Modeling experimental data in a Monte Carlo simulation

Gregory C. Rutledge
Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139
(Received 28 February 2000; revised manuscript received 23 October 2000; published 25 January 2001)

A method is presented for modeling the structure of disordered media consistent with a set of experimental observations, such as scattering data. The data are incorporated into a conventional semigrand canonical Monte Carlo simulation by introducing a generalized, polydisperse composition space. This approach improves upon previous reverse Monte Carlo procedures in that thermodynamic consistency is retained. By way of example, the structure of a Lennard-Jones fluid is derived solely from radial distribution data.

DOI: 10.1103/PhysRevE.63.021111 PACS number(s): 82.20.Wt, 05.10.Ln, 61.20.Ja

The so-called ''inverse problem'' is fundamental to the study of fluids [1]. It entails the deduction of an interaction potential from measurements of structure, such as scattering data. Several methods aimed at solving a part of this problem, the determination of detailed structure from experimental data, have appeared in recent years [2–5]. One reason for the popularity of these methods is that conventional Monte Carlo and molecular dynamics simulations often fail to reproduce known information about the system within acceptable accuracy, perhaps due to limitations in the method or in the interaction potential available, and offer little guidance when agreement between the experimental observation and its simulated equivalent is unsatisfactory. There is a need, then, to devise modeling or simulation methods which are consistent with the statistical mechanics of fluids or glasses, but which correctly reproduce known observations about the system. One method that accomplishes the latter is the reverse Monte Carlo (RMC) method [2], which has found applications in the study of liquids, glasses, polymers, and even imperfect crystals; for reviews, see [6–8]. RMC replaces the role of interaction potential in the conventional Metropolis Monte Carlo (MMC) method with a measure of error in a configuration relative to the observed structure factor, while the role of temperature is replaced by experimental uncertainty. However, the thermodynamic interpretation is lost, thus raising questions about uniqueness and disagreement as to how to improve upon the original procedure [7,9,10].

The goal of this article is to show that information from experiments can be incorporated simply and directly into a simulation through the introduction of a generalized, polydisperse composition space. This permits a general approach to modeling experimental data or other information about the system (i.e., not limited to scattering data), which retains the character of a thermodynamic system and which reverts to a conventional statistical mechanical procedure when the experimental data are absent. Experimental data can be modeled in the absence of an accurate interaction potential (illustrated here by the Lennard-Jones fluid example), but the inclusion of both data and interaction potential within a single simulation can also be handled, without adjustable parameters. Furthermore, thermal fluctuations are properly retained, allowing ensembles consistent with a canonical simulation to be realized. Finally, the general structure of the method suggests a wide range of possible uses, including the modeling of structure and orientation in materials that reflect the processing history. Data may be of either theoretical or empirical origin, creating the opportunity to refine structural models, to test one experimental result against another, or to identify redundant measurements.

In the physics of fluids, a polydisperse substance is a mixture of infinitely many components. The concept was first defined by de Donder [11], and has been invoked extensively in recent years in the theory [12–15] and simulation [16–18] of phase equilibria. Following Briano and Glandt [19], we begin by assuming that the components of the fluid interact through a pair-wise additive potential given by $\phi_{ij}=\phi(\mathbf{r}_i,\mathbf{r}_j,I_i,I_j)$, where I is a random, component-designating variable, distributed according to $p(I)$ such that $p(I)$ is normalized and everywhere non-negative. For a mixture of c components in the limit $c\to\infty$, the grand canonical partition function may be written

$$\Xi[V,T,\mu(I)]=1+\sum_{N=1}^{\infty}\frac{1}{N!}\int_{I_1}\cdots\int_{I_N}\prod_{i=1}^{N}\left[\frac{q_{\mathrm{int}}(I_i)}{\Lambda^3(I_i)}e^{\beta\mu(I_i)}\right]\times\int_{\mathbf{r}_1}\cdots\int_{\mathbf{r}_N}e^{-\beta U}\prod_{i=1}^{N}(d\mathbf{r}_i dI_i). \tag{1}$$

