptors identified in the current work, the octanol-water partition coefficient and the 2-dimensional shape of the molecule have significant effect on the γ^{∞} values for both cyclohexane and benzene systems.
6. The current model developed using the aqueous data set performs significantly better than the model by Giralt and coworkers [8] on an external test set of 45 compounds. The MAE value on the external test set for the model by Giralt et al. [8] is 0.52 as compared to a MAE value of 0.35 from the current model.
7. The resulting models from this work can be used to predict *a priori* the infinite-dilution activity coefficients of cyclohexane or benzene binary systems.

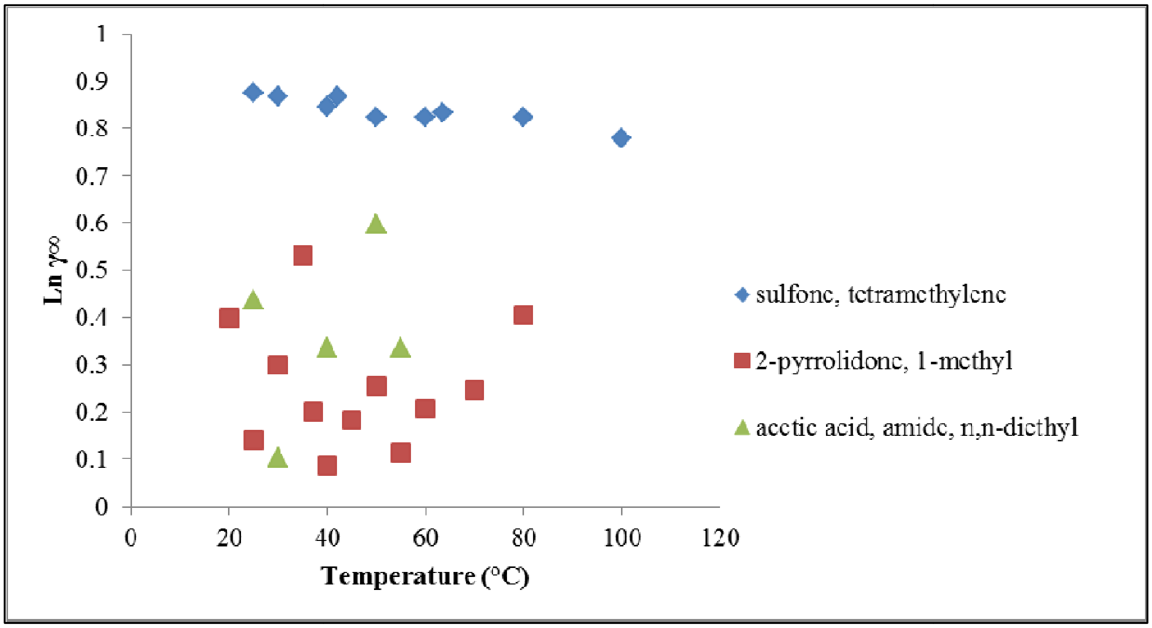


Figure 8.1: Variation in the ln γ^{∞} values with temperature, for benzene in three different solvents

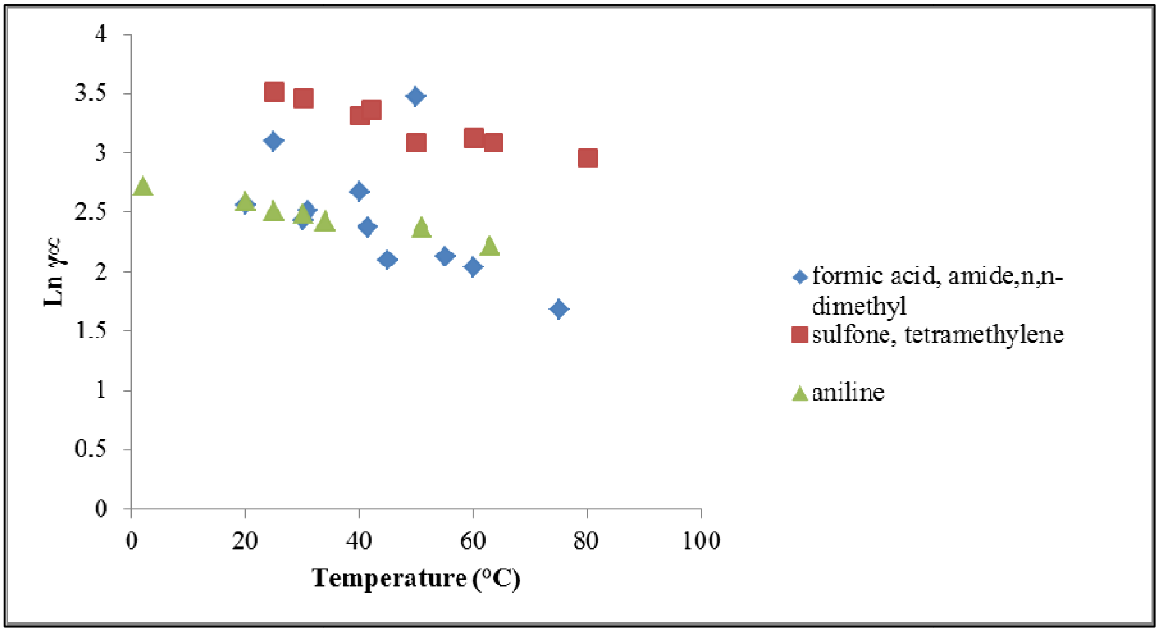


Figure 8.2: Variation in the ln γ^{∞} values with temperature, for cyclohexane in three different solvents

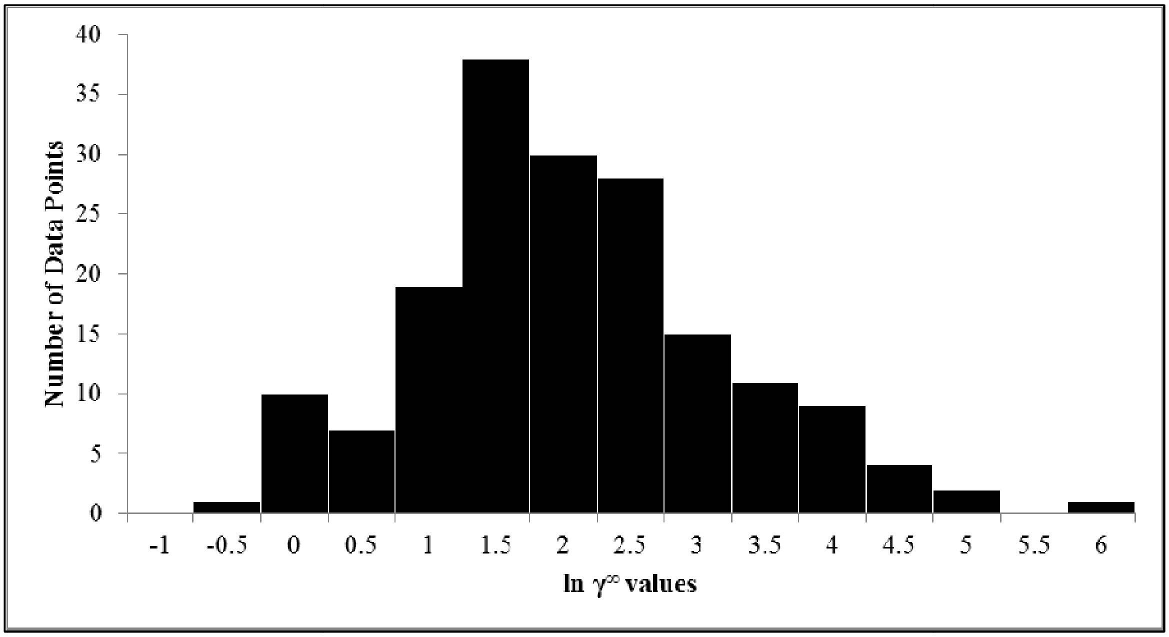


Figure 8.3: Distribution of the ln γ^∞ values in the final cyclohexane data set

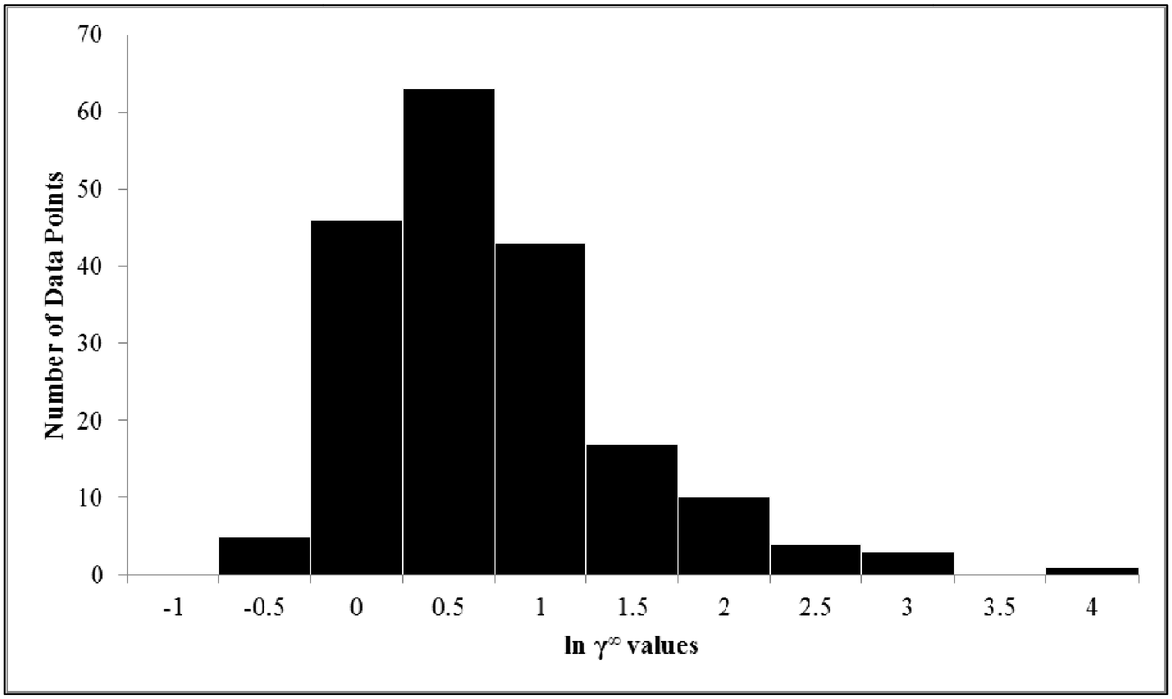


Figure 8.4: Distribution of the ln γ^∞ values in the final benzene data set

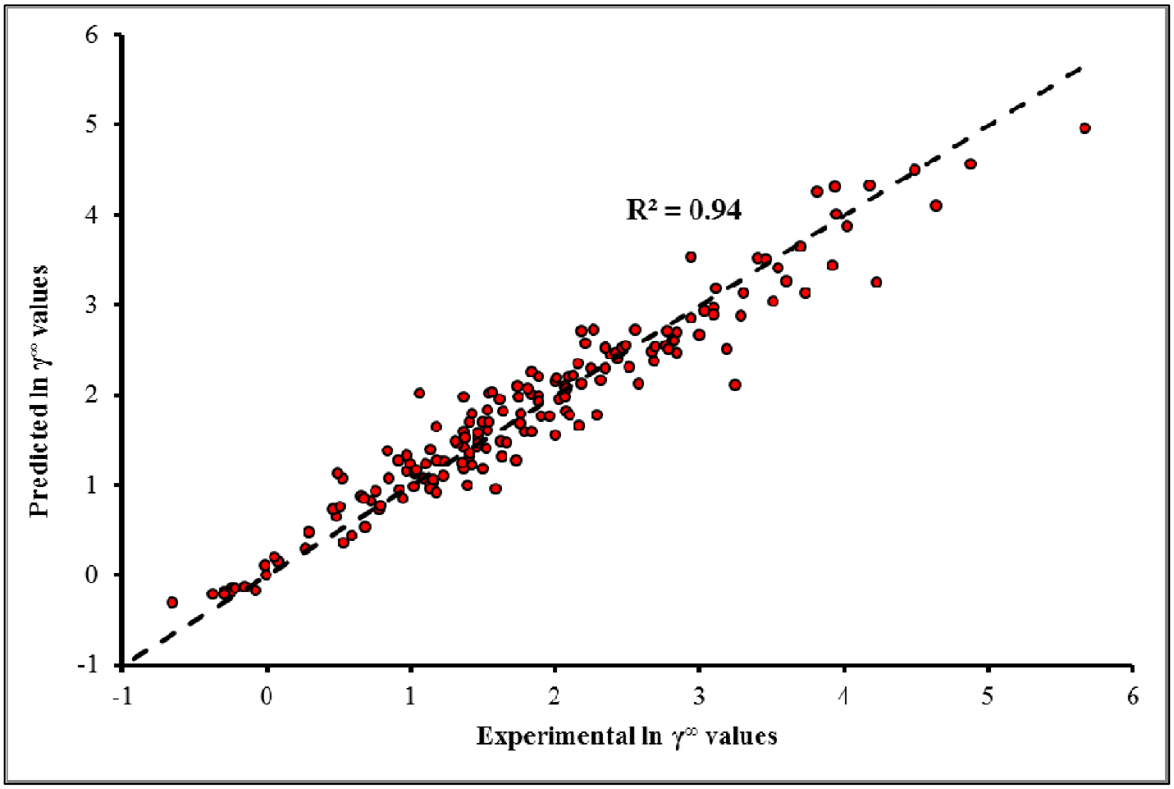


Figure 8.5: Comparison between the experimental and predicted ln γ^{∞} values by Model 1 for cyclohexane. The broken line represents perfect predictions

