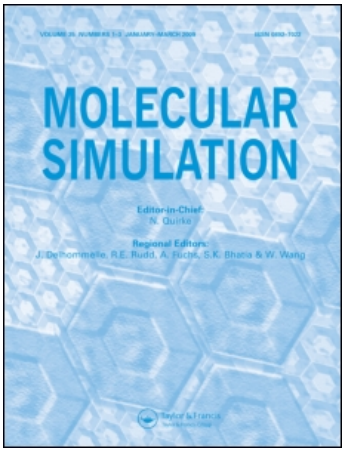

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Inferring transferable intermolecular potential models

Sinan Ucyigitler[a]; Mehmet C. Camurdan[a]; Metin Turkay[b]; J. Richard Elliott[c]
[a] Chemical Engineering Department, Bogazici University, Istanbul, Turkey [b] Industrial Engineering Department, Koc University, Sariyer, Turkey [c] Chemical and Biomolecular Engineering Department, University of Akron, Akron, USA

To cite this Article Ucyigitler, Sinan , Camurdan, Mehmet C. , Turkay, Metin and Richard Elliott, J.(2008) 'Inferring transferable intermolecular potential models', Molecular Simulation, 34: 2, 147 — 154
To link to this Article: DOI: 10.1080/08927020801930612
URL: http://dx.doi.org/10.1080/08927020801930612

Molecular Simulation
Vol. 34, No. 2, February 2008, 147–154

Taylor & Francis
Taylor & Francis Group

Inferring transferable intermolecular potential models

Sinan Ucyigitler[a], Mehmet C. Camurdan[a], Metin Turkay[b] and J. Richard Elliott[c]*

[a]*Chemical Engineering Department, Bogazici University, Istanbul, Turkey;* [b]*Industrial Engineering Department, Koc University, Sariyer, Turkey;* [c]*Chemical and Biomolecular Engineering Department, University of Akron, Akron, USA*

(Received 30 December 2007; final version received 7 January 2008)

Discontinuous molecular dynamics is combined with thermodynamic perturbation theory to provide an efficient basis for characterising molecular interactions based on vapour pressure and liquid density data. Several prospective potential models are discretised to permit treatment by Barker–Henderson perturbation theory. The potentials are characterised by 11 wells ranging over radial distances from the site diameter to three times that diameter. Considered potential models include the Lennard-Jones (LJ), square-well, Yukawa (Yuk) and multi-line potentials, and their combinations. The optimal model is found to be a combination of square-well and Yuk potentials, with the switch position and Yuk decay set to universal values. This model provides average vapour pressure deviations of less than 10% for a database of 86 aliphatic, aromatic and naphthenic compounds. The LJ potential provides the least competitive accuracy. Considering statistical information criteria facilitates the identification of the optimal model.

Keywords: transferable potentials; molecular dynamics; perturbation theory; vapour pressure

1. Introduction

Characterising interaction potentials is fundamental to molecular modelling and molecular design. In principle, all classical properties can be computed from Newton's laws once the interaction potentials are defined. Properties of interest include density, solubility parameter, vapour pressure, diffusivity, thermal conductivity, viscosity, volatility, adsorption and liquid distribution coefficient, among others. Since estimates of so many properties are affected, it is crucial to develop an accurate methodology for characterising the potentials. Furthermore, the methodology must be capable of evolving as more data and more capability become available. For example, the reliability of viscosity prediction by simulation is relatively limited at this time, but may be improved soon. The interaction potentials may need refinement when that capability is added. Other future capabilities might include absorption, distribution, metabolization and excretion (ADME) properties, Young's modulus, melting temperature, or environmental compliance properties.

Characterising and continuously improving intermolecular potential models requires efficient evaluation of the impacts of the potential on the properties. Several years ago, we observed that thermodynamic perturbation theory (TPT) yields quantitative accuracy when applied to discontinuous molecular dynamics (DMD) simulations of step potentials [9]. In this context, the step depths act as perturbations, so they can be varied instantly without repetitive simulations. At that time, we tested a relatively small number of stepwise characterisations for a relatively small number of compounds. Presently, we have a much larger database with the capability to test many more characterisations of the shape of the potential. Hence, we would like to revisit the analysis of optimised transferable potentials at this time. Specifically, we would like to consider the optimal step depths for 11 independent wells per site type ranging from σ to 3σ, for 18 site types characterising 86 compounds.

