OPTIMIZATION OF TRANSFERABLE MOLECULAR STEP POTENTIALS

by

Sinan Üçyiğitler

B. Sc. in Chemical Engineering, Boğaziçi University, 2003

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering
Boğaziçi University

2006

OPTIMIZATION OF TRANSFERABLE MOLECULAR STEP POTENTIALS

APPROVED BY:

Prof. Mehmet Cihan Çamurdan (Thesis Supervisor)	.....
Prof. Ahmet Erhan Aksoylu	.....
Prof. J. Richard Elliott	.....
Assoc. Prof. Metin Türkay	.....

DATE OF APPROVAL: 14.09.2006

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to Prof. Mehmet Çamurdan, Prof. J. Richard Elliott and Assoc. Prof. Metin Türkay for their guidance and help in preparation of this dissertation.

I need to mention Feyza Gökaliler, Murat Karadeniz and all members of CATREL for their help and understandings and I am also thankful to my family for their support and encouragements.

Financial support provided by Tübitak, NSF and Boğaziçi University through projects TBAG-U/99 104T097, OISE-0421849 and BAP-04HA501 is greatfully acknowledged.

ABSTRACT

OPTIMIZATION OF TRANSFERABLE MOLECULAR STEP POTENTIALS

To study thermodynamic properties for hydrocarbons and oxygenated compounds, discontinuous molecular dynamics and thermodynamic perturbation theory have been used. Accurate thermodynamic predictions require realistic and efficient transferable potential models for intermolecular interactions. Necessary parameters for these models can be obtained through reliable optimization methods. A stochastic search algorithm, namely Recursive Random Search, combined with a local search algorithm, namely Levenberg-Marquardt, is used together to optimize the parameters of 12 different potential models. These models are then compared using 5 statistical testing methods, which are cross-validation, Akaike's Information Criterion, Bayesian Information Criterion, Mallow's Information Criterion and F-Test. The optimum model is then selected to be the StepYukawa-Universal model (SYU) because of its predictive power on validation sets. Besides SYU, parameters for 3 additional models, Lennard Jones model (LJ), Yukawa-Universal model (YU) and Linear 2580 model, are also documented for their potential use and wide acceptance. The root mean squared percentage vapor pressure errors evaluated at reduced temperatures approximately between 0.480 and 0.900 for 102 compounds from 6 organic families, namely n-alkanes, branched alkanes, aromatics, naphthenics, alcohols and phenols, are 13.30, 13.23, 13.28 and 13.42 per cent for SYU, YU, LJ and Linear 2580 model, respectively. In addition to the accuracy in vapor pressure predictions, similar united atoms exhibit high similarities in terms of the parameters for molecular potentials which support the consistency and accuracy of the optimized parameters as well.

ÖZET

AKTARILABİLİR MOLEKÜLER POTANSİYELLERİN ENİYİLEMESİ

Hidrokarbonlar ve oksijenli bileşiklerin termodinamik özelliklerini incelemek için kesikli moleküler dinamiği ve termodinamik pertürbasyon teorisi kullanıldı. Hatasız termodinamik tahminler, moleküller arası etkileşim için gerçekçi ve etkili aktarılabilir potansiyel modelleri gerektirmektedir. Bu modeller için gerekli olan parametreler güvenilir eniyileme yöntemleri ile elde edildi. 12 farklı potansiyel modeli için gerekli olan parametreleri eniyilemek üzere, gelişigüzel arama yöntemi bölgesel arama yöntemlerinden Levenberg-Marquardt ile birleştirildi. Bu modeller daha sonra 5 farklı istatistiksel karşılaştırma yöntemiyle (PRESS, AIC, BIC, MIC ve F-Test) karşılaştırıldı. Sağlama kümesindeki tahmin gücü nedeniyle StepYukawa-Universal (SYU) en iyi model seçildi. SYU'nun yanı sıra, olası kullanılırlıkları ve yaygın bir şekilde kabul edilmiş olmaları nedeniyle Lennard Jones (LJ), Yukawa-Universal (YU) ve doğrusal 2580 modelleri de eniyilendi. SYU, YU, LJ ve doğrusal 2580 için, n-alkanlar, dallanmış alkanlar, aromatikler, naftenikler, alkoller ve fenollerden oluşan 102 bileşik için yaklaşık 0.480 ve 0.900 arasındaki indirgenmiş sıcaklıklarda elde edilmiş hata karelerinin ortalamalarının karekökleri sırasıyla yüzde 13.30, 13.23, 13.28 ve 13.42 olarak elde edildi. Buhar basıncı tahminlerindeki başarının yanı sıra, benzer birleşik atom gruplarının benzer moleküler potansiyel parametrelerine sahip olmaları , eniyileme ile elde edilen parametrelerin de tutarlılığını desteklemektedir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS/ABBREVIATIONS

A	Helmholtz Energy
a^{ref}	Helmholtz Departure
k_B	Boltzman Constant
N_{ijm}	number of interactions between i^{th} and j^{th} site in the m^{th} well
$P_{C,T}^{calc}$	Calculated Vapor Pressure of the compound C at temperature T
$P_{C,T}^{err}$	Error in Vapor Pressure of the compound C at temperature T
$P_{C,T}^{exp}$	Experimental Vapor Pressure of the compound C at temperature T
r	distance between sites
T	Temperature
u_{ijm}	Potential Energy between i^{th} and j^{th} site in the m^{th} well
u_{im}	Potential Energy of i^{th} site in the m^{th} well
X_{ij}	Trial Point
Z	Compressibility Factor
β	Reciprocal of the Boltzman Constant times Temperature
ϵ	Parameter to estimate well depths
γ	Dimensionless Distance in the Yukawa Functions
η	Packing Fraction
κ	Steepnes Factor in the Yukawa Functions
$\rho_{C,T}^{err}$	Error in Density of the compound C at temperature T
σ	distance where the potential is zero
BCONF	Quasi-Newton Algorithm
BFGS	Broyden-Fletcher-Goldfarb-Shanno
DMD	Discontinous Molecular Dynamics
LB_{ij}	Lower Boundary
LM	Levenberg-Marquardt Algorithm

nData	Number of Data Points
nFunCalls	number of Function Evaluations
nTrialsExploit	number of trials in the Explolitation Phase
nTrialsExplore	number of trials in the Exploration Phase
nVar	Number of Decision Variables
$pSSR$	Sum of Squares of Residuals in the Vapor Pressure
pAad	Absolute Average Deviation in Vapor Pressure
pBias	Average Deviation in Vapor Pressure
pMax	Maximum Deviation in Vapor Pressure
rAad	Absolute Average Deviation in Density
rand	Random Number between 0-1 from a uniform distribution
rBias	Average Deviation in Density
rMax	Maximum Deviation in Density
RMSE	Root Mean Squares of the Errors
rms(devEOK)	Root Mean Squares of deviations in the variables
RRS	Random Recursive Search
SSR	Sum of Squares of Residuals
TPT	Thermodynamic Perturbation Theory
UB_{ij}	Upper Boundary
$\%\ devP$	percentage deviation in the pSSR

1. INTRODUCTION

Estimating the physical properties like vapor pressure through Helmholtz energy necessitates molecular simulation for the compound of consideration. Instead of performing a traditional molecular dynamics (MD) simulation for the full potential, Cui and Elliott [1] showed that discontinuous molecular dynamics (DMD) can be used to simulate the repulsive part of the potential and attractive part can be incorporated to the system as perturbations. In that work it is concluded that second order thermodynamic perturbation theory (TPT) is sufficient to provide quantitative accuracy for the effect of the attractive potential on the physical properties. Owing to DMD/TPT formalism, a smooth interaction is established between simulation and optimization which eliminates repeating simulation phase after each iteration in the optimization calculations.

In this work, a database of hydrocarbons and oxygenated compounds from 6 organic families, namely n-alkanes, branched alkanes, aromatics, naphthenics, alcohols and phenols, is used to obtain a model to describe the attractive potential for 26 sites (united atoms) from which 102 different compounds are built. The algorithm for optimization is chosen to be a combination of a stochastic search algorithm, namely Random Recursive Search (RRS), and a gradient based local search algorithm, namely Levenberg-Marquardt (LM). Using this hybrid method, the attractive potential is fitted to 12 models which differ in the number of parameters to describe the potential. To compare the percentage errors in vapor pressures obtained from different models, and hence, to determine the optimum model with minimum number of parameters, 5 statistical tests are used. These are cross-validation (PRESS), Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallow's Information Criterion (MIC) and statistical F-Test. The optimum model according to these tests is the StepYukawa-Universal model. The parameters for this model and for Lennard-Jones, Yukawa-Universal and Linear-2580 models are documented. For these models, the root mean squared vapor pressure error (RMSE) is around 13 per cent.

2. LITERATURE SURVEY

Vapor pressure prediction using only the molecular structure is a very difficult challenge. A related approach would be through group contributions, but relatively few publications have appeared that attempt to predict vapor pressure. A typical example is the work of Fredenslund and coworkers [2, 3] based on UNIFAC groups. The UNIFAC effort, although claimed accuracy of 3 per cent for the training set, was restricted to pressures below 0.3MPa and the method was not validated beyond the training set. Recently, Asher *et al.* [4] adapted the UNIFAC model as the basis for a customized regression targeted at 76 compounds with the temperature range restricted to 290-320K. Despite these customization of dataset and the conditions, the validation set indicated deviations averaging near 300 per cent. Bureau *et al.* [5] performed a similar evaluation for seven esters and found deviations averaging near 80 per cent. An indirect approach to estimating vapor pressure with a group contribution method can be achieved with the Clausius-Clapeyron equation in combination with estimates of the normal boiling temperature and heat of vaporization. For example, such an approach could be applied with the Constantinou-Gani [6] or Joback [7] group contribution methods. As a specific example, Asher *et al.* [4] combined the Lee-Kesler [8] equation with the Joback method to predict the critical data and normal boiling temperature. They observed deviations averaging near 900 per cent for the same data set used in UNIFAC validation. Bureau *et al.* [5] also evaluated the Lee-Kesler method with several alternatives for estimating critical properties and found deviations near 60 per cent. One other group contribution approach is the method of Elliott and Natarajan [9] in which the parameters in equations of state are estimated based on group additivity. Once again, the method has not been validated as a means of estimating vapor pressure. Finally, molecular modeling has presented a new alternative that has been applied directly to predicting vapor pressure. Some examples are the works of Siepmann and coworkers, [10, 11] Fuchs *et al.* [12] and Ünlü *et al.* [13]. Siepmann and coworkers refer to their model as the TraPPE-UA (Transferable Potentials for Phase Equilibria - United Atom) method and report average deviations near 40 per cent. Fuchs *et al.* [12] apply a method known as the AUA (Anisotropic United Atom) approach. They

report average deviations closer to 20 per cent, but they employ three parameters per site-type whereas the TraPPE method has only two. The method of Ünlü *et al.* [13] is referred to as the SPEAD (Step Potentials for Equilibria And Dynamics) model, also applying three parameters per site type and reporting average errors for straight chain hydrocarbons, ethers, and alkanes of less than 10 per cent.

Apparently, the use of molecular simulation techniques is becoming a routine way to predict equilibrium thermodynamic properties of fluids and fluid mixtures [12]. Development of realistic and efficient potential models for molecular interactions is a crucial requirement to obtain accurate thermodynamic predictions via molecular simulation techniques which presents the advantage of providing a unified theoretical framework for the prediction of equilibrium and transport properties [14, 15]. However, exploiting this advantage requires that the potential models are transferable [14].

For the interaction models, continuous potentials like Lennard-Jones have been used. On the other hand, discontinuous models of the potentials can provide several benefits [16]. Among these are the clean separation between attractive and repulsive effects and the faster execution of dynamics described by the ultimate multiple time step methods [13]. Moreover, the algorithm for simulating DMD is fundamentally different from the algorithm for continuous potentials, and generally much more efficient [16]. The combination of DMD with TPT builds a bridge between theoretical and experimental work in the sense that it enables a detailed study of the molecular interactions by relating them directly to the macroscopic physical properties [13]. DMD/TPT breaks the molecular potential down into discrete steps, similar to a square-well potential but with more than one attractive well. The depths of the steps characterize the strength of the molecular attraction and its variation with distance. The structure of the molecule in terms of branching, rings, and bond-angles is represented with the same rigorous detail as any other molecular simulation method. Discretizing the potential, on the other hand, permits a clean separation of the repulsive and attractive effects. Cui and Elliott [1] demonstrated that second order TPT provides quantitative accuracy for the impact of the attractive potential on thermodynamic properties. This is quite significant because it means that the dynamics of the full potential need not

be simulated, only the repulsive potential. When combined with the inherently higher computational efficiency of DMD, DMD/TPT affords a quantum leap in the feasibility of molecular based process and product design. When further combined with global optimization, molecular modeling of all physical properties with a single methodology will quickly supersede empirical methods based on years of optimization for narrowly defined physical properties.

There are significant advantages in recognizing the accuracy of TPT for discontinuous potentials. These stem largely from noting that all attractive interactions are negligible for the reference fluids. This means that the exact shape of the disperse potential is not important during the computationally intensive simulation phase. To be specific, the step depths represent perturbations that are specified during the TPT application, so many candidate potentials can be evaluated in rapid succession after the simulation is over. Therefore, the precise characterization of the attractive potentials can be optimized to a high degree of accuracy in a short amount of time.

It may seem surprising that such a simple hybrid approach has not been previously considered for commercial scale molecular modeling. Many years ago, Barker and Henderson [17] observed that TPT for multi step potentials could be formulated such that the second order term was characterized rigorously from purely repulsive simulations. The essential equation is

$$\frac{(A - A^{ig})}{Nk_BT} = \frac{A_0 - A^{ig}}{Nk_BT} + f(\beta) + f(\beta^2) + f(\beta^3) \tag{2.1}$$

where

$$f(\beta) = \frac{\beta}{N} \sum_i \sum_j \sum_m \langle N_{ijm} \rangle \, u_{ijm} \tag{2.2}$$

and

$$f(\beta^2) = -\frac{\beta^2}{2N} \sum_i \sum_j \sum_m \sum_k \sum_l \sum_n \left(\langle N_{ijm} N_{kln} \rangle - \langle N_{ijm} \rangle \langle N_{kln} \rangle \right) u_{ijm} u_{kln} \tag{2.3}$$

where, for example, u_{ijm} designates the attractive energy in the m^{th} well between the i^{th} and j^{th} site types. N_{ijm} is the number of pairs of interactions obtained from the reference fluid simulation and N is the number of molecules. The $\langle\rangle$ denotes an ensemble average of the reference fluid.

Noting that $\beta = 1/k_BT$ where k_B is Boltzmann's constant, Equation (2.1) can be rewritten as a second order series in reciprocal temperature, with the coefficients $A1$ and $A2$ being density dependent, as shown in Equation (2.4).

$$\frac{(A - A^{ig})}{Nk_BT} = a^{ref} + \frac{A1(\eta)}{T} + \frac{A2(\eta)}{T^2} \tag{2.4}$$

where a^{ref} is the Helmholtz departure and is given as

$$a^{ref} = \frac{(A_0 - A^{ig})}{Nk_BT} = \int_0^{\eta} \frac{(Z^{ref} - 1)}{\eta} d\eta \tag{2.5}$$

The Z_{ref} in Equation (2.5) is the compressibility factor which is derived from simulations of the reference fluid over a range of packing fractions, η, and interpolated by regressing Z_i with $i = 1, 2, 3$. The necessary equation is

$$Z_{ref} = (1 + Z_1\eta + Z_2\eta^2 + Z_3\eta^3)/(1 - \eta)^3 \tag{2.6}$$

$A1$ and $A2$ are evaluated from simulations over a range of densities and then interpolated from state to state by polynomials in η. For associating compounds there is an additional contribution, A^{assoc}, given by Wertheim's theory. In the current work, DMD tabulates the value for a^{ref} and number of interactions for all sites in all distances at each density. Inserting the energy values to Equation (2.1), one obtains the Helmholtz energy consistent with the full potential simulation. Hence, time consuming simulation of trial values for the attractive potential is replaced by a function evaluation of a straightforward function as given in Equation (2.1).

