Journal of Optoelectronics and Advanced Materials Vol. 7, No. 3, June 2005, p. 1563 - 1571

EXPLORING PHASE SPACES OF BIOMOLECULES WITH MONTE CARLO METHODS

A. Buţu*

National Institute for Research and Development of Biological Sciences,
Splaiul Independenţei 296, Buchurest, Romania

Monte Carlo (MC) methods have become an important tool in biophysics, structural molecular biology and rational drug design. They are used to predict quantities that either cannot be measured directly or when accurate experimental data are difficult to obtain. They are also helpful for the interpretation of experimental results since in a simulation the system of interest can be studied in detail on the molecular level.

(Received January 18, 2005; accepted May 26, 2005)

Keywords: Monte Carlo method, Biomolecules conformations, Metropolis algoritm

1. Introduction

A central problem for molecular simulations is the sampling of conformational degree of freedom. The two most widely used methods for atomic-level modeling of fluids are the Monte Carlo (MC) methods and the molecular dynamics (MD). Both procedures could use the same molecular models, classical force fields for the potential energy terms, and the implementation of boundary conditions. The principal differences are the methods of sampling the configuration space available to the system. The conventional form of molecular dynamics represents a realization of Boltzmann's approach to statistical mechanics, whereas the Monte Carlo method is rooted on the Gibbs' formulation of the problem. MC serves as a very robust algorithm and can be applied to more types of models and potential functions. The advantages of the MD methods are the efficiency [1,2,3] of searching in the phase space for high density systems and the built-in parallellizable nature. However, in some situations and by means of optimization of algorithms the MC methods could be still more efficient [3].

The Monte Carlo method was developed by von Neumann, Ulam, and Metropolis at the end of the Second World War to study the diffusion of neutrons in fissionable material. The name, which derives from the famous Monaco casino, emphasizes the importance of randomness, or chance, in the method and was coined by Metropolis in 1947 in the title of a paper describing the early work at Los Alamos [4].

It was when the MANIAC computer in Los Angels became operational in March 1952, Metropolis was interested in having as broad a spectrum of problems as possible tried on the new machine, in order to evaluate its logical structure and to demonstrate the capabilities of the machine. Solving the classical statistical mechanical N-body problem via Monte Carlo technique was one of the first problems, done by Metropolis, in collaboration with the Tellers and the Rosenbluths, and led to the development of what is now known as the Metropolis Monte Carlo method [5].

[*] Corresponding author: alina_butu@yahoo.com

2. Importance of the sampling technique

Monte Carlo (MC) methods refer, in a very general sense, to any simulation of an arbitrary system which uses a computer algorithm explicitly dependent on a series of (pseudo)random numbers [6]. MC is particularly important in statistical physics, where systems have a large number of degrees of freedom and quantities of interest, such as thermal averages, cannot be computed exactly.

Consider the integral of interest:

$$I = \int_{x_1}^{x_2} f(x)\, dx \tag{1}$$

A naive way to calculate the above integral using the Monte Carlo approach would be the following quadrature formula:

$$I \approx \frac{x_2 - x_1}{N_{trial}} \sum_{i=1}^{N_{trial}} f(\xi_i) \tag{2}$$

where ξ_i is the i-th chosen random number from a uniform distribution at the interval $[x_1, x_2]$.

Importance sampling techniques choose random numbers from a distribution $\rho(x)$, which allows the function evaluation to be concentrated in the regions of space that make important contribution to the integral.

If we rewrite the above integral as:

$$I = \int_{x_1}^{x_2} \left(\frac{f(x)}{\rho(x)} \right) \rho(x)\, dx \tag{3}$$

where $\rho(x)$ is an arbitrary positive weight function, then the integral I could be evaluated as the expectation value of $f(x)/\rho(x)$ with the probability density function $\rho(x)$.

Such a reformulation makes it possible to speed up the efficiency of the sampling.

