Molecular modeling and simulation: model development, thermodynamic properties, scaling behavior and data management

Matthias Heinen, René S. Chatwell, Simon Homes, Gabriela Guevara-Carrion, Robin Fingerhut, Maximilian Kohns, Simon Stephan, Martin T. Horsch and Jadran Vrabec

Abstract We are outlining our most recent findings, covering: 1) A comparison of a micro- and macroscopic solution of a two-phase Riemann problem obtained from molecular dynamics simulations and finite volume schemes; 2) A novel equation of state for the bulk viscosity of liquid noble gases based on a multi-mode relaxation ansatz; 3) A detailed analysis of the evaporation process of simple fluids; 4) Diffusion coefficients of quaternary liquid mixtures obtained with the Green-Kubo formalism; 5) An analysis of the solid/fluid phase transition for the face centered cubic (fcc) lattice; 6) The relative permittivity of mixtures of water and acetone; 7) An assessment of the reliability and reproducibility of molecular simulation results; 8) Techniques for the data management in simulation workflows, including annotations of simulation outcomes with appropriate metadata standardized by an ontology.

Matthias Heinen, René S. Chatwell, Simon Homes, Gabriela Guevara-Carrion, Robin Fingerhut, Jadran Vrabec
Lehrstuhl für Thermodynamik und Thermische Verfahrenstechnik,
Technische Universität Berlin, Ernst-Reuter Platz 1, 10587 Berlin, Germany
e-mail: vrabec@tu-berlin.de

Maximilian Kohns, Simon Stephan
Lehrstuhl für Thermodynamik,
Technische Universität Kaiserslautern, Erwin-Schrödinger-Straße 44, 67663 Kaiserslautern, Germany

Martin T. Horsch
STFC Daresbury Laboratory,
UK Research and Innovation, Keckwick Ln, Daresbury WA4 4AD, UK

1 Two-phase shock tube scenario

Large scale molecular dynamics (MD) simulations of a two-phase shock tube scenario were conducted. These simulations were intended to serve as a benchmark for macroscopic solutions obtained from computational fluid dynamics (CFD) simulations employing finite volume (FV) schemes. Two macroscopic approaches were considered: the homogeneous equilibrium method (HEM) and the sharp interface method. Both are implemented in the discontinuous Galerkin spectral element method (DGSEM) framework *FLEXI* [24].

In contrast to the scenario in Ref. [25] that considered two supercritical states, known as the classical Riemann problem, the present scenario consisted of a liquid and a vapor phase, connected through a planar interface. The thermodynamic states of the liquid and vapor phases were specified such that they were out of equilibrium and hence phase transition occurred. This scenario is known as two-phase Riemann problem. While the classical Riemann problem is fully understood, the two-phase Riemann problem has implications for the system of equations that cannot be solved in a straightforward manner.

As in Ref. [25], the Lennard-Jones Truncated and Shifted (LJTS) fluid was assumed for two reasons: it is computationally cheap in MD simulations and an accurate equation of state (EOS) [21] is available for the macroscopic solution.

The initial configurations of the MD simulations were prepared in two steps, cf. Fig. 1. First, a vapor-liquid equilibrium (VLE) was maintained at a temperature of $T = 0.9$. The liquid phase was extracted and brought into contact with a vapor phase at a lower temperature $T = 0.8$ and a lower density in a symmetric setup. Three cases were considered with varying the density of the vapor phase ρ_v, i.e. specifying 50%, 70% and 90% of the saturated density at a temperature of $T = 0.8$. The system dimensions, specified in particle diameters σ, were defined as follows: The extent of the vapor phase $L_v = 1500$ was chosen to be wide enough so that the shock wave exerted from the liquid phase, which is a consequence of the global non-equilibrium, could be observed for a sufficiently long time period before it reached the periodic boundary. The specified width of the liquid phase $L_l = 200$ was a compromise between being sufficiently wide such that the opposite vapor-liquid interfaces do not interfere with each other, because of rapid state changes due to evaporation, and being small enough to keep the computational cost on an acceptable level. To check whether the specified width of L_l is appropriate, an exemplary simulation was repeated with a doubled width $L_l = 400$. From that, almost identical results were obtained.

The cross-sectional area had an extent of $10^6\sigma^2$ so that the sampling of the rapidly changing profiles, i.e. temperature, density and hydrodynamic velocity, employing a classical binning scheme with a high spatial and temporal resolution yielded an excellent statistical quality. Because of the large number of up to $N \approx 3 \cdot 10^8$ particles, all simulations were carried out with the massively parallel code *ls1 mardyn* [42]. This code is being continuously improved and was recently optimized with respect to its node-level performance and parallel efficiency [56].

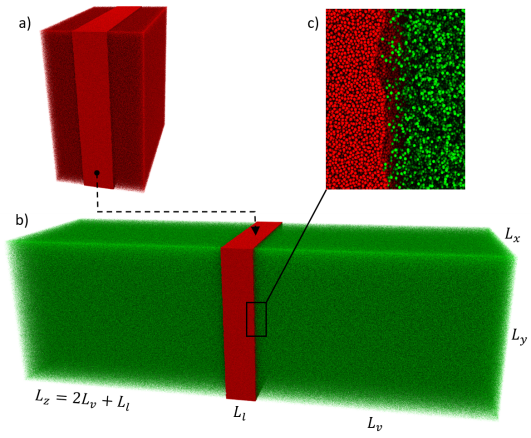


Fig. 1: Snapshots of the prepared molecular systems, rendered with the cross-platform visualization prototyping framework MegaMol [18] and the Intel OSPRAY plugin. a) Final configuration of the vapor-liquid equilibrium simulation from which the liquid phase was extracted to build the test case scenarios. b) One of the test case scenarios with a vapor phase (green) diluted to 70% of the saturated vapor density at the temperature $T = 0.8$. c) Close-up look at the interface.

Results for the case with 50% of the saturated vapor density are shown in Fig. 2. While the HEM approach is not able to reproduce the results of the MD simulation, the sharp interface approach showed a very good agreement with the MD data in the homogeneous bulk phases, except for the vicinity of the interface. It reproduced the propagation speed of the shock wave and the characteristic shapes of all profiles. Even the magnitudes of plateaus of the profiles were quantitatively matched.

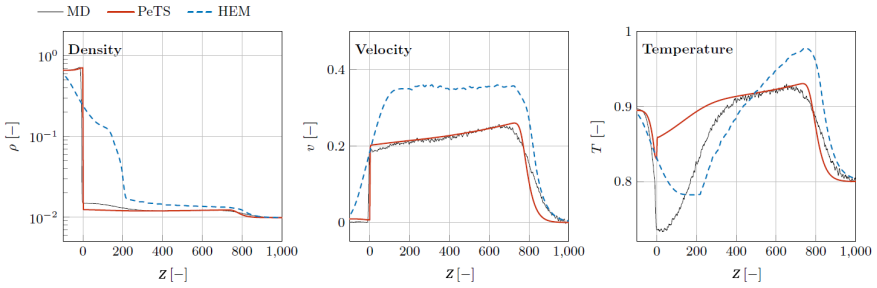


Fig. 2: Results for the case with a vapor density of 50% of its saturation value. The plots show profiles of density, velocity and temperature at $t = 200$ obtained from the sharp interface method, HEM and MD.

