MOLECULAR SIMULATION
2021, VOL. 47, NO. 5, 395–407
https://doi.org/10.1080/08927022.2020.1747617

Taylor & Francis
Taylor & Francis Group

Flat-histogram extrapolation as a useful tool in the age of big data

Nathan A. Mahynski [a], Harold W. Hatch [a], Matthew Witman [b], David A. Sheen [a], Jeffrey R. Errington [c] and Vincent K. Shen[a]

[a]Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD, USA; [b]Sandia National Laboratories, Livermore, CA, USA; [c]Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, NY, USA

ABSTRACT

Here we review recent work by the authors to revisit the concept of extrapolating thermodynamic properties of classical systems using statistical mechanical principles. Specifically, we discuss how the combination of these principles with biased sampling techniques enables the prediction of free energy landscapes and other detailed information, such as structural properties, of the system in question. Remarkably accurate estimates of physical properties across a broad range of conditions have been achieved using this approach, greatly reducing the number of simulations needed to explore a given system's behaviour. While approximate, these extrapolations significantly amplify the amount of reasonably accurate information that can be extracted from simulations enabling a small set of them to feed data-intensive regression algorithms such as neural networks. Thus, this extrapolation methodology represents a useful tool for performing tasks such as high-throughput screening of physical properties, optimising force field parameters, exploring equilibrium phase behaviour, and enabling theory-guided data science for these systems.

ARTICLE HISTORY
Received 19 November 2019
Accepted 17 March 2020

KEYWORDS
Data science; extrapolation; thermodynamics

1. Introduction

Advances in computational resources during the twentieth century have progressively made simulations a more standard part of the scientific toolbox in many disciplines including the chemical sciences [1–4]. The large quantities of data produced by these tools created fertile ground in which novel algorithms and statistical methods soon gave rise to the 'big data' paradigm in science and engineering [5]. This paradigm for scientific exploration relies on the intensive analysis of large amounts of data to reach conclusions. In contrast to previous scientific paradigms ('empirical', 'theoretical', and 'computational'), this fourth paradigm [6] relies as much on data quantity as quality. When developing a theory to model a system, it is expected that a governing equation might be solved exactly or at least numerically to high accuracy if an analytical solution cannot be found. The result is a model that accurately and precisely describes the physics of the system of interest using only a few physical constants or other parameters, such as viscosity or temperature.

In contrast, statistical methods used in data science take large amounts of observations and combine them into a (generally) probabilistic model that makes predictions about the system's behaviour [7, 8]; because of the nature of the model, there is some associated (often quantifiable) uncertainty associated with the accuracy of these predictions. Furthermore, they rely on being provided a sufficient amount of data to be well trained. Of course, if inaccurate measurements or observations are made, the resulting model will produce erroneous results. Regardless, many approaches such as neural networks [7], Gaussian process regression [9], and other machine learning techniques [7, 10, 11] are capable of finding coherence in large amounts of data, and users are often willing to sacrifice some degree of accuracy at each individual data point in exchange for many more points. As the utility and popularity of data-intensive approaches grows, it is important to have appropriate techniques to generate and collect the large amounts of data on which they rely [12].

Indeed, a number of hybrid approaches combining molecular simulation with data-driven modelling have appeared recently. It has been demonstrated that relatively small amounts of simulation data can be used to regress models such as neural networks, autoencoders, and Gaussian processes to predict properties of fluids [13–15] , adsorption isotherms [16, 17], fit intermolecular potential functions [18, 19], develop simulation biases [20], and design self-assembling systems [21–23]. In the context of free energy simulations, much effort in molecular simulation over the past decades has been devoted to developing advanced computational tools to more accurately calculate properties of chemical systems [24]. While this is an important research objective, here we review some recent developments [25–30] by the authors to instead develop relatively simplified techniques which are reasonably fast, accurate, and yield a high density of data relative to their computational requirements. The principal driving force for the development of these approaches is to produce more data at a reduced computational cost, and by doing so, to enable data-intensive analysis. We review the theory and applications of these methods to date, and briefly highlight how they can be integrated within, and further enable, data-driven science in the field of molecular thermodynamics. This integration of physical results with data science is part of a broader theory-guided data science

CONTACT Nathan A. Mahynski nathan.mahynski@nist.gov

paradigm [31] that seeks to incorporate physics into data science methods to learn scientifically consistent models that are sound, interpretable, and can be generalised to make new scientific discoveries.

The basic technique we focus on is extrapolation *via* Taylor series expansion. This is a foundational concept in statistical mechanics that has been explored by many authors since at least the mid-twentieth century [32]. It is premised on the fact that derivatives of the free energy with respect to intensive variables are given by fluctuations of extensive properties, which are observable. For example, a temperature expansion of the Helmholtz free energy of a system may be expressed as a Taylor series in energy, which can be measured over the course of a simulation. Similar expansions exist for open ensembles as well [33] and have been used to predict phase coexistence of fluids [34–38], order-disorder transitions [39], and free energy changes in biophysical processes [40]. Perturbation theories such as those premised on statistical associating fluid theory (SAFT) can also directly link macroscopic equations of state to the molecular level Hamiltonian of a system [41, 42].

A central theme in previous simulation work is the expansion around a single state point, for example, at a single chemical potential, temperature and volume in a grand canonical simulation. Due to statistical noise and truncation error, these extrapolations have a limited range over which a prediction has reasonable accuracy. Here, we reconsider this notion of extrapolation, not of a single state point, but of an entire free energy landscape. This landscape may be obtained from flat-histogram simulations, which are biased simulations that force the simulated system to explore different regions of order parameter space over time. Flat-histogram methods are already appreciated as being rich in information since they can be reweighted to compute properties continuously over a range of certain conditions depending on the order parameter used to define the landscape; by combining the two concepts, we illustrate further information amplification which enables the prediction of thermodynamic and structural properties over an even wider range of conditions.

Specifically, we consider molecular systems for which Monte Carlo methods may be employed to measure these landscapes and the appropriate fluctuations needed to define the coefficients in these series expansions. We demonstrate that, despite some inherent truncation error, the error is often quite small, enabling accurate predictions of physical properties over a broad range of conditions. This paper is organised as follows. First, we review flat-histogram simulations and the relevant statistical mechanics in Section 2, followed by a brief derivation of the relevant extrapolation equations in Section 3. In Section 4, we illustrate how this can be applied to simple, single component systems in both bulk and confinement. Section 5 extends this to multicomponent mixtures. Next, we show how these principles may be extended to predict structural properties of these systems in Section 6. Finally, in Section 7, we discuss extensions of these extrapolation principles to include systems with internal degrees of freedom and to predict changes in the virial coefficients of fluids as a function of model parameters (alchemical transformations). Throughout, we discuss how this approach leads to data amplification making extrapolation techniques amenable to high-throughput computational screens which can be used in conjunction with data-intensive techniques.

2. Review of flat histogram methods

Flat-histogram, or 'density of states', simulations seek to explore a set of (macro)states defined by some order parameter, $\vec{\Psi}$, with equal probability [1, 43]. This is done by recording the frequency at which each state is visited or the transition probabilities between different states. In the latter case, detailed balance allows the reconstruction of the landscape, or (logarithm of) probability, of observing each state at equilibrium, $\ln\Pi(\vec{\Psi})$. Through the appropriate reweighting equations, these landscapes can provide information about a broad range of conditions often very different from that of the original simulation. Although numerous variants exist, here we employ a hybrid technique known as Wang-Landau Transition Monte Carlo (WL-TMMC) [44, 45]. This approach begins with an initial Wang-Landau (WL) stage [46, 47] to quickly produce an initial estimate of the landscape, which is subsequently refined with Transition Matrix Monte Carlo (TMMC) [48]. The former technique has a low tunnelling time during the early stages of simulation and can construct an initial guess quickly, but suffers from slow convergence rates. The opposite is true of TMMC, so a combination of the two tends to be optimal [44].

Depending on the thermodynamic ensemble of interest, statistical mechanics allows us to express the probability of observing a given microstate, $\pi(s)$. The unbiased Metropolis acceptance criterion of moving from microstate 'a ' to 'b' is

$$p_{\mathrm{u}} = \min\left[1, \frac{\pi(b)}{\pi(a)}\right]. \tag{1}$$

Note that a macrostate is defined as some collection of microstates with the same order parameter value, e.g., different configurations with the same number of particles present. In a flat-histogram simulation, a biasing function, $\eta(\vec{\Psi})$, is introduced to create artificially flat sampling of all macrostates. The simulation proceeds according to p_{bias}:

$$p_{\mathrm{bias}} = \min\left[1, \frac{\exp\left(\eta\left[\vec{\Psi}(b)\right]\right)\pi(b)}{\exp\left(\eta\left[\vec{\Psi}(a)\right]\right)\pi(a)}\right]. \tag{2}$$

The biasing function is simply the inverse of the macrostate distribution, which would result in an equally probable sampling of the system's macrostates:

$$\eta(\vec{\Psi}) = -\ln\Pi(\vec{\Psi}). \tag{3}$$

Simulations proceed by updating $\eta(\vec{\Psi})$ on-the-fly; once converged, the landscape is computed from the bias function according to Equation (3). WL simulations proceed by setting $\eta(\vec{\Psi}) = 0$ for all values of $\vec{\Psi}$. After each move, the estimated macrostate distribution is incrementally updated by a factor, f, which is progressively reduced over time in cycles to converge. The established protocols for this are described in detail

elsewhere [46, 49]:

$$\ln\Pi(\vec{\Psi}) = \ln\Pi(\vec{\Psi}) + \ln f\,. \tag{4}$$

Once the system has undergone enough cycles to establish a reasonable guess of the true $\ln\Pi(\vec{\Psi})$, TMMC begins. This constructs an estimate of the macrostate distribution by instead measuring the transition probabilities between different macrostates rather than the frequency of observing them. A collection matrix, C, is recorded as a simulation proceeds and measures the unbiased probability of moving between states, even as the simulation actually proceeds according to the biased acceptance rates:

$$C\left[\vec{\Psi}(a) \to \vec{\Psi}(b)\right] = C\left[\vec{\Psi}(a) \to \vec{\Psi}(b)\right] + p_{\mathrm{u}}, \tag{5}$$

$$C\left[\vec{\Psi}(a) \to \vec{\Psi}(a)\right] = C\left[\vec{\Psi}(a) \to \vec{\Psi}(a)\right] + (1 - p_{\mathrm{u}}). \tag{6}$$

The probability, $P[\vec{\Psi}(a) \to \vec{\Psi}(b)]$, of moving between two macrostates can be computed by normalising the transition rates:

$$P\left[\vec{\Psi}(a) \to \vec{\Psi}(b)\right] = \frac{C\left[\vec{\Psi}(a) \to \vec{\Psi}(b)\right]}{\sum_i C\left[\vec{\Psi}(a) \to \vec{\Psi}(i)\right]}. \tag{7}$$

Detailed balance then provides the relationship between this transition probability and the macrostate probability.

$$\ln\Pi\left[\vec{\Psi}(b)\right] = \ln\Pi\left[\vec{\Psi}(a)\right] + \ln\left(\frac{P\left[\vec{\Psi}(a) \to \vec{\Psi}(b)\right]}{P\left[\vec{\Psi}(b) \to \vec{\Psi}(a)\right]}\right). \tag{8}$$

In what follows, the specific method by which the macrostate is obtained is immaterial, and we take the landscape, $\ln\Pi(\vec{\Psi})$, as a known quantity.

