Calculation of the entropy of random coil polymers with the hypothetical scanning Monte Carlo method

Ronald P. White and Hagai Meirovitch[a)]
Department of Computational Biology, University of Pittsburgh School of Medicine, Biomedical Science Tower, Pittsburgh, Pennsylvania 15261

(Received 5 July 2005; accepted 5 October 2005; published online 8 December 2005)

Hypothetical scanning Monte Carlo (HSMC) is a method for calculating the absolute entropy S and free energy F from a *given* MC trajectory developed recently and applied to liquid argon, TIP3P water, and peptides. In this paper HSMC is extended to random coil polymers by applying it to self-avoiding walks on a square lattice—a simple but difficult model due to strong excluded volume interactions. With HSMC the probability of a given chain is obtained as a product of transition probabilities calculated for each bond by MC simulations and a counting formula. This probability is exact in the sense that it is based on all the interactions of the system and the only approximation is due to finite sampling. The method provides rigorous upper and lower bounds for F, which can be obtained from a very small sample and even from a *single* chain conformation. HSMC is independent of existing techniques and thus constitutes an independent research tool. The HSMC results are compared to those obtained by other methods, and its application to complex lattice chain models is discussed; we emphasize its ability to treat any type of boundary conditions for which a reference state (with known free energy) might be difficult to define for a thermodynamic integration process. Finally, we stress that the capability of HSMC to extract the absolute entropy from a given sample is important for studying relaxation processes, such as protein folding. © *2005 American Institute of Physics*. [DOI: 10.1063/1.2132285]

I. INTRODUCTION

In spite of progress achieved in the last 50 years, calculation of the entropy and free energy remains a central problem in computer simulation, which affects physics, chemistry, biology, and engineering.[1,2] Recently, we have developed a new technique for calculating the entropy—the hypothetical scanning Monte Carlo (HSMC) method—and applied it to liquid argon, water,[3,4] and peptides in helical, extended, and hairpin states.[5] The aim of the present paper (as that of our preliminary study[6]) is to extend HSMC to lattice polymer models, and, in particular, to examine its applicability to random coil polymers.

It should be pointed out that lattice models have been utilized for studying a wide range of phenomena in polymer physics[7,8] as well as in structural biology, mainly related to protein folding and stability[9] (Refs. 7 and 8 present only very limited lists). Because of their simplicity, these models have been invaluable tools for understanding global properties that do not depend strongly on molecular details. Such models vary in complexity, ranging from self-avoiding walks on a square lattice to chain models on enriched three-dimensional (3D) lattices with a large effective coordination number.

Commonly, these systems are simulated by variants of Metropolis Monte Carlo (MC)—a method that enables one to generate samples of chain configurations i distributed according to their Boltzmann probability P_i^B from which equilibrium information can be extracted.[10] In many cases where the simulation moves are local, MC is referred to as a dynamical method. [It is noted, however, that the technique does not need to (and often does not) map the physical dynamics of the system.] Using MC it is straightforward to calculate properties that are measured directly from i, such as the potential energy E_i (that is obtained by summing up the atom-atom interactions) or geometrical quantities such as the radius of gyration. On the other hand, the *value* of P_i^B cannot be obtained in a straightforward manner, which makes it difficult to obtain the *absolute* entropy $S \sim -\ln P_i^B$ directly, i.e., as a by-product of the simulation (like E_i). There is a strong interest in S as a measure of order and as an essential ingredient of the free energy, $F=E-TS$, where T is the absolute temperature; F constitutes the criterion of stability, which is mandatory in the structure determination of proteins, for example. Furthermore, because MC simulations constitute models for dynamical processes, one would seek to calculate changes in F and S during a relaxation process, by assuming local equilibrium in certain parts along the MC trajectory; a classic example is the simulation of protein folding.[11]