2 Bulk viscosity of liquid noble gases

Stokes' hypothesis postulates that any form-invariant changes of a fluid's local volume, i.e. compression or dilatation, are not associated with the dissipation of linear momentum, which is synonymous with a vanishing bulk viscosity $\mu_b = 0$. Despite theoretical and experimental evidence to the contrary, the hypothesis is still widely applied throughout all branches of fluid mechanics. The bulk viscosity of liquid noble gases was studied here on the basis of a multi-mode relaxation ansatz and an equation of state is proposed [9]. The relaxation ansatz is based on a large data set that was generated by dedicated atomistic simulations, resulting in an EOS exposing the bulk viscosity as a two-parametric power function of density, with the parameters being functions of temperature. The noble gases' atomistic description rests on the Lennard-Jones potential.

The Green-Kubo formalism relates the bulk viscosity to time-autocorrelation functions (ACF) that were sampled in the microcanonical (*NVE*) ensemble, utilizing the fully open source program *ms*2 [44]. To reduce finite size effects and to gain better statistics, ensembles containing $N = 4096$ particles were placed in cubic volumes with periodic boundary conditions. To resolve the small-scale pressure fluctuations, a small integrator time step was used and each autocorrelation function was sampled over a substantial time period. Relatively large ensemble sizes in combination with long simulation times are computationally demanding and thus require modern HPC architectures. The bulk viscosity can be determined microscopically by ACF of local small-scale, transient pressure fluctuations that are intrinsic in any fluid under equilibrium. These pressure fluctuations have been observed to relax in different modes. Each mode decays exponentially over time, following a Kohlrausch-Williams-Watt function. For liquid noble gases, three superimposing relaxation modes were found to be present, leading to the relaxation model

$$B_R(t) = C_f \exp\left(-\left(\frac{t}{\delta_f}\right)^{\beta_f}\right) + C_m \exp\left(-\left(\frac{t}{\delta_m}\right)^{\beta_m}\right) + C_s \exp\left(-\left(\frac{t}{\delta_s}\right)^{\beta_s}\right) . \quad (1)$$

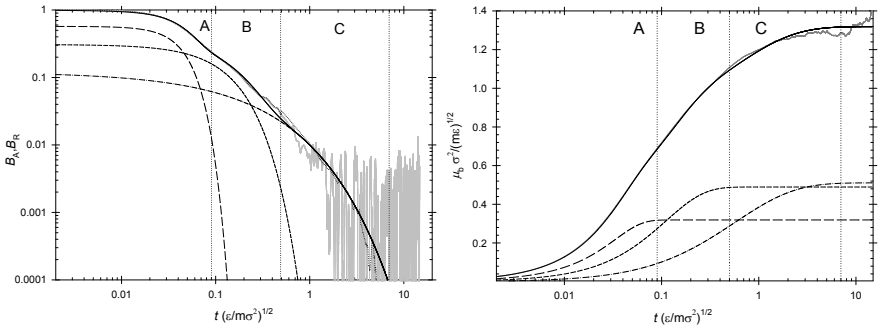
The first term describes the fast, and the subsequent terms the intermediate and slow modes, respectively. The weighting factors are constraint, i.e. $C_f + C_m + C_s = 1$, and the Kohlrausch parameters δ_i, β_i are a measure of relaxation time scale and distortion from the exponential function. The model's eight independent parameters $C_f, C_m, \delta_f, \delta_m, \delta_s, \beta_f, \beta_m, \beta_s$ were determined by fitting the relaxation model B_R to the data sampled by MD. Each mode's average relaxation time τ_i is defined as integral mean value of its respective relaxation model contribution $B_{R,i}$. As originally proposed by Maxwell, the bulk viscosity μ_b is proportional to the cumulative averaged relaxation time

$$\mu_b = K_r \sum_{i=1}^{3} \lim_{t\to\infty} \int_0^t dt \, (B_{R,i}) \ , \quad (2)$$

with the proportionality constant K_r being the fluid's relaxation modulus. The sampled ACF partitions into three segments, with each segment being dominated by a

different mode, cf. Fig. 3a. In contrast to the sampled ACF that is plagued by noise, the employed relaxation model's time integral properly converges to a definite value, thus allowing to determine μ_b unambiguously, cf. Fig. 3b.

Applying the relaxation ansatz to all sampled state points generates a large dataset from which the EOS emerges as a two-parametric power function with both parameters showing a conspicuous saturation behavior over temperature. After passing a temperature threshold, the bulk viscosity is found to vary significantly over density, a behavior that resembles the frequency response of a one pole low-pass filter.



(a) (b)

Fig. 3: (a) Comparison of a sampled ACF with the relaxation model including all three modes. While the gray line constitutes the sampled ACF, the solid black line represents the relaxation model and its fast (dashed), intermediate (short-dashed) and slow (dashed-dotted) modes, respectively. (b) Comparison of the integrated sampled autocorrelation function with the relaxation model. Due to noise contributions to the slow mode of the sampled autocorrelation function, the bulk viscosity is difficult to determine precisely by molecular dynamics simulation (gray line). In contrast, the employed relaxation model (solid black line) converges towards an unambiguous value at finite times. The dashed, short-dashed and dotted-dashed lines represent the fast, intermediate and slow modes, respectively.

3 Evaporation of simple fluids

Evaporation phenomena play a crucial role in process engineering and in many other fields, but they are not yet fully understood. In order to gain further knowledge about these basic phenomena, MD simulations were conducted building upon Ref. [23]. The simple and computationally efficient LJTS potential was applied to describe the interactions between the particles. A parallelepiped was employed as simulation

volume, consisting of one liquid and one vapor phase, respectively. A net evaporation flux was constrained by establishing a vaccuum boundary condition in the vapor. Particles reaching this vaccuum region were deleted. In order to maintain a constant number of particles N in the simulation domain, new particles had to be added. This was achieved by a method proposed in Ref. [22]. Numerous simulations were conducted for a range of temperatures of the bulk liquid as well as varying distances between the bulk liquid and the interface.
The systems under investigation contained between 1.3 and $8.3 \cdot 10^6$ particles. The utilized MD software was *ls1 mardyn* [42], which is designed for massive parallelization so that up to $7.8 \cdot 10^3$ cores could be used for one simulation.
It was found that the evaporation flux setting in led to a decrease of the interface temperature T_i. Since the particle flux j_p depends on T_i, a decreasing temperature induces a drop of j_p itself. Upon progress of simulation, the interface temperature and the particle flux converge towards a limit, cf. Fig. 4. Their value depends on the bulk liquid temperature among others. For all conducted simulations, the respective stationary values can be expressed by the dashed line in Fig. 4.