3. Thermodynamic ensembles and typical extrapolation equations

At this point, we have left the order parameter, $\vec{\Psi}$, general. In this paper, we discuss extrapolation primarily with respect to systems in the grand canonical ensemble, so we review this instance in detail. In this ensemble, the number of particles in the simulation serves as a convenient order parameter to study self-assembly, first-order phase separation, and other phenomena. For a multicomponent mixture with k components, the system is described by the set of chemical potentials, $\vec{\mu} = (\mu_1, \mu_2, \cdots, \mu_k)$, volume, V, and temperature, T ($\beta \equiv 1/k_{\mathrm{B}}T$, where k_{B} is the Boltzmann constant). The partition function, $\Xi(\vec{\mu}, V, \beta)$, is given by

$$\Xi(\vec{\mu}, V, \beta) = \sum_{N_1}\sum_{N_2}\cdots\sum_{N_k}\exp\left(\beta\sum_{i=1}^{k}\mu_i N_i\right)Q(\vec{N}, V, \beta), \tag{9}$$

where $Q(\vec{N}, V, \beta)$ is the canonical partition function, and $\vec{N} = (N_1, N_2, \cdots, N_k)$. It is possible to simply take $\vec{\Psi} = \vec{N}$; however, the order parameter is now a vector and its size grows with the number of components. Commensurately, the (hyper)volume of macrostate space encompassing all possible $\vec{N}$ grows exponentially. Furthermore, it is practically difficult to define the boundaries as sets of possible joint number observations that are possible in an arbitrary mixture , for example, all possible (N_1, N_2) in a size-asymmetric binary mixture of Lennard-Jones particles. Even when defined, it is computationally intensive to sample these states exhaustively [50–52]. These boundaries must be well defined for any flat-histogram simulation since it will attempt to visit each possible state. It is often simpler to recast $\Xi(\vec{\mu}, V, \beta)$ in terms of the isochoric semigrand partition function, $\Upsilon(N_{\mathrm{tot}}; \Delta\vec{\mu}, V, \beta)$ [53, 54], which provides a scalar order parameter, N_{tot}:

$$\Xi(\vec{\mu}, V, \beta) = \sum_{N_{\mathrm{tot}}}\exp(\beta\mu_1 N_{\mathrm{tot}})\Upsilon(N_{\mathrm{tot}}; \Delta\vec{\mu}, V, \beta), \tag{10}$$

where

$$\Upsilon(N_{\mathrm{tot}}; \Delta\vec{\mu}, V, \beta) = \sum_{N_2}\exp(\beta\Delta\mu_2 N_2) \times \ldots \sum_{N_k}\exp(\beta\Delta\mu_k N_k)Q(\vec{N}, V, \beta). \tag{11}$$

Here, $\Delta\mu_i \equiv \mu_i - \mu_1$. Consequently, one may write the probability of a macrostate as

$$\ln\Pi(N_{\mathrm{tot}}; \mu_1, \Delta\vec{\mu}, V, \beta) = \beta\mu_1 N_{\mathrm{tot}} + \ln\Upsilon - \ln\Xi. \tag{12}$$

This equation gives rise to a reweighting relationship that enables the macrostate distribution to be recalculated for any value of μ_1, at the same $(\Delta\vec{\mu}, V, \beta)$, if the distribution is already known at another value of μ_1^0:

$$\ln\Pi(N_{\mathrm{tot}}; \mu_1) = \ln\Pi(N_{\mathrm{tot}}; \mu_1) + \beta(\mu_1 - \mu_1^0)N_{\mathrm{tot}} + C, \tag{13}$$

where C is a constant related to the difference between the grand canonical partition functions at different conditions, which may be neglected in practice. In this way, only one simulation needs to be performed to obtain a plethora of data. Note that once the macrostate distribution is known, grand canonical ensemble average properties follow directly as weighted averages of the states:

$$\langle X^\alpha\rangle = \frac{\sum_{N_{\mathrm{tot}}\in\alpha}\Pi(N_{\mathrm{tot}})\widetilde{X}(N_{\mathrm{tot}})}{\sum_{N_{\mathrm{tot}}\in\alpha}\Pi(N_{\mathrm{tot}})}. \tag{14}$$

Here α refers to the collection of macrostates that belong to a given phase, and $\widetilde{X}$ is the (isochoric semigrand ensemble) average property collected at each macrostate. This can found by segmenting the macrostate distribution according to any local minima that occur in the distribution. Two phases are in coexistence if they have equal pressures, which also follows from the macrostate distribution [55]:

$$P^\alpha = \frac{\ln\Xi^\alpha}{\beta V} = \frac{\ln\sum_{N_{\mathrm{tot}}\in\alpha}[\Pi(N_{\mathrm{tot}})/\Pi(0)]}{\beta V}. \tag{15}$$

In summary, we have a method to perform a simulation at a single condition, using a scalar order parameter, and use the results to obtain all properties as a function of arbitrary μ_1. However, this 'exact' reweighting relationship does not allow one to predict what happens when, for example, the temperature changes. To obtain a reweighting relationship for that,

the energy must also be collected as part of the order parameter, negating the original premise of seeking a simple, scalar one. The same is true for multicomponent systems if we wish to change μ_i (or equivalently, $\Delta\mu_i$) where we would need to collect the conjugate, N_i, for each i.

This is where an approximation to these reweighting equations enables amplification of data without the need for additional programming effort. We may expand the probability distribution at each macrostate in a Taylor series with respect to the intensive variables of interest that are not being reweighted with respect to, in this case, $\vec{\phi} = (\beta, \mu_2, \mu_3, \cdots, \mu_k)$, or equivalently, $\vec{\phi} = (\beta, \Delta\vec{\mu})$. Consider a general function, $g(N_{\text{tot}}; \mu_1, \vec{\phi})$; a second-order Taylor series of this function is given by

$$g(N_{\text{tot}}; \mu_1, \vec{\phi}) = g(N_{\text{tot}}; \mu_1, \overrightarrow{\phi^0}) + \delta\vec{\phi} \cdot \nabla g(N_{\text{tot}}; \mu_1, \overrightarrow{\phi^0}) + \frac{1}{2!}\left[\delta\vec{\phi} \cdot \mathbf{H}(N_{\text{tot}}; \mu_1, \overrightarrow{\phi^0}) \cdot \delta\vec{\phi}^{\mathrm{T}}\right], \tag{16}$$

where $\delta\vec{\phi} = (\delta\beta, \delta\Delta\mu_2, \cdots, \delta\Delta\mu_k)$. The gradient, $\nabla g(N_{\text{tot}}; \mu_1, \vec{\phi^0})$, and symmetric Hessian matrix, $\mathbf{H}(N_{\text{tot}}; \vec{\phi^0})$, contain partial derivatives of $g(N_{\text{tot}}; \mu_1, \vec{\phi})$ with respect to the intensive variables in $\vec{\phi}$. It is well known that such partial derivatives may be expressed using fluctuations in their conjugate extensive variables [56, 57], and therefore, if these properties are recorded during the simulation, one may compute all necessary coefficients of the Taylor series. For example [27],

$$\frac{\partial \ln\Pi(N_{\text{tot}}; \vec{\phi^0})}{\partial\beta} = \mu_1 N_{\text{tot}} + \sum_{i=2}^{k} \Delta\mu_i \widetilde{N}_i - \widetilde{U} + C, \tag{17}$$

where C is a constant that can be neglected. Here, N_{tot} is a fixed value for each macrostate and $\widetilde{X}$ is the average quantity, X, observed at that macrostate (N_{tot}) over the course of the simulation. This can be measured on-the-fly or reconstructed from a log file as a post-processing step. The latter can prove very helpful in employing this extrapolation scheme to previous work that was not originally performed with this approach in mind.

Second derivatives follow naturally and contain terms that are derivatives of isochoric semigrand averages, such as

$$\frac{\partial^2 \ln\Pi(N_{\text{tot}}; \vec{\phi^0})}{\partial\beta^2} = \sum_{i=2}^{k} \Delta\mu_i \left(\frac{\partial \widetilde{N}_i}{\partial\beta}\right) - \frac{\partial \widetilde{U}}{\partial\beta}. \tag{18}$$

These derivatives have closed-form expressions in terms of fluctuations of extensive quantities, $f(X, Y) \equiv \widetilde{(XY)} - \widetilde{(X)}\widetilde{(Y)}$, which are derived and presented in more detail elsewhere [27]. Higher order terms involve fluctuations of fluctuations, and may be generally described by Ursell functions [58], or connected correlation functions, but ultimately simply require knowledge of averaged extensive quantities raised to certain powers. Consequently, we simply need to measure

$$Z(N_{\text{tot}}; \vec{\xi}) = \overbrace{\left[N_1^{\xi_1} N_2^{\xi_2} \cdots N_k^{\xi_k} U^{\xi_u}\right]}^{\text{Extensive Conjugates}} N_{\text{tot}}^{\xi_n} \tag{19}$$

over the course of the simulation and average the resulting entries to obtain these moments. Here, $\vec{\xi} = (\xi_1, \xi_2, \cdots, \xi_k, \xi_u, \xi_n)$, where ξ_i are integers ranging from 0 to $\xi_{\max}$ which sets the maximum order to which extrapolation may be performed [27]. Averaged over the course of the simulation, this provides all the necessary information to perform an extrapolation. We note that higher order moments generally take longer to converge than lower order ones, and the accuracy with which these are measured is usually the limiting factor in this extrapolation approach since higher order expansions enable accurate predictions to be made further away from the originally simulated conditions.