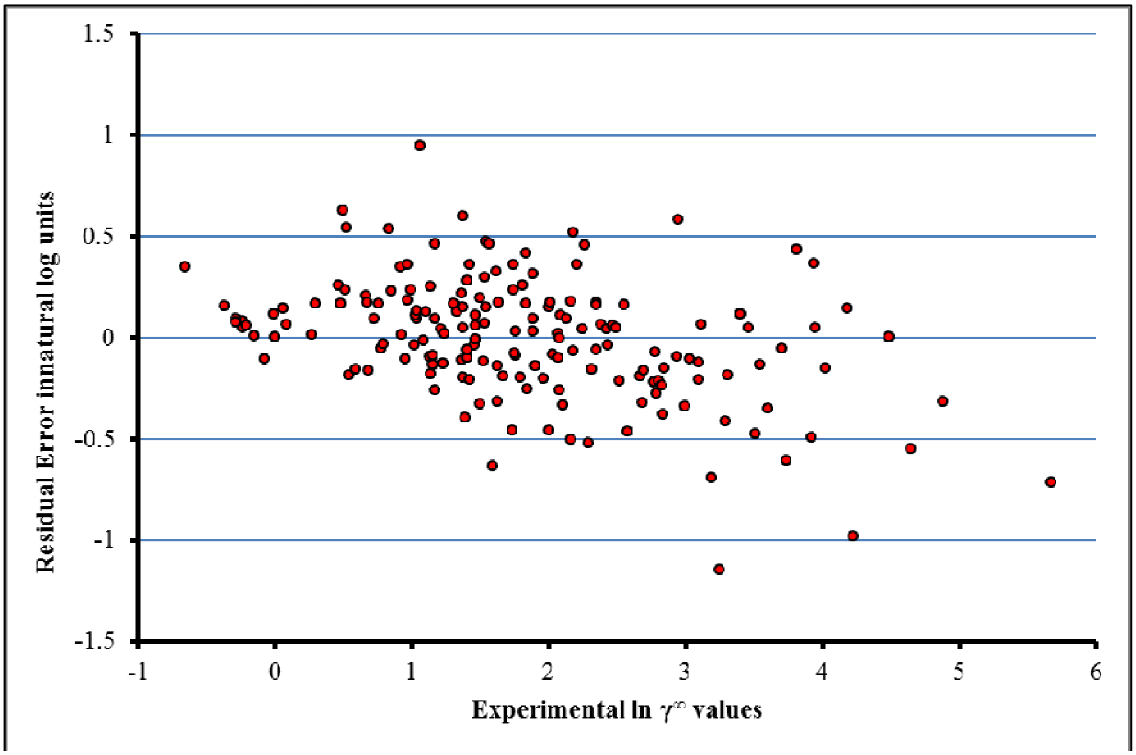


Figure 8.6: Residual error plot of the Model 1 predictions for cyclohexane

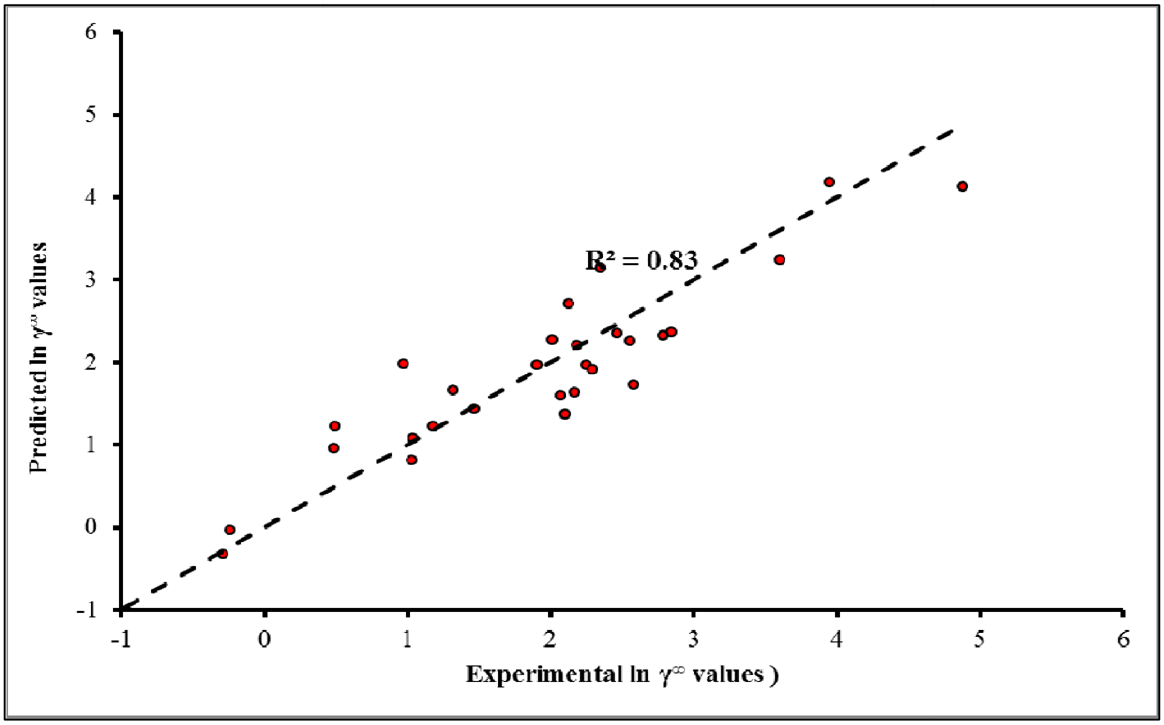


Figure 8.7: Comparison between the experimental and predicted ln γ^{∞} values for the external set data in Model 2 for cyclohexane. The broken line represents perfect predictions

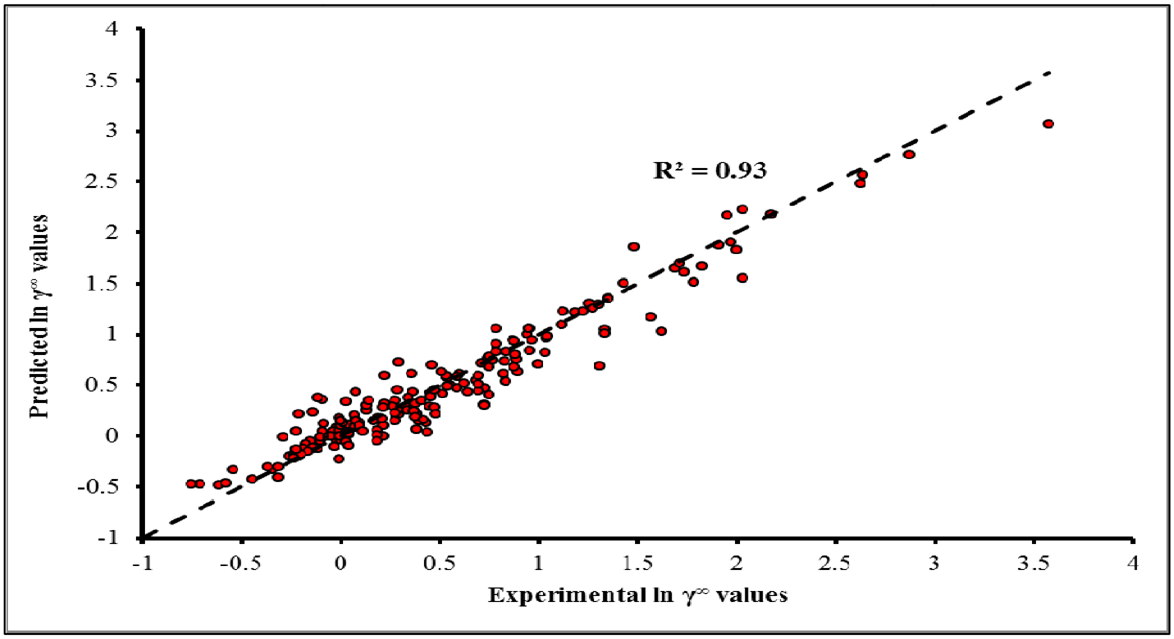


Figure 8.8: Comparison between the experimental and predicted ln γ^{∞} values for Model 1 for benzene. The broken line represents perfect predictions

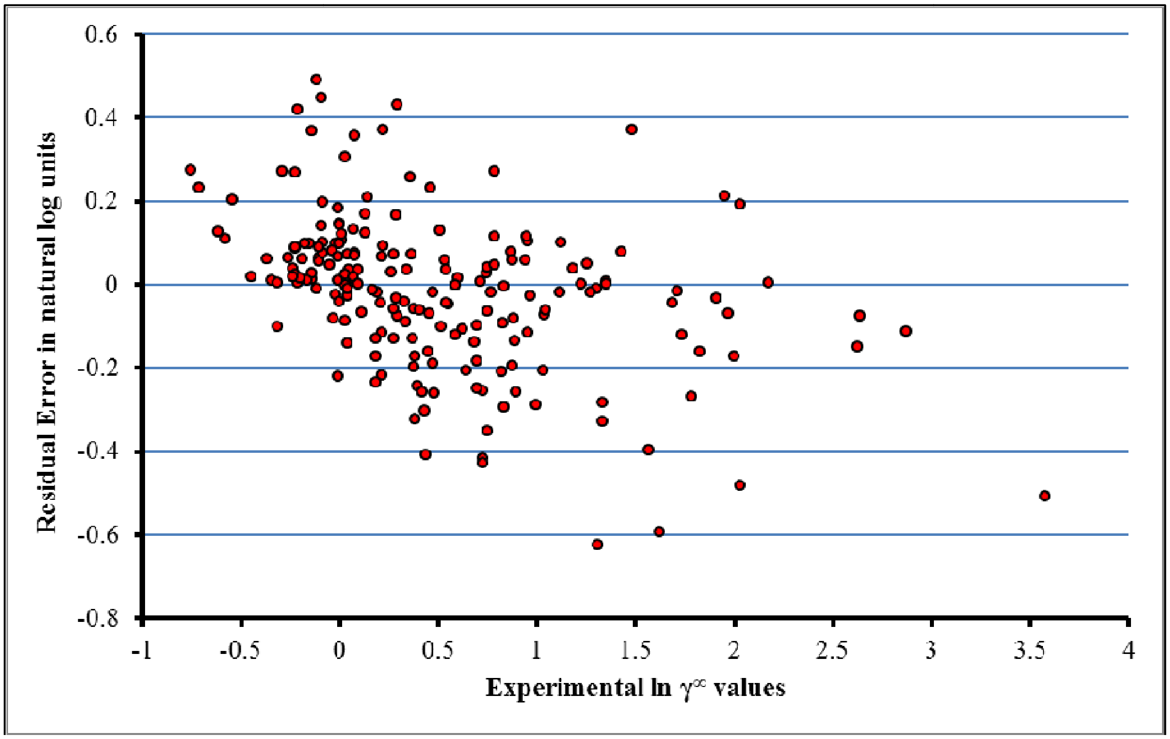


Figure 8.9: Residual error plot of the Model 1 predictions for benzene

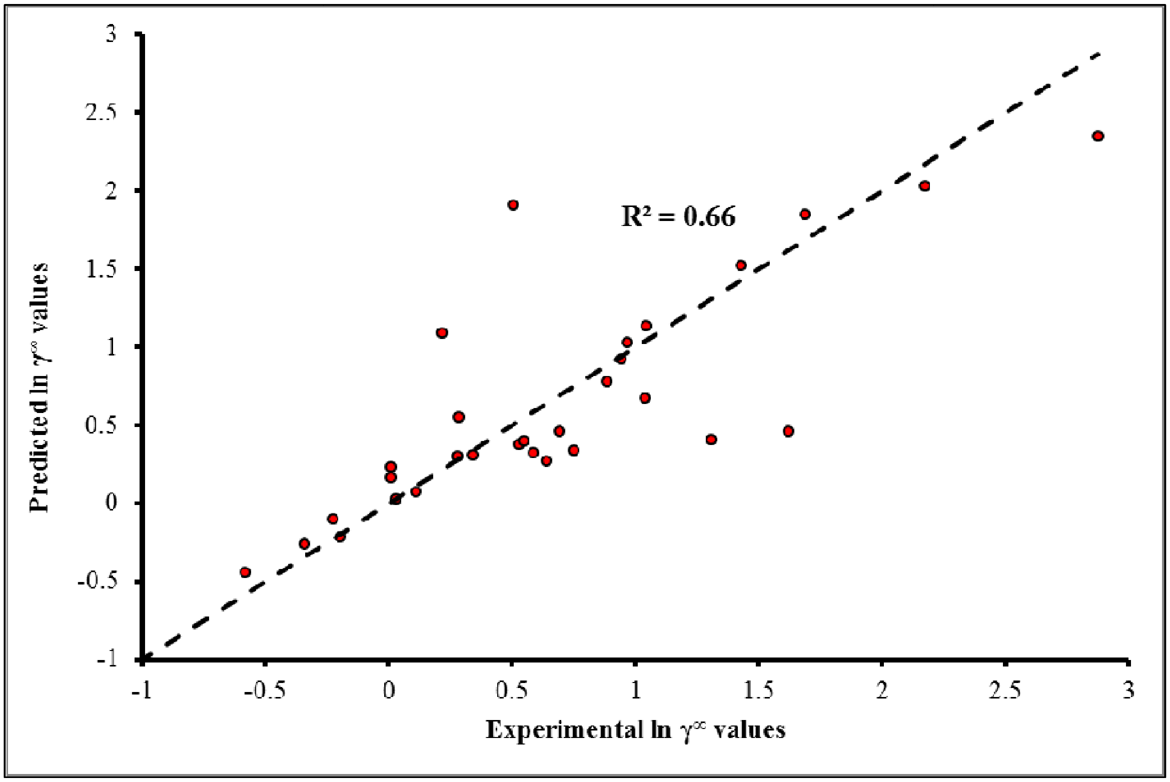


Figure 8.10: Comparison between the experimental and predicted ln γ^{∞} values for the external test set in Model 2 for benzene. The broken line represents perfect predictions

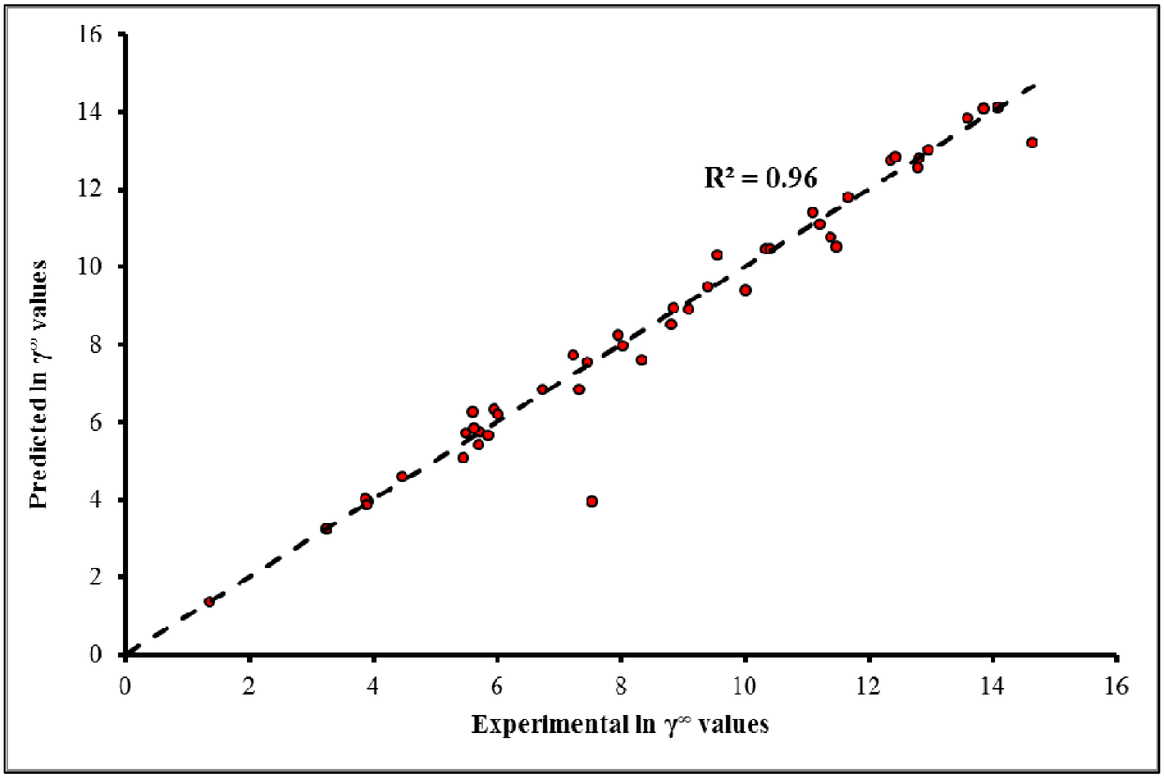


Figure 8.11: Comparison between the experimental and predicted ln γ^{∞} values for the external test set in the aqueous database. The broken line represents perfect predictions

