

RESEARCH ARTICLE

Automated development of force fields for the calculation of thermodynamic properties: Acetonitrile as a case study

Stephan Deublein^a, Patrick Metzler^b, Jadran Vrabec^c, Hans Hasse^a *

^a*Lehrstuhl für Thermodynamik, Universität Kaiserslautern, 67653 Kaiserslautern, Germany;*

^b*Institut für Automatisierungsinformatik, Hochschule RheinMain, 65428 Rüsselsheim, Germany;*

^c*Lehrstuhl für Thermodynamik und Energietechnik, Universität Paderborn, 33098 Paderborn, Germany;*

(Received 00 Month 200x; final version received 00 Month 200x)

Force fields for engineering applications are often parameterized using strategies based on quantum mechanical ab initio calculations and thermodynamic properties from experiment. An automated procedure for adjusting molecular model parameters to experimental thermodynamic property data is introduced. The process accelerates the development of molecular models by an efficient use of parallel computing power and an autonomous progress of the model development without any user interaction. As a case study, the procedure is applied to the parameterization of a molecular model for acetonitrile. The resulting model reproduces vapor-liquid equilibrium data of acetonitrile with an accuracy of 0.1% for the saturated liquid density, 4.9% for the vapor pressure and 3.7% for the enthalpy of vaporization. These accuracies are superior to data obtained with previously published force fields for acetonitrile

Keywords: molecular modeling, force field, optimization, automation, acetonitrile

1 Introduction

Molecular modeling and simulation is a promising route for the prediction of thermodynamic properties of pure substances and mixtures. The more widespread use of this technique in engineering applications was restricted for a long time by two shortcomings [1]: first, the lack of simulation programs that efficiently yield the thermodynamic properties of industrial interest; and second, the poor availability of suitable molecular models, which yield these properties with the necessary accuracy at a moderate simulation effort.

In the recent past, both issues were addressed by our group. With the release of the simulation program *ms2* [2] in 2011, a powerful simulation tool for the calculation of thermodynamic properties is now publicly available. The scope of thermodynamic properties that are accessible with this program ranges from basic static properties over Henry's law constant and entropic data to transport properties of bulk fluids. The accuracy of the calculated data is high, satisfying current requirements expressed by the industry for their applications. With respect to simulation time, the program makes efficient use of modern computing hardware for a large variety of different computer architectures. It is executed in parallel, which reduces the response time significantly [2]. The lack of accurate force fields for industrially relevant molecules was addressed by many groups [3–13]. Recent efforts of our group in that field [7] resulted in a systematic approach for the development of such models using Lennard-Jones (LJ) type force fields with superimposed electrostatic sites. Quantum mechanical (QM) ab initio calculations were employed to determine the geometry and the permanent electrostatics of a molecule. These data were directly passed on to the molecular force field. The dispersive and repulsive interactions were modeled by LJ potentials. Their parameters σ and ε were adjusted to vapor-liquid equilibrium

Corresponding author. Email: hans.hasse@mv.uni-kl.de

(VLE) data, namely saturated liquid density, vapor pressure and enthalpy of vaporization of the pure fluid. This strategy was successfully applied to the development of rigid, non-polarizable molecular force fields of comparatively small molecules up to the size of cyclohexanol [14].

For larger molecules, which require a larger number of molecular sites, however, the adjustment of the LJ parameters to VLE data becomes more and more complex and tedious. In such cases, the model parameter adjustment can be facilitated by automation procedures that execute all required steps without user interaction [15–17]. In the present work, an automation procedure was developed, which efficiently supports the user in the parameter adjustment. The parameterization strategy is introduced in Section 2. The process was validated by the development of a molecular force field for acetonitrile, which is presented in Section 3, while in Section 4, the work is concluded. Note that a routine that steers the present automated process is publicly available upon prior registration at <http://www.ms-2.de>. It is designed for the use with *ms2* [2], but can easily be adapted to other simulation engines.

2 Molecular model properties

The geometry as well as the electrostatics of molecules can routinely be determined by QM ab initio calculations. A detailed description of a QM based parameterization strategy for the geometry was reported by Eckl et al. [18], which is summarized here briefly.

The geometry of the molecular models, i.e. bond lengths, angles and dihedrals, is directly passed on from QM calculations. The geometry optimization is carried out using GAMESS(US) [19]. The Hartree-Fock level of theory is applied with a relatively small (6-31G) basis set. For determining the charge distribution of the molecule of interest, the Møller-Plesset 2 level of theory is used that takes into account electron correlation in combination with the polarizable 6-311G(d,p) basis set. The calculation of the electrostatic moments for the development of engineering molecular models is preferably done for a liquid-like state. This is achieved by placing the molecule within a dielectric continuum and assigning the experimental dielectric constant of the liquid to the continuum via the COSMO method [20]. From the resulting electron density distribution, point charges, point dipoles and point quadrupoles are estimated by a simple multipole expansion in a user-defined position, typically the molecular center of mass. The use of such a multipole expansion for modeling permanent electrostatics is advantageous, since it allows for a compact but nonetheless detailed description of the interaction energy from the charge distribution [21]. In many cases, the magnitudes and orientations of the resulting electrostatic interaction sites are such a good approximation for the charge distribution that they do not require any further modification.

The dispersive and repulsive interactions between the molecules are usually reduced to pairwise interactions, which are modeled by the LJ 12-6 potential. This potential relies on two parameters which cannot be suitably predicted by QM calculations. Typically, both parameters are optimized with respect to VLE data of the pure fluid, namely saturated liquid density, vapor pressure and enthalpy of vaporization. Many different procedures are available for a stable and efficient optimization, e.g. gradient based algorithms [22] or the Gauss-Newton least square estimator [6]. In our group, the optimization is carried out following a scheme by Stoll [23]. For a set of N_a data points a_i , the square of the relative deviations $a_{i,\text{rel}}$ of simulation results compared to experimental data is minimized

$$\frac{1}{N_a} \sum_i^{N_a} (G_{ii} a_{i,\text{rel}}(\mathbf{m}))^2 = \frac{1}{N_a} \sum_i^{N_a} \left(G_{ii} \frac{a_{i,\text{Exp}} - a_{i,\text{Sim}}(\mathbf{m})}{a_{i,\text{Exp}}} \right)^2 \stackrel{!}{=} \min . \quad (1)$$

Here, the vector \mathbf{m} represents the set of N_m model parameters that are subject to optimization. The relative deviations between experimental data and simulation results are weighted by the diagonal $N_a \times N_a$ matrix \mathbf{G} that individually scales the contributions of each considered property

a_i in the minimization function (1). The functional dependence of the relative error $a_{i,\text{rel}}$ on the model parameter m_j is approximated by a first order Taylor expansion developed in the vicinity of the original parameter set. The required, but a priori unknown, sensitivities $S_{ij} = \partial a_{i,\text{rel}} / \partial m_j$ are estimated from individual simulations, in which m_j is varied. Thus, the resulting solution $\mathbf{m}^{(s+1)}$ for the linearized optimization problem according to Eq. (1) based on parameter set $\mathbf{m}^{(s)}$ is given by

$$\mathbf{G}\mathbf{S}^{(s)}\Delta\mathbf{m}^{(s)} = \mathbf{G}\mathbf{a}_{\text{rel}}^{(s)}, \quad (2)$$

and

$$\mathbf{m}^{(s+1)} = \mathbf{m}^{(s)} + \Delta\mathbf{m}^{(s)}, \quad (3)$$

where $\mathbf{S}^{(s)}$ is the $N_a \times N_m$ matrix of the sensitivities S_{ij} of the optimized properties on the model parameters.

On the basis of experimental data \mathbf{a}_{Exp} , solving the optimization starting problem from a physically reasonable, initial model $\mathbf{m}^{(1)}$ is straightforward: molecular simulation is applied to determine $\mathbf{a}_{\text{Sim}}^{(1)}$ and hence $\mathbf{a}_{\text{rel}}^{(1)}$ as well as $\mathbf{S}^{(1)}$. From Eqs. (2) and (3), the solution $\mathbf{m}^{(2)}$ is determined. Repeating the scheme over a certain number of iteration steps results in an optimized molecular model. The iteration is terminated when a desired accuracy of the model is reached or no significant progress is achieved in the iteration scheme. The time required for one iteration, i.e. for generating the model $\mathbf{m}^{(s+1)}$ from model $\mathbf{m}^{(s)}$, is dominated by the molecular simulations that need to be performed for $\mathbf{a}_{\text{Sim}}^{(s)}$ and $\mathbf{S}^{(s)}$. It can be reduced by the use of a simulation program that efficiently exploits multicore computing resources, such as *ms2* [2], and by performing all necessary simulation runs in parallel.