S and F are commonly calculated by thermodynamic integration (TI) and perturbation techniques[1,2] that do not operate on a given MC sample but require conducting a *separate set* of MC simulations. This is a robust approach that enables one to calculate differences, $\Delta S_{\mathbf{ab}}$ and $\Delta F_{\mathbf{ab}}$, between two states **a** and **b** of a system; however, if the structural variance of such states is large (e.g., helical and hairpin states of a polypeptide) the integration from state **a** to **b** becomes difficult and in many cases unfeasible. On the other hand, if one could calculate the absolute $F_{\mathbf{a}}$ and $F_{\mathbf{b}}$

[a)] Author to whom correspondence should be addressed. Fax: 412-648-3163. Electronic mail: hagaim@pitt.edu directly from two separate sets of simulations carried out at states **a** and **b**, $\Delta F_{\mathbf{ab}}=F_{\mathbf{a}}-F_{\mathbf{b}}$ and the integration can be avoided. Still, the absolute F can also be obtained with TI provided that a reference state **r** is available, where the free energy is known exactly and the integration path between **r** and **a** (and **b**) is relatively short. A classic example is the calculation of F of liquid argon or water by integrating the free-energy change from an ideal-gas reference state. However, for nonhomogeneous systems such integration might not be trivial, and in models of peptides and proteins defining reference states that are close to the state of interest is a standing problem. It should be noted that F and S can always be obtained *approximately* from a given sample by harmonic or quasiharmonic methods[12] or by the local states method (see below).[13]

Another type of simulation method has been developed for polymers, where a chain is constructed step by step with transition probabilities (TPs).[14–17] The product of these TPs leads to P_i^B, hence S is known. To this category belongs the direct MC (DMC) (Ref. 14) procedure, the enrichment (Rosenbluth and Rosenbluth) and the dimerization methods,[15] the scanning method,[16] and other techniques.[17] However, these buildup procedures are not always the methods of choice mainly because they lack the dynamical aspects (and simplicity) of MC, which thus has become the commonly used method. Hence, it is important to develop methods for calculating the entropy from MC trajectories. Nonetheless, a hybrid of one buildup procedure, the scanning method,[16] with the dynamical MC approach has led to two approximate techniques, the local states[13] (LS) and hypothetical scanning (HS) methods.[18,19] These methods enable one to calculate S and F directly from a *given* sample generated by *any* simulation technique; they are general, and have been applied successfully to polymers, peptides, proteins, magnetic systems, and lattice-gas models.[2] Unlike the harmonic and quasiharmonic methods mentioned above,[12] HS and LS, in principle, can handle any chain flexibility, i.e., local fluctuations of a stable state (e.g., around an α-helix structure of a peptide), random coil fluctuations, as well as mixtures of these two extreme cases.

Recently, the HS method has been extended to fluids and has been further developed by replacing the deterministic partial future scanning used to calculate the TPs with a stochastic but complete scanning based on MC simulations;[3–5] this HSMC method has been applied very successfully to liquid argon, TIP3P water, and polyglycine molecules in helical, extended, and hairpin states.[3–5] HSMC is significantly more accurate than HS; it provides rigorous upper and lower bounds for F, which can be calculated from a relatively small sample and even from a *single* conformation.

As stated earlier, the aim of this paper is to extend the scope of HSMC to lattice polymer models, in particular, to random coil chains. For that we study self-avoiding walks (SAWs) on a square lattice—a difficult test case due to the strong excluded volume interactions occurring in two dimensions (2D). This paper is an extension of our recent Letter[6] in which part of Table I has been presented and discussed. We emphasize the generality of HSMC and discuss its unique aspects for lattice systems, which makes it a powerful research tool independent of existing techniques. The HSMC results are compared to those obtained some time ago by the scanning method,[20] to the results obtained by us using TI, to series-expansion values,[21] and to the results obtained by the HS method.

