Published in final edited form as:

J Chem Phys. 2018 May 21; 148(19): 194105. doi:10.1063/1.5026493.

Predicting structural properties of fluids by thermodynamic extrapolation

Nathan A. Mahynski^{1,a),b)}, Sally Jiao^{2,b)}, Harold W. Hatch¹, Marco A. Blanco^{1,3}, and Vincent K. Shen¹

¹Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8320, USA

²Department of Chemical and Biological Engineering, Princeton University, Princeton, New Jersey 08544, USA

³Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, Maryland 20850, USA

Abstract

We describe a methodology for extrapolating the structural properties of multicomponent fluids from one thermodynamic state to another. These properties generally include features of a system that may be computed from an individual configuration such as radial distribution functions, cluster size distributions, or a polymer's radius of gyration. This approach is based on the principle of using fluctuations in a system's extensive thermodynamic variables, such as energy, to construct an appropriate Taylor series expansion for these structural properties in terms of intensive conjugate variables, such as temperature. Thus, one may extrapolate these properties from one state to another when the series is truncated to some finite order. We demonstrate this extrapolation for simple and coarse-grained fluids in both the canonical and grand canonical ensembles, in terms of both temperatures and the chemical potentials of different components. The results show that this method is able to reasonably approximate structural properties of such fluids over a broad range of conditions. Consequently, this methodology may be employed to increase the computational efficiency of molecular simulations used to measure the structural properties of certain fluid systems, especially those used in high-throughput or data-driven investigations.

I. INTRODUCTION

The microscopic structure of a fluid is inexorably linked to its observable macroscopic (e.g., thermodynamic and dynamic) properties.^{1–3} Because theory and simulations provide the rigorous link between these disparate length scales, they figure prominently as design tools that enable the control of macroscopic behavior via deliberate manipulation of microscopic fluid structure. This control is particularly useful when designing composite materials, the self-assembly of soft matter systems, stabilization of protein solutions,^{4–9} and catalysis.^{10–18} Here, within the context of simulations, we consider the “structural properties” of a system

^{a)} nathan.mahynski@nist.gov.

^{b)} N. A. Mahynski and S. Jiao contributed equally to this work.

to be those which may be computed from a single configuration or “snapshot.” These microstate properties are generally trivial to obtain from particle-based simulations such as Monte Carlo (MC) or molecular dynamics (MD),^{19,20} if the appropriate interactions between particles are known. Alternatively, various statistical mechanical theories including integral equation theory,^{1,3,21} density functional theory,²² statistical associating fluid theory,^{23,24} and reference interaction site model theory^{25–28} allow these properties to be directly computed. However, such theories are limited to relatively simple systems; thus, molecular simulations are often needed to evaluate realistic systems and serve as a benchmark against which to compare these theories.

In most cases, individual molecular simulations pertain to only a single thermodynamic state point, e.g., temperature and chemical potential(s). However, histogram reweighting techniques enable the calculation of thermodynamic properties at similar state points.^{29–31} Underlying this approach is the fundamental concept of a “macrostate” which may be defined as the collection of microstates or configurations, with identical order parameter values; order parameters may be chosen arbitrarily, but the system’s joint instantaneous energy and particle number(s) is commonly employed.³¹ These variables emerge naturally due to their connection with thermodynamic potentials, such as entropy, in this case. The relative probability of observing a particular set of energy and particle number between two thermodynamic states is formally provided by statistical mechanics. Thus, if a macrostate distribution can be determined at one set of conditions, it can simply be recalculated at another set from the original distribution. One way to construct an initial macrostate distribution is by combining overlapping histograms from several independent, closely spaced simulations which span the conditions of interest.^{29–36} The combined macrostate distribution provides an accurate thermodynamic description over the range of simulated conditions which may be reweighted as needed.

Recently, a methodology has been proposed which revisits the principle^{37–46} of using thermodynamic fluctuations of a system to construct a Taylor series describing its macrostate distribution and other underlying properties.^{47–49} To this end, a flat-histogram Monte Carlo simulation is used to obtain the macrostate probability distribution and the moments of extensive properties across a predefined set of order parameter values at a thermodynamic state point of interest. It is well-known that pairwise fluctuations in a system are related to its thermodynamic derivatives.^{37,50,51} Thus, a Taylor series describing the macrostate distribution up to arbitrary order may be found by using the appropriate moments, measured over the course of a simulation, which provide the coefficients in this series. These Taylor series allow one to estimate the system’s thermodynamic and other underlying properties across a wide range of state points. In contrast to reweighting methods, which require that simulations sufficiently sample intermediate state points within the range of conditions of interest to obtain reliable statistics, this approach does not have to explicitly visit those states. Thus, it allows for the prediction of properties at conditions differing significantly from the original simulation. In other words, the methodology presented is extrapolative rather than interpolative.

Although truncating the Taylor series yields inherently approximate predictions, they have been shown to be quantitatively accurate for many fluids even when the series is truncated at

relatively low order in the grand-canonical ensemble.^{47–49} This stands somewhat in contrast to the canonical ensemble in which fluctuations in energy, which are required for temperature extrapolation, are often considered inadequate to accurately describe complex systems at significantly different temperatures.^{37,52} This is because the most probable regions of configurational phase tend to quickly become disjoint as temperature changes. This further restricts the degree to which changes due to the alchemical transformations may be expected to be reasonably extrapolated.^{52–54} However, this assessment depends on the complexity of the model being considered and the range of conditions over which one expects to extrapolate.

Here, we build upon this extrapolation methodology, demonstrating that structural properties of simple fluids may be reasonably predicted in a manner similar to their thermodynamic ones.^{49,54} This enables an efficient, simultaneous examination of a fluid's structural and thermodynamic properties across a broad range of state points. This can be particularly advantageous when comparing different metrics for fluid behavior which have both thermodynamic and structural definitions, such as the critical micelle concentration (CMC).^{55–60} The grand canonical ensemble is a convenient ensemble to use when locating first-order fluid phase transitions, since discontinuous changes in particle number(s) manifest at these transitions. While open systems are more amenable to Monte Carlo simulations, molecular dynamics methods are predominantly performed in closed systems where the number of molecules is fixed. However, this structural extrapolation approach can also be implemented in such thermodynamic ensembles. Consequently, the methodology may be used to enhance the computational efficiency of, or at least increase the amount of extractable information from, both MC and MD simulations.

This paper is organized as follows. In Sec. II, we discuss the thermodynamic ensembles that are employed here, followed by a brief review of previous work and the equations necessary to construct the requisite Taylor series of thermodynamic properties in these ensembles. We then explain how this may be adapted to extrapolate structural properties of fluids as well. In Sec. III, we discuss the application of this approach to various structural properties of simple or coarse-grained systems, including the radius of gyration of linear homopolymers (Sec. III A), radial distribution functions (RDFs) for pure and multicomponent square-well fluids (Sec. III B), cluster distributions and the RDF for a pure Lennard-Jones fluid (Sec. III C), and finally the critical micelle concentration of a model self-assembling system (Sec. III D). We conclude in Sec. IV.

II. METHODS

A. The canonical ensemble

First, consider a k -component fluid in the canonical ensemble. In this ensemble, the number of particles of each species, $\vec{N} = (N_1, \dots, N_k)$, the system volume, V , and the (inverse) temperature, $\beta \equiv 1/k_B T$, are fixed. Here, k_B is the Boltzmann constant. Often, one is interested in understanding how the properties of a fluid change with respect to β . From the perspective of statistical mechanics, a canonical ensemble-averaged property, $\hat{X}$, is given by

$$\hat{X} = \frac{\sum_{\nu} X(\nu) \Omega(\nu) \exp(-\beta U(\nu))}{\sum_{\nu} \Omega(\nu) \exp(-\beta U(\nu))}, \quad (1)$$

where ν is a set of microstates with degeneracy, $\Omega(\nu)$, and internal energy, $U(\nu)$; the sum in the denominator is the canonical partition function, $Q(\vec{N}, V, \beta)$. $\hat{X}$ may then be approximated at some arbitrary inverse temperature if it is known at some reference state, β_0 , using a Taylor series

$$\hat{X}(\beta) = \hat{X}(\beta_0) + \delta\beta \left. \frac{\partial \hat{X}}{\partial \beta} \right|_{\beta_0} + \frac{1}{2!} (\delta\beta)^2 \left. \frac{\partial^2 \hat{X}}{\partial \beta^2} \right|_{\beta_0} + \dots, \quad (2)$$

where $\delta\beta \equiv \beta - \beta_0$. Here, and throughout this paper, we state partial derivatives with respect to thermodynamic variables with the implied understanding that all other independent variables in the ensemble of interest are held constant. Expressions for the derivatives in this expansion can be obtained from fluctuations of extensive properties in the same ensemble.^{37,50,51} For example,

$$\frac{\partial \hat{X}}{\partial \beta} = -\widehat{XU} + \hat{X}\hat{U} = -f(\hat{X}, \hat{U}), \quad (3)$$

where U is the internal energy of the system and $f(\hat{X}, \hat{U})$ denotes the pairwise fluctuation in the canonical ensemble. Higher order derivatives follow from the application of the chain rule; consider that

$$\frac{\partial f(\hat{X}, \hat{U})}{\partial \beta} = \frac{\partial(\widehat{XU})}{\partial \beta} - \hat{X} \frac{\partial \hat{U}}{\partial \beta} - \hat{U} \frac{\partial \hat{X}}{\partial \beta} \quad (4)$$

which enables the calculation of the second-order derivative of $\hat{X}$ with respect to β . By recursively applying this formula, e.g., allowing $X \leftarrow (XU)$ or $X \leftarrow U$ as needed, third-order and higher derivatives may be computed. One simply requires the ensemble-averaged moments $\widehat{XU^{\xi_u}}$, where ξ_u is some non-negative integer, to compute the Taylor expansion up to order ξ_u . These ensemble averages can readily be obtained from traditional MC or appropriately thermostatted MD simulations.^{20,61}

In the canonical ensemble, kinetic energy contributions can be neglected since they depend only on the number of independent quadratic degrees of freedom in the system, which are fixed when the number of particles is constant. Thus, the (average) kinetic energy is simply a constant so we may take U as referring exclusively to the potential energy of the system.