Starting from a physically reasonable model, the present automated algorithm allows for an user independent execution of all operations required in the optimization process with respect to experimental VLE data, namely saturated liquid density ρ' , vapor pressure p and enthalpy of vaporization Δh_v . Note that these data can be reliably measured experimentally over a wide range of temperature and are hence available in the literature for many industrially relevant substances [24]. Note also that in molecular simulation, various efficient algorithms are available, e.g. Gibbs-Ensemble MC [25] and the Grand Equilibrium method [26], to determine VLE data with low statistical uncertainties. For each iteration step in the parameter adjustment, the automated tasks are the initiation and evaluation of the performed simulation runs as well as all actions that are required to prepare and perform the optimization. Technical details are given in the Appendix.

For the evaluation of the molecular model quality and the optimization of model parameters, all VLE data determined by molecular simulation are regressed over a temperature range between the triple point and the critical point. This regression is performed with temperature dependent fits for the saturated liquid density ρ' , dew density ρ'' and vapor pressure p following Lotfi et al. [27]

$$\rho' = \rho_c + D_1(T_c - T)^{1/3} + D_2(T_c - T) - D_3(T_c - T)^{3/2}, \quad (4)$$

$$\rho'' = \rho_c - E_1(T_c - T)^{1/3} + E_2(T_c - T) - E_3(T_c - T)^{3/2}, \quad (5)$$

$$\ln p = C_1 T - C_2/T - C_3/T^4, \quad (6)$$

where C_i , D_i and E_i are model specific constants, which are adjusted to the simulation data. Employing such functions for the description of simulation data has two advantages: first, they

allow for an inter- and extrapolation of simulation data $a_{i,\text{Sim}}$, which are typically determined for few discrete temperatures only, to a wide range of temperatures, where experimental reference data are available. This also includes an extrapolation of simulation data to properties that are not directly accessible, such as critical data, so that these can be included in the investigation of the fluids. The second advantage is that the scatter in the simulation data due to statistical uncertainties are smoothed by the regression functions. This allows for the determination of a continuous sensitivity $S_{ij}^{(s)}$ of property $a_{i,\text{rel}}^{(s)}$ on $m_j^{(s)}$ over the entire range of studied temperatures and hence, for a more accurate estimation of the optimized parameter set $\mathbf{m}^{(s+1)}$.

The automated execution of all tasks facilitates the molecular model parameter adjustment significantly. An alternative manual execution of the iteration steps is tedious, time consuming and, hence, often leads to an unsystematic exploration of the parameter space. The stopping criterion is often not the optimal solution of Eq. (1), but determined by a simple time out.

However, note that due to the automated execution of all required tasks for one iteration step, problems may occur which are related to the physical nature of the studied problem; e.g. it is not guaranteed that each parameter set $\mathbf{m}^{(s)}$ that is studied during the optimization yields all experimental target data \mathbf{a}_{Exp} at the conditions of interest. E.g., for a certain parameter set, VLE may not be present at a given temperature, if the critical temperature of the model is too low. During this automation, such data are excluded according to specific criteria.

3 Case study: Acetonitrile

A force field for acetonitrile was developed to test the applicability and efficiency of the present automation process introduced above. All required simulations were performed with the simulation program *ms2* [2], which determines VLE data with the Grand Equilibrium method [26]. Technical details on the simulations are given in the Appendix. However, note that these simulation details with respect to the sampling have to be specified appropriately, e.g. when molecules are regarded that are more demanding than acetonitrile.

Following the united-atom approach, acetonitrile was modeled by three LJ sites with one superimposed point dipole. The geometry of acetonitrile was taken from preceding work [18], cf. Figure 1 and Table 1, that used the QM based procedure mentioned above. The electrostatic interactions were determined from the electron density distribution at discrete positions and a multipole expansion. This led to a single point dipole located at a distance of 0.695 Å from the nitrogen atom shifted towards the carbon atom, cf. Table 1. The dipole was oriented with its positive end towards the methyl site. Throughout the study, the internal molecular degrees of freedom were neglected. This assumption is reasonable for acetonitrile, since the molecule is small enough to show only a minor dependence of its thermodynamic properties on its internal motions. The complete force field thus writes as

$$u_{ij} = \sum_{k=1}^{N_{s,i}} \sum_{l=1}^{N_{s,j}} 4\varepsilon_{kl} \left(\left(\frac{\sigma_{kl}}{r_{kl}} \right)^{12} - \left(\frac{\sigma_{kl}}{r_{kl}} \right)^6 \right) + \frac{1}{4\pi\epsilon_0} \frac{\mu_i \mu_j}{r_{ij}^3} (\sin \theta_i \sin \theta_j \cos \phi_{ij} - 2 \cos \theta_i \cos \theta_j), \quad (7)$$

where r_{kl} is the distance between two LJ sites k and l of the interacting molecules, θ_i is the angle between the dipole direction and the distance vector of the two interacting dipoles and ϕ_{ij} is the azimuthal angle of the two dipole directions. σ_{kl} and ε_{kl} denote the LJ parameters. Note that throughout this study, the Lorentz-Berthelot [28, 29] combining rules were applied for the interactions between unlike LJ sites.

The parameter set for the LJ interaction sites of the nitrogen atom and of the methyl group, i.e. $N_m = 4$ parameters, were subject to optimization with respect to VLE data of the pure fluid, namely saturated liquid density ρ' , vapor pressure p and enthalpy of vaporization Δh_v . The accuracy of the molecular model required after a successful optimization, i.e. the average of the absolute values of the relative deviations for each property between experimental data

and simulation results using the optimized model, were specified to be not larger than 1% for ρ' , 5% for p and 10% for Δh_v over the temperature range $0.6 < T/T_{c,\text{Exp}} < 0.99$, where T_c is the critical temperature of acetonitrile. Note that the LJ parameters of the third site, i.e. the carbon site, showed hardly any effect on the VLE properties in an earlier study [18] and were thus assumed constant here [18]. The value of $\sigma_C = 2.81 \text{ \AA}$ and $\varepsilon_C/k_B = 10.64 \text{ K}$ as proposed by Eckl et al. [18] were used.

In this study, experimental VLE data were obtained via correlations of experimental measurements from the literature [24]. VLE data were obtained by simulation at five temperatures ($N_T = 5$) explicitly, namely at 270, 360, 420, 490 and 518 K. These results were regressed according to Lotfi et al. [27] so that data in the temperature range of $0.6 < T/T_c < 0.99$ was available. The regression functions for the VLE data were evaluated in intervals of 2 K in the studied temperature range, i.e. for $N_{T,\text{corr}}$ temperatures, and compared to experimental data. The weighting factors for the individual properties were set to one for ρ' , four for p and 14 for Δh_v . These values were chosen to reflect the accuracies that were demanded on the optimized model for each investigated VLE property type, i.e. they define what can be expected from the model. They have been successfully employed in numerous studies, e.g. to describe cyclohexanol [14], and showed a fast convergence of the cost function.

Hence, the explicit form of the optimization problem that had to be solved in the present model development was

$$\begin{aligned} & \frac{1}{3N_{T,\text{corr}}} \sum_{k=1}^{N_{T,\text{corr}}} \left(\frac{\rho'_{\text{Exp}}(T_k) - \rho'_{\text{Sim}}(T_k)}{\rho'_{\text{Exp}}(T_k)} \right)^2 + \frac{1}{3N_{T,\text{corr}}} \sum_{k=1}^{N_{T,\text{corr}}} \frac{1}{4^2} \left(\frac{p_{\text{Exp}}(T_k) - p_{\text{Sim}}(T_k)}{p_{\text{Exp}}(T_k)} \right)^2 + \\ & + \frac{1}{3N_{T,\text{corr}}} \sum_{k=1}^{N_{T,\text{corr}}} \frac{1}{14^2} \left(\frac{\Delta h_{v,\text{Exp}}(T_k) - \Delta h_{v,\text{Sim}}(T_k)}{\Delta h_{v,\text{Exp}}(T_k)} \right)^2 \stackrel{!}{=} \min . \end{aligned} \quad (8)$$

Following the approach of Stoll [23] for solving Eq. (8), the sensitivity of the observables with respect to the model parameters $\partial a_{i,\text{rel}}/\partial m_j$ was determined by variations of 2% of the LJ size parameter σ and 5% of the LJ energy parameter ε . VLE data for the current model and all N_m model variations that were required for the calculation of the sensitivities S_{ij} were determined simultaneously throughout the model development. Hence, an overall of $N_T \cdot (N_m + 1)$ VLE simulation runs were executed in parallel for each iteration, i.e. $5 \times 5 = 25$ simulation runs.

