Multivariable extrapolation of grand canonical free energy landscapes

Nathan A. Mahynski, Jeffrey R. Errington, and Vincent K. Shen

Citation: The Journal of Chemical Physics **147**, 234111 (2017);
View online: https://doi.org/10.1063/1.5006906
View Table of Contents: http://aip.scitation.org/toc/jcp/147/23
Published by the American Institute of Physics

Multivariable extrapolation of grand canonical free energy landscapes

Nathan A. Mahynski,[1,a)] Jeffrey R. Errington,[2] and Vincent K. Shen[1]
[1]*Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, USA*
[2]*Department of Chemical and Biological Engineering, University at Buffalo, The State University of New York, Buffalo, New York 14260-4200, USA*

(Received 28 September 2017; accepted 1 December 2017; published online 21 December 2017)

We derive an approach for extrapolating the free energy landscape of multicomponent systems in the grand canonical ensemble, obtained from flat-histogram Monte Carlo simulations, from one set of temperature and chemical potentials to another. This is accomplished by expanding the landscape in a Taylor series at each value of the order parameter which defines its macrostate phase space. The coefficients in each Taylor polynomial are known exactly from fluctuation formulas, which may be computed by measuring the appropriate moments of extensive variables that fluctuate in this ensemble. Here we derive the expressions necessary to define these coefficients up to arbitrary order. In principle, this enables a single flat-histogram simulation to provide complete thermodynamic information over a broad range of temperatures and chemical potentials. Using this, we also show how to combine a small number of simulations, each performed at different conditions, in a thermodynamically consistent fashion to accurately compute properties at arbitrary temperatures and chemical potentials. This method may significantly increase the computational efficiency of biased grand canonical Monte Carlo simulations, especially for multicomponent mixtures. Although approximate, this approach is amenable to high-throughput and data-intensive investigations where it is preferable to have a large quantity of reasonably accurate simulation data, rather than a smaller amount with a higher accuracy. https://doi.org/10.1063/1.5006906

I. INTRODUCTION

Grand canonical Monte Carlo (GCMC) simulation is an effective tool for studying first order phase transitions in many fluids.[1–5] In this approach, a system with a fixed volume and temperature is exposed to multiple fictitious thermodynamic reservoirs with which heat and mass may be exchanged. By specifying the temperature and chemical potentials of these reservoirs, respectively, the system of interest reaches equilibrium based on these inputs. Because energy and particle number fluctuate in the system, it is easy to observe when the first order phase separation occurs. A single GCMC simulation normally pertains to only a single set of conditions (i.e., temperature and chemical potentials); however, when combined with histogram reweighting, it is possible to predict the system's properties at other conditions which are not too dissimilar.[3,6,7] A macrostate of a system may be defined as a collection of microstates, or configurations, with the same energy and particle number, although other definitions are also possible.[3] Reweighting works on the principle that when the probability of a certain macrostate is known at one set of external conditions, it can be predicted at another since the relative probability of the macrostate at each condition is known from statistical mechanics.[3,6,7] However, due to the finite length of a conventional GCMC simulation, typically only the most probable macrostates, consistent with the specified chemical potentials and temperature, are sampled.

Because of this limitation, the macrostates which become most probable at chemical potentials and temperatures differing significantly from those of the simulation are sampled infrequently, if at all, due to their extremely low probabilities of being observed. Thus, a conventional simulation provides little information about state points that differ significantly from its own. One way to calculate thermodynamic properties over a more broad range is to perform many simulations across a diverse set of conditions so that they sample different but overlapping portions of the phase space.[8,9] By combining many such simulations, a system's thermodynamic properties can be determined at very dissimilar conditions, at the cost of having to perform many simulations.[3,6,7,10–12] While potentially effective for single-component systems, this can quickly become intractable for multicomponent systems since the volume of the macrostate phase space grows geometrically with the number of components.

Recently, we developed an alternative approach which relies on flat-histogram biasing to explore a predefined set of macrostates determined exclusively by the number of particles present, either individually, or as a scalar sum.[13,14] By measuring the moments of extensive thermodynamic properties at each macrostate, one can expand the local free energy landscape in a Taylor series which can be used to extrapolate the landscape to arbitrary external conditions, e.g., temperature.[13] The logical basis for this approach is as follows. A Taylor series, truncated to a finite order, requires derivatives up to that order with respect to the thermodynamic variable(s) in which the extrapolation is to be performed. It is well known that thermodynamic derivatives may be expressed as

[a)]Electronic mail: nathan.mahynski@nist.gov

fluctuations of thermodynamic variables.[15–18] Since fluctuations may be computed from the appropriately averaged products of such variables, one may obtain the coefficients for the Taylor expansion by simply averaging the correct set of variables (moments) over the course of the simulation.[15,19] This enables the extrapolation of the landscape from one set of conditions to another, even when the relevant macrostates occupy a significantly different region of phase space compared to the original simulation.

The principles behind such an expansion are hardly new, especially in the canonical ensemble,[15,19] and similar concepts have been applied to estimate fluid phase coexistence from grand canonical and Gibbs ensemble Monte Carlo simulations in the past.[20–25] However, these efforts have focused on extrapolating unbiased simulations which only sample the phase space relatively near to the most likely macrostates at a given condition; consequently, these approaches only yield reasonable predictions relatively close to the original simulation's state point. By invoking flat-histogram methods to explore a broad range of predefined macrostates, we anticipate that these principles can be used to make predictions over an even more broad range of conditions. Extrapolations are approximate by nature; their inaccuracies naturally increase with the difference between the state point of interest and the reference point. We anticipate that this method should be less accurate than histogram reweighting approaches if the latter can make use of simulations which sampled all relevant regions of phase space at the initial and final states; however, the method described here does not require this. It simply relies on the straightforward accumulation of moments at the reference state. Thus, we expect this approach to be a helpful tool in high-throughput screening and data-driven investigations where the quantity of reasonably accurate simulation data, rather than an extremely high level of accuracy, is paramount. As we will demonstrate, this method is, indeed, capable of generating very reasonable predictions over a broad range of conditions from a relatively small set of initial simulations.

Previously, we exclusively focused on the temperature extrapolation of grand canonical landscapes to explore the phase behavior and self-assembly of fluids at low temperatures.[13,14] In this work, we generalize the approach, enabling one to extrapolate simultaneously in all the intensive thermodynamic variables which define the grand canonical ensemble for a multicomponent system, i.e., the temperature and chemical potential of each component. These extrapolations can be used to directly calculate thermodynamic properties in many cases or may be refined with additional flat-histogram Monte Carlo simulations. In the latter case, starting from a reasonable estimate of the macrostate distribution often greatly accelerates the convergence of these simulations. Although we focus exclusively on the grand canonical ensemble here, this approach may be applied to other ensembles as well.[13,26] This method often allows the equation of state for a multicomponent fluid to be easily generated from a relatively small number of simulations. We further demonstrate how to extrapolate and combine these simulations in a manner which enforces thermodynamic consistency required by the Gibbs-Duhem equation.[18,27]

This paper is organized as follows. We present the relevant statistical mechanics employed in our multicomponent sampling method in Sec. II. Then, we derive the necessary equations for the multivariable extrapolation which allows one to obtain an estimate of the macrostate probabilities (landscape) at arbitrary conditions given a macrostate distribution measured at other conditions. In Sec. III, we present representative results of this technique for binary fluids and demonstrate a way of combining multiple individual simulations to produce a self-consistent hypersurface defined by the component chemical potentials at a given temperature. We then show how this can be used to compute all relevant thermodynamic properties for a fluid and conclude in Sec. IV.

II. METHODS

A. Statistical mechanical ensemble

Consider a k-component system in the microcanonical ensemble, for which the entropy representation of the fundamental thermodynamic equation is given by

$$\mathrm{d}\left(S/k_\mathrm{B}\right) = \beta \mathrm{d}U + \beta P \mathrm{d}V - \sum_{i=1}^{\mathrm{k}} \beta\mu_i \mathrm{d}N_i, \tag{1}$$

where S, U, V, and N_i refer to the extensive entropy, internal energy, volume, and particle number of each species, i, respectively. Here, $\beta \equiv 1/k_\mathrm{B}T$, where k_B is the Boltzmann constant. The pressure, P, and chemical potentials, μ_i, are multiplied by β in this case. However, in what follows, we will separate them from this factor (which occurs naturally in the energy representation of the fundamental equation) in order to work with independent intensive variables $(\beta, \mu_1, \ldots, \mu_\mathrm{k})$.

Under the appropriate Legendre transforms, we obtain the expression relevant for the grand canonical ensemble, in which V, β, and the set of all chemical potentials, $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_\mathrm{k})$ are fixed,

$$\mathrm{d}\left(\ln \Xi\right) = -U\mathrm{d}\beta + \beta P \mathrm{d}V + \sum_{i=1}^{\mathrm{k}} N_i \mathrm{d}\left(\beta\mu_i\right), \tag{2}$$

where Ξ is the grand partition function. From a classical statistical mechanical perspective, this may be expressed as

$$\Xi\left(\vec{\mu}, V, \beta\right) = \sum_{N_1}\sum_{N_2}\cdots\sum_{N_\mathrm{k}} \exp\left(\beta \sum_{i=1}^{\mathrm{k}} \mu_i N_i\right) Q(\vec{N}, V, \beta), \tag{3}$$

where $Q(\vec{N}, V, \beta)$ is the canonical partition function, henceforth referred to simply as Q. For convenience, we assume that the contributions to Q from the kinetic energy of the system in question have been integrated out and neglected so that we work only with the configurational partition function (configurational integral).[13,14] This has no effect on the extensive thermodynamic properties of the system, except the internal energy. Consequently, and without loss of generality, we henceforth take U as referring to only the potential energy of the system. Furthermore, we emphasize that chemical potentials will not contain any contribution from the de Broglie wavelength as a result of this.

