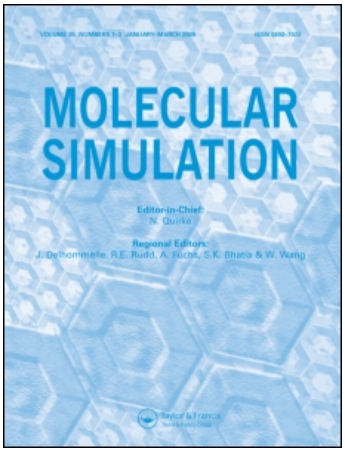

Molecular Simulation
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713644482

Multi-objective optimisation on the basis of random models for ethylene oxide

Astrid Maaß[a]; Lialia Nikitina[a]; Tanja Clees[a]; Karl N. Kirschner[a]; Dirk Reith[a]
[a] Department of Simulation Engineering, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, Sankt Augustin, Germany

First published on: 04 November 2010

To cite this Article Maaß, Astrid , Nikitina, Lialia , Clees, Tanja , Kirschner, Karl N. and Reith, Dirk(2010) 'Multi-objective optimisation on the basis of random models for ethylene oxide', Molecular Simulation, 36: 15, 1208 — 1218, First published on: 04 November 2010 (iFirst)
To link to this Article: DOI: 10.1080/08927020903483312
URL: http://dx.doi.org/10.1080/08927020903483312

Molecular Simulation
Vol. 36, No. 15, December 2010, 1208–1218

Taylor & Francis
Taylor & Francis Group

Multi-objective optimisation on the basis of random models for ethylene oxide

Astrid Maaß*, Lialia Nikitina, Tanja Clees, Karl N. Kirschner and Dirk Reith

Department of Simulation Engineering, Fraunhofer Institute for Algorithms and Scientific Computing SCAI, Schloss Birlinghoven, 53754 Sankt Augustin, Germany

(*Received 14 September 2009; final version received 12 November 2009*)

This paper is part of our pursuit to develop an efficient procedure for optimising parameters that provide a reliable foundation for highly predictive molecular simulations. We tested whether DesParO, a mathematical tool originally used in automotive design, is suitable for creating Lennard-Jones (LJ) parameters that accurately reproduce the experimental phase behaviour for our test compound ethylene oxide (EO). So, we created a multitude of diverse random parameter sets, performed Gibbs ensemble Monte Carlo simulations and collected the resulting physical properties. On that data basis, DesParO derived a meta-model through a multidimensional interpolation. We then explored, in an interactive fashion unique to DesParO, the LJ parameter space and selected some suitable parameter sets, which were then tested by simulations. For EO, the selected parameter sets were indeed superior to the initial parameters. Furthermore, the new parameters can be reliably used as input for further optimisation by other methods, resulting in extremely robust LJ parameters. Beyond the prediction of parameter sets, DesParO enabled us to examine the underlying parameter–property relationships that help us solve future optimisation problems by creating subordinate parameter optimisation tasks in a systematic manner; this ability makes DesParO a valuable tool in the overall optimisation process.

Keywords: parameter optimisation; force field; meta-model; DesParO; GROW

1. Introduction

In molecular simulations, it is highly desirable that a computational model most genuinely captures a real compound's behaviour over an extended range of experimental conditions. Our goal is to identify models that may be applied to a wide range of problems and conditions efficiently and reliably. Irrespective of the simulation method (i.e. molecular dynamics or Monte Carlo) and with a program-defined potential function, only the force field's parameters remain as adjustable variables for the reproduction of experimental observables.

Previously, optimising the parameters was addressed by either iterative manual adjustments [1–3] or employing systematic minimisation schemes (e.g. simplex optimisation or Newton optimisation [4–6]).

Instead of focusing the search to a minimum of the respective error function in a small area of parameter space, here, we wanted to establish a workflow that initially covers a broad range of parameter space prior to full optimisation, and then narrow that space down to the most promising region. Thus, in this study, we have created arbitrary models for ethylene oxide (EO), used as a test compound, and determined the physical properties, via simulations, inherent to each of these models. From the underlying information embedded in the given input parameters and their corresponding output properties, we may derive new parameter combinations that reproduce the experimental behaviour best. The final outcome is a set of well-suited parameter sets, which can be used for relatively minor subsequent optimisation steps (e.g. the gradient-based optimisation scheme that was recently developed by our group [7,8]). Since vapour–liquid coexistence curves are well accepted as a standard for assessing a model's ability to reflect experimental behaviour, we conducted Gibbs ensemble Monte Carlo (GEMC) simulations [9] for our test compound EO. Hence, the liquid density and the heat of vaporisation at different temperatures served as indicators for assessing the quality of any given parameter set.

In order to analyse the relations between the simulation's input and output data, we employed the DesParO program [10–12]. DesParO was originally developed to support the automotive design process by means of analysing a set of simulation data; the data themselves are a collection of individual data-sets, each characterising a model by specific parameters and a list of associated properties. This is a similar situation to what is seen in force-field optimisation, and thus we wanted to explore the use of DesParO to adjust molecular parameters for simulations.

Improving a multitude of parameters based on many criteria simultaneously is termed multi-objective optimisation, and within DesParO the problem is formulated as a mapping from the parameters' space (synonymous for

*Corresponding author. Email: astrid.maass@scai.fraunhofer.de

ISSN 0892-7022 print/ISSN 1029-0435 online

DOI: 10.1080/08927020903483312
http://www.informaworld.com

design variables) to the criteria (i.e. properties) space. In other words, via advanced interpolation techniques, a meta-model that captures the underlying trends of the data is being created. On the basis of this meta-model, the performance of an arbitrary set of parameter values for reproducing the target properties may be predicted immediately, by exploring the relationship between the parameter space and the criteria/property space interactively. Thereby, the user may identify favourable regions in the parameter space prior to an eventually following full optimisation, or – if not yet possible – might reprioritise the choice and/or weighting of criteria accordingly.

This new approach comprises several profitable features [10–12]: first, it allows gaining a global and 'coarse' view of promising parameter combinations by interactively exploring a comparatively large region of the parameter space. The number of criteria to consider may be arbitrarily restricted and the effects can be observed instantly. Moreover, tolerance intervals are displayed, indicating the reliability of the resulting value for the respective criterion/property. These tolerance intervals are determined on a leave-one-out basis and are thus locally resolved. This aids the user in identifying the most promising regions in the parameter space.

Finally, employing a non-linear correlation matrix instead of the more typical linear scheme to identify the system's features improves the overall accuracy in prediction. With automated optimisation procedures (e.g. minimisation methods such as GROW [7,8]), this freedom of choosing interesting parameter ranges or readjusting the number of criteria is lacking, since an automatism requires the error function to be known beforehand. Even more important, with DesParO, one can avoid becoming stuck in a local minimum, as might be the case with a gradient-based method, when being applied to a large search space.