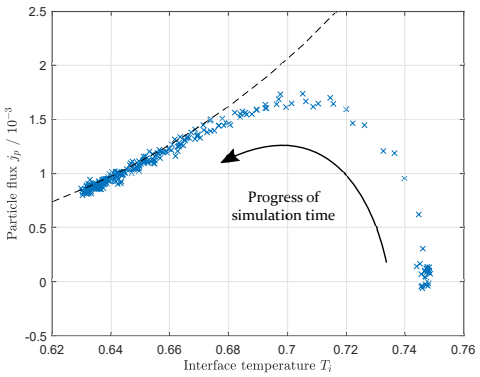


Fig. 4: Particle flux j_p over the interface temperature T_i over the course of one simulation. Throughout the simulation run, T_i declined until stationarity was reached. The particle flux first increased up to a maximum, then decreased, until it finally converged to a stationary value. The dashed line marks the fit through the respective stationary values of all simulations.

4 Diffusion in quaternary liquid mixtures

Most mass transfer processes occurring in nature and in technical applications involve liquid solutions with more than three components. Rate-based methods em-

ployed for modeling, design and control of separation unit operations in chemical engineering, such as distillation, rely on mass and energy transfer models which require reliable information on diffusion coefficient data for the regarded mixtures. Therefore, there is a significant interest in the improvement of experimental methodologies and the development of reliable methods for the prediction of mutual diffusion coefficients of liquid multicomponent mixtures.

Fick's law for diffusion in a quaternary mixture requires nine different diffusion coefficients that depend on temperature, pressure, composition and the regarded frame of reference. The presence of six cross diffusion coefficients makes interpretation and data processing in experimental work a challenging task which often leads to large experimental uncertainties. Despite the continuous improvement and development of experimental techniques during the last decades, the availability of diffusion coefficients of mixtures containing four components is still very poor. Thus, the growing need of accurate diffusion data for basic research and engineering applications cannot be satisfied by experimental measurements alone.

Most predictive equations for multicomponent diffusion of liquids rely on extensions of the Darken relation [3] and are therefore only truly valid for ideal mixtures. The underlying physical phenomena in non-ideal mixtures are still not well understood and the lack of data impedes the development and verification of new predictive equations. In this context, MD offers an alternative path not only to assess multicomponent diffusion coefficients, but also to gain insight into the underlying microscopic behavior.

Recently, the ability of MD to predict the Fick diffusion coefficient matrix of a quaternary liquid mixture has been demonstrated for water + methanol + ethanol + 2-propanol [19]. However, because of the lack of experimental data, only consistency test could be performed for the predicted diffusion data. Here, the Fick diffusion coefficient matrix of the mixture cyclohexane + toluene + acetone + methanol, which was recently studied with Raman spectroscopy [43], was successfully predicted solely with molecular simulation techniques.

In the framework of the generalized form of Fick's law, the molar flux of component i in a mixture of four components is written as a linear combination of concentration gradients ∇c_j [10]

$$J_i = -\sum_{j=1}^{3} D_{ij} \nabla c_j \,, \quad (i = 1,2,3)\,, \tag{3}$$

where D_{ii} are the main diffusion coefficients that relate the molar flux of component i to its own concentration gradient and D_{ij} are the cross diffusion coefficients that relate the molar flux of component i to the concentration gradient of component j. The Fick approach involves three independent diffusion fluxes and a 3×3 diffusion coefficient matrix, which is generally not symmetric, i.e. $D_{ij} \neq D_{ji}$. Further, the numerical values of D_{ij} depend both on the reference frame for velocity (molar-, mass- or volume-averaged) and on the order of the components.

The main shortcoming of Fick's law is the fact that concentration gradients are not the true thermodynamic driving forces for diffusion, which are rather given by

chemical potential gradients. Maxwell-Stefan theory follows this path, assuming that chemical potential gradients $\nabla\mu_i$ are balanced by friction forces between the components that are proportional to their mutual velocity [55]. The Maxwell-Stefan diffusion coefficient $Đ_{ij}$ plays the role of an inverse friction coefficient between components i and j [55] and its matrix is symmetric so that it has only six independent elements. Maxwell-Stefan diffusion coefficients are associated with chemical potential gradients and thus cannot directly be measured in the laboratory. However, they are accessible with equilibrium MD techniques, i.e. the Green-Kubo formalism or the Einstein approach.

This work employs the Green-Kubo formalism based on the net velocity autocorrelation function to obtain $n \times n$ phenomenological coefficients [37]

$$L_{ij} = \frac{1}{3N} \int_0^\infty \mathrm{d}t \left\langle \sum_{k=1}^{N_i} \mathbf{v}_{i,k}(0) \cdot \sum_{l=1}^{N_j} \mathbf{v}_{j,l}(t) \right\rangle , \tag{4}$$

in a mixture of n components. Here, N is the total number of molecules, N_i is the number of molecules of component i and $\mathbf{v}_{i,k}(t)$ denotes the center of mass velocity vector of the k-th molecule of component i at time t.

Starting from the phenomenological coefficients L_{ij}, the elements of a $(n-1) \times (n-1)$ matrix $\boldsymbol{\Delta}$ can be defined as [37]

$$\boldsymbol{\Delta}_{ij} = (1-x_i) \left(\frac{L_{ij}}{x_j} - \frac{L_{in}}{x_n} \right) - x_i \sum_{k=1 \neq i}^{n} \left(\frac{L_{kj}}{x_j} - \frac{L_{kn}}{x_n} \right) , \tag{5}$$

where x_i is the molar fraction of component i. Its inverse matrix $\mathbf{B} = \boldsymbol{\Delta}^{-1}$ is related to the Maxwell-Stefan diffusion coefficients $Đ_{ij}$.

On the other hand, experimental methods yield the Fick diffusion coefficients. Thus, to compare the predictions by molecular simulation with experimental values, a relation between Fick and Maxwell-Stefan diffusion coefficients is required [55]

$$\mathbf{D} = \boldsymbol{\Delta} \cdot \boldsymbol{\Gamma} , \tag{6}$$

in which all three symbols represent 3×3 matrices and the elements of $\boldsymbol{\Delta}$ are given in Eq. (5). $\boldsymbol{\Gamma}$ is the thermodynamic factor matrix

$$\Gamma_{ij} = \delta_{ij} + x_i \frac{\partial \ln \gamma_i}{\partial x_j} \bigg|_{T,p,x_{k,k \neq j=1\ldots3}} . \tag{7}$$

Therein, δ_{ij} is the Kronecker delta function and γ_i the activity coefficient of component i. Here, the thermodynamic factor matrix was estimated from information on the microscopic structure given by radial distribution functions $g_{ij}(r)$ based on Kirkwood-Buff theory. In the grand canonical (μVT) ensemble Kirkwood-Buff integrals G_{ij} are defined by [33]

$$G_{ij} = 4\pi \int_0^\infty \left(g_{ij}(r) - 1 \right) r^2 dr. \tag{8}$$

Because the canonical (NVT) ensemble was employed, possible convergence issues [41] were corrected with the method by Krüger et al. [38]. Moreover, corrections of the radial distribution functions are required. Therefore, Kirkwood-Buff integrals were calculated based on the methodology proposed by Ganguly and van der Vegt [16]. Extrapolation to the thermodynamic limit was not necessary because of the rather large ensemble size N = 8000.