In order to obtain a convenient, scalar order parameter to use for sampling multicomponent fluids, we perform a change of variables to express $\Xi(\vec{\mu}, V, \beta)$ in terms of the total number of molecules of each species present, i.e., $N_{\text{tot}} = \sum_{i=1}^{k} N_i$,

$$\begin{aligned}\Xi(\vec{\mu}, V, \beta) &= \sum_{N_{\text{tot}}} \exp(\beta\mu_1 N_{\text{tot}}) \\ &\quad \times \left[\sum_{N_2} \cdots \sum_{N_k} \exp\left(\beta \sum_{i=2}^{k} \Delta\mu_i N_i\right) Q\right] \\ &= \sum_{N_{\text{tot}}} \exp(\beta\mu_1 N_{\text{tot}})\, \Upsilon(N_{\text{tot}}; \Delta\vec{\mu}, V, \beta), \end{aligned} \tag{4}$$

where $\Delta\mu_i \equiv \mu_i - \mu_1$, $\Delta\vec{\mu}_i = (\Delta\mu_2, \Delta\mu_3, \ldots, \Delta\mu_k)$, and $\Upsilon(N_{\text{tot}}; \Delta\vec{\mu}, V, \beta)$ is the isochoric semigrand partition function which is given by

$$\Upsilon(N_{\text{tot}}; \Delta\vec{\mu}, V, \beta) = \sum_{N_2} \cdots \sum_{N_k} \exp\left(\beta \sum_{i=2}^{k} \Delta\mu_i N_i\right) Q. \tag{5}$$

For this order parameter, a macrostate is defined by the value of N_{tot}; the probability of each macrostate in the grand canonical ensemble may be expressed as

$$\ln \Pi(N_{\text{tot}}) = \beta\mu_1 N_{\text{tot}} + \ln \Upsilon - \ln \Xi. \tag{6}$$

Extensions to other thermodynamic order parameters have been described elsewhere previously and are beyond the scope of this work.[13] However, we point out that if the change of variables we performed in Eq. (4) to obtain N_{tot} had been omitted, and instead we employed a single particle number, e.g., N_1, as our order parameter, it follows that

$$\ln \Pi(N_1) = \beta\mu_1 N_1 + \ln \Upsilon_{N_1} - \ln \Xi. \tag{7}$$

The isochoric semigrand partition function is now defined at each fixed value of N_1,

$$\Upsilon_{N_1}(N_1; \vec{\mu}, V, \beta) = \sum_{N_2} \cdots \sum_{N_k} \exp\left(\beta \sum_{i=2}^{k} \mu_i N_i\right) Q. \tag{8}$$

Because of the structural similarity of Eqs. (6) and (7), in the equations that follow, the reader may substitute N_1 for N_{tot} and μ_i for $\Delta\mu_i$ to obtain the expressions necessary to perform the extrapolation of the macrostate distribution obtained by using N_1 as the sampling order parameter instead of N_{tot}. This is especially relevant for conditions where, e.g., liquid-liquid phase separation occurs.

B. Flat-histogram Monte Carlo simulations

To construct the (logarithm of the) macrostate distribution, $\ln \Pi(N_{\text{tot}})$, we performed flat-histogram Monte Carlo simulations. Specifically, we used a combination of Wang-Landau and Transition Matrix Monte Carlo (WL-TMMC) methods.[26,28–31] Our implementation has previously been described in detail elsewhere,[13,14] so we will not reproduce it here. The specific method used to obtain $\ln \Pi(N_{\text{tot}})$ is ultimately up to the practitioner, and the extrapolation methodology we derive here is independent of this choice. Therefore, we treat $\ln \Pi(N_{\text{tot}})$ as a known quantity.

In this work, we primarily focus on demonstrating the extrapolation method for a binary mixture which forms a negative azeotrope. In this mixture, all particles i and j interact with a square-well potential,

$$u_{i,j}(r) = \begin{cases} \infty, & r < \sigma, \\ -\epsilon_{i,j}, & \sigma_{i,j} \le r < \lambda_{i,j}\sigma_{i,j}, \\ 0, & r \ge \lambda_{i,j}\sigma_{i,j}. \end{cases} \tag{9}$$

TABLE I. Binary square-well parameters used in this work. All energies and lengths pertaining to these fluids are reported in units of $\epsilon_{1,1}$ and $\sigma_{1,1}$.

i	j	$\epsilon_{i,j}$	$\sigma_{i,j}$	$\lambda_{i,j}$
1	1	1.00	1.00	1.50
1	2	1.25	1.00	1.50
2	2	1.00	1.00	1.50

Table I gives the parameters used in this mixture, which result from using the Lorentz-Berthelot (LB) mixing rule for the molecular "diameters," $\sigma_{i,j}$, but not for their cohesive potential energy, $\epsilon_{i,j}$. The latter has been increased 25% over the LB result. All these simulations were performed in a periodic, cubic simulation cell with a volume, $V = 729\,\sigma_{1,1}^3$. Unless otherwise stated, all energy and length scales in this work are non-dimensionalized according to $\epsilon_{1,1}$ and $\sigma_{1,1}$, respectively. For instance, we defined the reduced temperature as $T^* = k_B T/\epsilon_{1,1}$.

We will also demonstrate our approach for a binary, linear force-shifted Lennard-Jones mixture of particles in Sec. III C. The interaction between these particles is given by

$$U_{i,j}(r) = \begin{cases} U_{i,j}^{\text{LJ}}(r) - U_{i,j}^{\text{LJ}}(r_{c,i,j}) \\ \qquad -(r - r_{c,i,j}) \left.\dfrac{dU_{i,j}^{\text{LJ}}}{dr}\right|_{r_{c,i,j}} & r < r_{c,i,j}, \\ 0 & r \ge r_{c,i,j}, \end{cases} \tag{10}$$

where $U^{\text{LJ}}(r)$ is the Lennard-Jones potential,

$$U_{i,j}^{\text{LJ}}(r) = 4\epsilon_{i,j}\left[\left(\frac{\sigma_{i,j}}{r}\right)^{12} - \left(\frac{\sigma_{i,j}}{r}\right)^{6}\right]. \tag{11}$$

For this mixture, we set the cutoff radius for each pair to be $r_{c,i,j} = 3\sigma_{i,j}$; other parameters used are given in Table II. Similar non-dimensionalization applies in this case as well. These simulations were performed in a larger periodic simulation cell with a volume, $V = 10^3\sigma_{1,1}^3$.

C. Taylor series approach

The key to our approach lies in expanding a known macrostate distribution as a Taylor series around the

TABLE II. Binary, linear force-shifted Lennard-Jones parameters used in this work. All energies and lengths pertaining to these fluids are reported in units of $\epsilon_{1,1}$ and $\sigma_{1,1}$.

i	j	$\epsilon_{i,j}$	$\sigma_{i,j}$
1	1	1.00	1.00
1	2	$2/\sqrt{3}$	1.25
2	2	4/3	1.50

reference conditions where the distribution is known. This is done at each individual macrostate. Truncating the Taylor series at finite order yields a Taylor polynomial that can be used to estimate this distribution at some other set of chemical potentials and temperature. The intensive thermodynamic variables which are fixed in the grand canonical ensemble are $\vec{\phi} = (\beta, \mu_1, \mu_2, \ldots, \mu_k)$, or equivalently, $\vec{\phi} = (\beta, \mu_1, \Delta\vec{\mu})$. We can express the multidimensional Taylor polynomial for any function, $g(N_{tot}; \vec{\phi})$, up to second order as

$$g(N_{tot}; \vec{\phi}) = g(N_{tot}; \vec{\phi}_0) + \delta\vec{\phi} \cdot \nabla g(N_{tot}; \vec{\phi}_0) + \frac{1}{2!}\left[\delta\vec{\phi} \cdot \mathbf{H}_g(N_{tot}; \vec{\phi}_0) \cdot \delta\vec{\phi}^T\right], \quad (12)$$

where the superscript "T" denotes the transpose of the vector. Here, $\vec{\phi}_0$ denotes the reference point, and

$$\delta\vec{\phi} = (\delta\beta, \delta\Delta\mu_2, \delta\Delta\mu_3, \ldots, \delta\Delta\mu_k), \quad (13)$$

where $\delta z \equiv z - z_0$, for each intensive variable, z. Note that μ_1 does not appear in $\delta\vec{\phi}$; this is because Eq. (6) enables "exact" histogram reweighting from one μ_1^a to another μ_1^b, when all other conditions are identical,[32]

$$\ln \Pi(N_{tot}; \mu_1^b) = \ln \Pi(N_{tot}; \mu_1^a) + \beta(\mu_1^b - \mu_1^a)N_{tot}. \quad (14)$$

