Determination of the chemical potential using energy-biased sampling

R. Delgado-Buscalioni[a)]
Departamento Ciencias y Técnicas Fisicoquímicas, Facultad de Ciencias, Universidad Nacional de Educación a Distancia (UNED), Paseo Senda del Rey 9, Madrid 28040, Spain

G. De Fabritiis[b)] and P. V. Coveney[c)]
Centre for Computational Science, Department of Chemistry, University College London, 20 Gordon Street, WC1H 0AJ London, United Kingdom

(Received 11 May 2005; accepted 22 June 2005; published online 10 August 2005)

An energy-biased method to evaluate ensemble averages requiring test-particle insertion is presented. The method is based on biasing the sampling within the subdomains of the test-particle configurational space with energies smaller than a given value freely assigned. These *energy wells* are located via unbiased random insertion over the whole configurational space and are sampled using the so-called Hit-and-Run algorithm, which uniformly samples compact regions of any shape immersed in a space of arbitrary dimensions. Because the bias is defined in terms of the energy landscape it can be *exactly* corrected to obtain the unbiased distribution. The test-particle energy distribution is then combined with the Bennett relation for the evaluation of the chemical potential. We apply this protocol to a system with relatively small probability of low-energy test-particle insertion, liquid argon at high density and low temperature, and show that the energy-biased Bennett method is around five times more efficient than the standard Bennett method. A similar performance gain is observed in the reconstruction of the energy distribution. © *2005 American Institute of Physics*. [DOI: 10.1063/1.2000244]

I. INTRODUCTION

The chemical potential is a central quantity underpinning many physical and chemical processes, such as phase equilibria, osmosis, thermodynamic stability, binging affinity, and so on.[1] However, its evaluation by computer simulation is more complicated and time consuming than for other intensive thermodynamic quantities, such as the pressure P or temperature T. While P and T can be evaluated from averages over mechanical properties of molecules (forces, velocities, and positions), the chemical potential is a thermal average and therefore it requires sampling of the phase space of the system. Indeed, computing the chemical potential is a special case of the more general problem of computing a free-energy difference A_1-A_0 between two states (labeled as 0 and 1), a problem for which the inherent difficulty is well understood.[1–4] Free-energy perturbation (FEP) is an important category of methods for free-energy calculation; we refer to the recent works by Lu *et al.*[1] and by Shirts and Pande[5] for review and comparisons. As explained by Lu *et al.*,[1] the general working equation for FEP methods can be cast as

$$\exp[-\beta(A_1-A_0)]=\frac{\langle w(u)\exp[-\beta u/2]\rangle_0}{\langle w(u)\exp[-\beta u/2]\rangle_1}, \tag{1}$$

with $\beta=1/k_BT$ and $u\equiv U_1-U_0$ the energy difference between both systems; k_B is the Boltzmann constant. The angular brackets denote ensemble averages performed on the system labeled by the subscript "0" or "1." The weighting function $w(u)$ is arbitrary and differs for each method introduced in the literature.

The chemical potential is the free-energy difference between two thermodynamic states differing by the presence of a single molecule. In other words, the chemical potential is A_1-A_0, where $A_1=A(N+1,V,T)$ and $A_0=A(N,V,T)$. Here $A(N,V,T)$ is the Helmholtz free energy of the system which depends on the number of molecules N, the volume V, and temperature of the system. In order to express the averages of Eq. (1) in terms of one-dimensional integrals of the energy difference u one can then introduce the following distribution functions.[6]

$$f(u)=\int\langle\delta(u-U_1+U_0)\rangle_0V^{-1}d\mathbf{r}, \tag{2}$$

$$g(u)=\langle\delta(u-U_1+U_0)\rangle_1, \tag{3}$$

where $\delta(.)$ is the Dirac delta function. In Eq. (2), $U_1=U_1(\mathbf{R}^N,\mathbf{r})$, where $\mathbf{R}^N$ is the configuration of the first N molecules and $\mathbf{r}$ denotes the configuration of the $N+1$ molecule. Note that in Eq. (2) the $N+1$ molecule acts as a "test molecule" which probes the system 0 (i.e., the system with N molecules), but does not interact with it. Therefore $f(u)$ is the probability density of the N-molecule ensemble increasing in potential energy by an amount u if this test molecule was randomly inserted into the ensemble. Conversely, $g(u)$ is the probability density of the $(N+1)$-molecule ensemble decreasing in potential energy by an amount u if a randomly selected real molecule was removed from the ensemble.

[a)]Electronic mail: r.delgado-buscalioni@ucl.ac.uk
[b)]Electronic mail: g.defabritiis@ucl.ac.uk
[c)]Electronic mail: p.v.coveney@ucl.ac.uk

From Eqs. (1)–(3) an expression for the excess chemical potential $\mu = A_1 - A_0 - \mu_{\mathrm{id}}$ (where μ_{id} is the ideal-gas chemical potential[3]) can be derived in terms of the f and g distributions,[1,6,7]

$$\exp(\beta\mu) = \frac{\int w(u)g(u)du}{\int w(u)f(u)\exp(-\beta u)du}. \tag{4}$$

A good choice of the weighting function $w(u)$ is a key for the efficiency of the method. For instance, the Widom method[2,3] ($w(u)=1$) is known to provide very poor convergence at large densities. The Widom method is a single-stage FEP, meaning that sampling is only performed in the reference system 0 [i.e., in the f distribution, see Eq. (4)]. As discussed by Lu *et al.*,[1] multiple staging provides much better efficiency. The efficiency is generally defined as the reciprocal of the product of the variance of the estimator multiplied by its cost n_{cost} (that is, the total number of energy evaluations performed by the algorithm)

$$\varepsilon = (n_{\mathrm{cost}}\mathrm{Var}[\beta\mu])^{-1}. \tag{5}$$

Bennett[8] showed that the variance of Eq. (4) is minimized if the weighting function is $w(u) = \mathcal{F}[\beta(u-c)]$, where $\mathcal{F}(x) = 1/(1+\exp(x))$ is the Fermi function and c is an arbitrary constant. The Bennett estimator is then

$$\beta\mu = \ln\left(\frac{\langle \mathcal{F}[-\beta(u-c)]\rangle_g}{\langle \mathcal{F}[\beta(u-c)]\rangle_f}\right) + \beta c, \tag{6}$$

where the subscripts g and f indicate (simple) averages over the distributions $g(u)$ and $f(u)$. The value of c providing the minimum variance and maximum overlap is $c=\mu$ and to evaluate μ using the optimum $c(=\mu)$ one requires to use a self-consistent procedure, iterating the value of c in Eq. (6) and resetting $c=\mu$ until $\langle \mathcal{F}[-\beta(u-c)]\rangle_g = \langle \mathcal{F}[\beta(u-c)]\rangle_f$. In practice, this step only requires a small number of iterations. Recent publications[1,5] demonstrate that the Bennett method remains the best general method to compute the chemical potential for many applications.