4. Single component systems

First, we consider the simplest case of a single component system, where $N_{\text{tot}} = N_1 = N$, to illustrate extrapolation in temperature. Reweighting in temperature (in practice, β) often requires that the energy of the system be discretised in order to create a histogram [60]; this presents its own challenges in determining the bin width, for instance, which has led to the development of binless variants [61, 62]. Consider a simple square-well fluid presented in Figure 1 from [25]. A single flat histogram simulation was performed to measure $\ln\Pi(N)$ at a supercritical reduced temperature of $T^* = 1.35$. This fluid has a critical point of $T_c^* \approx 1.22$ so the initial simulation represents a supercritical condition where sampling is typically much easier than when subcritical. First, the macrostate distribution was reweighted to different values of μ at $T^* = 1.35$, then extrapolated to lower temperature at fixed μ. From this landscape, we computed the average density of the most stable phase at each (T, μ) to create various isotherms. Subsequent comparison with direct simulation at those conditions reveals essentially perfect agreement even down to deeply subcritical temperatures as low as $T^* = 1.05$.

The macrostate distributions themselves are shown in Figure 1 (b) when the distributions have been reweighted to the value of μ corresponding to vapour–liquid coexistence at subcritical temperatures. It is clear that the extrapolated distributions match nearly exactly, except for the intermediate region ($100 \lesssim N \lesssim 300$) between the peaks. These low-likelihood regions of N correspond to densities within the binodal where bulk phase separation would occur. However, both the location and height of the peaks are well predicted by extrapolation. Since the properties of a given phase are a weighted average of the individual canonical states, whose weights are determined by the macrostate distribution (cf. Equation 14), the bulk properties are essentially perfectly recovered. The binodal and enthalpy of vaporisation obtained from the Clausius–Clapeyron equation are shown in Figure 1(c) and, again, illustrate essentially perfect agreement to within statistical error.

Extrapolation to higher temperatures is even more accurate and is not reproduced here. Similar results have been obtained for other simple fluids, such as Lennard-Jones. Therefore, for these simple fluids, thermodynamic extrapolation enables enough information to be extracted from a single flat-histogram simulation to effectively represent a complete equation of state with a valid range from supercritical conditions to well below the vapour–liquid critical point.

This is not limited to bulk systems. For example, in Figure 2, we applied this thermodynamic extrapolation procedure to a

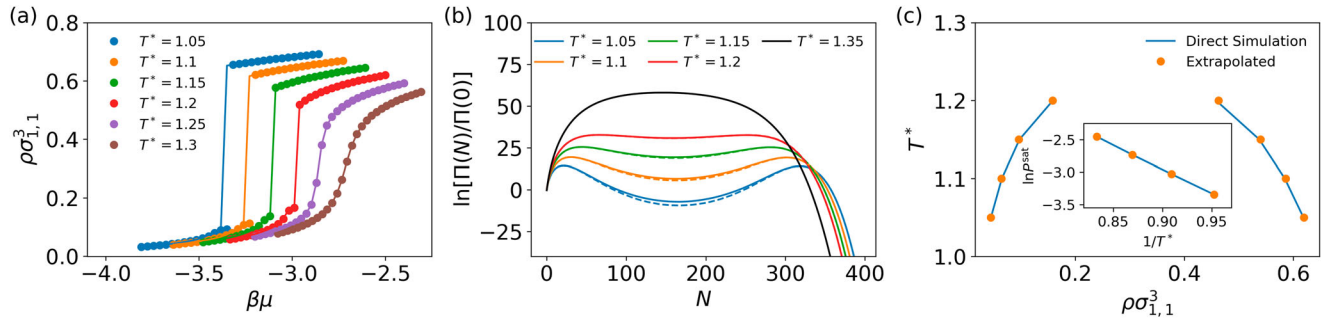

Figure 1. (Colour online) Comparison between extrapolation and direct simulation of the single component square-well system reported in [25]. (a) Isotherms obtained from direct simulations (lines) at discrete subcritical temperatures ($1.05 \leq T^* \leq 1.30$ in increments of 0.05) compared to those obtained from the extrapolation of a single supercritical simulation at $T^* = 1.35$ (points). (b) Representative macrostate distributions for this fluid from subcritical simulations (solid lines) compared to extrapolation (dashed lines) of the supercritical one (solid black line) used in (a). (c) Phase diagram (vapour–liquid coexistence) computed from direct, subcritical simulations compared to results obtained by extrapolating the supercritical one at $T^* = 1.35$.

model of argon condensing in a slit pore composed of solidified carbon dioxide [59]. This system exhibits a prewetting transition in which an initial low density film develops on the surface which can exhibit coexistence with a thicker film. This manifests in $\ln\Pi(N)$ as a pair of peaks occurring at low N, where each peak corresponds to different film thicknesses. Predictions were obtained by starting at the highest T^* reported then extrapolating to lower T^* using different maximum orders of terms in the underlying Taylor series. Second-order extrapolation seems to not capture certain features which ultimately manifest as systematic deviation in $\ln\Pi(N)$ at higher N values, whereas third-order extrapolation seems to capture this well but is more noisy. Running these simulations for a longer time to collect more accurate moments of the extensive properties would improve this; however, a natural trade-off would emerge between running at higher T^* to extrapolate better versus simply running a direct simulation at lower T^*. Achieving the balance that minimises computational cost will depend on the code efficiency (Monte Carlo moves, etc.) as well as the system itself, so we do not attempt to generalise any conclusions from this.

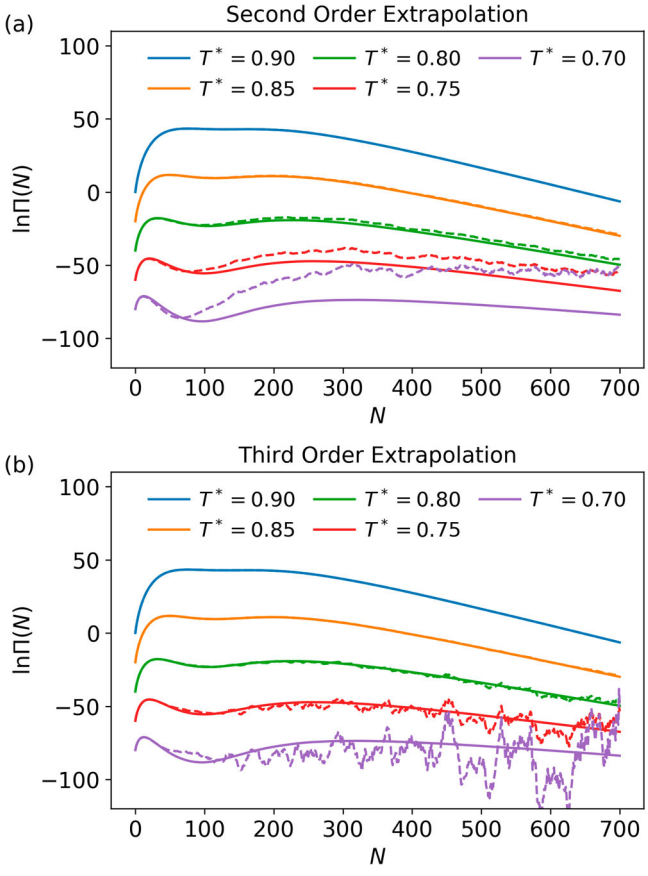

Figure 2. (Colour online) Comparison between direct simulations (solid lines) and extrapolation from $T^* = 0.90$ to $T^* = 0.70$ (dashed lines) in increments of 0.05, from top to bottom, for a model of argon on solidified carbon dioxide described in [59]. Results are reported at the chemical potentials corresponding to the prewetting transition. (a) Extrapolation using up to second-order terms in the Taylor series. (b) Extrapolation using up to third-order terms in the Taylor series [25].

We emphasise, however, that this extrapolation was performed using the original data that was deemed accurate enough to compute properties at a single T^*; that is, a separate flat-histogram simulation was performed at each different T^* in the original work [59]. The extrapolation we have performed in [25] and reproduced here was done by simply retrieving the log files from these previous simulations, constructing the appropriate averages, then expanding the macrostate distribution in a Taylor series at each N. Thus, extrapolation can be used to amplify the data of existing simulations without the need to add code or perform the simulations again. This re-use illustrates data amplification without the need for any additional effort beyond post-processing, making it a useful tool to extract much more information from existing simulation results.

5. Multicomponent mixtures

Extrapolation in temperature conveniently enables the computation of phase diagrams for pure component systems. This can also be performed for multicomponent systems [25]; however, binodals for multicomponent mixtures are determined by the equality of temperature, pressure, and chemical potentials of all components in each phase; therefore, to locate the phase envelope we must extrapolate in terms of the chemical potentials for which we do not have an explicit reweighting equation for ($\Delta\vec{\mu}$ values in $\vec{\phi}$).