N is the number of particles comprising the system, each with position $\mathbf{r}_i$ and component designation I_i. $\mu(I)$, $q_{\mathrm{int}}(I)$, and $\Lambda(I)$ are the chemical potential, internal partition function, and de Broglie wavelength, respectively, for a particle of type I. U is the potential energy of the system, $U=\Sigma_{ij>i}\phi_{ij}$, and $\beta=(k_BT)^{-1}$. We define the activity $a(I)=\exp[\beta\mu(I)]$, residual activity $a^*(I)=a(I)/p(I)$, and configurational integral $Z_N=\int\cdots\int\exp(-\beta U)\Pi_i d\mathbf{r}_i$. It is also convenient to define a reference μ_r and transform to the isomolar semigrand canonical ensemble, where N is fixed but composition is not. The resulting partition function is

$$Y_N[N,V,T,a^*(I)/a_r^*]=\frac{Z_N}{N!}\left[\frac{q_{\mathrm{int}}}{\Lambda^3}\right]^N\int_{I_1}\cdots\int_{I_N}\prod_{i=1}^{N}\left[\frac{a^*(I_i)p(I_i)}{a_r^*p_r}\right]\prod_{i=1}^{N}dI_i. \tag{2}$$

In Eq. (2), we have factored out the contributions from $q_{\mathrm{int}}(I)/\Lambda^3(I)$, which are independent of I in the examples to follow. Taking the functional derivative of Y_N with respect

to $a(I)$, one obtains the following important result for the probability distribution $p(I)$ of any polydisperse parameter [19]:

$$p(I)=\frac{1}{Y_N}\int_{I_2}\cdots\int_{I_N}\frac{Z_N}{N!}\left[\frac{q_{\rm int}}{\Lambda^3}\right]^N\left[\prod_{i=1}^{N}\frac{a^*(I_i)}{a_r^*}\frac{p(I_i)}{p_r}\right]\prod_{i=2}^{N}dI_i\,. \tag{3}$$

With regard to Eq. (3), several points are noteworthy. Speciation of the system is arbitrary, so long as each component is stipulated in a definite way and $\Sigma_{i=1}^{N}\delta(I_i-I)=n(I)$, where $n(I)$ is the number of ''particles'' of type I [20]. The species label I is not limited to indexing of chemically distinct entities; as early as 1949, Onsager introduced the artifice of treating (otherwise identical) anisometric particles of different orientation as being of different kind in order to evaluate solution imperfection [21]. Thus a system that is monodisperse in one sense, e.g., a crystal of identical Lennard-Jones particles, may at the same time be polydisperse if speciated in a different manner, e.g., by treating the phonons as the ''particles.'' Furthermore, Eq. (3) does not require that $a(I_i)$ be independent of the values $I_{j\neq i}$. In general, speciation may involve more than one parameter, resulting in the replacement of I by a vector $\mathbf{I}$ of component labels. If a set of observations can be formulated as $p(\mathbf{I})$, then the equations above may be used to construct a Monte Carlo procedure by which a molecular model that reproduces these observations is simulated directly.

Unlike a phase equilibrium calculation, whose objective is to determine coexisting compositions $p(I)$, given knowledge of the chemical potentials, modeling experimental data involves determining the chemical potential distribution $\mu(I)$, or equivalently $a^*(I)$, responsible for the observed data $p(I)$. At first glance, it might appear reasonable to construct a configuration that satisfies $p(I)$, and then perform a canonical simulation in which $p(I)$ is conserved. However, in many real applications, it is not a trivial matter to construct even an initial trial configuration that satisfies the desired distribution $p(I)$. Furthermore, imposing an experimentally observed distribution on each configuration *individually* (a ''quenched composition'') is too restrictive and usually not justified by the experimental data. For these reasons, the semigrand ensemble simulations suggested here are preferable, both in principle and in practice.