Slightly more care must be observed in the grand canonical ensemble where the particle number fluctuates, but it is possible to do the same there as well.⁴⁸ Consequently, here and throughout this work, we take U as always referring to the potential energy of the system. This is particularly relevant for MD simulations which are commonly performed in the canonical ensemble. For example, by taking $X = R_g^2$, the (square of the) radius of gyration for a polymer system, the extrapolation approach can predict the collapse transition from only a single, or small number, of either canonical MD or MC simulations (cf. Sec. III A).

B. The grand canonical ensemble

Now consider the k -component mixture in the grand-canonical ensemble, where N_i the number of particles of species i , can fluctuate while its chemical potential, μ_i , is fixed. This ensemble has been the focus of previous work,^{47–49} so we will only summarize the necessary results here. The grand partition function, $\Xi(\vec{\mu}, V, \beta)$, is given by

$$\Xi(\vec{\mu}, V, \beta) = \sum_{N_1} \sum_{N_2} \cdots \sum_{N_k} \exp \left(\beta \sum_{i=1}^k \mu_i N_i \right) Q(\vec{N}, V, \beta), \quad (5)$$

where $\vec{\mu} = (\mu_1, \mu_2, \dots, \mu_k)$. Since we have neglected kinetic energy, the canonical partition function, $Q(\vec{N}, V, \beta)$, is essentially just the configurational integral and we emphasize that chemical potentials do not contain any implicit contributions from the de Broglie wavelength.⁴⁸ In the grand canonical ensemble, there are multiple independent intensive variables, i.e., β and $\vec{\mu}$. Thus, extrapolation may, in principle, be performed in multiple dimensions simultaneously. For convenience, when simulating multicomponent mixtures, it is often easier to perform a change of variables such that

$$\begin{aligned} \Xi(\vec{\mu}, V, \beta) &= \sum_{N_{\text{tot}}} \exp(\beta \mu_1 N_{\text{tot}}) \times \left[\sum_{N_2} \cdots \sum_{N_k} \exp \left(\beta \sum_{i=2}^k \Delta \mu_i N_i \right) Q(\vec{N}, V, \beta) \right] \\ &= \sum_{N_{\text{tot}}} \exp(\beta \mu_1 N_{\text{tot}}) \Upsilon(N_{\text{tot}}; \vec{\Delta \mu}, V, \beta), \end{aligned} \quad (6)$$

where $\mu_i \equiv \mu_i - \mu_1$, $\vec{\Delta \mu} = (\mu_2, \mu_3, \dots, \mu_k)$, and $\Upsilon(N_{\text{tot}}; \vec{\Delta \mu}, V, \beta)$ is the isochoric semigrand partition function.⁶² Thus, the total particle number, $N_{\text{tot}} = \sum_i^k N_i$, is a scalar order parameter that describes the macrostates of the system. The probability of observing a macrostate is given by

$$\ln \Pi(N_{\text{tot}}) = \beta \mu_1 N_{\text{tot}} + \ln \Upsilon - \ln \Xi. \quad (7)$$

Therefore, if one macrostate probability distribution is known at a fixed set of conditions ($\mu_1^a, \vec{\mu}, V, \beta$), then it may be computed at any other μ_1^b via the reweighting expression⁶²

$$\ln \Pi(N_{\text{tot}}; \mu_1^b) = \ln \Pi(N_{\text{tot}}; \mu_1^a) + \beta(\mu_1^b - \mu_1^a)N_{\text{tot}} + C, \quad (8)$$

where C is a constant related to the grand canonical partition functions at the two states, which may be set to zero in practice. Similar relationships hold in the case where a species-specific order parameter, such as N_1 , is used instead.⁴⁹ It is straightforward to compute this distribution using flat-histogram Monte Carlo techniques,^{62–64} and the method used to obtain $\ln \Pi(N_{\text{tot}})$ does not affect the extrapolation procedure. When performing MC simulations in this work, we use a combination of Wang-Landau and transition matrix Monte Carlo methods to construct this macrostate distribution in the grand canonical ensemble.^{47–49,65}

Knowledge of this macrostate distribution, along with other appropriate isochoric semigrand averages, enables the calculation of the system's thermodynamic properties.^{62,64,65} This is why previous work focused on extrapolating this distribution and the corresponding moments. In general, we may wish to extrapolate these in any of the intensive variables, $\vec{\phi} = (\beta, \vec{\mu})$. Note that we have removed μ_1 from the set of “extrapolation” variables because Eq. (8) provides an exact formula for changing this variable. The multidimensional Taylor polynomial for any function, $\psi(N_{\text{tot}}; \vec{\phi})$ evaluated at $\vec{\phi}_0$, may be expressed as

$$\psi(N_{\text{tot}}; \vec{\phi}) = \psi(N_{\text{tot}}; \vec{\phi}_0) + \delta\vec{\phi} \cdot \nabla \psi(N_{\text{tot}}; \vec{\phi}_0) + \frac{1}{2!} \left[\delta\vec{\phi} \cdot \mathbf{H}_{\psi}(N_{\text{tot}}; \vec{\phi}_0) \cdot \delta\vec{\phi}^T \right] + \dots, \quad (9)$$

where the superscript “T” denotes the transpose of the vector and $\mathbf{H}_{\psi}(N_{\text{tot}}; \vec{\phi}_0)$ represents the symmetric Hessian matrix. This is simply a generalization of Eq. (2).

When we take $\psi(N_{\text{tot}}; \vec{\phi}_0) = \ln \Pi(N_{\text{tot}}; \vec{\phi}_0)$, we may express the necessary derivatives as follows:^{47,49}

$$\frac{\partial \ln \Pi(N_{\text{tot}}; \vec{\phi}_0)}{\partial \beta} = \mu_1 (N_{\text{tot}} - \langle N_{\text{tot}} \rangle) + \sum_{i=2}^k \Delta \mu_i (\bar{N}_i - \langle N_i \rangle) - (\tilde{U} - \langle U \rangle), \quad (10)$$

where an average property in the isochoric semigrand ensemble is denoted $\tilde{X}$ and a grand canonical ensemble-averaged property is denoted $\langle X \rangle$. Simply averaging a measured quantity over the course of our MC simulations yields $\tilde{X}$ which is defined from statistical mechanics [cf. Eq. (1)],

$$\begin{aligned}\tilde{X}(N_{\text{tot}}) &= \frac{1}{Y} \sum_{N_2} \cdots \sum_{N_k} \sum_{\nu} X(\nu) \Omega(\nu) \times \exp \left(\beta \left[\sum_{i=2}^k \Delta \mu_i N_i - U(\nu) \right] \right) \quad (11) \\ &= \frac{1}{Y} \sum_{N_2} \cdots \sum_{N_k} \hat{X} \exp \left(\beta \sum_{i=2}^k \Delta \mu_i N_i \right) \mathcal{Q}(\vec{N}, V, \beta).\end{aligned}$$

Note that the grand canonical and isochoric semigrand averages are related by the macrostate distribution

$$\begin{aligned}\langle X \rangle &= \frac{1}{\Xi} \sum_{N_{\text{tot}}} \exp(\beta \mu_1 N_{\text{tot}}) \bar{X}(N_{\text{tot}}) \quad (12) \\ &= \frac{\sum_{N_{\text{tot}}} \Pi(N_{\text{tot}}) \bar{X}(N_{\text{tot}})}{\sum_{N_{\text{tot}}} \Pi(N_{\text{tot}})}.\end{aligned}$$

Also note that for single component systems, the isochoric semigrand ensemble is equivalent to the canonical ensemble. In terms of the chemical potential difference between species i and 1, it can be shown that⁴⁹

$$\frac{\partial \ln \Pi(N_{\text{tot}}; \vec{\phi}_0)}{\partial \Delta \mu_i} = \beta [\tilde{N}_i - \langle N_i \rangle]. \quad (13)$$