Turning to the development of transferable potential functions, preliminary work demonstrated that a single square-well with λ =1.5 was inaccurate for correlating the

vapor pressure of n-alkanes [18]. Cui and Elliott [1] showed that a single well with $\lambda \approx 1.9\ \sigma$ was much better, yet could not fit all the data. Martin and Siepmann [10] have shown that the Lennard-Jones potential exhibits transferability that is more favorable. Chapela's work [19] has shown that a discrete LJ potential can provide porperties that are similar to the continuous potential for spheres. In more recent work, Cui and Elliott [16] established that much more accurate results are achieved when the CH_3 group is modeled independently from the CH_2 groups. These potentials are transferable in the sense that the same potentials are used for the CH_3 and CH_2 groups in ethane, butane, hexane, and octane. Demonstrating this transferability was a significant achievement since it removes any doubt about our ability to transfer step potentials in the same manner as that practiced for continuous potentials.

3. OPTIMIZATION

After accepting the hypothesis that the potentials are transferable and that they can be incorporated into the system through TPT, one is faced with an optimization problem. The optimization problem with one site would be convex and have an unique extremum; however, when more than one site is involved, the bilinear interaction term renders the problem non-convex and hence gives rise to a global optimiziation problem. To start the optimization, one should start with modeling the problem.

3.1. Mathematical Modeling of the Optimization Problem

To start with, the problem is formulated. Although the objective function is to minimize the sum of squares of residuals in vapor pressure ($pSSR$), the ultimate goal is to have the optimum $pSSR$ with the optimum number of variables. To determine the optimum number of variables, 12 models are used to fit the well depths. By fitting them to different models one obtains different $pSSR$'s with different number of variables. The results thereby obtained are then compared to each other using statistical comparisons like statistical F-Test, PRESS and different information criteria. The problem is to minimize $pSSR$ which is defined as

$$pSSR = \sum_{C} \sum_{T} (\%\ P_{C,T}^{err})^2 \tag{3.1}$$

where $\%\ P_{C,T}^{err}$ is the percentage vapor pressure error and is calculated as

$$\%\ P_{C,T}^{err} = \left(\frac{P_{C,T}^{calc} - P_{C,T}^{exp}}{P_{C,T}^{exp}} \right) \cdot 100 \tag{3.2}$$

where $P_{C,T}^{calc}$ is the calculated vapor pressure and is dependent on the well depth of the i^{th} site in the j^{th} well u_{ij}'s, $P_{C,T}^{exp}$ is the experimental data for vapor pressure available in the literature, and subscripts C and T denote the number of compounds and the number of temperature points, respectively.

At a given temperature, vapor pressure for a compound is calculated using the interactions between sites. These interactions are defined for each site separately and the strength of the interaction is thought to be changing with the changing distance between interacting sites. The attractive part of the potential is thus divided into 11 subintervals which are referred to as wells starting from r/σ ratio of 1.0 to 3.0. The potential at r/σ > 3.0 is assumed to be zero. The well depths correspond to the strength of interactions between sites. The absolute value of these depths are a monotone nonincreasing function of the distance between interacting sites.

In Equations (2.2) and (2.3) the interaction strength between two sites i and j at a certain distance, i.e. at the m^{th} well is defined as

$$u_{ijm} = \sqrt{u_{ij}u_{jm}} \tag{3.3}$$

Briefly, the pressure at a certain temperature can be thought as a function of interaction energies between sites and once these interactions are known, pressure and density can be easily calculated.

3.1.1. Models to Generate the Interaction Energies

The attractive part of the potential from r/σ of 1.0 to 3.0 is divided into eleven intervals at 1.10, 1.15, 1.20, 1.30, 1.40, 1.50, 1.70, 1.80, 2.00, 2.40 and 3.00. Then using these intervals, 12 different models are obtained with different number of wells.

3.1.1.1. The Linear-2580 Model. In this model the attractive potential is discretized into four wells from r/σ of 1.0 to 2.0. The outer part of the potential is assumed to be zero. The model is a two parameter model, ϵ_1 and ϵ_4. The four wells are located between r/σ of 1.0 to 1.2, 1.2 to 1.5, 1.5 to 1.8 and 1.8 to 2.0. The first well depth is equal to the first parameter ϵ_1, the last well depth is equal to the second parameter ϵ_4. The second and third well depths are equal to ϵ_2 and ϵ_3, respectively, and ϵ_2 and ϵ_3 are

obtained by a linear interpolation between ϵ_1 and ϵ_4.

$$\epsilon_2 = \epsilon_4 + \frac{2}{3}(\epsilon_1 - \epsilon_4) \tag{3.4}$$

$$\epsilon_3 = \epsilon_4 + \frac{1}{3}(\epsilon_1 - \epsilon_4) \tag{3.5}$$

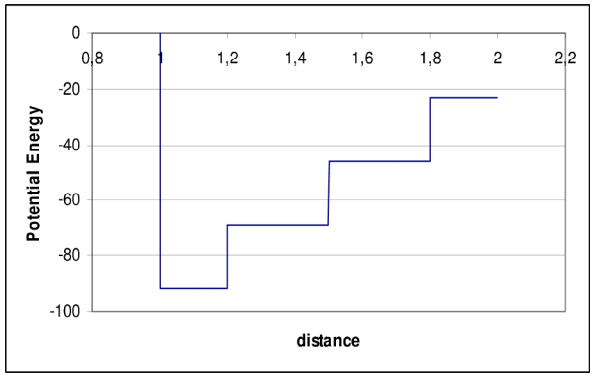

Figure 3.1. The Linear-2580 model

3.1.1.2. The Linear-2580+ Model. This model has three parameters and is similar to the Linear-2580 model in terms of the first four wells except that it contains one additional parameter, ϵ_5, which is located between r/σ of 2.0 to 3.0.

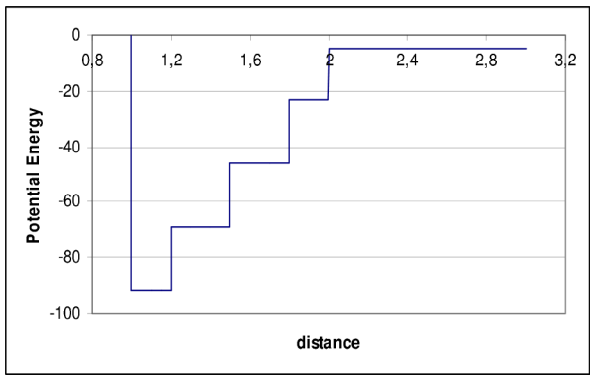

Figure 3.2. The Linear-2580+ model

3.1.1.3. The Unconstrained-2580+ Model. This model consists of five parameters and is developed from the Linear-2580+ model by having the interior dependent wells free. In other words, ϵ_2 and ϵ_3 are the additional parameters.

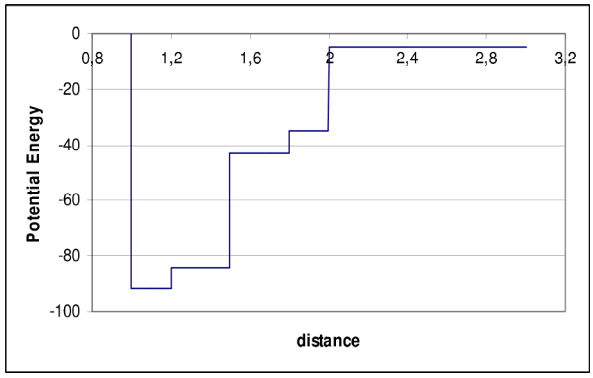

Figure 3.3. The unconstrained-2580+ model

3.1.1.4. The Linearly Interpolated 9 Wells Model. In this model the attractive potential is fitted to a line between r/σ of 1.0 and 2.0 and the outer part of the potential is assumed to be zero. Necessary number of parametes in this model is two, namely the first and last well depths. The non-zero part of the potential is discretized into nine wells at r/σ of 1.1, 1.15, 1.2, 1.3, 1.4, 1.5, 1.7, 1.8 and 2.0. Since there are nine non-zero wells obtained by linear interpolation, we refer to this model as Linear-9.

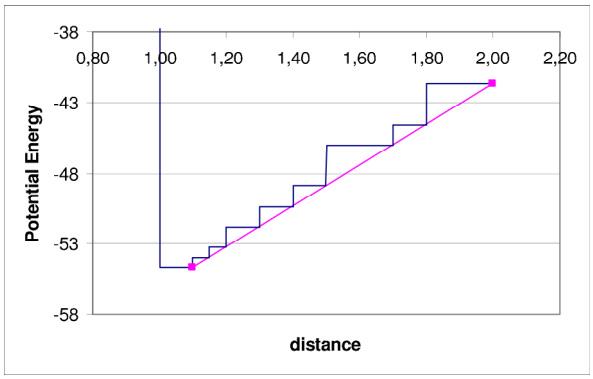

Figure 3.4. Linearly interpolated 9 wells model

3.1.1.5. The Two Jointed Lines Model. In this case two jointed lines are used in between the dimensionless distance r/σ of 1.0 and 2.0 on which nine wells are constructed.

This requires 4 parameters which are the ordinates of three points and abscissa of the middle point where the lines are jointed. The nine wells are then fitted to these lines and the outer part is assumed to be zero. Since there are 2 lines on which the wells are constructed, this model is referred to as 2Lines.

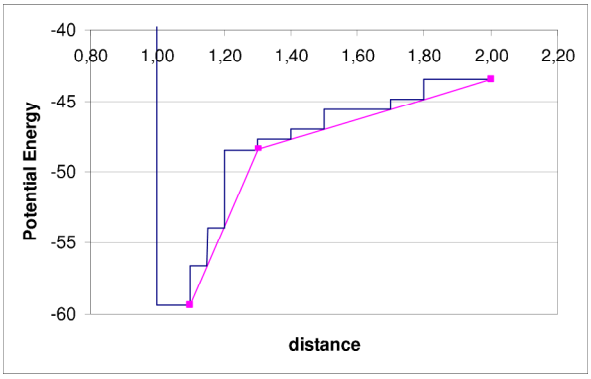

Figure 3.5. The two jointed lines model

3.1.1.6. The Independent 9 Wells. There are nine independent wells between r/σ of 1.0 and 2.0 located as in the Linear-9 model. Since all nine wells are independent, there are nine parameters and this model is referred to as 9 wells model.

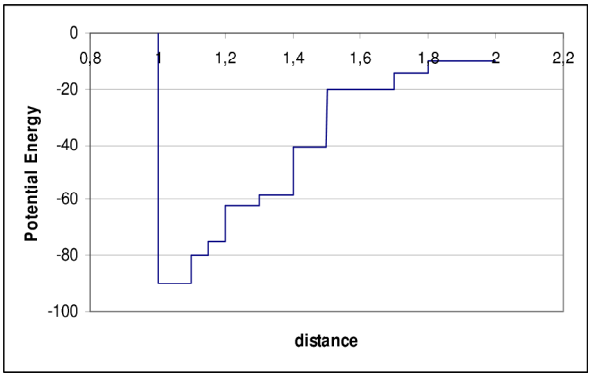

Figure 3.6. The independent 9 wells

3.1.1.7. The Independent 11 Wells. This model is derived from the independent 9 wells model by adding two additional wells between r/σ of 2.0-2.4 and 2.4-3.0. This increases the number of parameters of the model to eleven.

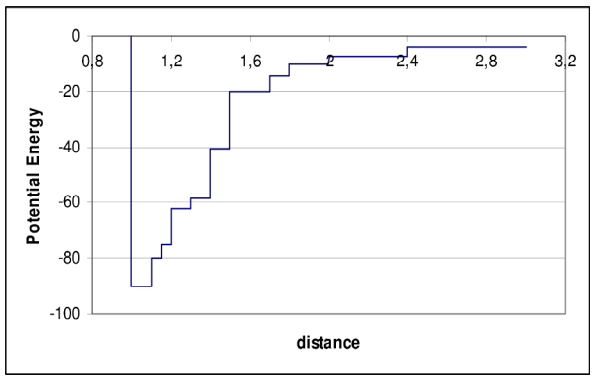

Figure 3.7. The independent 11 wells

3.1.1.8. The Lennard-Jones Model. The eleven wells as identified in the independent 11 wells model are fitted to the Lennard-Jones function which has one parameter, ϵ.

$$y = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right) \tag{3.6}$$

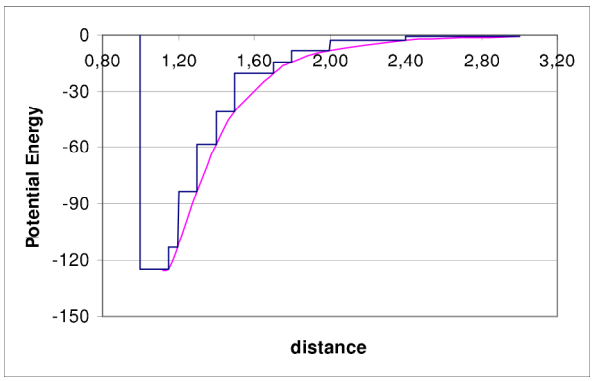

Figure 3.8. The Lennard-Jones model

3.1.1.9. The Yukawa Model. In this model eleven wells are fitted to Yukawa function, which is defined as

$$u = \epsilon \exp(-\kappa\gamma) \tag{3.7}$$

where ϵ is the depth of the first well, γ is the dimensionless distance and κ is the steepness factor. The two parameters of the model are ϵ and κ.

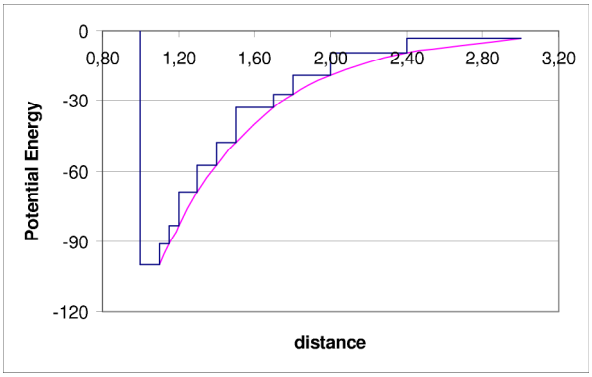

Figure 3.9. The Yukawa model

3.1.1.10. The Yukawa-Universal Model. Different from Yukawa model, the Yukawa-Universal model has a universal steepnes factor κ for all sites. Hence, it is only one parameter model with one additional universal parameter.

3.1.1.11. The StepYukawa Model. This model differs from the Yukawa model by its extended ability of lumping some wells together. For $\gamma \geq X_{lumped}$, the StepYukawa model is defined as

$$u = \epsilon \exp(-\kappa\gamma) \tag{3.8}$$

else

$$u = \epsilon \tag{3.9}$$

where ϵ is the depth of the first well, γ is the dimensionless distance, κ is the steepness of the Yukawa model and X_{lumped} is a dimensionless distance until which the potential is represented by a square well.