However, the practicality of our proposed study becomes impaired with increasing the number of adjustable parameters. Furthermore, utilising a program in an interactive manner can be regarded as a disadvantage, because the optimisation process is non-deterministic and the exact adjustment procedure might not be reproducible. Since the final results can be reproduced using traditional methods, we deem that to be a minor problem. Moreover, automation of the multi-objective optimisation is possible (see also Section 5).

To test the above concept for an exemplary compound, we chose EO as it has been thoroughly characterised by experiments [13,14] and by simulations [5,6,15–18]. For the optimisation, we restrict ourselves to modifying only the Lennard-Jones (LJ) parameters as variables. We want to emphasise again that the outlined approach requires recording data from many simulations. Thus, one of the challenges is the economic handling of resources, such that the number of simulations is kept reasonable, with each being performed for a reasonable amount of time. As stated above, the most important return is a profound understanding of the relationship between the parameter and the criteria spaces.

2. Methodology

In order to prepare the input for the DesParO tool, an automatism for executing a series of simulations for a given model was set up and tested for a reference model and a prototypic precursor model of EO. Based on the precursor model, 75 different LJ parameter sets were created using a random number generator and characterised by reconstructing their vapour–liquid coexistence curves via simulation at different temperatures. The overall study comprises the following steps:

(1) GEMC simulations for model 1 (reference model) at seven temperatures,
(2) computing a new set of partial atomic charges to create model 2,
(3) GEMC simulations for model 2 (precursor model) at seven temperatures,
(4) creating 75 new models by variation of ε- and σ-values for carbon and oxygen,
(5) GEMC simulations for models 3–77 (random models) at five temperatures,
(6) analysis of parameter–property relationships and identifying three new models and
(7) GEMC simulations for models 78–80 (predicted models) at seven temperatures.

2.1 Model creation

Throughout this study, EO is being represented as a simple united-atom model based on that published by Wielopolski and Smith [15]. Starting from this model (model 1, reference model) as a template, we created a second model (model 2, precursor model) by assigning a new set of partial atomic charges. The molecular parameters for model 1, model 2 and their descendants are summarised in Table 1. For creating model 2, we modified the charge set in such a way that the following quantum mechanical computations were conducted to yield the final partial atomic charges: first, a geometry optimisation was performed at the HF/6-31G(d) level of theory, while the molecular electrostatic potential, as computed by the CHELPG formalism, was obtained at the HF/aug-cc-pVTZ//HF/6-31G(d) level of theory. The partial atomic charges, given in Table 1, were generated using the RESP algorithm and a weighting factor of 0.01 [19]. The quantum mechanical calculations were performed using the program GAMESS [20,21].

Model 2 served as a template for creating 75 random models. These models for EO differed among each other and model 2 in the individual LJ parameters. Their values

Table 1. Data used to construct the EO models in this study.

Models 1–80	Model 1	Model 2	Models 3–80
r(C—C) 1.466 Å	LJ parameters		
r(C—O) 1.431 Å	σ_{CC} 3.7143 Å ε_{CC} 90.0 K σ_{OO} 2.6666 Å ε_{OO} 73.0 K		$3.588 < \sigma_{CC} < 3.777$ Å $79.20 < \varepsilon_{CC} < 111.60$ K $2.475 < \sigma_{OO} < 2.911$ Å $64.24 < \varepsilon_{OO} < 90.52$ K
α(C—O—C) 61.62°	Partial atomic charges		
α(O—C—C) 59.19°	q_C 0.1608 q_O −0.3218		q_C 0.1741 q_O −0.3482

were chosen randomly from the following intervals: σ_{CC}, 3.588–3.777 Å; σ_{OO}, 2.475–2.911 Å; ε_{CC}, 79.2–111.6 K; ε_{OO}, 64.24–90.52 K. The values for the hetero-atomic combinations were derived via the Lorentz–Berthelot mixing rules, i.e. $\varepsilon_{CO} = \sqrt{\varepsilon_{CC}\varepsilon_{OO}}$ and $\sigma_{CO} = (\sigma_{CC} + \sigma_{OO})/2$.

Both the reference and the precursor models were subjected to the simulation procedure outlined below in order to check its practicability before applying the procedure to the random models and for assessing the quality of the precursor model.

2.2 Simulation

For each individual model set, a system of 650 EO molecules was subjected to GEMC simulations [22] at 260, 300, 330, 375 and 400 K using the program Towhee (MCCCS Towhee, available at http://towhee.sourceforge.net). For selected sets (reference, precursor and predictions), additional simulations at 230 and 430 K were conducted.

All simulations were done under periodic boundary conditions. The total volume at a given temperature, which includes both a liquid-phase box and a vapour-phase box, was identical for all models. The dimensions were chosen such that the liquid box accommodated 450 molecules at the respective experimental liquid density, leading to accordant box lengths of 31.7, 32.3, 32.9, 34.0 and 34.8 Å for the above temperatures. The dimensions for the vapour boxes at low temperatures, however, were reduced significantly with respect to the values that would correspond to experimental vapour densities (i.e. 200, 120, 100, 85 and 75 Å), otherwise models that displayed very weak intermolecular attraction tended to evaporate too easily when exposed to larger volumes.

LJ interactions were truncated at 10.0 Å and analytical tail corrections were applied, while electrostatic interactions were computed via Ewald summation. In order to compensate for the low acceptance rates for volume moves and for two-box molecule transfer moves at low temperatures, their probability values were adapted to the given temperature by $1000/T^2$ and $10{,}000/T^2$, respectively. The remaining amount of probabilities was distributed equally among rotational and translational moves.

2.3 Equilibration

The time of an entire simulation was split into cycles. Each cycle consisted of 5000 steps, with each step involving 650 Monte Carlo moves. Each simulation consisted of 200–300 cycles, leading to a total of 1.0–1.5 million simulation steps per temperature. From the output of each cycle, the liquid density, vapour pressure, potential energy and the heat of vaporisation were continuously monitored. The evolution of the potential energy of both boxes served as an indicator for stopping the simulation when the equilibration criterion defined below was reached.

Since all simulated systems started from an artificial lattice structure, the first 10% of recorded data were discarded. The remaining 90% served as the production data-set and were split into 10 blocks. Simulations were stopped, if the following equilibration criterion, as implemented in a script for the statistics program R (the R-project for statistical computing http://www.r-project.org/), was fulfilled. Firstly, the difference between the mean of all blocks and the mean of an individual block was not allowed to exceed 70% of the standard deviation for the respective block, preventing largely undulating behaviour from being considered converged. Secondly, the slope of a straight-line fit through the valid data points was determined; the absolute value of the slope times 100 and divided by the mean of the valid data-set was not allowed to exceed a value of 0.001. This was to prevent also slightly drifting systems from being accepted as fully equilibrated. The combination of both criteria provides a rather rigorous equilibration stopping condition.

2.4 DesParO interface and use

Post-equilibration, the mean values for the liquid density and heat of vaporisation were collected from the latest 50% of the respective simulation. Finally, all results were collected in a single table listing per line the index of the individual data-set, which comprises the four variable

parameters ε_{CC}, σ_{CC}, ε_{OO} and σ_{OO} defining a model and the total of 10 associated properties obtained by simulation (liquid density and heat of vaporisation at five temperatures). This table serves as input for the pre-processing program newmodel that creates the input for the main program DesParO. Both newmodel and DesParO are described in [12] in detail. We summarise details on the use of DesParO in the following.