II. THEORY

A. Statistical mechanics of SAWs

Assume a single SAW of N steps (bonds), i.e., $N+1$ monomers starting from the origin on a square lattice. All the SAWs i are equally probable with the following Boltzmann probability:

$$P_i^B = 1/Z_{\mathrm{SAW}}, \tag{1}$$

where the partition function Z_{SAW} is the total number of different SAWs, and the free energy is thus

$$F/k_BT = -S/k_B = \sum_i P_i^B \ln P_i^B = -\ln Z_{\mathrm{SAW}} = \ln P_j^B, \tag{2}$$

where k_B is the Boltzmann constant, and j is *any* SAW. The summations (in i) in Eq. (2) and in the rest of the paper (except in Sec. II G) are over the *ensemble* of SAWs. Equation (2) demonstrates that F (and S for this particular model) has zero fluctuation, which is a general property of the *correct* free energy of any system. On the other hand, the fluctuation of a free-energy functional based on an *approximate* probability distribution (see below) is expected to be finite.[22] Equation (2) also shows that if the Boltzmann probability of any single SAW (j) is known, F (and S for this particular model) is known as well, which again is a general property satisfied by any system in equilibrium.

B. The direct Monte Carlo method

An unbiased sample of SAWs on a square lattice can be obtained by the DMC method.[14] With this method a nonreversal *random walk* (ideal chain) is generated step by step, where at step k the direction of the k^{th} bond is chosen at random (i.e., blindly) out of three possible directions (immediate chain reversal is not allowed). If the chosen site is unoccupied the bond is added to the existing chain, and the process continues; in the other case, the partial chain is discarded and a new one is started. This process is very inefficient for generating long SAWs due to strong (exponential) sample attrition. However, the entropy can be obtained from the relation $Z_{\mathrm{SAW}}/Z_{\mathrm{id}} \approx n_{\mathrm{suc}}/n_{\mathrm{start}}$ where Z_{id} is the known partition function of ideal chains, $Z_{\mathrm{id}}=4\times 3^{N-1}$, and n_{start} and n_{suc} are the number of chains started and the number of SAWs of N steps succeeded, respectively; this leads to an estimation S_{DMC} for the entropy of SAWs,

$$S_{\mathrm{DMC}}/k_B = \ln(4\times 3^{N-1} n_{\mathrm{suc}}/n_{\mathrm{start}}). \tag{3}$$

C. The scanning simulation method

The scanning method is a step-by-step construction (growth) procedure, where unlike DMC, the bonds are not selected blindly, but with TPs that are based on scanning possible SAWs in future steps;[16] thus, at step k of the pro-

TABLE I. HSMC results for the entropy per bond of N-bond SAWs. The results were obtained from n reconstructions of a straight chain. S^A [Eqs. (8) and (15)] is an upper bound, and σ_A is its fluctuation [Eq. (9)]. S^B [Eqs. (10), (11), and (16)] and its Gaussian approximation S_G^B [Eq. (17)] are lower bounds, and their averages with S^A are denoted by S^M [Eq. (12)] and S_G^M [Eq. (18)], respectively. S^D [Eq. (19)] is an exact entropy functional. n_{future} is related to the number of MC steps per bond (see text). $S_{\text{TI}}, S_{\text{scan}}, S_{\text{series}}$, and S_{HS} were obtained by thermodynamic integration [Eq. (26)], the scanning method [Eq. (6), Ref. 20], a series-expansion formula [Eq. (20)], and the HS method, respectively. The statistical error is defined by parentheses: 1.00(3)=1.00±0.03.