Note that the uncertainty of the cost function Eq. (8) is dependent on the accuracy of the performed simulations and hence the quality of the regression functions that correlate the data.

3.1 Optimization pathway

The initial parameter set $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(1)}$ for acetonitrile was taken from previous work by Eckl et al. [18]. However, since the electrostatic sites of the acetonitrile model were changed, these parameters were not expected to yield good results. The initial force field underpredicted the experimental saturated liquid density significantly, cf. Figure 2. Based on the results for model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(1)}$ and the sensitivity calculations, a new model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(2)}$ for acetonitrile was generated, which showed significant changes for the LJ size parameter of the methyl group and of the nitrogen atom (cf. Figure 3). The LJ energy parameters were altered to slightly smaller values.

VLE data calculated with $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(2)}$ show an improvement for all considered properties, especially for the saturated liquid density and the vapor pressure. This is most evident in terms of the critical data, which increased from 80% ($\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(1)}$) to 95% of the experimental critical temperature and from 25% ($\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(1)}$) to 83% of the experimental critical pressure. However, $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(2)}$ still underestimated the target data systematically. Looking at the main contributions for the further optimization, the cost function (8) of $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(2)}$ was still dominated by the lack of a phase

transition at high temperatures, cf. Table 2.

The subsequent model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(3)}$ reproduced the saturated liquid density well, while the vapor pressure and the enthalpy of vaporization showed deviations of 7% and 5%, respectively. Looking at the cost function, cf. Table 2, the pressure at low temperatures became the most outlying quantity. This was expected, since the vapor pressure is a very sensitive property, especially at low temperatures. Based on model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(3)}$, only slight changes of the parameters were made for minimizing the cost function (8) and hence obtaining a new molecular model.

Model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(4)}$ reproduced the vapor pressure with a higher accuracy. However, this improvement corrupted the preceding good agreement for the saturated liquid density. With $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(5)}$, a good agreement for the vapor pressure and the saturated liquid density at low temperatures was achieved, while the behavior at high temperatures was not well reproduced.

With model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(6)}$, the overall deviations from experimental data were determined to be 0.5% for the saturated liquid density, cf. Figure 4, 4.9% for the vapor pressure, cf. Figure 5, and 3.7% for the enthalpy of vaporization, cf. Figure 6. The accuracy of these data was already within the specification for the optimization. This shows that a successful model parameter optimization can be performed automatically. Note here that the cost function was monotonically falling, except for model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(4)}$, although the variations of the molecular model parameters seem to be arbitrary for each iteration step, cf. Figure 3.

To test whether even better results can be achieved, the automation was continued. This led to model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(7)}$, which showed large deviations for all considered VLE properties. Such a behavior was expected, since close to the minimum of the cost function, the sensitivity of the parameters on the reference observables is small. It is well known that gradient based algorithms show difficulties predicting better parameter sets under such conditions [22].

Performing an additional iteration, the model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(8)}$ again described the VLE data with the specified accuracy, cf. Figures 4 to 6. The quality of $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(8)}$ is similar to the one of model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(6)}$, although the model parameters differ significantly. This shows that there are various parameter sets that solve the optimization problem. Furthermore, it shows that the optimization scheme proposed here, is convergent from different physically reasonable parameter sets.

Note that all results discussed above were obtained with simulation runs containing 500 molecules in the liquid phase and 864 molecules in the vapor phase, cf. Appendix B. The models $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ and $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(8)}$, however, were further assessed with simulations containing 1372 molecules in the liquid phase and again 864 molecules in the vapor phase. The results obtained with these simulations confirmed the quality of these models.

3.2 Final molecular model

A molecular model for acetonitrile based on the LJ approach with a superimposed dipole was developed. The geometry of the molecule and its electrostatics were calculated by ab initio QM, while four LJ parameters were adjusted to experimental VLE data of the pure fluid [24] using an automated optimization procedure. This resulted in two parameter sets that reproduce the experimental data according to the specifications. Note that of both the parameter sets are given in Table 1.

The molecular model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ reproduces the experimental reference data with average deviations of 0.5%, 4.9% and 3.7% for saturated liquid density ρ' , vapor pressure p and enthalpy of vaporization Δh_v over a temperature range $0.6 < T/T_{c,\text{Exp}} < 0.99$. For model $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(8)}$, the according deviations are 0.1%, 4.7% and 3.9%. The agreement with the experiment for both of these molecular models is hence superior in comparison to the one reported by Eckl et al. [18].

The critical data for acetonitrile determined by molecular simulation with models $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ and $\mathbf{m}_{\text{C}_2\text{H}_3\text{N}}^{(8)}$ match with the experimental data excellently, cf. Figures 4 to 6. The critical density

from simulation deviates from measurements [30, 31] by roughly 2%, being 5.61 mol/l for both $m_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ and $m_{\text{C}_2\text{H}_3\text{N}}^{(8)}$, where the experimental value is 5.78 mol/l. The critical temperature is within 0.4% and 0.1% of the experimental value of 545.46 K [30, 31] and the vapor pressure within 1.6% and 1.4% of the experimental value of 4.83 MPa [31, 32] for model $m_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ and $m_{\text{C}_2\text{H}_3\text{N}}^{(8)}$, respectively.

Based on these results, both models can be applied in predictive calculations of thermodynamic properties. Nevertheless, the use of model $m_{\text{C}_2\text{H}_3\text{N}}^{(8)}$ is recommended, since it best solves the optimization problem, cf. Eq. (8).

4 Conclusion

An automated procedure was presented for the parameter optimization of molecular models based on the LJ approach with superimposed point charges to experimental VLE data. It was developed and tested in combination with the simulation program *ms2*, the Grand Equilibrium method for the calculation of VLE data and an optimization algorithm proposed by Stoll [23], but it can be easily adapted to other simulation programs and optimization schemes. The parameter optimization is fast in terms of model development time and requires no user interaction. The automation works in a Linux environment and requires no commercial software installation. The functionality of the automation and the convergence of the optimization scheme according to Stoll [23] was shown for acetonitrile. Thereby, a new molecular model for acetonitrile was developed that reproduces the VLE data of this pure fluid with a high accuracy.

The automation algorithm is publicly available upon prior registration on <http://www.ms-2.de>.

Acknowledgements

The authors gratefully acknowledge financial support by the BMBF "01H08013A - Innovative HPC-Methoden und Einsatz für hochskalierbare Molekulare Simulation" and computational support by the Steinbuch Centre for Computing under the grant LAMO and the High Performance Computing Center Stuttgart (HLRS) under the grant MMHBF. The present research was conducted under the auspices of the Boltzmann-Zuse Society of Computational Molecular Engineering (BZS).

Appendix A - Automation details

The automation procedure presented here was developed for the optimization of molecular model parameters with respect to VLE data, namely saturated liquid density, vapor pressure and enthalpy of vaporization of the pure fluid. During the optimization, no interaction with the user is required.

The employed algorithms were specifically designed for the use of external computing resources (ECR) for the time consuming molecular simulation runs, e.g. at computing centers. All remaining process steps are executed on a local computing resource (LCR) in order to avoid difficulties imposed by constraints on the ECR, such as maximum disc space, missing permissions for the automated execution of programs, maximum run time, etc. The connection between LCR and ECR was realized via the secure shell approach (ssh) and public key authorization, the common route for login on external information technology resources.