DesParO starts with a bipartite graphical user interface as the main user window. On the left-hand side, the parameters are represented as sliders that may be adjusted within a given range of values. On the right-hand side, the criteria are listed likewise, i.e. for each criterion a bar spanning the range as covered in the input is displayed. Here, however, only the sliders, denoting the upper and lower tolerable limits of the respective criterion, are movable by mouse clicks, whereas the actual value of a criterion solely depends on the position of the parameter sliders. Thus, by moving the parameter sliders, a certain combination of values may be picked, which results in an autonomous motion of the sliders on the criteria panel that is indicative of the predicted associated properties. In this manner, manually adjusted parameter sets were subjected to the same simulation procedure as detailed above, and thus checked for their actual performance versus the predicted behaviour.

Additionally, a correlation matrix can be viewed, which briefly summarises the influence of the isolated parameters on each criterion according to the deduced meta-model. The latter matrix can be used to guide the search for an optimal region in the parameter space.

3. Results

3.1 Performance of models 1 and 2

For model 1 [15], published values characterising the vapour–liquid behaviour are available [6,18], therefore recording GEMC results for this model were used to confirm that the simulation procedure as detailed above worked properly. For model 2 (the same as model 1 except for RESP charges), we wanted to assess its performance in order to check whether it was suited as a template model for creating random variants. Thus, both models have been subjected to the above outlined series of simulations, except for the fact that seven temperatures ranging from 230 to 430 K (instead of five ranging from 260 to 400 K) have been used to reconstruct the vapour–liquid coexistence curve by GEMC simulations.

As a typical example for the equilibration behaviour seen generally, Figure 1 presents the evolution of total energy, vapour pressure, density and heat of vaporisation as a function of time for model 1 at 260 and 400 K. Starting from an artificial lattice conformation of the system, after approximately 50 cycles, both the high- and the low-temperature simulations become equilibrated to their respective properties. The low-temperature simulations

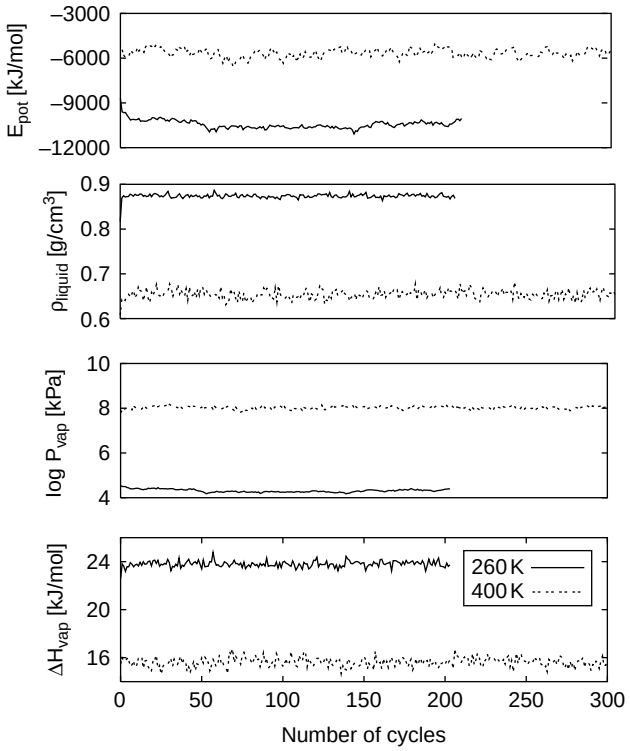

Figure 1. Representative equilibration behaviour as recorded for model 1 [15].

were associated with smaller fluctuation (favouring early termination), but had a tendency to reach equilibration later, due to the smaller MC acceptance rates, as compared to the high-temperature simulations. These tendencies lead to an unpredictable number of cycles for a given system to fulfil the equilibrium criterion.

The results obtained for the reference model are in good agreement with previously published data [4,5] as illustrated for the liquid density and the heat of vaporisation in Figure S1 of the Supplementary Material. This suggests that the above equilibration procedure indeed serves its purpose and may safely be applied to other models as well.

In order to assess the situation before modifying any LJ parameters, we also compared the computed values for models 1 and 2 to the experimental reference values of the liquid density, the vapour pressure and the heat of vaporisation [13,14], as shown in Figure 2. We observe that, for both models, the liquid density and the heat of vaporisation are underestimated over the whole temperature range, the latter being in accordance with elevated computed values for the vapour pressure. However, choosing the RESP charge set did improve the overall quality of the initial model with respect to the original one significantly. Therefore, we considered this model a promising starting point.

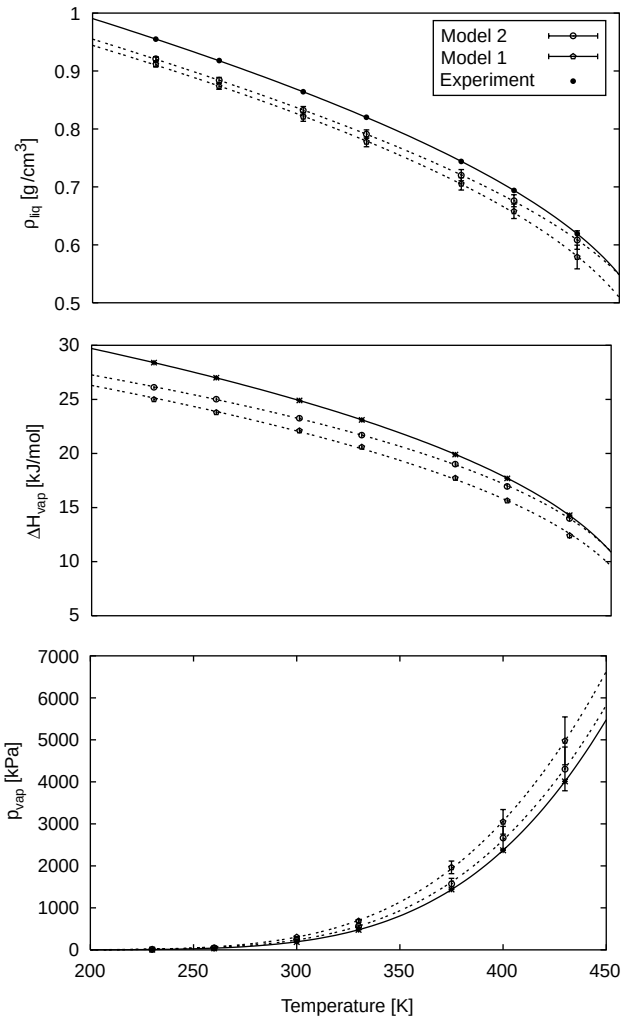

Figure 2. Comparison of experimental [13] versus computed values for the liquid density, heat of vaporisation and vapour pressure (top, middle, bottom) in order to illustrate the starting positions. Both models 1 and 2 yield too low densities over the entire range of temperatures. The fit function used to connect the data points for the liquid density was $\rho(t) = a/(b**(1+(1-T/T_{crit})**c))$. Likewise, both models underestimate the heat of vaporisation; the fit function used here was $\Delta H_{vap}(T) = a(1-T/T_{crit})**b$. Consistent with the underestimated heat of vaporisation, the vapour pressure is systematically overestimated; for plotting the fit function $p(T) = \exp(a+b/T)$ was used.

