1. Optimization of effective Lennard-Jones model
   1. General relations

To optimize the Lennard-Jones interaction parameters to match the experimentally determined constraints, we will use the recently developed principles of statistical distance minimization.[1](#_ENREF_1),[2](#_ENREF_2) In this approach the optimization loss function is based on comparing the statistics of measurement outcomes performed on the target experimental system and its model. The statistical distance metric *s* is defined as a measure of closeness of the probability distributions of measurement outcomes (*i.e.*, normalized histogram) performed on the target system *P* and its model *Q* according to

, (1)

where *pi* and *qi* are the probabilities of registering elementary outcome *i* upon measurement of systems *P* and *Q*, with the summation running over all such outcomes. The argument *CB*, known as the Bhattacharyya coefficient, can also be defined for continuum distributions as,

 , (2)

where for the continuum space  the summation is replaced by integration over the coordinates of this space , while  and  are probability densities at .

Equivalent but more convenient form of the loss function minimizing statistical distance *s* can be written as the negative logarithm of *CB*, also known as the Bhattacharyya distance,

 . (3)

If a series of *M* independent measurements are combined, the resulting probabilities *pi* of different outcomes of such compound measurements are given by product

. (4)

where *pm,im* is the probability of outcome *im* in measurement *m*. The corresponding compound coefficient *CM* is then obtained as

. (5)

where *CB,m* is the Bhattacharyya coefficient, Eq. (2), for measurement *m*.

In the following we will adjust the Lennard-Jones potential parameters to minimize *DB*, with the particular form of the loss function depending on the probability distributions describing the measurement outcomes. In the present case, our reference information comes from the equation of state and shear viscosity in the low-density limit at high temperatures. Specifically, the data are in the form of the second virial coefficients and collision integrals.

* 1. Second virial coefficient matching

As the first step we need to translate the available target data, i.e., second virial coefficients, into the language of the probability distributions of appropriate measurement outcomes. Here we will be approximating the equation of state (EOS) of water vapor in the limit of low density or pressure. The thermodynamic behavior of such fluid can be described by the virial EOS truncated at the second order, i.e.,

 , (3)

where *P* is the system pressure, *kB* is the Boltzmann constant, *T* is temperature, *B* is the second virial coefficient, and  is the density, with *N* being the number of particles and *V* the volume of the system.

Having this information, we can reconstruct the probability distribution of coarse-grained system states defining its thermodynamics. This can be achieved using the fluctuation relations.[3](#_ENREF_3) Given the EOS, Eq. (3), we can reconstruct the probability distribution of particle number *N* or density  in volume *V* at temperature *T* and chemical potential , which define the thermodynamic parameters of the grandcanonical ensemble. The particle number distribution attains the Gaussian form,

 (4)

Here  denotes the average number of particles in the simulation box, with the corresponding variance calculated as . The mean and variance of the Gaussian are related to density  and isothermal compressibility  as  and .

The goal of the model optimization can be thought as a way of trying to decide whether the measurement of a given particle number came from the target system or from the model. The more indistinguishable the two are based on such measurement, the better the model. The optimum is then achieved by minimization of the loss function Eq. (1). The Bhattacharyya coefficient, Eq. (2), for two Gaussian functions, Eq. (4), can be written as,

 (5)

If we require our model to represent the thermodynamic properties at different temperatures, we need to match simultaneously all the corresponding Gaussians in Eq. (4), represented by the multivariate Gaussian distribution,

. (6)

with the loss function based on Eqs. (3) and (5) acquiring the form,

. (8)

Note that the first term within the curly brackets is an extensive quantity and grows with the system size *V*, whereas the second term is system-size independent. This means that in the thermodynamic limit of , only the first term is significant and the loss function can be simplified as,

 . (9)

The densities and compressibilities of the target and model systems can be obtained from the EOS, Eq. (3), and then the loss function can be re-expressed as a function of the second virial coefficients of the target and model systems.

First, we use the EOS, Eq. (3), to express the isothermal compressibilities as functions of densities as,

 , (10)

to obtain

 . (11)

Then, we can express the densities  and  in reduced forms in terms of ideal gas density at the same conditions. Based the chemical equilibrium

 , (12)

were  and  are the activity coefficients, we can write

 . (13)

Eq. (11) then transforms into

 . (14)

The inverse activity coefficient can be expressed as an ensemble average

 . (15)

For  the ensemble average can be written as,

 . (16)

where  is potential energy between the pair of particles of system *P* at distance *r*. The coefficient given in Eq. (16) can be related to the second virial coefficient *Bp* as,

 . (17)

where we used the definition of the second virial coefficient,

 . (18)

After substitution to Eq. (14) we have,

 (19)

Where for the approximate equality we used the low the fact that in the thermodynamic limit  and the low density limit  the denominator of the last fraction goes to 1.

In Eq. (19) both the ideal gas density and virial coefficients are temperature dependent. If volume and pressure are fixed between the different temperatures, the ideal gas density is proportional to the inverse of temperature,

 (20)

Consequently, the loss function proportional to Eq. (19) can be written as

 (21)

In the case of a model defined by the Lennard-Jones (12-6) potential, the second virial coefficient can be calculated from,

 . (22)

In the optimization procedure the Lennard-Jones parameters are varied and the resulting values of the model virial coefficient, *Bq*, are used along the experimentally determined *Bp* in the loss function Eq. (21), whose minimum is sought.

* 1. Diffusion coefficient matching

Following the same principles, it is possible to match dynamic properties of a model to that of the target experimental system. In the present case, we can minimize the distance between distributions describing the probability of finding a particle at distance *r*(*t*) at time *t* from its origin *r(*0*)* =0 at time *t*=0. For sufficiently long times the distances will be determined by diffusive motion with the probability density described by the Gaussian distribution

 (23)

where *Dp* is the diffusion coefficient of particles in system *P*. Since both the target system and model distributions are centered at zero, their mutual statistical distance is determined only by the variance of the Gaussian distribution, which is proportional to the diffusion coefficient. The corresponding Bhattacharyya coefficient is then

 , (24)

with the Bhattacharyya distance for the series of measurements at different temperatures *T*

. (25)

The diffusion coefficient at the low-density limit can be determined from the low- density limit of Eq. (46) as,

. (26)

In the case of the experimental system, value of  can be derived from the value of the diffusion coefficient, and in the case of the model it can be calculated from the reduced collision integrals as . Consequently, the loss function, Eq. (25), can be expressed as a function of the reduced diffusion coefficients as,

. (27)

The values of the reduced collision integrals as a function of model LJ parameters can be obtained from the relation Eq. (D7) with parameters listed in Table D1.

(1) Vlcek, L.; Chialvo, A. A. Rigorous Force Field Optimization Principles Based on Statistical Distance Minimization. *J. Chem. Phys.* **2015**, *143*, 15.

(2) Vlcek, L.; Vasudevan, R. K.; Jesse, S.; Kalinin, S. V. Consistent Integration of Experimental and Ab Initio Data into Effective Physical Models. *J. Chem. Theory Comput.* **2017**, 10.1021/acs.jctc.7b00114, doi: 10.1021/acs.jctc.1027b00114.

(3) Landau, L. D.; Lifshitz, E. M. *Statistical Physics*; Elsevier, 1980; Vol. 1.