A second-order expansion requires knowledge of the components of the Hessian matrix, which are given by the following second-order derivatives:

$$\frac{\partial^2 \ln \Pi(N_{\text{tot}}; \vec{\phi}_0)}{\partial \beta^2} = -\mu_1 \frac{\partial \langle N_{\text{tot}} \rangle}{\partial \beta} + \sum_{i=2}^k \Delta \mu_i \left(\frac{\partial \tilde{N}_i}{\partial \beta} - \frac{\partial \langle N_i \rangle}{\partial \beta} \right) - \left(\frac{\partial \tilde{U}}{\partial \beta} - \frac{\partial \langle U \rangle}{\partial \beta} \right), \quad (14)$$

$$\frac{\partial^2 \ln \Pi(N_{\text{tot}}; \vec{\phi}_0)}{\partial \beta \partial \Delta \mu_i} = (\tilde{N}_i - \langle N_i \rangle) + \beta \left[\frac{\partial \tilde{N}_i}{\partial \beta} - \frac{\partial \langle N_i \rangle}{\partial \beta} \right], \quad (15)$$

$$\frac{\partial^2 \ln \Pi(N_{\text{tot}}; \vec{\phi}_0)}{\partial \Delta \mu_j \partial \Delta \mu_i} = \beta^2 [f(\tilde{N}_i, \tilde{N}_j) - f(\langle N_i \rangle, \langle N_j \rangle)], \quad (16)$$

where, following Eq. (3), we denote fluctuations in the isochoric semigrand ensemble as $f(\tilde{X}, \tilde{Y}) \equiv \tilde{X}\tilde{Y} - \langle \tilde{X} \rangle \langle \tilde{Y} \rangle$ and fluctuations in the grand canonical ensemble as $f(\langle X \rangle, \langle Y \rangle) \equiv \langle XY \rangle - \langle X \rangle \langle Y \rangle$.

Once again, it is clear that these derivatives may be evaluated by measuring the moments of extensive properties over the course of a simulation.^{47–49} For instance, in the isochoric semigrand ensemble, we have

$$\frac{\partial \tilde{X}}{\partial \beta} = \sum_{i=2}^k \Delta \mu_i f(\tilde{X}, \tilde{N}_i) - f(\tilde{X}, \tilde{U}) \quad (17)$$

and

$$\frac{\partial \tilde{X}}{\partial \Delta \mu_i} = \beta f(\tilde{X}, \tilde{N}_i). \quad (18)$$

Similarly, in the grand canonical ensemble,

$$\frac{\partial \langle X \rangle}{\partial \beta} = \mu_1 f(\langle X \rangle, \langle N_{\text{tot}} \rangle) + \sum_{i=2}^k \Delta \mu_i f(\langle X \rangle, \langle N_i \rangle) - f(\langle X \rangle, \langle U \rangle) \quad (19)$$

and

$$\frac{\partial \langle X \rangle}{\partial \Delta \mu_i} = \beta f(\langle X \rangle, \langle N_i \rangle). \quad (20)$$

Furthermore, higher order moments can be used to compute the derivatives of lower order ones if one exploits the chain rule as described in Sec. II A. For example, second-order derivatives of some quantity, X , in the isochoric semigrand ensemble are given by⁴⁹

$$\frac{\partial^2 \tilde{X}}{\partial \beta^2} = \sum_{i=2}^k \Delta \mu_i \frac{\partial f(\tilde{X}, \tilde{N}_i)}{\partial \beta} - \frac{\partial f(\tilde{X}, \tilde{U})}{\partial \beta}, \quad (21)$$

$$\frac{\partial^2 \tilde{X}}{\partial \beta \partial \Delta \mu_i} = \beta \left[\frac{\partial f(\tilde{X}, \tilde{N}_i)}{\partial \beta} \right] + f(\tilde{X}, \tilde{N}_i), \quad (22)$$

$$\frac{\partial^2 \tilde{X}}{\partial \Delta \mu_j \partial \Delta \mu_i} = \frac{\partial [\beta f(\tilde{X}, \tilde{N}_i)]}{\partial \Delta \mu_j} = \beta^2 [f(\tilde{X} \tilde{N}_i, \tilde{N}_j) - \tilde{X} f(\tilde{N}_i, \tilde{N}_j) - \tilde{N}_i f(\tilde{X}, \tilde{N}_j)], \quad (23)$$

where we have simplified the last expression for illustrative purposes.

C. Measuring the moments

This extrapolation methodology fundamentally relies on the measurement of ensemble-averaged moments of extensive properties, such as $\widetilde{U^2}$, $\widetilde{N_1^2 N_2 U}$, etc. For bookkeeping purposes, these quantities can be collected in a matrix, $Z(N_{\text{tot}}; \vec{\xi})$, whose elements enable one to calculate the coefficients in Eqs. (2) and (9),

$$Z(N_{\text{tot}}; \vec{\xi}) = \Gamma^\gamma \left[\overbrace{N_1^{\xi_1} N_2^{\xi_2} \dots N_k^{\xi_k} U^{\xi_u}}^{\text{Extensive Conjugates}} \right] \overbrace{N_{\text{tot}}^{\xi_n}}^{\text{O.P.}} \quad (24)$$

where $\vec{\xi} = (\xi_1, \xi_2, \dots, \xi_k, \xi_u, \xi_n)$ such that all ξ_i are non-negative integers. Here, the exponent γ may take the value of either 0 or 1. In the former instance, only extensive properties are contained within this matrix which enables the subsequent extrapolation of these properties, e.g., $\widetilde{N_1}$ and $\widetilde{N_2}$, as well as $\ln \Pi(N_{\text{tot}})$ in the grand canonical ensemble;⁴⁹ in the latter case, the matrix contains the structural property of interest and enables its extrapolation. The maximum value of ξ_i measured determines the highest order derivative we may evaluate and therefore the order to which extrapolation may be performed.⁴⁹ To extrapolate structural properties to a desired order, ξ_s , we must collect all moments such that $\sum \xi_i \leq \xi_s$. This matrix is the result of the product of three contributions: (1) the factor Γ^γ ; (2) the extensive conjugates of the intensive thermodynamic variables in which we intend to extrapolate (for the grand canonical ensemble, all N_i and U ; just U in the canonical ensemble); and (3) the order parameter, “O.P.” above (in this case, N_{tot}). The latter two enable one to extrapolate the thermodynamic properties of a system.

The factor (or function) Γ , appearing in Eq. (24), defines the current work. Using the equations presented thus far, it is possible to compute all the derivatives required to perform a Taylor series expansion of Γ from the $Z(N_{\text{tot}}; \vec{\xi})$ matrix; Eqs. (3) and (4) apply in the canonical ensemble, whereas Eqs. (17)–(23) apply in the isochoric semigrand and grand canonical ensembles, where $X = \Gamma$. As long as Γ does not explicitly depend on either β or any μ_i , its presence will not change the expressions for the coefficients (derivatives) in its

Taylor series expansion [cf. Eqs. (1) and (11), where $X = \Gamma$]. For example, $\Gamma = R_g^2$ is perfectly valid, whereas some function like $\Gamma = 1 + 2\beta$ would not be. While the latter example is contrived, this can be an issue when considering things related to kinetic energy. In this case, extrapolation is still possible but entails additional effort.⁴⁸

III. RESULTS

A. Linear homopolymer

We first consider the ability of this methodology to predict the radius of gyration of an isolated, linear homopolymer in an implicit solvent across a range of temperatures. In this case, the polymer is a fully flexible, force-shifted Lennard-Jones ($r_c = 4\sigma$) chain with five identical, attractive beads with rigid bonds of length σ . Here bonded neighbors have no interaction energy, while non-bonded pairs interact via

$$U(r) = \begin{cases} U_{\text{LJ}}(r) - U_{\text{LJ}}(r_c) - (r - r_c) \frac{dU_{\text{LJ}}}{dr} \Big|_{r_c} & r < r_c \\ 0 & r \geq r_c \end{cases}, \quad (25)$$

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

$$U_{\text{LJ}}(r) = 4\epsilon \left[\left(\frac{\sigma}{r} \right)^{12} - \left(\frac{\sigma}{r} \right)^6 \right]. \quad (26)$$

This model serves as a convenient benchmark since the behavior of these short Lennard-Jones chains is well understood theoretically.⁶⁶ We performed canonical metropolis MC simulations using FEASST.⁶⁷ A cubic simulation box of length $L = 8\sigma$ and the following configurational bias moves were utilized: regrowth of the entire chain, regrowth of four beads from a fixed end, and regrowth of two beads, while fixing three beads at one end. These were performed with a relative probability of 1:2:4, respectively. Following a short equilibration period, approximately 10^7 MC trials were performed during which the moments were collected; however, we observed that moments even up to fourth order converged well within 10^6 trials. We emphasize that the same results can be obtained with canonical MD simulations as well, if they are appropriately thermostatted.^{20,61}

After each step of the simulation, the gyration tensor of the polymer was computed and subsequently diagonalized. The diagonal elements, denoted as λ_1^2, λ_2^2 , and λ_3^2 , represent the principal semi-axes of the polymer and their sum yields the square of the radius of gyration, $R_g^2 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$. These semi-axes were then sorted according to their magnitude such that $\lambda_1^2 > \lambda_2^2 > \lambda_3^2$. By taking $\Gamma = \lambda_i^2$ and recording a moments matrix [Eq. (24)] for each semi-axis,

we constructed a Taylor series up to fourth order for each, which allowed us to extrapolate each $\widehat{\lambda}_i^2$ and compute $\widehat{R}_g^2(\beta) = \widehat{\lambda}_1^2(\beta) + \widehat{\lambda}_2^2(\beta) + \widehat{\lambda}_3^2(\beta)$.