Note that the Bennett method is a two-stage FEP and therefore it also requires sampling of the system 1. In the case of the determination of the chemical potential this system has $N+1$ molecules and $g(u)$ is obtained from its single-molecule energy distribution. However this extra requirement is not really a drawback. Lu *et al.*[1] showed that, provided $N > O(100)$, the g average can be evaluated in the same simulation as is used to sample the f distribution (system 0) without any noticeable loss in accuracy. The g distribution (constructed from the energy of the real particles) is thus a by-product of the simulation so the average $\langle \mathcal{F} \rangle_g$ does not demand any extra computational cost.

Another group of methods for the determination of the chemical potential are based on *biased* instead of uniform sampling. In particular, cavity-biased methods first select spherical cavities of minimum radius R_c (a free parameter) in which to insert the test molecule. This accelerates the evaluation of the ensemble average in dense phases because the low-energy configurations of the test molecule (with large Boltzmann factors) are usually located in larger cavities with less steric hindrance. Variations of this method have been proposed by several authors; these include the cavity insertion Widom method (CIW) by Jedlovszky and Mezei,[9] the excluded volume map sampling by Deitrick *et al.*,[6] and the method proposed by Pohorille and Wilson.[10] The cavities are located by a grid search over the whole simulation cell. A cavity center is assigned at each grid point whose distance to the closest particle is greater than R_c. In order to correct the bias introduced in sampling only inside the cavities one also has to calculate the probability of finding a cavity, which is obtained in the same grid-search step. A drawback of the cavity-biased method is that it is only indirectly related to the test-particle energy via the excluded volume. This fact introduces a certain inaccuracy in the estimation of the chemical potential, as it can depend on the value of the cavity radius R_c selected. For instance, the CIW has recently been used to calculate the chemical potential of several species across a lipid bilayer.[9] As a test calculation the authors estimated the chemical potential of water in water and reported variations of about 1 Kcal/mol as R_c was varied from 2.6 to 2.8 Å. Also, using $R_c \in [2.6, 2.9]$ Å resulted in uncertainties of about 2 Kcal/mol in estimates of the excess chemical potential of some species across the lipid layer. Note that the important region of the cavity-biased method is constructed over the translational degrees of freedom of a "coarse-grained" spherical molecule with an effective radius. This means that it can only be applied to small solutes with spherical or roughly spherical shapes.[6]

In this work we present an energy-biased method for the estimation of the chemical potential and reconstruction of the energy distribution $f(u)$ in dense phases. The idea is to restrict the sample to an important region defined by the set of bounded domains in the configurational space of the test molecule where the energy u is smaller than a given free parameter u_w. We denote as an *energy well* each compact subdomain within the test-molecule energy landscape for which $u < u_w$. Note that the present approach retains the main benefit of the cavity-biased method, but provides an exact evaluation of the energy distribution $f(u)$ and the chemical potential, because the energy wells are defined directly in terms of the energy landscape. Moreover our energy-biased method does not assume any particular molecular shape and therefore it may be used for nonspherical molecules and can coherently sample over rotational degrees of freedom as well.

We also note that the number of stages are not limited to two. When systems 0 and 1 are very different it may be impossible within the simulation time to sample the importance region of the two systems. In this case it is more efficient to compute the total free-energy difference by using a set of intermediate states. The energy-biased method can be applied on each of these intermediate state transitions at the cost of performing independent simulations for each state. Other approaches include, for instance, slow and fast growth methods where the system is changed from one state to another within a certain simulation time τ (large for slow growth). The fast growth method consists of sampling rapid

transformation from many simulations which are then combined by using the Jarzynski nonequilibrium work relation[11] to obtain the total free-energy difference.

The rest of the paper proceeds as follows. The energy-biased method is explained in Sec. II, while in Sec. III we derive an analytical expression for the efficiency of the method and estimate the optimal parameter u_w by maximizing the efficiency. In Sec. IV the method is tested in liquid argon at high density (modeled as Lennard-Jones atoms) where it is used to reconstruct the test-particle energy distribution $f(u)$ and the chemical potential. We also demonstrate the gain in efficiency obtained with energy-biased sampling with respect to uniform sampling. We conclude with a summary of our findings in Sec. V. Finally in Appendix A we briefly explain the Hit-and-Run algorithm which efficiently samples bounded regions of arbitrary shape immersed in an arbitrary number of dimensions.

II. OVERVIEW OF THE METHOD

As stated in the Introduction, energy-biased sampling consists of uniform sampling of the importance region defined by the set of subdomains in the test-molecule configurational space where its potential energy is less than u_w. The probability density is therefore given by

$$h(u) = \begin{cases} f(u)/F_w & u \leqslant u_w \\ 0 & u > u_w, \end{cases} \tag{7}$$

where the normalization factor $F_w \equiv \int_{-\infty}^{u_w} f(u)du$ is the cumulative probability of the unbiased distribution $f(u)$ and u_w is an arbitrary energy (free parameter).

Note that the energy-biased distribution of Eq. (7) can be straightforwardly combined with any of the popular methods to calculate the chemical potential from Eq. (4). We shall use the Bennett method due to its excellent performance. Introducing the weighting function $w(u)=\mathcal{F}[\beta(c-u)]$ in Eq. (6) and using Eq. (7), one obtains the energy-biased Bennett estimator for $\beta\mu$,

$$\beta\mu = \ln\left(\frac{\langle \mathcal{F}_c \rangle_g}{F_w \langle \mathcal{F}_c \rangle_h}\right) + \beta c, \tag{8}$$

where we have introduced the notation $\mathcal{F}_c \equiv \mathcal{F}[\beta(u-c)]$ to indicate that after the ensemble average we still have a function of c. As before, the subscript h indicates the average over the biased distribution of Eq. (7).