In principle, this optimisation involves 198 decision variables (a.k.a. adjustable parameters), all varying nonlinearly and highly coupled. Preliminary attempts to resolve the decision variables simultaneously led to variability in the optimised parameters depending on the initial guess, although, the average deviations were small in all cases [11]. In that study, the number of wells was restricted to four with step changes at $r/\sigma = \{1,1.2,1.5, 1.8,2.0\}$, where r is the radial distance and σ is the site diameter. We refer to this form of potential as the 2580 potential. This experience led to development of a linear relationship between step depths (Lin2580) in order to reduce the number of parameters to three per site type: the diameter, the depth of the inner well, and the depth of the outer well [24]. Recently, Ucyigitler et al. [23] have developed a combined stochastic/deterministic optimisation algorithm that converges more robustly. By applying the new optimisation algorithm to the larger database that

*Corresponding author. Email: jelliott@uakron.edu

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020801930612
http://www.informaworld.com

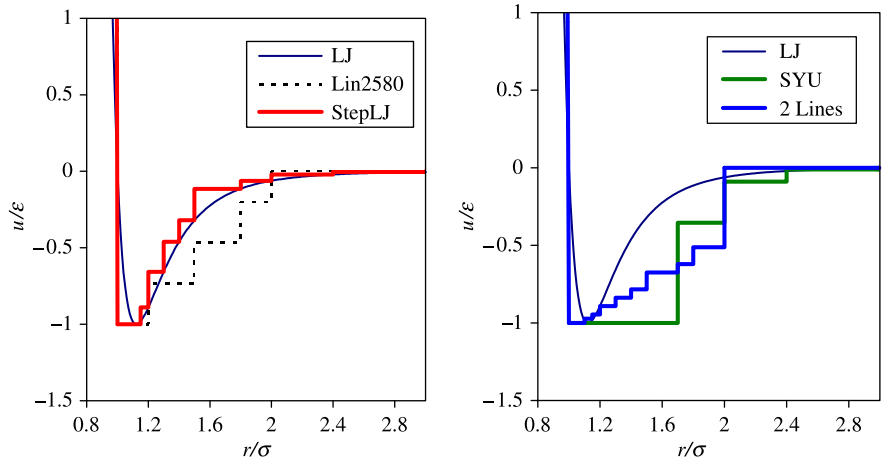

Figure 1. Optimal discretised potential functions compared to the continuous LJ potential. (a) Step LJ and Lin2580: ε and σ are decision variables for all potentials, ε_4 is an additional decision variable for Lin2580. (b) SYU and 2Lines: λ_w is an additional decision variable for SYU, but a single value applies to all site types; λ_w, ε_w, and $\varepsilon(2.0)$ are additional decision variables for the 2Lines model.

has developed over time, it is possible to unambiguously evaluate many alternative characterisations of the potential.

It is most instructive to analyse a large database with minimal decision variables when characterising potentials. In this context, we focus initially on a broad group of hydrocarbons, a database comprising 86 compounds with 18 distinguishable site types. We would like to test multiple conceptions of the potential function to see which provides the greatest accuracy. For example, the Lennard-Jones (LJ) potential is often applied in molecular modelling. The Yukawa (Yuk) and square well (SW) potentials also have their adherents. It is also possible to combine potentials. For example, a single step potential (or square well) combined with a Yuk decay (SwYukawa) could be conceived, with implications for analytical Fourier transformation. In the present work, we focus on the SwYukawa potential with universal value for the SW width, which we abbreviate as SYU. In addition to a linear step potential like the Lin2580, we would like to consider a combination of two lines (2line). For all decaying potentials, the steps are inscribed within the interpolated values of the potential, as illustrated in Figure 1. Finally, it is possible to allow all 11 wells to vary independently. In each case, we would like to evaluate the improvement in the average deviation relative to the added number of parameters.

Identifying the most effective potential model requires an objective statistical criterion. The obvious choice would be to simply choose the model with minimal root mean square error (RMSE). On the other hand, one should choose the model with fewer decision variables, among two models with roughly 'the same' RMSE. The question then becomes one of establishing what difference is statistically meaningful. This question must be analysed, while recognising that the assumption of transferability has inherent error of roughly 10%. We have considered five different criteria in our analysis. These include the *F*-test and information criteria (IC).

This particular study also focuses on the deviations with respect to vapour pressure, rather than a composite of all possible properties. Previous works have shown that the vapour pressure is more sensitive to assumptions about the potential than liquid density is. Furthermore, the liquid density is primarily a function of the site diameters. We do not vary the site diameters in this study and we do not include the liquid density in the RMSE. In this context, the RMSE is better designated by RMSEp, to emphasise that it simple the vapour pressure deviation. The diameters that we apply are those that were determined previously for the Lin2580 potential. The impact of assuming these constant diameters is apparent in the LJ potential, as discussed below. Further work beyond this present manuscript shows that the LJ RMSEp is not significantly altered by optimising the diameters specifically for the LJ model.

The paper is organised as follows. We begin by briefly outlining the theory and simulation methodology. This section is kept brief by referring extensively to previous work. The third section reviews the statistical criteria. The fourth section summarises the results and discussion. Finally, conclusions are reviewed.

2. Theory and simulation

We have been exploring the combination of DMD with TPT for several years [4,10,15,24]. The following discussion provides a brief outline of the results from this previous literature. To abbreviate, we refer to the methodology as step potential equilibria and discontinuous

Table 1. Details of the hydrocarbon database.

