To find a suitable loss function based on information from the second virial coefficients at different temperatures, we follow the model optimization principles based on statistical distance minimization outlined in our previous work.[1](#_ENREF_1) The basic idea of this approach rests on matching the outcomes of measurements performed on the target and model systems. We have shown that a suitable loss function, can be written as

 . (1)

Where *DB* and *CB* are known as the Bhattacharyya distance and coefficient, respectively. The Bhattacharyya coefficient is defined as

 , (2)

where the summation runs over all *k* states *i* of the system that can be distinguished by a measurement, while  and are the probabilities of state *i* in systems *P* (target) and *Q* (model). For a continuum phase space  the summation is replaced by integration over the coordinates of this space , while  and  are probability densities at .

The particular sets of distinguishable states *k* or the phase space  in Eq. (2) depend on a given set of experiments. 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 system states defining its thermodynamics. This can be achieved using the fluctuation relations.[2](#_ENREF_2) 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)

The Bhattacharyya coefficient between these multivariate Gaussians is defined as,

. (7)

with the loss function based on Eq. (1) 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 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) Vlcek, L.; Chialvo, A. A. Rigorous Force Field Optimization Principles Based on Statistical Distance Minimization. *J. Chem. Phys.* **2015**, *143*.

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