We performed simulations at three temperatures, $T^* = k_B T/e = 0.70, 0.20, 0.10$, and the corresponding extrapolations for both second and fourth orders are given in Fig. 1(a). Compared to the known behavior of the chain,⁶⁶ it is clear that higher order extrapolation produces more accurate predictions than the lower order ones. Nonetheless, even second-order extrapolations are at least qualitatively accurate over a similar range of temperature differences, $\delta\beta$, relative to the simulation temperature. Note that the simulation near the collapse temperature ($T^* \approx 0.50$) produced extrapolations accurate over the broadest temperature range, consistent with the behavior of this extrapolation method near critical points.⁴⁷ Following previous work,⁴⁹ we may combine the results of these local extrapolations into a single curve [cf. Fig. 1(b)] by “mixing” predictions from neighboring simulations such that

$$\widehat{R}_g^2(\beta; m) = \frac{\widehat{R}_g^2(\beta_l) |\beta - \beta_r|^m + \widehat{R}_g^2(\beta_r) |\beta - \beta_l|^m}{|\beta - \beta_r|^m + |\beta - \beta_l|^m}, \quad (27)$$

where $\beta_l < \beta < \beta_r$, and β_l and β_r correspond to the (inverse) simulation temperatures that bound the temperature of interest. Here we use $m = 20$, but this may be treated as a variable to be optimized if there exist known constraints that should also be satisfied by the resulting curve.⁴⁹

In Fig. 1(b), the combined curve eventually deviates at low temperature ($T^* \lesssim 0.05$), sufficiently far from the closest simulated temperature. Although this is clearly erroneous since the radius of gyration is being compared to known results, it is intuitive that the estimate is incorrect there. In fact, if one compares extrapolations of several different orders, one can infer where $\delta\beta$ becomes too large for the Taylor polynomial to be reasonably accurate. As a rule of thumb, we propose this to be a convenient way to cover β -space when the underlying quantity (in this case, $\widehat{R}_g^2$) is not known *a priori*. First, a single simulation is performed and extrapolated. From this, one may estimate at what temperature the next simulation should be performed, and so on. Following, one uses Eq. (27) to combine the results. This is the procedure we followed in this case, which allowed us to accurately capture $\widehat{R}_g^2$ for this polymer over more than two decades of temperature with only three simulations.

B. Square-well fluid mixtures

Next, we consider the structural properties of square-well fluids. For these simple fluids, the interparticle interactions are given by

$$U_{i,j}(r) = \begin{cases} \infty & r \leq \sigma_{i,j} \\ -\varepsilon_{i,j} & \sigma_{i,j} < r \leq \lambda_{i,j}\sigma_{i,j} \\ 0 & \lambda_{i,j}\sigma_{i,j} < r \end{cases} \quad (28)$$

for two species i and j . The parameters for the two fluids we consider are given in Table I. We will first consider a pure fluid in the canonical and grand canonical ensembles and then a binary mixture in the grand canonical ensemble.

1. Single component in the canonical ensemble—The single component square-well system we begin with is pure species 2. We performed canonical MetropolisMCsimulations⁶⁷ using simple translation moves on $N = 100$ particles contained within a cubic simulation box of length $L = 8\sigma_{2,2}$ ($V = 512\sigma_{2,2}^3$). These simulations were run for a duration on the order of 10^8 MC trials, over which the moments matrix was collected.

Extrapolation of $\widehat{g(r)}$ is actually based upon extrapolation of each bin in the unnormalized histogram $h(r_k; \beta_0)$ which records the number of times two particles are observed a distance r_k apart. Thus, $\Gamma = h(r_k; \beta_0)$, and a moments matrix must be collected for each bin centered at r_k . This enables the extrapolation of each bin to a target β using a Taylor polynomial of desired order. Here, we report the results of second-order extrapolation. The resulting $h(r_k; \beta)$ is then divided by N_c , the number of configurations collected in the histogram at β_0 , N , the total number of particles in each configuration (in this case, $N = 100$), and the number of particles that would be in each bin if the fluid were an ideal gas. In general, the radial distribution function of species j around species i in a multicomponent mixture is given by

$$\widehat{g_{i,j}(r_k)} = \frac{h_{i,j}(r_k; \beta)}{N_c \left[\left(\frac{N_j - \delta_{i,j}}{V} \right) V_{\text{bin}}(r_k) \right] N_i}, \quad (29)$$

where $\delta_{i,j}$ is the Kronecker delta function. For a pure component system, $i = j$ and we neglect the redundant species subscripts though in Sec. III B 3 we will move beyond this. This is the conventional way in which $\widehat{g(r)}$ is computed from $h(r; \beta)$,²⁰ but now we have simply obtained it via extrapolation instead of direct measurement. In other words,

$$\begin{aligned}\widehat{g(r_k)} &= \frac{h(r_k; \beta_0) + \delta\beta \frac{\partial h(r_k; \beta_0)}{\partial \beta} + \frac{1}{2!}(\delta\beta)^2 \frac{\partial^2 h(r_k; \beta_0)}{\partial \beta^2}}{N_c \left[\left(\frac{N-1}{V} \right) V_{\text{bin}}(r_k) \right] N} \quad (30) \\ &= \widehat{g(r_k)} \Big|_{\beta_0} + \delta\beta \frac{\partial \widehat{g(r_k)}}{\partial \beta} \Big|_{\beta_0} + \frac{1}{2!}(\delta\beta)^2 \frac{\partial^2 \widehat{g(r_k)}}{\partial \beta^2} \Big|_{\beta_0},\end{aligned}$$

where the partial derivatives of $h(r_k; \beta_0)$ may be evaluated from the relevant moments collected. Note that because the normalization factors are constant in the canonical ensemble, there is essentially no difference between setting $\Gamma = h(r_k; \beta_0)$ vs. $\Gamma = g(r_k)$; this will not be the case in the semigrand ensemble which will be explored later in Sec. III B 3.

Extrapolated radial distribution functions inherently contain numerical errors because they are approximations. Generally, this manifests as improper normalization,⁶⁸ so, as a final step, we scale the extrapolated radial distribution function uniformly across all bins so that it will satisfy mass conservation,⁶⁹

$$\frac{\hat{N}_j}{V} \int_0^L \frac{1}{2} [\widehat{g_{i,j}(r)} - 1] 4\pi r^2 dr = -\delta_{i,j}. \quad (31)$$

This equation is true in the limit that the system is homogeneous and contains no long-range spatial correlations. Following this procedure, we extrapolated the radial distribution function of this pure square-well fluid from $T^* = k_B T / \epsilon_{2,2} = 1.35$ to $T^* = 1.05$. The results are reported in Fig. 2. The critical temperature for this fluid is $T^* \approx 1.22$,^{70,71} and therefore, the final temperature is subcritical. In fact, the system's density falls within the binodal curves as well, and the system is phase-separated. As a result, the rescaled $\widehat{g(r)}$ fails to converge to unity at $r \rightarrow L/2 = 4\sigma_{2,2}$ for both extrapolations and direct simulations; regardless, they are consistent with each other. This lack of convergence to unity may be remedied by working in the grand canonical ensemble instead, where we may consider the system in only one phase at a time.

2. Single component in the grand canonical ensemble—At subcritical temperatures, the macrostate distribution, $\ln \Pi(N)$, contains multiple local maxima, each corresponding to different macroscopic phases. Macrostates bounded by the local minima on each side of maxima are considered to contribute to that phase.^{62,64,72} Phase coexistence is obtained by using Eq. (8) to locate the chemical potential where there are equal areas under these peaks, which corresponds to phases with equal pressures,

$$\langle P^\alpha \rangle V \beta = \ln \left(\sum_{N \in \alpha} \frac{\Pi(N)}{\Pi(0)} \right), \quad (32)$$

where P^α is the pressure of phase α , and the summation is restricted to those macrostates belonging to that phase. By performing this macrostate segmentation, it is possible to consider only those macrostates contributing to an individual phase of interest, which allows one to produce a grand canonical ensemble-averaged radial distribution function for a single phase,

$$\langle g^\alpha(r) \rangle = \frac{\sum_{N \in \alpha} \Pi(N) \widehat{g(r; N)}}{\sum_{N \in \alpha} \Pi(N)}. \quad (33)$$

To test this extrapolation approach, we performed grand canonical flat-histogram MC simulations⁶⁷ at various temperatures in the same simulation box volume as those in the canonical ensemble. Using up to second-order terms, we reweighted and then extrapolated the macrostate distribution, $\ln \Pi(N)$, from a supercritical state at $T^* = 1.35$ to subcritical temperatures as described in Sec. II B. To make use of Eq. (33), the radial distribution function at each N was extrapolated in a similar manner, and then each was rescaled with Eq. (31). As a final step, we chose to scale the resulting $\langle g(r) \rangle$ again to satisfy Eq. (31) since the macrostate distribution is also approximate and will affect the estimate of $\langle g(r) \rangle$ from $\widehat{g(r; N)}$. However, this was done for the sake of simplicity. For an open system, Eq. (31) is not strictly true, rather⁶⁹

$$\frac{\langle N^2 \rangle - \langle N \rangle^2}{\langle N \rangle} - 1 = \lim_{R \rightarrow \infty} \left(\frac{\langle N \rangle}{V} \int_0^R [\langle g(r) \rangle - 1] 4\pi r^2 dr \right), \quad (34)$$

where the moments of N required may be computed from the $Z(N_{\text{tot}}; \vec{\xi})$ matrix and the macrostate distribution. Rescaling to satisfy Eq. (34) instead of Eq. (31) as a final step would be a more accurate, albeit involved, approach. Regardless, this does not affect the qualitative comparison between direct simulations and extrapolations as long as they are normalized consistently.