Hence, there is no need to extrapolate in this variable. In Eq. (12), $\mathbf{H}_g(N_{tot}; \vec{\phi}_0)$ represents the symmetric Hessian matrix,

$$\mathbf{H}_g(N_{tot}; \vec{\phi}_0) = \begin{bmatrix} \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\beta^2}\right|_{\phi[\beta]} & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\beta\partial\Delta\mu_2}\right|_{\phi[\beta,\Delta\mu_2]} & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\beta\partial\Delta\mu_3}\right|_{\phi[\beta,\Delta\mu_3]} & \cdots & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\beta\partial\Delta\mu_k}\right|_{\phi[\beta,\Delta\mu_k]} \\ \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_2\partial\beta}\right|_{\phi[\Delta\mu_2,\beta]} & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_2^2}\right|_{\phi[\Delta\mu_2]} & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_2\partial\Delta\mu_3}\right|_{\phi[\Delta\mu_2,\Delta\mu_3]} & \cdots & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_2\partial\Delta\mu_k}\right|_{\phi[\Delta\mu_2,\Delta\mu_k]} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_k\partial\beta}\right|_{\phi[\Delta\mu_k,\beta]} & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_k\partial\Delta\mu_2}\right|_{\phi[\Delta\mu_k,\Delta\mu_2]} & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_k\partial\Delta\mu_3}\right|_{\phi[\Delta\mu_k,\Delta\mu_3]} & \cdots & \left.\frac{\partial^2 g(N_{tot}; \vec{\phi}_0)}{\partial\Delta\mu_k^2}\right|_{\phi[\Delta\mu_k]} \end{bmatrix}. \quad (15)$$

Here and throughout this manuscript, we state partial derivatives with respect to some intensive thermodynamic variable(s) with the implied understanding that all other intensive variables are held constant, in addition to volume. This is shown explicitly here with the subscript, e.g., $\phi[\beta]$, which denotes that all variables in ϕ are held constant except β. Subsequently, we neglect this for notational convenience. It is worth reiterating that we have chosen $\vec{\phi} = (\beta, \mu_1, \Delta\mu_2, \ldots, \Delta\mu_k)$ to reflect the intensive conjugates which more naturally manifest in the energy representation of the fundamental equation, rather than the entropy representation, $\vec{\phi} = (\beta, \beta\mu_1, \beta\Delta\mu_2, \ldots, \beta\Delta\mu_k)$. In the latter case, the fact that chemical potentials are multiplied by β means that this product (the "activity," or differences thereof) must be held constant when taking partial derivatives. In fact, this approach tends to produce simpler expressions for the partial derivatives we will derive next; however, when used to perform an extrapolation with the resulting Taylor polynomial, one must be cautious of the interdependence of these conjugate variables. Without loss of generality, we have elected to separate them for the sake of operational convenience, using the intensive conjugates more natural to the energy representation of the fundamental equation.

In what follows, we will take the function $g(N_{tot}; \vec{\phi}_0)$ either as the macrostate distribution, $\ln \Pi(N_{tot}; \vec{\phi}_0)$, or as the products of extensive thermodynamic properties (moments) which are conjugates of the intensive thermodynamic variables in which the Legendre transforms have been performed. For the grand canonical ensemble, this corresponds to particle number, N_i, and potential energy, U. These moments are collected in a matrix at each value of the order parameter over the course of the simulation,

$$Z(N_{tot}; \vec{\xi}) = N_1^{\xi_1} N_2^{\xi_2} \ldots N_k^{\xi_k} N_{tot}^{\xi_n} U^{\xi_u}, \quad (16)$$

where $\vec{\xi} = (\xi_1, \xi_2, \ldots, \xi_k, \xi_n, \xi_u)$ such that all ξ_i are non-negative integers, where $\xi_i \in [0, \xi_t + 1]$. Here, ξ_t refers to the maximum order of extrapolation we wish to achieve using Eq. (12). Note that we only require all terms such that $\sum \xi_i \leq \xi_t$ to extrapolate the macrostate distribution up to order ξ_t. However, one additional power is required to perform the same order of extrapolation on first order moments in the $Z(N_{tot}; \vec{\xi})$ matrix, which is required to compute extensive properties at the new conditions. As previously noted,[13] it is not necessary for the orders used in practice to be the same, but we generally keep them identical in this work. Note that strictly speaking, the order parameter, N_{tot}, is also included in the $Z(N_{tot}; \vec{\xi})$ matrix; however, this does not need to be explicitly measured during a simulation since its value is fixed at each "bin" in the matrix.

When members of $Z(N_{tot}; \vec{\xi})$, the matrix of moments, are averaged over the course of the simulation, they correspond to thermodynamic averages in the isochoric semigrand ensemble. Henceforth, we denote all such averages with a tilde, $\widetilde{X}$, where

X is some member of the $Z(N_{\text{tot}};\vec{\xi})$ matrix. Averages in the grand canonical ensemble may be obtained from these values by employing the macrostate distribution,

$$\langle X\rangle = \frac{1}{\Xi}\sum_{N_{\text{tot}}} \exp\left(\beta\mu_1 N_{\text{tot}}\right)\widetilde{X}(N_{\text{tot}}) = \sum_{N_{\text{tot}}} \bar{\Pi}(N_{\text{tot}})\widetilde{X}(N_{\text{tot}}). \tag{17}$$

Here, the macrostate probabilities have been normalized such that

$$\sum_{N_{\text{tot}}} \bar{\Pi}(N_{\text{tot}}) = 1. \tag{18}$$

At a first order phase transition between low- and high-density fluids, the macrostate distribution will display multiple peaks as a function of N_{tot}. Different peaks correspond to different phases, which are delineated by the local minimum in $\ln\Pi(N_{\text{tot}})$ between them. In this way, the macrostate distribution may be segmented into domains corresponding to different phases. At a given temperature, the coexistence between these phases is determined by searching for the chemical potential(s) that produce phases with equal pressure. The pressure of a phase, α, may be computed directly from the macrostate distribution,[32]

$$P^{\alpha}V\beta = \ln\left(\sum_{N_{\text{tot}}\in\alpha}\Pi(N_{\text{tot}})\right) - \ln\Pi(N_{\text{tot}}=0). \tag{19}$$

In a similar way, average grand canonical properties of an individual phase are computed by considering only those macrostates assigned to that phase,

$$\langle X\rangle^{\alpha} = \sum_{N_{\text{tot}}\in\alpha}\bar{\Pi}(N_{\text{tot}})\widetilde{X}(N_{\text{tot}}). \tag{20}$$

D. First order derivatives

We begin by considering the first order derivatives of the macrostate distribution and the moments matrix. Subsequently, in Sec. II E, we will demonstrate how higher order derivatives may be computed recursively using these results, due to fortuitous properties conferred by the application of the chain rule. Beginning with the macrostate distribution, the first derivative in terms of β follows easily from Eq. (6),

$$\frac{\partial\ln\Pi(N_{\text{tot}};\vec{\phi}_0)}{\partial\beta} = \mu_1 N_{\text{tot}} + \frac{1}{\Upsilon}\frac{\partial\Upsilon}{\partial\beta} - \frac{1}{\Xi}\frac{\partial\Xi}{\partial\beta} = \mu_1\left(N_{\text{tot}} - \langle N_{\text{tot}}\rangle\right) + \sum_{i=2}^{k}\Delta\mu_i\left(\widetilde{N}_i - \langle N_i\rangle\right) - \left(\widetilde{U} - \langle U\rangle\right), \tag{21}$$

which has already been reported in Refs. 13 and 14. Similarly, one can show that

$$\frac{\partial\ln\Pi(N_{\text{tot}};\vec{\phi}_0)}{\partial\Delta\mu_i} = \frac{1}{\Upsilon}\frac{\partial\Upsilon}{\partial\Delta\mu_i} - \frac{1}{\Xi}\frac{\partial\Xi}{\partial\Delta\mu_i} = \beta\left[\widetilde{N}_i - \langle N_i\rangle\right]. \tag{22}$$

Together, Eqs. (21) and (22) enable the first order extrapolation ($\xi_t = 1$) of the macrostate distribution. In order to extrapolate the matrix of moments, we require similar expressions for each member of the $Z(N_{\text{tot}};\vec{\xi})$ matrix. For an arbitrary moment in this matrix, Z, it can be shown[13] that in the semigrand ensemble,

$$\frac{\partial\widetilde{Z}}{\partial\beta} = \sum_{i=2}^{k}\Delta\mu_i\widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right) - \widetilde{f}\left(\widetilde{Z},\widetilde{U}\right), \tag{23}$$

while in the grand canonical ensemble,

$$\frac{\partial\langle Z\rangle}{\partial\beta} = \mu_1\hat{f}\left(\langle Z\rangle,\langle N_{\text{tot}}\rangle\right) + \sum_{i=2}^{k}\Delta\mu_i\hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right) - \hat{f}\left(\langle Z\rangle,\langle U\rangle\right). \tag{24}$$