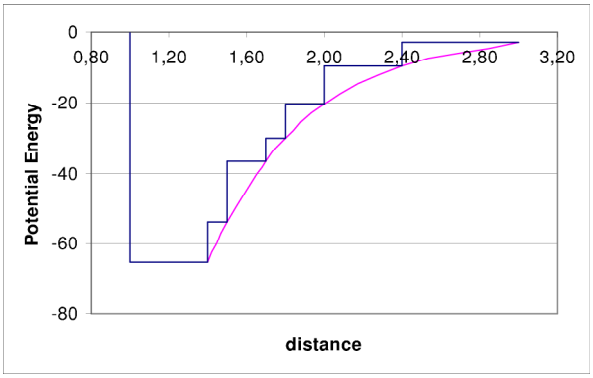

Figure 3.10. The StepYukawa model

3.1.1.12. The StepYukawa-Universal Model. The steepness factor κ and X_{lumped} in the StepYukawa model are the same for all sites. Due to that, the model is a one parameter model with two additional universal parameters.

3.2. Optimization Algorithms

There are several optimization algorithms developed to address chemical engineering optimization problems that can be classified as deterministic or stochastic. Altough deterministic algorithms guarantee the global optimality, they require specific formulations [20]. On the other hand, stochastic methods does not guarantee the global optimality, but they can be applied to many optimization problems. Fortunately, with appropriately selected parameters, they have a high probability of locating the globally optimal solution [20]. Among these stochastic methods are Recursive Random Search (RRS), Tabu Search (TS), Genetic Algortihm (GA) [21] and simulated annealing (SA) [22]. SA and GA are most widely used algorithms since they require little a priori information from the concerned problem and are generally applicable. The disadvantage of these algorithms is that they lack in efficiency. In general, these are combined with local search methods to increase their efficiency. TS is a heuristic approach for solving optimization problems by using a guided, local search procedure to explore the entire solution space without becoming easily trapped in a local optima [23]. It differs from other stochastic optimization techniques by maintaining lists of previous solutions that help guide the search process [23]. RRS is based on the initial high efficiency of random sampling. Besides this high efficiency, RRS algorithm is also robust to random noise

and trivial parameters in the objective function. There are also several local search algorithms such as Levenberg-Marquardt (LM) and quasi-Newton methods.

In this work three optimization algorithms are considered. These are RRS as stochastic method and LM and quasi-Newton as local search algorithms. Local search algorithms are used to decrease the computational effort after a promising area is identified using the stocastic method.

3.2.1. The Levenberg-Marquardt Algorithm

A general class of algorithms using gradient can be expressed as

$$x_{k+1} = x_k - \alpha M_k^{-1} \nabla f(x_k) \tag{3.10}$$

where M_k is a n×n matrix, α is a positive search parameter, and $\nabla f(x_k)$ is the gradient at the point x_k. When M_k is equal to identity matrix, I, the algorithm is steepest descent and when M_k is the Hessian, H, it is Newton's algorithm. Levenberg-Marquardt is a compromise between these two and is obtained by taking $M_k = (H + \lambda_k I)$, where λ_k is a positive number which makes the M_k positive definite [24]. The specific implementation is that of MINPACK [25].

3.2.2. Quasi-Newton Method

The BCONF (or DBCONF for double precision) is taken from the IMSL Math Library. It minimizes a function of N variables subject to bounds on the variables using a quasi-Newton method and a finite-difference gradient.

From a given starting point x_c, an active set IA, which contains the indices of the variables at their bounds, is built. A variable is called a free variable if it is not in the active set. The routine then computes the search direction, d, for the free variables

according to the formula

$$d = -B^{-1} g_c \tag{3.11}$$

where B is a positive definite approximation of the Hessian and g_c is the gradient evaluated at x_c; both are computed with respect to the free variables. The search direction for the variables in IA is set to zero. A line search is used to find a new point x_n;

$$x_n = x_c + \lambda d \qquad \lambda \in (0, 1] \tag{3.12}$$

such that

$$f(x_n) \leq f(x_c) + \alpha g^T d \qquad \alpha \in (0, 0.5) \tag{3.13}$$

Finally, the optimality conditions

$$||g(x_i)|| \leq \epsilon \qquad l_i < x_i < u_i \tag{3.14}$$

$$g(x_i) < 0 \qquad x_i = u_i \tag{3.15}$$

$$g(x_i) > 0 \qquad x_i = l_i \tag{3.16}$$

are checked, where ϵ is a gradient tolerance. When optimality is not achieved, B is updated according to the Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula

$$B \leftarrow B - \frac{Bss^T B}{s^T Bs} + \frac{yy^T}{y^T s} \tag{3.17}$$

where $s = x_n - x_c$ and $y = g_n - g_c$. Another search direction is then computed to

begin the next iteration. The active set is changed only when a free variable hits its bounds during an iteration or the optimality condition is met for the free variables but not for all variables in IA, the active set. In the latter case, a variable that violates the optimality condition will be dropped out of IA [26, 27].

3.2.3. The Random Recursive Search

In this work, RRS is used as the randomized sampling algorithm. Random sampling is the simplest and most widely used search technique. It takes random samples from a uniform distribution over the variable space. Despite its simplicity, random sampling is able to provide a strong probabilistic convergence guarantee. Furthermore, random sampling has surprisingly proved to be more efficient than deterministic exploration methods, such as, grid covering, in terms of some probabilistic criteria and it is especially so for high-dimensional problems [28].

As the number of function evaluations increases, bracketing the true optimum within r-percentile diminishes. However, this lessening levels off as shown in Figure 3.11. This means that random sampling is efficient initially, but loses efficiency as the number of function evaluations increases. The disadvantage of random sampling is its

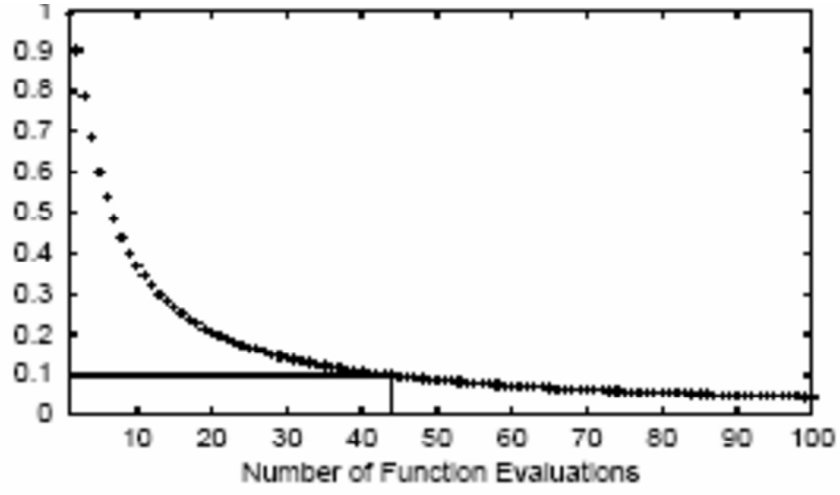

Figure 3.11. Converging curve vs. number of function evaluations

apparent lack of efficiency at the later steps. The basic idea of RRS is to keep the initial

efficiency of random sampling by restarting it before its efficiency becomes low. The restart of the random sampling is accomplished by changing its sample space. Basically, random sampling is performed for a number of times, and then the sample space is moved or resized according to the previous samples and another random sampling is restarted in the new sample space. At the beginning of the search, RRS performs sampling from the entire variable space and thus examines the overall structure of the objective function. With the search continuing and the sample space gradually shrinking, the search algorithm obtains more detailed information about the objective function until it finally converges to an optimum.

A stochastic search algorithm usually comprises two elements: exploration and exploitation. Exploration aims to identify promising areas in the parameter space, while exploitation attempts to exploit local information to quickly improve the solution. Basically, the RRS algorithm uses random sampling for exploration and recursive random sampling for exploitation. In the exploration part, RSS determines promising areas and executes the exploitation procedure only in these areas.

Restart of the random sampling is performed through resizing or moving of the sample space. This resizing is performed only if a better result than that in the last sampling could not be found and this procedure is basically a "shrinkage" of the space around the best set of variables by a shrinkage ratio. If a better result is found then the variable space is realigned such that the best set of variables is in the center of the new space with the same size. With realignment and shrinkage alternately performed, the sample space gradually moves to the local optimum. This local optimum is probably the global optimum if multiple explorations and exploitations converge to the same solution.

3.2.3.1. The Exploration Phase. Within the space assigned for variables, a random search point is generated using uniformly distributed random numbers in the interval [0, 1]. Then the trial point x_{ij} at which the interaction energies are obtained is determined

from the relation

$$x_{ij} = LB_{ij} + (UB_{ij} - LB_{ij}) \cdot rand \tag{3.18}$$

where *rand* stands for a uniformmly distributed random number, LB_{ij} is the lower bound and UB_{ij} is the upper bound for the j^{th} parameter of the i^{th} site.

With the trial set of interaction energies obtained at randomly selected points, vapor pressure is calculated using the necessary thermodynamic relations. This procedure is repeated R times and in each repetition S subiterations are carried out. The optimum function value and the decision variable vector are stored and passed to the exploitation phase.

3.2.3.2. The Exploitation Phase. The initial space of search in the exploration phase is resized by a shrinkage factor and relocated around the optimum decision variable point obtained in the exploration phase in the following manner

$$LB_{ij}^{new} = x_{ij} - (UB_{ij} - LB_{ij}) \cdot \frac{k}{2} \tag{3.19}$$

$$UB_{ij}^{new} = x_{ij} + (UB_{ij} - LB_{ij}) \cdot \frac{k}{2} \tag{3.20}$$

where k is the shrinkage ratio between 0 and 1 and usually taken around 0.6, x_{ij} is the best point obtained from the previous iteration.

After obtaining the new search space, R iterations with S subiteration in each are carried out and the optimum function value is compared to that of the previous one. If this newly obtained function value is smaller than the old one then the space of search is relocated around this new point. If this is not the case, the space is shrunk around the old point by a factor k. These alternating shrinkage and realignment procedures are performed until a termination criterion for a single exploitation phase is satisfied.

3.3. Evaluation at Different Solution Methods

The search for a suitable algorithm is started with the RRS, LM and BCONF algorithms individually. The analysis then turns to combinations of the RRS with the LM or the BCONF. Straight chain alkanes ranging from nC_3 to nC_{30} excluding nC_{28} and nC_{29} are used as the dataset of study. The model to fit the attractive potential is the Linear-2580. To evaluate the accuracy of each optimization algorithm, it was necessary to estimate the global optimum. Many runs with different optimization methods are performed and no $pSSR$ below 42440 is obtained. Hence, the global optimum is assumed to be at the point corresponding to this $pSSR$. These are given in Table 3.1. A more detailed explanation is given in the following sections.

Table 3.1. The assumed global optimum

Site	ϵ_1	ϵ_4
CH_3a	58.526	38.893
CH_2a	33.593	17.166

3.3.1. The Random Recursive Search

The RRS has its own parameters to enhance its efficiency. These are mainly

1. Number of iterations in the exploration phase (nTrialsExplore),
2. Number of iterations in the exploitation phase (nTrialsExploit),
3. Shrinkage Ratio,
4. Termination Criteria: nFails and tolerance for diffNormEOK

The termination criterion is set as $nFails$ consecutive failures of single exploitation results. A failure of a single exploitation means that the newly obtained result is not better than the previous results. Since several exploitations may be required, one additional termination criterion for single exploitaitons is necessary. It is defined such that the exploitaiton phase will continue unless the root mean square of the norms of the search space, $diffNormEOK$, is less than a given tolerance. $diffNormEOK$ is

given as

$$diffNormEOK = \sqrt{\frac{\sum_{i=1}^{p}(UB_i - LB_i)^2}{p}} \tag{3.21}$$

where p is the number of decision variables and UB_i and LB_i are the upper and lower boundaries of the i^{th} variable, respectively. To get the optimum with the least number of function evaluations the parameters should be adjusted.

3.3.1.1. Adjusting nTrialsExploit. To investigate the relation between nTrialsExploit and the number of variables, several runs for the number of variables of 1, 2 and 4 are done by gradually increasing nTrialsExploit. nTrialsExplore is set to two times of nTrialsExploit and the shrinkage ratio is set to 0.6 for all runs. For one run, 6 exploitations are performed and the best is accepted as the final result.

To start with only one variable, upper and lower boundaries of all variables except for the ϵ_4 of CH_2a are set to the global optimum values. The center of the interval of search for the ϵ_4 of CH_2a is the global optimum value and the width of the interval is 20. The values are given in Table 3.2. Three runs with nTrialsExploit of 3, 5 and 10

Table 3.2. Upper and lower boundaries for one decision variable

	CH_3a		CH_2a	
	ϵ_1	ϵ_4	ϵ_1	ϵ_4
Upper Bound	58.526	38.893	33.593	27.166
Lower Bound	58.526	38.893	33.593	7.166

were performed, each resulting in exactly the same global optimum with that of LM. It is concluded that having nTrialsExploit value 3 is sufficient for optimization of one variable. The results can be seen in Figures 3.12 and 3.13.