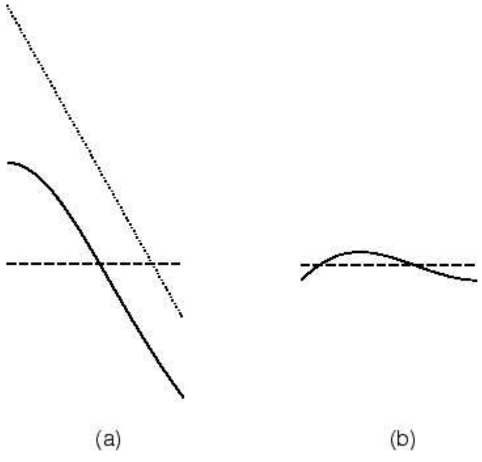

Fig. 1. Transformation of integrand in the importance sampling technique. The solid line and dotted line in (a) represent the original integrand *f*(x) and weighting function ρ (x), respectively. The solid line in (b) represents the transformed integrand *f*(x) / ρ (x). The dashed lines in both parts represent the average values.

If we choose a ρ (x), which behaves approximately as f (x) does (i.e., ρ (x) is large where f (x) is large and small where f (x)is small), then the integrand in eq.(3.), f(x) / ρ (x), can be made very smooth, see Fig. 1 with a consequent reduction in the standard deviation of the Monte Carlo estimate:

$$\sigma \approx \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}\left(f(\xi_i)-f\right)^{2}}{N}} \tag{4}$$

$$\geq \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}\left[\left(\frac{f(\zeta_i)}{\rho(\zeta_i)}\right)-\overline{\left(\frac{f}{\rho}\right)}\right]^{2}}{N}} \approx \sigma' \tag{5}$$

where ζ_i is the i-th chosen random number according to the probability distribution function $\rho(x)$. This is a result from the central limit theorem, and the values of σ and σ' defined above are the deviation of the integral of the function f from that of the Monte Carlo evaluation.

Common integrals we will encounter in statistical mechanics are as follows:

$$\langle A\rangle = \int d^{3N}p\,d^{3N}q\,\rho(\mathbf{p},\mathbf{q})\,A(\mathbf{p},\mathbf{q}) \tag{6}$$

where $(\mathbf{p},\mathbf{q}) = (p_1, p_2, \ldots, p_{3N}, q_1, q_2, \ldots, q_{3N})$, is the phase point in the 6N-dimensional phase space and the choice of $\rho(\mathbf{p},\mathbf{q})$ depends on the ensemble of interest. For the microcanonical ensemble, the $\rho(\mathbf{p},\mathbf{q})$ is:

$$\rho(\mathbf{p},\mathbf{q}) = \delta\left(\mathrm{H}(\mathbf{p},\mathbf{q}) - \varepsilon\right) \tag{7}$$

where ε is the total energy of the system, and for the canonical ensemble, the $\rho(\mathbf{p},\mathbf{q})$ reads:

$$\rho(\mathbf{p},\mathbf{q}) = e^{-\beta \mathrm{H}(\mathbf{p},\mathbf{q})} \tag{8}$$

For the isobaric-isothermal ensemble, the $\rho(\mathbf{p},\mathbf{q})$ reads:

$$\rho(\mathbf{p},\mathbf{q}) = e^{-\beta\left(\mathrm{H}(\mathbf{p},\mathbf{q})+\mathrm{PV}\right)} \tag{9}$$

3. Markovian chain and the master equation

It is clear now that we should generate a stochastic process which is distributed according to the desired distribution $\rho(\mathbf{p},\mathbf{q})$ in order to calculate integrals like eq.(6).

For simplicity we discuss the stochastic process with discrete time step. Suppose that the probability density function at time $\rho(\mathbf{p},\mathbf{q};t) = \rho(\mathbf{p},\mathbf{q})$, where $t = t_0 + \tau\Delta t, \tau = 0,1,2\ldots$. Then we have the following "master equation" for a Markovian stochastic process:

$$\rho_{r+1}(\mathbf{p_n},\mathbf{q_n}) = \rho_r(\mathbf{p_n},\mathbf{q_n}) + \int \left[\rho_r(\mathbf{p_m},\mathbf{q_m})P(m\to n) - \rho_r(\mathbf{p_n},\mathbf{q_n})P(n\to m)\right]d\mathbf{p_m}d\mathbf{q_m} \tag{10}$$

where $P(m\to n)$ is the probability that the system ("random walker") will make a transition from state *m* to state *n*. We aim at reaching the desired distribution, $\rho(\mathbf{p},\mathbf{q})$, for the ensemble of interest after long enough time, i.e.,

$$\lim_{r\to\infty}\rho_r(\mathbf{p},\mathbf{q}) = \rho(\mathbf{p},\mathbf{q}) \tag{11}$$