Here $\widetilde{f}(\widetilde{X},\widetilde{Y})$ and $\hat{f}(\langle X\rangle,\langle Y\rangle)$ denote fluctuations of moments X and Y in the semigrand and grand canonical ensembles, respectively. For instance, in the semigrand ensemble, a fluctuation may be expressed as

$$\widetilde{f}(\widetilde{X},\widetilde{Y}) \equiv \widetilde{XY} - \widetilde{X}\widetilde{Y}. \tag{25}$$

A similar definition holds for the grand canonical ensemble,

$$\hat{f}(\langle X\rangle,\langle Y\rangle) \equiv \langle XY\rangle - \langle X\rangle\langle Y\rangle. \tag{26}$$

It has been discussed previously[13,14] that since members of the moments' matrix are used to estimate the derivatives of the partition functions, Υ and Ξ, it is critical that the energy, U, being recorded in $Z(N_{\text{tot}};\vec{\xi})$ is consistent with the energy which defines the canonical partition function. Therefore, it is important to reiterate that we have chosen to work only with the configurational part of the canonical partition function. This means that the energy in the moments' matrix is only the potential (configurational) energy. Consequently, U is not a function of β nor does the de Broglie wavelength implicitly contribute to the chemical potential of a species. In the classical limit, kinetic and potential energies are separable in the system's Hamiltonian and are independent.[4] Therefore, it is always possible to incorporate kinetic effects after the simulations and extrapolations have been performed, if so desired. This is generally accomplished by adjusting the reference state of a species' chemical potential to include a contribution from the de Broglie wavelength.[14]

Following a similar procedure as in Refs. 13 and 14 to obtain the partial derivatives with respect to the difference in chemical potentials, for the semigrand ensemble, we have

$$\frac{\partial\widetilde{Z}}{\partial\Delta\mu_i} = \beta\widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right), \tag{27}$$

and in the grand canonical ensemble,

$$\frac{\partial\langle Z\rangle}{\partial\Delta\mu_i} = \beta\hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right). \tag{28}$$

E. Second order derivatives

With the equations we have now presented, all subsequent derivatives necessary to compute the coefficients of the Taylor series [Eq. (12)] up to an arbitrary order may be found. This follows from the fact that further derivatives involve invoking the chain rule on fluctuations,

$$\left.\frac{\partial f(X,Y)}{\partial\phi_i}\right|_{\phi[i]} = \left.\frac{\partial(XY)}{\partial\phi_i}\right|_{\phi[i]} - X\left.\frac{\partial Y}{\partial\phi_i}\right|_{\phi[i]} - Y\left.\frac{\partial X}{\partial\phi_i}\right|_{\phi[i]}, \tag{29}$$

where ϕ_i is some individual thermodynamic variable, such as β or $\Delta\mu_k$, and X and Y are moments in the $Z(N_{tot};\vec{\xi})$ matrix averaged according to the ensemble in which the fluctuation occurs.

To illustrate, we begin by considering the second derivative of the macrostate distribution. The second derivative in β follows from Eq. (21),

$$\frac{\partial^2 \ln \Pi(N_{tot};\vec{\phi}_0)}{\partial\beta^2} = -\mu_1 \frac{\partial\langle N_{tot}\rangle}{\partial\beta} + \sum_{i=2}^{k} \Delta\mu_i \left(\frac{\partial\widetilde{N}_i}{\partial\beta} - \frac{\partial\langle N_i\rangle}{\partial\beta}\right) - \left(\frac{\partial\tilde{U}}{\partial\beta} - \frac{\partial\langle U\rangle}{\partial\beta}\right). \quad (30)$$

Through Eqs. (23) and (24), it is clear how the partial derivatives which appear here may be expressed in terms of fluctuations; semigrand fluctuations may be computed directly from the $Z(N_{tot};\vec{\xi})$ matrix using Eq. (25), whereas grand canonical fluctuations may be computed through the application of Eqs. (26) and (17). Similarly, the cross derivatives between temperature and chemical potential differences follow from Eq. (22),

$$\frac{\partial^2 \ln \Pi(N_{tot};\vec{\phi}_0)}{\partial\beta\partial\Delta\mu_i} = \frac{\partial}{\partial\beta}\left[\beta\left(\widetilde{N}_i - \langle N_i\rangle\right)\right] = \left(\widetilde{N}_i - \langle N_i\rangle\right) + \beta\left[\frac{\partial\widetilde{N}_i}{\partial\beta} - \frac{\partial\langle N_i\rangle}{\partial\beta}\right]. \quad (31)$$

In general, it can be shown that changing the order of the differentiation produces the same expression, which we leave as an exercise for the reader in the interest of brevity. Once again, applying Eqs. (23) and (24) allows this to be expressed in terms of fluctuations and evaluated. Also following from Eq. (22), the final cross derivative we require is

$$\frac{\partial^2 \ln \Pi(N_{tot};\vec{\phi}_0)}{\partial\Delta\mu_j\partial\Delta\mu_i} = \frac{\partial}{\partial\Delta\mu_j}\left[\beta\left(\widetilde{N}_i - \langle N_i\rangle\right)\right] = \beta^2\left[\widetilde{f}\left(\widetilde{N}_i,\widetilde{N}_j\right) - \hat{f}\left(\langle N_i\rangle,\langle N_j\rangle\right)\right], \quad (32)$$

where we have explicitly substituted the fluctuation formulas from Eqs. (27) and (28).

First, note that all second order derivatives of the macrostate distribution involve fluctuations. Second, observe that derivatives of fluctuations may be expressed in terms of derivatives of moments [cf. Eq. (29)]. It follows from Eqs. (23) and (24) that the derivative of a moment involves its fluctuation with respect to other extensive variables, i.e., moments of one order higher. This explicitly illustrates why we must measure moments up to $\xi \in [0, \xi_t + 1]$ if we are interested in extensive thermodynamic properties at the extrapolated conditions. Furthermore, it is now evident that iteratively applying the chain rule [Eq. (29)] and Eqs. (23) and (24) enables one to obtain expressions for partial derivatives up to arbitrary order.

Second order derivatives of the moments' matrix follow similarly, which we state here for completeness,

$$\frac{\partial^2\widetilde{Z}}{\partial\beta^2} = \sum_{i=2}^{k}\Delta\mu_i\frac{\partial\widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right)}{\partial\beta} - \frac{\partial\widetilde{f}\left(\widetilde{Z},\widetilde{U}\right)}{\partial\beta}, \quad (33)$$

$$\frac{\partial^2\langle Z\rangle}{\partial\beta^2} = \mu_1\frac{\partial\hat{f}\left(\langle Z\rangle,\langle N_{tot}\rangle\right)}{\partial\beta} + \sum_{i=2}^{k}\Delta\mu_i\frac{\partial\hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right)}{\partial\beta} - \frac{\partial\hat{f}\left(\langle Z\rangle,\langle U\rangle\right)}{\partial\beta}, \quad (34)$$

$$\frac{\partial^2\widetilde{Z}}{\partial\beta\partial\Delta\mu_i} = \frac{\partial\left[\beta\widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right)\right]}{\partial\beta} = \beta\left[\frac{\partial\widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right)}{\partial\beta}\right] + \widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right), \quad (35)$$

$$\frac{\partial^2\langle Z\rangle}{\partial\beta\partial\Delta\mu_i} = \frac{\partial\left[\beta\hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right)\right]}{\partial\beta} = \beta\left[\frac{\partial\hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right)}{\partial\beta}\right] + \hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right). \quad (36)$$

For the sake of clarity, we once again illustrate how to simplify these derivatives for the case of the cross derivative involving chemical potential differences. Using Eqs. (27) and (29), we have

$$\frac{\partial^2\widetilde{Z}}{\partial\Delta\mu_j\partial\Delta\mu_i} = \frac{\partial\left[\beta\widetilde{f}\left(\widetilde{Z},\widetilde{N}_i\right)\right]}{\partial\Delta\mu_j} = \beta^2\left[\widetilde{f}\left(\widetilde{ZN}_i,\widetilde{N}_j\right) - \widetilde{Zf}\left(\widetilde{N}_i,\widetilde{N}_j\right) - \widetilde{N}_i\widetilde{f}\left(\widetilde{Z},\widetilde{N}_j\right)\right]. \quad (37)$$

In the grand canonical ensemble, we can use Eqs. (28) and (29) to obtain a structurally similar result,

$$\frac{\partial^2\langle Z\rangle}{\partial\Delta\mu_j\partial\Delta\mu_i} = \frac{\partial\left[\beta\hat{f}\left(\langle Z\rangle,\langle N_i\rangle\right)\right]}{\partial\Delta\mu_j} = \beta^2\left[\hat{f}\left(\langle N_iZ\rangle,\langle N_j\rangle\right) - \langle Z\rangle\hat{f}\left(\langle N_i\rangle,\langle N_j\rangle\right) - \langle N_i\rangle\hat{f}\left(\langle Z\rangle,\langle N_j\rangle\right)\right]. \quad (38)$$

As a final note, we point out that we have included the derivatives of averages in the grand canonical ensemble, $\langle X\rangle$, in our derivations merely for the sake of completeness. In fact, all terms in the Taylor series expansions originating from derivatives of the grand partition function, $\Xi(\vec{\mu}, V, \beta)$, may be neglected in practice since they are constant across all values of N_{tot}.[14] Consequently, these grand canonical derivatives merely shift $\ln\Pi(N_{tot})$ in log-space leaving the relative macrostate probabilities unaffected. These terms appear only in the expansion used to extrapolate the macrostate distribution, $\ln\Pi(N_{tot})$, not the matrix of moments, $Z(N_{tot},\vec{\xi})$.