For two variables, a similar procedure was performed. The upper and lower boundaries are arranged as in Table 3.3. After performing five runs with nTrialsExploit of 5, 10, 13, 15 and 20 it is found that for two variables it is sufficient to have

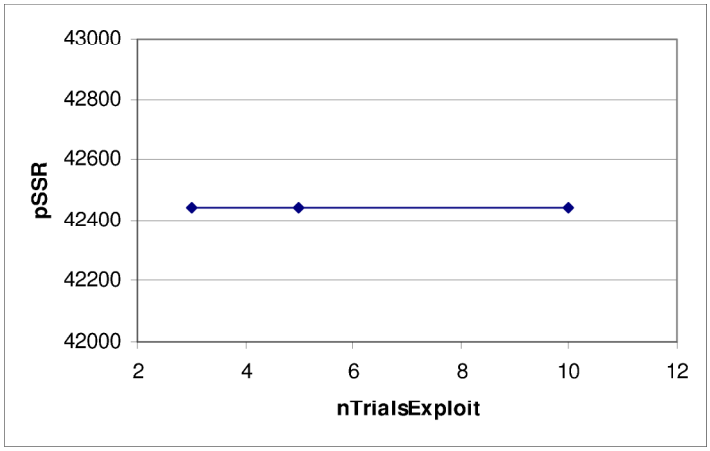

Figure 3.12. pSSR versus nTrialsExploit for one variable

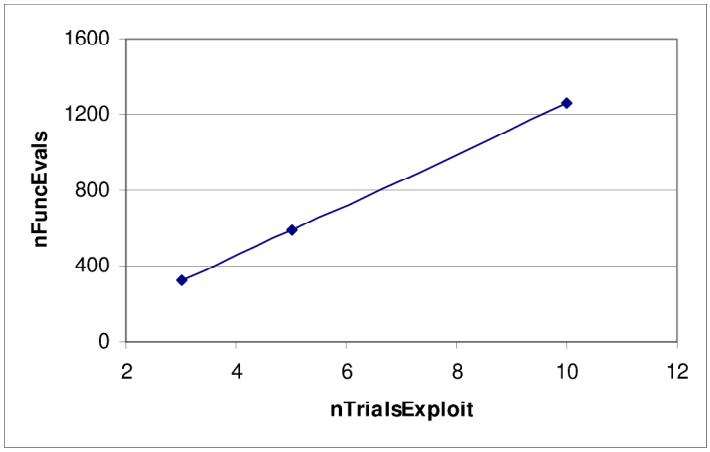

Figure 3.13. nFunCalls versus nTrialsExploit for one variable

nTrialsExploit as 15. The results are demonstrated in Figures 3.14 and 3.15. For four variables, 9 runs are done with nTrialsExploit of 5, 10, 20, 40, 50, 60, 80, 100 and 150. The algorithm could not converge to the assumed global optimum even with nTrialsExploit of 150. Required number of function evaluations are given in Figure 3.17. As can be seen in Figure 3.16, pSSR with nTrialsExploit of 60 is slightly higher than that of with 20. This type of rise and falls can be explained by the random nature of the search algorithm and flatness of the reponse surface. Because of the flatness of the surface demonstrated in Figures 3.18 and 3.19, the algorithm can terminate before converging to the optimum. Although it seems that the algorithm could never converge to the global optimum for four variables, it can be seen that the situation is not that dramatic if the root mean squares in differences in variables (rms(devEOK))

Table 3.3. Upper and lower boundaries for two decision variables

	CH_3a		CH_2a	
	ϵ_1	ϵ_4	ϵ_1	ϵ_4
Upper Bound	58.526	38.893	43.593	27.166
Lower Bound	58.526	38.893	23.593	7.166

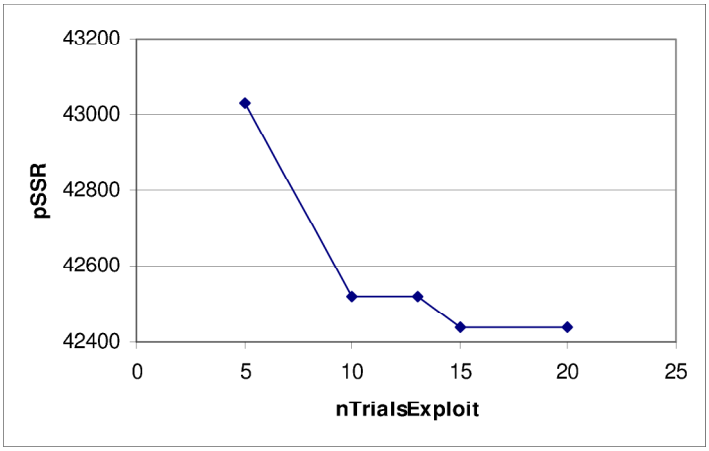

Figure 3.14. pSSR versus nTrialsExploit for two variables

are investigated. As can be seen in Table 3.4, for four decision variables, the RRS could converge to about 2 per cent neighborhood of the global optimum coordinates even with nTrialsExploit of 10. For two decision variables, it could converge to 1.33 per cent of the global optimum with nTrialsExploit of 5. In all cases the percentage deviation of the pSSR (% devPSSR) from that of the global optimum is less than 1.5. The significant rms(devEOK) despite small % devPSSR can be explained by a flat surface around the assumed global optimum. This behaviour is shown in Figures 3.18 and 3.19. The variables given in Tables 3.4, 3.8, 3.10, 3.12 and 3.14 are defined in

Table 3.4. Detailed investigation of the results for 4 variables

p	nTrialsExploit	rms(devEOK)	rms(% devEOK)	% devPSSR
4	10	0.989	2.064	1.013
4	20	0.640	1.418	0.353
4	40	0.933	2.104	0.778
2	5	0.368	1.337	1.390
2	10	0.202	0.735	0.189

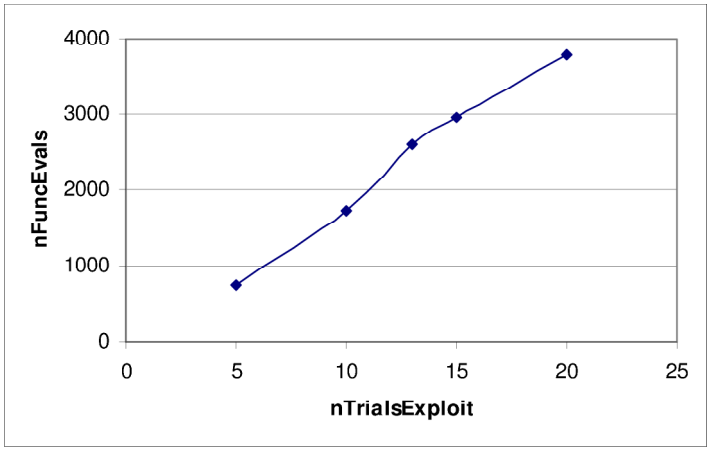

Figure 3.15. nFunCalls versus nTrialsExploit for two variables

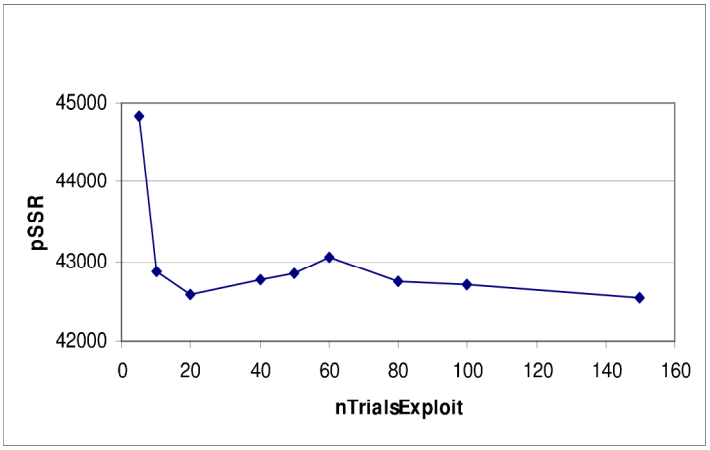

Figure 3.16. pSSR versus nTrialsExploit for four variables

Equations (3.22), (3.23), (3.24) and (3.25).

$$rms(devEOK) = \sqrt{\frac{\sum_{i=1}^{p}(EOK_i - EOK_i^{global})^2}{p}} \tag{3.22}$$

$$rms(\%\ devEOK) = \sqrt{\frac{\sum_{i=1}^{p}\left(\frac{EOK_i - EOK_i^{global}}{EOK_i^{global}}\right)^2}{p}} \cdot 100 \tag{3.23}$$

$$\%\ devP = \left(\frac{pSSR - pSSR^{global}}{pSSR^{global}}\right) \cdot 100 \tag{3.24}$$

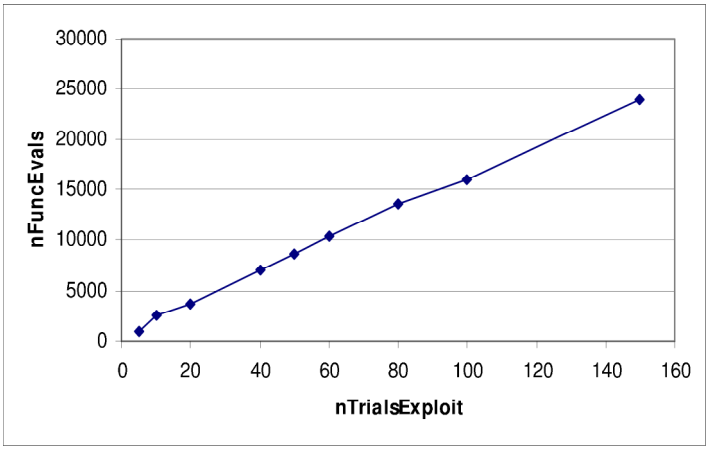

Figure 3.17. nFunCalls versus nTrialsExploit for four variables

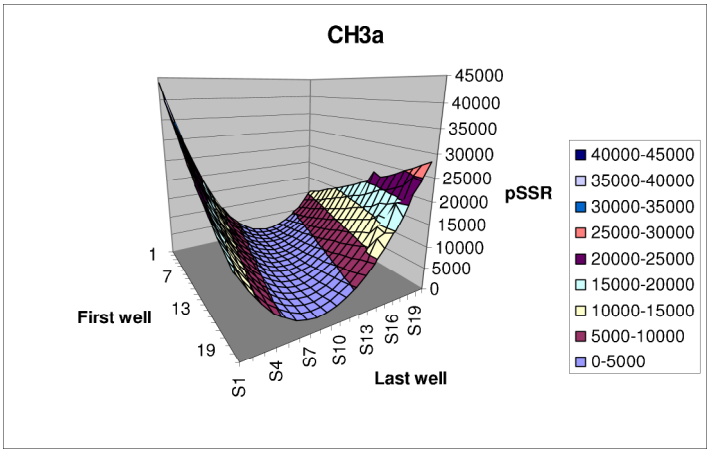

Figure 3.18. Flat surface around the optimum with respect to CH_3a

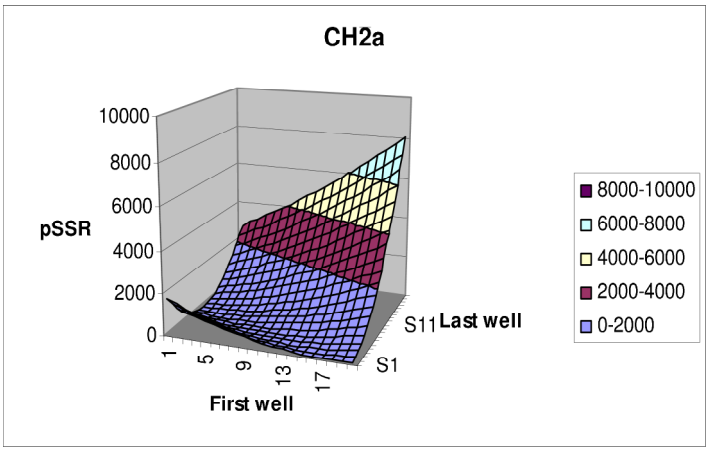

Figure 3.19. Flat surface around the optimum with respect to CH_2a

$$RMSE = \sqrt{\frac{pSSR}{n}} \tag{3.25}$$

where p is the number of decision variables and n is the number of data points where the vapor pressures are evaluated.

To conclude, the exact global optimum is hard to achieve using RRS even if the number of variables is increased to 4; nevertheless, very close solutions can be obtained with a small number of iterations in the exploration and exploitation sub-phases.

3.3.2. Direct Application of LM and BCONF Algorithms

It is not possible to use the LM and the BCONF directly since they cannot converge to the global optimum if started from an initial guess which is not close to the global optimum. This behavior is demonstrated with 16 different starting points which are obtained either by adding or subtracting 10 from the global optimum coordinates given in Table 3.1. "0" in the start index denotes a subtraction and "1" denotes an addition. As an example, "0100" denotes a starting point which has first, third and forth variable 10 less than the global optimum and second variable 10 more than the global optimum. The results are tabulated in Table 3.5. "no convergence" means that the pSSR is above 10^8.

3.3.3. Combining Different Optimization Algorithms

Since it is shown that the RRS necessitates exhaustive number of iterations to find the global optimum even if there are only 4 decision variables. Herein, 3 more procedures are proposed which avoid exhaustive number of iterations of the RRS.

1. To continue optimization with the LM after performing a crude optimization using the RRS,
2. To continue optimization with a 2^{nd} batch of the RRS after performing a crude optimization using the RRS,
3. To continue optimization with the BCONF after performing a crude optimization using the RRS.

Table 3.5. Different starting points for LM and BCONF

		BCONF		LM	
trial	startIndex	nFunCalls	result	nFunCalls	result
1	0000	25	no convergence	11	no convergence
2	1000	39	no convergence	53	no convergence
3	0100	32	no convergence	22	no convergence
4	0010	27	no convergence	18	no convergence
5	0001	5	no convergence	22	no convergence
6	1100	25	no convergence	7	no convergence
7	1010	34	no convergence	12	no convergence
8	1001	1	no convergence	6	no convergence
9	0110	20	no convergence	17	no convergence
10	0101	1	no convergence	6	no convergence
11	0011	8	no convergence	7	no convergence
12	1110	32	no convergence	7	no convergence
13	1101	1	no convergence	6	no convergence
14	1011	22	no convergence	7	no convergence
15	0111	39	no convergence	29	no convergence
16	1111	3	no convergence	7	no convergence

3.3.3.1. Performing a Crude Optimization with RRS. The RRS parameters for crude optimization are set such that nTrialsExplore value is 20, nTrialsExploit value is 10, shrinkRatio is 0.6, nFails is 5 nonconsecutive and tolerance is 0.001. Upper and lower boundaries for the crude optimization are given in Table 3.6. Single exploitation results and detailed investigation of the results from the crude optimization with the RRS are given in Tables 3.7 and 3.8, respectively.

3.3.3.2. LM after crude RRS. After the crude optimization with RRS, RRS is substituted with LM. Initial points for LM runs are single exploitation results from RRS. Hence, 6 LM results are obtained. These results and their detailed investigations are given in Tables 3.9 and 3.10. Although LM uses results of crude RRS as initial guess,

Table 3.6. Upper and lower boundaries for crude RRS

	CH_3a		CH_2a	
	ϵ_1	ϵ_4	ϵ_1	ϵ_4
Upper Bound	68.526	48.893	43.593	27.166
Lower Bound	48.526	28.893	23.593	7.166

Table 3.7. The results of the crude optimization

	CH_3a		CH_2a	
trial	ϵ_1	ϵ_4	ϵ_1	ϵ_4
1	56.036	40.005	36.225	16.025
2	65.497	35.996	26.549	20.250
3	52.384	44.983	32.005	17.627
4	53.764	41.784	35.306	16.396
5	60.205	37.964	34.028	16.961
6	51.140	47.033	28.554	19.025

second trial of LM lead to a result worse than that of crude RRS. Number of function evaluations of 8 in this trial explains that the LM algorithm has chrashed in that trial.

3.3.3.3. 2^{nd} batch of RRS. After the crude optimization with the RRS, the upper and lower boundaries are set to ± 2 of the single exploitation results wherein the RRS is again carried out which is referred to as the 2^{nd} batch. The 2^{nd} batch optimizations are performed with nTrialsExplore value of 40, nTrialsExploit value of 20 and a shrinkRatio of 0.6. The termination criterion is 5 consecutive fails of the exploitation results with tolerance of 0.001. The obtained results are given in Tables 3.11 and 3.12.

3.3.3.4. BCONF after crude RRS. Similar to LM after crude RRS, RRS is replaced with BCONF after a crude RRS. Initial points for BCONF runs are single exploitation results from crude RRS. Hence, 6 BCONF results are obtained. These are given in Tables 3.13 and 3.14. The trials 1, 2 and 4 lead to results worse than that of crude

Table 3.8. Detailed investigation of the crude optimization results

trial	nFunCalls	rms(devEOK)	pSSR	% devP	RMSE
1	217	1.979	43920	3.487	10.723
2	230	5.388	55120	29.877	12.012
3	576	4.403	53890	26.979	11.877
4	230	2.939	43740	3.063	10.701
5	254	0.989	42870	1.013	10.594
6	956	6.117	65410	54.123	13.086

Table 3.9. The results of LM

	CH_3a		CH_2a	
trial	ϵ_1	ϵ_4	ϵ_1	ϵ_4
1	59.359	38.081	34.611	16.744
2	65.504	35.996	26.555	20.250
3	58.541	38.823	33.816	17.072
4	61.225	36.895	34.030	17.012
5	58.710	38.780	33.575	17.175
6	59.478	38.354	33.103	17.384

RRS which may be because of crashing of BCONF algorithm before termination.