3.2 *Creating random models*

In order to generate new models, we had to define the intervals from which to pick new parameter values. The respective intervals were pre-estimated on the basis of the results obtained for models 1 and 2: considering that the reference model at 230 K computed the density $\sim 5\%$ too low, the atomic volume (which is proportional to σ^3) is likely overestimated. If one assumes that the density solely depends on σ, we can infer that the σ-parameters should be decreased by $\sim 1.7\%$, leading to expected values of ~ 3.66 Å for σ_{CC} and ~ 2.62 Å for σ_{OO}.

Similarly, the computed heat of vaporisation underestimates the experimental value by $\sim 12\%$; therefore, we anticipated that ε-values should grow by roughly this amount, yielding an ε_{CC}-value of 97 K and an ε_{OO}-value of 79 K. These estimates, together with the known starting points, guided us in defining the intervals from which to select the random parameter values. In order to cover a sufficiently broad range, the deviation between the original value and the anticipated value was used as an additional safety margin below the initial value and above the anticipated one. Thus, the actual random values were in the range between $(\sigma_{orig} - 1.7\%)$ and $(\sigma_{orig} + 2 \times 1.7\%)$, and $(\varepsilon_{orig} - 12\%)$ and $(\varepsilon_{orig} + 2 \times 12\%)$, respectively. As model 2 performed better (liquid density underestimated by 3.6%, heat of vaporisation underestimated by 8%), we concluded that the respective ranges of values were sufficient to allow searching for diverse solutions without having to consider (too many) nonsense combinations. For details please refer to Section 2.

In order to cover enough data points within the parameter space for interpolation, DesParO requires $C(2+n+n(n+1)/2)$ data-sets, where n is the number of variable parameters and C a constant that equals at least 1 (better 3–5). In our case, $n = 4$, implying that a minimum of 16 data-sets is required (i.e. 16 models with associated properties at five temperatures). Preferably, 48 or even 80 data-sets should be provided. Finally, we generated in total 75 diverse models for later analysis with DesParO.

3.3 *Analysis of random models*

The execution of all simulation threads took approximately 2 months of calendar time to accomplish. Finishing one thread did require 7–10 days.

In order to make sure that our input was indeed suited for DesParO, we checked the evenness of the distribution of values within the parameter ranges by plotting ε_{CC} values versus ε_{OO} values, σ_{CC} values versus σ_{OO} values and the ε and σ combinations. We find that the entirety of 75 parameter sets is scattered quite evenly over the respective areas (data not shown), except for σ_{OO}, where the upper and lower margins appear under-represented, as we increased the range during the course of data creation for reasons explained below.

3.3.1 *General view on the recorded data*

For surveying the large amount of data roughly and quickly, the following error function was computed for each model: error $= \sum_{i=1}^{n} w_i((x_i - x_{ref})/x_{ref})^2$, where x_i denotes the actual computed value for property i, x_{ref} is the

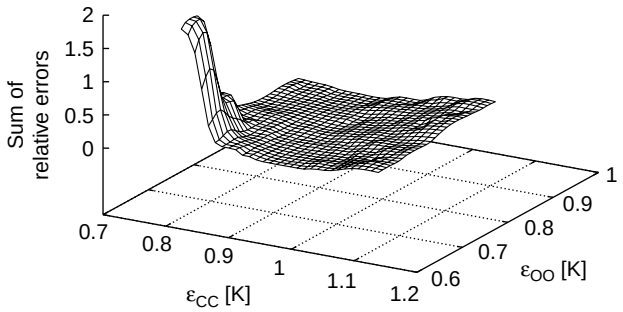

Figure 3. Survey of the performance landscape as a function of ε_{CC} and ε_{OO} values.

target value as observed experimentally and w_i is a weighting factor, which here equals 1 for all properties.

When plotting the ε-values versus the error function, as illustrated in Figure 3, a 'forbidden' corner, i.e. the region of low ε_{CC} and ε_{OO} values, is associated with extremely high scores. Low ε-values imply weak intermolecular interactions, leading to instantaneous evaporation of the liquid phase within the simulation volume if chosen too large. In that case, it is impossible to determine reasonable liquid density or the heat of vaporisation values at the highest temperatures (i.e. the corresponding parameter set is useless). Towards the opposite corner of the diagram (high ε_{CC} and ε_{OO} values), the error function rises moderately, thus leaving a roughly diagonal shallow valley, where a multitude of suitable ε_{CC} and ε_{OO} combinations may be found. Plotting the error function against both σ-values and the sum of ε-values and the sum of σ-values, in contrast, yielded very rugged landscapes with no trend information (data not shown). In conclusion of Figure 3, the ε-values dominate the vaporisation behaviour while atomic size does not matter in this respect.

Importantly, some randomly created models already performed quite well according to the above defined error function, and significantly better than the precursor model. The respective parameter combinations were:

(1) $\varepsilon_{CC} = 94.38\,\text{K}$, $\sigma_{CC} = 3.682\,\text{Å}$, $\varepsilon_{OO} = 70.24\,\text{K}$, $\sigma_{OO} = 2.551\,\text{Å}$ (model 61),
(2) $\varepsilon_{CC} = 95.06\,\text{K}$, $\sigma_{CC} = 3.695\,\text{Å}$, $\varepsilon_{OO} = 66.26\,\text{K}$, $\sigma_{OO} = 2.692\,\text{Å}$ (model 41),
(3) $\varepsilon_{CC} = 91.34\,\text{K}$, $\sigma_{CC} = 3.721\,\text{Å}$, $\varepsilon_{OO} = 90.06\,\text{K}$, $\sigma_{OO} = 2.529\,\text{Å}$ (model 67).

The resulting properties of these (almost equally) good models were compared to model 2 as well as to the experimental reference values, as shown in Figure 4.

Considering the uncertainties of the computed observables, as measured by the standard deviations and indicated by the error bars, the models are almost inseparable in terms of quality from each other even though they comprise a broad range of parameter values.