In this limit we have:

$$\int \rho(\mathbf{p_m},\mathbf{q_m})P(m\to n)\,d\mathbf{p_m}d\mathbf{q_m} = \int \rho(\mathbf{p_n},\mathbf{q_n})P(n\to m)\,d\mathbf{p_n}d\mathbf{q_n} \tag{12}$$

A sufficient but not necessary "detailed balancing" condition (also called "microscopic reversibility") could satisfy above equation with great simplicity:

$$\rho(\mathbf{p_m},\mathbf{q_m})P(m\to n) = \rho(\mathbf{p_n},\mathbf{q_n})\,p(n\to m) \tag{13}$$

or

$$\frac{P(m\to n)}{P(n\to m)} = \frac{\rho(\mathbf{p_n},\mathbf{q_n})}{\rho(\mathbf{p_m},\mathbf{q_m})} \tag{14}$$

4. Metropolis algorithm

Usually the transition probability $P(m\to n)$ consists of two parts:

$$P(m\to n) = T(m\to n)\,A(m\to n) \tag{15}$$

where $T(m\to n)$ is the probability of making a trial from state *m* to state *n*, and $A(m\to n)$ is the probability of accepting that step. If state *n* can be reached from state *m* in a single step, then

$$T(m\to n) = T(n\to m) \tag{16}$$

so that the equilibrium distribution of the random walkers satisfies

$$\begin{aligned}\frac{\rho(\mathbf{p_n},\mathbf{q_n})}{\rho(\mathbf{p_m},\mathbf{q_m})} &= \frac{T(m\to n)\,A(m\to n)}{T(n\to m)\,A(n\to m)}\\ &= \frac{A(m\to n)}{A(n\to m)}\end{aligned} \tag{17}$$

It is easy to prove that the following choice for $A(m\to n)$ which was first proposed by Metropolis et al. [5]

$$A(m\to n) = \min\left[1, \rho(\mathbf{p_n},\mathbf{q_n})/\rho(\mathbf{p_m},\mathbf{q_m})\right] \tag{18}$$

satisfies the above stochastic equation.

In the canonical ensemble, we have:

$$A(m \to n) = \min\left[1, \frac{e^{-\beta \mathrm{H}(\mathbf{p_n},\mathbf{q_n})}}{e^{-\beta \mathrm{H}(\mathbf{p_m},\mathbf{q_m})}}\right] = \min\left[1, e^{-\beta \Delta E}\right] \tag{19}$$

where $\Delta E = \mathrm{H}(\mathbf{p_n},\mathbf{q_n}) - \mathrm{H}(\mathbf{p_m},\mathbf{q_m})$.

It should be noted that eq.(19) is not the only solution for the stochastic equation eq.(18).

In systems containing large potential energy barriers, the Monte Carlo methods may not be able to search the whole configuration space, leading to the so-called *quasi-ergodicity*. This can be avoided by jumping over the barriers by coupling the conventional Metropolis sampling to the Boltzmann distribution generated by another random walker at higher temperature [7]. On the other hand, if kinetic mechanisms or continuous dynamics are still of interest, the efficient moves based on the informations of the local properties of potential functions, e.g., forces, torques and Hessian matrices, have been proposed in the force bias methods [8, 9], smart Monte Carlo [10], or energy biased method [11].

In a MC simulation a segment of a molecule moves stochastically through random moves of a small number of monomers at a time. In general MC serves as a very robust algorithm with well-defined ensembles. Unfortunately the MC methods suffers from two disadvantages. First, it is very difficult to implement an efficient algorithm in parallel computer architectures. Second, in dense systems, the motions are mostly reduced to cooperative fluctuations. However, the MC method has been very successful and will remain very important at low and moderate densities.

5. Monte Carlo methods of simulating polymer systems

In the simple sampling algorithm the polymer chains are each time grown from scratch. This is a good method for Random Walks (but Random Walks are trivial anyway). For Self Avoiding Walks this would mean that each time a self-intersection occurs the chain should be grown anew from the beginning. All the effort put in growing the chain is lost. For large *N* the probability that a walk is grown without self-intersection is very low. Most of the attempts to grow a Self Avoiding Walks fail. This is why simple sampling is ineffective for large *N*. This problem can be partially remedied by choosing at each step only from the unoccupied lattice sites. However, this introduces a bias which should be compensated for. This is done by giving a weight to the conformation. If *si* is the number of unoccupied neighbors (i.e. the number of possible steps) at step *i* the conformation should be weighted by:

$$w = \prod_i s_i \tag{20}$$

This method is also known as the Rosenbluth method [12]. A drawback of the method is that one spends a lot of time generating conformations which have a low weight.