Table 8.1: Characteristics of the final cyclohexane data set made up of 175 solvent data

Molecular Property	Minimum	Maximum	Mean	Standard Deviation
Ln (γ^{∞})	-0.65	5.70	1.80	1.1
Molecular weight (g/mol)	32.05	426.76	133.28	62.8
Octanol-water partition coeff. (Log K_{ow})	-1.7	11.3	1.5	2.1

Table 8.2: Characteristics of the final benzene data set made up of 192 solvent data

Molecular Property	Minimum	Maximum	Mean	Standard Deviation
Ln (γ^{∞})	-0.76	3.60	0.50	0.7
Molecular weight (g/mol)	32.05	426.76	137.75	68.4
Octanol-water partition coeff. (Log K_{ow})	-1.7	11.3	1.6	2.2

Table 8.3: List of the descriptors used in the final eight ensembles in Model 1 for cyclohexane

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	R1s	R1s	R1s	R1s	R1s	R1s	R1s	R1s
2	ALOGP	ALOGP	ALOGP	ALOGP	ALOGP	ALOGP	ALOGP	ALOGP
3	MATS2s	MATS1s	Chi0_EA (dm)	Chi0_EA (dm)	Chi0_EA (dm)	Chi0_EA (dm)	Chi0_EA (dm)	Chi0_EA (dm)
4	SM07_EA(bo)	SM07_EA(bo)	HyWi_B (m)	HyWi_B (m)	HyWi_B (m)	HyWi_B (m)	HyWi_B (m)	HyWi_B (m)
5	EE_B(m)	EE_B(m)	G (N..N)	G (N..N)	CATS2D_04_NL	CATS2D_04_NL	T (N..N)	T (N..N)
6	SpMAD_B (s)	SpMAD_B (s)	SM09_EA (bo)	SM09_EA (bo)	ALOGP2	ALOGP2	NsOH	NsOH
7	ZM1Kup	ZM1Kup	ATSC6m	ATSC6m	HOMA	AVS_D/ Dt	VR1_X	VR1_X
8	SpMax_Dt	SpPosLog_Dz (Z)	SM3_Dz (i)	SM3_Dz (i)	RDF130e	RDF130e	SpPosLog_Dz (Z)	SpPosLog_Dz (m)
9	Eig01_EA(ri)	GATS1m	VE3_G/D	QZZm	Eig12_AEA(ed)	Eig12_AEA(ed)	Eta_epsi	Eta_epsi
10	---	---	---	---	AVS_D/ Dt	---	VR3_D	---

Table 8.4: Physical meaning of the commonly occurring descriptors across the ensembles in Model 1 for cyclohexane

Descriptor	Descriptor Type	Physical Meaning
R1s	GETAWAY descriptor	Influence/distance matrix R, autocorrelation of lag 1 / weighted by I-state [26]
ALOGP	Molecular property	Ghose-Crippen octanol-water partition coefficient
Chi0_EA **(dm)**	Edge-adjacency index	Connectivity-like index of order 0 from edge adjacency matrix weighted by dipole moment. The edge adjacency matrix is derived from the H-depleted molecular graph and encodes the connectivity between graph edges. The entries of the matrix equal one if the considered bonds are adjacent and zero otherwise.
HyWi_B **(m)**	2D-matrix based descriptor	Hyper-Wiener-like index (log function) from Burden matrix weighted by mass
SpPosLog_Dz (Z)	2D-matrix based descriptor	Logarithmic spectral positive sum from Barysz matrix weighted by atomic number 2D [27]

Table 8.5: List of the descriptors used in the final eight ensembles in Model 1 for benzene

Descriptor #	Ensemble 1	Ensemble 2	Ensemble 3	Ensemble 4	Ensemble 5	Ensemble 6	Ensemble 7	Ensemble 8
1	HATS4i	HATS4i	VR1_L	VR1_L	VR1_H2	VR1_H2	Chi_Dz (m)	Chi_Dz (m)
2	MLOGP	MLOGP	MLOGP	MLOGP	H-048	H-048	P_VSA_LogP_6	P_VSA_LogP_6
3	SAdon	SAdon	R7e	R7e	SAdon	SAdon	SAacc	SAacc
4	P_VSA_m_1	P_VSA_m_1	SpPos_B (e)	SpPos_B (e)	SpAbs_B (e)	SpAbs_B (e)	SpPosLog_B (e)	SpPosLog_B (e)
5	SM08_EA (ed)	SM08_EA (ed)	SsOH	SsOH	J_B(i)	J_B (i)	piPC10	CATS2D_02_AL
6	TDB04i	TDB04i	VR3_B (v)	SpPosLog_Dt	SpAD_B (e)	SpPos_Dz(Z)	Mor14u	Mor14u
7	AVS_B (s)	AVS_B (s)	SpMin6_Bh (e)	AVS_B (p)	ATSC1m	ATSC1m	C-009	C-009
8	VE3_Dz (v)	---	HyWi_Dz(v)	AVS_Dz (v)	BLTD48	BLTD48	NdO	NdO
9	---	---	Chi0_AEA (dm)	---	ON0V	ATSC2p	---	---

Table 8.6: Physical meaning of the commonly occurring descriptors across the ensembles in Model 1 for benzene

Descriptor	Descriptor Type	Physical Meaning
MLOGP	Molecular property	Moriguchi octanol-water partition coefficient
SAdon	Molecular property	Surface area of donor atoms from P_VSA-like descriptors
Sp_Abs_B (e)	2D-matrix based descriptor	Graph energy from Burden matrix weighted by Sanderson electronegativity

Table 8.7: The errors for all models developed in this work

Model	Training Set			External Test Set		
	RMSE	MAE	R^2	RMSE	MAE	R^2
Model 1 for cylcohexane	0.29	0.22	0.94	---	---	---
Model 2 for cylcohexane	0.32	0.23	0.93	0.48	0.39	0.83
Model 1 for benzene	0.19	0.14	0.93	---	---	---
Model 2 for benzene	0.19	0.15	0.93	0.45	0.29	0.83
Model for aqueous data	0.38	0.28	0.99	0.67	0.35	0.96

Table 8.8: Comparison of the current model with literature models on the aqueous data set

Researchers	Model Type	Training Set MAE	Number of Molecules in External Test Set	External Test Set MAE
This work	Stochastic optimization and ANNs	0.28	45	0.35
Giralt et al. [8]	Neural classifiers and self-organizing maps	0.02	45	0.52
Mitchell and Jurs* [16]	ADAPT and neural networks	0.28	25	0.33

* The external test set used in the referenced work is different from the one employed by the other models

REFERENCES

1. Sandler, S.I., *Infinite Dilution Activity Coefficients in Chemical, Environmental and Biochemical Engineering.* Fluid Phase Equilibria, 1996. **116**(1-2): p. 343-353.

2. Fredenslund, A., R.L. Jones, and J.M. Prausnitz, *Group-Contribution Estimation of Activity Coefficients in Nonideal Liquid Mixtures.* AIChE Journal, 1975. **21**(6): p. 1086-1099.

3. 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.

4. Gmehling, J., et al., *A Modified UNIFAC (Dortmund) Model. 3. Revision and Extension.* Industrial & Engineering Chemistry Research, 1998. **37**(12): p. 4876-4882.

5. Hansen, H.K., et al., *Vapor-Liquid Equilibria by UNIFAC Group Contribution. 5. Revision and Extension.* Industrial & Engineering Chemistry Research, 1991. **30**(10): p. 2352-2355.

6. Voutsas, E.C. and D.P. Tassios, *Prediction of Infinite-Dilution Activity Coefficients in Binary Mixtures with UNIFAC. A Critical Evaluation.* Industrial & Engineering Chemistry Research, 1996. **35**(4): p. 1438-1445.

7. Tochigi, K., et al., *Determination of New ASOG Parameters.* Journal of Chemical Engineering of Japan, 1990. **23**(4): p. 453-463.

8. Giralt, F., et al., *Estimation of Infinite Dilution Activity Coefficients of Organic Compounds in Water with Neural Classifiers.* AIChE Journal, 2004. **50**(6): p. 1315-1343.

9. Sherman, S.R., et al., *Compilation and Correlation of Limiting Activity Coefficients of Nonelectrolytes in Water.* Industrial & Engineering Chemistry Research, 1996. **35**(4): p. 1044-1058.

10. Lazaridis, T. and M.E. Paulaitis, *Activity Coefficients in Dilute Aqueous Solutions from Free Energy Simulations.* AIChE Journal, 1993. **39**(6): p. 1051-1060.

11. Klamt, A., et al., *Refinement and Parametrization of COSMO-RS.* The Journal of Physical Chemistry A, 1998. **102**(26): p. 5074-5085.

12. Putnam, R., et al., *Prediction of Infinite Dilution Activity Coefficients Using COSMO-RS.* Industrial & Engineering Chemistry Research, 2003. **42**(15): p. 3635-3641.

13. Mackay, D. and W.Y. Shiu, *Aqueous Solubility of Polynuclear Aromatic Hydrocarbons.* Journal of Chemical & Engineering Data, 1977. **22**(4): p. 399-402.

14. Medir, M. and F. Giralt, *Correlation of Activity Coefficients of Hydrocarbons in Water at Infinite Dilution with Molecular Parameters.* AIChE Journal, 1982. **28**(2): p. 341-343.

15. Neely, B.J., et al. *Improved Quantitative Structure Property Relationship Models of Infinite-Dilution Activity Coefficients of Aqueous Systems*. in *Proceedings of the Sixth International Petroleum Environmental Conference*. 2004. Albuquerque, NM.

16. Mitchell, B.E. and P.C. Jurs, *Prediction of Infinite Dilution Activity Coefficients of Organic Compounds in Aqueous Solution from Molecular Structure.* Journal of Chemical Information and Computer Sciences, 1998. **38**(2): p. 200-209.

17. Rani, K.Y. and N.V.K. Dutt, *Estimation of Activity Coefficients at Infinite Dilution of Halocarbons in Water and Organic Compounds in Hydrofluoroparaffins Using Neural Networks.* Chemical Engineering Communications, 2002. **189**(3): p. 372-390.

18. Schult, C.J., *Design of Solvents for Extractive Distillation*. PhD. Thesis. 2000, Oklahoma State University, Stillwater, Oklahoma.

19. Tiegs, D., *Activity Coefficients at Infinite Dilution: C1-C9*. 1986: DECHEMA.

20. Tiegs, D., *Activity Coefficients at Infinite Dilution: C10-C36*. 1986: DECHEMA.

21. *Dragon Professional 6*. 2010, Talete SRL.

22. Tropsha, A., 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, 2003. **22**(1): p. 69-77.

23. Tetteh, J., et al., *Quantitative Structure−Property Relationships for the Estimation of Boiling Point and Flash Point Using a Radial Basis Function Neural Network.* Journal of Chemical Information and Computer Sciences, 1999. **39**(3): p. 491-507.

24. Hall, L.H. and C.T. Story, *Boiling Point and Critical Temperature of a Heterogeneous Data Set:□ QSAR with Atom Type Electrotopological State Indices Using Artificial Neural Networks†.* Journal of Chemical Information and Computer Sciences, 1996. **36**(5): p. 1004-1014.

25. Tse, G. and S.I. Sandler, *Determination of Infinite Dilution Activity Coefficients and 1-Octanol/Water Partition Coefficients of Volatile Organic Pollutants.* Journal of Chemical & Engineering Data, 1994. **39**(2): p. 354-357.

26. Consonni, V., R. Todeschini, and M. Pavan, *Structure/Response Correlations and Similarity/Diversity Analysis by GETAWAY Descriptors. 1. Theory of the Novel 3d Molecular Descriptors.* Journal of Chemical Information and Computer Sciences, 2002. **42**(3): p. 682-692.

27. Ivanciuc, O., *QSAR Comparative Study of Wiener Descriptors for Weighted Molecular Graphs.* Journal of Chemical Information and Computer Sciences, 2000. **40**(6): p. 1412-1422.

CHAPTER 9

COMPUTER-AIDED MOLECULAR DESIGN (CAMD): METHODOLOGY AND APPLICATIONS

9.1. Introduction

The demand for newly designed molecules that enhance existing processes and satisfy more stringent operating requirements in technology has been increasing. However, the rational design of molecules with desired properties poses a significant challenge to engineers attempting to meet the needs of various industries, including pharmaceuticals, polymers, petrochemicals and construction [1-3]. The traditional approach of identifying molecules with desired properties involves testing thousands of molecules for their chemical and physical properties, which is an expensive and laborious undertaking. Hence, rational design techniques, such as computer-aided molecular design (CAMD), have found wide application in recent years [4, 5]. CAMD methods have been employed successfully to identify novel molecules with superior properties for a wide range of applications, including solvent design/selection [6] and design of chloro-fluro-carbon substitutes, alternative process fluids, polymers [2] and drugs [7]. In pharmaceutical industries, CAMD is used to discover novel drugs for targeted applications, while meeting health constraints, such as minimal side effects and toxicity.

In contrast to traditional methodologies, CAMD methods expedite the design process by predicting the behavior of potential molecules using reliable property models. CAMD involves the design of new molecules based on a specified set of desired properties and can be classified as (a) forward CAMD, which involves computation of chemical, physical and biological properties from the molecular structure, and (b) inverse CAMD, which involves generation of a molecular structure with the desired properties [8, 9].

A typical CAMD design algorithm utilizes two key components: (a) a method for generating candidate molecules, and (b) accurate models to predict the pertinent physicochemical properties of the newly generated molecules. Property predictions for the generated molecules are usually completed using group-contribution methods, equation-of-state approaches and quantitative structure-property relationship (QSPR) models. Figure 9.1 presents a simplified view of the various stages involved in CAMD.

9.2. State of the Art in CAMD

In general, CAMD techniques can be divided into the following categories: knowledge-based generation and test methods [10, 11], mathematical optimization methods [12, 13] and combinatorial optimization methods [6, 14, 15]. Knowledge-based methods utilize expert rules that guide the design process; however, many non-linear structure property relationships are not easily simplified to rules. Mathematical optimization methods utilize mixed-integer, non-linear programming (MINLP) approaches and are computationally expensive to perform and have a high probability of being trapped in local minima (sub-optimal molecular structures) for a highly non-linear system of equations. Recently, combinatorial approaches that involve stochastic optimization methods such as simulated annealing and genetic algorithms have been applied successfully to CAMD. These methods have several advantages which makes them widely applicable, such as ease of applicability and independent implementation with respect to the property prediction portion of the algorithm. However, they are highly dependent on the parameters used for the various mutation operations. Inverse QSPR methods for generating new structures have also been implemented [16]. These methods involve the use of specific descriptor types leading to accurate property prediction, as well as allowing for molecules to be designed based on these same descriptors; however, the design of feasible molecules using these specific descriptors is usually difficult [17]. Another disadvantage of this method is the inability to account for 3-dimensional (3D) molecular descriptors, which in most cases lead to better predictions than 2-dimensional (2D) descriptors alone for many physical properties. Further, limitations are usually placed on the types of descriptors that can be used in this approach such as the monotonically increasing or decreasing descriptors employed by Miyao et al. [18] in their inverse-QSPR approach.