Family	# Compounds	Nature
n-Alkanes	26	C3—C30 (omitting C28,C29)
Branched alkanes	26	C4—C9: 15-Methyl, 9-DiMethyl, 2-Ethyl
Alkenes	17	C4—C8: 4-*cis*, 4-*trans*, 5-áOlefins, 3-dienes
Aromatics	10	C6—C9: 5-mono,di,trimethyl, 3-alkyl, 3-fused rings
Naphthenics	7	C6—C9: 5-alkyl, 2-dimethyl

Summary: 86 compounds, 18 site types, 1500 vapour pressures and 1500 densities.

molecular dynamics (SPEADMD). The TPT component was pioneered by the classic work of Barker and Henderson (BH; [2]). The basic idea of the BH perturbation theory is that the attractive potential can be divided into a series of wells, each well small enough that the energy of each step can be treated as a constant. Similarly, the Helmholtz free energy can be decomposed into a reference contribution and an attractive contribution in the form of a power series in β, where $\beta \equiv 1/k_B T$, $k_B \equiv$ Boltzmann's constant. The coefficients of β^n are referred to as the nth order perturbation terms. Therefore, the Helmholtz free energy can be written as:

$$\begin{aligned}\frac{A - A^{ig}}{Nk_BT} &= \frac{A_0 - A^{ig}}{Nk_BT} + \frac{\beta}{N}\sum_i\sum_j\sum_m \langle N_{ijm}\rangle u_{ijm} \\ &\quad - \frac{\beta^2}{2N}\sum_i\sum_j\sum_k\sum_l\sum_m\sum_n (\langle N_{ijm}N_{k\ln}\rangle \\ &\quad - \langle N_{ijm}\rangle\langle N_{k\ln}\rangle) u_{ijm}u_{k\ln} + O(\beta^3) \\ &= \frac{A_0 - A^{ig}}{Nk_BT} + \frac{A_1}{T} + \frac{A_2}{T^2} + O(\beta^3) \end{aligned} \quad (1)$$

where, for example, u_{ijm} designates the attractive energy in the mth well between the ith and jth site types. N_{ijm} is the number of pairs of interactions obtained from the reference fluid simulation. $\langle\ \rangle$ Denotes an ensemble average of the reference fluid. In all cases, we tabulate the averages in 11 wells with step transitions at $r/\sigma = 1.0$, 1.1, 1.15, 1.2, 1.3, 1.4, 1.5, 1.7, 1.8, 2.0, 2.4, 3.0.

Note that $A_i = A_i(\eta)$ where η is the packing fraction (i.e. the volume fraction occupied by molecules). The functional forms for interpolating the simulation results in terms of the $\{A_i\}$ are given by Gray et al. and references cited there [15]. They are reiterated below:

$$\frac{A_0 - A^{ig}}{Nk_BT} = \int_0^\eta \frac{Z_0 - 1}{\eta}\,d\eta \quad (2)$$

$$Z_0 = \frac{1 + z_1\eta + z_2\eta^2 + z_3\eta^3}{(1-\eta)^3} \quad (3)$$

$$A_1 = a_{11}\eta + a_{12}\eta^2 + a_{13}\eta^3 + a_{14}\eta^4 \quad (4)$$

$$A_2 = \frac{a_{21}\eta + a_{22}\eta^2 + a_{23}\eta^3 + a_{24}\eta^4}{(1 + 500\eta^4)}. \quad (5)$$

The BH formalism can be highly accurate as long as the packing fraction is greater than $\sim$0.3. This was observed for spherical molecular models long ago [3]. Cui and Elliott's work with square-well chains showed that the BH formalism also provided quantitative accuracy for vibrating chain molecular models over the same range of packing fraction [10]. Fundamentally, the implication is that the distributions and fluctuations in the reference fluid are sufficient to provide an accurate representation of the thermodynamics of the full potential at high density. Practically, the configurational internal energy varies linearly with reciprocal temperature according to the TPT result. The accuracy of this assertion has been tested repeatedly for many chain lengths. It is reliable as long as the packing fraction (η) is greater than 0.3.

To clarify the definition of each potential, it is necessary to specify the procedure for identifying the step depths and widths at each position in radial distance. The step depths are given for each potential in Table 2. For the Lin2580 potentials, ε_2, ε_3 are interpolated:

$$\varepsilon_2 = \frac{2\varepsilon_1 + \varepsilon_4}{3}; \quad \varepsilon_3 = \frac{\varepsilon_1 + 2\varepsilon_4}{3}. \quad (6)$$

For the Lin9well potential, intermediate wells are interpolated as:

$$\varepsilon_m = \varepsilon_9 + \frac{[(\varepsilon_1 - \varepsilon_9)\times(2 - \lambda_m)]}{2 - \lambda_1}. \quad (7)$$

For the 2Line potential switching at $\lambda_w = 1.4$, intermediate wells are interpolated according to

$$\begin{aligned}\varepsilon_m &= \varepsilon_5 + \frac{[(\varepsilon_1 - \varepsilon_5)\times(1.4 - \lambda_m)]}{(1.4 - \lambda_1)}, \quad m < 5; \\ \varepsilon_m &= \varepsilon_9 + (\varepsilon_5 - \varepsilon_9)\times\frac{(2 - \lambda_m)}{(2 - \lambda_5)}, \quad m > 4.\end{aligned} \quad (8)$$

For the unconstrained potentials, all well depths are treated as freely varying.