It is desirable if one could generate conformations with the correct probability so no reweighting is necessary. This is called “Importance sampling” [13]: one avoids to sample conformations which have a low weight.

The Metropolis method [5] is a frequently used method to achieve importance sampling. It consists of repeatedly trying to change the conformation a little bit (moving a few atoms or so) and then deciding whether to accept the new conformation or to retain the previous one. Such a prospective change in the conformation is called a “move”. The acceptance criterion should be chosen in such a way that the conformations are sampled with the desired probability (usually the Boltzmann distribution).

The method of generating a new conformation (on the basis of the previous one) is stochastic and should satisfy “detailed balance”: the probability of generating state *j* when in state *i* should be the same as the probability of generating state *i* when in state *j*. Furthermore ‘ergodicity’ is required: one should be able to reach each microscopic state by one or more of these sampling steps.

If the sampling procedure satisfies detailed balance and ergodicity the conformations are sampled with uniform probability. A popular acceptance criterion for achieving Boltzmann statistics is [13]:

$$P(\text{accept}) = \min(1, \exp(-\beta\Delta E)) \tag{21}$$

where

$$\Delta E = E(\text{new}) - E(\text{previous}) \tag{22}$$

A convenient way to implement this in a computer program is:
1. compute the change in energy ΔE
2. if $\Delta E < 0$ accept the new conformation always.
3. if $\Delta E > 0$ compute a random number *r* from a uniform distribution on
the interval [0, 1) and compare it to $\exp(-\beta\Delta E)$. If $r < \exp(-\beta\Delta E)$
accept the new conformation. Otherwise retain the previous conformation.

It is essential that, when a move is rejected, the previous conformation is taken into account once more in the statistics (for instance in the averages of physical quantities). Not doing so would result in biased statistics and incorrect ensemble averages.

Note that a move which leads to excluded volume overlap should be handled as if $\Delta E = \infty$ and should always be rejected.

The Metropolis procedure described above generates a sequence of conformations, some identical to the previous one, some slightly altered. Ensemble averages can be calculated simply by averaging over the generated conformations. No weights are needed.

One of the drawbacks of the method is that subsequent conformations are very much alike or identical (if the move was rejected). It takes a number of steps to obtain a conformation which is uncorrelated. This number is called the “correlation time” [14]. Something similar happens when one starts a simulation with an initial conformation that is designed by hand (for example a completely stretched chain). Usually such a conformation is artificial and a number of steps is needed to reach a conformation that is uncorrelated to the initial conformation. This is called “equilibration". The conformations that were generated during equilibration should not be included in the statistics; they must be discarded.

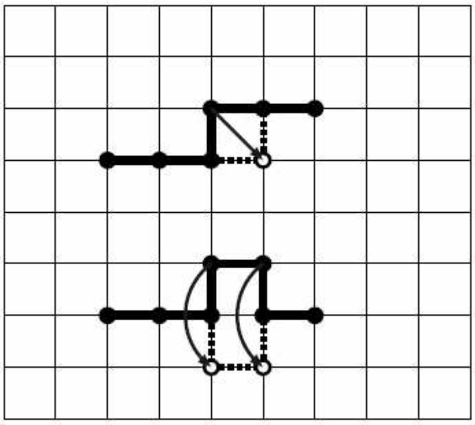

Fig. 2. Monte Carlo moves: kink jump and crankshaft moves.

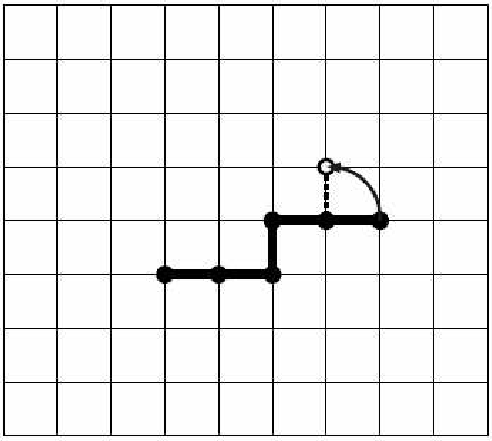

Fig. 3. Monte Carlo moves: end move.