Predictive equilibrium MD simulations of diffusion coefficients and the thermodynamic factor of the quaternary mixture cyclohexane (1) + toluene (2) + acetone (3) + methanol (4) were carried out at 298.15 K and 0.1 MPa for one composition that was studied experimentally [43], i.e. $x_1 = x_2 = x_3 = 0.05$ mol mol^{-1}. A cubic simulation volume of containing 8000 molecules with a cut-off radius of 24.5 Å was employed for this purpose. The resulting phenomenological coefficients L_{ij} were averaged from more than 10^5 correlation functions with a length of 20 ps. These coefficients were employed together with the thermodynamic factor matrix to calculate the Fick diffusion coefficient matrix in the molar frame of reference. In order to compare simulation results with the experimental data, the Fick diffusion coefficient matrix was transformed into the volume-averaged frame [55].

A comparison between present simulation results and experimental data is given in Fig. 5. The simulation results for all elements of the diffusion matrix agree with the experimental data within the reported uncertainties.

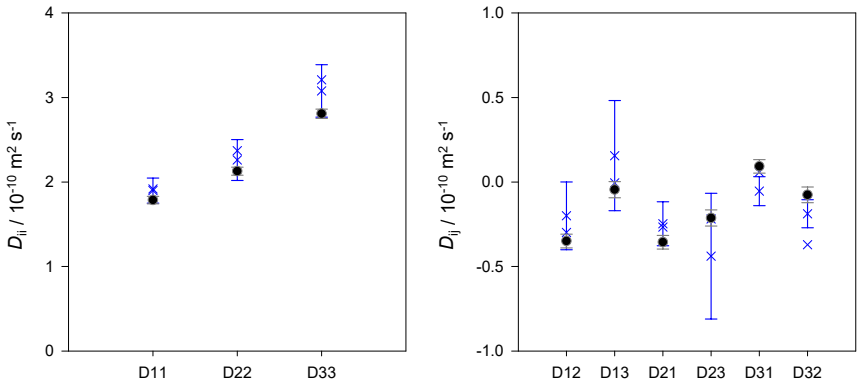
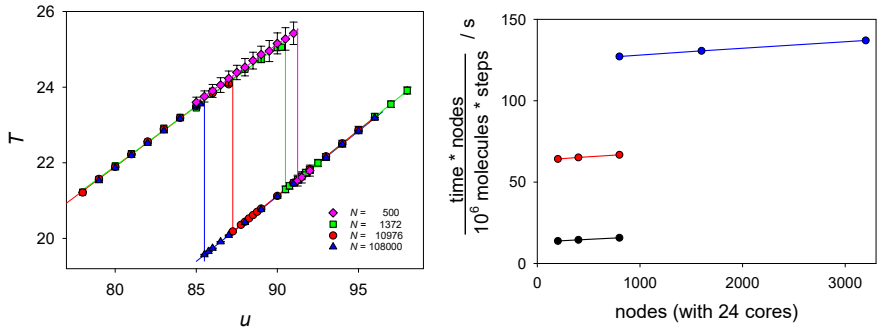


Fig. 5: Main (left) and cross (right) elements of the Fick diffusion coefficient matrix of the mixture cyclohexane (1) + toluene (2) + acetone (3) + methanol (4) at 298.15 K and 0.1 MPa. Present simulation results (black bullets) are compared with the experimental data [43] estimated from three and twelve experiments (blue crosses).

5 Solid/fluid phase transition and strong scaling of *ms*2

In a recent study [40], the solid/fluid (S/F) phase transition was analyzed for the face centered cubic (fcc) lattice utilizing *ms*2 [44]. The LJ potential was applied in MD simulations such that the solid was heated at constant volume up to its phase transition. The Z method [5] was applied to determine the limit of superheating (LS) and the melting point (MP). For this purpose, total energy u (potential + kinetic energy) and temperature T were evaluated, cf. Fig. 6a. First, the fcc lattice remains for a specific total energy range in the metastable solid state, which is limited by u_{LS} when the solid melts. Beyond this range (u reaches values slightly above u_{LS}), the temperature drops to its melting temperature T_m since kinetic energy supplies the internal energy of fusion. At the S/F transition, the total energy $u^{solid}(v,T_{LS}) = u^{fluid}(v,T_m)$ holds with the maximum T_{LS} and minimum temperature T_m when constant volume is maintained [40].



(a) (b)

Fig. 6: (a) Z method sampled by MD simulations using *ms*2; pink: $N = 500$; green: $N = 1372$; red: $N = 10976$; blue: $N = 108000$ atoms; vertical lines indicate the temperature drop from the LS to the MP. (b) Strong scaling efficiency of MD simulations with *ms*2 for a fcc LJ solid measured on CRAY XC40 (*Hazel Hen*) with hybrid MPI + OpenMP parallelization (each MPI process had two OpenMP threads); black: $N = 64,000$ with a cutoff of $r_c = 6\sigma$; red: $N = 64,000$ with $r_c = 29\sigma$; blue: $N = 120,000$ with $r_c = 36\sigma$.

The melting process is strongly dependent on the system's structure and dynamics, particularly when a perfect fcc lattice without defects is considered. Thus, finite size effects were hypothesized and supported by a recent study [40]. Fig. 6a clearly shows a substantial system size dependence of T_{LS} and T_m. Moreover, this behavior reinforces how well molecular simulations are able to tackle physical phenomena from a theoretical point of view, where experiments are challenging or even impossible because of extreme conditions.

In this context, the strong scaling efficiency of *ms*2 was analyzed for its hybrid MPI + OpenMP parallelization. Combining MPI and OpenMP, the memory demand of *ms*2 was optimized such that simulations with a larger particle number N can be achieved. In Fig. 6b, the vertical axis shows the computing power (nodes) times computing time per computing intensity (problem size), thus, horizontal lines would show a strong scaling efficiency of 100 %. From Fig. 6b, it becomes clear that *ms*2 is close to optimal strong scaling. However, the computing intensity of traversing the particle matrix is proportional to N^2, but intermolecular interactions are calculated for particles that are in the cutoff sphere only. As a result, the scaling of *ms*2 should be in between N to N^2. Fig. 6b indicates that almost doubling the number of particles ($120000/64000 = 1.875$) in *ms*2 leads to an increase of computational cost of a factor around 1.91 if the number of nodes was chosen appropriately so that the overhead is small (comparison of red and blue symbols for 800 nodes).

6 Relative permittivity of mixtures

The relative permittivity of a fluid ε, also known as the dielectric constant, indicates how that fluid weakens an external electric field compared to vacuum. While experimental data on the relative permittivity are available for many pure fluids (at least under ambient conditions), measurements of the relative permittivity for mixtures have rarely been reported. However, such information is important for chemical engineering, e.g. for electrolyte solutions with mixed solvents or solutions of weak electrolytes [35].