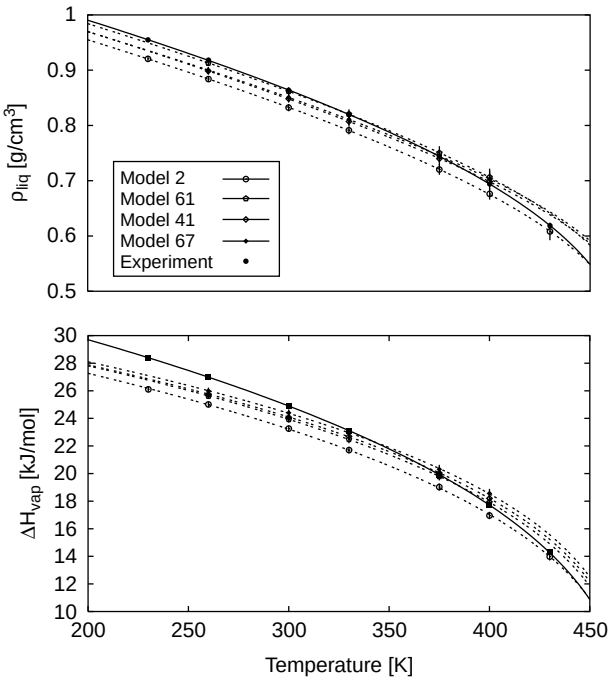

Figure 4. Computed values for liquid density (top) and heat of vaporisation (bottom) for model 2 and the three best performing, randomly created models (dotted lines) in comparison to the experimental references (solid line [13]).

The qualitative shapes of the curves are similar to those of model 2. However, they appear to be shifted towards the target curve. Similarly, the computed curves for the heat of vaporisation display a common overall shape and steepness, but do not exactly reproduce the curve defined by the experimental reference values. Overall, we found that only the points of intersection (between the experimental and the computed curves) appear to be adjustable by the four LJ parameters, but not the overall course of the curves.

3.3.2 Analysis with DesParO: properties of the meta-model

We applied DesParO to the whole set of models 3–77. From this input, DesParO derives a meta-model, which captures the relations inherent to the model parameters and their associated physical properties. Before advancing to the search of new parameter combinations, we wanted to study the meta-model itself. This was accomplished by first reading the above data-set collection, which was used for calibrating the meta-model, into the program. We then specified additional and equally spaced parameter values, whose corresponding property values were unknown. We then requested DesParO to print out the missing values for these properties as predicted by the meta-model.

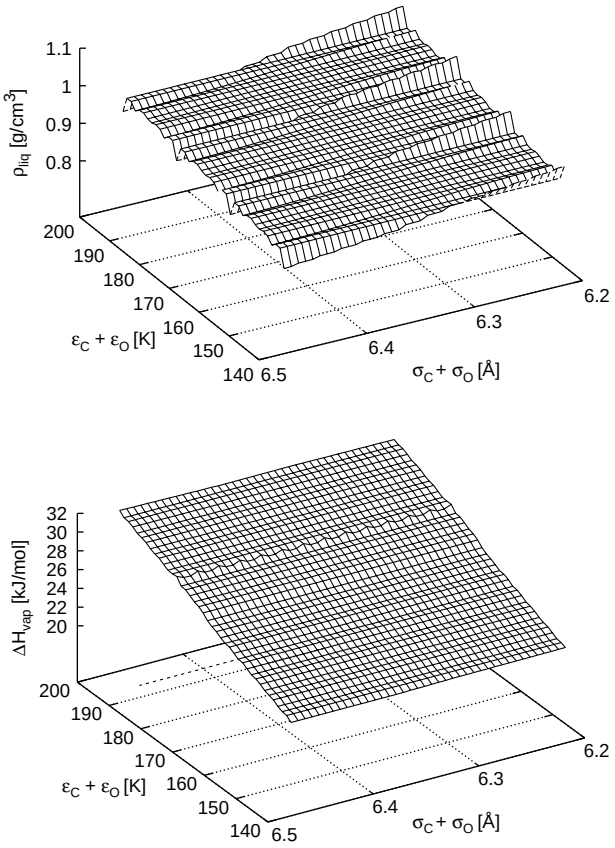

Figure 5. Predicted values for liquid density (top) and heat of vaporisation (bottom) at 260 K, illustrating the trends as identified by DesParO for the meta-model.

Some of these predictions are illustrated in Figure 5, which displays the dependence of the liquid density according to DesParO at 260 K from $\varepsilon_{CC}+\varepsilon_{OO}$ and $\sigma_{CC}+\sigma_{OO}$ as well as the heat of vaporisation, respectively. We find that the density depends on both parameter types, whereas the heat of vaporisation is only influenced by ε, which is consistent with the previous results from plotting the sum of relative square errors versus the ε-values. The ripples on the surface are artefacts generated by the plotting software and reflect the fact that the displayed 'surface' has a certain thickness. The useful information we gain from such plots is the sign and the value of the surface's slopes. This is also summarised in DesParO's correlation matrix that can be displayed optionally and will be discussed in more detail below (screenshots of the correlation matrices are given in Figure S2 of the Supplementary Material). Qualitatively these results could be expected, but, to our knowledge, this is the first time that we have access to quantitative information for each parameter and criterion separately.

3.3.3 *Analysis with DesParO: validation*

In order to estimate how many data-sets are needed to generate a meaningful meta-model, we tested three different data collections. They included 26, 55 and all 75 individual parameter sets plus simulation results, further referred to as D26, D55 and D75. As a first means of validation, for each data-set, we adjusted the parameter sliders to the initial values corresponding to model 2, which was *not* included in any of the data-sets, and put down the according results as calculated by DesParO in Table 2. The agreement between the computed references and the interpolated estimates turned out to be very good for temperatures ranging from 260 to 375 K. In this temperature region, we find that the DesParO-predicted values, within their confidence intervals, matched the simulated values. Likewise, as for simulations, the size of the confidence intervals increases with temperature, with a significant rise at the highest temperature. Note that raising the number of parameter sets does not necessarily improve the accuracy of a prediction in this application, as data-sets D26 and D75 perform quite comparably, whereas D55 happens to reproduce the expected values best. This might relate to the fact that for recording the latest 20 data-sets, the allowed interval for σ_{OO} had been increased, thus D75 covers a larger parameter space than D26 and D55 and

Table 2. Values and uncertainties or confidence intervals, respectively, for liquid density and heat of vaporisation as computed by GEMC simulation or by DesParO on the basis of the three collections of data-sets for model 1.

	Simulation	DesParO (D26)	DesParO (D55)	DesParO (D75)
Liquid density (g/cm^3)				
260 K	0.884 ± 0.005	0.884 ± 0.001	0.884 ± 0.001	0.880 ± 0.003
300 K	0.832 ± 0.007	0.832 ± 0.001	0.832 ± 0.001	0.840 ± 0.007
330 K	0.791 ± 0.010	0.791 ± 0.001	0.790 ± 0.001	0.798 ± 0.006
375 K	0.720 ± 0.010	0.720 ± 0.002	0.720 ± 0.002	0.720 ± 0.002
400 K	0.676 ± 0.020	0.699 ± 0.08	0.675 ± 0.028	0.700 ± 0.027
Heat of vaporisation (kJ/mol)				
260 K	25.01 ± 0.18	24.97 ± 0.03	24.95 ± 0.03	24.74 ± 0.17
300 K	23.26 ± 0.23	23.29 ± 0.03	23.28 ± 0.03	23.15 ± 0.13
330 K	21.70 ± 0.23	21.80 ± 0.08	21.80 ± 0.03	21.73 ± 0.10
375 K	19.02 ± 0.26	18.99 ± 0.07	18.99 ± 0.04	19.06 ± 0.07
400 K	16.96 ± 0.27	17.52 ± 1.74	17.03 ± 0.48	17.70 ± 0.65

therefore may not be strictly comparable to the smaller collections. As mentioned before, the increased range was not equally sampled, which might account for the reduced accuracy of prediction despite the increased collection of data-sets.