There is a lot of freedom in choosing the Monte Carlo moves, provided they satisfy detailed balance and ergodicity. This freedom of choice should be used to optimize the efficiency of the sampling: the phase space (the space of microscopic states) should be traversed as fast as possible.

We will now give a few examples of the Monte Carlo algorithms used in polymer simulation.

For polymers on a simple cubic lattice several Monte Carlo moves have been suggested. The kink jump and crankshaft moves are illustrated in Fig. 2. The kink jump is a single atom move while the crankshaft move moves two atoms at a time. For a linear polymer also moves that reposition a loose end should be included in the sampling algorithm (Fig. 3). The combination of kink jumps, crankshaft moves and end moves is called the Verdier Stockmayer algorithm.

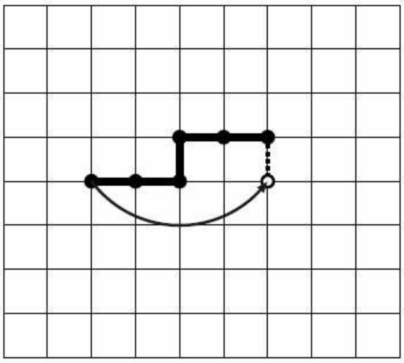

Fig. 4. Monte Carlo moves: reptation move.

A reptation move [15] consists of removing a monomer from one chain end and randomly adding a monomer to the other end. (Fig. 4). It is also called the “slithering snake” move for obvious reasons. The growth direction can be chosen anew at every attempted step, but it can be shown that it is also correct to let the polymer grow consistently in one direction reversing the growth direction only when a move is rejected.

Another more involved lattice model for polymers is the Bond Fluctuation Model [16]. Monomers occupy a square of 4 lattice sites in 2 dimensions or a cube of 8 sites in 3 dimensions. There is a set of allowed bond vectors between the monomers. This set has been carefully chosen such that no crossing bonds will form during the simulation (provided that there were no crossing bonds in the initial conformation). Only one type of Monte Carlo move is used: the chosen monomer is moved by one lattice constant. The advantage of the Bond Fluctuation Model is that the chains have more flexibility: the number of allowed bond lengths and bond angles is larger than the simple cubic lattice model where only bonds of length 1 and bond angles of 90◦ and 180◦ are allowed. The Bond Fluctuation Model can be viewed as an intermediate between lattice models and off-lattice models.

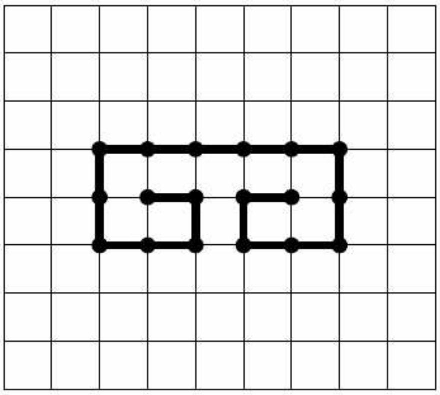

Fig. 5. Trapped conformation.

There are a number of off-lattice models that are used in polymer simulations.The models that will be described below are again coarse grained models.

In the ‘bead and spring’ model spherical beads represent the monomers. They are not allowed to overlap (excluded volume interaction). The beads are connected by springs: there is a quadratic or FENE potential between bonded monomers. A Monte Carlo move consists of randomly choosing a monomer and moving it in some convenient way that satisfies detailed balance.

In the ‘freely jointed’ model [17] the length of the bonds is held fixed. This can be achieved by starting in a conformation with the correct bond lengths and using Monte Carlo moves that do not change the bond lengths. The ‘tangent spheres’ model consists of impenetrable spheres for monomers and fixed bond lengths which are chosen such that the bonded beads touch one another.