In Fig. 3, we report the binodal curve and the associated radial distributions functions for the coexisting vapor and liquid phases at representative temperatures. Clearly, the extrapolated predictions match quite well with those obtained from direct simulations at subcritical temperatures, consistent with the corresponding thermodynamic quantities over a similar range of $\delta\beta$.⁴⁷ Although we have chosen to present the results at coexistence, these radial distribution functions may also be computed along isotherms at essentially any desired density, even away from coexistence. At very low density (chemical potential), some care must be taken computationally because $\widehat{g(r)}$, a fundamentally pairwise distribution, is not

defined when fewer than two particles are present. We address this further in Sec. III B 3; however, this is rarely of great interest since fluids approach the well-known expression in this limit,⁶⁹

$$\lim_{\rho \rightarrow 0} \langle g(r) \rangle = \exp[-\beta U(r)]. \quad (35)$$

3. Binary mixture in the grand canonical ensemble—A similar procedure may be followed to predict the structural properties of fluid mixtures. For many structural metrics, there is no additional complexity involved, other than that Eq. (24) now contains more $N_i^{\xi_i}$ terms pertaining to additional components. However, in the case of extrapolating the radial distribution functions, an additional bookkeeping step is required due to our choice of N_{tot} as the sampling order parameter.

The issue is that $g_{i,j}(r)$ is strictly defined in the canonical ensemble and, by virtue of its normalization, is an explicit function of individual particle numbers, N_1 and N_2 , not N_{tot} ; therefore, in order to compute $g_{i,j}(r)$ in the first place, one must separately accumulate an unnormalized $h_{i,j}(r)$ for each unique set of (N_1, N_2) observed during the simulation and then divide by factors involving these numbers after the simulation has finished to correctly normalize the distance histogram into the radial distribution function [cf. Eq. (29)]. Similarly, the moments used to extrapolate $g_{i,j}(r)$ will require this independent accumulation. Once the simulation is complete, all instances where $N_1 + N_2 = N_{\text{tot}}$ for each value of N_{tot} can be combined to provide the semigrand ensemble-averaged $\widetilde{g}_{i,j}(r; N_{\text{tot}})$ and other necessary moments.

We emphasize that this is a special consequence of the disparity between the definition of the structural metric we are concerned with and our choice for an order parameter, N_{tot} . It is not a complication that arises directly from the extrapolation methodology itself. Algorithm 1 (see the Appendix) summarizes the appropriate steps which enable the collection and normalization of the necessary moments. A further caveat addressed in Algorithm 1 is that neither the sum pertaining to each unnormalized moment nor the counts value by which the sum is divided can include states for which $g_{i,j}(r)$ is undefined. This occurs when one of the normalization factors becomes zero. Specifically, there are two cases when $g_{i,j}(r)$ becomes undefined: (1) when $\min[N_1, N_2] = 0$ for $i \neq j$ and (2) when $N_i = 1$ for $i = j$. Indeed, all such invalid states must not contribute to the $h_{i,j}^{\text{Eu}}(r; N_1, N_2)$ matrix so that derivatives (fluctuations) of $h_{i,j}(r; N_1, N_2)$ are subsequently computed self-consistently. Note that this does not affect the collection of the macrostate distribution, $\ln \Pi(N_{\text{tot}})$, for which $N_1 = 0, 1$ are valid states.

Once a simulation has been completed at some $\vec{\phi} = (\beta, \mu_2)$, we first reweight the reference macrostate distribution to the desired μ_1 and then perform a second-order extrapolation in $\vec{\phi}$ to reach the final state. Again, we obtain a prediction for each $\widetilde{g}_{i,j}(r; N_{\text{tot}})$ via extrapolation, followed by rescaling to satisfy Eq. (31). Subsequently, Eq. (36) is applied to compute the

grand canonical ensemble-averaged radial distribution functions, followed by another rescaling of $\langle g_{i,j}^\alpha(r) \rangle$ for each phase α ,

$$\langle g_{i,j}^\alpha(r; \mu_1, \vec{\phi}) \rangle = \frac{\sum_{N_{\text{tot}} \in \alpha} \Pi(N_{\text{tot}}; \mu_1, \vec{\phi}) \widetilde{g}_{i,j}(r; N_{\text{tot}})}{\sum_{N_{\text{tot}} \in \alpha} \Pi(N_{\text{tot}}; \mu_1, \vec{\phi})}. \quad (36)$$

Using parameters similar to those in the single component case, a simulation⁶⁷ of the binary square-well mixture was performed at $T^* = 1.35$, $\mu_2/\epsilon_{2,2} = 0$ and then extrapolated to other conditions. At the simulation temperature, the binary mixture exists as a single phase⁷¹ but undergoes vapor-liquid phase separation at the lower ones. We extrapolated the high temperature simulation down to $T^* = 1.10$ at various points along the binodal (different μ_2 values) using a second-order Taylor polynomial for the macrostate distribution and a first-order polynomial for $\widetilde{g}_{i,j}(r; N_{\text{tot}})$. For comparison, simulations were performed at $T^* = 1.10$ and $\mu_2/\epsilon_{2,2} = [\pm 3, \pm 2, \pm 1, 0]$ which may be combined to reconstruct the phase diagram.⁴⁹ Figure 4 illustrates the representative results of this extrapolation, which once again show good agreement between extrapolated predictions and direct simulations.

C. Lennard-Jones fluid

Having only applied the extrapolation method to systems governed by relatively short-ranged interactions thus far, we next turn our attention to a single component Lennard-Jones fluid. For this fluid, the particle interactions are given by Eq. (26) where interactions were truncated at $r_c = 3\sigma$ but were compensated with long-range corrections.²⁰ The critical temperature for this fluid is approximately $T_c^* = k_B T_c / \epsilon \approx 1.291$.⁷³ To evaluate this extrapolation approach, we carried out grand canonical flat-histogram MC simulations in a cubic box of length $L = 9\sigma$ across a range of temperatures.

Similar to the radial distribution function, we also extrapolated the cluster size distribution, $\widehat{N}(s)$, collected at each (canonical) macrostate; these canonical distributions are then combined according to Eq. (12) to produce the grand canonical distribution, $\langle N(s) \rangle$. To construct this distribution, a particle is considered to be part of a cluster of size s if it is within a cutoff distance of 1.5σ from at least one of the other $s - 1$ particles belonging to the cluster. This cutoff distance ensures that the separation between two particles is within the well of the pair potential, following traditional clustering criteria.⁷⁴ $\widehat{N}(s)$ represents the probability that a randomly chosen sphere resides in a cluster of size s , and it is normalized so that integrating over the entire distribution yields the system's number density, $\hat{\rho}$,

$$\widehat{N}(s) = \left[\frac{\widehat{P}(s) \times s}{\sum_s \widehat{P}(s) \times s} \right] \hat{\rho}, \quad (37)$$

where $\widehat{P(s)}$ is the probability of observing a cluster of size s when the system's density is fixed at $\hat{\rho}$.

In Fig. 5, we report the phase diagram computed from direct simulations and via extrapolation as well as the radial distribution functions along the coexistence curve. Here, we report the results of second-order extrapolation from $T^* = 1.20$. As in the case of the square-well fluid (Sec. III B 2), we find excellent agreement across the entire range of temperatures investigated, both in terms of the thermodynamic and structural properties. Similarly, the cluster size distribution along the coexistence curve is remarkably accurate (cf. Fig. 6) even down to low temperatures. We are able to extrapolate with little to no noticeable error from $T^* = 1.20$, which is subcritical but near the critical point, down to as low as $T^* = 0.85$. This is particularly remarkable since the triple point temperature for this fluid is $T_t^* \approx 0.70$.^{74,75}

The most significant deviations of the extrapolated $\langle N(s) \rangle$ in the liquid phase are observed at large $\delta\beta$. These deviations are more noticeable for the less populated clusters. Nonetheless, they are still relatively small and behave as random error, oscillating around the true value of $\langle N(s) \rangle$. Unlike in the liquid phase, the extrapolated $\langle N(s) \rangle$ of the vapor phase is indistinguishable from the simulated distribution for most of the tested cluster-size range and at all temperatures. Differences are only noticeable for $\langle N(s) \rangle < 10^{-6}$ which is around the statistical limit for the number of sampled configurations. Observe that the extrapolation approach is able to accurately capture that certain cluster sizes are essentially forbidden at low temperatures—cf. $\langle N(s) \rangle$ at $T^* = 1.20$ vs. $T^* = 0.85$ in Fig. 6(a). Clusters of preferential size are known to be associated with the nucleation of solid-like particles within the fluid as the system approaches the triple point temperature.⁷⁴ The fact that signatures of this complex process can be predicted from much higher temperature data with only second-order extrapolation demonstrates the utility of this method.