Currently, the automation algorithm makes use of the simulation program *ms2* [2], which determines VLE data with the Grand Equilibrium method [26]. *ms2* exploits current multicore computing hardware by an efficient parallelization of simulations with the molecular dynamics (MD) as well as the Monte-Carlo (MC) technique. However, an adaption of the automation

to other optimization schemes, e.g. steepest gradient [22], other simulation methods, e.g. the widely used Gibbs-Ensemble-MC method [33], or other programs, e.g. *MCCCS Towhee* [34], is straightforward.

The structure of the automation is illustrated in Figure 7. The different tasks are divided into two executables: *Autochecker* and *Optimization.sh*. In both files, intrinsic Linux functions were used, thus no proprietary software is required.

The first executable *Autochecker* contains all algorithms that are required to transfer data from LCR to ECR and vice versa. The execution of these tasks relies on the *scp* command, which is based on the secure shell approach. In addition, *Autochecker* contains algorithms to manage the molecular simulation runs that are executed on the ECR remotely triggered from the LCR, i.e. molecular simulations are initiated and checked with respect to their execution status. The checks are performed in regular time intervals, taking advantage of the Linux concept of crontabs¹. The time span between two checks is recommended to be less than one hour. Using *ms2* for the molecular simulations, a simulation run has terminated, if the simulation run restart file (*.rst) has been created. Applying other simulation programs, the tag for termination has to be replaced, which can straightforwardly be done in *Autochecker*.

The second executable *Optimization.sh* is a bash script that focuses on the optimization of the model parameters based on the simulation results and the experimental target data. This script employs resources on the LCR only so that no connection with the ECR has to be established. The most important tasks initiated by *Optimization.sh* are: extraction of the relevant data from the simulation result files, evaluation of that data with respect to the experimental data \mathbf{a}_{Exp} , data storage, optimization of the model parameters and termination of the entire process. Note that most of the algorithms required for these tasks are stored in separate files so that the *Optimization.sh* file has a modular structure. This approach allows for a straightforward modification, which is necessary when using other simulation programs rather than *ms2* or when using a different optimization algorithm.

VLE data are extracted from the simulation result files using the Perl script *res2vledata*, which is designed for the simulation result file format generated by *ms2*. All data are summarized in a tabular form in an ASCII file labeled *VLEDATA*. For each force field sampled during optimization, such a file is generated. The pooled data as well as the raw information are stored in a directory for documentation.

The *VLEDATA* files form the basis for the evaluation of the quality of the employed parameter set as well as the parameter optimization. The result of the optimization, i.e. the optimized force field parameters, is written to the potential model file in *ms2* format (*.pm). Additionally, all parameter sets required for the subsequent optimization step are generated.

The optimization algorithm itself is implemented in MATLAB R2009b². For an automated execution initiated by the script *Optimization.sh*, the algorithm has to be translated into standard C code using the commercial MATLAB-C converter of the MathWorks group, since native m-code can not be called by Linux bash scripts. Throughout the ongoing optimization process, the quality of the current parameter set as well as the expected quality of the optimized molecular model are plotted along with the experimental target data and stored in encapsulated postscript format for analysis and documentation. Note that for the use of the automation procedure, the freely accessible MATLAB runtime environment is sufficient. A MATLAB license is only required for changes of the optimization scheme.

Appendix B - Automation input and output

Only at the beginning of the automation process, user interaction is required. The user has to set up the simulation conditions, to specify the set of molecular model parameters that are consid-

¹<http://www.crontab.org>

²The MathWorks, Inc.

ered for optimization, together with their initial values, and to assign the computing resources where the simulation runs are to be executed. The according basic data are specified in form of two ASCII files:

ControlData.dat: This file contains all information with respect to the optimization process, i.e. the number of parameters that are considered for parameterization, the current status of the parameterization, the current iteration step and the number of temperatures at which VLE data are simulated.

DirectoryPath: This file contains the information about the ECR, i.e. login name and directory path, where the simulations are executed.

The parameter set considered for adjustment is specified directly in the potential model file. The parameters are marked by the tag “#adjust”.

Apart from these data, the model optimization runs fully independently. User interference may occur, however, at any time during the parameter optimization if wanted. I.e., the user may extend the scope of adjustable parameters within an ongoing optimization process by simple declarations.

The automation logs all important information throughout the process into one output file called *auto_log*. I.e., the current status of the parameterization is written to that file as well as the number of terminated simulation runs over time. Furthermore, detailed information is given upon possible errors that occurred in a simulation step as well as after the successful termination of the entire parameterization process.

Visually, the optimization process is logged in the form of plots of the simulation results together with the experimental target data. These plots cover the full range of simulated temperatures for saturated liquid density, vapor pressure and enthalpy of vaporization. The plots are generated during the optimization process in MATLAB format and are converted automatically into encapsulated postscript format. In addition to the simulation data, each plot contains the expected quality of the optimized parameter set.

Appendix C - Simulation details

In this work, the Grand Equilibrium method [26] was used for VLE calculations. To determine the chemical potential in the liquid, gradual insertion [35, 36] was used for temperatures $T < 360$ K, while for higher temperatures, Widom’s test molecule method [37] was applied. For gradual insertion, MC simulations in the NpT ensemble were performed using 500 molecules. Starting from a face-centered cubic lattice, 3,000 MC cycles were sampled for equilibration with the first 1,000 time steps in the canonical (NVT) ensemble and 20,000 for production, each cycle containing 500 displacement moves, 500 rotation moves and 1 volume move. Every 100 cycles, 15,000 fluctuating state change moves, 15,000 fluctuating particle translation/rotation moves and 75,000 biased particle translation/rotation moves were performed to determine the chemical potential. For Widom’s test molecule method, MD simulations were performed. Again starting from a face-centered cubic lattice, 25,000 time steps were sampled for equilibration with the first 5,000 time steps in the canonical (NVT) ensemble. The production run was performed for 200,000 steps. The time step was set to 1.2 fs, the integrator used in this study was the Gear-predictor corrector. The chemical potential using Widom’s test molecule method was determined by inserting 2,000 virtual molecules into the simulation volume and averaging over all results. For the corresponding vapor, MC simulations in the pseudo- μVT ensemble were carried out. The simulation volume was adjusted to lead to an average number of 500 molecules in the vapor phase. After 1,000 initial NVT MC cycles, starting from a face centered cubic lattice, 5,000 equilibration cycles in the pseudo- μVT ensemble were performed. The length of the production run was 40,000 cycles. One cycle is defined here to be a number of attempts to displace and rotate molecules equal to the actual number of molecules plus two insertion and two deletion attempts.

Thermodynamic properties were determined in the production phase of the simulation on the

fly. The statistical uncertainties of all results were estimated by block averaging according to Flyvbjerg and Petersen [38] and the error propagation law.