III. RESULTS

We now present a representative, systematic study of how this extrapolation approach may be used to compute the properties of binary fluid mixtures across a broad range of conditions, from a small number of initial simulations. Unless otherwise stated, we focus exclusively on the binary square-well mixture described in Sec. II B. Thus far, we have provided all the necessary equations to extrapolate both the macrostate distribution, $\ln\Pi(N_{tot})$, and the moments' matrix, $Z(N_{tot};\vec{\xi})$, obtained from a flat-histogram simulation at one set of

temperature and chemical potentials, $\vec{\phi}_0$, to another set of conditions, $\vec{\phi}$. Recall that replacing N_{tot} with N_1 and $\Delta\mu_i$ with μ_i enables the same approach to be used when N_1 is used as the sampling order parameter rather than N_{tot}. Although mathematically the order does not matter,[14] here we first reweighted a simulation using Eq. (14) to a desired μ_1 and then performed the extrapolation to reach a desired $\delta\vec{\phi}$ relative to the original simulation.

A. Multivariable macrostate extrapolation

First, we assess the ability of this method to accurately predict the macrostate distribution at chemical potentials and temperatures far from those of an initial simulation. Previous work[13] has show that temperature extrapolation can be quantitatively accurate over ranges on the order of $\delta\beta = \delta(1/T^*) \approx 0.25$ when using up to second order extrapolation. In Fig. 1(a), we consider the extrapolation exclusively in $\Delta\mu_2$ at fixed $\beta = 1/1.20 \approx 0.83$ for the square-well fluid mixture. This is a subcritical temperature where converging both the macrostate distribution and moments' matrix tends to be more difficult than at supercritical conditions. Regardless, these results were obtained after approximately 48 hours of wall clock time on an Intel Xeon 2.4 GHz processor; to improve efficiency, each simulation was broken into 10-20 smaller overlapping "windows" which were self-consistently combined after this amount of time had passed, as described in more detail elsewhere.[33,34]

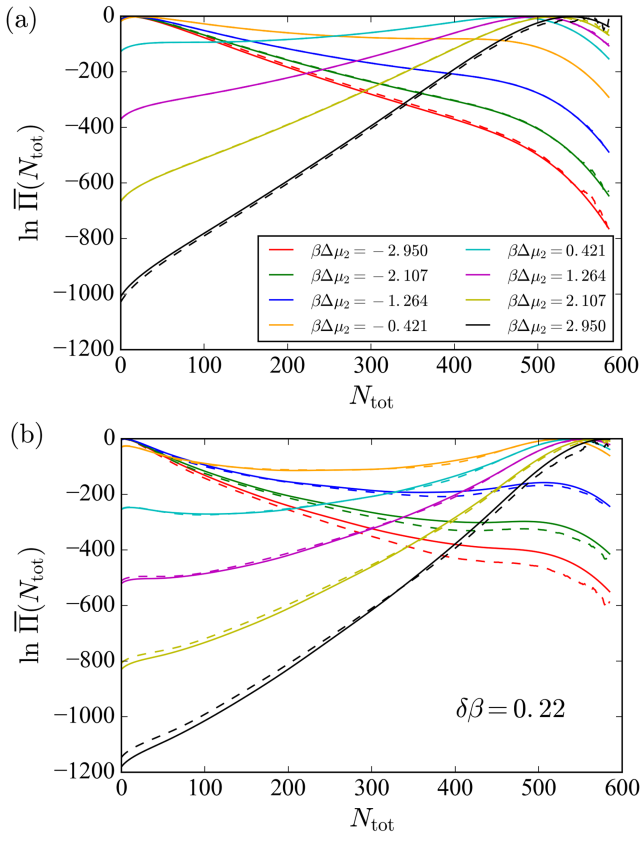

FIG. 1. Comparison between $\ln \Pi(N_{\text{tot}})$ for the square-well fluid resulting from the second order extrapolation of a single simulation performed at $\beta = 1/T^* = 1/1.20 \approx 0.83$, $\Delta\mu_2 = 0.00$ and simulations performed directly at other conditions. (a) Normalized macrostate distributions from direct simulation (solid lines) compared to those obtained via extrapolation (dashed) to different $\Delta\mu_2$ at fixed $\beta \approx 0.83$. (b) Same comparison as in (a), but when extrapolation has now been performed in both β (from $T^* = 1.20$ to $T^* = 0.95$, $\delta\beta \approx 0.22$) and $\Delta\mu_2$.

As we will illustrate later, this square-well fluid mixture forms a negative azeotrope which occurs at $\Delta\mu_2 = 0.00$. This is the initial condition we focus on extrapolating to other $\Delta\mu_2$ values, although similar results were obtained regardless of the initial value of $\Delta\mu_2$. The value of μ_1 is also inconsequential. We generally consider a range from $\beta\Delta\mu_i \in [-2.95, +2.95]$ because it corresponds to a fluid with a mole fraction of species one, x_1, ranging from approximately $0.05 \lesssim x_1 \lesssim 0.95$ which is estimated by assuming the fluid behaves as an ideal gas,

$$\beta\Delta\mu_2 = \ln\left(\frac{1}{x_1} - 1\right). \tag{39}$$

As shown in Fig. 1(a), extrapolated predictions deviate only weakly from those obtained from direct simulations over the composition range at the reported conditions. This suggests that simulations corresponding to $x_1 = 0.50$ ($\Delta\mu_2 = 0.00$) contain sufficient information to predict thermodynamic properties of mixtures with very different compositions, e.g., $x_1 = 0.05, 0.95$ ($\delta(\beta\Delta\mu_2) \approx \pm 3$). We emphasize that there is nothing unique about initiating the extrapolation from the azeotrope. In fact, similar or better extrapolations can be obtained from $\Delta\mu_2 \neq 0.00$ and are fully capable of reproducing these macrostate distributions, even on the opposite side of the azeotrope from which the initial simulation is chosen.

Next, we consider the simultaneous extrapolation in both the temperature and chemical potential, $\delta\vec{\phi} = (\delta\beta, \delta\Delta\mu_2)$. In Fig. 1(b), we extrapolate this single simulation at $\Delta\mu_2 = 0.00$ from $T^* = 1.20$ to $T^* = 0.95$ ($\delta\beta \approx 0.22$) across the same range of $\beta\Delta\mu_2$ values as before. As expected, the differences between these extrapolations and direct simulations are larger but are still relatively small. The qualitative shape of the curve is well predicted which is arguably the most significant aspect. This is because the comparison has been performed at identical (μ_1, $\Delta\mu_2$) values; however, we observe that the differences between an extrapolated distribution and the one obtained from direct simulation can be further minimized by slightly adjusting the value of μ_1.[14] This observation indicates that there will be some error in the predicted chemical potential corresponding to the state point in question, even though the estimation of the other thermodynamic properties will be generally accurate. We emphasize that this chemical potential difference is quite small.

Given the results presented thus far, it would appear that a similar magnitude of $\delta\beta$ can be employed in multivariable extrapolation as in scalar β extrapolation ($\delta\beta \approx 0.25$) over which the computation of thermodynamic properties remains quantitatively accurate. However, this is not entirely true. To estimate properties such as the mole fraction of species one in a given phase, $x_1 = \langle N_1\rangle/\langle N_{\text{tot}}\rangle$, the moments' matrix must be extrapolated so that N_1 as a function of N_{tot} is known at the new conditions, $\vec{\phi}$, enabling the use of Eq. (20). In this work, we chose to use a second order extrapolation, $\xi_t = 2$, to estimate the moments' matrix at the new conditions, which is the same order of extrapolation used to estimate the new macrostate distribution. As discussed in Sec. II E, the collection of terms in the $Z(N_{\text{tot}}; \vec{\xi})$ matrix that are of one order higher, i.e.,

$\xi_t + 1 = 3$, is thus required. We will explicitly demonstrate in Sec. III C that this tends to represent the principal source of error, rather than the extrapolated macrostate distribution itself. This is because a simulation must be run progressively longer to numerically converge each higher order moment. Therefore simulations must run longer than what is necessary to converge the moments used to extrapolate $\ln \Pi(N_{tot})$ to the same order ($\xi_t = 2$) as first order moments ($\xi_t + 1 = 3$). It then becomes an issue of diminishing returns to continue running a single simulation in an effort to extrapolate to incrementally larger $\delta\vec{\phi}$.

Alternatively, one could simply perform a small number of different simulations at relatively disparate values of $\vec{\phi}$ and then use this "extrapolation" approach to, in fact, "interpolate" to any condition bounded by these simulations. In order to accomplish this, one must be able to self-consistently combine these simulations in a way that predicts the properties (macrostate distribution and moments' matrix) at intermediate values of $\vec{\phi}$. Next, we describe one way in which this might be achieved and illustrate its effectiveness at predicting the thermodynamic properties of this mixture over large regions of $\vec{\phi}$ parameter space.