D. Self-assembling trimer

Finally, we consider the ability of this approach to predict the structure of a self-assembling system. Here we focus on a model trimer system^{76,77} composed of three equally sized, tangent spheres (of diameter σ) arranged in a rigid equilateral triangle. One of the three beads is considered of type “A” and the other two of type “B.” All interactions are given by the force-shifted Lennard-Jones [cf. Eq. (25)], where the cutoff for each pair was set to $r_c^{A,B} = r_c^{B,B} = 2^{1/6}\sigma$ and $r_c^{A,A} = 3\sigma$. Thus, all A–A interactions are attractive, while A–B and B–B interactions are purely repulsive. This system is known to self-assemble into micelles,⁷⁶ so we employ this extrapolation approach to predict two different metrics for the critical micelle concentration (CMC).

In self-assembling systems, the CMC can be estimated directly from the pressure vs. density curve.^{55,59} Below the CMC, the system may be approximated as an ideal gas of monomers, whereas immediately above the CMC, it can be approximated as an ideal gas of micelles. As a result, the pressure will be a linear function of density but will change its slope as the CMC is crossed. A common approach, which we adopt here, is to take the CMC as the density at which these lines intersect, as shown by the dashed gray lines in Fig. 7(a).^{55,59}

Therefore, by extrapolating $\ln \Pi(N)$ across β , we can create a pressure vs. density curve at each temperature by reweighting with respect to chemical potential and thus calculate the CMC as a function of temperature. We performed grand canonical Wang-Landau simulations⁶⁷ with up to 50 trimers in a cubic box of length $L = 16\sigma$ for 10^8 MC trials. Statistics required for extrapolation were collected after the flatness parameter had been reduced 20 times. Implementation details for the geometric cluster algorithm, rigid cluster moves, and configurational bias employed in this work are further described in Refs. 76 and 77.

As an alternative to this thermodynamic approach, we may compute the CMC from the system's structural properties, namely, its cluster size distribution, $\widehat{N}(s)$. It is easier to measure this in the canonical ensemble rather than in the grand canonical ensemble, so we chose to perform simulations in the former. Two trimers were part of the same cluster if the center of their attractive beads of type "A" were within a distance of $4\sigma/3$ of each other. For a micellizing system, $\widehat{N}(s)$ will exhibit a monomer peak at $s = 1$ and another peak at the preferred micelle size. The minimum between these peaks can be used as a cutoff to distinguish between premicellar aggregates and micelles. Therefore, another way to estimate the CMC is to integrate the cluster size distribution from $s = 0$ until the first minimum in $\widehat{N}(s)$.⁵⁸ This should only be done at low densities as the concentration of premicellar aggregates can show a significant decrease with respect to density past the CMC. For this trimer system, it has been shown that the concentration of premicellar aggregates only weakly decreases for $\hat{\rho}\sigma^3 < 0.05$.⁷⁶ We collected the cluster size histogram and the necessary moments during canonical MC simulations using cubic boxes of length $L = 20\sigma$ (containing 400 particles) and $L = 22\sigma$ (containing 500 particles) for which similar results were obtained and the latter is presented here. Simulations again lasted on the order of 10^8 MC trials. Following our extrapolation procedure, each bin in the cluster size distribution histogram can be extrapolated across β . The CMC can then be calculated at each temperature from the resulting distributions.

Figure 7(b) shows a representative first-order extrapolation of the cluster size distribution compared to direct simulations. As mentioned in Sec. III A, it is also possible to combine extrapolation results for a given property from different temperatures analogous to Eq. (27) for $\widehat{R}_g^2$. Figure 7(c) shows the result of this for both the thermodynamic and structural extrapolation methods. Predictions from the two approaches agree well with those obtained from direct simulations⁷⁶ and with each other. The slight deviation that exists between the two metrics is a consequence of their definitions and not from error in the extrapolation method. At high temperatures, the CMC calculated from the pressure curves starts to deviate from the CMC calculated from the cluster size distributions. This is because $T^* = 0.28$ is close to the "critical micelle temperature," above which the equation of state no longer exhibits two different pressure regimes at low densities.⁷⁶

IV. CONCLUSIONS

In summary, we investigated the feasibility of reasonably extrapolating the structural properties of fluids by exploiting fluctuations in extensive properties to construct a Taylor

series expansion in both the canonical and grand canonical ensembles. This follows in the same spirit as previous investigations which assessed this approach's ability to predict thermodynamic properties by extrapolating the fluid's macrostate distribution and underlying moments of extensive properties.^{42–49} Here we find that this approach is also capable of providing a well-informed estimate of structural properties over a broad range of conditions. Although we have limited our investigation to relatively simple fluid models, we were able to obtain accurate extrapolations of structural properties such as the radial and the cluster-size distribution functions germane to complex processes including homogeneous phase separation and nucleation.

Algorithm 1

Computing moments to extrapolate $g_{i,j}(r)$.

```

1: procedure compute rdf moments( $i, j, \xi_1, \xi_2, \xi_u, h_{i,j}^{U\xi_u}(r; N_1, N_2), \text{counts}(N_1, N_2), V$ )
2:   for all ( $N_1, N_2$ ) do
3:     if  $g_{i,j}(r)$  is defined for ( $N_1, N_2$ ) then
4:        $N_{\text{tot}} \leftarrow N_1 + N_2$ 
5:       for all  $r$  do
6:          $\text{binv} \leftarrow 4/3\pi [(r + r/2)^3 - (r - r/2)^3]$ 
7:          $N_{\text{id}} \leftarrow \text{binv} * (N_j - \delta_{i,j})/V$ 
8:          $A \leftarrow h_{i,j}^{U\xi_u}(r; N_1, N_2)/N_{\text{id}}/N_i$ 
9:          $g_{i,j}^{U\xi_u} N_1^{\xi_1} N_2^{\xi_2}(r; N_{\text{tot}}) + = A * N_1^{\xi_1} * N_2^{\xi_2}$ 
10:      end for
11:       $\text{counts}(N_{\text{tot}}) += \text{counts}(N_1, N_2)$ 
12:    end if
13:  end for
14:  for all  $N_{\text{tot}}$  do
15:     $Z(N_{\text{tot}}; \vec{\xi}) \leftarrow g_{i,j}^{U\xi_u} N_1^{\xi_1} N_2^{\xi_2}(r; N_{\text{tot}})/\text{counts}(N_{\text{tot}})$ 
16:  end for
17: end procedure

```

Therefore, this extrapolation method may be used to investigate the thermodynamic and structural properties of multi-component fluids simultaneously. The breadth of conditions over which these extrapolations provide reasonable predictions suggests that this methodology may provide a route to high-throughput screening of coarse-grained fluid behavior. Furthermore, we anticipate that this approach may accelerate the development and validation of coarse-grained models based on, e.g., the potential of mean force which is derived from the radial distribution function.

Acknowledgments

Contribution of the National Institute of Standards and Technology (NIST), not subject to U.S. Copyright. S.J. gratefully acknowledges support from a NIST Summer Undergraduate Research Fellowship, Grant No. 70NANB17H102.