The majority of applications of CAMD employ connectivity indices or fragment-based QSPRs, which decrease the execution time for the algorithm but are not as accurate as the 3D descriptor-based QSPRs for many important molecular properties. The performance of CAMD techniques, however, relies heavily on the accuracy of the underlying predictive models. Korichi et al. [19] have used 2D and 3D descriptors for computational design of aromatic molecules and reported that 3D descriptors perform better than 2D descriptors in their design framework. Further, in most studies, the search space is limited to a certain family of molecular functional groups. This leads to a reduction in computational time at the cost of failing to discover better molecules that may be present outside the search space. Therefore, there is a need to develop generalized molecular search algorithms for CAMD.

In summary, the CAMD literature suggests that a need exists for a highly accurate but reasonably fast algorithm that searches for globally-optimal structures of molecules satisfying a certain set of molecular property constraints. Further, this algorithm should be capable of handling non-linear constraints and be generally applicable for a wide range of molecular design problems. In addition, an algorithm that can be fully automated would be much more efficient, and it would also help in reducing the errors associated with human involvement.

9.3. CAMD Methodology

A genetic algorithm (GA) based approach was used in the current work to identify the optimal molecular structures that satisfy specific molecular design constraints. The basic premise of the GA approach relates to the theory of natural selection, as famously proposed by Charles Darwin [20], which asserts that individuals that respond better to environmental stresses or changes in a given population have a better chance of transferring their genetic material to future generations. Over a large number of generations, this process leads to elimination of the weaker individuals and proliferation of the stronger individuals in a specific population. In biological evolution, the two aspects of change through reproduction from the parents to the offspring and the selective survival of the offspring are sufficient to produce generations of individuals that are progressively better suited to the existing environment.

The same concepts of natural selection can be extended to CAMD, where the molecules represent individuals, and the selection pressure is applied using an objective function and design constraints. The fitness of the molecules is assessed in terms of a specified objective function and the number of design constraint violations. The reproduction of individuals is simulated using crossover and mutation operations, and natural selection is simulated using various selection procedures such as tournament selection and roulette wheel selection. The overview of the GA-based CAMD algorithm employed in the current work is provided in Figure 9.2. The details of the algorithm are presented in the following sections.

9.3.1 Problem Formulation: According to Achenie et al. [21], the basic CAMD problem can be defined as: "Given a set of building blocks and a specified set of target properties, determine the molecule or molecular structure that matches these properties." Therefore, identifying the desired target properties of the chemical compounds to be designed is the first step in CAMD processes. A knowledge-based system is required to identify target properties, as well as their corresponding property values. A typical CAMD problem would need the following information:

1. The desired application of the designed molecules
2. The relevant operating conditions of the process
3. The design criteria based on molecular descriptors or properties
4. A property prediction method to predict the relevant molecular properties
5. A quantitative measure of the fitness of the generated molecules.

The CAMD problem is then to design molecules that have the optimal value of a particular fitness function, and at the same time adhere to specific design criteria. In mathematical form, this can be expressed as:

$$
\begin{aligned}
&\min_{\mathbf{x}} \mathrm{F}(\bar{\mathrm{x}}) \\
&\text{Subject to:} \\
&\mathrm{g}(\bar{\mathrm{x}}) = 0 \\
&\mathrm{h}(\bar{\mathrm{x}}) \leq 0
\end{aligned}
\tag{9.1}
$$

where, *F* is the fitness function that is dependent upon the vector of molecular properties denoted as x□$_{i}$. The design criteria, operating conditions and logical constraints are represented using *g* and *h*, respectively. The above formulation can be used to treat both linear and non-linear objective functions and constraints and allows for analytical or numerical techniques of evaluation.

9.3.2 Fitness Function and Constraint Handling: The fitness function is a key component of a GA, and the value of this function determines the probability of the individuals in the population to participate in the reproductive operations such as crossover and mutation. There are many different variations of fitness functions in the literature which are tailored to specific CAMD problems. Due to the stochastic nature of the GA-based CAMD algorithm, the fitness function does not have to be a well-characterized equation with calculable derivatives; however, the fitness function does need to act as a relevant guideline that can help distinguish between two molecules in a population [22]. Therefore, the fitness function should be a function of the desired target property values, the individual molecular property values, the user-specified tolerance level based on the confidence in property prediction methods, and some user-specified tuning parameters to set penalties for molecules whose properties deviate from the target property values. In general, the choice of fitness function is based on the user's prior knowledge of the important design criteria for the specific problem. Although many forms of fitness functions can be used with GAs, the majority of the CAMD algorithms in the literature employ a function that varies continuously between 0 and 1. Molecules with a fitness function value close to 0 are considered closer to the optimal molecular structure than molecules whose fitness values are closer to 1. Venkatasubramanian et al. [1, 5] were the first to use a Gaussian-like function to calculate the fitness value in a CAMD algorithm. When designing for target properties with both lower and upper bounds, the fitness function proposed by Venkatasubramanian et al. [1, 5] had the following Gaussian-like form:

$$\mathrm{F} = \exp\left(-\alpha\left[\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{(\mathrm{P_i}-\overline{\mathrm{P}_\imath})}{(\mathrm{P_{i,max}}-\mathrm{P_{i,min}})}\right]\right) \tag{9.2}$$

where, P_i is the i^th^ property value, $\mathrm{P_{i,max}}$, $\mathrm{P_{i,min}}$ and $P\square_i$ are the maximum, minimum and average values of the ith property, respectively, and α is the fitness decay factor.

When designing for target properties with only a lower bound or an upper bound, the following sigmoidal form of the fitness function is commonly used:

$$\mathrm{F_i} = \frac{1}{1+\exp\left(-\alpha\left[\frac{\mathrm{P_i}-\mathrm{P_{i,r}}}{\mathrm{P_{i,range}}}\right]\right)} - 0.5 \tag{9.3}$$

where, $\mathrm{F_i}$ is the contribution of the ith property to the overall fitness F, $\mathrm{P_i}$ is the ith property value, $\mathrm{P_{i,r}}$ is the lower or upper bound on the ith property and $\mathrm{P_{i,range}}$ is the overall possible range of the ith property, which is used to normalize the contribution of each property toward the fitness function. The 0.5 term in Equation 9.3 ensures, that the $\mathrm{F_i}$ value is zero when the property value $\mathrm{P_i}$ is equal to the lower or upper bound value $\mathrm{P_{i,r}}$. This value is a matter of convenience, and the use of any other value would not affect the shape of the fitness function, but it would shift the fitness function higher or lower with respect to a base line value.

The magnitude of the decay factor, α, determines how strictly an individual molecule is penalized for not meeting the desired property values. A large value of the decay factor would mean that small deviations from the target value are penalized heavily leading to small fitness values; whereas a small decay factor would be more lenient, and large deviations from the desired property values would lead to moderate fitness scores. The decay factor plays a significant role in influencing the selection pressure of the algorithm. For instance, in the case of a large decay factor, small differences in the deviations from target values between two molecules are amplified leading to widely different fitness scores. In the extreme case, this may lead to premature convergence of the GA to a population of similar and suboptimal solutions. When a small decay factor is used, the fitness function is more forgiving, and the GA may accept solutions with large deviations from a desired target which would lead to an increase in the diversity of the sampled solutions. The disadvantage in this case would be the additional computational time required for algorithm convergence.

The fitness function must also account for the various design constraints associated with the problem. One approach is to devise a molecular generation scheme, which generates only those molecules that satisfy all constraints. In other words, the constraints are handled in the crossover and mutation stages, instead of in the selection stage of the GA algorithm. This approach may work for simple constraints but is impossible to implement for constraints based on complex molecular properties. Another method of constraint satisfaction in GA is to reject individuals that violate constraints, i.e., the infeasible individuals. Infeasible individuals can appear as the result of the genetic operators, but these individuals are not admitted to the new generation. This method may work when the feasible region of the search space for molecules is reasonably large; however, when this feasible search region is small, rejection of infeasible individuals may lead to the loss of important genetic information that might be useful when coupled with genetic information from other individuals in future generations. Another common technique for handling constraints is to penalize the infeasible molecules [23]. Venkatasubramanian et al. [23] used the following fitness function, F_{total}, in their work for a minimization problem:

$$F_{\mathrm{total}} = F + \delta \sum_{i=1}^{P} \emptyset_{i} \qquad (9.4)$$

where, F is the fitness function associated with the molecular property that needs to be optimized, *P* is the total number of design constraints, *δ* is the penalty coefficient and $\square_i$ is the penalty weight associated with the ith penalty term.

The magnitude of the penalties in Equation 9.4 depends on the extent of constraint violation. The selection of molecules for crossover or mutation is based on the total fitness, which is the weighted sum of fitness and penalty. The infeasible individuals participate in the genetic process, as they are still considered capable of delivering useful offspring; however, a careful adjustment of the penalty weights is required. If the penalty weights are too low, infeasible individuals could be preferred to slightly less fit but much more feasible individuals, or application of high penalty weights may result in the loss of useful genetic information, and the process may converge to feasible but sub-optimal individuals.

Another method of property constraint handling is the inclusion of constraints in the fitness function, where a property constraint is treated as an additional contribution to the fitness function (Equation 9.3). However, this method is only suitable when the constraints are simple inequalities, leading to lower or upper bounds in a particular molecular property value. Since all constraints encountered in the current work belong to this category, the sigmoidal form of the fitness function shown in Equation 9.3 is used to account for both the actual fitness function and the constraints. For example, consider the design of molecules having a normal boiling point (NBP) of less than 400 K. The fitness values calculated using Equation 9.3 are plotted as a function of NBP values in Figure 9.3 for four different fitness decay values. For values of α close to one, there is a better distinction in fitness values between molecules having slightly different NBP values; however, this would lead to longer computational times for the GA to reach convergence. At higher values, convergence can be achieved faster, but there is not enough distinction between molecules, including even those that have NBP values separated by 500 K. A middle value of 5 for α is a compromise between premature convergence and long computational times and is used in this work.

Although Equation 9.3 is suited perfectly to represent the minimization or maximization of property values, the handling of inequality constraints in property values requires a small modification. According to Figure 9.3, among molecules that satisfy the NBP constraint, those molecules with NBP values close to zero are preferred over molecules with NBP values slightly lower than 400 K. From a design perspective, however, molecules with NBP close to zero might not offer any practical advantage over molecules whose NBP values are closer to 400 K. In such circumstances, using Equation 9.3 as written, would drive the GA algorithm to search for molecules with NBP values close to zero and would lead to a loss of diversity in the population. To avoid this, the following modification has been made to Equation 9.3 in the current work:

$$\mathrm{F_i} = \frac{1}{1 + \exp\left(-\alpha\left[\frac{\mathrm{P_i} - \mathrm{P_{i,r}}}{\mathrm{P_{i,range}}}\right]\right)} - 0.5 \tag{9.5}$$

$$\mathrm{F_i} = \begin{cases} \mathrm{F_i} & \text{if } \mathrm{F_i} > 0 \\ 0 & \text{if } \mathrm{F_i} \leq 0 \end{cases}$$

The above modification would lead to an equal probability of selection for all molecules that satisfy a particular inequality constraint. Of course, this modification is only applicable when there is no design advantage for molecules that have relatively lower or higher property values, as long as these values are within the constrained property range. For cases where there is a design advantage for lower or higher values within the constrained range, Equation 9.3 in its original form is applied.

The total fitness function in the current work is now calculated by summing the weighted fitness functions for each property constraint:

$$\mathrm{F_{total}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \emptyset_{\mathrm{i}} \mathrm{F_i} \tag{9.6}$$

where, $\square_i$ is the weight term associated with the ith constraint, F_i is the fitness contribution of the ith constraint and n is the total number of constraints in the problem.

The weights for each constraint are decided by the user based on experience and specific requirements.

9.3.3 Genetic Representation: When designing a GA, choosing a representation scheme is an important step. Genetic algorithms traditionally operate using bit string encoding of the chromosomes in the population. Employing a bit string representation for GA-based CAMD algorithms would involve constructing large binary matrices for all possible functional groups or atomic fragments in the molecules. Dealing with these matrices would quickly become impractical even for moderately sized molecules [1]. A practical alternative is to use the representations used commonly by chemists, where molecular and atomic fragments are represented using symbols. For CAMD, one can imagine a number of such molecular representation schemes, ranging from simple strings (line notation) to more complicated 3D structures. Despite using the same underlying principles of inheritance and evolution, the results obtained with different schemes can vary widely. These differences are due to the representation scheme along with the recombination operators limiting the exploration of the search space to certain regions. The 3D representations are most commonly used in protein-docking search algorithms, where the 3D structures of the protein and ligand are significant factors. In the current work, the property prediction QSPR models are based on 3D molecular descriptors, and theoretically, the CAMD algorithm must also be based on a 3D representation of the molecules. However, dealing with the higher level 3D representation of the molecules is far from simple, and could prove computationally cumbersome even for medium-sized molecules. Therefore, a lower level representation scheme based on line notations is used in this work.

Line notations are a popular method for representing chemical formulas. The simplified molecular input line entry system (SMILES) is the most popular line notation, which is based on rules derived from molecular graph theory. The SMILES notation allows rigorous structure specification by use of natural grammar and is well suited for high-speed machine processing. SMILES have four basic rules which apply to 98% of the molecules typically encountered in solvent design [24].

Rule 1: Atom Specifications (see Table 9.1)

a. Use ordinary atomic symbols C, N, O, S, F, Cl, and Br.

b. Suppress hydrogen except on pyrrole nitrogen where it is [nH].

c. Other atoms and any charges are placed in brackets; e.g. [N+].

d. Use lower case for sp2-hybridized atoms and upper case for all other atoms.

Rule 2: Bond Specifications (see Table 9.2)

a. Bonds are always assumed to be single bonds unless specified otherwise (example: ethane, represented as CC).

b. Double bonds are represented by an equal symbol (example: acetaldehyde, represented as O=CC).

c. Triple bonds are represented by a pound symbol (example: hydrogen cyanide, represented as C#N).

Rule 3: Branching Specification (see Table 9.3)

a. A branched group is placed in parentheses (example: isobutyric acid, represented as O=C(O)C(C)C).

b. Branches can be stacked (example: fluroform, represented as C(F)(F)F) or nested (example: 4-heptanoic acid, represented as CCCC(C(=O)O)CCC).

c. No predefined limit to how deep branching may be nested.

d. Most implementations, however, define such a limit, typically between 10 and 50.