In a Monte Carlo simulation of polydisperse components, one satisfies detailed balance by accepting a trial configuration according to the following importance criterion (m is a pseudorandom number between 0 and 1):

$$m\leqslant\min\left\{1,\frac{\left\{e^{-\beta U}\prod_{i=1}^{N}a^*(I_i)p_{\rm tar}(I_i)\right\}_{\rm new}}{\left\{e^{-\beta U}\prod_{i=1}^{N}a^*(I_i)p_{\rm tar}(I_i)\right\}_{\rm old}}\right\}. \tag{4}$$

$p_{\rm tar}(I)$ is the target distribution for the parameter I. If $p_{\rm tar}(I)$ is independent of I, then $a^*(I)=a_r=1$, and one recovers the conventional Metropolis Monte Carlo procedure. If the I values of the particles are independent, then again $a^*(I)=a_r=1$, and one immediately obtains a simulation that samples the desired distribution $p_{\rm tar}(I)$ in a single run. The more general and interesting case occurs when the I value of each particle is coupled to that of the other particles of the system. In this case, $a^*(I)$ is an unknown function, to be determined iteratively. One suitable procedure, employed here, uses a trial function $a_k^*(I)$ to obtain the estimate $p_k(I)$ from a short simulation. An improved $a_k^*(I)$ may then be obtained using

$$a_{k+1}^*(I)=a_k^*(I)\left[\frac{p_{\rm tar}(I)}{p_{r,{\rm tar}}}\frac{p_{r,k}}{p_k(I)}\right]^{\alpha}, \tag{5}$$

and the process repeated until $p_k(I)$ converges to $p_{\rm tar}(I)$ for all I to within statistical uncertainty. k indexes the iteration. The exponent $0<\alpha<1$ is a damping factor to improve convergence. $a_0^*(I)=1$ has been used here, although other choices are possible.

By way of illustration, we consider the standard problem of determining the structure of a Lennard-Jones fluid from scattering data. Identifying the polydispersity index I with the scattering vector q, one first observes that the original formulation of RMC satisfies the more general semigrand ensemble Monte Carlo formulation presented here, with an activity function for reciprocal space, $a(q)$, defined as

$$a(q)=\frac{1}{N_A}\exp\left(-\frac{1}{2}\left[\frac{A(q)-A_{\rm tar}(q)}{\sigma(q)}\right]^2\right). \tag{6}$$

$A(q)$ is the structure factor and $\sigma(q)$ is the experimental uncertainty in the data $A_{\rm tar}(q)$. N_A is a constant required to normalize $p(q)$, but its exact value is inconsequential to implementing the simulation. Defining the activity directly in this manner eliminates the thermodynamic temperature from the simulation, replacing it with a nonthermodynamic quantity $\sigma(q)$. This has the undesirable consequence of altering the fluctuations in the ensemble from their thermodynamic values—as the experimental uncertainty decreases, the effective simulation temperature decreases, regardless of the true experimental temperature.

Instead, we associate $p(\mathbf{I}_i)$ with a normalized N-particle distribution function $g^{(N)}(r_{ij,j\neq i})$ specific to particle i, i.e., each particle is speciated according to its position relative to the rest of the system. Invoking the approximation $g^{(N)}(r_{ij,j\neq i})\approx\Pi_j g(r_{ij})$, where $g(r)$ is the radial distribution function, one can rewrite Eq. (3) with $p(I)$ equal to $(\rho/N)g(r)$ and the product taken over all pairs. $g(r)$ is readily obtained by inverse Fourier transformation of $A(q)$, truncation issues notwithstanding, and contains in principle the same information. Grouping contributions of similar r, the number of pairwise interactions between r and $r+\delta r$ is $n(r)=\rho N(2\pi r^2\delta r)g(r)$, resulting in the following inequality for importance sampling [22]:

$$m\leqslant\min\left\{1,e^{-\beta\Delta U}\prod_{r=0}^{r_{\rm cut}}\left[a_k^*(r)g_{\rm tar}(r)\right]^{\Delta n(r)}\right\}, \tag{7}$$

where $g_{\rm tar}(r)$ is the target (e.g., experimental) distribution function, $a_k^*(r)$ is the kth estimate of the residual activity