After investigating number of function evaluations (nFunCalls) and RMSE for three procedures, it is concluded that the 2^{nd} batch of the RRS still requires exhaustive number of function evaluations. Even though BCONF has lower number of function evaluations than LM, it is decided to use LM over BCONF for it has lower RMSE. The RRS, in this formulation, finds promising locations for the LM to start and the LM finds the minimum around each promising location.

3.3.4. RRS Revisited

At this point, it has been concluded to use a second batch optimization with the LM after each exploitaion obtained by the RRS. Thanks to that, the exhaustive

Table 3.10. Detailed investigation of LM results

trial	nFunCalls	rms(devEOK)	pSSR	% devP	RMSE
1	51	0.801	42860	0.990	10.592
2*	8	5.388	55260	30.207	12.027
3	43	0.126	42510	0.165	10.549
4	41	1.695	42630	0.448	10.564
5	41	0.108	42440	0.000	10.540
6	61	0.609	42480	0.094	10.545

Table 3.11. The results of 2^{nd} batch of RRS

	CH_3a		CH_2a	
trial	ϵ_1	ϵ_4	ϵ_1	ϵ_4
1	58.452	38.693	34.395	16.831
2	66.761	33.997	31.361	18.196
3	52.465	42.850	34.702	16.637
4	55.609	40.519	35.535	16.309
5	59.473	38.304	33.457	17.230
6	52.444	43.201	33.147	17.298

number of function evaluations of the RRS is decreased significantly. To decrease it further without worsening the results, one can adjust other parameters of the RRS.

3.3.4.1. Adjusting tolerance for diffNormEOK. The stopping criterion for single exploitations is satisfied if $diffNormEOK$ in Equation (3.21) is less than a tolerance. By increasing tolerance for the stopping criterion, it is expected to have smaller nFunCalls in the expense of higher $pSSR$ values. As expected, nFunCalls is decreased for single exploitations with increasing tolerance; however, overall nFunCalls did not change significantly because the number of single exploitations is independent of the tolerance for $diffNormEOK$. In Table 3.15, the mean of the number of function evaluations per single exploitation, overall number of function evaluations and the $pSSR$s obtained from the crude RRS and the LM are given. As can be seen in Table 3.15, it

Table 3.12. Detailed investigation of 2^{nd} batch of RRS results

trial	nFunCalls	rms(devEOK)	pSSR	% devP	RMSE
1	5668	0.447	42610	0.401	10.561
2	5044	4.945	44400	4.618	10.781
3	3384	3.671	43580	2.686	10.681
4	5492	1.979	43360	2.168	10.654
5	2909	0.563	42510	0.165	10.549
6	4133	3.734	44720	5.372	10.820

Table 3.13. The results of BCONF

	CH_3a		CH_2a	
trial	ϵ_1	ϵ_4	ϵ_1	ϵ_4
1	56.036	40.005	36.224	16.019
2	65.497	35.996	26.549	20.243
3	53.244	42.220	35.073	16.472
4	53.764	41.784	35.306	16.390
5	60.016	37.836	33.997	17.001
6	53.877	41.898	33.992	16.964

is concluded that LM can converge to global optimum or to a very close point regardless of the magnitude of the tolerance for $diffNormEOK$. To have more reasonable nFunCalls, tolerances between 0.5 and 2.0 can be taken.

3.3.4.2. Adjusting nFails. To decrease nFunCalls in the crude optimization with the RRS further, number of consecutive failures of the single exploitation can be decreased. A failure in the single exploitation means that the result obtained in that exploitation is worse than the best result obtained since the beginning of the run. The termination criterion is satisfied if the number of consecutive failures of single exploitations exceeds a given integer value, $nFails$.

Previously, it was determined that tolerance for diffNormEOK could be taken

Table 3.14. Detailed investigation of BCONF results

trial	nFunCalls	rms(devEOK)	pSSR	% devP	RMSE
1*	25	1.980	43940	3.534	10.725
2*	18	5.387	55170	29.995	12.018
3	42	3.226	43660	2.875	10.691
4*	17	2.940	43750	3.087	10.702
5	30	0.939	42550	0.259	10.554
6	42	2.777	43610	2.757	10.685

Table 3.15. Comparison of tolerances

	nFunCalls		pSSR	
Tolerance	mean	overall	crude RRS	LM
0.1	159.5	1097	43680	42440
0.5	94.2	702	43890	42600
0.9	79.4	627	44090	42890
1.5	61.8	511	46260	42510
2.0	54.7	692	47080	42440

between 0.5-2.0. In this work, $nFails$ is changed for 3 different tolerances (0.5, 0.9 and 1.5) and pSSR and nFunCalls are compared to each other. Crude optimization results obtained with RRS and LM results are the same for nFails of 5 and 1 for all tolerances. However, as can be seen in Table 3.16, nFunCalls in RRS are quite different for nFails of 5 and 1 with different tolerances. Although the results obtained with $nFails$ of 1

Table 3.16. Comparison of nFails

	nFunCalls	
Tolerance	nFails = 5	nFails = 1
1.5	511	167
0.9	627	191
0.5	702	214

and 5 are the same, for engineering safety, it is always better to have it around 5.

3.3.4.3. Adjusting Shrinkage Ratio. The shrinkage ratio can be adjusted to increase the efficiency and decrease the nFunCalls of the RRS. It is expected that as the shrinkage ratio approaches to 1, the number of function evaluations and the probability of finding the optimum increases. As determined previously, nFails and tolerance for diffNormEOK are taken as 1 and 1.5, respectively. nTrialsExplore is 20 and nTrialsExploit is 10. As can be seen in Table 3.17, the LM can converge to the assumed

Table 3.17. Comparison of shrinkage ratios

		pSSR	
shrink Ratio	nFunCalls	crude RRS	LM
0.1	208	47370	42510
0.3	300	47390	42440
0.5	206	56050	42430
0.6	167	46260	42430
0.7	231	44480	42440
0.9	837	43430	42770

global optimum regardless of the starting point obtained by the RRS. Although higher shrinkage ratios increase the success of the crude RRS, it does not influence the overall result. Hence, any shrinkage ratio can be taken, preferably one which does not lead to high nFunCalls. Hence the value of 0.6 for shrinkRatio has been assumed throughout this study for RRS implementation.

3.4. Conclusion for Optimization Algorithm Selection

This systematic study of stochastic and gradient algorithms leads to the conclusion that the optimal regression model for the SPEADMD model should include elements of both algorithms. A crude application of RRS with a single exploitation failure, followed by LM iteration to the optimum provides accuracy comparable to an exhaustive search for global optimum while minimizing number of function evaluations. Substitution of BCONF algorithm in place of LM algorithm provided similar performance, but slightly reduced accuracy relative to the assumed global optimum.

To ensure that a reliable estimate of the global optimum is achieved, it is recommended that multiple trials be performed with independent explorations being followed by exploitation and LM application. Based on RMSE and the $diffNormEOK$ for roughly three trials being within tolerance of each other, the global optimum can be identified. The reason for this performance is probably because the contours of the response surface are reasonably smooth while the contours near the optimum are fairly flat along the direction of steepest descent. This behavior is evident in the 2 variable study performed here for n-alkanes at least, although it is difficult to visualize for more complex optimizations. Despite the smoothness of the contours near the optimum, a direct application of a either gradient algorithm was found to fail when initial guesses were generated along the boundaries of the search without the benefit of the stochastic search. This behavior derives from the strong sensitivity of the vapor pressure to the potential and the divergence of the thermodynamic calculation when poor guesses are made. The gradient becomes flat in that case because all trials in the vicinity of a poor guess result in virtually infinite error.

While more optimal algorithms may be possible, this particular combination of stochastic and gradient algorithms seems reasonable for a comprehensive study of different models for a large number of site types. Examples of different models like the LJ and StepYukawa are given in the previous chapters. Examples of site types represent branched alkane, alcohol, and aromatic characters in molecules. The following chapter describes the undertaking of such a comprehensive study.

4. MODEL SELECTION

The main objective of the model selection is to find a model which predicts the vapor pressure with a minimal error and maximal reliability. Instead of using all available data for optimizing the variables of a model, the dataset is divided into two equal parts, namely the training set and the validation set. For a model, the variables are obtained using the training set and the comparison is carried out using the validation set, since a decrease in the pSSR of the training set may be because of fitting the noise, which has no (or worsening) effect on the validation set. For model selection, five methods are used. These are:

1. Cross-Validation (PRESS),
2. Akaike's Information Criterion (AIC),
3. Bayesian Information Criterion (BIC),
4. Mallow's Information Criterion (MIC),
5. Statistical F-Test

In the cross-validation, sum of $pSSR$s of the validation sets are used, whereas the $pSSR$ with other methods is obtained for the whole dataset and then used to compare the models.

4.1. Cross-Validation (PRESS)

The PRESS is a prediction-based model selection statistic. It is the summation of the sum of squares of residuals of the validation sets. The model with a smallest PRESS statistic is the best model. The PRESS statistic is an example of the general idea of using cross-validation to assess the predictive power of models. Dividing the entire set into two subsets is known as 2-fold cross-validation.

For PRESS analysis, the dataset is divided into two subsets such that both sets have equal numbers of all sites and these sites are given in Table 4.1.

Table 4.1. Number of sites in the subsets

Site	Number	
CH_3a	28	28
CH_3b	15	15
CH_3c	1	1
CH_3d	2	2
CH_3e	1	1
CH_2a	28	28
CH_2b	2	2
CH_2c	6	6
CH_2d	1	1
CH_2e	1	1
CHa	11	11
CHb	2	2
CHc	4	4
C	1	1
Ar-CHa	8	8
Ar-Ca	6	6
Ar-Cb	1	1

4.2. Akaike's Information Criterion (AIC)

The AIC statistic is defined as

$$AIC = n \ln\left(\frac{SSR}{n}\right) + 2p \tag{4.1}$$

where SSR is the sum of squares of residuals, n is the number of the observations and p is the number of decision variables.

AIC can be used only to compare models and it is not an absolute measure of the fit of the model. Therefore, to compare models, Δ AIC should be used rather than

its absolute values. All models having Δ AIC less than 2 are considered equal.

4.3. Bayesian Information Criterion (BIC)

BIC is similar to AIC but the penalty on the number of decision variables is greater. The BIC statistic for a model is

$$BIC = n \ln\left(\frac{SSR}{n}\right) + p \ln(n) \tag{4.2}$$

The penalty on the variables is $p \ln(n)$ which was $2p$ in the former method. The purpose of having the penalty dependent on the n is to reduce the likelihood that small and relatively unimportant parameters are included (which is more likely with large n).

4.4. Mallow's Information Criterion (MIC)

The MIC statistic for a given model can be calculated by

$$MIC = (n - q)\frac{SSR_{reduced}}{SSR_{full}} + 2p - n \tag{4.3}$$

where SSR is the sum of squares of residuals, n is the number of the observations, p is the number of decision variables of the reduced model and q is the number of decision variables of the full model.

MIC is closely related to AIC. AIC has come to be preferred by many statisticians in recent years. BIC, which is also similar to AIC, is motivated by a Bayesian approach to model selection and is said not to tend to overfit like AIC. Since the RMSE with all models are quite close to each other, the number of decision variables in the model becomes more significant in model selection. With these in mind, the main method for model selection is chosen as BIC, which can be supported by AIC and MIC.

4.5. Statistical F-Test

In regression work, the question often arises as to whether or not it was worthwhile to include certain terms in the model. This question can be investigated by considering the extra portion of the regression sum of squares which arises due to the fact that the terms under consideration were in the model. The mean square derived from this extra sum of squares can then be compared with the estimate, s^2, of σ^2 to see if it appears significantly large. If it does, the terms should have been included; if it does not, the terms will be judged unnecessary and could be removed [29].

F^* can be calculated in terms of the reduction in the residual sum of squares as

$$F^* = \frac{SSR_{Red} - SSR_{Full}}{df_{Red} - df_{Full}} \cdot \frac{df_{Full}}{SSR_{Full}} \tag{4.4}$$

where df is the degrees of freedom.

This F^* can then be compared to $F(q-p, q, 1-\alpha)$; however, this test will not be exact since the models used to calculate SSR are nonlinear, hence our level of significance that was specified is not exact. This kind of application of F-test for nonlinear models will only give an approximate level of significance. One possible way may be to compare the residual mean squares visually as suggested in [29]; another way may be to accept the uncertainty in the significant level.

The optimizations are carried for both subsets and for the entire set separately and RMSE for all training and validation sets are tabulated in Table 4.2. The obtained statistics with different comparison methods are given in Table 4.3.

According to PRESS analysis, 2Lines, Yukawa, 9 wells and 11 wells models are more promising than other models. AIC favors 2Lines and Yukawa as well and also favors StepYukawa and StepYukawa-Universal models. BIC favors LJ, Yukawa, Yukawa-Universal and StepYukawa-Universal whereas MIC suggests StepYukawa-Universal, Yukawa, 2Lines and StepYukawa as more promising.

Table 4.2. RMSE with different models

	RMSE				
Model	Tr-1	Val-1	Tr-2	Val-2	overall
Linear-2580	10.55	12.16	10.10	13.11	11.26
Linear-2580+	10.52	13.14	9.94	40.00	11.25
uncon-2580+	10.03	11.79	9.77	12.73	10.84
Linear-9	11.01	12.44	10.37	13.63	11.65
2Lines	10.16	10.77	9.21	12.15	10.45
LJ	10.54	11.77	9.77	12.78	11.04
Yukawa	10.14	11.44	9.57	12.43	10.72
Yukawa-Univ	10.28	11.88	9.88	15.85	11.11
StepYukawa	9.96	10.94	9.07	13.30	10.36
StepYukawa-Univ	10.39	11.54	9.66	12.72	10.94
9wells	9.89	10.53	8.94	12.26	10.23
11wells	9.83	10.65	8.87	12.66	10.19

In the application of F-Test, 11 wells model is considered as the full model. According to the first test in Table 4.4, 9 wells is concluded to be satisfactory compared to 11 wells model. In the next application of F-Test in Table 4.5, where 9 wells model is considered as the full model this time, the StepYukawa model is better than the 9 wells model and hence it is assumed to be the full model in the next application of F-Test. According to the final application in Table 4.6, the StepYukawa model is chosen as the best model with respect to F-Test.

Taking all selection methods and physical reliabilities and aceptances of the models into account, it is concluded that the StepYukawa-Universal is the best model. However, because of the historical reasons of Linear-2580 and wide acceptance of LJ and physical background of Yukawa-Universal, the results for these models will also be documented.