Sampling from the energy probability distribution $h(u)$ requires a more careful consideration of the energy landscape of the system. We indicate by $\mathbf{r}$ a configuration of the $(N+1)$th molecule and by $\mathbf{R}$ the configuration of the remaining N molecules. For a simple argon fluid $\mathbf{r} \in D$ where $D \subset R^3$, while for a three site flexible water model like TIP3P, $D \subset R^9$, which includes the three Euler angles determining the molecule orientation, the H–O–H angle, and the two H–O distances. As shown in Fig. 1, the region

$$A_{u_w} = \{\mathbf{r} \in D : u(\mathbf{r},\mathbf{R}) < u_w\} \tag{9}$$

is composed of many disconnected bounded regions of different sizes such that $A_{u_w} = \cup_\alpha A^\alpha_{u_w}$, where each $A^\alpha_{u_w}$ is now a connected region. Of course, for $u_w \to \infty$ we have that all the regions $A^\alpha_{u_w}$ connect and $A^\alpha_\infty = D$, the entire domain. The sampling algorithm must reproduce a uniform probability distribution,

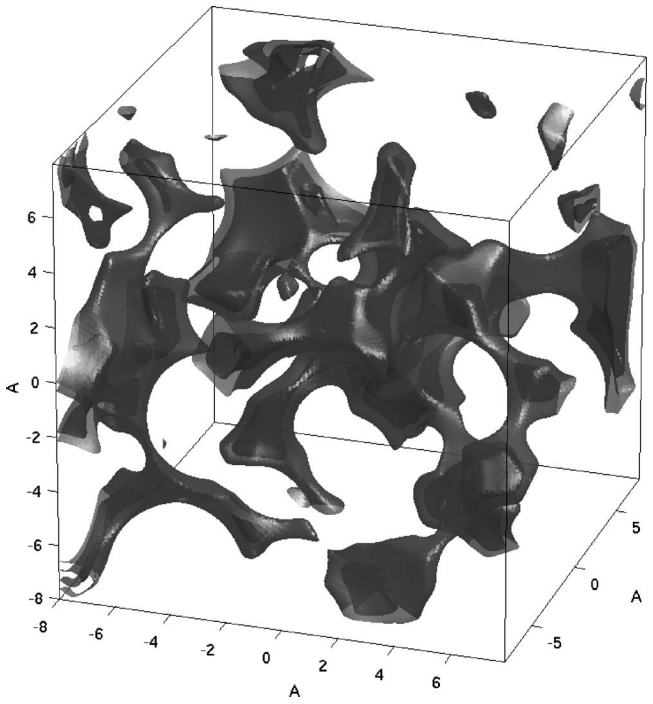

FIG. 1. Energy landscape for the three-dimensional configurational space generated by inserting an argon atom in a cube of side 16 Å of argon fluid. The isosurfaces of regions A_{u_w} are shown for u_w equal to 1 (dark grey) and to 10 Kcal/mol (light gray).

$$p_{u_w}(\mathbf{r}) = \frac{1}{\Omega(A_{u_w})}, \tag{10}$$

where $\Omega(A_{u_w})$ is the volume of the region.

For a given energy bias u_w, the algorithm for selecting configurations $\mathbf{r}$ according to Eq. (10) can be described in terms of two main steps which are applied iteratively:

(1) Locate a compact energy-well $A^\alpha_{u_w}$ in the configurational space D, where $u < u_w$.
(2) Sample the energy-well $A^\alpha_{u_w}$ with a uniform probability density.

The simplest procedure for locating energy wells in step (1) is to perform a random search over the whole configurational space until a fixed number of cavities is found. This procedure, however, does not avoid the probability of exploring the same well more than once, and we observed that it can easily lead to highly correlated data. Instead we perform step (1) by choosing points on a grid within the whole configurational space of the test molecule. In the case of the Lennard-Jones fluid, the three-dimensional configurational space is probed at the nodes of a Cartesian grid of size $n_x \times n_y \times n_z$, where n_α is the number of nodes along the coordinate α. We observed that the minimum distance between nodes that guarantees statistically independent samples is around 0.5σ.

An energy well is found at each node where the energy of the test molecule is $u < u_w$. Then, the locations of each of

these nodes are used as starting configurations for independent well samplings. In this way we ensure that we are sampling different cavities for each explored configuration (snapshot) of the system. Note that using grid sampling the number of cavities found per snapshot is a fluctuating quantity.

The search requires an average of $n_0=1/F_w$ energy evaluations to locate one well (i.e., one configuration with energy $u<u_w$.) During this same step (1) one can calculate the cumulative probability F_w from the estimator m/n_0, with n_0 being the total number of samples (Bernoulli trials) and m the number of successful trials with $u<u_w$, i.e., the total number of energy wells found. This number m/n_0 converges to F_w as $n_0\to\infty$ and, for a finite number of statistically independent trials n_0, its variance is $(1-F_w)F_w/n_0$. In practice, the estimation of F_w requires the number of unbiased samples to be $n_0 \gg 1/F_w$; this condition also ensures that a significant number of energy wells ($m>0$) are to be found.

Step (2) of the loop mentioned above requires a procedure to sample in an unbiased way the interior of each energy well. This is a delicate step because any bias incurred in sampling the importance region will be transfered to the estimator for $\beta\mu$, resulting in inaccuracy of the method. To tackle this problem we use the so-called Hit-and-Run algorithm,[12] which is explained in Appendix A.

III. EFFICIENCY AND OPTIMAL PARAMETERS OF THE METHOD

We now calculate the efficiency of the method and provide a way of choosing the optimal value of the parameter u_w by maximizing the efficiency. We also compare the efficiency of the estimator in Eq. (8) based on energy-biased sampling with that of the standard Bennett algorithm of Eq. (6).

A. Energy-biased Bennett method

The variance of the Bennett method can be cast in terms of the probability densities $f(u)$ and $g(u)$. Starting from Eq. (6), after some algebra the variance of the Bennett method assumes the form

$$\mathrm{Var}_B[\beta\mu]=\frac{1}{n_0\langle\mathcal{F}[\beta(u-c)]\rangle_f}, \tag{11}$$

where n_0 is the number of insertions used to sample the complete configurational space of the test particle. Note that the computational cost of the standard Bennett method is n_0, so according to (5) and Eq. (11) its maximum efficiency is given by

$$\varepsilon_B=\langle\mathcal{F}_c\rangle_f. \tag{12}$$

Let us now consider the variance of the estimator in Eq. (8), which is the sum of the variance of the estimator for F_w and the estimator for the ensemble average,

$$\mathrm{Var}_{\mathrm{EB}}[\beta\mu]=\mathrm{Var}[\ln F_w]+\frac{1}{n_w\langle\mathcal{F}_c\rangle_h}\simeq\frac{1}{n_0F_w}+\frac{1}{n_w\langle\mathcal{F}_c\rangle_h}, \tag{13}$$

where we have used the relation $\mathrm{Var}[\ln(F_w)]\simeq\mathrm{Var}[F_w]/F_w^2=(1-F_w)/(n_0F_w)\simeq 1/(n_0F_w)$, for $F_w\ll 1$. Here n_0 is the number of random insertions in the entire configurational space and n_w is the number of independent samples within the importance region $u<u_w$.