References

1. Chandler D. Introduction to Modern Statistical Mechanics. Oxford University Press; 1987.
2. Hansen J-P, McDonald IR. Theory of Simple Liquids. Elsevier; 1990.
3. McQuarrie DA. Statistical Mechanics. University Science Books; 2000.
4. Blanco MA, Shen VK. Effect of the surface charge distribution on the fluid phase behavior of charged colloids and proteins. *J Chem Phys*. 2016; 145(15):155102. [PubMed: 27782465]
5. Raut AS, Kalonia DS. Pharmaceutical perspective on opalescence and liquid-liquid phase separation in protein solutions. *Mol Pharmaceutics*. 2016; 13:1431–1444.
6. 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. [PubMed: 27314827]
7. Castellanos MM, Clark NJ, Watson MC, Krueger S, McAuley A, Curtis JE. Role of molecular flexibility and colloidal descriptions of proteins in crowded environments from small-angle scattering. *J Phys Chem B*. 2016; 120(49):12511–12518. [PubMed: 27973814]
8. Woldeyes MA, Calero-Rubio C, Furst EM, Roberts CJ. Predicting protein interactions of concentrated globular protein solutions using colloidal models. *J Phys Chem B*. 2017; 121:4756–4767. [PubMed: 28422503]
9. Corbett D, Hebditch M, Keeling R, Ke P, Ekizoglou S, Sarangapani P, Pathak J, Van Der Walle CF, Uddin S, Baldock C, Avendaño C, Curtis RA. Coarse-grained modeling of antibodies from small-angle scattering profiles. *J Phys Chem B*. 2017; 121(35):8276–8290. [PubMed: 28796519]
10. van der Schoot P. Protein-induced collapse of polymer chains. *Macromolecules*. 1998; 31(14):4635–4638.
11. van den Berg B, Wain R, Dobson CM, Ellis RJ. Macromolecular crowding perturbs protein refolding kinetics: Implications for folding inside the cell. *EMBO J*. 2000; 19(15):3870–3875. [PubMed: 10921869]
12. Yin Y, Lu Y, Gates B, Xia Y. Template-assisted self-assembly: A practical route to complex aggregates of monodispersed colloids with well-defined sizes, shapes, and structures. *J Am Chem Soc*. 2001; 123(36):8718–8729. [PubMed: 11535076]
13. Lai Z, Bonilla G, Diaz I, Nery JG, Sujaoti K, Amat MA, Kokkoli E, Terasaki O, Thompson RW, Tsapatsis M, Vlachos DG. Microstructural optimization of a zeolite membrane for organic vapor separation. *Science*. 2003; 300:456–460. [PubMed: 12624179]
14. Pouget E, Dujardin E, Cavalier A, Moreac A, Valéry C, Marchi-Artzner V, Weiss T, Renault A, Paternostre M, Artzner F. Hierarchical architectures by synergy between dynamical template self-assembly and biomineralization. *Nat Mater*. 2007; 6:434–439. [PubMed: 17515916]
15. Russo J, Tanaka H. The microscopic pathway to crystallization in supercooled liquids. *Sci Rep*. 2012; 2:505. [PubMed: 22792437]
16. Mahynski NA, Irick B, Panagiotopoulos AZ. Structure of phase-separated athermal colloid-polymer systems in the protein limit. *Phys Rev E*. 2013; 87(2):022309.
17. Vogel N, Retsch M, Fustin C-A, del Campo A, Jonas U. Advances in colloidal assembly: The design of structure and hierarchy in two and three dimensions. *Chem Rev*. 2015; 115:6265–6311. [PubMed: 26098223]
18. Rimer JD, Tsapatsis M. Nucleation of open framework materials: Navigating the voids. *MRS Bull*. 2016; 41(05):393–398.
19. Allen MP, Tildesley DJ. Computer Simulation of Liquids. Oxford University Press; 1989.
20. Frenkel D, Smit B. Computational Science Series. 2. Vol. 1. Academic Press, Inc; 2001. Understanding Molecular Simulation: From Algorithms to Applications.
21. Baxter RJ. Ornstein–Zernike relation and Percus–Yevick approximation for fluid mixtures. *J Chem Phys*. 1970; 52(9):4559–4562.
22. Davis HT. Statistical Mechanics of Phases, Interfaces and Thin Films. Wiley; 1996.
23. Chapman WG, Jackson G, Gubbins KE. Phase equilibria of associating fluids: Chain molecules with multiple bonding sites. *Mol Phys*. 1988; 65(5):1057–1079.

24. Economou IG. Statistical associating fluid theory: A successful model for the calculation of thermodynamic and phase equilibrium properties of complex fluid mixtures. *Ind Eng Chem Res.* 2002; 41(5):953–962.
25. Lowden LJ, Chandler D. Solution of a new integral equation for pair correlation functions in molecular liquids. *J Chem Phys.* 1973; 59(12):6587–6595.
26. Schweizer KS, Curro JG. Integral-equation theory of the structure of polymer melts. *Phys Rev Lett.* 1987; 58(3):246. [PubMed: 10034880]
27. Ishizuka R, Chong S-H, Hirata F. An integral equation theory for inhomogeneous molecular fluids: The reference interaction site model approach. *J Chem Phys.* 2008; 128:034504. [PubMed: 18205507]
28. Keži B, Perera A. Towards a more accurate reference interaction site model integral equation theory for molecular liquids. *J Chem Phys.* 2011; 135:234104. [PubMed: 22191861]
29. Ferrenberg AM, Swendsen RH. New Monte Carlo technique for studying phase transitions. *Phys Rev Lett.* 1988; 61(23):2635–2638. [PubMed: 10039183]
30. Ferrenberg AM, Swendsen RH. Optimized Monte Carlo data analysis. *Phys Rev Lett.* 1989; 63(12):1195–1198. [PubMed: 10040500]
31. Panagiotopoulos AZ. Monte Carlo methods for phase equilibria of fluids. *J Phys: Condens Matter.* 2000; 12(3):R25–R52.
32. Wu D, Kofke DA. Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation. *J Chem Phys.* 2005; 123(5):054103. [PubMed: 16108627]
33. Wu D, Kofke DA. Phase-space overlap measures. II. Design and implementation of staging methods for free-energy calculations. *J Chem Phys.* 2005; 123(8):084109. [PubMed: 16164284]
34. Shirts MR, Chodera JD. Statistically optimal analysis of samples from multiple equilibrium states. *J Chem Phys.* 2008; 129:124105. [PubMed: 19045004]
35. Tan Z, Gallicchio E, Lapelosa M, Levy RM. Theory of binless multi-state free energy estimation with applications to protein-ligand binding. *J Chem Phys.* 2012; 136:144102. [PubMed: 22502496]
36. Ferguson AL. BayesWHAM: A Bayesian approach for free energy estimation, reweighting, and uncertainty quantification in the weighted histogram analysis method. *J Comput Chem.* 2017; 38:1583–1605. [PubMed: 28475830]
37. Zwanzig RW. High temperature equation of state by a perturbation method. I. Nonpolar gases. *J Chem Phys.* 1954; 22:1420–1426.
38. Lotfi A, Vrabec J, Fischer J. Vapour liquid equilibria of the Lennard-Jones fluid from the NpT plus test particle method. *Mol Phys.* 1992; 76(6):1319–1333.
39. 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.
40. Boda D, Liszi J, Szalai I. Dielectric constant of a Stockmayer fluid along the vapour-liquid coexistence curve. *Mol Phys.* 1995; 85(3):429–434.
41. Boda D, Winkelmann J, Liszi J, Szalai I. Vapour-liquid equilibrium of Stockmayer fluids in applied field. *Mol Phys.* 1995; 87(3):601–624.
42. Escobedo FA. Simulation and extrapolation of coexistence properties with single-phase and two-phase ensembles. *J Chem Phys.* 2000; 113(19):8444–8456.
43. Boda D, Kristóf T, Liszi J, Szalai I. A new simulation method for the determination of phase equilibria in mixtures in the grand canonical ensemble. *Mol Phys.* 2001; 99(24):2011–2022.
44. 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.
45. Boda D, Kristóf T, Liszi J, Szalai I. The extrapolation of the vapourliquid equilibrium curves of pure fluids in the isothermal Gibbs ensemble. *Mol Phys.* 2002; 100(12):1989–2000.
46. 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. [PubMed: 24606348]