Rule 4: Ring Specification

a. Ring closure is specified by appending matching digits to the joined atoms.

b. Pick one bond in each ring numbering them in any order. Break the numbered bonds, appending the bond number to the atoms on the ends of the bonds. This leaves an acyclic structure which can always be specified using the rules for specifying atoms and branching (Figure 9.4).

c. There are usually many different, but equally valid descriptions of the same structure (see Figure 9.5).

d. A single atom may have more than one ring closure.

e. A ring closure digit can be reused (see Figure 9.6).

A molecule always has a unique SMILES structure but can have multiple 3D conformations of which one conformation will possess the lowest conformational energy. In this work, an automated algorithm that searches for this minimum energy conformation starting from the line notation of the molecule was developed and combined with the CAMD algorithm. This conformational energy search algorithm ensures a one-to-one mapping between the lower level line notations (which allows for easy crossover and mutation operations) and the higher level 3D representation (which is essential for molecular property predictions).

9.3.4 Initial Population: In creating an initial population, two decisions need to be made: the size and the source of the initial population. The size of the population for GAs is usually proportional to the number of adjustable parameters for the specific molecular design of interest. While a larger population would increase the required computational power and ensure globally optimum solutions, a smaller population would require lower computational power and could lead to sub-optimal solutions. The size of the initial population is governed largely by the type of attachments used for new structure generation. If molecular fragments or functional groups are used instead of atomic fragments, then a larger population size would be needed to ensure global optimal solutions. When functional groups are used, the algorithm has a tendency to polarize the results, which means that if a molecule generated in the initial generation shows high fitness value, then the probability of such a molecule being involved in future reproduction operations is high. This process results in the generation of molecules that are similar to the high fitness molecule, and consequently molecules with low fitness that could potentially lead to a better candidate molecule after a few generations are eliminated. The selection of the initial population should reflect a wide range of structural diversity, while considering the design constraints. The initial population can arise from random structures, which satisfy some minimal fitness criteria, or from the results of other calculations or studies. Similar to choosing the population size, there is no single best method for generating initial populations for different applications of CAMD. Of note, GAs with initial populations that are fitter but not sufficiently diverse would most probably yield inferior final solutions.

9.3.5 Genetic Operators: In each generation, individuals from the current population are selected and processed using genetic operators to create a new population. The selected individuals are referred to as the mating pool individuals. In GAs, several types of genetic operators such as crossover, mutation, elitism and reproduction are used. Crossover and mutation operators must be carefully designed since their choice contributes highly to the performance and convergence speed of the GA. In this work, elitism, crossover and mutation are employed as operators and are described in greater detail below.

Elitism: A part of the new population of structures can be created by simply copying, without change, selected individuals from the present population. This gives a possibility of survival for already developed fit solutions. In the current work, the best two individual molecules from each generation are allowed to pass over to the next generation without any modifications to their structures.

Crossover: Crossover is a mechanism that promotes interbreeding of molecules. The genetic material of the parents is combined to form new molecules that retain some characteristics of the parent molecules. The first step of the crossover operation is identification of structural fragments which are suitable for crossover. Two types of crossover are possible. They are (a) single-point crossover, where the fragment in the terminal position is cut and connected to the similar terminal position from another molecule, and (b) multipoint-crossover which involves the excision of an internal portion of a molecule and insertion into a molecule with a similar region removed. A diagrammatic representation of one-point and two-point crossover is presented in Figures 9.7 and 9.8, respectively. Both methods begin with the random selection of crossover type, followed by the selection of a random pair of distinct parents from the previous generation.

In later phases of evolution, adaptively changing the crossover rate might be beneficial (higher crossover rates in early phases and a lower rate at the end of the genetic algorithm), to keep the fitter individuals intact. Sometimes, using several different types of crossover at different stages of evolution might be beneficial. This so-called "knowledge-augmented" crossover operation constructs offspring from the parents by making use of domain knowledge related to a given problem.

In the current work, the molecules can only undergo only the simple one-point and two-point crossover operations; each of these operations occurs with a probability of 0.25 for a pair of selected molecules.

Mutation: In CAMD algorithms, mutation is an important operator, which performs local search around a molecular structure. The mutation operator has to be implemented carefully when dealing with chemical systems, due to the constraints imposed by the valency rules and the requirement to keep all the atoms in a molecule connected. The mutation operator applied is usually selected at random. In the current work, eight different mutation operations are performed (Table 9.4). These operators are similar to those used by Lameijer et al. [25]. The probability of a mutation operation for a selected molecule is 0.5, and the different mutation operations have uniform probabilities of being selected. Further, for mutations that involve adding or inserting a new atom into the molecule, the probabilities of the new atom being a certain type are tabulated in Tables 9.5 and 9.6. In addition, the mutation operation is not carried out if the selected molecule does not meet the requirements for the particular mutation operation. For example, if a molecule without any rings is considered for the ‘break-ring’ mutation operation, the operation is not performed, and the algorithm returns to selecting another molecule and another genetic operator.

The different mutation operations are described briefly as follows:

1. *Add Atom:* An atom in the molecule whose implicit valence is not satisfied is picked randomly, and one of the ‘new’ atoms from Table 9.5 is bonded to it using a single bond. The second column in the table lists the probabilities of selection for the corresponding ‘new’ atoms.
2. *Insert Atom:* An atom in the molecule whose implicit valence is not satisfied is picked randomly, and one of the ‘new’ atoms from Table 9.6 is bonded to it using a single bond. The second column in the table lists the probabilities of selection for the corresponding ‘new’ atoms. The current algorithm cannot insert atoms in a ring.
3. *Delete Atom:* An atom in the molecule that is bonded to only one other non-hydrogen atom (through a single bond) in the molecule is deleted.
4. *Uninsert Atom:* An atom in the molecule that is bonded to exactly two other non-hydrogen atoms in the molecule is deleted. A single bond is now created between the two neighboring non-hydrogen atoms.
5. *Increase Bond-Order:* Two neighboring atoms in a molecule, whose implicit valence is not satisfied, are selected and an additional bond is created between them (a single bond is updated to a double bond, and a double bond is updated to a triple bond).
6. *Decrease Bond-Order:* Two neighboring atoms in a molecule, whose implicit valence is not satisfied, are selected and the bond-order is decreased by one (a double bond is updated to a single bond, and a triple bond is updated to a double bond). *Create Ring:*

Similar to increase bond-order operation, but operates between two unconnected atoms in the molecule. A single bond is created between two randomly selected unconnected atoms.

7. *Break Ring:* A single bond in the molecule that is inside a ring is chosen and deleted.

9.3.6 Selection: In each generation of a GA, some individuals are selected to the mating pool, where these individuals exchange genetic material and produce offspring that comprise the next generation population. A "good" mating pool of individuals can be ensured by employing an effective selection strategy, which enforces a high selective pressure leading to the selection of the best individuals in the population and consequently, to faster convergence of the algorithm. However, while a high selection pressure may lead to premature convergence to sub-optimal solutions, low selection pressure leads to an increase in population diversity, but slower convergence. Therefore, an effective selection strategy must strike a balance between convergence speed and diversity.

Selection strategies commonly include proportionate-based selection and ordinal-based selection [26]. In proportionate-based selection, the individuals are selected based on their fitness values when compared to other individuals in the population. Examples include proportionate selection [27] and stochastic universal sampling [28]. In ordinal-based selection, the individuals are selected based on their relative fitness ranking with respect to other individuals and not on the basis of their absolute fitness values. Common examples of ordinal-based selection strategies are tournament selection [29] and linear ranking [28]. Ordinal-based strategies are usually preferred over proportionate-based strategies for many reasons, including stochastic sampling errors and scaling problems associated with the latter methods [26].

Tournament selection was chosen as the selection strategy in the current work, because of its advantages over proportionate-based selection and simplicity of implementation. In tournament selection, a specified number of parents, known as the tournament size, are chosen in each generation and are allowed to enter a competition. The winner is decided based on the fitness values of the individuals. The process is repeated until the desired number of offspring molecules has been generated. This method is useful if the population has some individuals with high fitness, and it biases the selection toward the above-average individuals while at the same time not allowing the super-fit individuals to dominate the search. This differs from other selection schemes in that the selection probability is fairly static; therefore, no update of selection probabilities is required. Binary tournament selection, where only two individuals compete in each tournament, was implemented in the current algorithm.

9.3.7. Property Prediction: A typical CAMD algorithm utilizes two key components, which are a search method for generating candidate molecules, and models to predict the pertinent physiochemical properties of the generated candidate molecules. Property predictions for the generated molecules are usually done using group-contribution methods, equation-of-state approaches and QSPR models. The present state of CAMD is heavily reliant on fragment-based QSPR models for property predictions. This leads to inaccurate predictions when the generated structures have fragments that are not included in the training phase of the models. Models based on complete 3D information of molecules do not suffer from this problem and can be used to predict properties for unknown structures with reasonable accuracy. Further, a majority of the QSPR efforts in the literature are based on linear models, which can fail when a strong non-linear relationship exists between the target property and molecular structure. In this work, a novel non-linear QSPR modeling methodology was developed and applied to predict the various molecular properties for the CAMD algorithm.

9.4. Applications

9.4.1. Chemical Penetration Enhancers for Transdermal Delivery of Insulin: Traditional insulin delivery techniques, such as intravenous administration, are often associated with problems relating to over- and under-dosing, interactions with the harsh gastro-intestinal environment and/or the production of toxic by-products through metabolism in the liver. Recently, transdermal drug delivery (TDD) has gained popularity due to its ability to overcome most of the above problems with conventional delivery techniques.

Human skin is considered to be one of the most efficient natural polymers and serves as a barrier to the transport of chemicals both in and out of the human body [30, 31]. Each of the different layers of the skin offers a varying resistance to permeation [32, 33], and for large hydrophilic molecules like insulin, this resistance is significantly higher. Several physical and chemical alternatives are currently being investigated for possible improvement of TDD of insulin [34] and other drugs. However, the economic viability and technical feasibility of using chemicals as penetration enhancers (CPEs) makes them the most attractive option [35].

Problem Formulation: Only a few knowledge systems that discuss problem formulation for novel drug design exist. Lipinski’s ‘rule of 5’ is one such expert system that predicts the solubility and permeability of the drug molecules based on four target properties [36], which are the molecular weight, count of hydrogen bond donors, count of hydrogen bond acceptors and octanol/water partition coefficient (log K_{ow}). Since our target is the identification of novel potential CPEs, extensive knowledge of the properties of the CPEs and their corresponding functionalities is needed. The target molecules should be able to enhance the permeation of a selected drug through the skin without causing any harmful side-effects. After thorough analysis of the currently available CPEs and their properties, Golla et al. [37] [38] have identified the following property constraints as significant for transdermal drug delivery. This is a subjective list based on knowledge acquired from the open literature and our previous experience with CPEs:

1. Molecular weight: Molecules with low molecular weights easily penetrate the skin due to their small size. Hence, an upper limit of 500 was imposed on the molecular weight of potential CPEs [36, 39-41].
2. Octanol/water partition coefficient (K_{ow}): Drugs with very low or high partition coefficient fail to reach systemic circulation [36, 40, 41]. Several ranges of log K_{ow} values have been proposed in the literature for effective permeation enhancement. In this work, molecules with log K_{ow} values in the range of 1-3 were accepted and considered to indicate good permeation enhancement [39].
3. Melting point: Molecules with high melting points, due to their low solubility both in water and fat, are ineffective in transdermal drug delivery (TDD) [40], and only molecules with melting points less than 200°C were considered as good CPEs [39].
4. Skin sensitization: The CPE should not cause any skin irritation or sensitization upon application [39]. All the newly generated molecules are scored using three

independent skin sensitization QSPR models, based on the Federal Institute for Health Protection of Consumers and Veterinary Medicine (BgVV) database, the guinea pig maximization test (GPMT) database and the local lymph node assay (LLNA) database.

5. Number of hydrogen donor groups: The sum of the hydrogen atoms linked to oxygen and nitrogen atoms in the molecule determines the total number of hydrogen-bond donor groups in a molecule. The permeability across the lipid bi-layer has been identified to be significantly lower for drugs with an excessive number of these groups [36, 39]. Hence, a hydrogen-bond donor number upper limit of five was specified for acceptance of a molecule as a CPE.
6. Number of hydrogen acceptor groups: The total number of nitrogen, oxygen and fluorine atoms in the molecule (excluding nitrogen atoms with a formal positive charge, higher oxidation states and pyrrolyl forms) determines the total number of hydrogen-bond acceptor groups in a molecule. Presence of too many acceptor groups has been identified as a hindrance to the permeability across the lipid bi-layer [36]; therefore, an upper limit of 10 was used for the hydrogen-bond acceptor number.

In addition to the above constraints, two more constraints are imposed on the current design algorithm, based on experimental measurements of the reduction in skin resistance and the enhancement of insulin permeation in the presence of more than 100 different compounds:

7. Combined number of hydrogen donors and acceptors: All the compounds that have been proven experimentally to enhance the permeation of insulin had at least one hydrogen-bond donor or acceptor group. The hydrogen bonding capacity of some

compounds is known to temporarily disrupt the structure of the skin and thereby enhance the permeation of the drug molecules [42, 43]. Therefore, the potential CPEs are constrained to have a minimum of one hydrogen donor or acceptor.

8. Permeability coefficient of the CPE: The CPEs that enhanced insulin permeation had a permeability coefficient (Log K_p) of greater than -2.5. This suggests that the higher permeability allows the CPEs to permeate into the inner layers of the skin easily, which is essential for the CPEs to ultimately disrupt the internal structure of the skin through hydrogen bonding. Therefore, only those CPEs with a permeability coefficient (Log K_p) greater than -2.5 were preferred.

In addition to the above constraints, the potential CPE has to be a stable molecule at room temperature and atmospheric pressure; thus, the following constraint was added to the design algorithm to account for thermodynamic stability, where Gibbs energy of formation was included to quantify the stability of the designed molecules:

9. Standard Gibbs free energy of formation: The standard Gibbs free energy of formation for any molecule has to be lower than zero for stability at room temperature and pressure. The lower the Gibbs free energy, the more stable the molecule is relative to its elements. Therefore, only compounds with negative Gibbs free energy in reference to their elements were preferred in the algorithm.

Initial Population: In earlier CPE design work, Golla et al. [37, 38] compiled an Oklahoma State University (OSU) CPE database composed of over 400 CPE molecules from diverse chemical classes such as fatty alcohols, fatty acids, fatty acid esters, fatty alcohol ethers, alkanones, sulfoxides, biologics, enzymes, amines, amides, complexing agents, macrocyclics, classical surfactants, pyrrolidones, ionic compounds, solvents and azone-related compounds. One hundred CPEs from this list were randomly selected and used as the initial population in the current design algorithm.