B. Combining simulations in a thermodynamically consistent manner

Conventional histogram reweighting was developed to enable individual Monte Carlo simulations performed at specific values of temperature, chemical potentials, etc., to produce a macrostate distribution over a continuous range of values in their neighborhood, increasing the computational efficiency of GCMC simulations by several orders of magnitude.[6,7] In the same spirit, we seek to take a small number of individual flat-histogram simulations at different $\vec{\phi}$ and combine them to produce a continuous surface accurate at (nearly) all $\vec{\phi}$. To do so, we formulate the problem as one where we seek to construct a hypersurface (macrostate distribution and moments) which satisfies the Gibbs-Duhem equation. For binary systems, we have

$$\psi = x_1 \left.\frac{\partial \mu_1}{\partial x_1}\right|_{\beta,P} + (1 - x_1) \left.\frac{\partial \mu_2}{\partial x_1}\right|_{\beta,P}. \tag{40}$$

If this surface is thermodynamically consistent, then $\psi = 0$. Since numerical errors will always persist, we formulate this as an optimization problem where we collect various isothermal isobars across the domain of interest and then minimize ψ^2.

In the grand canonical ensemble, it is natural to define our hypersurface in terms of the variables in $\vec{\phi}$. At a fixed temperature, the coordinates of the surface are defined by the chemical potentials of each species. For a binary mixture, the surface is constructed as follows. First, we perform a series of simulations at a fixed temperature with different values of $\Delta\mu_2$, e.g., $\beta_0\Delta\mu_2 = (\pm 2.95, \pm 1.1, 0.0)$. If we want to compute the properties of the system at some new state point via extrapolation, then we first determine the two "closest" direct simulations that bound the state point of interest in terms of $\Delta\mu_2$ using a distance, $d_i \equiv |\Delta\mu_2^{\text{target}} - \Delta\mu_2^{\text{neighbor},i}|$. In Fig. 2(a), the simulations which bound the point on the "left" and "right" are denoted d_l and d_r, respectively. For these two bounding direct simulations (or nearest neighbors), the macrostate distribution and the moments' matrix can then be extrapolated to the desired state point. Because extrapolating from the two neighbors will produce different results, the extrapolations are combined using a weighting function, $w(d_i)$,

$$X = \frac{X_l w(d_l) + X_r w(d_r)}{w(d_l) + w(d_r)}, \tag{41}$$

where X represents either $\ln \Pi(N_{tot})$ or the $Z(N_{tot}; \vec{\xi})$ matrix. Thus, the simulations performed to the left and right of the desired state point are extrapolated to the conditions of interest and then "mixed" using this weighting function to produce the final result. Although it is possible to define a procedure which mixes all the $\Delta\mu_2$ simulations, we purposefully choose not to do so. Since we are using Taylor polynomials to estimate the intermediate properties, it is expected *a priori* that the simulations which are performed at the most similar conditions will yield the most accurate estimates. Incorporating data obtained from increasingly dissimilar conditions will only introduce unnecessary error. Similarly, we expect that at state points which lie much closer to one neighbor than the other, e.g., $d_l << d_r$, the closer neighbor should principally determine the properties at that point.

To accommodate these expectations, we propose the following form of a normalized weighting function inspired by the Minkowski distance:

$$\bar{w}(d_i) = 1 - \frac{d_i^m}{d_l^m + d_r^m}, \tag{42}$$

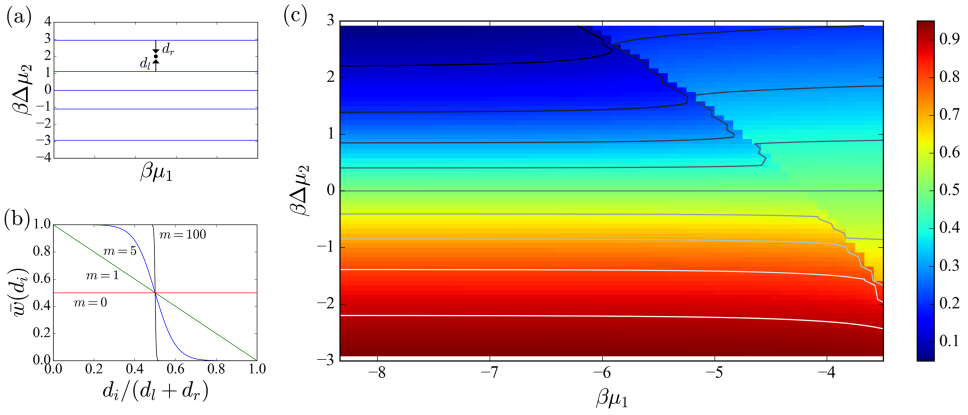

FIG. 2. Construction of a finely spaced $(\beta\mu_1, \beta\Delta\mu_2)$-grid from a discrete set of disparate simulations at a fixed temperature for the square-well fluid mixture. (a) Distance, d_i, for a target point, which is computed from an example set of simulations performed at the $\beta\Delta\mu_2$ indicated by the blue lines. (b) Normalized weighting function for different exponents, m. (c) Mole fraction of species one, x_1, at each point in the grid when $m = 2.5$ and $\beta = 0.83$. Only the properties of the most thermodynamically stable phase are reported. Isopleths have been traced out at $x_1 = (0.1, 0.2, \ldots, 0.9)$.

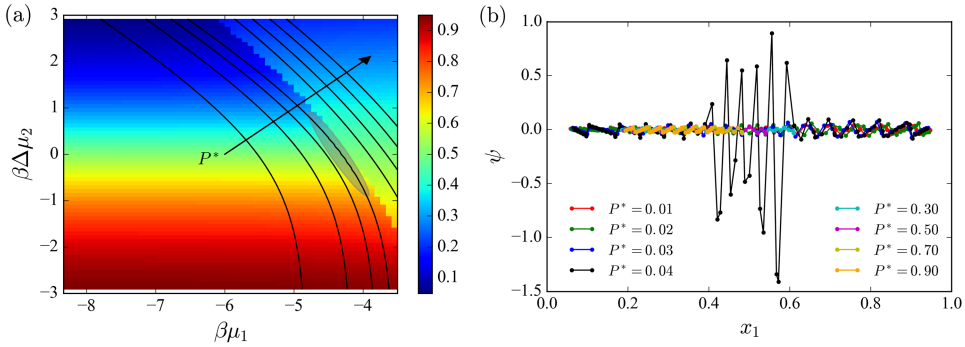

FIG. 3. Computing the deviation from the Gibbs-Duhem equation. (a) Sample isobars [black lines, values given in (b)] along which ψ is computed when m = 2.5 for the square-well fluid mixture. The underlying mole fraction, x_1, over the chemical potential grid at $\beta = 0.83$ is shown for reference. (b) Signed error, ψ, along each isopleth. For this system, $P^* = 0.40$ (black curve) passes near the azeotrope and corresponds to the gray region indicated in (a).

where m is some non-negative exponent and $i \in [l, r]$. This function is graphically depicted in Fig. 2(b). When $m = 0$, each of the neighboring simulations makes an equal contribution, regardless of a point's proximity to one or the other in $\Delta\mu_2$-space. In contrast, when $m \to \infty$, only the nearest neighbor is used to predict the properties at such a point. The case of $m = 1$ represents an intuitive "linear" scheme, which has been widely used in the past.[34,35]

Therefore, if we choose m to be a single value, valid throughout all of $\Delta\mu_2$-space, we may simply seek the value which minimizes ψ^2 for some predefined set of discrete isobars. In principle, one may allow m to vary throughout different regions of $\Delta\mu_2$-space; however, we find that using a single value is both simple and accurate. In fact, for the systems we report on here, we found that a value of $m \approx 2.5$ was generally optimal. However, for this square-well fluid, $m = 1$ did not yield substantially different results. We note that Eqs. (41) and (42) may be trivially extended to systems with an arbitrary number of components. However, in such a case, we anticipate that a separate exponent, m_i, would be necessary for each dimension of $\Delta\mu_i$.

Figure 2(c) depicts the x_1-surface over a range of chemical potentials when $m = 2.5$. A few isopleths (constant mole fraction) are drawn as lines to guide the eye which generally show the fluid to be rather ideal at sufficiently low μ_1, as expected; a phase transition occurs at the abrupt boundary where the isopleths change sharply. Although we have focused on presenting the x_1-surface because it is linked to the calculation of m, in principle, all thermodynamic properties may be computed at every point in this $(\beta\mu_1, \beta\Delta\mu_2)$-space. A sketch of how we computed ψ is outlined in Fig. 3(a). Several example isopleths have been extracted and plotted as black solid lines. In general, we choose to use isobars which are above or below coexistence since numerical instabilities tend to arise in the vicinity of a phase transition. The isobars chosen should span the region(s) of interest where further computation is intended. Here, we used linear interpolation on the $(\beta\mu_1,\beta\Delta\mu_2)$-grid to quickly parameterize the chemical potentials at which pressure is fixed at our desired value. Along these parameterized isobars, we computed the mole fraction, x_1, so that the derivatives in Eq. (40) could be evaluated numerically with cubic splines.