A typical application of the extrapolation procedure for a binary system would be as follows. First, a few

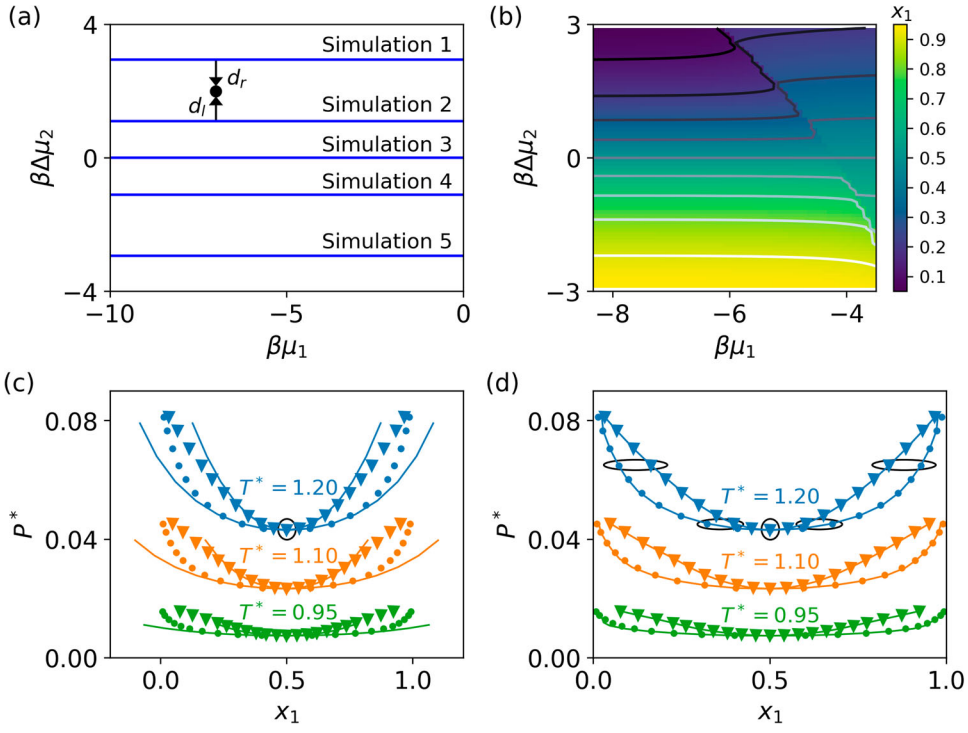

Figure 3. (Colour online) Thermodynamic extrapolation of multicomponent mixture properties. (a) Qualitative construction of a ($\beta\mu_1$, $\beta\Delta\mu_2$)-grid from a discrete set of five different binary simulations. (b) Mole fraction of species 1, x_1, in the most stable phase for the binary square-well mixture described in [27] over a grid as in (a). (c) Symbols represent coexistence points for the binary square-well mixture obtained from direct simulation, whereas lines denote the predictions made by second-order extrapolation of the simulation performed at $T^* = 1.20$, $\Delta\mu_2 = 0.00$ (circled in black). (d) The same predictions now using a combination of simulations at $\Delta\mu_2 \in (\pm 3.54, \pm 1.01, 0.0)$ and $T^* = 1.20$ (circled in black) [27].

simulations are performed at a given temperature and different values of $\Delta\mu_2$. The order parameter used is simply $\vec{\Psi} = N_{\text{tot}} = N_1 + N_2$. Note that following Equation (13) the value of μ_1 used to perform the simulations is irrelevant, and $\ln\Pi(N_{\text{tot}})$ can simply be recalculated at essentially any other value desired *via* histogram reweighting. Figure 3(a) illustrates this concept; because $\ln\Pi(N_{\text{tot}})$ can be reweighted along the abscissa (μ_1), we can compute all properties along a given blue line. However, each simulation is discrete along the ordinate ($\Delta\mu_2$). To obtain a numerical equation of state, we would like to have $\ln\Pi(N_{\text{tot}}; \mu_1, \Delta\mu_2, \beta)$; by extrapolating in $\Delta\mu_2$ and β we can achieve this. Each simulation at a different $\Delta\mu_2$ can be extrapolated most accurately in its neighbourhood, bounded by its nearest neighbour simulation. To predict the properties at a value of $\Delta\mu_2$ between two different simulations, each can be extrapolated to that point, and their individual results combined in a linear fashion with a different weight assigned to each. The weight given to each extrapolated macrostate distribution can be computed by optimisation to satisfy the Gibbs–Duhem equation for thermodynamic consistency [27], or simply approximated as a function of the distance each contributing simulation had to be extrapolated. As the spacing between $\Delta\mu_2$ values of the simulations is reduced, this weight becomes progressively less relevant as extrapolations from both neighbours predict nearly the same values even if one is extrapolated further than the other. For the example in Figure 3(b) which used five different simulations, these weights were found to be of the order of unity and the results were relatively insensitive to them.

Thus, when properties of the system at a certain condition (μ_1, $\Delta\mu_2$, β) are desired, this reweighting plus extrapolation procedure can be used to combine a small number of discrete simulations to make predictions over a broad range of conditions. To illustrate the representative accuracy of this procedure for simple fluids, we consider a binary square-well mixture that forms an azeotrope in Figure 3(c,d) [27]. Here we have extrapolated a single simulation performed at a reduced temperature of $T^* = 1.20$ and $\Delta\mu_2 = 0$ to different temperatures and chemical potentials. From these extrapolated $\ln\Pi(N_{\text{tot}})$ arrays, we can compute the binodal for this system. When extrapolating only this single simulation [cf. Figure 3(c)], we are able to get reasonable predictions of the binodal for mole fractions near $x_1 = 0.5$ (corresponding to where $\Delta\mu_2 = 0$) across the temperatures of interest ($T^* \in [1.20, 1.10, 0.95]$); however, the predictions (solid lines) quickly diverge from the true answer when $x_1 \lesssim 0.4$ or $x_1 \gtrsim 0.6$. By instead combining the results of five different simulations over a range of different $\Delta\mu_2$ values at the same original temperature, we can achieve quantitatively accurate predictions across the entire range of mole fractions [cf. Figure 3(d)].

Similar quality results can be obtained for Lennard-Jones systems as well [27]. While the choice of $\Delta\mu_2$ spacing may vary, this suggests that only a small handful of different simulations are necessary to predict thermodynamic properties of a binary mixture across a wide range of conditions, even for non-trivial systems such as azeotropes. For systems with more components, a similar method can be used to combine extrapolations from a grid of different simulations at different $\Delta\mu_i$.

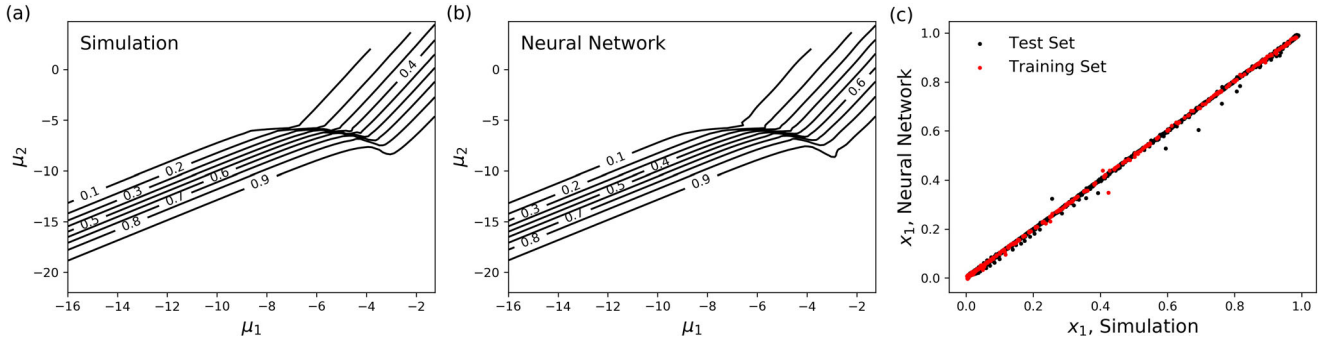

Figure 4. (Colour online) Isopleths (constant x_1) for the binary Lennard-Jones fluid described in [27]. (a) Isopleths in chemical potential space at a reduced temperature of $T^* = 1.3$ created by extrapolating and combining five different simulations at various $\Delta\vec{\mu}$. (b) Predictions from an example neural network composed of 4 dense, fully connected layers of 15 nodes using the hyperbolic tangent activation function; 15,000 data points were collected spanning 5 temperatures with different combinations of chemical potentials, as in (a); only 20 % of these data were used to train the network. (c) Predicted mole fractions from the neural network compared to those computed directly from simulation. The root mean squared error (RMSE) on the training set was RMSE $\approx 6.3 \times 10^{-3}$, while for the test set RMSE $\approx 6.4 \times 10^{-3}$.

Moreover, even if one decides to only trust the extrapolation locally around each simulation and a combination procedure is not used, a large amount of data is still produced; although it contains gaps in $\Delta\mu_i$ or β, these can be used in machine learning models such as neural networks to generate accurate, predictive models.

For example, in Figure 4, we present a simple 4-layer, fully connected neural network that has been fit to data obtained by combining extrapolations for this binary Lennard-Jones fluid [27]. Clearly even this relatively simple network can produce reasonable estimators of the original data. Deeper networks, more advanced layers, and other hyperparameter optimisation *via* cross-validation can produce even better results [7, 18]. However, this illustrates how data from five simulations can be amplified *via* extrapolation to feed data-intensive machine learning algorithms which can be used to explore the properties of these fluids even further.

6. Structural properties

As previously mentioned, the concept of thermodynamic extrapolation may also be extended to enable the extrapolation of a system's equilibrium structural properties, i.e. the average spatial arrangement of atoms or molecules which constitute a system. This includes properties such as radial distribution functions, cluster size distributions, or radii of gyration for polymeric systems. Recall that Equation (14) describes how to average a property, $\tilde{X}(N_{\text{tot}})$, measured at each macrostate (in this case, N_{tot}) to yield an ensemble average at the conditions simulated, while Equation (16) describes how to extrapolate that result to other conditions. As discussed in [29], even if this is not a conventional thermodynamic variable such as particle number, it will not affect the equations defining the derivatives in the Taylor series. The only caveat is that the desired property will end up multiplying the moments matrix which defines terms in the series. So we need to measure a new matrix [27]:

$$Z'(N_{\text{tot}}; \vec{\xi}) = Z(N_{\text{tot}}; \vec{\xi}) \times \Gamma(N_{\text{tot}}), \tag{20}$$

where $\Gamma(N_{\text{tot}})$ is the structural property of interest.