TABLE I. Result of Lennard-Jones fluid simulations. $\langle U\rangle/N$ is the average potential energy per particle and $\langle P\rangle$ is the pressure (not including long range corrections). $\sigma(X)=\langle X^2\rangle-\langle X\rangle^2$. MMC: Metropolis Monte Carlo. RMC: reverse Monte Carlo of Ref. [2]. SGMC: method based on the semigrand canonical ensemble.

	$\rho^*=0.84,\ T^*=0.75$	$\rho^*=0.55,\ T^*=1.35$
$\langle U/N\rangle_{\rm MMC}\pm\sigma(U/N)$	-6.044 ± 0.048	-3.727 ± 0.053
$\langle U/N\rangle_{\rm RMC}\pm\sigma(U/N)$	-4.270 ± 0.560	-1.962 ± 0.643
$\langle U/N\rangle_{\rm SGMC}\pm\sigma(U/N)$	-6.07 ± 0.059	-3.710 ± 0.061
$\langle P\rangle_{\rm MMC}\pm\sigma(P)$	0.20 ± 0.24	0.36 ± 0.17
$\langle P\rangle_{\rm RMC}\pm\sigma(P)$	6.833 ± 2.00	4.73 ± 1.51
$\langle P\rangle_{\rm SGMC}\pm\sigma(P)$	-0.07 ± 0.32	0.39 ± 0.19
$\chi^2_{g,\rm RMC}$	0.0012	0.0014
$\chi^2_{g,\rm SGMC}$	0.0011	0.0002

function, and $\Delta U=U_{\rm new}-U_{\rm old}$ and $\Delta n(r)=n_{\rm new}(r)-n_{\rm old}(r)$ are the differences in energy and number of pairwise interactions at distance r, respectively, between successive configurations. The resulting procedure in this case is very similar to the empirical potential structure refinement method [4] and to the method of Lyubartsev and Laaksonen for determining ''effective potentials'' [23]. Using experimental data with $a^*(r)=1$ introduces a mean field potential of the form $-\beta^{-1}\ln g_{\rm tar}(r)$. Successive updates of the estimate of $a^*(r)$ converge on an estimate of the effective potential $\phi_{\rm eff}(r)$:

$$\phi_{\rm eff}(r)=\lim_{k\to\infty}\left\{\phi_0(r)-k_BT\ln\left[g_{\rm tar}(r)\,\frac{a_k^*(r)}{a_r}\right]\right\}. \qquad (8)$$