Property Constraints and Fitness Function: The property constraints that need to be satisfied for a compound to be an insulin CPE were described previously. Table 9.7 lists the property constraints and the fitness function weights used in the CPE design algorithm, along with the mean average error (MAE) associated with the QSPR models for property prediction when applicable. Some of these properties were calculated using DRAGON [44] software, while other properties were estimated using QSPR models developed by the Molecular Design Group at OSU. The constraints were adjusted to account for the model prediction uncertainties when applicable. The fitness contribution of each constraint is calculated using Equation 9.5 and varies between 0 and 0.5, where a value of 0 implies that the constraint has been satisfied. This fitness contribution of each constraint is then multiplied by the particular fitness weight for that constraint, and the resulting values for all constraints are summed to give the total fitness function value, as shown in Equation 9.6. For the CPE design case, a penalty of magnitude 5 was further added to the total fitness function if the molecule has no hydrogen-bond donors or acceptors. Therefore, the total fitness function was now modified as follows:

$$\mathrm{F_{total}} = \begin{cases} \displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \emptyset_{\mathrm{i}}\mathrm{F_i} \ \text{ if } (\mathrm{nHDon} + \mathrm{nHAcc}) > 0 \\ 5 + \displaystyle\sum_{\mathrm{i}=1}^{\mathrm{n}} \emptyset_{\mathrm{i}}\mathrm{F_i} \ \text{ if } (\mathrm{nHDon} + \mathrm{nHAcc}) = 0 \end{cases} \tag{9.7}$$

where, $\square_i$ is the weight term associated with the i^{th} constraint, F_i is the fitness contribution of the i^{th} constraint, n is the total number of constraints in the problem and *nHDon* and *nHAcc* denote the number of hydrogen-bond donors and acceptors in the molecule, respectively.

Results: A total of 62 iterations of the design algorithm were completed, during which 6,200 molecules were generated. The molecular properties listed in Table 9.7 were computed for these molecules, and only 627 of the original 6,200 molecules satisfied the constraints listed in Table 9.7. A self-organizing map was developed to identify clusters among the best 627 molecules based on functional group descriptors calculated using DRAGON [45]. Five major clusters composed of at least 15 molecules were identified. Table 9.8 lists the structure of an example compound from each cluster accompanied with the relevant molecular properties.

The results from the current CAMD approach for designing CPEs can be compared with the results obtained by Golla et al. [38]. Specifically, the predicted K_p values of the majority of the CPEs identified in this work are comparable to the compounds that were identified by Golla et al. [38] and Godavarthy et al. [14], and were experimentally tested at OSU to be good enhancers. However, the current CAMD approach has several advantages over the methodology adopted by Golla et al. [38] and Godavarthy et al. [14]. First, the QSPR models employed in the current approach are more accurate, and second, the entire CAMD algorithm has been automated to minimize human intervention, and therefore the implementation of a large number of generations was possible. The design approaches by Golla et al. [38] and Godavarthy et al. [14] were not automated and therefore, the execution of the algorithm was limited to less than ten generations. Also, the inclusion of the Gibbs free energy of formation models in the current design algorithm ensures that only stable molecules are identified by the algorithm. Such stability models were not included in the previous work by our group [14, 38].

9.4.2. Solvents for Extractive Distillation of Cyclohexane and Benzene: In conventional distillation, chemical mixtures are separated into constituent components to yield products with greater commercial value. However, mixtures frequently contain molecular species that are similar in their physical properties and behavior ("close-boiling" mixtures), which makes their separation by conventional distillation extremely difficult and cost-prohibitive. One well-established method to deal with such situations is to use extractive distillation (ED), where an additional component (or components) is introduced to alter the behavior of the mixture in such a way that the original components become easier to separate. The ability of a given component or solvent to improve the separability of the components in the original mixture depends on the molecular interactions between the original species and the solvent added. From a process view point, the technical and economic feasibility of ED is, to a large degree, decided by the solvent used. In addition, some of the economic benefits and motivations for designing new solvents are as follows, where the statistics are projections from Phillips Petroleum [46, 47]:

1. A successful new solvent can provide multi-million dollars of annual sales
2. Efficiencies of current processes can be increased
3. Capital costs for new processes can be decreased
4. Recovery of specialty chemicals is profitable (~$40/gallon for these chemicals compared to ~$1.98/gallon for gasoline)

5. The replacement of a currently used solvent in an existing petrochemical process with an improved solvent could result in substantial operating cost savings.

In this work, the solvent design methodology is exemplified using the cyclohexane/benzene system. The methodology, after some minor changes, can be extended to the design of solvents for any system.

Problem Formulation: The design of solvents for extractive distillation involves consideration of various properties, among which three are of major significance [48]:

1. Selectivity: The manner in which an extractive solvent affects the separability of close-boiling substances may be explained in terms of its relative volatility. The relative volatility, α_{ij} of a mixture represents a measure of the ease with which two chemicals (species i and j, species i being the more volatile species) may be separated:

$$\alpha_{ij} = \frac{(y/x)_i}{(y/x)_j} \tag{9.8}$$

   where, y and x are the mole fractions of the component in the vapor and liquid phases, respectively. The higher the value of α_{ij}, the easier the substances are to separate by distillation. The value of α_{ij} can be expressed in thermodynamic terms as follows:

$$\alpha_{ij} = \left(p_i^{\circ}/p_j^{\circ}\right)\left(\gamma_i/\gamma_j\right) \tag{9.9}$$

   where, p° is the vapor pressure and γ is the activity coefficient. For typical close-boiling species, both the vapor pressures and the activity coefficients of the two components are very similar, leading to a relative volatility near unity and, thus, a difficult separation. However, by introducing a suitable solvent, which has a higher affinity for one component, the activity coefficient ratio (γ_i / γ_j) can be changed significantly and separation of the components becomes easier. The ratio of the activity coefficients at infinite dilution (selectivity) of species i and j is given by

$$S_{ij} = {\gamma_i^\infty}/{\gamma_j^\infty} \tag{9.10}$$

Here, γ_i^∞ is the infinite-dilution activity coefficient (IDAC) of a species i, which is defined as

$$\gamma_i^\infty = \lim_{x_i \to 0} \gamma_i \tag{9.11}$$

Equation 9.11 may be written in a similar manner for species j. A higher selectivity leads to a larger relative volatility, a smaller reflux ratio and lower capital costs in a distillation column [49]. For these reasons, the solvent with the highest selectivity is always considered the most promising candidate for a given separation process [50]. In this work, a lower limit of 4 was imposed on the selectivity, and molecules with selectivity lower than this value were considered unfit.

2. Normal boiling point: The normal boiling point (NBP) of the solvent must be significantly higher than the mixture components to avoid possible formation of a solute-solvent azeotrope and to ensure easy recovery of the solvent. A minimum difference of 25-50 K is usually desired [11]. Therefore, a solvent for the cyclohexane

(NBP = 354 K)/benzene (NBP=353 K) system must have a NBP value of at least 380 K (107°C).

3. Melting point: Melting point (MP) of the solvent is significant in order to avoid any crystallization problems at ambient temperature. The general tendency in industry has been to use solvents that are liquids at room temperature. Therefore, the potential solvents must have a MP value lower than 300 K (27°C).

In addition to the above constraints, a potential solvent has to be a stable molecule at room temperature and atmospheric pressure. The following constraint was added to the design algorithm to account for thermodynamic stability:

4. Standard Gibbs free energy of formation: The standard Gibbs free energy of formation for any molecule has to be lower than zero for it to be stable at room temperature and pressure. The lower the Gibbs free energy, the more stable the molecule is relative to its elements.

Further, to avoid the identification of complex molecules that might be difficult to synthesize, a limit is placed on the maximum molecular weight of the solvent.

5. Molecular weight: Solvent molecules with molecular weight lower than 150 g/mol are preferred. This number was chosen after carefully analyzing the reported solvent molecules in the literature.

Other considerations such as cost, safety, availability and environmental toxicity of candidate solvents are also important, but they have not been considered in this work. As an alternative, these constraints could be imposed on the final population of the best structures identified by the design algorithm, to further narrow the number of potential solvents.

Initial Population: Solvent molecules for which the γ^{∞} values of cyclohexane or benzene are available from the DECHEMA chemistry data series were extracted. A hundred molecules from this extracted database were selected randomly and used as the initial population for the solvent design algorithm. The initial database was made up of diverse chemical classes such as fatty alcohols, fatty acids, fatty acid esters, fatty alcohol ethers, alkanones, sulfoxides, amines, amides, pyrrolidones, pyridines, classical surfactants, chlorides, bromides, nitriles, and azone-related compounds.

Property Constraints and Fitness Function: The property constraints selected for the design of a suitable cyclohexane/benzene solvent were described previously. Table 9.9 lists the property constraints and the fitness function weights used in the solvent design algorithm accompanied with the mean average error (MAE) associated with the QSPR models for property prediction when applicable. Some of these properties were calculated using DRAGON [44] software, while other properties were estimated using QSPR models developed by the Molecular Design Group at OSU. The constraints were adjusted to account for the model prediction uncertainties when applicable. The fitness contribution of each constraint is calculated using Equation 9.5, and varies between 0 and 0.5, where a value of 0 implies that the constraint has been satisfied. The fitness contribution of each constraint is then multiplied by the particular fitness weight for that constraint, and the resulting values for all constraints are summed to provide the total fitness function value as shown in Equation 9.6. All constraints are in the form of Equation 9.5, except for the selectivity constraint, which is in the form of Equation 9.3. Therefore, molecules that have the highest selectivity values are preferred over other molecules that have slightly lower selectivity values but still meet the selectivity constraint. This sometimes leads to the generation of molecules that have very high selectivity values but are unstable (positive Gibbs energy of formation values). To avoid this, a higher weight was given to the Gibbs energy of formation in the solvent design case in comparison to the CPE design case.

Results: A total of 63 iterations of the design algorithm were completed, during which 6,300 molecules were generated. The molecular properties listed in Table 9.9 were computed for these molecules, and only 407 of the original 6,300 molecules satisfied all the constraints listed in Table 9.9. A self-organizing map was developed to identify clusters among the best 407 molecules based on functional group descriptors calculated using DRAGON [45]. Five major clusters composed of at least 40 molecules were identified, and Table 9.10 lists the structure of an example compound from each cluster along with the relevant molecular properties.

The results from the current CAMD approach for designing solvents can be compared with previous results obtained by our group. In his dissertation work, Godavarthy [6] limited his design methodology to search only for nitrogen- and sulphur-containing compounds, based on experimental knowledge. In the current CAMD methodology, such restrictions were not placed; nevertheless, the majority of the potential solvents that were identified are nitrogen- and sulphur-containing compounds. This proves the ability of the current CAMD approach to identify the best solvents, starting from random chemical structures. Also, the predicted selectivity values of the majority of the solvents identified in this work are 2 to 4 times better than the best solvents identified earlier [6]. Further, the current CAMD approach has several additional advantages. First, the QSPR models employed in the current work are more accurate and were developed using larger data sets. Second, the entire CAMD algorithm has been automated to minimize human intervention, and therefore the implementation of a large number of generations was possible. In fact, the current automated approach allowed for significantly greater number of generations during the execution of the algorithm. Third, the inclusion of the Gibbs free energy of formation models in the current design algorithm ensures that only stable molecules are identified by the algorithm. Such stability models were not included n the earlier work [6].

9.4.3. Additional Selection Criteria: At the end of the design algorithm, hundreds of potential candidate molecules that have similar fitness function values are identified. However, validating all these molecules experimentally is impractical. Instead, additional criteria such as ease of synthesis, cost of manufacturing, safety, and toxicity should be imposed on the initial list of potential candidate molecules to select the best candidates for immediate attention.

9.4.4. Experimental Validation: As a final validation, the best candidate molecules should be experimentally tested for their efficacy. The potential CPEs must be tested *in vitro* for reduction in skin resistance [51] and enhancement of insulin permeation [52]. Also, the toxicity potential of the CPEs must be experimentally determined *in vitro*. The CPEs that perform well in the *in vitro* experiments must then be put through *in vivo* experimentation using mouse/rat models.

Similarly, the best candidate solvent molecules are experimentally validated. Specifically, infinite-dilution activity coefficient measurements must be conducted to validate the selectivity of the solvent molecules. This should be followed up by process-simulation studies to estimate the cost and other process parameters associated with the utilization of the solvent molecules in the separation process. The final steps of validation must include lab-scale and pilot scale studies of the envisioned separation process.

9.5. Conclusions

1. A robust algorithm combining genetic algorithms and QSPR techniques was developed for the design of novel molecules with desired properties.
2. The current algorithm is the only completely automated design tool in the literature that is based on accurate 3-dimensional structure-property relationship models. The algorithm was applied to two separate case studies: identification of new CPEs for enhancing insulin transdermal delivery and identification of solvents for the extractive distillation of cyclohexane /benzene mixtures.
3. A total of 627 molecules that meet all the specifications of a good insulin CPE have been identified. The identified molecules are categorized into five different clusters based on their functional groups.
4. A total of 407 molecules that meet all the specifications of a good cyclohexane/benzene solvent have been identified. The identified molecules are categorized into five different clusters based on their functional groups.
5. Further, the algorithm in this work is generalized and so could be adapted to any design problem, where there exists a need for new molecules.