References

- [1] S. Gupta and J.D. Olson, *Industrial Needs in Physical Properties*, Industrial & Engineering Chemistry Research 42 (2003), pp. 6359–6374.
- [2] S. Deublein, B. Eckl, J. Stoll, S.V. Lishchuk, G. Guevara-Carrion, C.W. Glass, T. Merker, M. Bernreuther, H. Hasse, and J. Vrabec, *ms2: A Molecular Simulation Tool for Thermodynamic Properties*, Computer Physics Communications 182 (2011), pp. 2350–2367.
- [3] A. Poncela, A.M. Rubio, and J.J. Freire, *Determination of the potential parameters of a site model from calculations of second virial coefficients of linear and branched alkanes*, Molecular Physics 91 (1997), pp. 189–201.
- [4] T. Kristof, J. Vorholz, J. Liszi, B. Rumpf, and G. Maurer, *A simple effective pair potential for the molecular simulation of the thermodynamic properties of ammonia*, Molecular Physics 97 (1999), pp. 1129–1137.
- [5] M.H. Ketko, J. Rafferty, J.I. Siepmann, and J.J. Potoff, *Development of the TraPPE-UA force field for ethylene oxide*, Fluid Phase Equilibria 274 (2008), pp. 44–49.
- [6] E. Bourasseau, M. Haboudou, A. Boutin, A.H. Fuchs, and P. Ungerer, *New optimization method for intermolecular potentials: Optimization of a new anisotropic united atoms potential for olefins: Prediction of equilibrium properties*, The Journal of Chemical Physics 118 (2003), pp. 3020–3034.
- [7] B. Eckl, J. Vrabec, and H. Hasse, *Molecular modelling and simulation for the process design*, Chemie Ingenieur Technik 80 (2008), pp. 25–33.
- [8] B. Eckl, J. Vrabec, and H. Hasse, *An optimised molecular model for ammonia*, Molecular Physics 106 (2008), pp. 1039–1046.
- [9] 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 Equilibria 274 (2008), pp. 16–26.
- [10] Y.L. Huang, J. Vrabec, and H. Hasse, *Prediction of ternary vapor-liquid equilibria for 33 systems by molecular simulation*, Fluid Phase Equilibria 287 (2009), pp. 62–69.
- [11] B. Eckl, M. Horsch, J. Vrabec, and H. Hasse, *Molecular Modeling and Simulation of Thermophysical Properties: Application to Pure Substances and Mixtures*, High Performance Computing in Science and Engineering '08, Springer, Berlin (2009), pp. 119–133.
- [12] T. Merker, G. Guevara-Carrion, J. Vrabec, and H. Hasse, *Molecular modeling of hydrogen bonding fluids: New cyclohexanol model and transport properties of short monohydric alcohols*, High Performance Computing In Science And Engineering '08, Springer, Berlin (2009), pp. 529–541.
- [13] J. Vrabec, Y.L. Huang, and H. Hasse, *Molecular models for 267 binary mixtures validated by vapor-liquid equilibria: a systematic approach*, Fluid Phase Equilibria 279 (2009), pp. 120–135.
- [14] T. Merker, J. Vrabec, and H. Hasse, *Engineering Molecular Models: Efficient Parametrization Procedure and Cyclohexanol as Case Study*, Soft Materials 10 (2012), pp. 3–24.
- [15] J.R. Errington and A.Z. Panagiotopoulos, *Phase equilibria of the modified Buckingham exponential-6 potential from Hamiltonian scaling grand canonical Monte Carlo*, The Journal of Chemical Physics 109 (1998), pp. 1093–1100.
- [16] R. Faller, H. Schmitz, O. Biermann, and F. Muller-Plathe, *Automatic parameterization of force fields for liquids by simplex optimization*, Journal of Computational Chemistry 20 (1999), pp. 1009–1017.
- [17] P. Ungerer, C. Beauvais, J. Delhommelle, A. Boutin, B. Rousseau, and A.H. Fuchs, *Optimization of the anisotropic united atoms intermolecular potential for n-alkanes*, Journal of Chemical Physics 112 (2000), pp. 5499–5510.
- [18] B. Eckl, J. Vrabec, and H. Hasse, *Set of Molecular Models Based on Quantum Mechanical Ab Initio Calculations and Thermodynamic Data*, Journal of Physical Chemistry B 112 (2008), pp. 12710–12721.
- [19] 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*, Journal of Computational Chemistry 14 (1993), pp. 1347–1363.
- [20] A. Klamt, *Conductor-like Screening Model for Real Solvents - A New Approach to the Quantitative Calculation of Solvation Phenomena*, Journal of Physical Chemistry 99 (1995), pp. 2224–2235.
- [21] A.J. Stone, *Intermolecular Potentials*, Science 321 (2008), pp. 787–789.
- [22] M. Hulsmann, J. Vrabec, A. Maass, and D. Reith, *Assessment of numerical optimization algorithms for the development of molecular models*, Computer Physics Communications 181 (2010), pp. 887–905.
- [23] J. Stoll, *Molecular Models for the Prediction of Thermophysical Properties of Pure Fluids and Mixtures*, University of Stuttgart, 2005.
- [24] R. Rowley, W. Wilding, J. Oscarson, Y. Yang, N. Zundel, T. Daubert, and R. Danner Design Institute for Physical Properties, AIChE, 2003.
- [25] A.Z. Panagiotopoulos, *Direct Determination of Phase Coexistence Properties of Fluids by Monte-Carlo Simulation in a New Ensemble*, Molecular Physics 61 (1987), pp. 813–826.
- [26] J. Vrabec and H. Hasse, *Grand Equilibrium: vapour-liquid equilibria by a new molecular simulation method*, Molecular Physics 100 (2002), pp. 3375–3383.
- [27] A. Lotfi, J. Vrabec, and J. Fischer, *Vapor liquid equilibria of the Lennard-jones Fluid From the Npt Plus Test Particle Method*, Molecular Physics 76 (1992), pp. 1319–1333.
- [28] H. Lorentz, *Über die Anwendung des Satzes vom Virial in der kinetischen Theorie der Gases*, Annalen der Physik 248 (1881), pp. 127–136.
- [29] D. Berthelot, *Sur le melange des gaz*, Comptes Rendues de l'Academie des Sciences 126 (1898), pp. 1703–1706.
- [30] G. Christou, C.L. Young, and P. Svejda, *Gas-liquid critical-temperatures of mixtures of propane, butane, pentane, sulfur-hexafluoride, dichlorodifluoromethane and chlorotrifluoromethane with less volatile compounds of a range of varying polarities*, Fluid Phase Equilibria 67 (1991), pp. 45–53.
- [31] K.N. Marsh, C.L. Young, D.W. Morton, D. Ambrose, and C. Tsonopoulos, *Vapor-liquid Critical Properties of Elements and Compounds. 9. Organic Compounds Containing Nitrogen*, Journal of Chemical and Engineering Data 51 (2006), pp. 305–314.
- [32] M.B. Ewing and J.C.S. Ochoa, *Vapor Pressures of Acetonitrile Determined by Comparative Ebulliometry*, Journal of Chemical and Engineering Data 49 (2004), pp. 486–491.
- [33] A.Z. Panagiotopoulos, N. Quirke, M. Stapleton, and D.J. Tildesley, *Phase equilibria by simulation in the Gibbs ensemble - Alternative derivation, generalization, and application to mixture and membrane equilibria*, Molecular Physics 63 (1988), pp. 527–545.
- [34] Towhee, <http://www.towhee.sourceforge.org/>; (2008), .

- [35] A.P. Lyubartsev, A.A. Martinovski, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov, *New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles*, The Journal of Chemical Physics 96 (1992), pp. 1776–1783.
- [36] I. Nezbeda and J. Kolafa, *A New Version of the Insertion Particle Method for Determining the Chemical Potential by Monte Carlo Simulation*, Molecular Simulation 5 (1991), pp. 391–403.
- [37] B. Widom, *Some Topics In Theory of Fluids*, The Journal of Chemical Physics 39 (1963), pp. 2808–2812.
- [38] H. Flyvbjerg and H.G. Petersen, *Error estimates on averages of correlated data*, The Journal of Chemical Physics 91 (1989), pp. 461–466.

Table 1. Molecular model parameters for acetonitrile. The dipole moment magnitude was set to $\mu = 4.04$ D, its negative end was oriented towards the nitrogen atom.

		Unit	CH ₃	C	N	Dipole
	x	Å	0	0	0	0
	y	Å	0	0	0	0
	z	Å	-1.29	0.05	1.15	0.46
	σ	Å	3.88	2.81	3.27	-
$m_{\text{C2H3N}}^{\langle 6 \rangle}$	ε/k_{B}	K	184.31	10.64	43.19	-
	σ	Å	3.82	2.81	3.37	-
$m_{\text{C2H3N}}^{\langle 8 \rangle}$	ε/k_{B}	K	180.58	10.64	58.58	-

Table 2. Contributions to the cost function (8) at various temperatures for the individual models $m_{\text{C}_2\text{H}_3\text{N}}^{(s)}$. The last row contains the total cost function over the entire property and temperature range from $0.6 < T/T_c < 0.99$.