Table 4.3. Comparison of models

Model	PRESS	AIC	BIC	MIC
Linear-2580	210735	5893	6066	104.4
Linear-2580+	1186262	5926	6186	138.0
uncon-2580+	198413	5905	6338	116.9
Linear-9	224473	5975	6148	192.3
2Lines	173905	5783	6129	1.7
LJ	198921	5813	5899	23.7
Yukawa	188132	5774	5948	-11.8
Yukawa-Univ	259790	5830	5922	41.2
StepYukawa	195934	5727	5987	-51.0
StepYukawa-Univ	194580	5795	5892	6.5
9wells	172449	5902	6681	127.4
11wells	180841	5960	6912	187.0

Table 4.4. First application of F-Test

Model	SSR_{Full}	df_{Full}	SSR_{Red}	df_{Red}	F^*	Acceptance
9 wells	124964	1016	126001	1050	0.25	accepted
11 wells	Full Model					

Table 4.5. Second application of F-Test

Model	SSR_{Full}	df_{Full}	SSR_{Red}	df_{Red}	F^*	Acceptance
StepYukawa	126001	1050	129142	1152	0.26	accepted
9 wells	Full Model					

Table 4.6. Final application of F-Test

Model	SSR_{Full}	df_{Full}	SSR_{Red}	df_{Red}	F^*	Acceptance
Linear-2580	129142	1152	152446	1169	12.23	rejected
Linear-9	129142	1152	163258	1169	17.90	rejected
LJ	129142	1152	146698	1186	4.60	rejected
Yukawa-Univ	129142	1152	148598	1185	5.26	rejected
Yukawa	129142	1152	138144	1169	4.72	rejected
StepYukawa-Univ	129142	1152	144087	1184	4.16	rejected
StepYukawa	Full Model					

5. RESULTS and DISCUSSION

The results are obtained by sequentially optimizing 17 sites from n-alkanes, branched alkanes, aromatics and naphthenics with StepYukawa-Universal model. After that, the optimization is continued with the LM to find an optimum for 17 sites simultaneously. The steepness factor κ and X_{lumped} are determined as 2.849 and 1.15, respectively. These universal parameters are used when optimizing other organic families. Thereafter, alcohols and phenols are optimized in which some sites are transferred directly. The ultimate vapor pressure errors and the parameters of the StepYukawa-Universal model for the whole database are tabulated. Although the descriptions for sites and some examples are given in Table C.1, for convenience, some examples are also tabulated in Table 5.1. The placeholders in the following tables are explained in Equations (5.1), (5.2), (5.3), (5.4), (5.5) and (5.6).

$$pAad = \frac{\sum_T |\%\ P^{err}_{C,T}|}{T} \tag{5.1}$$

$$pBias = \frac{\sum_T \%\ P^{err}_{C,T}}{T} \tag{5.2}$$

$$pMax = max(\%\ P^{err}_{C,T}) \tag{5.3}$$

$$rAad = \frac{\sum_T |\%\ \rho^{err}_{C,T}|}{T} \tag{5.4}$$

$$rBias = \frac{\sum_T \%\ \rho^{err}_{C,T}}{T} \tag{5.5}$$

$$rMax = max(\%\ \rho^{err}_{C,T}) \tag{5.6}$$

Table 5.1. The parameters for StepYukawa-Universal

Site	ϵ	stdErr	Example
CH_3a	54.801	1.709	*n*-Butane, *n*-Octane
CH_3b	50.747	1.651	*iso*-Butane
CH_3c	55.745	4.892	*neo*-Pentane, 2-2-diMe-Hexane
CH_3d	118.970	3.020	*p*-Xylene, Toluene
CH_3e	113.213	5.214	*o*-Xylene
CH_2a	27.077	0.070	*n*-Butane, *n*-Octane
CH_2b	72.186	5.457	Benzene-*n*-Propyl
CH_2c	29.193	1.477	*cyc*Hexane
CH_2d	79.128	5.191	Benzene-*o*-MeEt
CH_2e	29.951	4.934	*cyc*Hexane-Et, *cyc*Hexane-Propyl
CHa	6.996	1.031	*iso*-Butane
CHb	7.112	1.719	2-3-diMe-Hexane
CHc	9.317	1.659	Decalin, Me-*cyc*Hexane
C	0.020	2.250	1-1-diMe-*cyc*Hexane
Ar-CHa	29.161	1.354	Benzene
Ar-Ca	0.000	0.160	Toluene
Ar-Cb	23.880	4.480	Phenol

There are 382 different evaluations of vapor pressure for 26 straight chain alkanes. Their vapor pressure and density results are given in Table 5.2.

24 different compounds from branched alkanes are evaluated at 456 conditions and their results are given in Table 5.3.

There are 13 compounds from aromatics at 219 different points. The results are as in Table 5.4.

9 compounds from naphthenics are evaluated at 146 points and results are given in Table 5.5.

Table 5.2. n-alkanes with StepYukawa-Universal

Compound	pAad	pBias	pMax	rAad	rBias	rMax
nC_3	13.52	-2.82	33.64	4.68	-4.68	-6.37
nC_4	14.23	-14.18	-20.98	1.22	-0.50	3.23
nC_5	12.53	-12.39	-17.40	1.56	1.56	5.46
nC_6	7.33	-6.41	-11.11	2.36	2.36	5.98
nC_7	3.79	-3.65	-8.57	2.85	2.85	6.46
nC_8	6.50	4.66	21.23	3.23	3.23	6.56
nC_9	5.24	4.96	13.61	3.92	3.92	7.38
nC_{10}	6.72	6.72	22.42	4.27	4.27	7.84
nC_{11}	10.54	10.54	26.78	3.94	3.94	4.93
nC_{12}	9.65	9.65	18.36	4.25	4.25	5.28
nC_{13}	7.60	7.60	17.85	5.20	5.20	7.79
nC_{14}	8.75	8.75	17.63	4.74	4.74	6.81
nC_{15}	6.22	6.22	10.04	5.34	5.34	7.86
nC_{16}	5.01	5.01	11.60	5.47	5.47	8.14
nC_{17}	3.41	3.41	11.51	5.80	5.80	8.80
nC_{18}	2.80	0.44	7.70	5.79	5.79	8.22
nC_{19}	4.72	-0.80	10.66	6.08	6.08	9.12
nC_{20}	5.96	-4.46	-12.75	6.28	6.28	8.71
nC_{21}	5.51	-2.43	8.62	6.09	6.09	8.33
nC_{22}	10.04	-7.39	-16.89	6.28	6.28	8.99
nC_{23}	9.48	-8.05	-17.97	6.81	6.81	9.50
nC_{24}	15.48	-15.48	-21.57	5.70	5.70	6.93
nC_{25}	12.61	-12.25	-21.88	6.33	6.33	8.57
nC_{26}	13.23	-12.95	-24.46	7.07	7.07	9.47
nC_{27}	22.15	-22.15	-30.36	5.94	5.94	8.09
nC_{30}	16.55	-16.55	-31.40	6.64	6.64	8.11

There are 20 compounds from alcohols evaluated at 395 points. For different alcohols types, different hydroxyl sites are used and checked for their significance in

Table 5.3. Branched alkanes with StepYukawa-Universal

Compound	pAad	pBias	pMax	rAad	rBias	rMax
2-methyl-C_4	18.11	-18.11	-21.26	0.93	0.46	4.64
2-methyl-C_5	28.75	-28.75	-30.27	2.64	2.64	7.35
2-methyl-C_6	2.23	2.22	9.51	2.13	2.13	6.65
2-methyl-C_7	5.63	5.63	12.24	2.95	2.95	6.29
2-methyl-C_8	10.68	10.68	27.33	3.74	3.74	8.17
2-methyl-C_9	10.53	10.53	29.08	4.45	4.45	9.17
3-methyl-C_5	16.37	-16.37	-20.81	1.32	1.32	5.53
3-methyl-C_6	11.14	-11.14	-14.29	1.80	1.80	4.63
3-methyl-C_7	1.92	-1.92	-3.03	2.65	2.65	5.73
3-methyl-C_8	6.60	6.60	11.95	3.37	3.37	7.77
3-methyl-C_9	6.42	6.42	11.72	3.51	3.51	7.64
4-methyl-C_7	3.86	-3.86	-5.63	3.17	3.17	7.25
4-methyl-C_8	4.33	4.33	7.97	3.99	3.99	8.62
4-methyl-C_9	5.32	5.32	7.13	4.56	4.56	9.62
5-methyl-C_9	9.53	9.53	16.72	4.45	4.45	8.92
23-diMethyl-C_4	6.83	-6.83	-11.00	1.00	0.20	4.43
23-diMethyl-C_5	7.69	-7.69	-12.49	0.72	0.34	2.80
23-diMethyl-C_6	1.40	0.18	2.01	1.72	1.72	5.22
23-diMethyl-C_8	20.14	20.14	34.78	2.93	2.93	6.81
24-diMethyl-C_5	11.03	-11.03	-17.49	2.44	2.44	6.28
24-diMethyl-C_6	7.26	-7.26	-15.29	5.31	5.31	17.96
25-diMethyl-C_6	2.75	2.75	6.23	2.72	2.72	6.84
25-diMethyl-C_8	13.11	13.11	14.74	3.73	3.73	8.23
26-diMethyl-C_7	13.49	13.49	26.33	2.72	2.72	5.89

the reduction of RMSE. In Table 5.6, "Full" model has additional sites for different hydroxyls and "Reduced" model has only one universal hydoxyl site for all type of alcohols. As in Table 5.6, hydroxyl groups having ϵ values ranging from 100 to 310 does not make sense in terms of their physical meanings. Further, ϵ values of CH_3f

Table 5.4. Aromatics with StepYukawa-Universal

Compound	pAad	pBias	pMax	rAad	rBias	rMax
Benzene	7.18	7.18	23.89	4.22	-4.19	-5.50
Benzene-Ethyl	1.73	1.73	5.19	1.86	-0.41	5.04
Benzene-isoPropyl	18.42	-11.50	-35.59	31.27	31.27	45.89
Benzene-nPropyl	7.50	-7.50	-11.21	3.11	3.09	10.88
Benzene-nButyl	5.79	5.79	8.14	3.31	3.31	9.99
Benzene-*o*-diEt	4.89	-4.10	-7.80	2.27	-0.63	5.70
Benzene-*o*-MeEt	7.27	7.27	14.51	3.08	-2.52	-4.34
Benzene-135triMe	6.10	-5.35	-12.54	5.46	-5.46	-7.10
Toluene	9.05	-9.05	-9.88	2.95	-2.39	-4.02
o-Xylene	4.65	-4.54	-6.39	2.88	-2.40	-3.90
m-Xylene	11.15	11.15	24.40	5.84	-5.84	-6.89
p-Xylene	5.50	-5.50	-5.94	4.03	-3.92	-5.28
Naphthalene	6.41	1.21	11.66	1.89	1.24	7.18

and CH_2f of the full model are not consistent with that of other CH_3 and CH_2 sites. Although the IC statistics given in Table 5.7 slightly favor the full model, investigating this improper empiricism in the ϵ values of the full model and only a slight worsening

Table 5.5. Naphthenics with StepYukawa-Universal

Compound	pAad	pBias	pMax	rAad	rBias	rMax
*cyc*Hexane	0.66	0.08	2.20	3.74	3.74	8.72
*cyc*Hexane-Me	6.53	-6.36	-13.92	4.99	4.99	10.86
*cyc*Hexane-Ethyl	4.67	-1.11	-8.90	3.15	3.15	7.97
*cyc*Hexane-nPropyl	5.06	3.30	13.09	4.17	4.17	8.73
*cyc*Hexane-nButyl	6.90	6.72	19.92	5.28	5.28	10.30
*cyc*Hexane-11diMe	5.55	0.63	10.30	5.50	5.50	6.48
*cyc*Hexane-12cis-diMe	8.74	8.74	17.01	2.20	2.20	7.77
*cyc*Hexane-13tr-diMe	5.91	5.89	14.29	2.69	2.69	8.60
*cyc*Hexane-14tr-diMe	12.77	-12.66	-28.62	5.52	5.52	11.48

Table 5.6. Comparing the RMSE with additional sites for hydroxyls

Site	Full	Reduced
CH_3f	3.423	52.536
CH_2f	0.000	15.757
CH_2g	23.172	23.695
CHd	1.718	9.373
OHa	308.254	100.181
OHb	244.021	100.181
OHc	101.842	100.181
OHd	170.696	100.181
RMSE	18.251	19.371

Table 5.7. IC statistics for additional sites for hydroxyls

Model	AIC	BIC	MIC
Reduced	2351	2371	51.0
Full	2310	2342	8.0

in RMSE, it is concluded to have one universal site for hydroxyl groups. Hence, the parameters for sites describing the alcohols are as in Table 5.8. As can be seen in

Table 5.8. The parameters for sites in alcohols

Site	ϵ	stdErr	Example
CH_3f	52.536	4.584	Methanol
CH_2f	15.757	3.779	Ethanol
CH_2g	23.695	3.559	*n*-Butanol, *n*-Octanol
CHd	9.373	1.956	Propanediol, Pentanetriol
OH	100.181	2.764	Phenol, *n*-Butanol

Table 5.9, presence of hydroxyl groups in higher molecular weight compounds leads to an increase in the vapor pressure deviations. This behaviour suggests that an additional optimization for hydrogen bonding sites is necessary.

Table 5.9. Alcohols with StepYukawa-Universal

Compound	pAad	pBias	pMax	rAad	rBias	rMax
nC_1OH	6.12	-1.08	27.67	4.35	-4.35	-5.55
nC_2OH	7.08	-2.07	15.80	1.79	-1.75	-2.42
nC_3OH	13.02	-8.69	22.44	0.72	0.31	3.59
nC_4OH	24.74	-24.74	-28.09	3.07	3.07	7.54
nC_5OH	8.57	8.57	42.97	1.22	-0.72	2.49
nC_6OH	9.21	-9.21	-13.65	2.34	2.34	5.14
nC_7OH	6.85	2.53	33.21	1.88	1.88	5.31
nC_8OH	8.65	-5.36	17.46	4.12	4.12	8.86
nC_9OH	25.64	25.64	102.23	0.47	-0.07	1.38
$nC_{10}OH$	32.47	32.47	85.23	0.96	0.96	3.16
C_3-2-(OH)	12.05	-12.05	-15.14	2.40	-2.40	-2.55
C_4-2-(OH)	17.58	-17.58	-21.41	1.04	-0.59	1.74
C_5-2-(OH)	4.18	-4.11	-6.78	0.22	0.14	0.50
C_6-2-(OH)	10.95	10.95	15.74	0.57	0.57	1.19
C_7-2-(OH)	7.43	7.43	9.31	1.01	1.01	1.54
C_8-2-(OH)	29.52	29.52	42.85	0.66	0.66	1.34
C_9-2-(OH)	36.79	36.79	41.21	1.54	1.54	1.57
C_5-3-(OH)	28.35	-22.80	39.98	0.72	-0.72	-0.92
C_6-3-(OH)	20.24	-20.24	-26.14	0.82	0.82	1.40
C_7-3-(OH)	3.27	-1.11	-5.55	1.56	1.56	1.62

There are 200 points for 10 phenolic compounds. In Table 5.10, it is checked whether an additional hydroxyl to describe phenols are necessary. Since IC statistics in Table 5.11 supports the equivalency of both model and the ϵ values and RMSE are very close to each other as shown in Table 5.10, it is concluded to use just one universal hydroxyl site for both alcohols and phenols. Hence, the parameters for sites describing the phenols are as in Table 5.12. With these parameters, the resulting vapor pressure and density errors are as in Table 5.13.