6. Molecular Monte Carlo

Monte Carlo simulations of molecules are performed in a similar manner to those of atomic systems. However, there are different degrees of freedom to consider and so additional Monte Carlo moves are required. For rigid molecules it is necessary to consider orientational as well as translational degrees of freedom. These are usually combined into a single trial move consisting of a centre of mass translation and a rotation around the centre-of-mass. A translational move is performed by randomly displacing the centre of mass of the molecule. An orientational move is then performed by randomly changing the orientation of the molecule. In order to perform rotational moves, it is necessary to have a set of coordinates to describe the molecular orientation. One such set are the Euler angles [18]. These are described in terms of a sequence of rotations of a set of Cartesian axes about the origin. The first is through an angle ϕ about the z axis. This is followed by a rotation of θ about the new x axis and then a final rotation of ψ about the new z axis. The change in orientation can then be achieved through random changes in these angles. For flexible molecules, changes to the internal coordinates occur along side translational and rotational moves. Again we need a set of coordinates to describe the molecular conformations. This is commonly done in terms of internal coordinates, the bond lengths, bond angles, and dihedral angles in each molecule. As for rigid molecules, these are usually combined into a single trial move. A Monte Carlo move for a flexible molecule could then consist of a centre of mass displacement, and rotation of the molecule about its centre of mass, and then a change in an internal dihedral angle, bond angle and bond length.

7. Hybrid methods

It is tempting to combine the nice features of both Monte Carlo method and molecular dynamics methods to reach higher efficiency for searching in the phase space. A hybrid Monte Carlo method was proposed [19] in the field of quantum chromodynamics which contains fermion degree of freedom, and was later applied to condensed-matter systems [20, 21, 22]. However, this algorithm does not reach higher efficiency although it was claimed to be more robust.

For the lipid bilayer simulation, a new equilibration procedure for the atomic level simulation was proposed recently by Chiu et al. [23], which is also a hybrid algorithm with configurational bias Monte Carlo moves searching in the configuration space.

8. Conclusion

The importance of the simulation Monte Carlo methods in the field of structural molecular biology, as well as the peculiarities of different procedures are pointed out.

References

[1] S. H. Northrup, J. A. McCammon, Biopolymers **19**, 1001 (1980).

[2] G. S. Grest, K. Kremer, Phys. Rev. A **33**, 3628 (1986).

[3] W. L. Jorgensen, J. Tirando-Rives, J. Phys. Chem. **100**, 14508 (1996).

[4] N. Metropolis, S. Ulam, J. Am. Stat. Ass. **44**, 335 (1949).

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[6] K. Binder, D. Heermann, Monte Carlo simulation in statistica physics, Springer-Verlag, 1988.

[7] D. D. Frantz, D. L. Freeman, J. D. Doll, J. Chem. Phys. **93**, 2769 (1990).

[8] M. Rao, C. Pangali, B. J. Berne, Mol. Phys. **37**, 1773 (1979).

[9] Jianshu Cao, B. J. Berne, J. Chem. Phys. **92**, 1980 (1990).

[10] P. J. Rossky, J. D. Doll, H. L. Friedman, J. Chem. Phys. **69**, 4628 (1978).

[11] E. Leontidis, U. W. Suter, Mol. Phys. **83**, 489 (1994).

[12] M. N. Rosenbluth, A.W. Rosenbluth. J. Chem. Phys. **23**, 356 (1955).

[13] D. Frenkel, B. Smit, Understanding Molecular Simulation: From Algorithms to Applications. Academic Press, 1996.

[14] H. Flyvbjerg, H. G. Petersen, J. Chem. Phys. **91**(1), 461 (1989).

[15] F. T. Wall, F. Mandell, J. Chem. Phys. **63**(11), 4592 (1975).

[16] H. P. Deutsch, K. Binder, J. Chem. Phys. **94**(3), 2294 (1991).

[17] A. Baumgaertner, J. Chem Phys. **72**(2), 871 (1980).

[18] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Phys. Lett. **195**, 216 (1987).

[19] M. Creutz, A. Gocksch, Phys. Rev. Lett. **63**, 9 (1989).

[20] B. Mehlig, D. W. Heermann, B. M. Forrest, Phys. Rev. B **45**, 679 (1992).

[21] B. Mehlig, D. W. Heermann, B. M. Forrest, Mol. Phys. **76**, 1347 (1992).

[22] F. A. Brotz, J. J. de Pablo, Chem. Eng. Sci. **49**, 3015 (1994).

[23] S. W. Chiu, M. M. Clark, E. Jakobsson, S. Subramaniam, H. L. Scott, J. Comput. Chem. **20**, 1153 (1999).