It is then possible to simply compute the ψ^2 error arising from each isobar and iteratively minimize the sum by changing the value of m used in the weighting function,

$$\min\left[\psi^2(m)\right] = \min\left[\sum_{P^*}\psi^2_{P^*}(m)\right]. \tag{43}$$

If the initial set of simulations performed at different $\Delta\mu_2$ values is not too sparse for the extrapolation order chosen and the simulations are sufficiently converged, then ψ tends to be relatively insensitive to m. In Fig. 3(b), we report the error along each isobar we considered. The weak oscillating deviations in Fig. 3(b) are due to our linear interpolation of the grid. However, we did not observe any significant differences in thermodynamic properties computed using more advanced interpolation schemes, so this level of error was found to be tolerable. The magnitude of these oscillations may also be decreased by using a finer grid. For most isobars, the oscillations are very small but begin to grow as the phase boundary is approached. It is clear, since we have the signed error, that fluctuations oscillate around zero, but are enhanced. As we will show in Sec. III C, the isobar of $P^* \equiv P\sigma^3_{1,1}/\epsilon_{1,1} = 0.40$ runs just below the azeotrope for the square-well mixture at this temperature, which is why the signed error exhibits this behavior.

C. Phase diagrams

Finally, to demonstrate the versatility and efficiency of this extrapolation approach, we compute the fluid-phase diagram for two different binary mixtures over a range of temperatures using only a small set of initial simulations. First, we consider the binary square-well fluid we have examined in detail thus far. For comparison, we performed direct simulations of this mixture at several representative temperatures, $T^* \in [1.20, 1.10, 0.95]$, across a range of $\Delta\mu_2$ values to obtain a phase diagram at each of these temperatures. At these subcritical temperatures, each $\Delta\mu_2$ corresponds to a single coexistence point. In Fig. 4, we report each of these points, where the lower density (vapor-like) phase is given by the filled circles, whereas the higher density (liquid-like) phase is given by inverted triangles. Clearly, the fluid forms a negative azeotrope.

Initially, we consider the predictions when only a single simulation is extrapolated. As a representative example, here we present the results from the simulation performed at $T^* = 1.20$ and $\Delta\mu_2 = 0.00$ ($x_1 = 0.5$). The corresponding point on the phase diagram is circled in black in Fig. 4(a). When $\Delta\mu_2$ is not too different, the extrapolated phase diagram is reasonably accurate over $0.3 \lesssim x_1 \lesssim 0.7$ even down to $T^* = 0.95$. However,

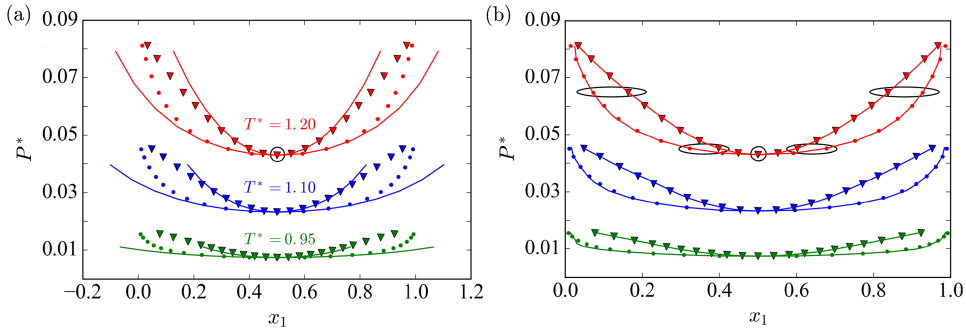

FIG. 4. Phase diagram for the binary square-well mixture studied. (a) Symbols represent coexistence points at each $T^* \in [1.20, 1.10, 0.95]$ (red, blue, and green, respectively) obtained from direct simulation. The lines correspond to the predictions made by second order extrapolation of the simulation performed at $T^* = 1.20$, $\Delta\mu_2 = 0.00$. The coexistence point this corresponds to is circled in black. (b) The same predictions now using a grid of $\Delta\mu_2 \in [\pm 3.54, \pm 1.01, 0.0]$ obtained at $T^* = 1.20$. The corresponding coexistence points are again circled in black.

beyond this range, it is clear that the estimates systematically diverge from the direct simulations, even to the point of predicting negative mole fractions. Since we employed Taylor polynomials to extrapolate the underlying moments' matrix, the unbounded growth of this error is to be expected as the magnitudes of $\delta\Delta\mu_2$ and $\delta\beta$ increase. In this case, we employed a second order extrapolation to estimate the macrostate distribution but only a first order extrapolation to predict the moments' matrix. There is simply nothing to prevent the underlying polynomial of, e.g., $\widetilde{N_1}(N_{\text{tot}})$, from violating a physical bound, such as $\widetilde{N_1} \geq 0$ or $\widetilde{N_1}(N_{\text{tot}}) + \widetilde{N_2}(N_{\text{tot}}) = N_{\text{tot}}$; the former of which results in $x_1 < 0$, while the latter can cause $x_1 > 1$.

This error can be reduced by combining several different simulations, especially those which correspond to states near the "edges" of the physically meaningful states. We chose to use $\Delta\mu_2 \in [\pm 3.54, \pm 1.01, 0.0]$ at $T^* = 1.20$ since they roughly correspond to a mixture of $x_1 \approx 0.05, 0.3, 0.50, 0.7, 0.95$ in the ideal gas limit ($\mu_1 \to -\infty$) at this temperature. These coexistence points are circled in Fig. 4(b). Using the method discussed in Sec. III B ($m = 2.5$), we computed these phase diagrams again. By using several simulations which more completely span the compositions of interest, the error is greatly suppressed. Notably, the introduction of the simulations corresponding to ideal gas mole fractions of $x_1 = 0.05, 0.95$ helps prevent the predicted mole fraction from violating its natural bounds, $0 \leq x_1 \leq 1$. Figure 4(b) shows excellent agreement between direct simulations and those obtained from the combined extrapolation of these five simulations. In fact, very similar results were obtained with only simulations corresponding to $\Delta\mu_2 \in [\pm 3.54, 0.0]$.

To further demonstrate the utility of this approach, we also consider a binary, linear force-shifted Lennard-Jones mixture (cf. Sec. II B). Compared to the azeotropic binary square-well fluid, this mixture has both longer ranged interactions, and particle size asymmetry. For the parameters chosen (cf. Table II), the two pure fluids have critical temperatures at $T^*_{c,1} \approx 1.08$ and $T^*_{c,2} \approx 1.44$.[36] We again performed five simulations at a supercritical temperature of $T^* = 1.50$ at $\beta\Delta\mu_2 = [\pm 2.95, \pm 1.10, 0.0]$ and combined the results using $m = 2.5$. Using second order extrapolation of both the macrostate

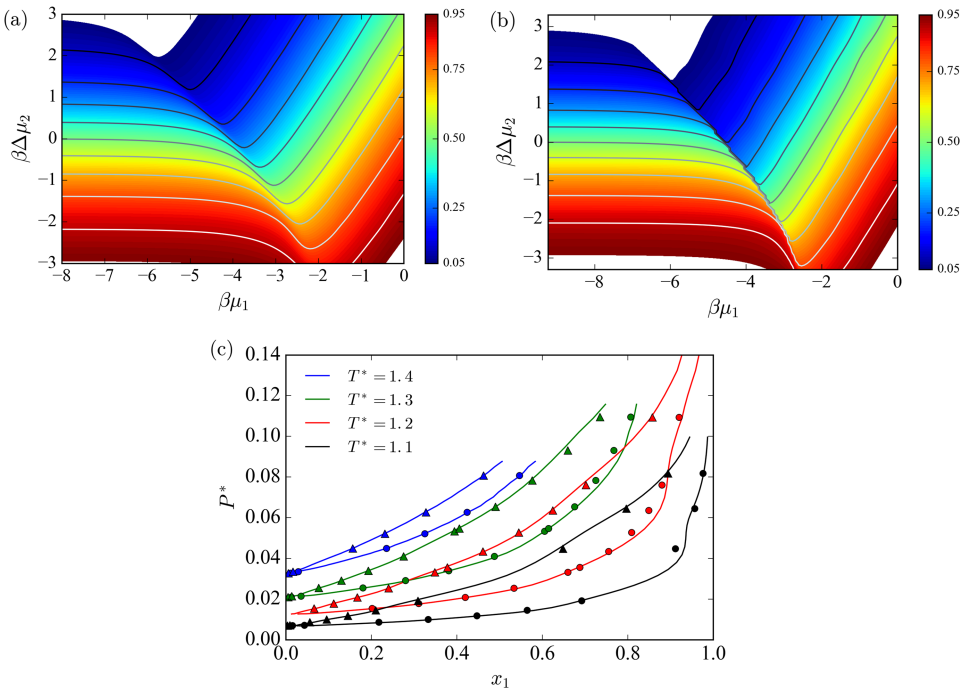

FIG. 5. Second order extrapolation of a linear force-shifted Lennard Jones mixture. (a) The mole fraction of species one, x_1, present in the most thermodynamically stable phase at $T^* = 1.50$. (b) x_1 present in the most thermodynamically stable phase at $T^* = 1.30$ predicted by extrapolating simulations from $T^* = 1.50$ ($m = 2.5$). (c) Phase diagrams as predicted by the extrapolation of supercritical simulations from $T^* = 1.50$ (lines) compared to direct subcritical simulations (circles, triangles).

distribution and moments matrix, we predicted the binodal curves over a range of temperatures from $T^* = 1.40$ to $T^* = 1.10$. The chemical potential phase space is depicted in Figs. 5(a) and 5(b) which correspond to the original supercritical state and a representative extrapolation to a subcritical one ($T^* = 1.30$), respectively. Figure 5(c) depicts a comparison between direct simulations and supercritical extrapolations across a range of subcritical temperatures. We observed excellent agreement across this range of conditions as we extrapolated to lower temperatures; moreover, extrapolating to higher temperature produced isotherms with a similar level of quantitative accuracy. We emphasize that this can be achieved with only second order moments, which are simple and efficient to collect. This level of agreement is representative of what we have found across a wide range of bulk and confined systems previously explored,[13,14,34] as well as those which are the subject of ongoing investigation.