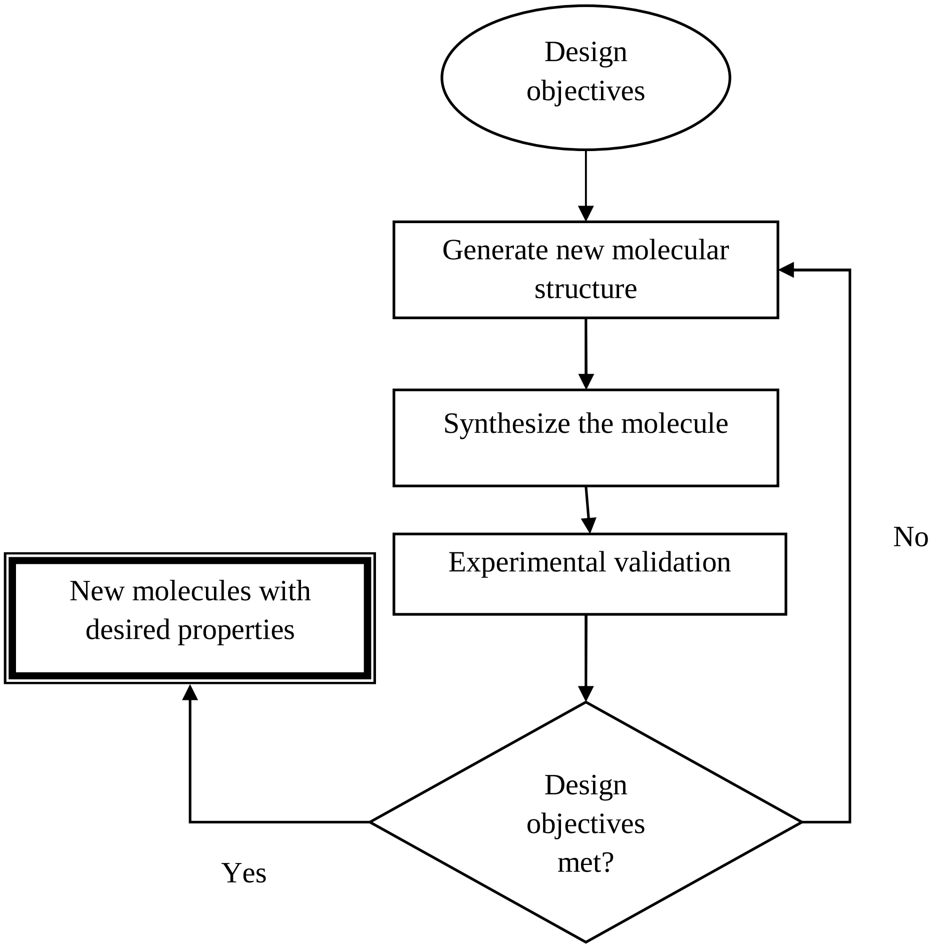


Figure 9.1: The various stages in CAMD

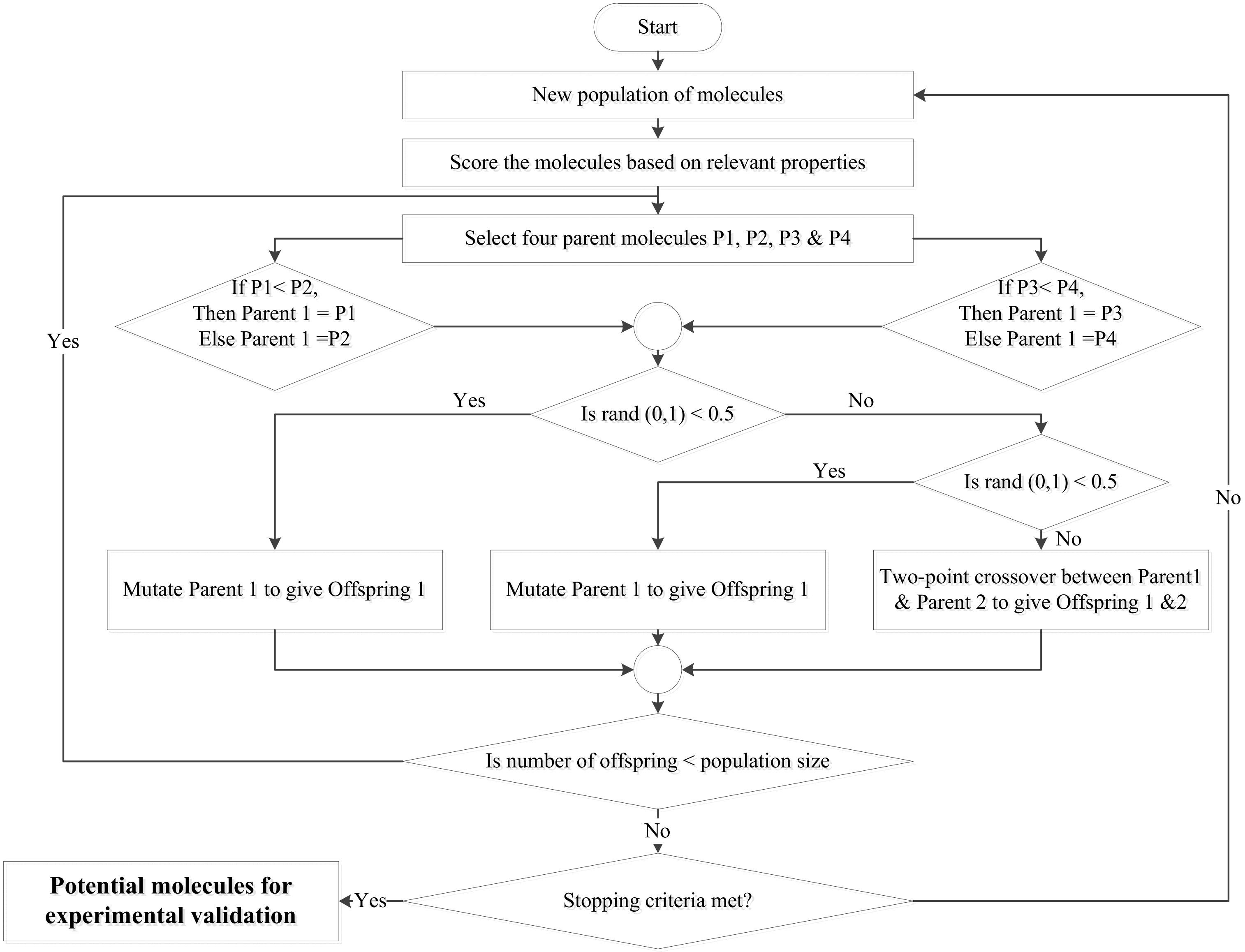

Start
New population of molecules
Score the molecules based on relevant properties
Select four parent molecules P1, P2, P3 & P4
If P1< P2,
Then Parent 1 = P1
Else Parent 1 =P2
If P3< P4,
Then Parent 1 = P3
Else Parent 1 =P4
Yes
Is rand (0,1) < 0.5
No
Yes
Is rand (0,1) < 0.5
No
Mutate Parent 1 to give Offspring 1
Mutate Parent 1 to give Offspring 1
Two-point crossover between Parent1 & Parent 2 to give Offspring 1 &2
Yes
Is number of offspring < population size
No
Potential molecules for experimental validation
Yes
Stopping criteria met?
No

Figure 9.2: Flow-diagram for the design algorithm used in this work

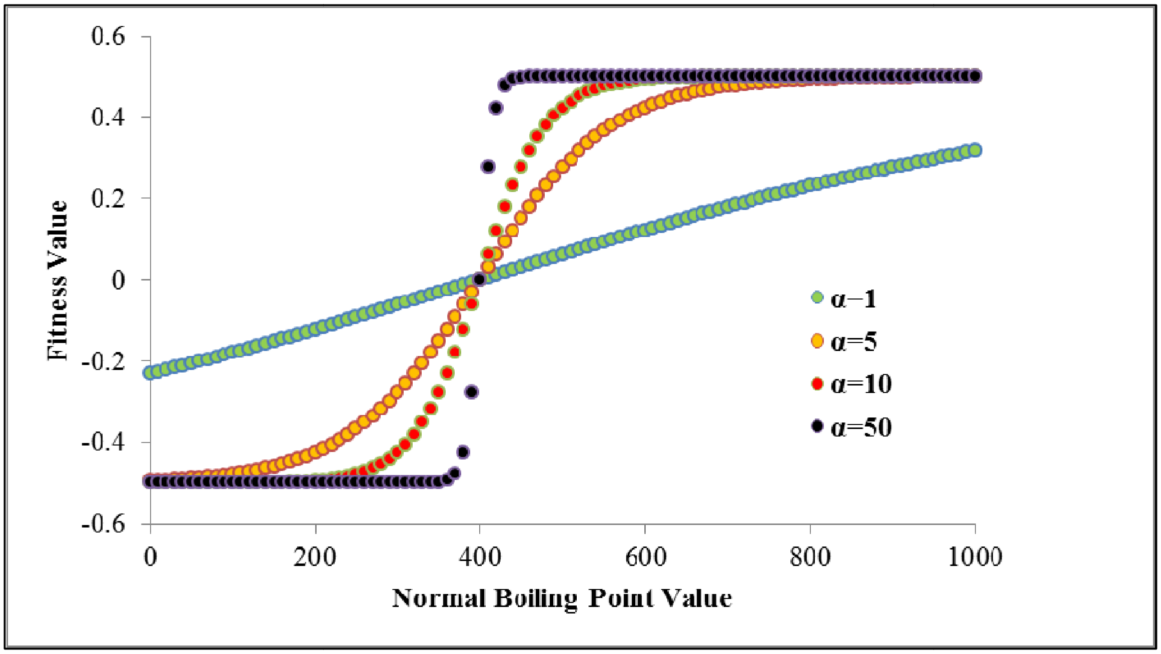


Figure 9.3: The influence of fitness decay on the fitness values calculated using Equation 9.3

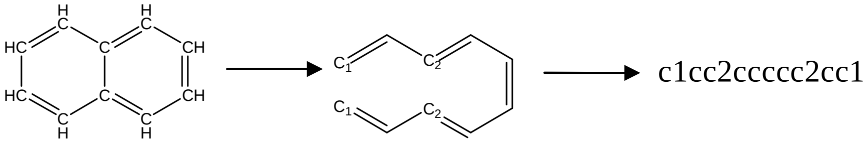


Figure 9.4: SMILES string generation for naphthalene

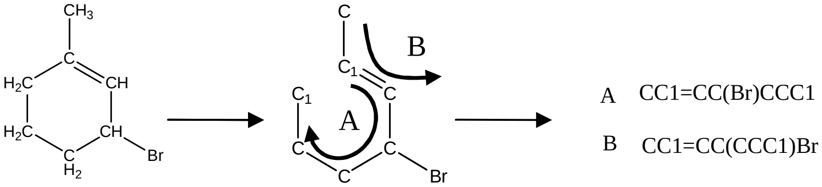


Figure 9.5: SMILES string generation for 1-methyl-3-bromo-cyclohexene

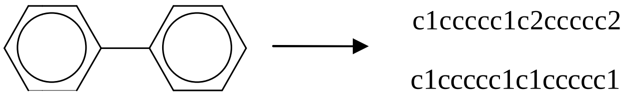


Figure 9.6: SMILES string generation for biphenyl

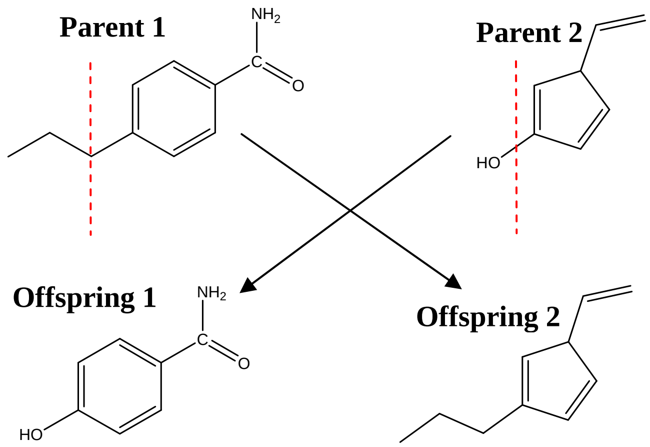


Figure 9.7: One-point crossover operation

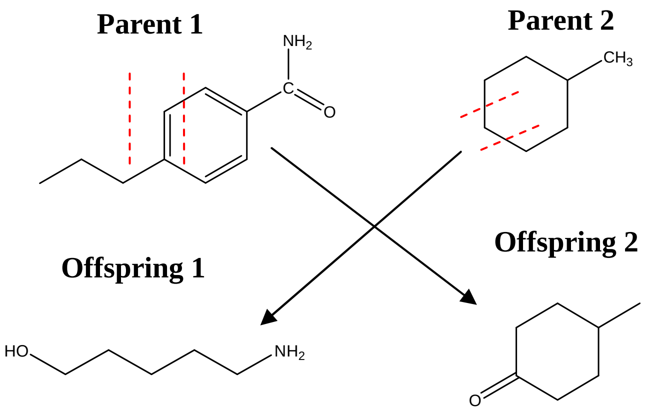


Figure 9.8: Two-point crossover operation

Table 9.1: Atomic specifications for SMILES

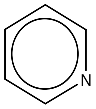
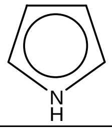
Chemical Name	SMILES	Structure
Methane	C	CH_4
Pyridine	n1ccccc1	
Pyrrole	c1c[nH]cc1	

Table 9.2: Bond specifications for SMILES

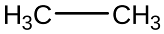
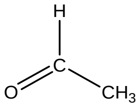
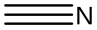
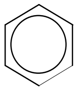
Chemical Name	SMILES	Structure
Ethane	CC	
Acetaldehyde	CC=O	
Hydrogen Cyanide	C#N	
Benzene	c1:c:c:c:c:c1	

Table 9.3: Branching specifications for SMILES

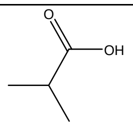
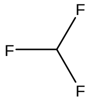
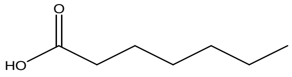
Chemical Name	SMILES	Structure
Iso Butyric Acid	CC(C)C(=O)O	
Fluroform	C(F)(F)F	
Heptanoic Acid	CCCC(C(=O)O)CCC	

Table 9.4: The different mutation operations used in this work

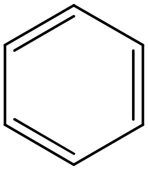
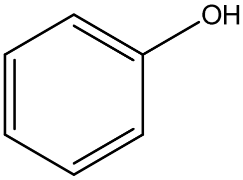
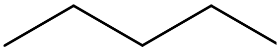
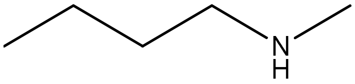
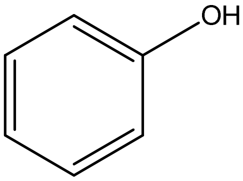
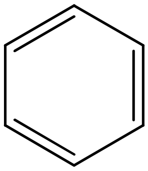
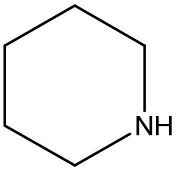
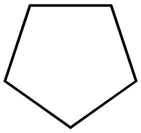
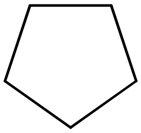
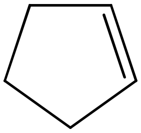
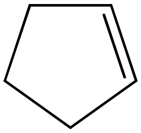
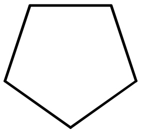
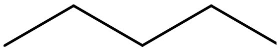
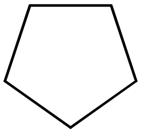
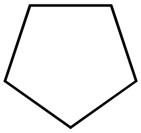
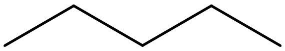
Mutation Operator	Initial Structure	Final Structure	Initial SMILES	Final SMILES
Add Atom			c1ccccc1	c1c(O)cccc1
Insert Atom			CCCCC	CCCCNC
Delete Atom			c1c(O)cccc1	c1ccccc1
Uninsert Atom			C1CCNCC1	C1CCCC1
Increase Bond-Order			C1CCCC1	C1CCC=C1
Decrease Bond-Order			C1CCC=C1	C1CCCC1
Create Ring			CCCCC	C1CCCC1
Break Ring			C1CCCC1	CCCCC

Table 9.5: The different atoms that can be added to a molecule, with the probability of selection

Atom	Probability of Selection
B	0.01
Br	0.04
C	0.36
Cl	0.05
N	0.15
O	0.20
P	0.075
S	0.075
F	0.04
I	0.01

Table 9.6: The different atoms that can be inserted in a molecule, with the probability of selection

Atom	Probability of Selection
B	0.01
C	0.39
N	0.2
O	0.2
P	0.1
S	0.1

Table 9.7: The different property constraints and fitness weights used in the CPE design algorithm and, when applicable, the mean average error (MAE) values for QSPR models used to predict the property values