Consider the case of the radial distribution function for a canonical system (fixed N_{tot}, V, β), which is typically computed by collecting a histogram, $h_{i,j}(r; \beta)$, of the pairwise distances, r, between particles of types i and j over the course of a simulation. In general,

$$g_{i,j}(r_k; \beta) = \frac{h_{i,j}(r_k; \beta)}{N_c\left[\left(\frac{N_j - \delta_{i,j}}{V}\right)V_{\text{bin}}(r_k)\right]N_i}, \tag{21}$$

where r_k is the centre of a given histogram bin, $\delta_{i,j}$ is the Kroenecker delta, V_{bin} is the volume of a given bin, and N_c is the number of configurations collected over the course of the simulation. For a single component system which we will consider here, $i = j$ which simplifies the notation. Instead of performing a new simulation to measure $g(r_k; \beta)$ at a new temperature, we can rewrite the radial distribution function as a Taylor series to predict it from one previously measured:

$$g(r_k; \beta) = \frac{h(r_k; \beta_0) + \delta\beta\frac{\partial h(r_k; \beta_0)}{\partial\beta} + \frac{1}{2!}(\delta\beta)^2\frac{\partial^2 h(r_k; \beta_0)}{\partial\beta^2} + \cdots}{N_c\left[\left(\frac{N-1}{V}\right)V_{\text{bin}}(r_k)\right]N}. \tag{22}$$

Thus, to extrapolate $g(r)$, we simply set $\Gamma = h(r_k)$ and measure a $Z'(N; \vec{\xi})$ matrix for each bin. In the end, we will extrapolate the value of each individual bin, then renormalise $g(r)$ to satisfy the known relationship [63]:

$$\frac{\langle N^2\rangle - \langle N\rangle^2}{\langle N\rangle} - 1 = \lim_{R\to\infty}\left(\frac{\langle N\rangle}{V}\int_0^R [\langle g(r)\rangle - 1]4\pi r^2 \mathrm{d}r\right). \tag{23}$$

The average N and N^2 values can be computed by extrapolating the thermodynamic properties as previously described. Thus, the extrapolated structural and thermodynamic properties will yield consistent results. A similar procedure may be followed for multicomponent radial distribution functions as well with a few more bookkeeping steps [29].

However, at sufficiently low temperatures, systems often undergo phase separation. When this occurs, only the states contributing to a given phase should be averaged to yield a prediction of, for example, the radial distribution function (cf. Equation (14)). Taking this into account allows us to extrapolate and reweight the system to locate coexistence conditions

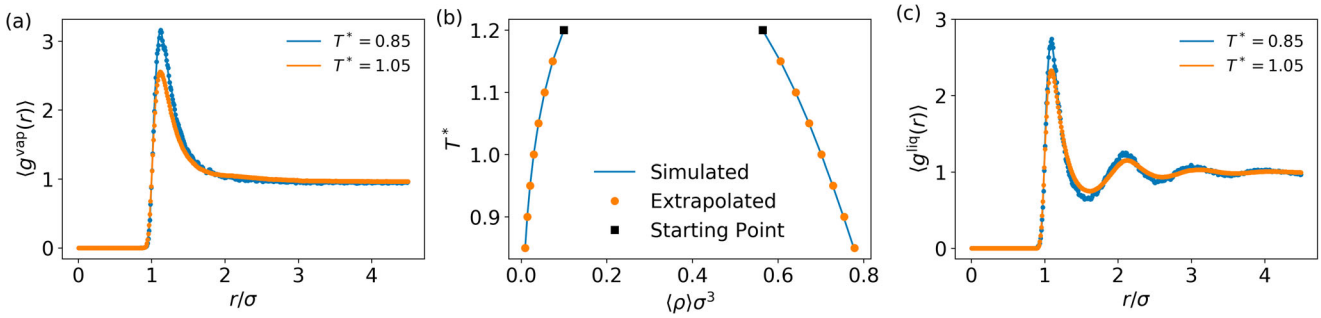

Figure 5. (Colour online) Second-order extrapolation of a pure Lennard-Jones fluid's properties from $T^* = 1.20$ (points) compared to direct simulations (solid lines) [29]. (a) The radial distribution function of the vapour phase at coexistence for two representative temperatures. (b) Vapour–liquid binodal computed from direct simulations compared to extrapolation of the simulation at $T^* = 1.20$ (black squares). (c) The radial distribution function of the liquid phase coexisting with (a).

and then examine the structural properties of each coexisting phase. Figure 5 illustrates some examples of grand canonical average radial distribution functions for a pure Lennard-Jones fluid with long-range corrections [29] along its vapour–liquid coexistence curve. The distribution was originally measured at a reduced temperature of $T^* = 1.20$ and extrapolated using up to second-order terms to lower temperatures. Extrapolation provides results nearly indistinguishable from direct simulations even down to $T^* = 0.85$, approaching the triple point temperature of $T_t \approx 0.70$ for this model [64].

In addition, we can also measure the cluster size distribution, defined as

$$N(s) = \left[\frac{P(s) \times s}{\sum_s P(s) \times s}\right]\rho, \tag{24}$$

where $P(s)$ is the probability of observing a cluster containing s particles when the system's density is ρ. More details on how clusters are defined are available in [29]. We collected this for the same Lennard-Jones fluid and extrapolated (second order) along the coexistence curves to examine the differences between the liquid and vapour phases. Figure 6 shows the results. Deviations increase with $\delta\beta$ and are most notable in the liquid phase; however, even there this manifests as apparently random noise oscillating around the correct mean value. The vapour phase extrapolations are remarkably accurate, even predicting the essential absence of certain cluster sizes at low enough temperatures associated with the nucleation behaviour of the fluid as the triple point temperature is approached [65].

While we have highlighted calculations along the binodal for this system, this extrapolation approach enables us to predict properties such as $g(r)$ and $N(s)$ at any temperature or set of chemical potentials. While accuracy is always a concern, these results suggest that the range over which these extrapolations can be expected to be reasonable is quite large. Thus, we can make predictions across a wide and continuous range of state points, rather than just a single, discrete point which is yielded by a direct simulation or other conventional approaches. Consequently, a broad range of conditions can be explored with only a small set of (sometimes even a single) simulations significantly amplifying the amount of reliable data that can be extracted.

7. Extensions

So far we have reviewed the use and performance of the extrapolation of macrostate distributions obtained by flat histogram Monte Carlo simulations of relatively simple fluids. However, the underlying principles can be extended to more complex systems and applied to other approaches as well. Next, we review several examples of these extensions.

7.1. Internal degrees of freedom and Rosenbluth sampling

While Monte Carlo insertion of rigid molecules simply requires a random sampling of their centre of mass and orientation, the same is not sufficient for the simulation of flexible molecules (*i.e.*, containing intramolecular potentials) which must typically be 'grown' in a biased manner, such as with Rosenbluth sampling, to be computationally efficient [66]. Extending the temperature extrapolation of macrostate probabilities to systems that contain molecules with intramolecular degrees of freedom requires us to consider an additional temperature-dependent contribution within the reference state chemical potential [30]. More details on why this can be neglected for simple, monatomic fluids can be found in [26].

The reference state chemical potential of an ideal gas of rigid particles, $\beta\mu^o_{IG} = -\ln q(\beta)$, only contains a term corresponding to the integration over all kinetic degrees of freedom, $q(\beta)$. Now consider a chain molecule with intramolecular potentials. Even in the ideal gas limit, the chain will have some intramolecular configurational energy, and thus integrating over all degrees of freedom results in an additional term for the ideal chain (IC) reference state chemical potential,

$$\beta\mu^o_{\mathrm{IC}}(\beta) = -\ln q(\beta) - \ln Q_{\mathrm{IC,c}}(\beta). \tag{25}$$

Here $Q_{\mathrm{IC,c}}(\beta)$ is the configurational partition function of an ideal chain with bonded degrees of freedom, $\vec{\theta}$, and intramolecular *bonded* energy, U^{int},

$$Q_{\mathrm{IC,c}}(\beta) = \int \mathrm{d}\vec{\theta} \exp\left[-\beta U^{\mathrm{int}}(\vec{\theta})\right]. \tag{26}$$

This reservoir of ideal chains constitutes the reference state for Rosenbluth sampling [1]. The chemical potential of a real gas can finally be written as $\beta\mu = \beta\mu^o + \ln(\beta\phi P)$, where ϕ is the fugacity coefficient and P is the gas pressure. In practice,

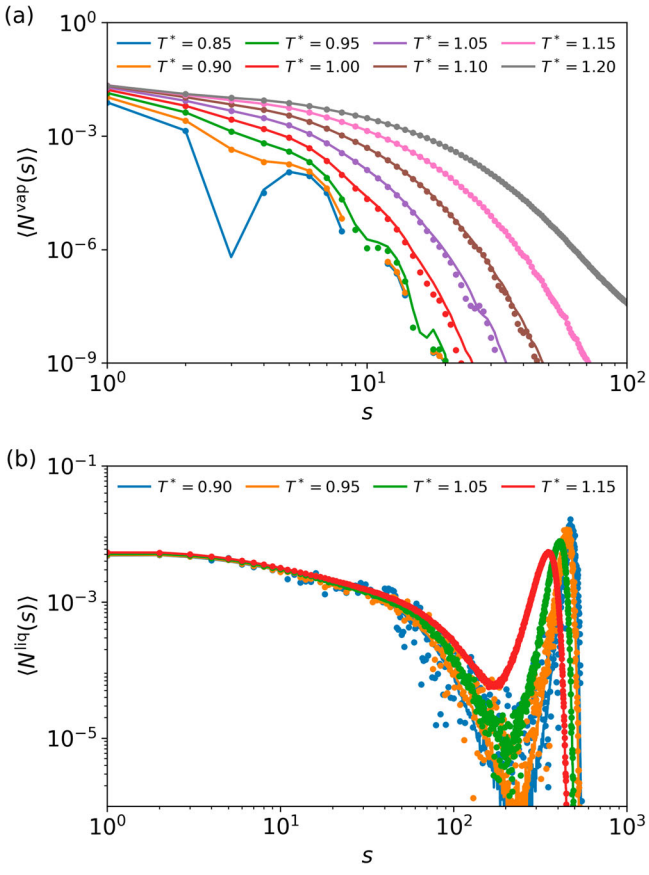

Figure 6. (Colour online) Extrapolated grand canonical cluster size distributions, $\langle N(s)\rangle$, along the binodal for the pure Lennard-Jones fluid [29]. Solid lines correspond to $\langle N(s)\rangle$ obtained from direct simulations; points correspond to second-order extrapolations from $T^* = 1.20$. (a) $\langle N(s)\rangle$ for the coexisting vapour phase. (b) $\langle N(s)\rangle$ for the coexisting liquid phase.

the fugacity coefficient can be found either from a known equation of state or from direct simulation of the bulk fluid.