The probability of finding an energy well with $u<u_w$ using uniform sampling over the whole configurational space is F_w, so the number of cavities found after n_0 trials is $m=F_wn_0$. If the number of statistically independent samples per well is s, the total number of independent samples within the restricted configurational space $u<u_w$ is

$$n_w=n_0sF_w. \tag{14}$$

We note that the number of independent samples per well s depends on the fluid considered and, of course, on the biasing energy u_w. In Appendix B we provide a way of estimating s from the outcome of the data obtained from the Hit-and-Run sampling. Inserting Eq. (14) into Eq. (13) one obtains for the energy-biased algorithm,

$$\mathrm{Var}_{\mathrm{EB}}[\beta\mu]=\frac{1}{n_0}\left(\frac{1}{F_w}+\frac{1}{s\langle\mathcal{F}_c\rangle_f}\right). \tag{15}$$

In deriving Eq. (15) we used that $\langle\mathcal{F}_c\rangle_f=F_w\langle\mathcal{F}_c\rangle_h$ up to a negligible amount. This can be seen by noticing that the function $\mathcal{F}[\beta(u-c)]$ in the integrand of $\langle\mathcal{F}_c\rangle_f=\int_{-\infty}^{\infty}f(u)\mathcal{F}[\beta(u-c)]du$ decays exponentially for $u>c$. Hence, in any practical case ($u_w>c$) most of the integral weight comes from $u<u_w$, for which the energy-biased reconstruction of the energy profile $f(u)$ is exact (see Fig. 3).

We now evaluate the cost, which is given by the total number of energy evaluations of the test molecule needed to obtain n_w samples,

$$n_{\mathrm{cost}}=n_0+n_w/\mathrm{a}, \tag{16}$$

where $\mathrm{a}<1$ is the acceptance ratio of the Hit-and-Run sampling algorithm, defined in Appendix A. Introducing Eq. (14) into Eq. (16) we obtain

$$n_{\mathrm{cost}}=n_0\left(1+\frac{sF_w}{\mathrm{a}}\right). \tag{17}$$

For the energy-biased algorithm the efficiency is $\varepsilon=(n_{\mathrm{cost}}\mathrm{Var}_{\mathrm{EB}}[\beta\mu])^{-1}$. Using Eq. (15) and Eq. (17) one obtains

$$\varepsilon_{\mathrm{EB}}^{-1}=\frac{1}{F_w}+\frac{1}{s\langle\mathcal{F}_c\rangle_f}+\frac{s}{\mathrm{a}}+\frac{F_w}{\mathrm{a}\langle\mathcal{F}_c\rangle_f}. \tag{18}$$

By maximizing the efficiency $\varepsilon=\varepsilon(F_w)$ in Eq. (18) with respect to F_w, one obtains the optimal value F_w^{opt} and the maximum efficiency $\varepsilon_{\mathrm{EB}_{\max}}=\varepsilon_{\mathrm{EB}}(F_w^{\mathrm{opt}})$,

$$F_w^{\mathrm{opt}}=\sqrt{\mathrm{a}\langle\mathcal{F}_c\rangle_f} \tag{19}$$

$$\varepsilon_{\mathrm{EB}_{\max}}^{-1} = 2\frac{1}{\sqrt{\mathrm{a}\langle\mathcal{F}_c\rangle_f}} + \frac{s}{\mathrm{a}} + \frac{1}{s\langle\mathcal{F}_c\rangle_f}. \tag{20}$$

Finally, we compare the efficiency of the energy-biased algorithm with that provided by the Bennett algorithm, given by $\varepsilon_B = \langle\mathcal{F}_c\rangle_f$. According to Eq. (20) the ratio of efficiencies is given by

$$\frac{\varepsilon_B}{\varepsilon_{\mathrm{EB}_{\max}}} = 2\sqrt{\frac{\langle\mathcal{F}_c\rangle_f}{\mathrm{a}}} + \frac{s\langle\mathcal{F}_c\rangle_f}{\mathrm{a}} + \frac{1}{s}. \tag{21}$$

Equation (21) yields the range of values of $\langle\mathcal{F}_c\rangle_f$ for which the energy-biased Bennett estimator for $\beta\mu$ method is more efficient than the standard (unbiased) Bennett algorithm. Note that for $s=\sqrt{\mathrm{a}/\langle\mathcal{F}_c\rangle_f}$ the efficiency ratio given by Eq. (21) reaches its minimum value, $\varepsilon_B/\varepsilon_{\mathrm{EB}_{\max}} = 4\sqrt{\langle\mathcal{F}_c\rangle_f/\mathrm{a}}$, and therefore $\varepsilon_B < \varepsilon_{\mathrm{EB}}$ if $\langle\mathcal{F}_c\rangle_f > \mathrm{a}/16$. Hence the energy-biased method is suited for fluids at high densities or low temperatures or for molecular fluids with low insertion probability. In this regime $\langle\mathcal{F}_c\rangle_f \ll \mathrm{a}/16$ and the dominant term in Eq. (21) is $1/s$, hence $\varepsilon_{\mathrm{EB}_{\max}} \simeq s\varepsilon_B$. In other words, the maximal efficiency of the present energy-biased method is limited by the average number s of independent samples that can be obtained within one energy well. As shown in Appendix B, for the Lennard-Jones fluid we have observed that in the most unfavorable case (high density and low temperature) $s \sim [5-10]$.