Property	Constraint	Calculated using DRAGON	MAE	Fitness Weight
Molecular weight (MW)	MW < 500 g/mol	Yes	N/A	10
Octanol-water partition coeff. (Log K_{ow})	0.5 < Log K_{ow} < 3.5	Yes	0.5	10
Melting Point (MP)	MP < 250°C	No	34°C	15
Federal Institute for Health Protection of Consumers and Veterinary Medicine (BgVV)	BgVV < 0.5	No	0.45	10
Guinea pig maximization test (GPMT)	GPMT < 0.33	No	0.30	10
Local lymph node assay (LLNA)	LLNA < 0.25	No	0.25	10
Number of hydrogen donors (nHDon)	nHDon < 5	Yes	N/A	5
Number of hydrogen acceptors (nHAcc)	nHAcc < 10	Yes	N/A	5
Skin permeability coefficient (Log K_p)	Log Kp > -3	No	0.5	10
Gibbs free energy of formation (ΔG_f)	ΔG_f < -20 kJ/mol	No	16 kJ/mol	15

Table 9.8: The properties of an example CPE from each cluster

Cluster #	Log (K_P)	BgVV	GPMT	LLNA	Log K_{ow}	MW (g/mol)	nHDon	nHAcc	ΔG_f (KJ/mol)	MP (°C)
1	-1.4	0.4	0.2	0.2	3.4	144	1	1	-115	-67
2	-1.5	0.4	0.1	0.2	2.4	148	1	2	-307	-20
3	-1.5	0.4	0.2	0.2	3.2	158	1	2	-312	-18
4	-1.7	0.5	0.2	0.2	3.0	156	1	1	-42	-6.0
5	-1.8	0.3	0.2	0.2	3.2	158	1	1	-118	-8.0

Table 9.9: The different property constraints and fitness weights used in the solvent design algorithm, and when applicable, the mean average error (MAE) values for QSPR models used to predict the property values

Property	Constraint	Calculated using DRAGON	MAE	Fitness Weight
Selectivity (S)	S > 5	No	0.9	25
Normal boiling point (NBP)	NBP > 135°C	No	28°C	15
Melting point (MP)	MP < -14°C	No	34°C	15
Gibbs free energy of formation (ΔG_f)	ΔG_f < -20 kJ/mol	No	16 kJ/mol	35
Molecular weight (MW)	MW < 150 g/mol	Yes	N/A	10

Table 9.10: The properties of an example solvent from each cluster

Cluster #	Selectivity	NBP (°C)	MP (°C)	ΔG_f (KJ/mol)	MW (g/mol)
1	38	235	-31	-292	137.22
2	26	271	-21	-317	112.01
3	15	245	-27	-376	148.25
4	24	159	-30	-233	76.05
5	46	200	-41	-291	137.13

REFERENCES

1. Venkatasubramanian, V., K. Chan, and J.M. Caruthers, *Computer-Aided Molecular Design Using Genetic Algorithms.* Computers and Chemical Engineering, 1994. **18**(9): p. 833–844.

2. Sundaram, A. and V. Venkatasubramanian, *Parametric Sensitivity and Search-Space Characterization Studies of Genetic Algorithms for Computer-Aided Polymer Design.* Journal of Chemical Information and Computer Science, 1998. **38**(6): p. 1177-1191.

3. Devillers, J., *Genetic Algorithms in Molecular Modeling*. 1996: Academic press.

4. Harper, P.M., et al., *Computer-Aided Molecular Design with Combined Molecular Modeling and Group Contribution.* Fluid Phase Equilibria, 1999. **158**: p. 337-347.

5. Venkatasubramanian, V., K. Chan, and J.M. Caruthers, *Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm.* Journal of Chemical Information and Computer Science, 1995. **35**: p. 188-195.

6. Godavarthy, S.S., *Design of Improved Solvents for Extractive Distillation.* Ph.D. Dissertation, School of Chemical Engineering. 2004, Oklahoma State University: Stillwater, Oklahoma.

7. Li, J., *CAMD in Modern Drug Discovery.* Drug Discovery Today, 1996. **1**(8): p. 311-312.

8. Nachbar, R.B., *Molecular Evolution: A Hierarchical Representation for Chemical Topology and Its Automated Manipulation.* Proc. of the Third Annual Genetic Programming Conference, 1998: p. 246-253.

9. Leardi, R., *Genetic Algorithms in Chemometrics and Chemistry: A Review.* Journal of Chemometrics, 2001. **15**(7): p. 559-569.

10. Brignole, E.A., S. Bottini, and R. Gani, *A Strategy for the Design and Selection of Solvents for Separation Processes.* Fluid Phase Equilibria, 1986. **29**: p. 125-132.

11. Pretel, E.J., et al., *Computer-Aided Molecular Design of Solvents for Separation Processes.* AIChE Journal, 1994. **40**(8): p. 1349-1360.

12. Churi, N. and L.E.K. Achenie, *Novel Mathematical Programming Model for Computer Aided Molecular Design.* Industrial & Engineering Chemistry Research, 1996. **35**(10): p. 3788-3794.

13. Ostrovsky, G.M., L.E.K. Achenie, and M. Sinha, *On the Solution of Mixed-Integer Nonlinear Programming Models for Computer Aided Molecular Design.* Computers & Chemistry, 2002. **26**(6): p. 645-660.

14. Godavarthy, S.S., et al., *Design of Improved Permeation Enhancers for Transdermal Drug Delivery.* Journal of Pharmaceutical Sciences, 2010. **99**(1): p. 563-563.

15. Song, J. and H.H. Song, *Computer-Aided Molecular Design of Environmentally Friendly Solvents for Separation Processes.* Chemical Engineering & Technology, 2008. **31**(2): p. 177-187.

16. Brown, W.M., et al., *Designing Novel Polymers with Targeted Properties Using the Signature Molecular Descriptor.* Journal of Chemical Information and Modeling, 2006. **46**(2): p. 826-835.

17. Wong, W. and F. Burkowski, *A Constructive Approach for Discovering New Drug Leads: Using a Kernel Methodology for the Inverse-QSAR Problem.* Journal of Cheminformatics, 2009. **1**(1): p. 4.

18. Miyao, T., M. Arakawa, and K. Funatsu, *Exhaustive Structure Generation for Inverse-QSPR/QSAR.* Molecular Informatics, 2010. **29**(1-2): p. 111-125.

19. Korichi, M., et al., *Computer-Aided Aroma Design. II. Quantitative Structure-Odour Relationship.* Chemical Engineering and Processing: Process Intensification, 2008. **47**(11): p. 1912-1925.

20. Darwin, C., *The Origin of Species.* Br Med J, 1958. **1**(5086): p. 1527-8.

21. Achenie, L.E.K., R. Gani, and V. Venkatasubramanian, *Computer Aided Molecular Design: Theory and Practice*. 2003: Elsevier.

22. Douguet, D., E. Thoreau, and G. Grassy, *A Genetic Algorithm for the Automated Generation of Small Organic Molecules: Drug Design Using an Evolutionary Algorithm.* Journal of Computer-Aided Molecular Design, 2000. **14**(5): p. 449-466.

23. Venkatasubramanian, V., K. Chan, and J. Caruthers, *Evolutionary Design of Molecules with Desired Properties Using the Genetic Algorithm.* J Chem Inf Comput Sci, 1995. **35**: p. 188 - 195.

24. Weininger, D., *SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules.* Journal of Chemical Information and Computer Sciences, 1988. **28**(1): p. 31-36.

25. Lameijer, E.-W., et al., *The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules.* Journal of Chemical Information and Modeling, 2006. **46**(2): p. 545-552.

26. Miller, B.L. and D.E. Goldberg, *Genetic Algorithms, Selection Schemes, and the Varying Effects of Noise.* Evolutionary Computation, 1996. **4**(2): p. 113-131.

27. Holland, J.H., *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*. 1992: MIT Press.

28. Baker, J.E., *Reducing Bias and Inefficiency in the Selection Algorithm*, in *Proceedings of the Second International Conference on Genetic Algorithms on Genetic algorithms and their application*. 1987, L. Erlbaum Associates Inc.: Cambridge, Massachusetts, United States. p. 14-21.

29. Goldberg, D.E., *Zen and the Art of Genetic Algorithms.* International Conference on {G}enetic {A}lgorithms '89, 1989: p. 80-85.

30. Monteiro-Riviere, N.A., *A Anatomical Factors Affecting Barrier Function.* Marzulli and Maibach's Dermatotoxicology, 2008.

31. Monteiro-Riviere, N.A., *Comparative Anatomy, Physiology, and Biochemistry of Mammalian Skin.* Dermal and Ocular Toxicology: Fundamentals and Methods, 1991: p. 3–71.

32. Elias, P.M., *Epidermal Lipids, Barrier Function, and Desquamation.* Journal of Investigative Dermatology, 1983. **80**: p. 44-49.

33. Chang, S.K. and J.E. Riviere, *Percutaneous Absorption of Parathion in Vitro in Porcine Skin: Effects of Dose, Temperature, Humidity, and Perfusate Composition on Absorptive Flux.* Toxicological Sciences, 1991. **17**(3): p. 494-504.

34. Rao, V.U. and A.N. Misra, *Enhancement of Iontophoretic Permeation of Insulin across Human Cadaver Skin.* Pharmazie, 1994. **49**(7): p. 538-9.

35. Guy, R.H., *Current Status and Future Prospects of Transdermal Drug Delivery.* Pharmaceutical Research, 1996. **13**(12): p. 1765-1769.

36. Lipinski, C.A., et al., *Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings.* Adv Drug Deliv Rev, 2001. **46**(1-3): p. 3-26.

37. Golla, S., *Virtual Design of Chemical Penetration Enhancers*. M.S. Vol. Oklahoma State University. 2008, Oklahoma State University: United States -- Oklahoma.

38. Golla, S., et al., *Virtual Design of Chemical Penetration Enhancers for Transdermal Drug Delivery.* Chemical Biology & Drug Design, 2011 (Accepted).

39. Finnin, B.C. and T.M. Morgan, *Transdermal Penetration Enhancers: Applications, Limitations, and Potential.* Journal of Pharmaceutical Sciences, 1999. **88**(10): p. 955-958.

40. Kumar, R. and A. Philip, *Modified Transdermal Technologies: Breaking the Barriers of Drug Permeation Via the Skin.* Tropical Journal of Pharmaceutical Research, 2007. **6**(1): p. 633-644.

41. Brown, L. and R. Langer, *Transdermal Delivery of Drugs.* Annual Review of Medicine, 1988. **39**(1): p. 221-229.

42. Barry, B.W., *Mode of Action of Penetration Enhancers in Human Skin.* Journal of Controlled Release, 1987. **6**(1): p. 85-97.

43. Hadgraft, J., et al., *Mechanisms of Action of Skin Penetration Enhancers/Retarders: Azone and Analogues.* International Journal of Pharmaceutics, 1996. **141**(1-2): p. 17-25.

44. SRL, T. *Dragon for Windows and Linux*. [cited 2010; Available from: http://www.talete.mi.it/help/dragon_help/index.html.

45. *Dragon Professional 6*. 2010, Talete SRL.

46. Gasem, K.A., et al., *Annual Progress Report: Improved Solvents for Extractive Distillation: Infinite-Dilution Activity Coefficient Measurements*. 1999, Phillips Petroleum Company and the State of Oklahoma.

47. Schult, C.J., *Design of Solvents for Extractive Distillation*. PhD. Thesis. 2000, Oklahoma State University, Stillwater, Oklahoma.

48. van Dyk, B. and I. Nieuwoudt, *Design of Solvents for Extractive Distillation.* Industrial & Engineering Chemistry Research, 2000. **39**(5): p. 1423-1429.

49. Momoh, S.O., *Assessing the Accuracy of Selectivity as a Basis for Solvent Screening in Extractive Distillation Processes.* Separation Science and Technology, 1991. **26**(5): p. 729-742.

50. Chen, B., et al., *Application of CAMD in Separating Hydrocarbons by Extractive Distillation.* AIChE Journal, 2005. **51**(12): p. 3114-3121.

51. Rachakonda, V., et al., *Screening of Chemical Penetration Enhancers for Transdermal Drug Delivery Using Electrical Resistance of Skin.* Pharmaceutical Research, 2008. **25**(11): p. 2697-2704.

52. Yerramsetty, K.M., et al., *Effect of Different Enhancers on the Transdermal Permeation of Insulin Analog.* International Journal of Pharmaceutics, 2010. **398**(1-2): p. 83-92.

VITA

Krishna M. Yerramsetty

Candidate for the Degree of

Doctor of Philosophy

Thesis: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELING & COMPUTER-AIDED MOLECULAR DESIGN: IMPROVEMENTS & APPLICATIONS

Major Field: Chemical Engineering

Biographical:

Education:

Completed the requirements for Bachelor of Engineering (Honors) Degree in Chemical Engineering at Birla Institute of Technology and Science, Pilani, India in 2005.

Completed the requirements for Doctor of Philosophy in Chemical Engineering at Oklahoma State University, Stillwater, Oklahoma in May, 2012.

Experience:

Employed by the School of Chemical Engineering, Oklahoma State University as a graduate research assistant from January, 2007 to December, 2011

Professional Memberships:

American Institute of Chemical Engineers

Name: Krishna M. Yerramsetty Date of Degree: May, 2012

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: QUANTITATIVE STRUCTURE-PROPERTY RELATIONSHIP MODELING & COMPUTER-AIDED MOLECULAR DESIGN: IMPROVEMENTS & APPLICATIONS

Pages in Study: 290 Candidate for the Degree of Doctor of Philosophy

Major Field: Chemical Engineering

Scope and Method of Study: The objective of this work was to develop an integrated capability to design molecules with desired properties. An automated robust genetic algorithm (GA) module has been developed to facilitate the rapid design of new molecules. The generated molecules were scored for the relevant thermophysical properties using non-linear quantitative structure-property-relationship (QSPR) models. The descriptor reduction and model development for the QSPR models were implemented using evolutionary algorithms (EA) and artificial neural networks (ANNs). QSPR models for octanol-water partition coefficients (K_{ow}), melting points (MP), normal boiling points (NBP), Gibbs energy of formation, universal quasi-chemical (UNIQUAC) model parameters, and infinite-dilution activity coefficients of cyclohexane and benzene in various organic solvents were developed in this work. To validate the current design methodology, new chemical penetration enhancers (CPEs) for transdermal insulin delivery and new solvents for extractive distillation of the cyclohexane + benzene system were designed.

Findings and Conclusions: A robust general framework for designing new molecules and an improved framework for building accurate models for thermophysical properties have been developed. In general, the use of non-linear QSPR models developed in this work provided predictions better than or as good as existing literature models. In particular, the current models for NBP, Gibbs energy of formation, UNIQUAC model parameters, and infinite-dilution activity coefficients have lower errors on external test sets than the literature models. The current models for MP and K_{ow} are comparable with the best models in the literature. The GA-based design framework implemented in this work successfully identified new CPEs for transdermal delivery of insulin, with permeability values comparable to the best CPEs in the literature. Also, new solvents for extractive distillation of cyclohexane/benzene with selectivities two to four times that of the existing solvents were identified. These two case studies validate the ability of the current design framework to identify new molecules with desired target properties.

ADVISER'S APPROVAL: Dr. Khaled A. M. Gasem