3.3.4 Analysis with DesParO: correlation matrices

Listing the specific model properties as criteria instead of computing the sum of relative square errors (which is common for creating a smooth and continuous error function for gradient-based minimisation schemes) implies that the meta-model retains the information about the qualitative trends (under-/overestimation). Therefore, DesParO's summarising correlation matrix is very illustrative (see Figure S2 of the Supplementary Material). DesParO's global correlation matrix can be seen as a non-linear version of the standard global linear Pearson correlation matrix and is particularly adapted to the non-linear multidimensional interpolation technique (radial basis functions) used [12]. Hence, it can be seen as an optimal global correlation estimation for the meta-modelling technique used.

When consulting this matrix first for the smallest data collection, D26, we found that ε_{CC} and σ_{CC} did effect all criteria (liquid density and heat of vaporisation at five temperatures) in an expected manner: with increasing ε_{CC}, the heat of vaporisation increases, as well as the density, for all temperatures. For the heat of vaporisation, this trend is most pronounced at low temperatures, but weaker at high temperatures and vice versa for the liquid density. At the same time, σ_{CC} has no impact on the heat of vaporisation, but a growing σ-value generally reduces the liquid density for all temperatures, again less effectively with rising temperatures. Unexpectedly, we found that the parameters describing the oxygen atom had almost no influence at all, although the oxygen atom carries the highest absolute partial atomic charge. Only the ε_{OO} parameter showed a correlation similar to that of the ε_{CC} parameter, but to a much lesser extent. For the parameter σ_{OO}, no influence was seen at all. To further explore the lack of influence that the oxygen parameters have (especially σ_{OO}), we performed the same analysis on D55, which contained almost twice as many individual data-sets. Despite the increased collection of data-sets, the correlation matrix did not show a more detailed pattern. Again, the parameter σ_{OO} apparently has no impact on the system's behaviour.

For recording more data-sets, therefore, we increased the interval for picking random σ-values to generate new models, in order to allow this parameter to become a more significant descriptor. As a consequence, D76's correlation matrix did show a small correlation pattern for both oxygen parameters, but is still significantly smaller than that of the methylene atom type.

3.3.5 DesParO analysis: predicting new parameter sets

Since all results were in line with our basic understanding of the problem and apparently displaying a satisfactory reliability, we used DesParO to generate new improved parameter sets. A single promising parameter set at a time was extracted for each data collection, yielding: model 78, based on D26 ($\varepsilon_{CC} = 93.16$ K, $\sigma_{CC} = 3.687$ Å, $\varepsilon_{OO} = 78.32$ K, $\sigma_{OO} = 2.625$ Å); model 79, based on D55 ($\varepsilon_{CC} = 94.63$ K, $\sigma_{CC} = 3.661$ Å, $\varepsilon_{OO} = 70.76$ K, $\sigma_{OO} = 2.621$ Å) and model 80, based on D75 ($\varepsilon_{CC} = 94.65$ K, $\sigma_{CC} = 3.687$ Å, $\varepsilon_{OO} = 74.06$ K, $\sigma_{OO} = 2.498$ Å).

Note again that the process of parameter selection is interactive. Hence, the actual procedure and subsequent results are not 100% reproducible, and the quality of the new parameter set may depend on the user's experience. Figure 6 summarises the simulation results for models 78–80. Concerning the predicted performance versus the actual computed performance, the results appear related to the previous validation: at lower temperatures, the agreement was very good, for the higher temperatures (375 and 400 K), the predicted values were better than observed by the confirming simulation.

The overall quality of the three predicted models with respect to the above-mentioned error function, or with respect to the general appearance of the graphs, is very similar to the three best models created randomly (cf. Figure 4). In fact, the ε-values of all six models are

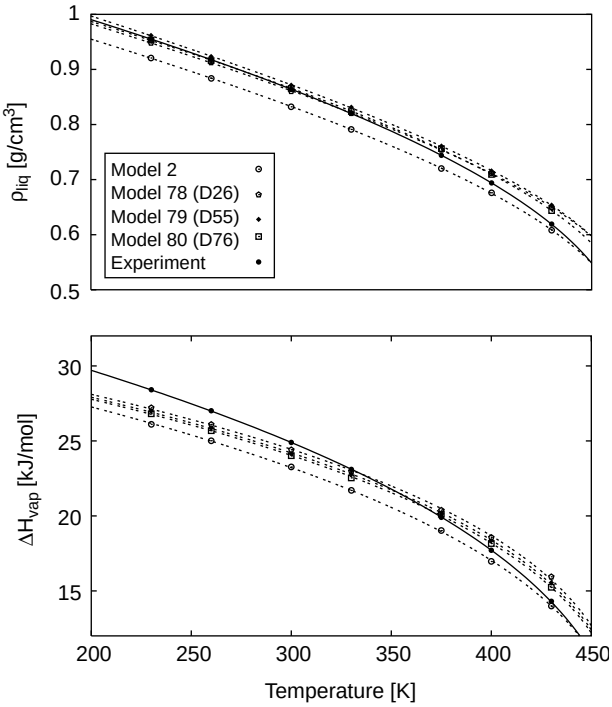

Figure 6. Computed values for liquid density (top) and heat of vaporisation (bottom) for model 2 and the models (dotted lines) predicted on the basis of three differently sized data-sets in comparison with the experimental references (solid line [13]).

located in the above-mentioned valley and lie very close to a straight line that stretches almost over the entire range of allowed ε_{OO} values (maximal difference is 23.8 K), but is confined to a quite small interval of ε_{CC} values (maximal difference is 3.72 K). The σ-values generally vary much less, the deviation from the average value amounting to 0.5% in the case of σ_{CC} and to 2.8% in the case of σ_{OO}. This is consistent with the summaries as presented by the correlation matrices, which display clear trends concerning the impact of the methylene parameters, but only a weak influence of the oxygen parameters.

4. Discussion

In our attempt to understand the trends inherent to the optimisation problem at hand, we computed physical properties for 79 models of EO, which differed in their LJ parameter values. The molecule was modelled as a rigid three-membered ring with fixed partial charges. For this fixed geometry, the properties chosen for assessing a model's performance (liquid density and heat of vaporisation) are a function of the non-bonded interactions between the molecules that are described by the LJ and the Coulomb potentials. As the partial atomic charges are uniform among the models, it is the LJ parameters, potential well depth ε and zero-potential distance σ, that decide about the exact location and steepness of the repulsive branch of the LJ potential.