Table 2. Detailed specifications of each potential model.

Model	Vars	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}
Lin2580	$\sigma,\varepsilon_1,\varepsilon_4$	ε_1	ε_1	ε_1	ε_2	ε_2	ε_2	ε_3	ε_3	ε_4	0	0
LJ	σ,ε	ε_1	ε_1	ε_3	ε_4	ε_5	ε_6	ε_7	ε_8	ε_9	ε_{10}	ε_{11}
SYU	$\sigma,\varepsilon,(\lambda^U,\kappa^U)$	ε_1	ε_1	ε_3	ε_4	ε_5	ε_6	ε_7	ε_8	ε_9	ε_{10}	ε_{11}
Lin2580 +	$\sigma,\varepsilon_1,\varepsilon_4,\varepsilon_5$	ε_1	ε_1	ε_1	ε_2	ε_2	ε_2	ε_3	ε_3	ε_4	ε_5	ε_5
Lin9well	$\sigma,\varepsilon_1,\varepsilon_9$	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ε_7	ε_8	ε_9	0	0
2Line	$\sigma,\varepsilon_1,\varepsilon_4,\varepsilon_9$	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ε_7	ε_8	ε_9	0	0
Uncon2580 +	$\sigma,\varepsilon_1-\varepsilon_5$	ε_1	ε_1	ε_1	ε_2	ε_2	ε_2	ε_3	ε_3	ε_4	ε_5	ε_5
Uncon11well	$\sigma,\varepsilon_1-\varepsilon_{11}$	ε_1	ε_2	ε_3	ε_4	ε_5	ε_6	ε_7	ε_8	ε_9	ε_{10}	ε_{11}

For the other potentials (LJ, SYU), the well depths after the first well are set at the value of the potential where the range of the step takes its maximum value. For example, ε_3 for the LJ potential is $u_{\mathrm{LJ}}(1.2\sigma)$. The step Yuk potentials retain the depth of the first step until the wth step. From there, they exhibit a Yuk decay as specified by κ.

$$u_m = \varepsilon_1 \exp\left[-\kappa(\lambda_m - \lambda_w)\right] \tag{9}$$

where $\kappa = \ln(0.3/\varepsilon_1)/(\lambda_w - 3)$, such that $u_{11} = 0.3$ in all cases. The difference between the SwYukawa Universal (SYU) potential and the SwYukawa potential is that a universal value of λ_w has been applied for all site types. This effectively reduces the number of decision variables per site type to $2^+(\sigma$ and $\varepsilon)$ for the SYU potential since there are 18 site types (Table 2).

The DMD formalism is as old as molecular simulation itself. Since Alder and Wainwright [1] first used MD to simulate spherical systems, this method has been steadily improved. Rapaport extended the DMD method to chain molecules improved its efficiency [20, 21]. Chapela and co-workers contributed extensively to the modelling of repulsive molecular models by DMD, including a study of the discretised LJ model [7,8]. Smith et al. demonstrated the trends in transport properties by the DMD method [22]. Details of the DMD algorithm as applied in the present work have been described previously [14,15,18,24].

3. Statistical measures of model effectiveness

Characterisation of transferable potential models is a challenging situation from a statistical perspective. To begin with, the primary objective is to minimise vapour pressure deviations, with minimal deviations for density and other properties as secondary objectives. Vapour pressures vary over 2–3 orders of magnitude in the range of interest. Since temperature and potential energy are closely related through the perturbation contributions, changes in the potential impact the entire vapour pressure curve with exponential sensitivity through the Clausius–Clapeyron relation. On the other hand, the deviations in liquid density vary almost linearly with the site diameters and are weakly correlated with vapour pressure deviations. With this in mind, we define the 'optimal model' as minimising the RMSEp as for a given set of site diameters. These considerations indicate a strongly nonlinear optimisation. Unfortunately, most rigorous statistical methods have been developed with linear situations in mind.

Another challenge relates to model error. The ideal situation corresponds to negligible model error. In that case, all the deviations are meaningful discrepancies that should be eliminated. In the present case, the transferability assumption implies that, for example, the potential of a CH_2 site in n-pentane is identical to that of a CH_2 site in isopentane. This transferability assumption is an approximation, and it results in substantial model error ($\sim 10\%$) owing to the sensitivity of the vapour pressure.

One point that would seem to work in favour of the present analysis is that we have much data, roughly 20 points per compound ranging from a reduced temperature ($T/T_c \equiv T_r$) of 0.9–0.45. Unfortunately, the interaction between the model error and the large database actually complicates the situation. Statistical tests based on zero model error lead to the conclusion that virtually any reduction in RMSE is significant when applied to a large database. For example, the difference between 10.6 and 10.1% RMSE might be deemed significant by an F-test, even if the number of decision variables increased from 5 to 50. It is difficult to justify such a large increase in decision variables from a practical perspective.