To simplify notation consider a single component system of chain molecules, $N_{tot} = N_1 = N$. We have already shown how to extrapolate the macrostate distributions to a new temperature using Equations (17) and (18) while holding μ constant. However, the reference state for μ is a function of temperature and so if we extrapolate in β leaving the numerical value of μ constant, then P must change to compensate. For simulated adsorption or phase equilibria corresponding to real systems, P, not μ, is controlled. If we are interested in a transition from $\mu \to \mu'$ and $\beta \to \beta'$, we can use the temperature extrapolation equations developed thus far, followed by Equation (13), to compute the new macrostate probability at μ' *via*

$$\ln \Pi(N;\mu',\beta') = \ln \Pi(N;\mu,\beta') + (\beta'\mu' - \beta\mu)N$$
$$= \ln \Pi(N;\mu,\beta') + N \ln\left[\frac{q(\beta)}{q(\beta')}\frac{Q_{IC,c}(\beta)}{Q_{IC,c}(\beta')}\frac{\beta'\phi'}{\beta\phi}\right]. \quad (27)$$

Thus, we can extrapolate the macrostate probabilities involving chain molecules by accounting for the temperature dependence of the reference state. Practically, $Q_{IC,c}(\beta)$ can be directly integrated for sufficiently small chain molecules or when $|\vec{\theta}|$ (the stiff degrees of freedom) is small. Otherwise, the ratio of $Q_{IC,c}(\beta)/Q_{IC,c}(\beta')$ can be obtained from a simulation of a single, isolated chain [30].

Using this methodology, flat histogram simulations have been performed for both rigid and chain molecules in an adsorption system of practical interest, MOF-950 [30]. Subsequent temperature extrapolation was used to vastly increase the amount of thermodynamic information obtained. For all three adsorbates studied (rigid CO_2, CH_4, and flexible C_3H_8), the isotherms computed at a given temperature were extrapolated over a total temperature range of 100 K. Figure 7 shows how temperature extrapolation can accurately reproduce the behaviour of the flexible propane adsorbate over a range of at least 100 K using data collected from only a single simulation.

The isotherm extrapolated from simulation data at 400 K predicts the isotherms at 350 K and 450 K exactly. The closed circles of Figure 7 show the propane loadings predicted by grand canonical Monte Carlo (GCMC) with Rosenbluth sampling. Each discrete T and P point requires a separate GCMC simulation, whereas the isotherms predicted by flat-histogram extrapolation can be made continuous in T and P space by simple post-processing of the original simulation data.

The power of a flat-histogram technique lies in its ability to generate significantly more thermodynamic data than would otherwise be obtained from standard GCMC simulations. For example, suppose one wishes to optimise temperature/pressure swing operating conditions for an adsorbent bed in a chemical separation process. To model the adsorption properties at different temperatures requires either brute force grand canonical Monte Carlo (GCMC) simulations over the entire possible temperature operating range (high computational cost) or fitting the simulation data to an analytical adsorption model often containing simplifying assumptions (potential loss of fidelity). Temperature extrapolated flat histogram simulations avoid both of these concerns. The computational cost of post-processing simulation data to perform temperature extrapolation, which can generate macrostate probabilities on an arbitrarily fine temperature grid, is negligible compared to running more GCMC simulations. And in the case of complex,

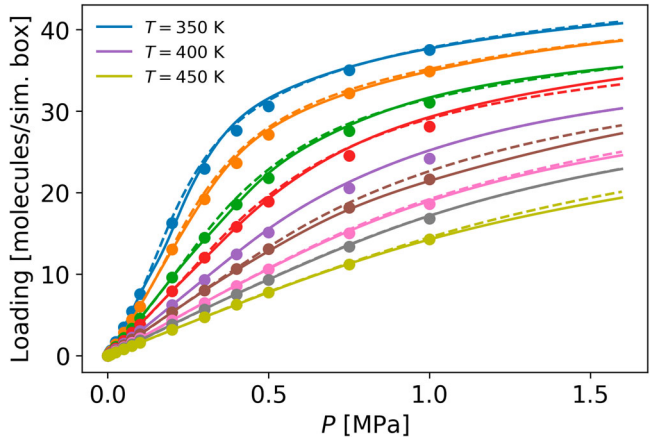

Figure 7. (Colour online) Adsorption isotherms of propane in MOF-950 as reported in [30]. Filled circles were obtained from standard GCMC simulations, colour-coded by the simulation temperature T[K] $\in$ (350, 360, 375, 385, 400, 410, 425, 435, 450). Dashed isotherms correspond to flat-histogram simulations calculated at a fixed temperature, whereas solid isotherms correspond to the extrapolation of the 400 K isotherm data.

non-Langmuir adsorption behaviour, such as the 350 K propane isotherm in Figure 7, the accuracy of the temperature extrapolation does not require selecting, deriving, or fitting a sufficiently complex adsorption model. Finally, we note that, in addition to flexible chain molecules, such extrapolation techniques could be applied to flat-histogram simulations of fully flexible adsorbents [67]. Extrapolation of fully atomistic adsorption simulations of flexible molecules in flexible adsorbents will have the power to greatly increase the amount of information obtained from simulations that are very expensive to execute in the first place.

7.2. Virial coefficients and Mayer sampling

Virial coefficients are of great interest for both fundamental and practical reasons. These coefficients can directly relate the interactions between particles to an experimental bulk measurement with very few assumptions [69]. For models with short-range interactions, the extended law of corresponding states [70] has shown how the second virial coefficient can be used to compare different models to each other or to compare models to experiments [71, 72]. In practice, virial coefficients are used to inform industrial processes, separations, and determine the stability of colloidal suspensions [73–76]. They have also been used to measure the aggregation propensity of protein solutions [77–79]. Thus, virial coefficient data obtained by computational methods over a large range of conditions [28, 80] has the potential to greatly reduce the number of experiments required for optimising chemical processes and performing pharmaceutical research.

While lower order virial coefficients of relatively simple models may be computed routinely [69], higher order coefficients and complicated models require specialised techniques [81]. For example, Mayer-sampling Monte Carlo (MSMC) [82] is a highly effective method for computing virial coefficients of complex models (e.g. proteins) [83, 84]. MSMC utilises free-energy perturbation and importance sampling methods such as umbrella-sampling to bias a Monte Carlo simulation toward configurations which contribute to the ensemble average of interest. Note that since histograms are one way to compute an importance sampling bias over a range of order parameter values [85], extrapolation principles [25, 34–37, 56, 57] may be extended to importance sampling Monte Carlo simulations.

Here we briefly review the extrapolation equations for MSMC simulations [82]; a more detailed derivation may be found in [28]. The MSMC [82] formula to compute the kth order virial coefficient, B_k, *via* biased sampling is

$$B_k^* = \frac{B_k}{\hat{B}_k} = \frac{\left\langle \frac{\gamma_k}{\pi} \right\rangle}{\left\langle \frac{\hat{\gamma}_k}{\pi} \right\rangle} = \frac{\int \mathrm{d}\mathbf{r}\gamma_k}{\int \mathrm{d}\mathbf{r}\hat{\gamma}_k}, \tag{28}$$

where the 'hat' (e.g. $\hat{B}_k$) refers to a reference potential with a known virial coefficient, $\langle M \rangle = \frac{\left(\int \mathrm{d}\mathbf{r} M\pi\right)}{\int \mathrm{d}\mathbf{r}\pi}$, π is the chosen sampling distribution, and $\int \mathrm{d}\mathbf{r}$ is an integration over the pairwise translational and orientational coordinates. For the second and third virial coefficients, k=2 and 3, respectively, $\gamma_2 = f_{12}$ and $\gamma_3 = f_{12}f_{13}f_{23}$, $f_{ij} = \exp(-\beta U_{ij}) - 1$, where U_{ij} is the pairwise intermolecular potential. Typically $\pi = |\gamma_k|$ is found to be effective but the following derivation does not depend upon this choice. In order to perform extrapolations with respect to some arbitrary variable, η, we now turn our attention to taking derivatives of Equation (28). Using the quotient rule and Leibniz integral rule, we obtain

$$\left\langle \frac{\hat{\gamma}_k}{\pi} \right\rangle^2 \frac{\partial B_k^*}{\partial \eta} = \left\langle \frac{\partial \gamma_k}{\partial \eta} \frac{1}{\pi} \right\rangle \left\langle \frac{\hat{\gamma}_k}{\pi} \right\rangle - \left\langle \frac{\gamma_k}{\pi} \right\rangle \left\langle \frac{\partial \hat{\gamma}_k}{\partial \eta} \frac{1}{\pi} \right\rangle. \tag{29}$$

If the reference potential is a function of η, i.e. $\hat{\gamma}_k = \hat{\gamma}_k(\eta)$, then higher order derivatives are obtained *via* the general Leibniz (product) rule. This equation may be greatly simplified by choosing a reference system, $\hat{\gamma}_k \neq \hat{\gamma}_k(\eta)$ (e.g. a hard sphere), to obtain the following equation for the nth order derivative:

$$\left\langle \frac{\hat{\gamma}_k}{\pi} \right\rangle \frac{\partial^n B_k^*}{\partial \eta^n} = \left\langle \frac{\partial^n \gamma_k}{\partial \eta^n} \frac{1}{\pi} \right\rangle. \tag{30}$$

The only remaining quantities we need to evaluate are $\frac{\partial^n f_{ij}}{\partial \eta^n}$, which depend upon what variable(s) we are extrapolating, e.g. temperature or model parameters in the case of alchemical transformations. The derivatives of the virial coefficients computed at the simulation conditions may then be used to extrapolate to other conditions that were not simulated by using a Taylor series or Padé approximants.

One demonstration of the ability of extrapolation to provide large amounts of quality data at reduced computational cost is the MSMC simulation of the SPC/E water model [28]. In this example, a single simulation using the extrapolation technique yields more data than hundreds of traditional simulations. As shown in Figure 8, a single MSMC simulation accurately extrapolates the second virial coefficient over the temperatures range of 250 K to 10^4 K. The same simulation used for temperature extrapolation was also used to accurately extrapolate the point charge of the model from 40% higher charge down to zero charge, as shown in Figure 9. This is possible because derivatives with respect to multiple variables may be obtained simultaneously from the same simulation. One may also compute mixed derivatives to perform multivariate extrapolation. The accuracy of the extrapolations may also be improved with Padé approximants [28].

Extrapolation with respect to model or forcefield parameters, also referred to as an alchemical transformation, is a promising way to improve and screen computational models. In the previous example, the charges of the SPC/E point charge water model were scaled by a constant. The development of coarse-grained models is another area where these extrapolation methods can be of great utility [78, 84, 86, 87]. For example, the derivatives of the second osmotic virial coefficient with respect to the model variables may be computed and used to improve the model. This may lead to a significant improvement in fitting the model to experimental data while reducing the computational expense and number of simulations required. Indeed, some minimisation and machine learning algorithms may make explicit use of derivatives obtained by the equations above to further improve efficiency.