IV. CONCLUSIONS

We have derived a method to extrapolate the grand canonical macrostate distribution of a multicomponent mixture from one set of temperature and chemical potential(s) to another. This is achieved by expanding the distribution as a multivariable Taylor series at the conditions originally simulated. Truncating the expansion to a finite order yields a Taylor polynomial which can then be used to predict the system's properties at any arbitrary condition. We explicitly derived the coefficients in this polynomial from classical statistical mechanics for a single sampling order parameter, N_{tot}; however, we emphasize that the form of these equations remains largely the same if a species-specific particle number (e.g., N_1) is used instead. Here, we have presented expressions for these coefficients up to second order, but we have also demonstrated how expressions for higher order derivatives follow directly from the equations provided.

Furthermore, we illustrated how extrapolations from multiple individual simulations may be combined to produce a continuous surface, defined by the chemical potentials of the mixture, in a way that optimizes numerical satisfaction of the Gibbs-Duhem equation. In concert, this extrapolation approach enables multicomponent fluid phase behavior to be computed over a broad, continuous range of conditions using data generated from a systematic, small number of flat-histogram GCMC simulations at different conditions. Consequently, the number of overall simulations necessary to compute fluid-phase diagrams and other thermodynamic properties of multicomponent fluid mixtures is greatly reduced compared to other computational approaches.

This multidimensional extrapolation approach is more approximate than approaches such as histogram reweighting but has the benefit of allowing one to easily infer properties at conditions far away from those of the original simulation(s). Although approximate, these extrapolations are accurate enough to provide reliable thermodynamic information over a broad range of conditions. As a result, we expect this approach to serve as a valuable tool especially when performing data-driven studies, such as high-throughput screenings, which require the generation of large amounts of reasonably accurate thermodynamic information quickly and in a computationally efficient manner.

ACKNOWLEDGMENTS

Contribution of the National Institute of Standards and Technology, not subject to US Copyright. N.A.M. gratefully acknowledges support from a National Research Council postdoctoral research associateship at the National Institute of Standards and Technology. J.R.E. acknowledges support from the National Science Foundation, Grant No. CHE-1362572.

[1] M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 1989).
[2] J. J. de Pablo, Q. Yan, and F. A. Escobedo, "Simulation of phase transtitions in fluids," Annu. Rev. Phys. Chem. **50**, 377–411 (1999).
[3] A. Z. Panagiotopoulos, "Monte Carlo methods for phase equilibria of fluids," J. Phys.: Condens. Matter **12**(3), R25–R52 (2000).
[4] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Volume 1 of Computational Science Series, 2nd ed. (Academic Press, Inc., 2001).
[5] D. P. Landau and K. Binder, *A Guide to Monte Carlo Simulations in Statistical Physics* (Cambridge University Press, 2009).
[6] A. M. Ferrenberg and R. H. Swendsen, "New Monte Carlo technique for studying phase transitions," Phys. Rev. Lett. **61**(23), 2635–2638 (1988).
[7] A. M. Ferrenberg and R. H. Swendsen, "Optimized Monte Carlo data analysis," Phys. Rev. Lett. **63**(12), 1195–1198 (1989).
[8] D. Wu and D. A. Kofke, "Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation," J. Chem. Phys. **123**(5), 054103 (2005).
[9] D. Wu and D. A. Kofke, "Phase-space overlap measures. II. Design and implementation of staging methods for free-energy calculations," J. Chem. Phys. **123**(8), 084109 (2005).
[10] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).
[11] Z. Tan, E. Gallicchio, M. Lapelosa, and R. M. Levy, "Theory of binless multi-state free energy estimation with applications to protein-ligand binding," J. Chem. Phys. **136**, 144102 (2012).
[12] A. L. Ferguson, "BayesWHAM: A bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method," J. Comput. Chem. **38**, 1583–1605 (2017).
[13] N. A. Mahynski, M. A. Blanco, J. R. Errington, and V. K. Shen, "Predicting low-temperature free energy landscapes with flat-histogram Monte Carlo methods," J. Chem. Phys. **146**(7), 074101 (2017).
[14] N. A. Mahynski, J. R. Errington, and V. K. Shen, "Temperature extrapolation of multicomponent grand canonical free energy landscapes," J. Chem. Phys. **147**, 054105 (2017).
[15] R. W. Zwanzig, "High temperature equation of state by a perturbation method. I. Nonpolar gases," J. Chem. Phys. **22**, 1420–1426 (1954).
[16] A. Z. Panagiotopoulos and R. C. Reid, "On the relationship between pairwise fluctuations and thermodynamic derivatives," J. Chem. Phys. **85**(8), 4650 (1986).
[17] P. G. Debenedetti, "On the relationship between principal fluctuations and stability coefficients in multicomponent systems," J. Chem. Phys. **84**(3), 1778 (1986).
[18] J. W. Tester and M. Modell, *Thermodynamics and Its Applications*, 3rd ed. (Prentice Hall, 1997).
[19] G. Hummer and A. Szabo, "Calculation of free-energy differences from computer simulations of initial and final states," J. Chem. Phys. **105**, 2004–2010 (1996).
[20] F. A. Escobedo, "Novel pseudoensembles for simulation of multicomponent phase equilibria," J. Chem. Phys. **108**(21), 8761–8772 (1998).
[21] F. A. Escobedo, "Simulation and extrapolation of coexistence properties with single-phase and two-phase ensembles," J. Chem. Phys. **113**(19), 8444–8456 (2000).
[22] D. Boda, T. Kristóf, J. Liszi, and I. Szalai, "A new simulation method for the determination of phase equilibria in mixtures in the grand canonical ensemble," Mol. Phys. **99**(24), 2011–2022 (2001).

[23]D. Boda, T. Kristóf, J. Liszi, and I. Szalai, "The extrapolation of the vapour-liquid equilibrium curves of pure fluids in the isothermal Gibbs ensemble," Mol. Phys. **100**(12), 1989–2000 (2002).
[24]T. Kristóf, J. Liszi, and D. Boda, "The extrapolation of phase equilibrium curves of mixtures in the isobaric-isothermal Gibbs ensemble," Mol. Phys. **100**(21), 3429–3441 (2002).
[25]F. A. Escobedo, "Mapping coexistence lines via free-energy extrapolation: Application to order-disorder phase transitions of hard-core mixtures," J. Chem. Phys. **140**, 094102 (2014).
[26]K. S. Rane, S. Murali, and J. R. Errington, "Monte Carlo simulation methods for computing liquid–vapor saturation properties of model systems," J. Chem. Theory Comput. **9**(6), 2552–2566 (2013).
[27]D. A. Kofke, "Gibbs-Duhem integration: A new method for direct evaluation of phase coexistence by molecular simulation," Mol. Phys. **78**(6), 1331–1336 (1993).
[28]J.-S. Wang, T. K. Tay, and R. H. Swendsen, "Transition matrix Monte Carlo reweighting and dynamics," Phys. Rev. Lett. **82**(3), 476 (1999).
[29]M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos, "An improved Monte Carlo method for direct calculation of the density of states," J. Chem. Phys. **119**, 9406 (2003).
[30]D. P. Landau, S.-H. Tsai, and M. Exler, "A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling," Am. J. Phys. **72**(10), 1294–1302 (2004).
[31]C. Desgranges and J. Delhommelle, "Evaluation of the grand-canonical partition function using expanded Wang-Landau simulations. III. Impact of combining rules on mixtures properties," J. Chem. Phys. **140**, 104109 (2014).
[32]J. R. Errington and V. K. Shen, "Direct evaluation of multicomponent phase equilibria using flat-histogram methods," J. Chem. Phys. **123**, 164103 (2005).
[33]J. R. Errington, "Direct calculation of liquid-vapor phase equilibria from transition matrix Monte Carlo simulation," J. Chem. Phys. **118**(22), 9915–9925 (2003).
[34]N. A. Mahynski and V. K. Shen, "Multicomponent adsorption in mesoporous flexible materials with flat-histogram Monte Carlo methods," J. Chem. Phys. **145**, 174709 (2016).
[35]V. K. Shen and D. W. Siderius, "Elucidating the effects of adsorbent flexibility on fluid adsorption using simple models and flat-histogram sampling methods," J. Chem. Phys. **140**, 244106 (2014).
[36]B. Smit and C. P. Williams, "Vapour-liquid equilibria for quadrupolar Lennard-Jones fluids," J. Phys.: Condens. Matter **2**(18), 4281 (1990).