To express this practical perspective quantitatively, statistics have been developed in addition to the F-test. We consider four such statistics in the analysis below: (1) F-test (2) Mallow's information criterion (MIC) (3) Akaike's information criterion (AIC) (4) Bayesian information criterion (BIC). The goal of all the additional statistics is to eliminate the overfitting that is inherent in the F-test. The various criteria place different emphasis on the factors that complicate this analysis. For example, the AIC provides better compensation for deviations from normality in the residual distributions than the MIC. In general, it is difficult to judge which emphasis is the

Table 3. Model comparison based on a database of 86 hydrocarbons.

Model	# DecVar	RMSEp	F^*	AIC	BIC	MIC	%ρRMSE
Lin2580	36	9.72	0.91	230	6895	7154	4.2
Lin2580 +	54	9.70	0.98	259	6924	7313	3.9
uncon2580 +	90	9.70	1.31	331	6996	7645	3.8
Lin9	36	9.69	0.85	219	6885	7145	4.3
2Lines	54	9.68	0.94	252	6918	7307	4.2
LJ	18	37.22	110.83	22840	10887	11016	14.9
SYU7	19	9.33	0.19	63	6736	6865	4.0
Full11wells	198	9.21	NA	385	7058	8484	4.0

most important for any particular analysis, including this one. Therefore, it is favoured to tabulate all the criteria and to consider them collectively in assessing the 'best' model overall.

3.1 F-*test statistic*

The *F*-test is designed to investigate whether the extra reduction of the regression sum of squares arises due to the terms under consideration being in the model. The value of $F*$ is calculated in terms of the reduction in the residual sum of squares as

$$F^* = \frac{\mathrm{SSR_{red}} - \mathrm{SSR_{full}}}{df_{\mathrm{red}} - df_{\mathrm{full}}} \cdot \frac{df_{\mathrm{full}}}{\mathrm{SSR_{full}}} \tag{10}$$

where *df* is the degrees of freedom (e.g. $n - q$), *SSR* is the sum of squared 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. The additional terms should be included if $F* > F(q - p;\ q;\ 1 - \alpha)$ where α is the significance level [13]. A small value for $F*$ indicates that the change in SSR is small. Hence, a minimal value of $F*$ is preferred. Note that the SSR and difference in SSR are proportional to the number of points. So, a large database yields a large value for $F*$ even if the *RMSE* reduction is small.

3.2 *Mallows' information criterion*

MIC is the first of three IC, that are treated in the present work. ICs penalise added decision variables more than the *F*-test. MIC is designed to diminish the role of bias error in the RMSE. Mallows noted that the RMSE includes bias error in addition to its most direct relation to the variance. For very large datasets, this bias error can play an artificially dominant role in the RMSE. Mallows formulated the MIC within the context of certain assumptions about linearity and normality in order to eliminate this artefact. The MIC statistic for a given model can be calculated by

$$\mathrm{MIC} = (n - q)\left(\frac{\mathrm{SSR_{red}}}{\mathrm{SSR_{full}}}\right) + 2p - n. \tag{11}$$

Note that the MIC might be negative for a large database. The least positive value of MIC indicates the best model within the context of this criterion.

3.3 *Akaike's information criterion*

The AIC statistic is defined as

$$\mathrm{AIC} = n \ln\left(\frac{\mathrm{SSR}}{n}\right) + 2p. \tag{12}$$

To use AIC for model selection, one simply chooses the model giving the smallest AIC over the set of models considered [13,16]. AIC is closely related to MIC, but includes a correction that compensates for non-normality in the residual distributions and takes the approach of maximising the log-likelihood of optimal model identification. AIC has come to be preferred by many statisticians in recent years.

3.4 *Bayesian information criterion*

BIC is similar to AIC, especially in the application of a maximum likelihood approach, but the penalty on the number of decision variables is greater, giving preference to simpler models in selection [16]. It can be shown that 'the BIC may be interpreted as a first approximation to the Bayes factor for comparing two models [6]'. The BIC statistic for a model is

$$\mathrm{BIC} = n \ln\left(\frac{\mathrm{SSR}}{n}\right) + p \ln(n). \tag{13}$$

The penalty on the variables is $p \ln(n)$ which was $2p$ in the AIC method. BIC is motivated by a Bayesian approach to model selection and is said to overfit less than AIC. Once again, the minimal BIC indicates the best model.

Table 4. The RMSEp according to family for the SYU7 and Lin2580 models.