B. Reconstruction of the energy distribution

We now show that the reconstruction of $f(u)$ using the energy-biased procedure (EB) is faster and more efficient than that obtained using any unbiased sampler which uniformly explores the whole configurational space. To that end we consider the evaluation of the cumulative probability $F(u)=\int_{-\infty}^{u} f(u')du$ for $u<u_w$ [i.e., for $F(u)<F_w$]. We shall compare the variance of two estimators for F: one based on uniform insertion over the whole domain and the other based on the energy-biased procedure. The variance of the unbiased estimator is simply $\mathrm{Var}(F)=F(1-F)/n_0$ and for low energies ($F \ll 1$) its efficiency is $1/F$. The expected value of the energy-biased estimator is HF_w, where $H(u)=\int_{-\infty}^{u} h(u')du'$ is the cumulative probability of the biased distribution in Eq. (7). This estimator is constructed as a product of two statistically independent fluctuating variables and its variance is[13]

$$\begin{aligned}\mathrm{Var}_{\mathrm{EB}}(F) &= \mathrm{Var}(HF_w) \\ &= F_w^2\mathrm{Var}(H) + H^2\mathrm{Var}(F_w) + \mathrm{Var}(F_w)\mathrm{Var}(H).\end{aligned} \tag{22}$$

Using $\mathrm{Var}(H)=H(1-H)/n_w$ and $\mathrm{Var}(F_w)=F_w(F_w-1)/n_0$ one obtains

$$\mathrm{Var}_{\mathrm{EB}} = \frac{F_w(F_w-1)H^2}{n_0} + \frac{H(1-H)F_w^2}{n_w} + \frac{F_wH(1-H)}{n_0n_w}. \tag{23}$$

Note that, as expected, for $H \simeq 1$ one recovers the variance of the unbiased insertion method. The interesting part of the energy distribution is the importance region, located in the low-energy range, where $H \ll 1$. In this regime one can make the approximation $1-H \simeq 1$. Using $F=HF_w$ and $n_w=n_0sF_w$, one gets

$$\mathrm{Var}_{\mathrm{EB}} = \frac{F}{n_0}\left(\frac{F}{F_w} + \frac{1}{s} + \frac{1}{n_0sF_w}\right). \tag{24}$$

Note that the term in brackets is the reduction in variance with respect to uniform-unbiased sampling. Because F_w is evaluated from n_0 probes, this means that necessarily $n_0 \gg 1/F_w$ so the third term inside the brackets is much smaller than unity. On the other hand, for the low energy range considered $F \ll F_w$ and one finally concludes that $\mathrm{Var}_{\mathrm{EB}} \simeq \mathrm{Var}(F)/s$, where $\mathrm{Var}(F) \simeq F/n_0$ is the variance obtained in the unbiased uniform sampling of the whole domain.

The cost associated with the energy-biased procedure is $n_{\mathrm{cost}}=n_0(1+sF_w/\mathrm{a})$. In the case of a Lennard-Jones liquid we have found that $\mathrm{a} \simeq 0.17$ and $s \sim O(10)$, while the optimal cumulative probability is $F_w \lesssim 10^{-3}$. This means that, in practical situations, $sF_w/\mathrm{a} \lesssim 1$ and $n_{\mathrm{cost}} \gtrsim n_0$. Thus, according to Eq. (24) the energy-biased sampling procedure is around s

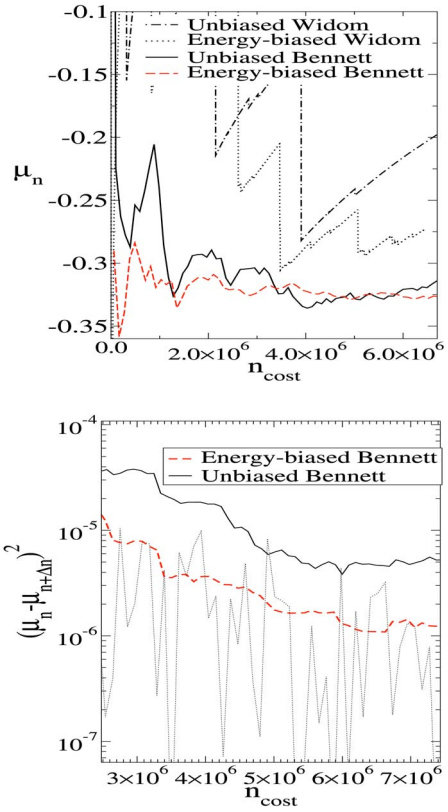

FIG. 2. (Top) The estimation of the chemical potential plotted against the overall number of energy probes (n_{cost}). We compare the standard (uniform sampling) Bennett and Widom methods with the corresponding energy-biased versions of these methods. The calculations correspond to a Lennard-Jones fluid with $\rho=0.0236$ Å^{-3} and $T=84$ K ($\rho=0.92\sigma^3$ and $T=0.7\epsilon/k_B$ in reduced LJ units); the energy-biased sampling was done using $u_w=14.19$ Kcal/mol and $d=15$ samples per well. (Bottom) The convergence measured as the squared difference between consecutive estimations with increasing cost ($\Delta n=10^5$). In the energy-biased method the cost is given by $n_{\mathrm{cost}}=n_0(1+dF_w/\mathrm{a})$, where n_0 is the number of random probes used to evaluate $F_w[=F(u_w)=0.001\,22]$ and $\mathrm{a}=0.165$ is the acceptance ratio of the Hit-and-Run sampler. The circles are the successive values of $(\mu_n-\mu_{n+\Delta n})^2$ (shown only for the energy-biased case) and the solid lines are the average over 20 consecutive differences.

TABLE I. Comparison of the chemical potential (in Kcal/mol) calculated via the standard Bennett method (i.e., using uniform-unbiased sampling) and the energy-biased Bennett (subscript EB). The inefficiency of both methods (reciprocal of efficiency) is also shown. In the case of the standard Bennett method we write the minimum inefficiency ($\varepsilon_B^{-1}=\langle\mathcal{F}_\mu\rangle_f$) while the inefficiency of the energy-biased method was obtained from numerical calculation of the variance of $\beta\mu$, using block analysis (see Appendix B or, e.g., Refs. 2 and 3) and agrees within error bars with the theoretical expression of Eq. (18) (see text). The error in $\varepsilon_{\mathrm{EB}}/\varepsilon_B$ comes mainly from the uncertainty in the numerical calculation of $\mathrm{Var}_{\mathrm{EB}}$.

ρ (Å^{-3})	T (K)	μ_{EB}	μ_B	ε_B^{-1}	$\varepsilon_{\mathrm{EB}}^{-1}$	$\varepsilon_{\mathrm{EB}}/\varepsilon_B$	F_w	d
0.023 60	84	−0.336	−0.323	0.9×10^5	$(1.2\pm0.1)\times10^4$	7 ± 1	0.001 22	15

times faster than a uniform-unbiased (grid or random) sampler in reconstructing the low-energy range of $f(u)$. As before, s is the average number of independent samples taken per well.