When beginning our study, we assumed that increasing σ would enlarge the atomic volume and would thus lower the liquid density. Similarly, increasing ε would result in stronger intermolecular attraction and then lead to higher values for the heat of vaporisation. However, the physical properties do not react in an independent way to changes of the LJ parameters. With increasing kinetic energy, the response becomes weaker. This behaviour was clearly illustrated in the correlation matrices created by DesParO (cf. Figure S2 of the Supplementary Material). Quantitatively these effects were difficult to predict.

4.1 Role of non-bonded interactions

In order to gain a more quantitative understanding, we plotted the sum of the LJ and Coulomb potentials for all possible interatomic combinations of model 2, as shown in Figure 7. We find that the net potential describing the non-bonded interactions between methylene atoms resembles almost a hard-sphere potential, although it is entirely repulsive; due to its small partial atomic charge, the electrostatic contribution is comparatively small and almost constant over a wide range of interatomic distances. In consequence of the high ε-value, the repulsive LJ branch at small distances becomes very steep. This leads to a dramatic increase in potential energy

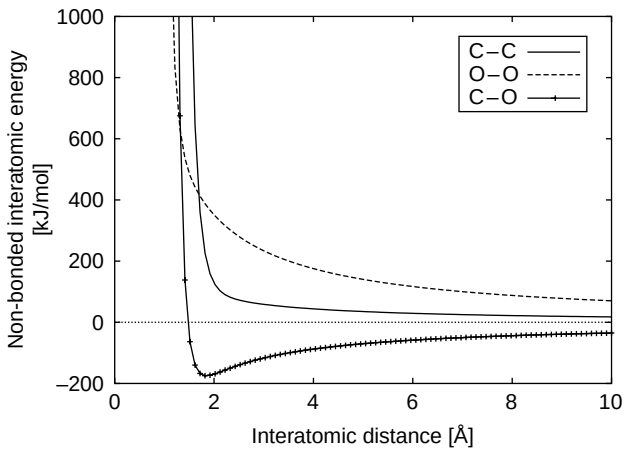

Figure 7. For model 1 as an example, the net non-bonded energy, i.e. the sum of the LJ energy and the Coulomb energy between the different atom types, clearly discriminates between the different atom-type combinations.

when the distance between two methylene atoms falls below $\sim$2.0 Å.

Similarly, the homo-nucleic interactions among oxygen atoms are purely repulsive, but as the atomic charge is doubled for the oxygen atoms, the electrostatic interactions are four times higher than those seen for the methylene–methylene combination. This leads to a much stronger repulsion between the oxygen atoms, which is growing continuously with decreasing interatomic distance. As the value for ε_{OO} is lower, the repulsive LJ branch is less steep. These effects together result in a comparatively smooth curve causing only small relative changes in energy with small changes in the interatomic distance.

Contrarily, the hetero-nucleic potential curve for oxygen and methylene atoms displays a long-range attractive part that sharply bends into a steep repulsive branch with decreasing interatomic distances. One could imagine that the oppositely charged, hetero-atomic interactions provide the most relevant part in parameterising our EO systems, as they cause an electrostatic attraction between the methylene and the oxygen atoms that stabilises intermolecular contacts in a liquid phase. The correlation matrices suggest otherwise; according to DesParO, the methylene atoms clearly dominate the system's behaviour.

This may be due to the fact that any atom of an EO molecule will encounter a methylene or an oxygen atom of a neighbouring EO molecule with a 2:1 chance. In case a particular atom were of the methylene type, this implies that it will, in two times out of three, experience the 'almost-hard-sphere' potential, i.e. nearly no interaction at all unless the atoms collide. As the molecule possesses two methylene atoms, the chance for such a kind of interaction is even doubled. In one case out of three, however, a single methylene atom will contact an oxygen atom, which may

result in an attraction between the two molecules. Again for a molecule, the chance for this combination is twofold.

In case the respective atom were of the oxygen type, it would have a methylene atom as a partner with a 2:1 chance, again leading to intermolecular attraction. Only in one time out of three, it would encounter the repulsive contact to another oxygen atom. Briefly, the methylene atoms participate with much higher probability in intermolecular interactions, which easily explains their dominance in the pattern illustrated in the correlation matrices. Additionally, the slope of the potential curve for the oxygen–oxygen combination is comparatively shallow, which is mainly caused by the contribution of the electrostatic potential and uniform among all models. Changes in σ_{OO} or ε_{OO} will therefore have no significant impact. Consequently, the ranges of methylene parameters are limited to comparatively small intervals, whereas the oxygen parameters allow larger modifications. Thus, our initial assumption that the LJ parameters of both atom types had an equivalent impact on the system's behaviour was obviously not confirmed. In retrospect, the potential curves for the non-bonded interactions explain straightforwardly the behaviour as summarised in the correlation matrix. In future, we will be able to interpret the potential curves more carefully when developing a suitable strategy for optimisation. Still, the balance between LJ and electrostatic interactions is very delicate, and we doubt it can be predicted merely on the basis of the potential curves.

4.2 Creation of suitable parameter sets

Another important result is that we were indeed able to identify several equally good parameter sets, either by coincidence in a stochastic data-set or by interactive parameter selection guided by DesParO based on said data. The inability to find a single best solution may be a natural feature of the chosen system, but most certainly originates from the natural indeterminacy of simulation results. Therefore, we expected mainly to identify a promising region in the parameter space, which was indeed achieved. DesParO created a useful meta-model already for the smallest data-set containing only 26 individual parameter sets. This implies that, for similar systems, a sample number falling below the demand of $C(2+n+n(n+1)/2)$ data-sets, where $C = 5$, may be sufficient to study the problem's key features.

We were, however, not able to find a set of LJ parameters that would reproduce the experimental curves perfectly. The selected best solutions relate to parameter values that lie quite central within the parameter space. Thus, it is unlikely we missed better solutions if we assume that the margins of the considered parameter space have been well defined.

If we have reached the limit of LJ parameter optimisation, as we believe, then this is indicative that error resides in another aspect of the force field. One possibility is that the partial atomic charges are themselves not validated with experiment. It is clear that the non-bonded interactions between molecules are governed by the LJ parameters and the partial atomic charges when they are derived to reproduce an experimental non-bonded observable. The exact extent of this dependency is currently unknown. However, from previous experience, we have seen that the LJ parameters have a larger impact on calculating observables, while the partial atomic charges can be adjusted to 'fine tune' the intermolecular interaction. Due to the dependency, the optimisation of both the LJ parameters and the partial atomic charges requires an iterative optimisation procedure.

Another possibility is that the error arises from the functional form of the force-field equation. It is possible that an alternative van der Waals equation (e.g. Buckingham potential or a 9-6 potential) would allow for parameter optimisation that better fits the experimental curves. The current force field is also lacking polarisation, which also may allow for better reproduction of the curves. These are issues that we intend to investigate more in the future.

In cases where neither approach is an option, DesParO presents us with the facility of pragmatically choosing temperature-dependent models for the particular temperature range of interest.