Name	NComps	Ndata	SYU7	2580
n-Alkanes	26	382	10.17	10.24
Br-Alkanes	26	488	7.77	9.17
Alkenes	14	329	5.46	4.68
Aromatics	10	182	9.94	9.80
Overall	86	1500	9.33	9.72

4. Results and discussion

Table 3 shows a comparison of several potential models for our hydrocarbon database. 'nDecVar' corresponds to the number of decision variables for a given potential model. For example, the Lin2580 + model has three decision variables for each site-type: the depth of the first well, the depth of the well at $r/\sigma = 1.8–2.0$, and the depth of a lumped outer well ranging from $r/\sigma = 2.0$ to 3.0. The SYU model has 19 decision variables because the position for switching from the square-well to the Yuk decay was varied independently. Prior to compiling Table 3, a simple study was performed, trying different values of λ_w and tabulating the optimal RMSEp for each case. The optimal value was $\lambda_w = 1.7$, so we refer to this model as SYU7. Note that the Yuk decay exponent was pre-set by Equation (9), so this was not a decision variable.

The values in Table 3 correspond to vapour pressure deviations. Note that the diameters of the potential models are not varied at this stage because they are determined primarily by minimising the liquid density deviations in a separate optimisation. The RMSEp values vary over a surprisingly narrow range for most models. As one would expect, the error generally decreases with increasing number of decision variables. The optimisations were conducted for both subsets and for the entire set separately with RMSEp values in Table 3 based on optimisation for the entire set. In applying the *F*-Test, the 11 well model was taken as the full model. For 1500 data points, every *F*∗ indicates roughly 0% probability that the reduced model is acceptable.

With the above analysis, two potential models remain in close contention: the SYU model and the full model. The SYU potential is favoured by all IC's, however. The SYU potential also has an additional advantage, which it shares with the LJ model. Looking forward to extended databases, it is conceivable that two site types may be similar, but it may be unclear whether they should be declared as identical. If characterised by a single decision variable, as in the LJ or SYU models, a simple *t*-statistic can often be used to resolve this issue. For example, the strength of the CH_3a interaction (50.7 ± 1.6) is distinguishable from the CH_3b interaction (54.8 ± 1.7) because their confidence intervals do not overlap. For higher order models, like the Lin2580 model, coupling may lead to a high value for one decision variable and a low value for the other, with the situation reversed for the other similar site type, such that the effects tend to cancel in the overall potential. In this scenario, comparing regressed values within their standard error may lead to a false conclusion that the site types are dissimilar. Taking all criteria and practical considerations into account, it is concluded that the SYU model should be favoured. However, owing to prior development of the Lin2580, the results for both models should be documented for the present. Over time, the Lin2580 model should be dropped. The final results are summarised in Table 4. Detailed results for each compound are given in the appendix. The characterisation of the SYU model for

Table 5. Parameters of the SYU7 potential.

Site type	ε/k_BT (K)	Standard Error	σ (nm)	Frequency	Description
CH_3a	57.7	4.2	0.363	65	1′ CH_3, e.g. *n*-butane, *n*-octane
CH_3b	53.1	5.4	0.363	27	2′ Branch e.g. 23dimethylbutane
CH_3d	120.8	4.4	0.363	4	Aromatic appendage, eg. toluene or pxylene
CH_3de	64.1	3.0	0.363	10	2-Alkene bonded, e.g. *cis*-2butene, *trans*-2butene
CH_2a	26.9	1.5	0.357	66	In *n*-alkanes, e.g. *n*-butane, *n*-octane
CH_2e	27.8	8.8	0.357	9	Bonded to CH = a, e.g. 3hexene
CH_2ar	67.8	1.0	0.357	3	Bonded to an aromatic ring, e.g. Propylbenzene
CH_2r	29.8	4.1	0.357	7	In a non-aromatic ring, e.g. cyclohexane
CH_2rb	30.9	2.5	0.357	3	Bonded to a non-aromatic ring, EtCyHexane
CHa	6.5	1.8	0.357	22	In branched alkanes, e.g. isopentane
CHbb	6.8	2.6	0.390	4	Adjacent to another CHbb, e.g. 23DiMeHexane
CHr	8.1	3.6	0.390	6	In reply to: a non-aromatic ring, e.g. MeCyHexane
ACHb	29.9	2.9	0.3425	11	In an aromatic ring, e.g. Benzene
ACt	0.1	0.9	0.330	7	In an aromatic ring, e.g. Toluene
ACfa	23.3	3.6	0.330	3	In fused aromatic rings, e.g. Naphthalene
CH_2 = a	51.9	3.2	0.350	8	Alpha-olefin, e.g. 1butene
CH = a	23.6	7.2	0.350	15	Simple olefin, e.g. 2pentene
CH = de	27.8	1.6	0.350	3	Conjugated olefin (bound to another CH = de)

Table 6. Summary of deviations with the SYU potential model.

Compound	$\%AAD_P$	$\%BIAS_P$	$\%MAX_P$	$\%AAD_\rho$	$\%BIAS_\rho$	$\%MAX_\rho$
n-Alkanes	9.27	−2.36	23.39	5.12	4.96	8.83
Br-Alkanes	7.39	0.08	25.12	3.07	2.98	17.26
Alkenes	4.50	−0.61	20.95	2.46	−1.08	5.32
Aromatics	7.87	−0.95	29.58	3.35	−1.18	10.64
Naphthenics	6.79	0.65	22.27	3.98	3.98	9.50
Overall	7.42	−0.87	29.58	3.70	2.50	17.26

each site type is given in Table 5. The frequency column in Table 5 indicates the frequency of occurrence of each site type in the 86 compounds of the database.