IV. RESULTS

In order to confirm the foregoing theoretical relations about efficiency and variance reduction, we performed molecular dynamics simulations of a Lennard-Jones liquid at high density and low temperature ($\rho=0.0236$ Å^{-3} and $T=84$ K). These simulations were performed in a cubic periodic box of side $L=10\sigma$. We used the standard Verlet method[2] to integrate Newton's equations of motion, incorporating a Langevin thermostat[14] to keep the system in the *NVT* ensemble.

During the simulation, the iterative loop (1)+(2) explained in Sec. II was performed m times per time interval $\delta t_{\mathrm{samp}}=0.5\tau$, which corresponds to about three times the collision time. The search for wells performed in step (1) was done by probing at the nodes of a Cartesian grid comprising 15^3 nodes. This ensured that the explored cavities are independent. All the cavities found in step (1) were sampled using the Hit-and-Run algorithm (see Appendix A).

A. Estimation of the chemical potential

One way to measure the efficiency of the method is to evaluate the convergence of the estimated value of the chemical potential for an increasing number of test-particle probes n_{cost}. Convergence can be calculated from the difference between successive values of μ_n, where $n(=n_{\mathrm{cost}})$ indicates the total number of evaluations of the test-particle energy. Figure 2 shows how this difference decreases in calculations based on both the energy-biased and the unbiased samples. These calculations correspond to liquid argon with number density of $\rho=0.0236$ Å^{-3} and temperature of $T=84$ K (these values correspond to $\rho=0.92\sigma^3$ and $T=0.7$ in Lennard-Jones units), for which the average of the Fermi function is $\langle\mathcal{F}_c\rangle_f=8.9\times10^{-6}$. According to Eq. (19) the optimum value of F_w is 0.0012, which corresponds to $u_w\simeq14.19$ Kcal/mol. We selected the predicted optimum parameter ($u_w=14.19$ Kcal/mol) and performed $d=15$ samples per well. As can be seen in Fig. 2, for equal numbers of energy probes ($n=n_{\mathrm{cost}}$), the average difference between successive estimates of the chemical potential via the energy-biased method is about five times smaller than that obtained with the unbiased sampler. As predicted by Eq. (21), such a gain in efficiency is consistent with the average number s of independent samples per well (see Table II), which for this simulation was $s\simeq5$.

Evaluations of the chemical potential for Lennard-Jones (LJ) fluids are shown in Table I together with the estimated efficiency of each calculation. For a LJ fluid with $\rho=0.023\,60$ Å^{-3} and $T=84$ K the numerically obtained net gain is around 7, which coincides with the prediction in Eq. (21) using $s=7$. For illustrative purposes we also analyzed a case for which the efficiency of our implementation of the energy-biased sampling is similar to the uniform-unbiased Bennett method. For instance, $\langle\mathcal{F}_c\rangle_f=0.0102$ for $\rho=0.017\,55$ Å^{-3} and $T=178.5$ K. Using $a=0.165$ and the (optimum) number of samples $s=\sqrt{a/\langle\mathcal{F}_c\rangle_f}\simeq4$ in Eq. (21) one obtains $\varepsilon_B/\varepsilon_{\mathrm{EB}_{\max}}\simeq1$; our numerical calculations, with $u_w=7.33$ and $d=8$, confirmed this conclusion. We note that for any value of u_w considered the energy-biased estimation of the chemical potential μ agrees within about 0.01 Kcal/mol with the unbiased Bennett result. This is illustrated in Table II where we show the estimated μ for the higher-density liquid, using several values of u_w.

B. Reconstruction of the energy distribution $f(u)$

In Fig. 3 we compare the reconstructed energy distribution $f(u)$ at energies $u<u_w$ with that computed from an unbiased method, which consists of a large number of random insertions within the entire configurational space. Figure 3 clearly illustrates that the energy-biased method *exactly* reproduces the unbiased distribution $f(u)$ for energies smaller that u_w. This attractive feature is a consequence of the fact that it is easy to exactly correct for the bias in terms of the cavity energies. This is not true for the accessible volume of the molecule, as in cavity-biased procedures.[6,9]

In order to illustrate the above conclusion we show in Fig. 4 the estimation of the cumulative probability $F(u)$ *versus* the total number of test-particle energy probes used for the evaluation. The particular case shown corresponds to $u=5$ Kcal/mol, for a LJ liquid at $\rho=0.0236$ Å^{-3} and $T=84$ K. The energy-biased sampling was done using $u_w=14.19$ Kcal/mol and $d=15$ samples per well, and for this calculation we obtained $s\simeq5$ (see Appendix B and Table II). Compared with the unbiased procedure, the reduction of variance provided by the energy-biased sampler is immediately apparent on inspection of Fig. 4. A numerical evaluation of the variance of each data set in Fig. 4 provides: $\mathrm{Var}_{\mathrm{EB}}=4.14\times10^{-5}/n_0$, while the (best) result for the algo-

TABLE II. Details of the energy-biased calculations in a Lennard-Jones (LJ) liquid at density $\rho=0.0236$ Å^{-3} and temperature $T=84$ K ($\rho=0.92$ and $T=0.7$ in LJ units). We compare the results for varying values of the energy parameter u_w, samples per cavity d, and varying number n_0 of energy probes within the unbiased distribution. The cumulative probabilities of the unbiased distribution $[F_w=F(u_w)=\int_{-\infty}^{u_w} f(u)du]$ are $F_w(165.53)=0.0704$, $F_w(28.38)=0.004\,58$, $F_w(14.19)=0.001\,22$. The average of the Fermi function in the biased distribution $\langle \mathcal{F}_c \rangle_h$ is defined in Eq. (8). The average number of independent samples per cavity s is obtained from $s=d/\tau_c$, where the correlation number τ_c is calculated from the correlation between the whole chain of data. The overall number of energy probes in the energy-biased method is $n_{\text{cost}}=n_0(1+dF_w/\text{a})$, where a is the acceptance ratio obtained for the Hit-and-Run sampler, $\text{a}\simeq 0.17$. The estimation of the chemical potential using the standard (unbiased) Bennett method with 1.1×10^7-energy samples is $\mu=-0.323$ Kcal/mol.

u_w (Kcal/mol)	d	n_0	$\langle \mathcal{F}_c \rangle_h$	s	n_{cost}/n_0	μ (Kcal/mol)
28.38	20	4.2×10^5	1.1×10^{-3}	7	1.5	−0.32
28.38	100	2.95×10^6	1.08×10^{-3}	7	3.8	−0.353
14.19	100	4.3×10^6	2.66×10^{-3}	12	1.7	−0.335
14.19	15	1.0×10^7	2.77×10^{-3}	5	1.1	−0.334
165.53	200	2.76×10^5	1.3×10^{-5}	25	85.8	−0.357

rithm based on uniform-unbiased sampling is $\text{Var}(F)=F/n_0=1.9\times10^{-4}/n_0$. Hence the net gain in efficiency is about 4.6, in agreement with the value of $s=5$ obtained from the independent correlation analysis explained in Appendix B. As shown in Table I, the estimated net gain in the evaluation of the chemical potential compared with the unbiased Bennett method is 7 ± 1, which is close to the estimate $s\simeq5$ obtained from the analysis of the cumulative probability.