47. Mahynski NA, Blanco MA, Errington JR, Shen VK. Predicting low-temperature free energy landscapes with flat-histogram Monte Carlo methods. *J Chem Phys.* 2017; 146(7):074101. [PubMed: 28228029]
48. Mahynski NA, Errington JR, Shen VK. Temperature extrapolation of multicomponent grand canonical free energy landscapes. *J Chem Phys.* 2017; 147:054105. [PubMed: 28789543]
49. Mahynski NA, Errington JR, Shen VK. Multivariable extrapolation of grand canonical free energy landscapes. *J Chem Phys.* 2017; 147:234111. [PubMed: 29272947]
50. Panagiotopoulos AZ, Reid RC. On the relationship between pairwise fluctuations and thermodynamic derivatives. *J Chem Phys.* 1986; 85(8):4650.
51. Debenedetti PG. On the relationship between principal fluctuations and stability coefficients in multicomponent systems. *J Chem Phys.* 1986; 84(3):1778.
52. Chipot C, Pohorille A, editors *Free Energy Calculations: Theory and Applications in Chemistry and Biology.* Springer; 2007.
53. Hummer G, Szabo A. Calculation of free-energy differences from computer simulations of initial and final states. *J Chem Phys.* 1996; 105:2004–2010.
54. Hatch HW, Jiao S, Mahynski NA, Blanco MA, Shen VK. Communication: Predicting virial coefficients and alchemical transformations by extrapolating Mayer-sampling Monte Carlo simulations. *J Chem Phys.* 2017; 147:231102. [PubMed: 29272929]
55. Floriano MA, Caponetti E, Panagiotopoulos AZ. Micellization in model surfactant systems. *Langmuir.* 1999; 15(9):3143–3151.
56. Panagiotopoulos AZ, Floriano MA, Kumar SK. Micellization and phase separation of diblock and triblock model surfactants. *Langmuir.* 2002; 18:2940–2948.
57. Salaniwal S, Kumar SK, Panagiotopoulos AZ. Competing ranges of attractive and repulsive interactions in the micellization of model surfactants. *Langmuir.* 2003; 19:5164–5168.
58. LeBard DN, Levine BG, Mertmann P, Barr SA, Jusufi A, Sanders S, Klein ML, Panagiotopoulos AZ. Self-assembly of coarse-grained ionic surfactants accelerated by graphics processing units. *Soft Matter.* 2012; 8(8):2385–2397.
59. Jusufi A, Panagiotopoulos AZ. Explicit- and implicit-solvent simulations of micellization in surfactant solutions. *Langmuir.* 2015; 31:3283–3292. [PubMed: 25226280]
60. Santos AP, Pękalski J, Panagiotopoulos AZ. Thermodynamic signatures and cluster properties of self-assembly in systems with competing interactions. *Soft Matter.* 2017; 13:8055–8063. [PubMed: 29052681]
61. Harvey SC, Tan RK-Z, Cheatham TE III. The flying ice cube: Velocity rescaling in molecular dynamics leads to violation of energy equipartition. *J Comput Chem.* 1998; 19(7):726–740.
62. Errington JR, Shen VK. Direct evaluation of multicomponent phase equilibria using flat-histogram methods. *J Chem Phys.* 2005; 123:164103. [PubMed: 16268677]
63. Wang J-S, Tay TK, Swendsen RH. Transition matrix Monte Carlo reweighting and dynamics. *Phys Rev Lett.* 1999; 82(3):476.
64. Errington JR. Direct calculation of liquid-vapor phase equilibria from transition matrix Monte Carlo simulation. *J Chem Phys.* 2003; 118(22):9915–9925.
65. 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.
66. Taylor MP. Collapse transition of isolated Lennard-Jones chain molecules: Exact results for short chains. *J Chem Phys.* 2001; 114(14):6472–6484.
67. Hatch HW, Mahynski NA, Shen VK. FEASST: Free energy and advanced sampling simulation toolkit. *J Res Natl Inst Stand Technol.* 2018; 123:123004.
68. For convenience, we generally neglected the right hand side of this equation. This causes the rescaling factor on the radial distribution function needed to satisfy this equation to become independent of system density. For large systems this makes only a small difference, but the most accurate results will be obtained if this equation is strictly obeyed.
69. Ben-Naim A. The Kirkwood-Buff integrals for one-component liquids. *J Chem Phys.* 2008; 128:234501. [PubMed: 18570504]

70. Orkoulas G, Panagiotopoulos AZ. Phase behavior of the restricted primitive model and square-well fluids from Monte Carlo simulations in the grand canonical ensemble. *J Chem Phys.* 1999; 110(3): 1581–1590.
71. Mahynski NA, Shen VK. Multicomponent adsorption in mesoporous flexible materials with flat-histogram Monte Carlo methods. *J Chem Phys.* 2016; 145:174709. [PubMed: 27825240]
72. 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. [PubMed: 15740389]
73. Shen VK, Siderius DW, Krekelberg WP, Hatch HW, editors NIST Standard Reference Database Number 173. National Institute of Standards and Technology; Gaithersburg, MD: 2017. NIST Standard Reference Simulation Website.
74. 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.
75. Mastny EA, de Pablo JJ. Melting line of the Lennard-Jones system, infinite size, and full potential. *J Chem Phys.* 2007; 127:104504. [PubMed: 17867758]
76. Hatch HW, Mittal J, Shen VK. Computational study of trimer self-assembly and fluid phase behavior. *J Chem Phys.* 2015; 142(16):164901. [PubMed: 25933785]
77. Hatch HW, Yang SY, Mittal J, Shen VK. Self-assembly of trimer colloids: Effect of shape and interaction range. *Soft Matter.* 2016; 12(18):4170–4179. [PubMed: 27087490]

APPENDIX: COMPUTING THE MOMENTS OF THE RADIAL DISTRIBUTION FUNCTION

For a binary system, Algorithm 1 describes how to compute the moments matrix, $Z(N_{\text{tot}}; \vec{\xi})$ [cf. Eq. (24)], necessary to extrapolate the radial distribution function for a pair of species, i and j , from the unnormalized histogram of their pairwise distances and energy, $h_{i,j}U^{\xi_u}(r; N_1, N_2)$, the number of times each entry in this histogram was measured over the course of the simulation, counts(N_1, N_2), and the box volume, V . Recall that we consider undefined behavior to occur when either N_i or N_{id} become zero, and thus, they cannot divide into $h_{i,j}U^{\xi_u}(r; N_1, N_2)$ as indicated in the algorithm.

FIG. 1.

Predicted $\widehat{R}_g^2$ for a linear, 5-bead homopolymer as a function of temperature. Theoretical results from Ref. 66 are shown in gray. (a) Simulations were performed at $T^* = 0.70, 0.20, 0.10$, which are labeled with colored dots. Second- and fourth-order extrapolations from these points are shown with solid and dashed lines, respectively. (b) The combined result of fourth-order extrapolations from these simulations using Eq. (27) with $m = 20$ is shown in black. Symbols are otherwise the same as in (a).

FIG. 2.

Radial distribution of the pure component square-well fluid (species 2, Table I) in the canonical ensemble for $\hat{\rho} = N/V \approx 0.20$. Direct simulations were performed at $T^* = 1.35$, 1.05 (blue and red, respectively); then, the former was extrapolated using a second-order expansion to the latter for comparison. Both direct simulations and extrapolations have been rescaled according to Eq. (31).

FIG. 3.

Extrapolated properties of a pure component square-well fluid in the grand canonical ensemble. A simulation at $T^* = 1.35$ was extrapolated to $T^* = 1.20, 1.15, 1.10, 1.05$ (cyan, blue, green, and red, respectively) using second-order terms (points) then compared to direct simulations (solid lines). (a) Radial distribution function of the vapor phase. (b) The binodal curve computed from direct simulations (solid gray lines) and extrapolation (colored points). In the inset, the extrapolated $\ln \Pi(N)$ (green points) is compared with the result from direct simulations (green line) at $T^* = 1.10$; the solid black curve corresponds to the original simulation at $T^* = 1.35$. (c) Radial distribution function of the liquid phase.

FIG. 4.

Radial distribution functions, $g_{1,1}(r)$, $g_{1,2}(r)$, and $g_{2,2}(r)$ for the binary square-well mixture at coexistence depicted in red, green, and blue, respectively. Direct simulations are indicated by solid lines and extrapolations are indicated by points. (a) Extrapolation from $T^* = 1.35$, $\mu_2/\epsilon_{2,2} = 0$ to coexistence at $T^* = 1.10$, $\mu_2/\epsilon_{2,2} = -1$ for the vapor phase. (b) A portion of the phase diagram at $T^* = 1.10$. The coexistence point corresponding to (a) and (c) is circled in red. $\langle y_1 \rangle$ and $\langle x_1 \rangle$ are the mole fraction of species 1 in the vapor and liquid phases, respectively. (c) Extrapolation from $T^* = 1.35$, $\mu_2/\epsilon_{2,2} = 0$ to coexistence at $T^* = 1.10$, $\mu_2/\epsilon_{2,2} = -1$ for the liquid phase.

FIG. 5.

Structural and thermodynamic properties of a Lennard-Jones fluid using second-order extrapolation from $T^* = 1.20$ (points) and direct simulations (solid lines) in the grand canonical ensemble. (a) Radial distribution function of the vapor phase at coexistence. (b) Coexistence curve computed from direct simulations compared to extrapolation, where the reference state is indicated by black squares. (c) Radial distribution function of the liquid phase at coexistence.

FIG. 6.

Extrapolated grand canonical cluster size distributions, $\langle N(s) \rangle$, along the coexistence curve for the Lennard-Jones fluid we investigated. Solid lines correspond to $\langle N(s) \rangle$ computed from direct simulations, whereas points represent the results extrapolated from $T^* = 1.20$ using up to second-order terms. (a) $\langle N(s) \rangle$ for the vapor phase at coexistence, from $T^* = 0.85$ to $T^* = 1.20$ in increments of $\delta T^* = 0.05$. (b) $\langle N(s) \rangle$ for the liquid phase at coexistence.

FIG. 7.

Computing the CMC for a model trimer. (a) Calculation of pressure vs. density curves via second-order extrapolation of $\ln \Pi(N)$ from $T^* = 0.22$ to lower T^* (points), compared to direct simulations (solid lines). (b) $\widehat{N}(s)$ extrapolated from $T^* = 0.30$ to lower T^* (dashed-dotted lines), compared to direct simulations (solid lines). (c) CMCs computed by integrating the extrapolated $\widehat{N}(s)$ curves (structural method, blue points) compared to those obtained from the extrapolated $\ln \Pi(N)$ (thermodynamic method, red points). Direct simulations used for these extrapolations are indicated by black triangles. The red and blue lines indicate the results of each method obtained from direct simulations in Ref. 76.

TABLE I

Parameters describing the two square-well fluids and their cross interaction.

i	j	$\epsilon_{i,j}$	$\sigma_{i,j}$	$\lambda_{i,j}$
1	1	1.20	1.50	1.33
1	2	1.10	1.25	1.40
2	2	1.00	1.00	1.50