Quantity	T / K	$m_{\text{C}_2\text{H}_3\text{N}}^{(s)}$							
		$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$	$s = 8$
$\left(\frac{\rho'_{\text{Exp}} - \rho'_{\text{Sim}}}{\rho'_{\text{Exp}}}\right)^2$	270	0.3	1.6	0.1	0.3	0.0	0.0	1.1	0.0
	360	0.9	2.5	0.2	1.1	0.1	0.0	1.3	0.0
	420	3.6	5.6	0.1	1.3	0.1	0.0	2.0	0.0
	490	∞	16.6	0.2	3.1	0.5	0.0	3.5	0.0
	518	∞	∞	0.6	0.7	2.3	0.2	6.3	0.0
$10^3 \left(\frac{p_{\text{Sim}} - p_{\text{Exp}}}{4p_{\text{Exp}}}\right)^2$	270	1.9	1.8	10.3	6.2	5.6	10.2	16.9	4.9
	360	12.0	5.9	0.5	0.1	0.5	0.2	3.4	0.5
	420	12.7	7.1	0.2	0.2	0.0	0.0	1.5	0.3
	490	∞	4.5	0.1	0.1	0.3	0.0	0.7	0.0
	518	∞	∞	0.0	0.0	0.2	0.1	0.2	0.0
$10^5 \left(\frac{\Delta h_{v,\text{Sim}} - \Delta h_{v,\text{Exp}}}{14\Delta h_{v,\text{Exp}}}\right)^2$	270	0.3	0.7	7.8	9.8	3.9	4.7	8.1	4.3
	360	0.7	0.4	1.3	0.2	0.6	1.3	4.8	2.0
	420	8.6	6.3	0.1	0.3	0.0	0.0	3.1	0.8
	490	∞	23.6	1.0	2.2	2.8	0.1	3.2	0.0
	518	∞	∞	1.5	1.0	6.1	0.8	5.1	0.0
10 * Total		-	-	0.7	1.6	1.4	0.3	5.1	0.3

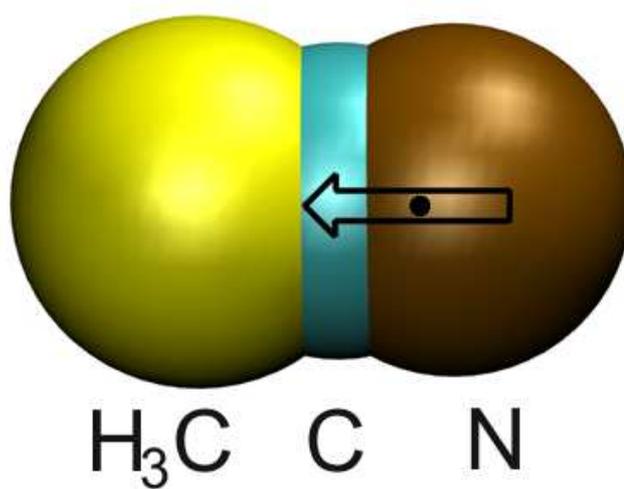


Figure 1. Structure of the molecular acetonitrile model. The arrow indicates the point dipole, located at the bullet.

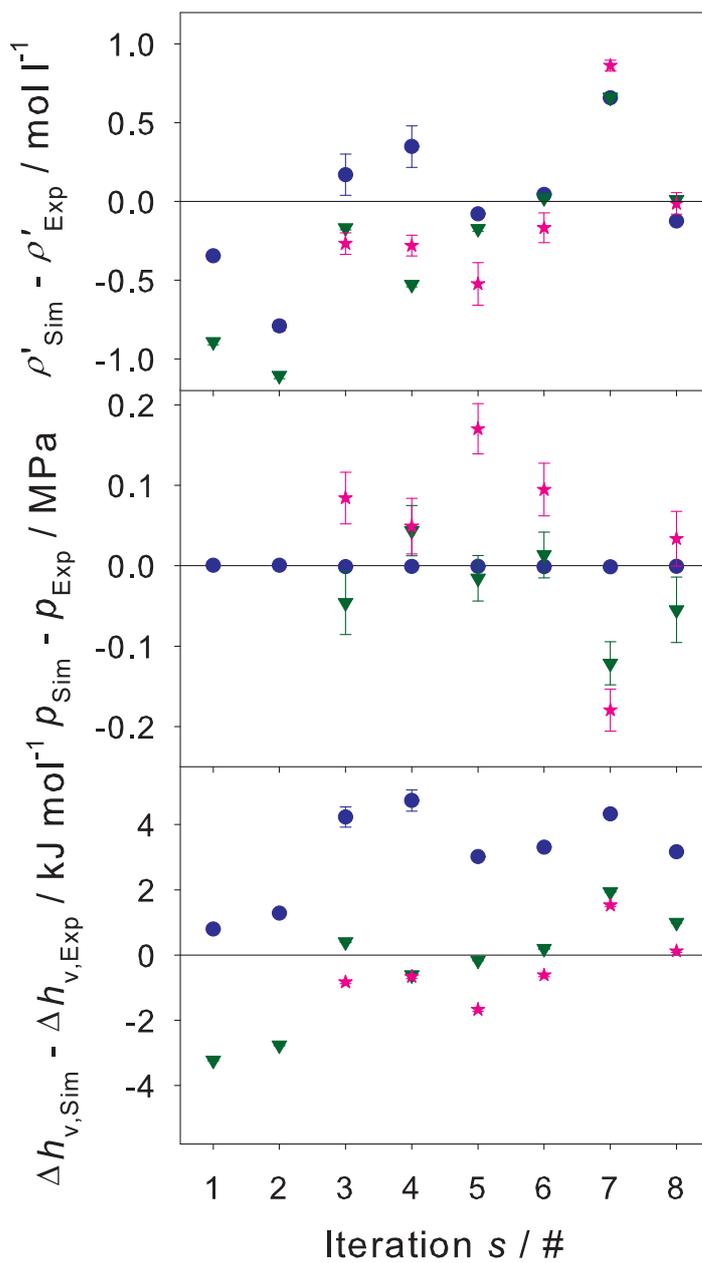


Figure 2. Absolute deviations between simulation results and experimental data [24] as a function of iterations for various temperatures: (●) 270 K, (▼) 420 K and (★) 518 K. Top: saturated liquid density ρ^l , center: vapor pressure p , bottom: enthalpy of vaporization Δh_v . Error bars indicate the statistical uncertainty of the simulation data, if they exceed symbol size. Data points that are out of scale are not shown.