V. CONCLUSION

We have presented a new method for sampling the energy of a test molecule in order to calculate single-particle ensemble averages and, in particular, the chemical potential. The method, called *energy-biased* sampling, restricts the important region to the bounded domains in the test-molecule energy landscape where the test-molecule energy u is smaller than a given free parameter u_w. This energy-biased sampling retains the principal benefit of cavity-biased methods[6,9] in the sense that, by sampling only within regions with a significant Boltzmann factor, convergence is greatly accelerated with respect to uniform sampling. Furthermore, because the energy-biased sampling is accurately defined in terms of the test-particle energy it has some important benefits: first, it allows accurate reproduction of the test-particle energy distribution $f(u)$ and the chemical potential; second, it is possible to sample cavities of arbitrary shape (not only spherical ones) and to generalize the cavity dimensionality to include the rotational degrees of freedom in the energy-well reconstruction; and finally, and rather importantly, it enables one to combine the sampling results with standard free-energy perturbation (FEP) formulas. In particular, we combined it with the Bennett method[8] which minimizes the variance of the estimator and has proved to be the best method in the literature.[1,5] Energy-biased sampling is a general protocol to bias the sampling and consists of two sequential steps: (1) searching and (2) sampling the interior of energy wells. In this work we have implemented these two steps using rela-

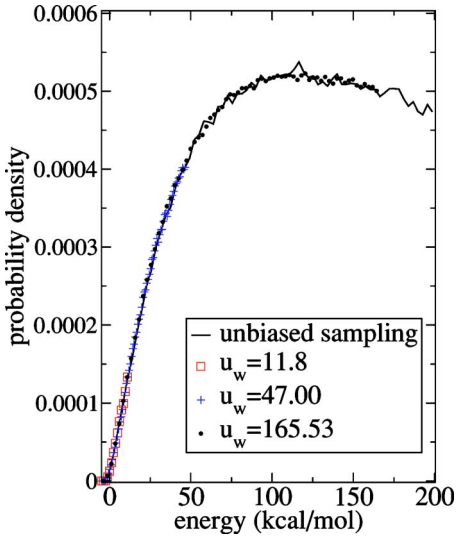

FIG. 3. The energy distribution $f(u)$ obtained from 4×10^6 random insertions over the whole configurational domain is compared with energy-biased sampling in the restricted configurational space $u<u_w$. The calculations correspond to the same case as in Fig. 2. The energy cavities are sampled using the Hit-and-Run algorithm, which provides an unbiased reconstruction of the energy distribution for any value of u_w chosen.

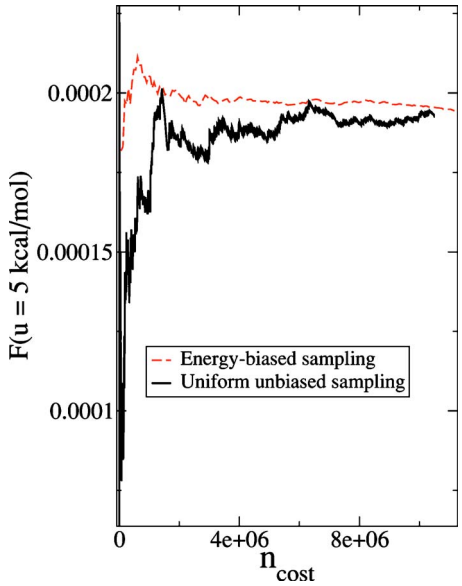

FIG. 4. The cumulative probability $F(u)=\int_{-\infty}^{u} f(u')du'$ for $u=5$ Kcal/mol vs the total number of energy evaluations of the test particle n_{cost}. The liquid is the same as in Fig. 2. We compare the estimations of $F(u)$ for grid sampling (with a regular mesh of 36^3 nodes) and for energy-biased samplings within $u<u_w=14.19$ Kcal/mol, performing $d=15$ samples per well. The cumulative probability at u_w is $F_w=F(u_w)=0.001\,22$.

tively simple algorithms: uniform-unbiased search and Hit-and-Run sampling. However, we note that other solutions are also possible. For instance, nonuniform sampling of the importance region may surely increase the efficiency of the present method. In dense systems, the searching step becomes the most difficult one and a more effective extension of this method could be to perform a biased search (using, for instance, some variation of the USHER algorithm[15,16]) so as to significantly increase the probability of finding favorable cavities for insertion of the test particle. These extensions are left for future studies.

ACKNOWLEDGMENTS

This research was supported by the EPSRC Integrative Biology Project No. GR/S72023 and by the EPSRC RealityGrid Project No. GR/67699. One of the authors (R.D.-B.) acknowledges support from the European Commission via Grant No. MERG-CT-2004-006316 and from the Spanish research Grant Nos. FIS2004-01934 and CTQ2004-05706/BQU.

APPENDIX A: SAMPLING BOUNDED REGIONS WITH THE HIT AND RUN ALGORITHM

There exists a relatively large literature on sampling a bounded connected region (see, for instance, Ref. 17 and references therein). In this work we have used the so-called Hit-and-Run algorithm for its simplicity and good performance.[17] The Hit-and-Run sampler is a special Markov chain Monte Carlo which draws numbers from an assigned distribution[12,17] $p(\mathbf{r})$, where $\mathbf{r} \in A$ lies within a bounded connected region of an n-dimensional space $A \subset R^n$. In our case, $p(\mathbf{r})$ is a uniform probability density over the region $A^{\alpha}_{u_w}$ such that

$$p(\mathbf{r}) = \frac{1}{\Omega(A^{\alpha}_{u_w})}. \tag{A1}$$

The Hit-and-Run algorithm starts from a point $\mathbf{r}_0$ within the bounded region A and performs the following steps:

(i) Choose a random direction $\mathbf{e}$ and find the intersections of the cavity border with the line $\mathbf{r}(\lambda)=\mathbf{r}_0+\lambda\mathbf{e}$, where λ is a real number. As the cavity A is bounded the intersection is composed by two points $\mathbf{r}(\lambda^+)$ and $\mathbf{r}(\lambda^-)$ (here $\lambda^+>0$ and $\lambda^-<0$).