On the molecular scale, the relative permittivity is directly related to the mutual orientation of the molecular dipoles via Kirkwood's theory [32]. Thus, the relative permittivity can be sampled straightforwardly with molecular simulations in the canonical (NVT) ensemble via

$$\varepsilon - 1 = \frac{4\pi}{3k_\mathrm{B}TV}(\langle \boldsymbol{M}^2 \rangle - \langle \boldsymbol{M} \rangle^2), \tag{9}$$

where k_B is Boltzmann's constant, T the temperature, V the volume and $\boldsymbol{M}$ the total dipole moment of the simulation volume that is obtained by summing up all molecular dipole moment vectors

$$\boldsymbol{M} = \sum_{i=1}^{N} \boldsymbol{\mu}_i. \tag{10}$$

Hence, molecular simulations are an ideal tool to study the relative permittivity of mixtures. It has recently been shown that with existing molecular models for mixtures of molecular fluids [34] and electrolyte solutions [45], at least qualitative agreement with experimental data can be obtained. To further demonstrate the power of this predictive approach also in a quantitative manner, MD simulations of the relative permittivity of the mixture acetone + water were carried out with

the molecular simulation tool *ms*2 [44], using the TIP4P/ε water model [15] and the acetone model by Windmann et al. [58]. These models are known to yield the pure component permittivities excellently. The Lorentz-Berthelot combining rules were applied so that the simulation results for the mixture are strictly predictive. First, the mixture density was obtained with isothermal-isobaric (NpT) runs and then the relative permittivity was sampled with NVT simulations. The results in Fig. 7 demonstrate that a good prediction of the mixture permittivity was obtained.

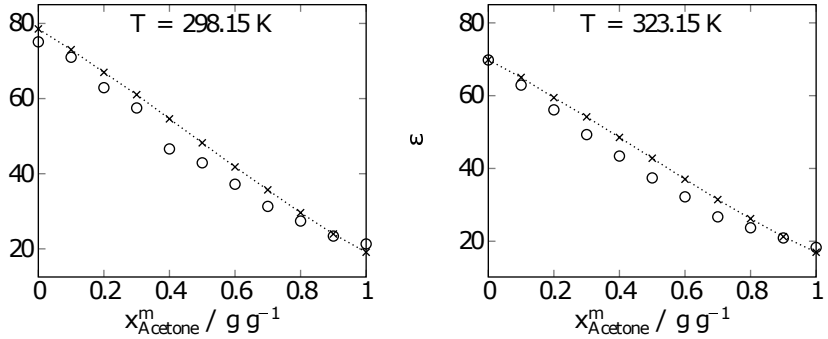


Fig. 7: Relative permittivity of mixtures of water and acetone as a function of the acetone mass fraction at two different temperatures and 1 bar. Crosses show the experimental data by Åkerlöf [1], open circles denote present simulation results. Simulation uncertainties are within symbol size, dotted lines are guides to the eye.

7 Reliability and reproducibility of simulation data

Molecular simulations have become a well established alternative to laboratory experiments for predicting thermophysical properties of fluids [50, 51, 53]. Evidently, the reliability and reproducibility of such predictions is of fundamental importance.

To sample thermophysical properties of a given molecular model, computer experiments can be carried out. In general, the simulation result of a given observable x^{sim} will not agree with the true model value x^{mod} [20]. Like in laboratory experiments, errors can also occur in computer experiments [46] that can cause deviations between the true value x^{mod} and the value observed in simulation x^{sim} [20, 46]. Both stochastic and systematic errors may in general occur in computer experiments. While techniques to assess statistical errors are well established for computer simulations [2, 13, 14], it is more difficult to deal with systematic errors, which have a significant influence on the reliability of the results. Systematic errors may be a consequence of erroneous algorithms, user errors, differences due to different simulation methods, finite size effects, erroneous evaluation of long-range interactions, insufficient equilibration or production periods, compilers, paralleliza-

tion, hardware architecture etc. [46]. As in laboratory experiments, round Robin studies can be made for quantifying systematic errors, in which the same simulation task is carried out by different groups with different programs.

The detection and assessment of outliers in large datasets is a standard task in the field of data science, but has to the best of our knowledge not yet been applied to thermophysical property data obtained by molecular simulation. The assessment of experimental thermophysical property data is a well-established field in chemical engineering [11], especially for phase equilibrium data [30, 31].

The accuracy with which properties of a simple (Lennard-Jones) model fluid can currently be determined by molecular simulation was assessed. The Lennard-Jones potential is often used as a starting point for the development of many force fields for complex molecules [52]. It is often taken as a benchmark for the validation of simulation codes and the test of new simulation techniques. Accordingly, a large number of computer experiment data are available for this fluid. Molecular simulations were performed both for homogeneous state points and for the vapor-liquid equilibrium to complement the data in regions that were only sparsely investigated in the literature. This database (cf. Fig. 8) allows for a systematic data evaluation and determination of outliers. In total, about 35,000 data points were evaluated [54]. The VLE properties: vapor pressure, saturated densities, enthalpy of vaporization and surface tension were investigated; for homogeneous state points, the investigated properties were: pressure, thermal expansion coefficient, isothermal compressibility, thermal pressure coefficient, internal energy, isochoric heat capacity, isobaric heat capacity, Grüneisen parameter, Joule-Thomson coefficient, speed of sound, Helmholtz energy and chemical potential.

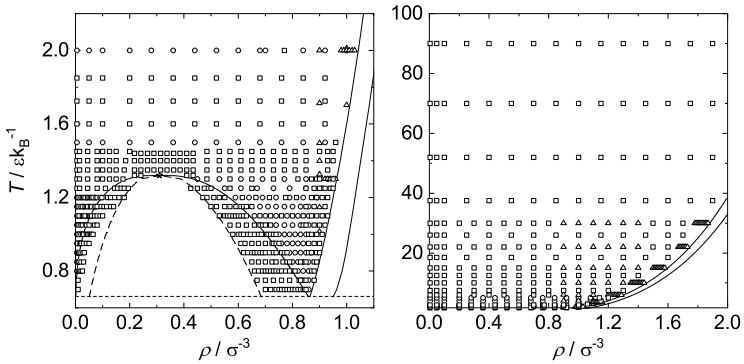


Fig. 8: Overview of about 1000 state points that were studied with the Lustig formalism [39] by different authors. Circles: Thol et al. [57]; triangles: Köster et al. [36]; squares: this work. Data for the Helmholtz energy and its density and inverse temperature derivatives up to second-order were available for the stable state points. For the metastable state points, the derivatives were available up to first-order.

Different consistency tests were applied to assess the accuracy and precision and thereby the reliability of the data. The data on homogeneous states were evaluated point-wise using data from their respective vicinity and EOS. Approximately 10% of all homogeneous bulk data were identified as gross outliers. VLE data were assessed by tests based on the compressibility factor, the Clausius-Clapeyron equation and by an outlier test. First, consistency tests were used to identify unreliable datasets. In a subsequent step, the mutual agreement of the remaining datasets was evaluated. Seven particularly reliable VLE data sets were identified. The mutual agreement of these data sets is approximately $\pm 1\%$ for vapor pressure, $\pm 0.2\%$ for saturated liquid density, $\pm 1\%$ for saturated vapor density and $\pm 0.75\%$ for enthalpy of vaporization – excluding the extended critical region. In most cases, the results from different datasets were found to differ by more than the combined statistical uncertainty of the individual data. Hence, the magnitude of systematic errors often exceeds that from stochastic errors.