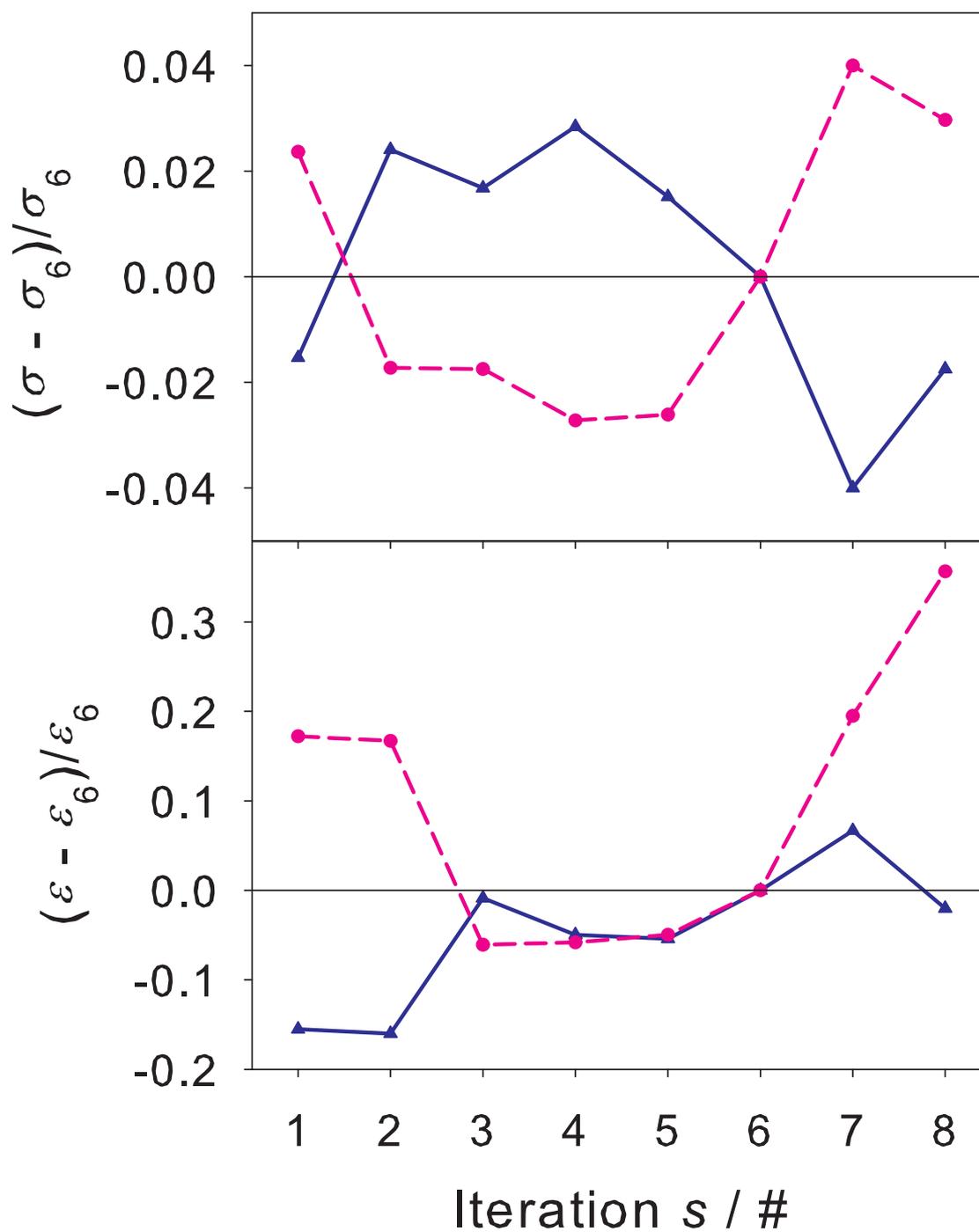


Figure 3. Relative deviations between the LJ size parameter σ and the LJ energy parameter ε from the values of model $m_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ over iteration steps: (▲) CH₃ site and (●) nitrogen site. The lines are guides for the eye.

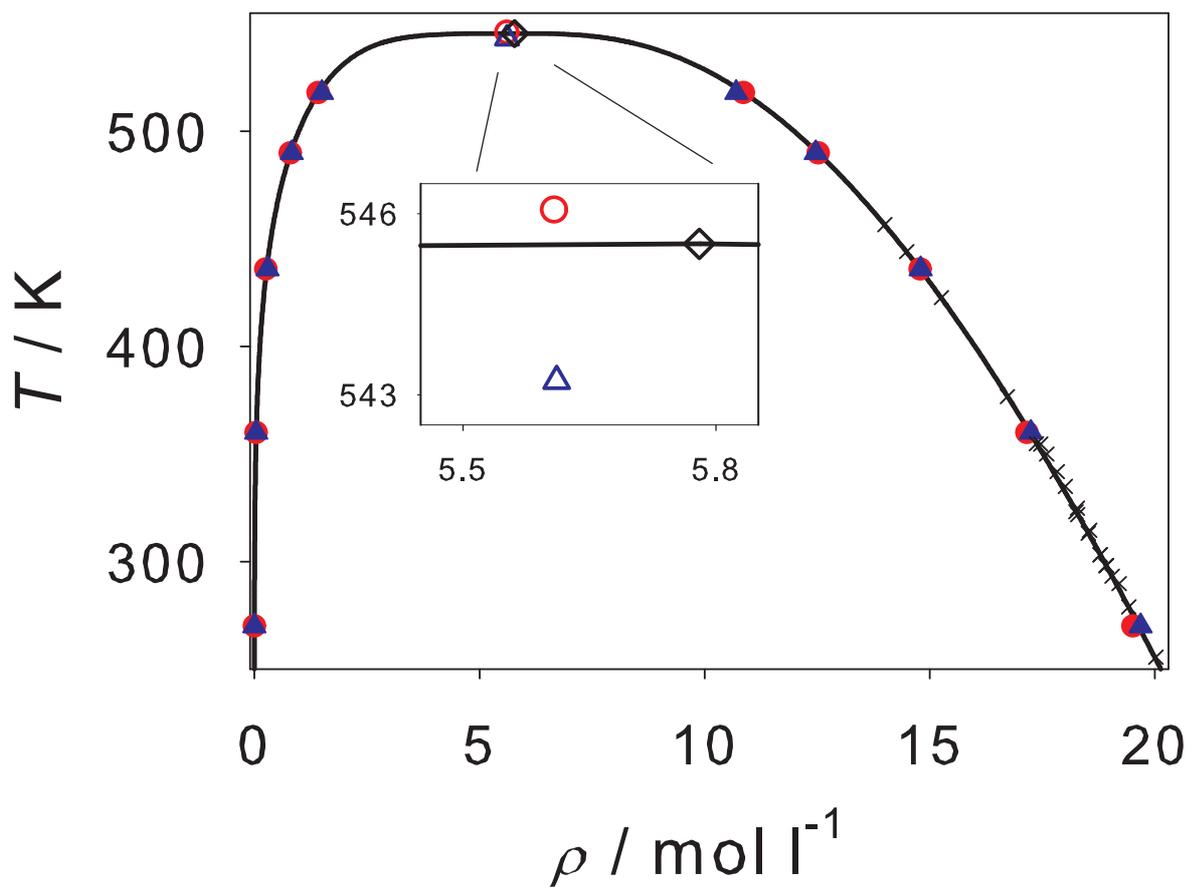


Figure 4. Saturated densities of acetonitrile for various temperatures. Simulation data for (\blacktriangle) model $m_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ and (\bullet) model $m_{\text{C}_2\text{H}_3\text{N}}^{(8)}$ are compared to (\times) experimental data and (—) the DIPPR correlation [24]. The critical point is denoted by empty symbols, the experimental value is denoted by \diamond . The statistical uncertainties of the simulation data are within symbol size.

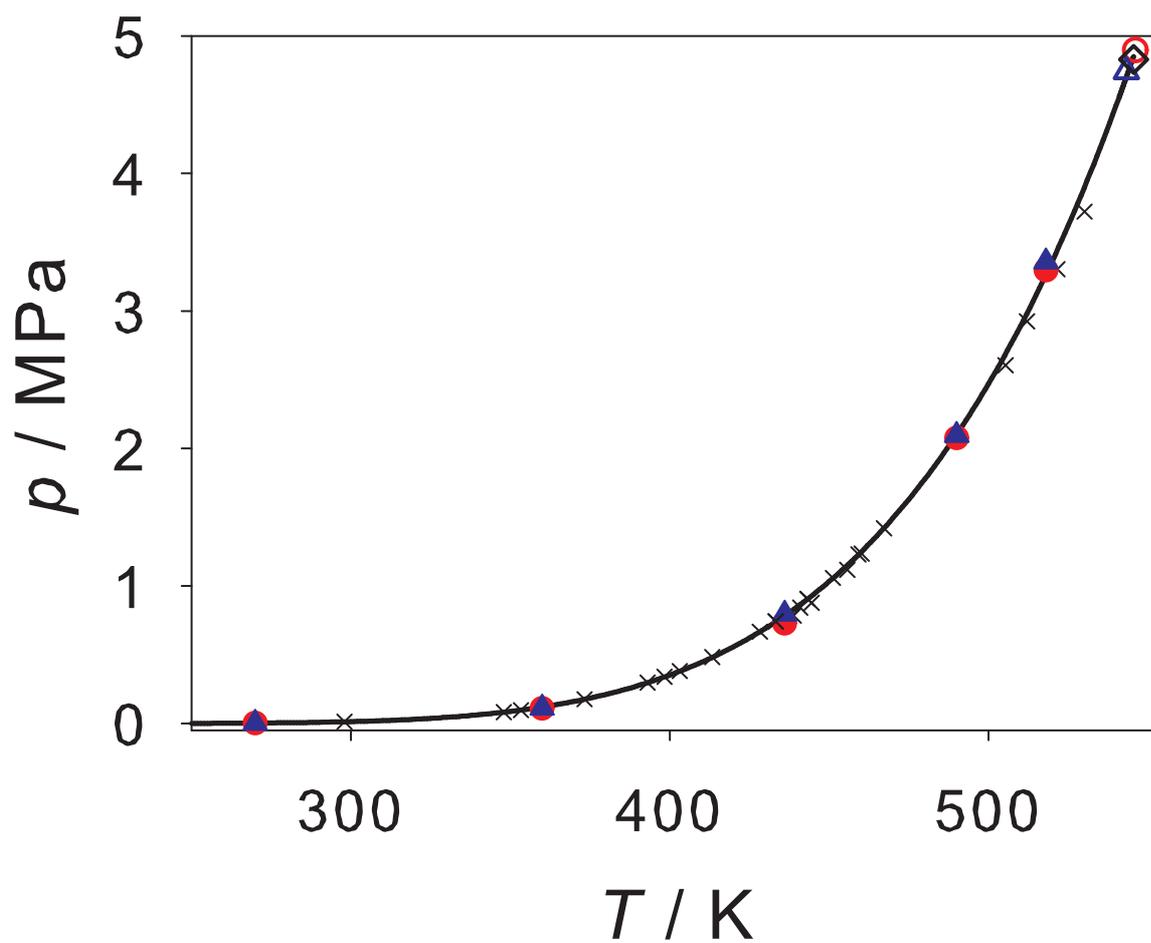


Figure 5. Vapor pressure of acetonitrile for various temperatures. Simulation data for (\blacktriangle) model $m_{C_2H_3N}^{(6)}$ and (\bullet) model $m_{C_2H_3N}^{(8)}$ are compared to (\times) experimental data and (—) the DIPPR correlation [24]. The critical point is denoted by empty symbols, the experimental value is denoted by \diamond . The statistical uncertainties of the simulation data are within symbol size.