5. Conclusions

Applying DesParO to our optimisation problem provided us first with valuable insights about the EO model. As a consequence, these insights may suggest how to proceed efficiently both with the current system and with the overall optimisation strategy. From the correlation matrices, we learned that the influence of the LJ parameters of one atom type dominates over the other ones. The number of methylene atoms exceeds the oxygen atoms by a factor of 2; it was not expected beforehand that they almost completely control the system's behaviour. Hence, the methylene parameters appear to be quite well defined, as we were able to limit the parameter ranges to comparatively small intervals. Secondly, we found that the heat of vaporisation was largely independent from the σ-values. That was unexpected and can be utilised for a new optimisation strategy. On the basis of such information, further optimisation might be divided into sub-tasks, especially when facing a large number of adjustable parameters. A sequential procedure may be more efficient than an attempt to optimise all variable parameters at once. For instance, we might prioritise certain atom types over others (e.g. trim the methylene atom type first, and then do fine-tune by customising the parameters for oxygen). Likewise, we might try a hierarchical procedure that reflects the diverse criteria (e.g. first adjusting ε-values

to match the desired heat of vaporisation, and afterwards optimising σ to correct for the resulting density).

When splitting up a complex optimisation task into smaller sub-tasks and narrowing down the respective parameter ranges, optimisation methods such as GROW [7,8] that directly search for the global optimum in a confined parameter space should finally lead to reproducible and satisfying final parameter sets.

Alternatively, also avoiding too much manual intervention, we might evaluate the meta-model in an automated fashion by Pareto front extraction as implemented in DesParO as well. By way of that algorithm, one identifies and ranks the best compromises among arbitrary criteria. Moreover, one could facilitate visualisation of dynamic sensitivity analyses (a feature of the upcoming DesParO version, cf. [11]). This would allow for a local resolution of the correlation matrix, thus enhancing the predictive power of multi-objective optimisation methods and tools such as DesParO.

References

[1] M.H. Ketko and J.J. Potoff, *Effect of partial charge parameterization on the phase equilibria of dimethyl ether*, Mol. Sim. 33 (2007), pp. 769–776.

[2] T. Köddermann, D. Paschek, and R. Ludwig, *Molecular dynamic simulations of ionic liquids: A reliable description of structure, thermodynamics and dynamics*, Chem. Phys. Chem. 8 (2007), pp. 2464–2470.

[3] K.N. Kirschner, A.B. Yongye, S.M. Tschampel, J. Gonzalez-Outeirino, C.R. Daniels, B.L. Foley, and R.J. Woods, *GLYCAM06: A generalizable biomolecular force field*, Carbohydrates. J. Comput. Chem. 29 (2008), pp. 622–655.

[4] R. Faller, H. Schmitz, O. Biermann, and F. Müller-Plathe, *Automatic parameterization of force fields for liquids by simplex optimization*, J. Comp. Chem. 20 (1999), pp. 1009–1017.

[5] T.J. Müller, S. Roy, W. Zhao, A. Maaß, and D. Reith, *Economic simplex optimization for broad range property prediction: Strengths and weaknesses of an automated approach for tailoring of parameters*, Fluid Phase Equilibr. 274 (2008), pp. 27–35.

[6] B. Eckl, J. Vrabec, and H. Hasse, *On the application of force fields for predicting a wide variety of properties: Ethylene oxide as an example*, Fluid Phase Equilibr. 274 (2008), pp. 16–26.

[7] M. Hülsmann, T. Köddermann, J. Vrabec, and D. Reith, *GROW: A gradient-based optimization workflow for the automated development of molecular models*, Comp. Phys. Comm. (2009), doi: 10.1016/j.cpc.2009.10.024.

[8] M. Hülsmann, J. Vrabec, A. Maaß, and D. Reith, *Assessment of numerical optimization algorithms for the development of molecular models*, Comp. Phys. Comm. (2009), submitted for publication.

[9] M.G. Martin, *Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor–liquid coexistence curves and liquid densities*, Fluid Phase Equilibr. 248 (2006), pp. 50–55.

[10] A. Stork, C.A. Thole, S. Klimenko, I. Nikitin, L. Nikitina, and Y. Ashtakhov, *Towards interactive simulation in automotive*, Vis. Comput. 24 (2008), pp. 947–953.

[11] T. Clees, *Computer-aided robust design for multi-disciplinary processes*, Proceedings of the 10th MpCCI User Forum, Sankt Augustin, Germany, 2009.

[12] T. Clees, N. Hornung, I. Nikitin, L. Nikitina, and D. Steffes-lai, *DesParO User's Manual*, Release1.5, Fraunhofer SCAI, 2009.

[13] C. Buckles, P. Chipman, M. Cubillas, M. Lakin, D. Slezak, D. Townsend, K. Vogel, and M. Wagner, *Ethylene oxide user's guide*, 1999. Available at http://www.ethyleneoxide.com.

[14] J.D. Olson and L.C. Wilson, *Benchmarks for the fourth industrial fluid properties simulation challenge*, Fluid Phase Equilibr. 274 (2008), pp. 10–15.

[15] P.A. Wielopolski and E.R. Smith, *Molecular dynamics study of dielectric behaviour and orientational correlations of liquid ethylene oxide (oxirane)*, Mol. Phys. 54 (1985), pp. 467–478.

[16] R.D. Mountain, *A polarizable model for ethylene oxide*, J. Phys. Chem. B 109 (2005), pp. 13352–13355.

[17] M.H. Ketko, J. Rafferty, J.I. Siepman, and J.J. Potoff, *Development of the TraPPE-UA force field for ethylene oxide*, Fluid Phase Equilibr. 274 (2008), pp. 44–49.

[18] X. Li, L. Zhao, T. Cheng, L. Liu, and H. Sun, *One force field for predicting multiple thermodynamic properties of liquid and vapor ethylene oxide*, Fluid Phase Equilibr. 274 (2008), pp. 36–43.

[19] C.I. Bayly, P. Cieplak, W.D. Cornell, and P.A. Kollman, *A well-behaved electrostatic potential based method using charge restraints for determining atom-centered charges: The RESP model*, J. Phys. Chem. 97 (1993), pp. 10269–10280.

[20] M.W. Schmidt, K.K. Baldridge, J.A. Boatz, S.T. Elbert, M.S. Gordon, J.H. Jensen, S. Koseki, N. Matsunaga, K.A. Nguyen, S.J. Su, T.L. Windus, M. Dupuis, and J.A. Montgomery, *General atomic and molecular electronic structure system*, J. Comput. Chem. 14 (1993), pp. 1347–1363.

[21] M.S. Gordon and M.W. Schmidt, *Advances in electronic structure theory: GAMESS a decade later*, in *Theory and Applications of Computational Chemistry, the first forty years*, C.E. Dykstra, G. Frenking, K.S. Kim, and G.E. Scuseria, eds., Elsevier, Amsterdam, 2005, pp. 1167–1189.

[22] A. Panagiotopoulos, *Direct determination of phase coexistence properties of fluids by Monte Carlo simulation in a new ensemble*, Mol. Phys. 61 (4) (1987), pp. 813–826.