As an illustration, a 256-particle Lennard-Jones (LJ) fluid with periodic boundaries was simulated at several values of reduced density $\rho^*=\rho\sigma^3$ and temperature $T^*=k_BT/\epsilon$. A hard sphere potential was imposed for $r<0.7\sigma$ and the interaction potential truncated at $r_{\rm cut}=2.5\sigma$, consistent with previous studies [24]. The particles were initially placed on the fcc lattice and equilibrated for 2048 Monte Carlo cycles, followed by 2048 cycles at each iteration for $a^*(r)$. Statistical uncertainties of 0.01 and 0.001 for $\langle U\rangle/N$ and $g(r)$, respectively, were estimated by the ''blocking'' method [25]. For each state point, a conventional MMC simulation was performed first to obtain target distributions $A_{\rm tar}(q)=A^{\rm MMC}(q)$ and $g_{\rm tar}(r)=g^{\rm MMC}(r)$, discretized between 0 and $r_{\rm cut}$. The average energies and pressures are reported in Table I and are in good agreement with previous simulations [24]. For the RMC calculation, the potential was turned off and the simulation repeated using the activity function defined by Eq. (6), with $\sigma(q)=0.001$. For the semigrand Monte Carlo method presented here (SGMC), Eq. (7) was used and convergence of $a^*(r)$ judged when the mean squared error in $g(r)$, $\chi_g^2=(1/n)\Sigma[g(r_i)-g^{\rm MMC}(r_i)]^2$, was less than the statistical uncertainty in $g(r)$. $\frac{1}{4}<\alpha<\frac{1}{2}$ yielded the best convergence. Figure 1 shows results obtained for $\rho^*=0.84$ and $T^*=0.75$, close to the triple point for the LJ fluid [26]; this state point represents a fairly stringent test of the robustness of the Monte Carlo procedure. The $g(r)$ shown was obtained by the method described here. Also plotted are the errors $\Delta g(r)=g(r)-g^{\rm MMC}(r)$ for both RMC and SGMC methods. These are of comparable magnitude and within the statistical uncertainty of $g^{\rm MMC}(r)$. The relative error in $g^{\rm SGMC}(r)$ is largest near $r=r_{\rm cut}$, probably due to truncation. Using the known LJ potential, we report in Table I the potential energies and pressures that would be obtained for the ensembles produced by RMC and SGMC near the triple point and at $\rho^*=0.55$, $T^*=1.35$. For RMC, the error in $\langle U\rangle/N$ and $\langle P\rangle$ is significant, presumably due to the treatment of fluctuations by this method. The agreement between MMC and SGMC methods in $g(r)$, the thermodynamic averages $\langle U\rangle/N$ and $\langle P\rangle$, and the fluctuations $\sigma(U/N)$ and $\sigma(P)$, is very good. Lastly, Fig. 1 shows the final distribution obtained for $a^*(r)$. This may be compared to a ''true''

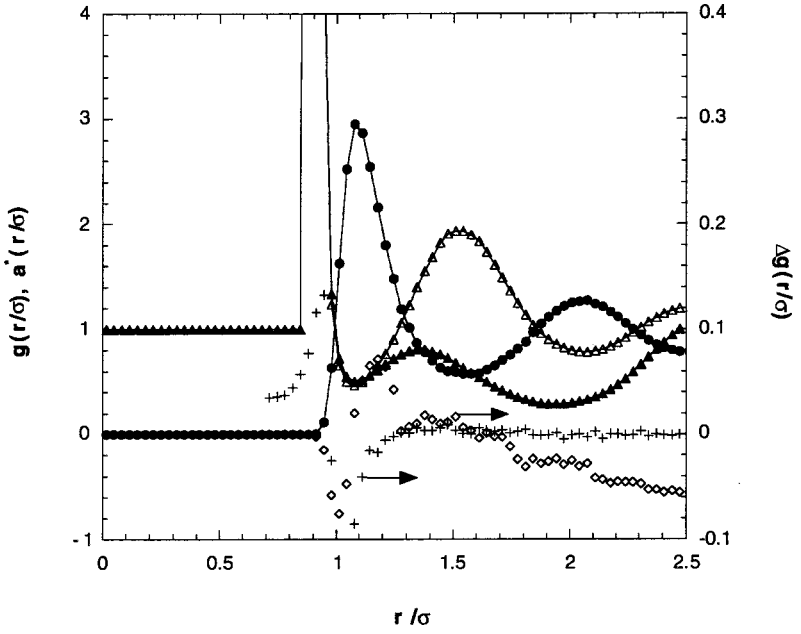

FIG. 1. Simulation results for Lennard-Jones fluid at $\rho^*=0.84$, $T^*=0.75$. Filled circles, radial distribution function $g^{\rm SGMC}(r)$; (+), error $\Delta g(r)=g^{\rm RMC}(r)-g^{\rm MMC}(r)$ for RMC method; open diamonds, error $\Delta g(r)=g^{\rm SGMC}(r)-g^{\rm MMC}(r)$ for SGMC method; filled triangles, residual activity profile $a^*(r)$ from SGMC simulation [for $g^{\rm MMC}(r)=0$, the magnitude of $a^*(r)$ is unimportant, and set arbitrarily to 1.0]; open triangles, ''true'' residual activity profile, $a^*_{\rm true}(r)=\exp[-\beta\phi(r)]/g(r)$.