Table 5.10. Comparing the RMSE with additional hydroxyl for phenols

Site	Full	Reduced
Ar-CHb	31.040	31.073
Ar-CHc	28.071	28.164
Ar-CHd	33.199	33.254
Ar-Cc	19.651	21.467
OH-phenol	106.067	100.181
RMSE	11.312	11.372

Table 5.11. IC statistics for additional hydroxyl for phenols

Model	AIC	BIC	MIC
Reduced	980	994	5.1
Full	980	997	5.0

Table 5.12. The parameters for sites in phenols

Site	ϵ	stdErr	Example
Ar-CHb	31.073	2.927	Phenol
Ar-CHc	28.164	2.674	Me-Phenol
Ar-CHd	33.254	3.585	*meta*-Me-Phenol
Ar-Cc	21.467	3.026	Phenol

Investigating the vapor pressure errors for a single compound, it is observed that the errors are not random but systematic. This behaviour is displayed in Figure 5.1.

The consistency of the parameters for StepYukawa-Universal is demonstrated in Table 5.14. All methyl groups are around 50-55 whereas CH_3d and CH_3e are around 113-118. This behaviour can be explained through the benzylic position of CH_3d and CH_3e. Apparently, they are much more attractive. A similar situation can be observed with the benzylic methylenes. Whereas all methylenes are around 23-29, the sites in the benzylic position are around 72-79 demonstrating stronger attractions. The methylene group in ethanol is less attractive than all other methylenes probably because of the strongly attractive hydroxyl group next to it. Such a behaviour is also observ-

Table 5.13. Phenols with StepYukawa-Universal

Compound	pAad	pBias	pMax	rAad	rBias	rMax
Phenol	10.95	-10.95	-22.17	1.43	-1.19	-1.84
Phenol-2Me	5.06	-5.04	-17.89	6.14	-6.14	-7.45
Phenol-3Me	3.87	-2.52	-10.81	5.58	-5.58	-6.81
Phenol-4Me	4.56	0.95	11.37	5.88	-5.88	-7.33
Phenol-23$diMe$	19.14	18.51	30.13	4.73	-4.73	-6.66
Phenol-24$diMe$	5.01	-4.73	-16.93	8.78	-8.78	-10.31
Phenol-25$diMe$	15.26	-15.26	-27.44	7.14	-7.14	-8.63
Phenol-26$diMe$	6.05	-0.69	-21.63	2.42	2.42	5.85
Phenol-34$diMe$	9.95	5.83	19.96	7.89	-7.89	-8.79
Phenol-35$diMe$	6.49	0.35	14.49	6.90	-6.90	-8.19

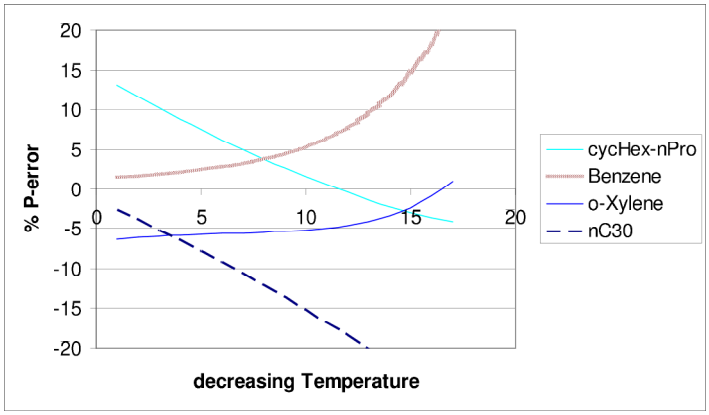

Figure 5.1. Systematic error

able with the sp^2 hybridized aromatic carbon, labeled as 'Ar-Ca', to which strongly attractive benzylic methyl and methylenes are attached. As can be seen in Table A.2, Yukawa-Universal and Lennard Jones exhibit such a consistency in parameters as well. Although some parameters of Linear-2580 seem to be improper in terms of the decay of attraction with increasing intermolecular distance, i.e. some sites like CH_3c, CH_2e, CH_2f and Ar-Cb exhibit a square well type attraction, a similar consistency to that of StepYukawa-Universal, Yukawa-Universal and LJ is also observable. This consistency of parameters is also supported by the estimated standard errors on parameters. As can be seen in Table 5.14 and in Figures 5.2 to 5.8, one-standard-error-ranges of pa-

rameters of similar sites are intersecting. These intersections are encircled by boxes in the corresponding figures. In Figure 5.2, intersecting region can be moved downwards

Table 5.14. Standard errors of SYU-parameters

Site	ϵ	stdErr
CH_3a	54.801	1.709
CH_3b	50.747	1.651
CH_3c	55.745	4.892
CH_3d	118.970	3.020
CH_3e	113.213	5.214
CH_3f	52.536	4.584
CH_2a	27.077	0.070
CH_2b	72.186	5.457
CH_2c	29.193	1.477
CH_2d	79.128	5.191
CH_2e	29.951	4.934
CH_2f	15.757	3.779
CH_2g	23.695	3.559
CHa	6.996	1.031
CHb	7.112	1.719
CHc	9.137	1.659
CHd	9.373	1.956
C	0.020	2.250
Ar-CHa	29.161	1.354
Ar-CHb	31.073	2.927
Ar-CHc	28.164	2.674
Ar-CHd	33.254	3.585
Ar-Ca	0.000	0.160
Ar-Cb	23.880	4.480
Ar-Cc	21.467	3.026
OH	100.181	2.764

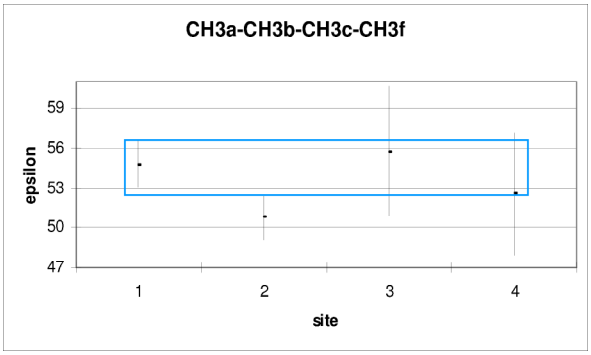

Figure 5.2. Intersection for methyls

in order to include CH_3b and ϵ for an universal methyl would be then probably around 52-53.

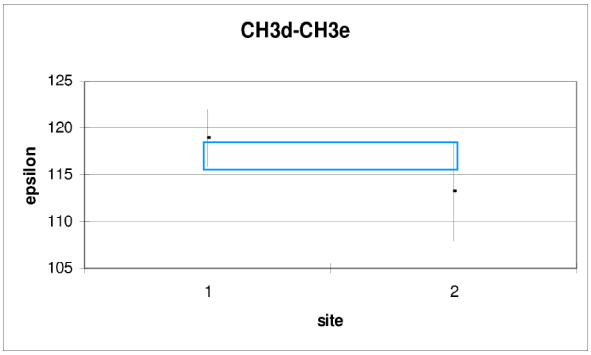

Figure 5.3. Intersection for benzylic methyls

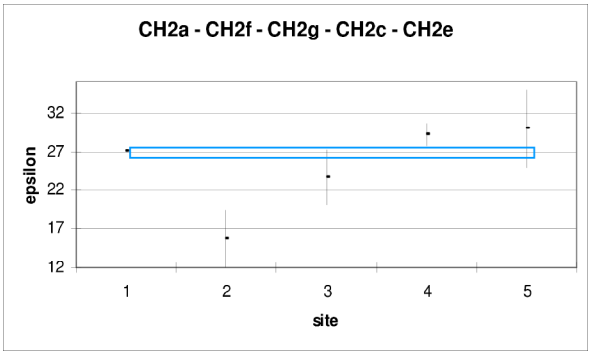

Figure 5.4. Intersection for methylenes

As can be seen in Figure 5.4, methylene site in the ethanol, namely CH_2f, is not in the intersection of other methylenes. Another important observation is that the accuracy of vapor pressure prediction is highly sensitive to CH_2a.

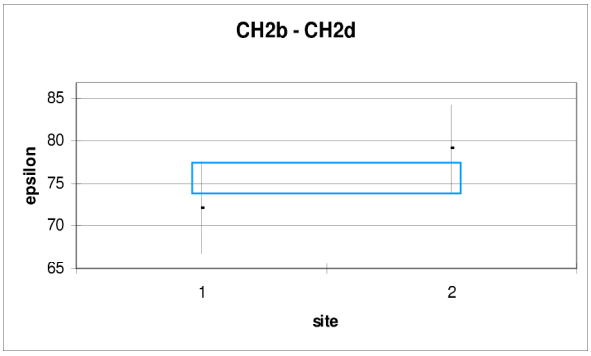

Figure 5.5. Intersection for benzylic methylenes

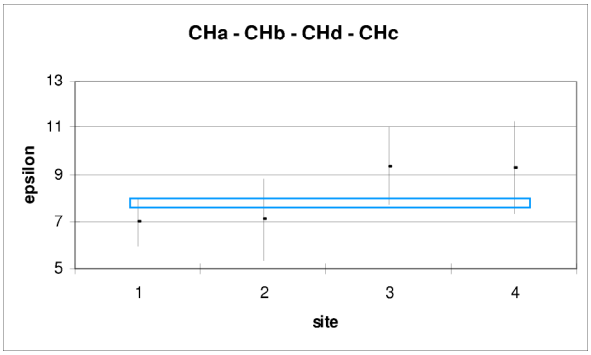

Figure 5.6. Intersection for aliphatic CH

In Table 5.15, one can observe that the decreasing tendency in the attraction from methyl (CH_3) to methyne (CH) in hydrocarbons is consistent with the LJ parameters of

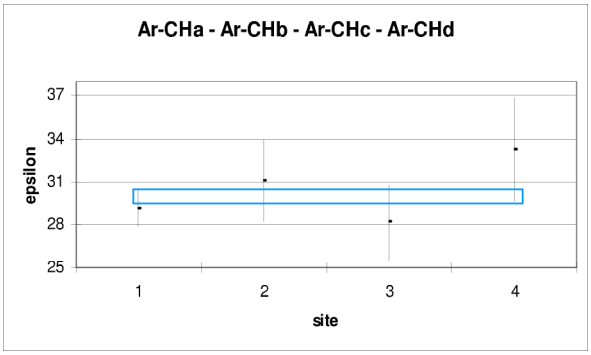

Figure 5.7. Intersection for aromatic CH

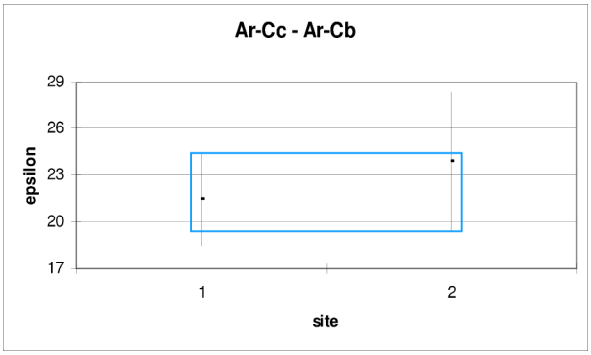

Figure 5.8. Intersection for aromatic C

other force fields like TraPPE, OPLS, PRF and AUA, even though they report higher attractions. The larger magnitudes in other potentials is likely due to larger diameters used in those models. In other words, larger diameters cause larger repulsions.

Table 5.15. LJ parameters for different force fields

Site	OPLS	TraPPE	PRF	AUA	SPEAD
CH_3	88.1	98	96	120.15	54.625
CH_2	59.4	46	57	86.29	26.842
CH	N/A	10	36	50.98	7.003

6. CONCLUSIONS and RECOMMENDATIONS

The methodology used to study thermodynamic properties of hydrocarbons and oxygenated compounds is DMD/TPT. DMD is advantageous in terms of being at least an order of magnitude faster than any conventional molecular dynamics simulation. TPT incorporates the attractive contribution to the system which was neglected during the simulation part. Using a hybrid optimization method, a combination of RRS and LM, a detailed study is carried out to obtain a model to fit the attractive potential with maximal reliability and minimal number of parameters. Besides the selected model, namely StepYukawa-Universal, parameters for 3 promising models are also obtained. Although the LJ parameters of SPEAD force field are smaller than that of OPLS, AUA, PRF and TraPPE, the decreasing tendency in the attraction from methyl to methyne in hydrocarbons is also observable.

The accuracy of SPEAD can be better understood if the scope of the database of compounds and the range of the reduced temparatures are taken into account. The vapor pressure predictions are performed at reduced temperatures approximately between 0.480 and 0.900. Such a temperature range covers a very broad range of industrial applications. Altough the accuracy in the vapor prediction is an indication of the consistency of the parameters for models, the consistency is further supported by the similarities in the parameters for similar sites.

The SPEAD force field with SYU, YU, LJ or 2580 is currently superior to other vapor prediction methods in terms of the scope and accuracy. An additional optimization for hydrogen bonding will improve the accuracy further. If similar sites can be reduced to single universal sites as it is done for hydroxyls in alcohols and phenols, necessary number of parameters can drastically decrease without a significant loss in accuracy. The optimization method applied to current database can also be applied to new organic families like amines, amides, ketones, esters, ethers and sulfides to extend the database.

APPENDIX A: Results for Other Promising Models

Lennard-Jones is a widely accepted and used model. Hence, the variables for this model is documented in this part. The variables for Yukawa-Universal and Linear-2580 are also reported here because of their potential use. SYU, YU, LJ and 2580 in Tables A.1 and A.2 denote StepYukawa-Universal, Yukawa-Universal, Lennard Jones and Linear-2580, respectively.

Table A.1. The RMSE for different models

Organic Family			RMSE			
name	nComps	nData	SYU	YU	LJ	2580
n-Alkanes	26	382	11.51	11.74	11.78	12.01
br-Alkanes	24	456	11.76	11.90	11.78	11.97
Aromatics	13	219	9.42	9.82	9.58	10.41
Naphthenics	9	146	8.66	8.37	8.39	7.53
Alcohols	20	395	19.37	19.04	19.12	19.37
Phenols	10	200	11.37	10.66	11.38	10.77
Overall	102	1798	13.30	13.23	13.28	13.42

Table A.2. The parameters for different models

	SYU			YU		LJ	2580	
Site	ϵ	X_{lumped}	κ	ϵ	κ	ϵ	ϵ_1	ϵ_4
CH_3a	54.801	1.15	2.849	57.947	2.104	54.625	51.921	0.030
CH_3b	50.747	1.15	2.849	53.593	2.104	50.329	47.771	0.324
CH_3c	55.745	1.15	2.849	57.897	2.104	53.941	44.297	44.282
CH_3d	118.970	1.15	2.849	122.244	2.104	118.529	100.553	79.066
CH_3e	113.213	1.15	2.849	115.875	2.104	112.781	101.235	22.229
CH_3f	52.536	1.15	2.849	53.992	2.104	51.250	47.118	11.105
CH_2a	27.077	1.15	2.849	28.504	2.104	26.842	25.206	0.254
CH_2b	72.186	1.15	2.849	73.489	2.104	71.772	64.270	9.313
CH_2c	29.193	1.15	2.849	30.752	2.104	28.972	26.476	6.473
CH_2d	79.128	1.15	2.849	80.358	2.104	78.636	67.776	29.899
CH_2e	29.951	1.15	2.849	31.531	2.104	29.568	24.611	24.605
CH_2f	15.757	1.15	2.849	15.736	2.104	14.842	12.982	11.239
CH_2g	23.695	1.15	2.849	23.962	2.104	22.646	20.895	0.022
CHa	6.996	1.15	2.849	7.234	2.104	7.003	5.914	4.246
CHb	7.112	1.15	2.849	7.350	2.104	7.142	5.888	5.814
CHc	9.137	1.15	2.849	9.692	2.104	9.255	7.559	7.559
CHd	9.373	1.15	2.849	3.115	2.104	8.907	7.472	5.398
C	0.020	1.15	2.849	0.038	2.104	0.120	0.171	0.075
Ar-CHa	29.161	1.15	2.849	30.676	2.104	28.931	27.184	0.183
Ar-CHb	31.073	1.15	2.849	33.066	2.104	30.906	29.619	0.256
Ar-CHc	28.164	1.15	2.849	29.418	2.104	27.821	25.974	2.721
Ar-CHd	33.254	1.15	2.849	35.190	2.104	33.017	31.617	0.613
Ar-Ca	0.000	1.15	2.849	0.030	2.104	0.000	0.028	0.002
Ar-Cb	23.880	1.15	2.849	25.152	2.104	23.685	19.555	19.549
Ar-Cc	21.467	1.15	2.849	20.783	2.104	20.346	17.382	1.129
OH	100.181	1.15	2.849	109.146	2.104	102.761	98.474	1.597

APPENDIX B: Reduced Temperature Ranges

Accuracy of the vapor pressure prediction depends highly on the temperature ranges under consideration. The maximum deviations in the vapor pressure predictions are observed at temperatures near to critical and melting temperatures. Generally, organic compounds are in solid state below reduced temperatures of 0.45. Hence, the reduced temperatures approximately from 0.480 to 0.900 cover a very broad range for industrial applications. The range of reduced temparatures for each compound in the database are given in the following tables.