More detailed statistics for the SYU model are summarised in Table 6. The column headings in Table 6 are defined below.

$$\%AAD_Q = \sum \frac{|Q_i^{calc} - Q_i^{expt}|}{Q_i^{expt}} \times 100, \qquad (14)$$

$$\%BIAS_Q = \sum \frac{(Q_i^{calc} - Q_i^{expt})}{Q_i^{expt}} \times 100, \qquad (15)$$

where Q is the quantity being evaluated (P or ρ). Note that the %AAD values are consistently smaller than the RMSE values because the squared error places greater emphasis on large deviations. Also, note that the density deviations are larger than those for the Lin2580 potential, because the site diameters have yet to be optimised specifically for the SYU7 potential. Larger site diameters would make the density bias more negative. The density deviations are also influenced by the inclusion of density data that have been extrapolated outside the measured range [12]. These extrapolated data should be eliminated before performing a detailed optimisation of site diameters.

The characterisation of the CH_2 = e site provides an illustration of model reduction based on comparing site types. The full model includes distinct CH_2 = e sites and the differences in well-depth are checked for their significance in the reduction of RMSE since their standard errors appear to indicate an overlap in confidence intervals. Since IC statistics in Table 7 do not support the equivalency of both models, it is concluded to use two values. Hence, the parameters for sites describing the SYU7 model are as in Table 5. In contrast, a clear distinction is evident between the benzylic carbon sites. For example, the aliphatic CH_2a is characterised by $\varepsilon/k_B = 26.9 \pm 1.5$ (K) while, the value for the benzylic CH_2b is characterised by $\varepsilon/k_B = 67.8 \pm 1.0$. These distinctions correlate with traditional qualitative observations about the attractive and reactive nature of these site types.

Noting the inaccuracy of the LJ potential model as observed in the present study, it is natural to consider how the parameters of the present work compare to those derived in previous works. In Table 8, we observe decreasing attractive strength from methyl (CH_3) to methyne (CH) in hydrocarbons. This trend is consistent with the LJ parameters of other force fields like TraPPE [19], OPLS [17] and AUA [5]. The LJ parameters of other force fields provide greater accuracy in liquid density. We can systematically increase the diameters and optimise the potentials. Preliminary results for the *n*-alkanes indicate that the deviations in vapour pressure change little during such a variation while density deviations approach those of the TraPPE model. In our opinion, a sacrifice of 2–3% accuracy in liquid density is worthwhile in comparison to a 300% improvement in vapour pressure accuracy. Unfortunately, a detailed study with variable diameters necessitates re-simulation of the entire database, repeating for each alternate diameter. Therefore, such a study is beyond the scope of the current work.

Table 7. Model reduction test of the SYU7 model for CH_2=e.

Site	Full	Standard Error	Reduced	Standard Error
CH_2a	26.9	1.5	26.9	1.6
CH_2e	27.8	8.8	NA	NA
RMSEp	9.33	–	9.35	–

Following Equation (10), $SSR_{full} = 9.33^2 \times 1500 = 130573$; $SSR_{red} = 9.352 \times 1500 = 131134$
$F^* = (131134 - 130573)/(1482 - 1481) \times 1481/130573 = 6.36$
MIC = 1481/130573 × 131134 + 2 × 18 − 1500 = 23; AIC and BIC are similar.

Model	F^*	MIC	AIC	BIC
Reduced	6.4	23	6742	6838
Full	NA	19	6838	6832

Table 8. Comparison of LJ well depths of alkanes for various force fields.

Site	OPLS	TraPPE	AUA	SPEADMD
CH_3	88.1	98	120.2	128.1
CH_2	59.4	46	86.3	96.0
CH	NA	10	51.0	58.4

5. Conclusions

This work has demonstrated the application of systematic statistical analysis to infer transferable intermolecular potential functions from molecular simulations and experimental data on vapour pressure and density. The combination of DMD/TPT permits efficient regression of potential parameters for large databases, raising the prospect of answering two questions: (1) what description of the disperse interactions is statistically optimal (e.g. LJ, linear or Yuk)? (2) what are the globally optimal values of the potential descriptors for the optimal potential?

Answering the first question is aided by considering statistical information criteria. The IC, apply penalties for increasing the number of decision variables. This is especially important for a large database with significant model error, as in the present case. We conclude that SYU potential is optimal overall, but other potentials are competitive. The SYU potential characterises the attractive strength in terms of a single decision variable, facilitating evaluations of whether isomeric interaction sites are distinguishable (e.g. *ortho*-, *meta*-, or *para*-methyl sites).