8 Data management

While it is generally always advisable to follow good practices of data management when dealing with research data, this becomes even more expedient in cases where the data have been obtained by accessing dedicated facilities, as it is the case in scientific high-performance computing: Simulation results without annotation become *dark data*, making their meaning and purpose unintelligible to others, in particular, to automated processing on repositories and computing environments [47]. HPC and other facilities are not employed adequately if they are used to generate dark data. Obversely, annotating the simulation outcome with appropriate metadata enhances its value to the community and ensures that data become and remain FAIR, i.e., findable, accessible, interoperable, and reusable, permitting their preservation far beyond the immediate circumstances that motivated their creation originally [6, 7, 48].

Data infrastructures, such as repositories, digital marketplaces or modelling and simulation environments, often follow a multi-tier design with an explicit logical or semantic layer, as illustrated by Fig. 9. In these cases, the underlying semantic technology, which may include non-relational databases, mechanisms for checking constraints, or handling digital objects, requires mechanisms for knowledge representation. This technical requirement is the main underlying cause of the increasing pressure on scientific communities to develop standardized schema metadata definitions or ontologies. Such metadata standards are known as semantic assets; an agreement on semantic assets establishes semantic interoperability.

In materials modelling, understood here as roughly comprising the fields of computational molecular engineering (CME) for soft matter [29] and integrated computational materials engineering (ICME) for solid materials [49], a major community-governed effort towards metadata standardization is conducted by the European Materials Modelling Council (EMMC), specifically the EMMC focus areas on digital-

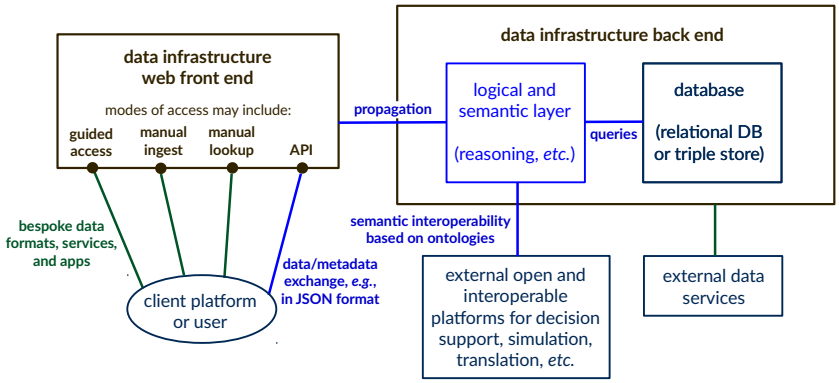


Fig. 9: Role of semantic technology within interoperable data infrastructures, illustrated for the case of a typical multi-tier architecture.

ization and interoperability, supported by a series of projects funded from the Horizon 2020 research and innovation programme, including VIMMP [28]. This approach, which is implemented by the present data management concept, is based on a system of ontologies that permits the characterization of CME/ICME data provenance at multiple levels of abstraction:

To facilitate *technical-level reproducibility*, metadata documenting all boundary conditions, technical parameters of the employed software and circumstances related to workflow execution need to be provided, including details on the hardware architecture and the mode of parallelization. Semantic assets that can be used for this purpose include the VIMMP ontologies MACRO, OSMO, VISO and VOV [28] in combination with the PaaSPort ontology [4]. Documenting a workflow manually in this way, at full detail, is not recommended except in the case of very straightforward scenarios; the complete viability of such an approach would require the automated annotation by an integration of ontologies with workflow management systems [29].

In a *logical representation* of a simulation workflow, details of the technical implementation are left out of consideration; instead, the workflow is described in terms of the involved use cases, models, solvers and processors, which are defined by their function rather than by their practical realization. This approach was introduced by the EMMC through the development of the MODA (Model Data) workflow description standard [8]; this is an adequate level of annotation for most purposes, as it permits documenting the provenance of simulation results as well as the intended use cases in a similar way as it is usually done in a scientific journal article. However, metadata provided in this way are machine-processable and standardized by an ontology – in the present case, by the Ontology for Simulation, Modelling, and Optimization (OSMO), i.e., the ontology version of MODA [29]. Logical data transfer (LDT) notation is a graph-based visualization of such workflow descriptions, which was also developed on the basis of a similar notation from MODA. The

LDT graph corresponding to the scenario from Section 1 is shown in Fig. 10; this graph corresponds to a workflow description in terms of OSMO and to a collection of digital objects that can be ingested into (and extracted from) an interoperable data infrastructure, e.g., in JSON or JSON-LD format, cf. Fig. 9.

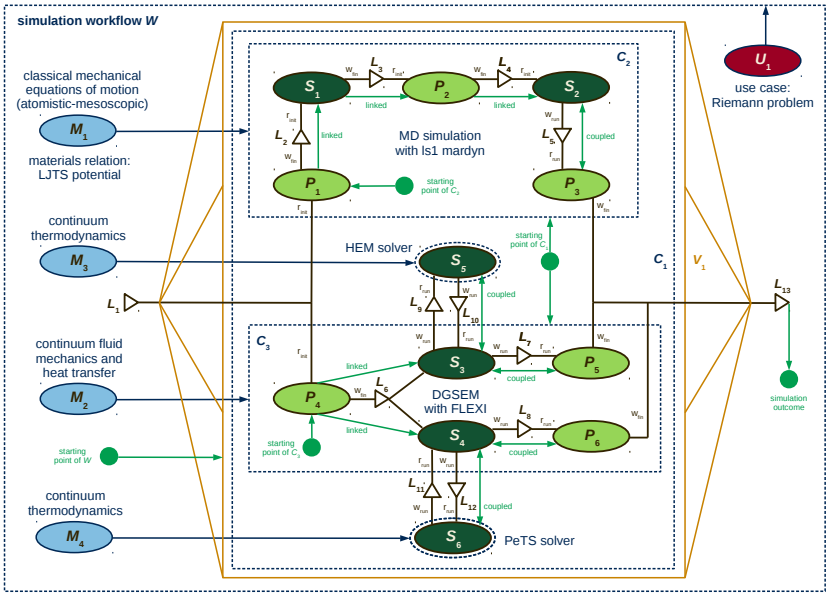


Fig. 10: Logical data transfer (LDT) provenance documentation for simulations by Hitz et al. [26] addressing the Riemann problem use case from Section 1. The LDT notation was presented in detail in previous work [29].

For a high-level representation of CME/ICME scenarios, a conceptualization of *modelling and simulation workflows as semioses* is developed on the basis of the European Materials and Modelling Ontology (EMMO) [12, 17]. As a top-level ontology, the main purpose of the EMMO consists in establishing the foundations for a coherent architecture of semantic assets at the highest possible degree of abstraction. Due to the nature of this work, technical and philosophical requirements need to be reconciled; the present stage of these developments is discussed in a recent report [27].

Acknowledgements The co-authors M.H., R.S.C., S.H., G.G.-C., R.F., M.K., S.S. and J.V. acknowledge funding by Deutsche Forschungsgemeinschaft (DFG) through the Project SFB-TRR 75, Project number 84292822 - "Droplet Dynamics under Extreme Ambient Conditions", and the co-author M.T.H. acknowledges funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 760907 (VIMMP). This work was carried out under the auspices of the Boltzmann-Zuse Society of Computational Molecular Engineering (BZS), and it was facilitated by activities of the Innovation Centre for Process Data Technology (Inprodat e.V.),

Kaiserslautern. The simulations were performed on the CRAY XC40 (*Hazel Hen*) at the High Performance Computing Center Stuttgart (HLRS). Discussions with S. Chiacchiera, B. Andreon, E. Bayro Kaiser, W. L. Cavalcanti, A. Fiseni, G. Goldbeck, A. Scotto di Minico, M. A. Seaton, S. Stephan, and I. T. Todorov are acknowledged.