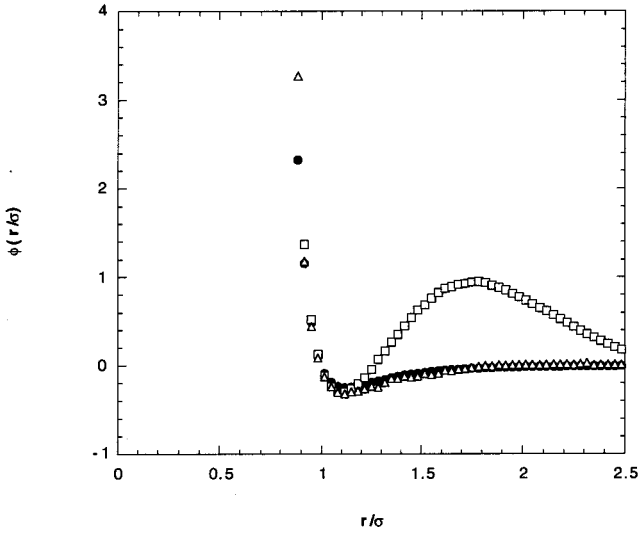

FIG. 2. Comparison of true and effective potentials estimated for the Lennard-Jones fluid at two different state points. Filled circles, true potential, $\phi_{\text{true}}(r)$; open squares, effective potential $\phi_{\text{eff}}(r)$ obtained at $\rho^*=0.84$, $T^*=0.75$; open triangles, effective potential obtained at $\rho^*=0.55$, $T^*=1.35$.

$a^*(r)$ obtained from knowledge of the LJ interaction potential: $a^*_{\text{true}}(r)=\exp[-\beta\phi(r)]/g_{\text{tar}}(r)$, also shown in Fig. 1. At this state point, $a^*(r)$ differs significantly from $a^*_{\text{true}}(r)$ for r values beyond the first-neighbor peak in $g(r)$, which is indicative of the insensitivity of $g(r)$ to the long range portion of $a^*(r)$ at such a high density; nevertheless, this does not detract from the quality of $g(r)$ or the thermodynamic quantities estimated. One anticipates that $a^*(r)$ may be a function of density and temperature. Figure 2 shows estimates of $\phi_{\text{eff}}(r)$ obtained from $a^*(r)$ by Eq. (8) for the LJ fluid at the two different state points, along with the original potential used to generate $g_{\text{tar}}(r)$ in each case. At $\rho^*=0.55$, $T^*=1.35$, the agreement between $\phi_{\text{true}}(r)$ and $\phi_{\text{eff}}(r)$ obtained is already quite good. A fuller investigation of $a^*(r)$ and $\phi_{\text{eff}}(r)$ is warranted, but beyond the scope of this communication.

The general utility of this procedure goes beyond the classical inverse problem. Many problems in materials physics involve the study of systems that are not in thermodynamic equilibrium, where even knowledge of the true interaction potential is generally insufficient to simulate their structure. The case of long chain molecules, where the longest relaxation time of the molecule scales with the length of the chain, is a particularly egregious example; polymers are commonly processed in the melt and then partially crystallized or quenched below the glass transition point, thereby fixing the structure of the material in an oriented, metastable state. It is this metastable state for which structure-property relationships are often desired. Experimental data on the orientation distribution of individual bonds of the chain in the real material may be obtained, for example, by solid state NMR spectroscopy; however, calculating properties requires a model for the conformations of the long chain molecules derived from this bond-level information. By equating I with $\cos\Theta$, the azimuthal orientation angle for the bonds with respect to the material's principal axis of anisotropy, a simulation of long chains that reproduces the experimentally observed bond orientation distribution can be obtained. We have tested this approach with some relatively simple chain models. Treating the chains as freely jointed with no excluded volume interactions (equivalent to a random walk) is particularly straightforward, since each bond of the chain is mutually independent: $a^*(\cos\Theta)=1$. Slightly more complicated models such as the freely rotating chain or chains with torsion potentials are also tractable, requiring only that one iterate to obtain $a^*(\cos\Theta)$. From the distribution of conformations obtained by simulation, one can deduce the extent of stretching and reorientation of the chain conformations away from their undeformed state, important factors that affect the properties of polymeric materials.