Optimal values for the potential descriptors are a natural outcome of statistical analysis. Noting the success of the SYU potential, it is natural to compare the values of the descriptors from this work to those from previous works based on continuous potentials. We note that the attractive strengths are higher in the present work, but the trends among site types are consistent (e.g. $\varepsilon_{CH_3} > \varepsilon_{CH2} > \varepsilon_{CH}$). This observation suggests a systematic study to determine whether a direct correspondence can be established between optimal continuous potentials and the optimal discretised version. Noting that the SYU7 model provides substantially greater accuracy for vapour pressure than previous works based on the LJ potential, we favour the present characterisation for now and plan to study the correspondence in future work.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant no. OISE-0421849. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Part of this project was also supported by TUBITAK grant TBAG-U/99 104T097 and Bogazici University research grant 04HA501, which are gratefully acknowledged. JRE would like to thank Dr Richard L. Einsporn in the University of Akron Statistics Department for his guidance on the subject of information criteria.

References

[1] B.J. Alder and T.E. Wainright, *Studies in molecular dynamics. I. General methods*, J. Chem. Phys. 31 (1959), p. 459.
[2] J.A. Barker and D. Henderson, *Perturbation theory and equation of state for fluids: the square-well potential*, J. Chem. Phys. 47 (1967), p. 2856.
[3] ———, *What is liquid?*, Rev. Modern Phys. 48 (1976), pp. 587–671.
[4] F.S. Baskaya et al., *Transferable step potentials for amines, amides, acetates, and ketones*, Fluid Phase Equ. 236 (2005), pp. 42–52.
[5] E. Bourasseau et al., *New optimization method for intermolecular potentials: optimization of a new anisotropic united atoms potential for olefins: prediction of equilibrium properties*, J. Chem. Phys. 118 (2003), pp. 3020–3034.
[6] P.J. Brown, *Measurement, Regression and Calibration*, Clarendon Press, Oxford, 1993.
[7] G.A. Chapela, S.E. Martinez-Casas, and J. Alejandre, *Molecular dynamics for discontinuous potentials. I. General method and simulation of hard poly atomic molecules*, Mol. Phys. 53 (1984), p. 139.
[8] G.A. Chapela, L.E. Scriven, and H.T. Davis, *Molecular dynamics for discontinuous potentials. IV. Lennard-Jonesium*, J. Chem. Phys. 91 (1989), p. 4307.
[9] J. Cui, *Step potential molecular models of pure components and mixtures by DMD/TPT*, Ph.D. diss., The University of Akron, 2001.
[10] J. Cui and J.R. Elliott, *Phase diagrams for multi-step potential models of n-alkanes by discontinuous molecular dynamics/thermodynamic perturbation theory*, J. Chem. Phys. 114 (2001), p. 7283.
[11] J. Cui and J.R. Elliott, Jr., *Phase envelopes for variable well-width square well chain fluids*, J. Chem. Phys. 116 (2002), p. 8625.
[12] T.E. Daubert and R.P. Danner, *Physical and Thermodynamic Properties of Pure Chemicals: Data Compilation*, AIChE, New York, 1989–2003.
[13] N.R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1981.
[14] J.R. Elliott, A. Vahid, and A.D. Sans, *Transferable potentials for mixed alcohol–amine interactions*, Fluid Phase Equ. 256 (2007), pp. 4–13.
[15] N.H. Gray, Z.N. Gerek, and J.R. Elliott, *Molecular modeling of isomer effects in napthenic and aromatic hydrocarbons*, Fluid Phase Equ. 228–229C (2005), pp. 147–153.
[16] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer-Verlag, New York, 2001.
[17] W.L. Jorgensen, J.D. Madura, and C.J. Swenson, *Optimized intermolecular potential functions for liquid hydrocarbons*, J. Am. Chem. Soc. 106 (1984), p. 6638.
[18] J.-X. Liu and J.R. Elliott, *Screening effects on Hydrogen bonding in chain molecular fluids: thermodynamics and kinetics*, IEC Res. 35 (1996), pp. 2369–2377.
[19] M.G. Martin and J.I. Siepmann, *Transferable potentials for phase equilibria. 1. United-atom description of n-alkanes*, J. Phys. Chem. B 102 (1998), pp. 2569–2577.
[20] D.C. Rapaport, *Molecular dynamics study of a polymer chain in solution*, J. Chem. Phys. 71 (1979), p. 3299.
[21] ———, *The event scheduling problem in molecular dynamic simulation*, J. Comput. Phys. 34 (1980), p. 184.
[22] S.W. Smith, C.K. Hall, and B.D. Freeman, *Molecular dynamics study of transport coefficients for hard-chain fluids*, J. Chem. Phys. 102 (1995), p. 1057.
[23] S. Ucyigitler et al., *Optimal selection of the most effective intermolecular potential model*, in preparation (2007).
[24] O. Unlu et al., *Transferable step potentials for the straight chain alkanes, alkenes, alkynes, ethers, and alcohols*, Ind. Eng. Chem. Res. 43 (2004), pp. 1788–1793.