References

1. Åkerlöf, G.: Dielectric constants of some organic solvent-water mixtures at various temperatures. J. Am. Chem. Soc. **54**, 4125–4139 (1932)
2. Allen, M.P., Tildesley, D.J.: Computer Simulation of Liquids. Oxford University Press, Oxford (1989)
3. Allie-Ebrahim, T., Russo, V., Ortona, O., Paduano, L., Tesser, R., Di Serio, M., Singh, P., Zhu, Q., Moggridge, G.D., D'Agostino, C.: A Predictive Model for the Diffusion of a Highly Non-Ideal Ternary System. Phys. Chem. Chem. Phys. **20**, 18436–18446 (2018)
4. Bassiliades, N., Symeonidis, M., Gouvas, P., Kontopoulos, E., Meditskos, G., Vlhavas, I.: PaaSPort semantic model: An ontology for a platform-as-a-service semantically interoperable marketplace. Data Knowl. Eng. **113**, 81–115 (2018)
5. Belonoshko, A.B., Davis, S., Skorodumova, N.V., Lundow, P.H., Rosengren, A., Johansson, B.: Properties of the fcc lennard-jones crystal model at the limit of superheating. Phys. Rev. B **76**, 064121 (2007)
6. Bicarregui, J.: Building and sustaining data infrastructures: Putting policy into practice. Technical report, Wellcome Trust (2016)
7. Bicarregui, J., Gray, N., Henderson, R., Jones, R., Lambert, S., Matthews, B.: Data management and preservation planning for Big Science. Int. J. Digit. Curation **8**, 29–41 (2013)
8. CEN-CENELEC Management Centre: Materials modelling: Terminology, classification and metadata. CEN workshop agreement, Brussels (2018)
9. Chatwell, R.S., Vrabec, J.: Bulk viscosity of liquid noble gases. J. Chem. Phys. **152**, 094503 (2020)
10. Cussler, E.L.: Mass Transfer in Fluid Systems, 2nd edn. Cambridge University Press, Cambridge (1997)
11. Dong, Q., Yan, X., Wilhoit, R.C., Hong, X., Chirico, R.D., Diky, V.V., Frenkel, M.: Data quality assurance for thermophysical property databases: Applications to the TRC SOURCE data system. J. Chem. Inf. Comput. Sci. **42**, 473–480 (2002)
12. EMMC Coordination and Support Action: European Materials and Modelling Ontology (EMMO). https://github.com/emmo-repo/ and https://emmc.info/emmo-info/ (2020)
13. Flyvbjerg, H., Petersen, H.G.: Error estimates on averages of correlated data. J. Chem. Phys. **91**, 461–466 (1989)
14. Frenkel, D.: Simulations: The dark side. Eur. Phys. J. Plus **128**, 10 (2013)
15. Fuentes-Azcatl, R., Alejandre, J.: Non-polarizable force field of water based on the dielectric constant: TIP4P/ε. J. Phys. Chem. B **118**, 1263–1272 (2014)
16. Ganguly, P., van der Vegt, N.F.A.: Convergence of Sampling Kirkwood–Buff Integrals of Aqueous Solutions with Molecular Dynamics Simulations. J. Chem. Theory Comput. **9**, 1347–1355 (2013)
17. Goldbeck, G., Ghedini, E., Hashibon, A., Schmitz, G.J., Friis, J.: A reference language and ontology for materials modelling and interoperability. In: Proceedings of the NAFEMS World Congress 2019, Québec, p. NWC_19_86. NAFEMS, Knutsford, UK (2019)
18. Grottel, S., Krone, M., Müller, C., Reina, G., Ertl, T.: Megamola prototyping framework for particle-based visualization. IEEE Trans. Vis. Comput. Graph. **21**, 201–214 (2015)
19. Guevara-Carrion, G., Fingerhut, R., Vrabec, J.: Fick Diffusion Coefficient Matrix of a Quaternary Liquid Mixture by Molecular Dynamics. J. Phys. Chem. B **124**, 4527–4535 (2020)

20. Hasse, H., Lenhard, J.: Boon and bane: On the role of adjustable parameters in simulation models. In: Mathematics as a Tool: Tracing New Roles of Mathematics in the Sciences, pp. 93–115. Springer International Publishing (2017)
21. Heier, M., Stephan, S., Liu, J., Chapman, W.G., Hasse, H., Langenbach, K.: Equation of state for the lennard-jones truncated and shifted fluid with a cut-off radius of 2.5 based on perturbation theory and its applications to interfacial thermodynamics. Mol. Phys. **116**, 2083–2094 (2018)
22. Heinen, M., Vrabec, J.: Evaporation sampled by stationary molecular dynamics simulation. J. Chem. Phys. **151**, 044704 (2019)
23. Heinen, M., Vrabec, J., Fischer, J.: Communication: Evaporation: Influence of heat transport in the liquid on the interface temperature and the particle flux. J. Chem. Phys. **145**, 081101 (2016)
24. Hindenlang, F., Gassner, G.J., Altmann, C., Beck, A., Staudenmaier, M., Munz, C.D.: Explicit discontinuous galerkin methods for unsteady problems. Comput. Fluids **61**, 86–93 (2012)
25. Hitz, T., Heinen, M., Vrabec, J., Munz, C.D.: Comparison of macro- and microscopic solutions of the riemann problem i. supercritical shock tube and expansion into vacuum. J. Comput. Phys. **402**, 109077 (2020)
26. Hitz, T., Jöns, S., Heinen, M., Vrabec, J., Munz, C.D.: Comparison of macro- and cicroscopic solutions of the Riemann problem. II. Two-phase shock tube. Preprint manuscript (2020)
27. Horsch, M.T., Chiacchiera, S., Seaton, M.A., Todorov, I.T.: Multiscale modelling and simulation of physical systems as semiosis. Technical report, Innovation Centre for Process Data Technology, Kaiserslautern (2020)
28. Horsch, M.T., Chiacchiera, S., Seaton, M.A., Todorov, I.T., Šindelka, K., Lísal, M., Andreon, B., Kaiser, E.B., Mogni, G., Goldbeck, G., Kunze, R., Summer, G., Fiseni, A., Brüning, H., Schiffels, P., Cavalcanti, W.L.: Ontologies for the Virtual Materials Marketplace. KI - Künstliche Intelligenz **34**, 423–428 (2020)
29. Horsch, M.T., Niethammer, C., Boccardo, G., Carbone, P., Chiacchiera, S., Chiricotto, M., Elliott, J.D., Lobaskin, V., Neumann, P., Schiffels, P., Seaton, M.A., Todorov, I.T., Vrabec, J., Cavalcanti, W.L.: Semantic interoperability and characterization of data provenance in computational molecular engineering. J. Chem. Eng. Data **65**, 1313–1329 (2020)
30. Kang, J.W., Diky, V., Chirico, R.D., Magee, J.W., Muzny, C.D., Abdulagatov, I., Kazakov, A.F., Frenkel, M.: Quality assessment algorithm for vapor-liquid equilibrium data. J. Chem. Eng. Data **55**, 3631–3640 (2010)
31. Kang, J.W., Diky, V., Chirico, R.D., Magee, J.W., Muzny, C.D., Kazakov, A.F., Kroenlein, K., Frenkel, M.: Algorithmic framework for quality assessment of phase equilibrium data. J. Chem. Eng. Data **59**, 2283–2293 (2014)
32. Kirkwood, J.: The dielectric polarization of polar liquids. J. Chem. Phys. **7**, 911–919 (1939)
33. Kirkwood, J.G., Buff, F.P.: The statistical mechanical theory of solutions. I. J. Chem. Phys. **19**, 774–777 (1951)
34. Kohns, M.: Molecular simulation study of dielectric constants of pure fluids and mixtures. Fluid Phase Equilib. **506**, 112393 (2020)
35. Kohns, M., Lazarou, G., Kournopoulos, S., Forte, E., Perdomo, F.A., Jackson, G., Adjiman, C.S., Galindo, A.: Predictive models for the phase behaviour and solution properties of weak electrolytes: nitric, sulphuric, and carbonic acids. Phys. Chem. Chem. Phys. **22**, 15248–15269 (2020)
36. Köster, A., Mausbach, P., Vrabec, J.: Premelting, solid-fluid equilibria, and thermodynamic properties in the high density region based on the Lennard-Jones potential. J. Chem. Phys. **147**, 144502 (2017)
37. Krishna, R., van Baten, J.M.: The Darken Relation for Multicomponent Diffusion in Liquid Mixtures of Linear Alkanes: An Investigation Using Molecular Dynamics (MD) Simulations. Ind. Eng. Chem. Res. **44**, 6939–6847 (2005)
38. Krüger, P., Vlugt, T.J.H.: Size and shape dependence of finite-volume kirkwood-buff integrals. Phys. Rev. E **97**, 051301 (2018)
39. Lustig, R.: Statistical analogues for fundamental equation of state derivatives. Mol. Phys. **110**, 3041–3052 (2012)