Funding for this work was provided in part by the National Science Foundation (Grant No. CTS-9457111).

[1] D. Levesque, J. J. Weis, and L. Reatto, Phys. Rev. Lett. **54**, 451 (1985).
[2] R. L. McGreevy and L. Pustzai, Mol. Simul. **1**, 359 (1988).
[3] D. A. Keen and R. L. McGreevy, Nature (London) **344**, 423 (1990).
[4] A. K. Soper, Chem. Phys. **202**, 295 (1996).
[5] Y. Rosenfeld and G. Kahl, J. Phys.: Condens. Matter **9**, L89 (1997).
[6] R. L. McGreevy and M. A. Howe, Annu. Rev. Mater. Sci. **22**, 217 (1992).
[7] R. L. McGreevy, Nucl. Instrum. Methods Phys. Res. A **354**, 1 (1995).
[8] R. L. McGreevy and L. Pustzai, Electrochim. Acta **43**, 1349 (1998).
[9] G. Tóth and A. Baranyai, J. Chem. Phys. **107**, 7402 (1997); G. Toth, L. Pusztai, and A. Baranyai, *ibid.* **111**, 5620 (1999).
[10] F. L. Da Silva *et al.*, J. Chem. Phys. **109**, 2624 (1998); **111**, 5622 (1999).
[11] T. de Donder, *L'Affinité*, 2nd ed. (Gauthier-Villars, Paris, 1931).
[12] P. Sollich and M. E. Cates, Phys. Rev. Lett. **80**, 1365 (1998).
[13] P. B. Warren, Phys. Rev. Lett. **80**, 1369 (1998).
[14] R. M. L. Evans, D. J. Fairhurst, and W. C. K. Poon, Phys. Rev. Lett. **81**, 1326 (1998).
[15] R. M. L. Evans, Phys. Rev. E **59**, 3192 (1999).
[16] D. A. Kofke and E. D. Glandt, J. Chem. Phys. **87**, 4881 (1987); Mol. Phys. **64**, 1105 (1988).
[17] M. R. Stapleton, D. J. Tildesley, and N. Quirke, J. Chem. Phys. **92**, 4456 (1990).
[18] P. G. Bolhuis and D. A. Kofke, Phys. Rev. E **54**, 634 (1996).
[19] J. G. Briano and E. D. Glandt, J. Chem. Phys. **80**, 3336 (1984).
[20] T. Morita and K. Hiroike, Prog. Theor. Phys. **25**, 537 (1961).
[21] L. Onsager, Ann. N. Y. Acad. Sci. **51**, 627 (1949).
[22] Alternatively, the importance criterion can be computed as the

product over all pairs of particles, using interpolation to estimate values of $a^*(r_{ij})$ and $g_{\mathrm{tar}}(r_{ij})$ for particles i and j.

[23] A. P. Lyubartsev and A. Laaksonen, Phys. Rev. E **52**, 3730 (1995).

[24] J. J. Nicolas *et al.*, Mol. Phys. **37**, 1429 (1979).

[25] H. Flyvbjerg and H. G. Petersen, J. Chem. Phys. **91**, 461 (1989).

[26] J. P. Hansen and L. Verlet, Phys. Rev. **184**, 151 (1969).