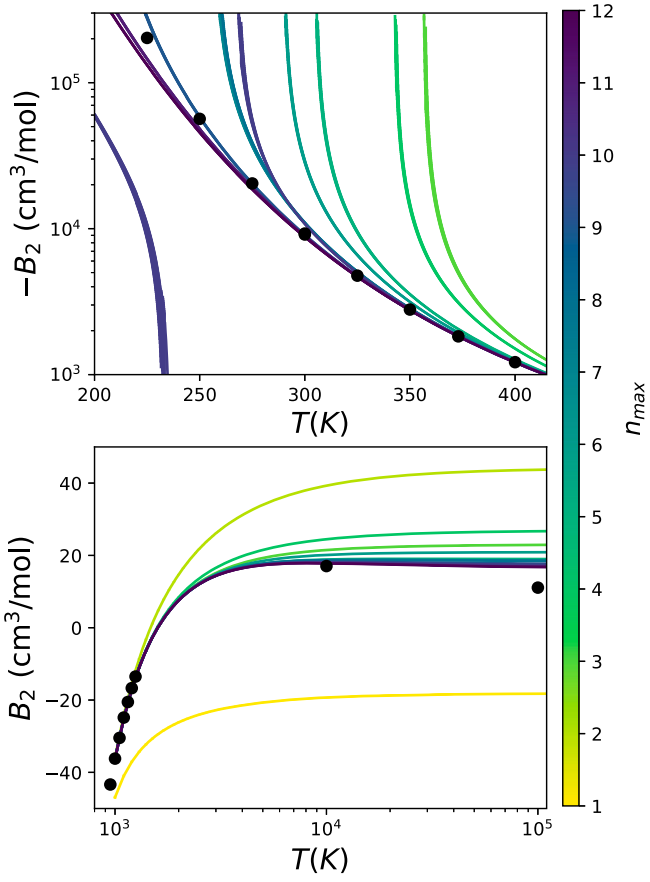

Figure 8. (Colour online) The second virial coefficient of the SPC/E water model obtained *via* MSMC (black circles) and extrapolations from $T = 800$ K (lines) with an $[n_{max} - 1/1]$ Padé approximant. The maximum order derivative, n_{max}, is shown in colour. Reproduced from [68].

While the increasing power and speed of computers allow researchers to perform an increasing number of independent simulations over a range of conditions, fluctuation formulas like those presented herein allow us to connect the dots more efficiently. The computational cost to obtain derivatives of ensemble averages of interest, in addition to the ensemble average themselves, is often negligible. Thus, statistical mechanical simulations may provide a significant amount of data on variations in parameters that are not often directly sought after.

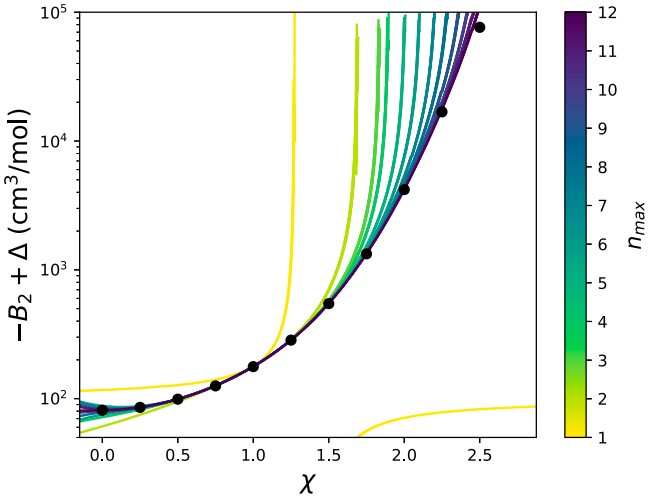

Figure 9. (Colour online) The second virial coefficient of the SPC/E water model with units of cm^3/mol at T=800 K obtained from MSMC (black circles) and extrapolation with respect to site charges, $q_i \rightarrow \sqrt{\chi} q_i$ from a single simulation at $\chi = 1$ (lines). The results are shifted by a parameter, $\Delta = 100$ cm^3/mol, so results can be plotted on a log scale. Reproduced from [68].

8. Conclusions

Extrapolating flat-histogram simulations enables the accurate prediction of both structural and thermodynamic properties of many systems across a broad range of conditions. This approach simply requires the collection of moments of extensive, observable quantities (such as energy) in simulations in order to extrapolate in their intensive thermodynamic conjugates (such as temperature). As a result, these moments can be either constructed during a simulation or *via* post-processing after the simulation has completed. Thus, this technique can be used to amplify the amount of data extracted from not only new simulations, but can be used to re-analyse the log files of previously performed ones. Extrapolation employs a series expansion to approximate a property of interest, which fundamentally implies that we may obtain that property as a continuous function of variables such as temperature or chemical potential; however, it also enables one to quickly obtain a large amount of discrete data points that can be regressed using data-intensive techniques such as neural networks. As the fourth paradigm becomes progressively more mainstream, it is important to have computational methodologies in place that can not only produce highly accurate estimates of physical properties, but also a large number of predictions in order to exploit these emerging techniques. The extrapolation approaches we have briefly reviewed here are capable of both, and are expected to be a useful tool in many aspects of computational chemical analysis and design including high throughput screening, force-field parameterisation, and process optimisation.

Acknowledgments

Contribution of the National Institute of Standards and Technology (NIST), not subject to U.S. Copyright. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Nathan A. Mahynski http://orcid.org/0000-0002-0008-8749
Harold W. Hatch http://orcid.org/0000-0003-2926-9145
Matthew Witman http://orcid.org/0000-0001-6263-5114
David A. Sheen http://orcid.org/0000-0003-1958-1848
Jeffrey R. Errington http://orcid.org/0000-0003-0365-0271

References

[1] Frenkel D, Smit B. Understanding molecular simulation: from algorithms to applications. 2nd ed. San Diego: Academic Press Inc.; 2001. (Computational science series; 1).

[2] de Pablo JJ, Escobedo FA. Molecular simulations in chemical engineering: present and future. AIChE J. 2002;48:2716–2721.

[3] Krylov A, Windus TL, Barnes T, et al. Perspective: computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science. J Chem Phys. 2018;149(18):180901.

[4] Winsberg E. Computer simulations in science. In: Zalta EN, editor. The stanford encyclopedia of philosophy (Winter edition); 2019. https://plato.stanford.edu/archives/win2019/entries/simulations-science/.

[5] Bell G, Hey T, Szalay A. Beyond the data deluge. Science. 2009;5919:1297–1298.

[6] Hey T, Tansley S, Tolle K, editors. The fourth paradigm: data-intensive scientific discovery. Vol. 1. Redmond (WA): Microsoft Research; 2009.

[7] Hastie Tz, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. Vol. 1. Springer; 2009.

[8] Jadrich RB, Lindquist BA, Truskett TM. Recent advances in accelerated discovery through machine learning and statistical inference; 2017. arXiv preprint arXiv:1706.05405.

[9] Williams CKI, Rasmussen CE. Gaussian processes for machine learning. Vol. 2. Cambridge (MA): MIT Press; 2006.

[10] Coifman RR, Lafon S. Diffusion maps. Appl Comput Harmon Anal. 2006;21:5–30.

[11] Jolliffe I. Principal component analysis. Berlin: Springer Berlin Heidelberg; 2011; p. 1094–1096.

[12] Buchanan M. The power of machine learning. Nat Phys. 2019;15:1208.

[13] Desgranges C, Delhommelle J. A new approach for the prediction of partition functions using machine learning techniques. J Chem Phys. 2018;149:044118.

[14] Groven SD, Desgranges C, Delhommelle J. Prediction of the boiling and critical points of polycyclic aromatic hydrocarbons via Wang-Landau simulations and machine learning. Fluid Phase Equilib. 2019;484:225–231.

[15] Desgranges C, Delhommelle J. Determination of mixture properties via a combined expanded Wang-Landau simulations-machine learning approach. Chem Phys Lett. 2019;715:1–6.

[16] Sun Y, DeJaco RF, Siepmann JI. Deep neural network learning of complex binary sorption equilibria from molecular simulation data. Chem Sci. 2019;10:4377–4388.

[17] Gopalan A, Bucior BJ, Bobbitt NS, et al. Prediction of hydrogen adsorption in nanoporous materials from the energy distribution of adsorption sites. Mol Phys. 2019;117(23–24):3683–3694.

[18] Berressem F, Khadilkar MR, Nikoubashman A. Boltzmann: Deriving effective pair potentials and equations of state using neural networks; 2019. arXiv preprint arXiv:1908.02448.

[19] Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J Chem Phys. 2011;134:074106.

[20] Sidky H, Whitmer JK. Learning free energy landscapes using artificial neural networks. J Chem Phys. 2018;148:104111.

[21] Long AW, Ferguson AL. Rational design of patchy colloids via landscape engineering. Mol Syst Design Eng. 2018;3:49–65.

[22] Seo B, Kim S, Lee M, et al. Driving conformational transitions in the feature space of autoencoder neural network. J Phys Chem C. 2018;122(40):23224–23229.

[23] Adorf CS, Moore TC, Melle YJU, et al. Analysis of self-assembly pathways with unsupervised machine learning algorithms. J Phys Chem B. 2020;124:69–78.

[24] Chipot C, Pohorille A. Free energy calculations: theory and applications in chemistry and biology. Berlin: Springer; 2007.

[25] Mahynski NA, Blanco MA, Errington JR, et al. Predicting low-temperature free energy landscapes with flat-histogram Monte Carlo methods. J Chem Phys. 2017;146(7):074101.

[26] Mahynski NA, Errington JR, Shen VK. Temperature extrapolation of multicomponent grand canonical free energy landscapes. J Chem Phys. 2017;147:054105.

[27] Mahynski NA, Errington JR, Shen VK. Multivariable extrapolation of grand canonical free energy landscapes. J Chem Phys. 2017;147:234111.

[28] Hatch HW, Jiao S, Mahynski NA, et al., Communication: predicting virial coefficients and alchemical transformations by extrapolating Mayer-sampling Monte Carlo simulations. J Chem Phys. 2017;147:231102.