40. Mausbach, P., Fingerhut, R., Vrabec, J.: Structure and dynamics of the lennard-jones fcc-solid focusing on melting precursors. J. Chem. Phys. **153**, 104506 (2020)
41. Milzetti, J., Nayar, D., van der Vegt, N.F.A.: Convergence of Kirkwood–Buff Integrals of Ideal and Nonideal Aqueous Solutions Using Molecular Dynamics Simulations. J. Phys. Chem. B **122**, 5515–5526 (2018)
42. Niethammer, C., Becker, S., Bernreuther, M., Buchholz, M., Eckhardt, W., Heinecke, A., Werth, S., Bungartz, H.J., Glass, C.W., Hasse, H., Vrabec, J., Horsch, M.: ls1 mardyn: The massively parallel molecular dynamics code for large systems. J. Chem. Theory Comput. **10**, 4455–4464 (2014)
43. Peters, C., Thien, J., Wolff, L., Koß, H.J., Bardow, A.: Quaternary Diffusion Coefficients in Liquids from Microfluidics and Raman Microspectroscopy: Cyclohexane + Toluene + Acetone + Methanol. J. Chem. Eng. Data **65**, 1273–1288 (2020)
44. Rutkai, G., Köster, A., Guevara-Carrion, G., Janzen, T., Schappals, M., Glass, C.W., Bernreuther, M., Wafai, A., Stephan, S., Kohns, M., Reiser, S., Deublein, S., Horsch, M., Hasse, H., Vrabec, J.: ms2: A molecular simulation tool for thermodynamic properties, release 3.0. Comput. Phys. Commun. **221**, 343–351 (2017)
45. Saric, D., Kohns, M., Vrabec, J.: Dielectric constant and density of aqueous alkali halide solutions by molecular dynamics: A force field assessment. J. Chem. Phys. **152**, 164502 (2020)
46. Schappals, M., Mecklenfeld, A., Kröger, L., Botan, V., Köster, A., Stephan, S., Garcia, E.J., Rutkai, G., Raabe, G., Klein, P., Leonhard, K., Glass, C.W., Lenhard, J., Vrabec, J., Hasse, H.: Round robin study: Molecular simulation of thermodynamic properties from models with internal degrees of freedom. J. Chem. Theory Comput. **13**, 4270–4280 (2017)
47. Schembera, B., Durán, J.M.: Dark data as the new challenge for big data science and the introduction of the scientific data officer. Philos. Tech. **33**, 93–115 (2019)
48. Schembera, B., Iglezakis, D.: EngMeta: metadata for computational engineering. Int. J. Metadata Semant. Ontol. **14**, 26 (2020)
49. Schmitz, G.J.: Microstructure modeling in integrated computational materials engineering (ICME) settings: Can HDF5 provide the basis for an emerging standard for describing microstructures? JOM **68**(1), 77–83 (2015). DOI 10.1007/s11837-015-1748-2
50. Stephan, S., Becker, S., Langenbach, K., Hasse, H.: Vapor-liquid interfacial properties of the binary system cyclohexane + CO2: Experiment, molecular simulation and density gradient theory. Fluid Phase Equilib. **518**, 112583 (2020)
51. Stephan, S., Hasse, H.: Interfacial properties of binary mixtures of simple fluids and their relation to the phase diagram. Phys. Chem. Chem. Phys. **22**, 12544–12564 (2020)
52. Stephan, S., Horsch, M., Vrabec, J., Hasse, H.: MolMod - an open access database of force fields for molecular simulations of fluids. Mol. Simul. **45**, 806–814 (2019)
53. Stephan, S., Langenbach, K., Hasse, H.: Interfacial properties of binary Lennard-Jones mixtures by molecular simulations and density gradient theory. J. Chem. Phys. **150**, 174704 (2019)
54. Stephan, S., Thol, M., Vrabec, J., Hasse, H.: Thermophysical properties of the Lennard-Jones fluid: Database and data assessment. J. Chem. Inf. Model. **59**, 4248–4265 (2019)
55. Taylor, R., Krishna, R.: Multicomponent Mass Transfer. John Wiley & Sons, New York (1993)
56. Tchipev, N., Seckler, S., Heinen, M., Vrabec, J., Gratl, F., Horsch, M., Bernreuther, M., Glass, C.W., Niethammer, C., Hammer, N., Krischok, B., Resch, M., Kranzlmüller, D., Hasse, H., Bungartz, H.J., Neumann, P.: Twetris: Twenty trillion-atom simulation. Int. J. High Perform. Comput. Appl. **33**, 838–854 (2019)
57. Thol, M., Rutkai, G., Köster, A., Lustig, R., Span, R., Vrabec, J.: Equation of state for the Lennard-Jones fluid. J. Phys. Chem. Ref. Data **45**, 023101 (2016)
58. Windmann, T., Linnemann, M., Vrabec, J.: Fluid phase behavior of nitrogen + acetone and oxygen + acetone by molecular simulation, experiment and the pengrobinson equation of state. J. Chem. Eng. Data **59**, 28–38 (2014)