Table B.1. Reduced temperature ranges for phenols

Compound	nData	T_r^{min}	T_r^{max}
Phenol	20	0.500	0.900
Phenol-2Me	20	0.502	0.888
Phenol-3Me	20	0.510	0.880
Phenol-4Me	20	0.510	0.880
Phenol-23$diMe$	20	0.505	0.885
Phenol-24$diMe$	20	0.510	0.890
Phenol-25$diMe$	20	0.510	0.890
Phenol-26$diMe$	20	0.515	0.898
Phenol-34$diMe$	20	0.500	0.875
Phenol-35$diMe$	20	0.503	0.895

Table B.2. Reduced temperature ranges for n-alkanes

Compound	nData	T_r^{min}	T_r^{max}
nC_3	21	0.400	0.900
nC_4	21	0.400	0.900
nC_5	19	0.450	0.900
nC_6	19	0.450	0.900
nC_7	14	0.575	0.900
nC_8	19	0.475	0.900
nC_9	18	0.475	0.900
nC_{10}	18	0.475	0.900
nC_{11}	14	0.475	0.800
nC_{12}	13	0.500	0.800
nC_{13}	14	0.500	0.825
nC_{14}	13	0.500	0.800
nC_{15}	13	0.525	0.825
nC_{16}	14	0.525	0.850
nC_{17}	14	0.525	0.850
nC_{18}	12	0.540	0.835
nC_{19}	13	0.535	0.850
nC_{20}	12	0.550	0.825
nC_{21}	11	0.565	0.820
nC_{22}	13	0.560	0.865
nC_{23}	13	0.578	0.880
nC_{24}	8	0.572	0.746
nC_{25}	12	0.567	0.837
nC_{26}	12	0.586	0.855
nC_{27}	8	0.580	0.840
nC_{30}	24	0.536	0.858

Table B.3. Reduced temperature ranges for branched alkanes

Compound	nData	T_r^{min}	T_r^{max}
2-methyl-C_4	19	0.450	0.900
2-methyl-C_5	19	0.450	0.900
2-methyl-C_6	19	0.473	0.900
2-methyl-C_7	19	0.460	0.900
2-methyl-C_8	19	0.450	0.900
2-methyl-C_9	19	0.473	0.900
3-methyl-C_5	19	0.482	0.900
3-methyl-C_6	19	0.495	0.900
3-methyl-C_7	19	0.482	0.900
3-methyl-C_8	19	0.481	0.900
3-methyl-C_9	19	0.482	0.900
4-methyl-C_7	19	0.473	0.900
4-methyl-C_8	19	0.473	0.900
4-methyl-C_9	19	0.473	0.900
5-methyl-C_9	19	0.473	0.900
23-diMethyl-C_4	19	0.450	0.900
23-diMethyl-C_5	19	0.450	0.900
23-diMethyl-C_6	19	0.473	0.900
23-diMethyl-C_8	19	0.450	0.900
24-diMethyl-C_5	19	0.480	0.900
24-diMethyl-C_6	19	0.483	0.965
25-diMethyl-C_6	19	0.450	0.900
25-diMethyl-C_8	19	0.450	0.900
26-diMethyl-C_7	19	0.450	0.900

Table B.4. Reduced temperature ranges for aromatics

Compound	nData	T_r^{min}	T_r^{max}
Benzene	17	0.500	0.900
Benzene-*Ethyl*	17	0.500	0.900
Benzene-*isoPropyl*	17	0.500	0.900
Benzene-*nPropyl*	17	0.500	0.900
Benzene-*nButyl*	17	0.500	0.900
Benzene-*o-diEt*	17	0.500	0.900
Benzene-*o-MeEt*	17	0.500	0.900
Benzene-135*triMe*	15	0.550	0.900
Toluene	18	0.475	0.900
o-Xylene	17	0.500	0.900
m-Xylene	17	0.500	0.900
p-Xylene	17	0.500	0.900
Naphthalene	16	0.525	0.900

Table B.5. Reduced temperature ranges for naphthenics

Compound	nData	T_r^{min}	T_r^{max}
*cyc*Hexane	17	0.500	0.900
*cyc*Hexane-*Me*	17	0.500	0.900
*cyc*Hexane-*Ethyl*	17	0.500	0.900
*cyc*Hexane-*nPropyl*	17	0.500	0.900
*cyc*Hexane-*nButyl*	17	0.500	0.900
*cyc*Hexane-11*diMe*	10	0.500	0.725
*cyc*Hexane-12*cis-diMe*	17	0.500	0.900
*cyc*Hexane-13*tr-diMe*	17	0.500	0.900
*cyc*Hexane-14*tr-diMe*	17	0.500	0.900

Table B.6. Reduced temperature ranges for alcohols

Compound	nData	T_r^{min}	T_r^{max}
nC_1OH	21	0.400	0.900
nC_2OH	19	0.450	0.900
nC_3OH	18	0.475	0.900
nC_4OH	19	0.450	0.900
nC_5OH	18	0.475	0.900
nC_6OH	18	0.475	0.900
nC_7OH	18	0.475	0.900
nC_8OH	19	0.450	0.900
nC_9OH	19	0.450	0.900
$nC_{10}OH$	19	0.450	0.900
C_3-2-(OH)	33	0.557	0.818
C_4-2-(OH)	33	0.593	0.865
C_5-2-(OH)	33	0.575	0.700
C_6-2-(OH)	27	0.543	0.708
C_7-2-(OH)	9	0.555	0.698
C_8-2-(OH)	33	0.582	0.751
C_9-2-(OH)	2	0.551	0.560
C_5-3-(OH)	24	0.438	0.695
C_6-3-(OH)	7	0.500	0.693
C_7-3-(OH)	6	0.498	0.707

APPENDIX C: Descriptions of the Sites

Table C.1. Description of sites

Site	Description	Example
CH_3a	methyl end groups of n-Alkanes	*n*-Butane, *n*-Octane
CH_3b	methyl bonded to a CH	*iso*-Butane
CH_3c	methyl bonded to a C	*neo*-Pentane, 2-2-diMe-Hexane
CH_3d	methyl bonded to an aromatic ring	*p*-Xylene, Toluene
CH_3e	methyl ortho-bonded to an aromatic ring	*o*-Xylene
CH_3f	methyl in Methanol	Methanol
CH_2a	methylene in n-Alkanes	*n*-Butane, *n*-Octane
CH_2b	methylene bonded to an aromatic ring	*n*-PropylBenzene
CH_2c	methylene in a non-aromatic ring	*cyc*Hexane
CH_2d	methylene ortho-bonded to an aromatic ring	*ortho*-MeEt-Benzene
CH_2e	methylene bonded to a non-aromatic ring	Et-*cyc*Hexane, Propyl-*cyc*Hexane
CH_2f	methylene in Ethanol	Ethanol
CH_2g	methylene in n-Alcohols	*n*-Butanol, *n*-Octanol
CHa	CH in branched alkanes	*iso*-Butane
CHb	CH next to another CH in branched alkanes	2-3-diMe-Hexane
CHc	CH bonded to a non-aromatic ring	Decalin, Me-*cyc*Hexane
CHd	CH in a branched alcohol	Propanediol, Pentanetriol
C	C in a non aromatic ring	1-1-diMe-*cyc*Hexane
Ar-CHa	CH in an aromatic ring	Benzene
Ar-CHb	CH ortho to OH in Phenols	Phenol
Ar-CHc	CH ortho to CH_3 in Phenols	Me-Phenol
Ar-CHd	CH between CH_3 and OH in Phenols	*meta*-Me-Phenol
Ar-Ca	C in an aromatic ring	Toluene
Ar-Cb	C in fused aromatic rings	Napthalene
Ar-Cc	C in an aromatic ring bonded to an oxygen	Phenol
OH	hydroxyl group in alcohols and phenols	Phenol, *n*-Butanol

REFERENCES

1. Cui, J. and J. R. Elliott Jr., "Phase envelopes for variable width square well chain fluids", *Journal of Chemical Physics*, Vol. 114, No. 16, pp. 7283-7290, April 2001.

2. Jensen, T., A. Fredenslund and R. Rasmussen, "Pure Component vapor pressures using UNIFAC group contribution", *Ind. Eng. Chem. Fundam.*, Vol. 20, No. 3, pp. 239-246, 1981.

3. Yair, O. B. and A. Fredenslund, "Extension of the UNIFAC group contribution method for the prediction of pure-component vapor pressures", *Ind. Eng. Chem. Process Des. Dev.*, Vol. 22, No. 3, pp. 433-436, 1983.

4. Asher, W. E., J. F. Pankow, G. B. Erdakos and J. H. Seinfeld, "Estimating the vapor pressures of multi-functional oxygen-containing organic compounds using group contribution methods", *Atmospheric Envi.*, Vol. 36, No. 9, pp. 1483-1498, 2002.

5. Bureau, N., J. Jose, I. Mokbel and J.-C. deHemptinne, "Vapor pressure measurements and prediction for heavy esters", *J. Chem. Thermo.*, Vol. 33, No. 11, pp. 1485-1498, November 2001.

6. Constantinou, L. and R. Gani, "New group contribution method for estimating properties of pure compounds", *AIChE J.*, Vol. 40, No. 10, pp. 1697-1710, October 1994.

7. Joback, K. G. and R. C. Reid, "Estimation of pure-component properties from group contributions", *Chem. Eng. Commun.*, Vol. 57, pp. 233-243, 1987.

8. Lee, B. I. and M. G. Kesler, "A generalized thermodynamic correlation based on three-parameter corresponding states", *AIChE J.*, Vol. 21, No. 3, pp. 510-527, May 1975.

9. Elliott, J. R. Jr. and R. N. Natarajan, "Extension of the Elliott-Suresh-Donohue Equation of State to Polymer Solutions", *Ind. Chem. Eng. Res.*, Vol. 41, No. 5, pp. 1043-1050, 2002.

10. Martin, M. G. and J. I. Siepmann, "Transferable Potentials for Phase Equilibria. 1. United-Atom Description of n-Alkanes", *J. Phys. Chem. B*, Vol. 102, No. 14, pp. 2569-2577, 1998.

11. Chen, B., J. J. Potoff and J. I. Siepmann, "Monte Carlo Calculations for Alcohols and Their Mixtures with Alkanes. Transferable Potentials for Phase Equilibria. 5. United-Atom Description of Primary, Secondary, and Tertiary Alcohols", *J. Phys. Chem. B*, Vol. 105, No. 15, pp. 3093-3104, 2001.

12. Bourasseau, E., M. Haboudou, A. Boutin, A. H. Fuchs and P. Ungerer, "New optimization method for intermolecular potentials: Optimization of a new anisotropic united atom potential for olefins: Prediction of equilibrium properties", *Journal of Chemical Physics*, Vol. 118, No. 7, pp. 3020-3034, February 2003.

13. Ünlü, Ö., N. H. Gray, Z. N. Gerek and J. R. Elliott, "Transferable Step Potentials for the Straight-Chain Alkanes, Alkenes, Alkynes, Ethers and Alcohols", *Ind. Eng. Chem. Res.* Vol. 43, No. 7, pp. 1788-1793, 2004.

14. Ungerer, P., C. le Beauvais, J. Delhommelle, A. Boutin, B. Rousseau and A. H. Fuchs, "Optimization of the anisotropic united atoms intermolecular potential for n-alkanes", *Journal of Chemical Physics*, Vol. 112, No. 12, pp. 5499-5510, March 2000.

15. Ahunbay, M. G., S. Kranias, V. Lachet and P. Ungerer, "Prediction of thermodynamic properties of heavy hydrocarbons by Monte Carlo simulation", *Fluid Phase Equilibria*, Vol. 228-229, pp. 311-319, February 2005.

16. Cui, J. and J. R. Elliott, "Phase diagrams for a multistep potential model of n-alkanes by discontinuous molecular dynamics and thermodynamic perturbation

theory", *Journal of Chemical Physics*, Vol. 116, No. 19, pp. 8625-8631, May 2002.

17. Barker, J. A. and D. Henderson, "Perturbation Theory and Equation of State for Fluids: The Square-Well Potential", *Journal of Chemical Physics*, Vol. 47, No. 8, pp. 2856-2861, October 1967.

18. Hu, L., J. Cui, H. Rangwalla and J. R. Elliott, "Vapor-liquid equilibria of vibrating square well chains", *Journal of Chemical Physics*, Vol. 111, No. 3, pp. 1293-1301, July 1999.

19. Chapela, G. A., L. E. Scriven and H. T. Davis, "Molecular dynamics for discontinuous potential. IV.Lennard-Jonesium", *Journal of Chemical Physics*, Vol. 91, No. 7, pp. 4307-4313, October 1989.

20. Lin, B. and D. C. Miller, "Tabu search algorithm for chemical process optimization", *Computers and Chemical Enigineering*, Vol. 28, pp. 2287-2306, 2004.

21. Mitchell, M., *An Introduction to Genetic Algorithms*, The MIT Press, Cambridge, 1998.

22. Kirkpatrick, S., D. C. Gelatt, and M. P. Vechhi, "Optimization by Simulated Annealing", *Science*, Vol. 220, No. 4598, pp. 671-680, May 1983.

23. Lin, B., S. Chavali, K. Camarda and D. C. Miller, "Computer-aided molecular design using Tabu search", *Computers and Chemical Enigineering*, Vol. 29, pp. 337-347, 2005.

24. Luenberger, D. G., *Introduction to Linear and Nonlinear Programming*, Addison-Wesley Pub. Co., Reading, 1973.

25. Garbow, B. S., K. E. Hillstrom and J. J. More, MINPACK Project, Argonne National Laboratory, Chicago, IL, 1980.

26. Gill, P. E. and W. Murray, *Minimization subject to bounds on the variables*, Report

NAC 71, National Physics Laboratory, England, 1976.

27. Microsoft Developer Studio, *Fortran PowerStation 4.0*, Microsoft Corporation, 1994-1995.

28. Kan, A. H. and G. T. Timmer, "Stochastic global optimization methods part I: Clustering methods", *Mathematical Programming*, Vol. 39, pp. 27-56, 1987.

29. Draper, N. R. and H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1981.