(ii) Select a point $\mathbf{r}_1$ within the segment $[\mathbf{r}(\lambda^+),\mathbf{r}(\lambda^-)]$, i.e.,

$$\mathbf{r}_1 = \mathbf{r}(\lambda^-) + \xi(\mathbf{r}(\lambda^+) - \mathbf{r}(\lambda^-)), \tag{A2}$$

where $\xi \in (0,1)$ is a uniformly distributed random number.

(iii) Sample at $\mathbf{r}_1$, set $\mathbf{r}_1 \rightarrow \mathbf{r}_0$ as the new starting point and go to (i).

The above procedure is repeated to obtain the desired number of samples d. In our case the starting point for the sample chain $\mathbf{r}_0$ is the test-particle configuration returned by the algorithm for energy-well searching $[U(\mathbf{r}_0,\mathbf{R})<u_w]$. In order to locate the borders of the energy well $\mathbf{r}(\lambda^+)$ and $\mathbf{r}(\lambda^-)$ we use the following procedure. Starting from $\mathbf{r}_0$ we cross the well along the line defined by the random unit vector $\mathbf{e}$ moving in steps of size δs, i.e., according to

$$\mathbf{r}(k) = \mathbf{r}_0 + k\delta s\mathbf{e}, \tag{A3}$$

with k being an integer starting from $k=\pm 1$. The energy is computed at each point $\mathbf{r}(k)$ until one crosses the edges of the well at $k=k^+$ and $k=k^-$ [for which $u(\mathbf{r}(k^{\pm}),\mathbf{R})>u_w$]. An approximate location of the cavity borders is provided by setting $\lambda^{\pm}=k^{\pm}$. We used typically $\delta s \simeq 0.3$ Å and required, on average, about five iterations to cross the well in one random direction (this value depends on the density and u_w). Note that the acceptance ratio is $a=\langle k^+-k^-\rangle^{-1}$ and for the high-density cases considered here $a \simeq 0.17$.

APPENDIX B: OPTIMAL NUMBER OF SAMPLING DIRECTIONS

It is possible to reduce the cost without increasing the variance by setting the number of samples per cavity d equal to or somewhat larger than s, the average number of independent samples per cavity. Note that the number of statistically independent samples within one cavity is $s=d/\tau_c$, where τ_c is an empirically estimated autocorrelation length of the whole chain of data. This number τ_c can be estimated from the large m limit of the quantity $m\mathrm{Var}[\mathcal{F}^{(m)}]/\mathrm{Var}[\mathcal{F}]$, where $\mathcal{F}_c=\mathcal{F}[\beta(u-c)]$ is the Fermi function evaluated at a single energy u and $\mathcal{F}^{(m)}$ denotes the mean of m consecutive $\mathcal{F}$ values.

The value of s can be estimated by performing several Hit-and-Run samplings with an increasing number of directions per cavity $d>s$, then computing τ_c for the chain of samples and evaluating d/τ_c, which should be nearly independent of d. We carried out this evaluation of s for varying values of u_w within the same system and for fixed u_w and varying density. The results of this study, reported in Table II, clearly indicate that s does not greatly vary for a broad range of values of the cavity-border energy u_w. In fact, at low and moderate values of u_w the energy cavities are isolated and their average size (in Å) grows quite slowly with u_w. This is due to the steepness of the hard-core part of the Lennard-Jones potential. Above a certain energy u_w the cavities become connected and a steep rise in the average size of the energy cavities is observed. This is reflected in the value of s. As shown in Table II for $u_w=14.19$ Kcal/mol we obtained $s \simeq 4.5$ and $s \simeq 11$ for two calculations using $d=15$ and $d=100$, respectively. We obtained a relatively close value $s \simeq 7$ for twice as large an energy limit $u_w=28.38$ Kcal/mol. However, using $u_w=165.53$ Kcal/mol the average number of independent samples increased up to 25, reflecting the more complex shape and larger volume of these energy cavities. In summary, for the optimum range of values of $u_w \sim [10-30]$ Kcal/mol we find $s \simeq [5-10]$ in the case of the Lennard-Jones liquid.

[1] N. Lu, J. K. Singh, and D. A. Kofke, J. Chem. Phys. **118**, 2977 (2003).
[2] M. Allen and D. Tildesley, *Computer Simulations of Liquids* (Oxford University Press, Oxford, 1987).
[3] D. Frenkel and B. Smith, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic, San Diego, 2002).

[4] P. Kollman, Chem. Rev. (Washington, D.C.) **93**, 2395 (1993).
[5] M. R. Shirts and V. S. Pande, J. Chem. Phys. **122**, 144107 (2005).
[6] G. L. Deitrick, L. E. Scriven, and H. T. Davis, J. Chem. Phys. **90**, 2370 (1989).
[7] K. S. Shing and K. E. Gubbins, Mol. Phys. **46**, 1109 (1982).
[8] C. H. Bennett, J. Comput. Phys. **22**, 245 (1976).
[9] P. Jedlovszky and M. Mezei, J. Am. Chem. Soc. **122**, 5125 (2000).
[10] A. Pohorille and M. A. Wilson, J. Chem. Phys. **104**, 3760 (1996).
[11] C. Jarynski, Phys. Rev. Lett. **78**, 2690 (1997).
[12] R. L. Smith, Oper. Res. **32**, 1296 (1984).
[13] L. Goodman, J. Am. Stat. Assoc. **55**, 708 (1960).
[14] K. Kremer and G. Grest, J. Chem. Phys. **92**, 5057 (1990).
[15] R. Delgado-Buscalioni and P. V. Coveney, J. Chem. Phys. **119**, 978 (2003).
[16] G. De Fabritiis, R. Delgado-Buscalioni, and P. V. Coveney, J. Chem. Phys. **121**, 12139 (2004).
[17] J. S. Liu, *Monte Carlo Strategies in Scientific Computing* (Springer, New York, 2001).