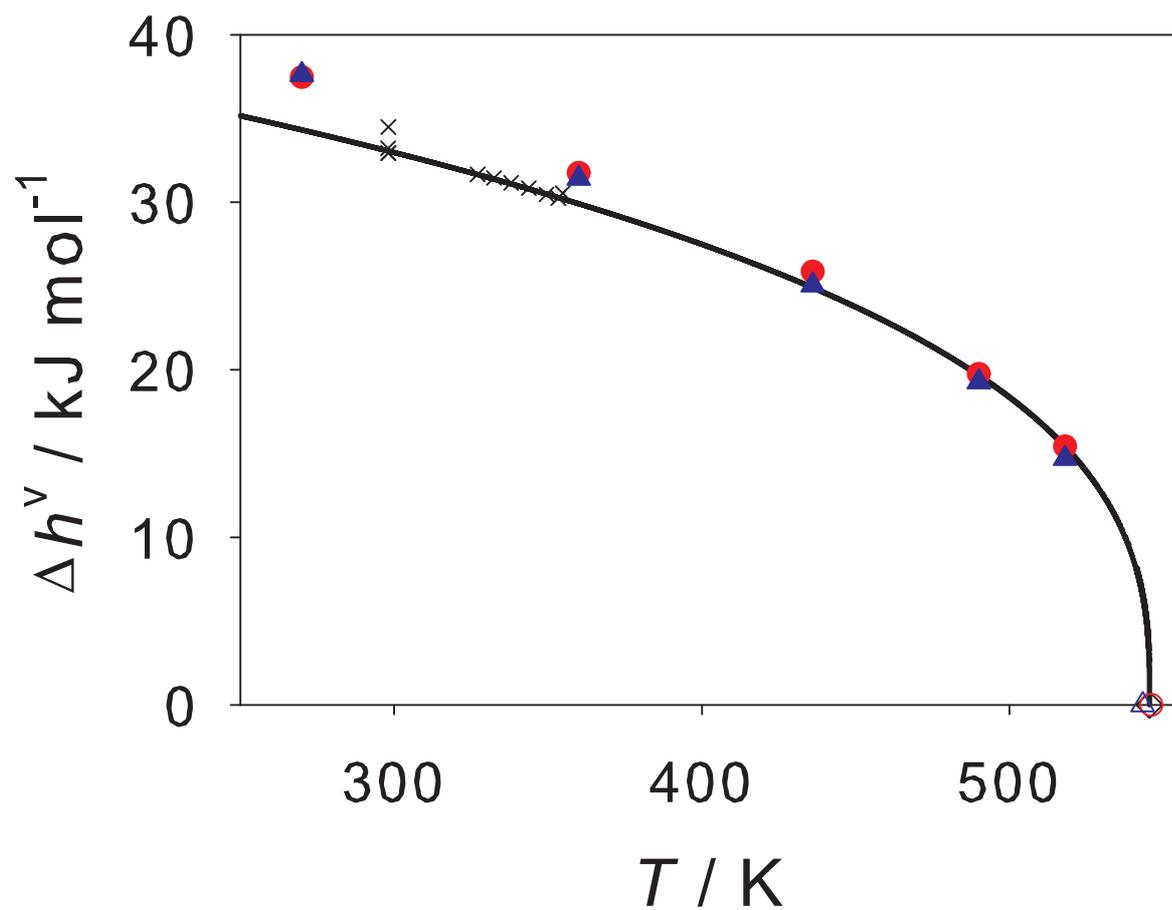


Figure 6. Enthalpy of vaporization of acetonitrile for various temperatures. Simulation data for (\blacktriangle) model $m_{\text{C}_2\text{H}_3\text{N}}^{(6)}$ and (\bullet) model $m_{\text{C}_2\text{H}_3\text{N}}^{(8)}$ are compared to (\times) experimental data and (—) the DIPPR correlation [24]. The critical point is denoted by empty symbols, the experimental value is denoted by \diamond . The statistical uncertainties of the simulation data are within symbol size.

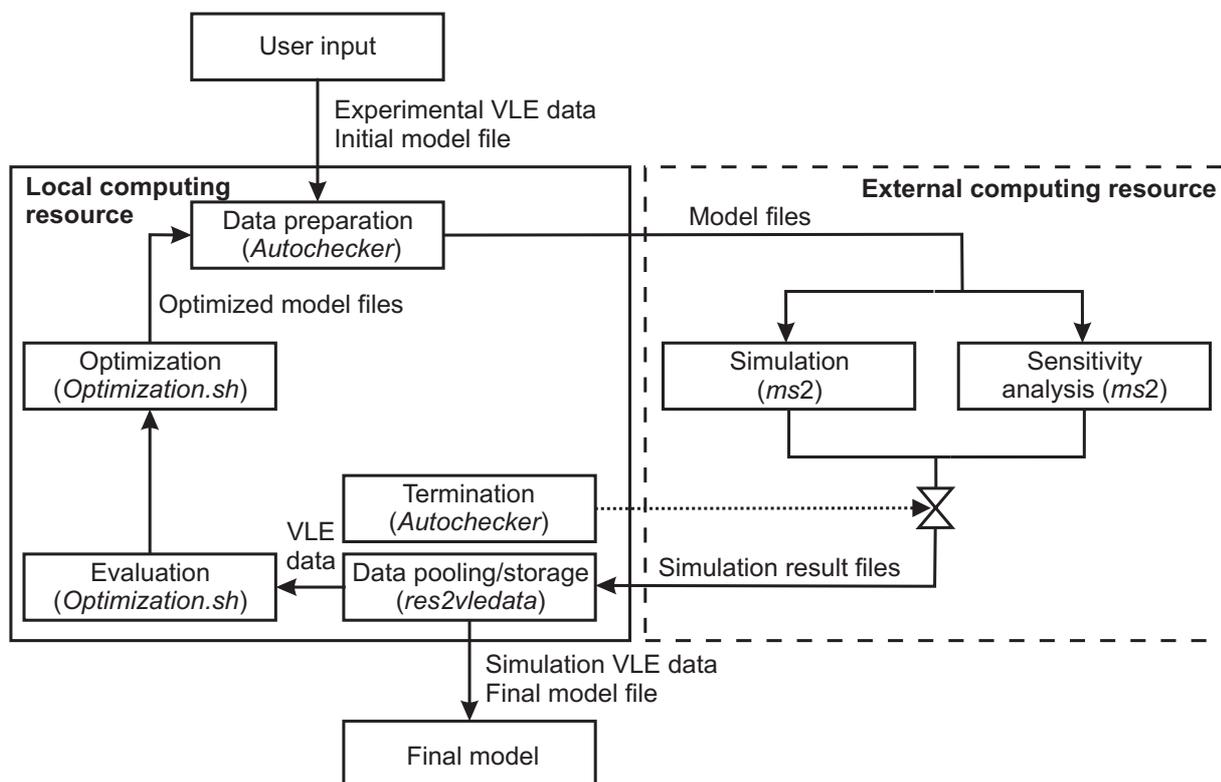


Figure 7. Schematic of the individual steps required for the optimization of a molecular model. The tasks in the boxes are performed by the automation, the employed programs are denoted in italics. The text along the arrows indicates data that are transferred between the steps.