[29] Mahynski NA, Jiao S, Hatch HW, et al. Predicting structural properties of fluids by thermodynamic extrapolation. J Chem Phys. 2018;148:194105.

[30] Witman M, Mahynski NA, Smit B. Flat-histogram Monte Carlo as an efficient tool to evaluate adsorption processes involving rigid and deformable molecules. J Chem Theory Comput. 2018;14:6149–6158.

[31] Karpatne A, Atluri G, Faghmous JH, et al. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng. 2017;29:2318–2331.

[32] Zwanzig RW. High temperature equation of state by a perturbation method. I. Nonpolar gases. J Chem Phys. 1954;22:1420–1426.

[33] Escobedo FA. Novel pseudoensembles for simulation of multicomponent phase equilibria. J Chem Phys. 1998;108 (21):8761–8772.

[34] Szalai I, Liszi J, Boda D. The NVT plus test particle method for the determination of the vapour-liquid equilibria of pure fluids. Chem Phys Lett. 1995;246:214–220.

[35] Boda D, Liszi J, Szalai I. An extension of the NpT plus test particle method for the determination of the vapour-liquid equilibria of pure fluids. Chem Phys Lett. 1995;235:140–145.

[36] Boda D, Liszi J, Szalai I. A new simulation method for the determination of the vapour-liquid equilibria in the grand canonical ensemble. Chem Phys Lett. 1996;256:474–482.

[37] Boda D, Kristf T, Liszi J, et al. The extrapolation of the vapour-liquid equilibrium curves of pure fluids in the isothermal Gibbs ensemble. Mol Phys. 2002;100:1989–2000.

[38] Kristóf T, Liszi J, Boda D. The extrapolation of phase equilibrium curves of mixtures in the isobaric-isothermal Gibbs ensemble. Mol Phys. 2002;100(21):3429–3441.

[39] Escobedo FA. Mapping coexistence lines via free-energy extrapolation: application to order-disorder phase transitions of hard-core mixtures. J Chem Phys. 2014;140:094102.

[40] Hummer G, Szabo A. Calculation of free-energy differences from computer simulations of initial and final states. J Chem Phys. 1996;105:2004–2010.

[41] Chapman WG, Gubbins KE, Jackson G, et al. SAFT: equation-of-state solution model for associating fluids. Fluid Phase Equilib. 1989;52:31–38.

[42] Müller EA, Jackson G. Force-field parameters from the SAFT-γ equation of state for use in coarse-grained molecular simulations. Annu Rev Chem Biomol Eng. 2014;5(1):405–427.

[43] Landau DP, Binder K. A guide to Monte Carlo simulations in statistical physics. Cambridge: Cambridge University Press; 2009.

[44] Shell MS, Debenedetti PG, Panagiotopoulos AZ. An improved Monte Carlo method for direct calculation of the density of states. J Chem Phys. 2003;119:9406.

[45] Rane KS, Murali S, Errington JR. Monte Carlo simulation methods for computing liquid–vapor saturation properties of model systems. J Chem Theory Comput. 2013;9(6):2552–2566.

[46] Wang F, Landau DP. Efficient, multiple-range random walk algorithm to calculate density of states. Phys Rev Lett. 2001;86 (10):2050–2053.

[47] Landau DP, Tsai S-H, Exler M. A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling. Am J Phys. 2004;72(10):1294–1302.

[48] Wang J-S, Tay TK, Swendsen RH. Transition matrix Monte Carlo reweighting and dynamics. Phys Rev Lett.. 1999;82(3):476.

[49] Wang F, Landau DP. Determining the density of states for classical statistical models: a random walk algorithm to produce a flat histogram. Phys Rev E. 2001;64:056101.
[50] Shen VK, Errington JR. Determination of fluid-phase behavior using transition-matrix Monte Carlo: binary Lennard-Jones mixtures. J Chem Phys. 2005;122(6):064508.
[51] Shen VK, Errington JR. Determination of surface tension in binary mixtures using transition-matrix Monte Carlo. J Chem Phys. 2006;124(2):024721.
[52] Mittal J, Shen VK, Errington JR, et al. Confinement, entropy, and single-particle dynamics of equilibrium hard-sphere mixtures. J Chem Phys. 2007;127:154513.
[53] Kofke DA, Glandt ED. Nearly monodisperse fluids. I. Monte Carlo simulations of Lennard-Jones particles in a semigrand ensemble. J Chem Phys. 1987;87:4881–4890.
[54] Kofke DA, Glandt ED. Monte Carlo simulation of multicomponent equilibria in a semigrand canonical ensemble. Mol Phys. 1988;64:1105–1131.
[55] Potoff JJ, Panagiotopoulos AZ. Critical point and phase behavior of the pure fluid and a Lennard-Jones mixture. J Chem Phys. 1998;109:10914.
[56] Panagiotopoulos AZ, Reid RC. On the relationship between pairwise fluctuations and thermodynamic derivatives. J Chem Phys. 1986;85(8):4650.
[57] Debenedetti PG. On the relationship between principal fluctuations and stability coefficients in multicomponent systems. J Chem Phys. 1986;84(3):1778.
[58] Ursell HD. The evaluation of Gibbs' phase-integral for imperfect gases. Math Proc Camb Philos Soc. 1927;23(6):685–697.
[59] Errington JR. Prewetting transitions for a model argon on solid carbon dioxide system. Langmuir. 2004;20:3798–3804.
[60] Errington JR. Direct calculation of liquid-vapor phase equilibria from transition matrix Monte Carlo simulation. J Chem Phys. 2003;118(22):9915–9925.
[61] Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. J Chem Phys. 2008;129:124105.
[62] Tan Z, Gallicchio E, Lapelosa M, et al. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. J Chem Phys. 2012;136:144102.
[63] Ben-Naim A. The Kirkwood-Buff integrals for one-component liquids. J Chem Phys. 2008;128:234501.
[64] Mastny EA, de Pablo JJ. Melting line of the Lennard-Jones system, infinite size, and full potential. J Chem Phys. 2007;127:104504.
[65] Yasuoka K, Matsumoto M. Molecular dynamics of homogeneous nucleation in the vapor phase. I. Lennard-Jones fluid. J Chem Phys. 1998;109(19):8451–8462.
[66] Rosenbluth MN, Rosenbluth AW. Monte Carlo calculation of the average extension of molecular chains. J Chem Phys. 1955;23(2):356–359.
[67] Witman M, Wright B, Smit B. Simulating enhanced methane deliverable capacity of guest responsive pores in intrinsically flexible MOFs. J Phys Chem Lett. 2019;10(19):5929–5934.
[68] Hatch HW, Jiao S, Mahynski NA, et al. Communication: predicting virial coefficients and alchemical transformations by extrapolating Mayer-sampling Monte Carlo simulations. J Chem Phys. 2017;147:231102.
[69] Hansen J-P, McDonald IR. Theory of simple liquids. Burlington (MA): Elsevier; 1990.
[70] Noro MG, Frenkel D. Extended corresponding-states behavior for particles with variable range attractions. J Chem Phys. 2000;113:2941–2944.
[71] Foffi G, Sciortino F. On the possibility of extending the Noro-Frenkel generalized law of correspondent states to nonisotropic patchy interactions. J Phys Chem B. 2007;111(33):9702–9705.
[72] Hatch HW, Yang S-Y, Mittal J, et al. Self-assembly of trimer colloids: effect of shape and interaction range. Soft Matter. 2016;12:4170–4179.
[73] Benjamin KM, Singh JK, Schultz AJ, et al. Higher-order virial coefficients of water models. J Phys Chem B. 2007;111:11463–11473.
[74] Shaul KRS, Schultz AJ, Kofke DA, et al. Semiclassical fifth virial coefficients for improved ab initio helium-4 standards. Chem Phys Lett. 2012;531:11–17.
[75] Feng C, Schultz AJ, Chaudhary V, et al. Eighth to sixteenth virial coefficients of the Lennard-Jones model. J Chem Phys. 2015;143:044504.
[76] Krekelberg WP, Mahynski NA, Shen VK. On the virial coefficients of confined fluids: analytic expressions for the thermodynamic properties of hard particles with attractions in slit and cylindrical pores to second order. J Chem Phys. 2019;150:044704.
[77] O'Brien CJ, Blanco MA, Costanzo JA, et al. Modulating non-native aggregation and electrostatic protein–protein interactions with computationally designed single-point mutations. Protein Eng Des Sel. 2016;29:231–243.
[78] Blanco MA, Hatch HW, Curtis JE, et al. Evaluating the effects of hinge flexibility on the solution structure of antibodies at concentrated conditions. J Pharm Sci. 2019;108:1663–1674.
[79] Hung JJ, Dear BJ, Karouta CA, et al. Protein-protein interactions of highly concentrated monoclonal antibody solutions via static light scattering and influence on the viscosity. J Phys Chem B. 2019;123:739–755.
[80] Schultz AJ, Kofke DA. Interpolation of virial coefficients. Mol Phys. 2009;107:1431–1436.
[81] Wheatley RJ. Calculation of high-order virial coefficients with applications to hard and soft spheres. Phys Rev Lett. 2013;110:200601.
[82] Singh JK, Kofke DA. Mayer sampling: calculation of cluster integrals using free-energy perturbation methods. Phys Rev Lett. 2004;92:220601.
[83] Shaul KRS, Schultz AJ, Kofke DA. Mayer-sampling Monte Carlo calculations of uniquely flexible contributions to virial coefficients. J Chem Phys. 2011;135:124101.
[84] Blanco MA, Sahin E, Robinson AS, et al. Coarse-grained model for colloidal protein interactions, B22, and protein cluster formation. J Phys Chem B. 2013;117:16013–16028.
[85] Paluch AS, Shen VK, Errington JR. Comparing the use of Gibbs ensemble and grand-canonical transition-matrix Monte Carlo methods to determine phase equilibria. Ind Eng Chem Res. 2008;47:4533–4541.
[86] Chaudhri A, Zarraga IE, Kamerzell TJ, et al. Coarse-grained modeling of the self-association of therapeutic monoclonal antibodies. J Phys Chem B. 2012;116:8045–8057.
[87] Calero-Rubio C, Saluja A, Roberts CJ. Coarse-grained antibody models for 'weak' protein-protein interactions from low to high concentrations. J Phys Chem B. 2016;120:6592–6605.