n_{future}	S^A/k_B	σ_A/k_B	S^B/k_B	S_G^B/k_B	S^M/k_B	S_G^M/k_B	S^D/k_B	n
			$N=29$	$S_{\text{DMC}}=1.016\,147(5)$				
500	1.020 84(2)	0.018 07(2)	1.011 48(5)	1.01 137(3)	1.016 16(3)	1.016 10(2)	1.016 14(3)	1 250 000
5000	1.016 62(2)	0.005 68(2)	1.015 68(3)	1.015 68(2)	1.016 15(2)	1.016 15(2)	1.016 15(2)	125 000
50 000	1.016 18(2)	0.001 81(2)	1.016 09(3)	1.016 09(2)	1.016 14(2)	1.016 14(2)	1.016 14(2)	125 00
S_{TI}	1.016 145(3)		1.016 145(3)	1.016 145(3)	1.016 145(3)	1.016 145(3)	1.016 145(3)	
S_{series}	1.016 15(1)		1.016 15(1)	1.016 15(1)	1.016 15(1)	1.016 15(1)	1.016 15(1)	
			$N=49$	$S_{\text{scan}}=1.000\,904(4)$				
500	1.005 83(1)	0.014 24(2)	0.996 02(5)	0.995 90(3)	1.000 93(3)	1.000 86(2)	1.000 91(3)	1 250 000
5000	1.001 40(1)	0.004 48(2)	1.000 42(3)	1.000 42(1)	1.000 91(2)	1.000 91(1)	1.000 91(2)	125 000
50 000	1.000 95(1)	0.001 42(2)	1.000 85(3)	1.000 85(1)	1.000 90(2)	1.000 90(1)	1.000 90(2)	12 500
S_{HS}	1.001 49(1)	0.004 34(1)	1.000 26(2)	1.000 57(1)	1.000 88(2)	1.001 03(1)	1.000 94(1)	250 000
S_{TI}	1.000 897(3)		1.000 897(3)	1.000 897(3)	1.000 897(3)	1.000 897(3)	1.000 897(3)	
S_{series}	1.000 899(5)		1.000 899(4)	1.000 899(4)	1.000 899(4)	1.000 899(4)	1.000 899(4)	
			$N=99$	$S_{\text{scan}}=0.987\,726(5)$				
500	0.992 94(2)	0.010 30(3)	0.9826(1)	0.982 43(6)	0.987 75(5)	0.987 69(4)	0.987 73(5)	250 000
5000	0.988 26(2)	0.003 24(3)	0.987 22(5)	0.987 22(3)	0.987 74(3)	0.987 74(2)	0.987 74(3)	25 000
50 000	0.987 77(2)	0.001 01(3)	0.987 67(4)	0.987 67(2)	0.987 72(2)	0.987 72(2)	0.987 72(3)	2500
S_{HS}	0.989 94(1)	0.005 07(1)	0.9855(2)	0.98 74(1)	0.9878(1)	0.9887(1)	0.988 17(5)	250 000
S_{TI}	0.987 727(3)		0.987 727(3)	0.987 727(3)	0.987 727(3)	0.987 727(3)	0.987 727(3)	
S_{series}	0.987 730(3)		0.987 730(3)	0.987 730(3)	0.987 730(3)	0.987 730(3)	0.987 730(3)	
			$N=149$	$S_{\text{scan}}=0.982\,740(3)$				
500	0.988 06(2)	0.008 52(3)	0.9774(2)	0.977 25(8)	0.9827(1)	0.982 65(4)	0.9827(1)	250 000
5000	0.983 29(2)	0.002 67(3)	0.982 22(5)	0.982 23(3)	0.982 76(3)	0.982 76(2)	0.982 76(3)	25 000
50 000	0.982 81(2)	0.000 85(3)	0.982 70(4)	0.982 70(2)	0.982 75(2)	0.982 75(2)	0.982 75(3)	2500
S_{TI}	0.982 742(3)		0.982 742(3)	0.982 742(3)	0.982 742(3)	0.982 742(3)	0.982 742(3)	
S_{series}	0.982 740(2)		0.982 740(2)	0.982 740(2)	0.982 740(2)	0.982 740(2)	0.982 740(2)	
			$N=249$	$S_{\text{scan}}=0.978\,36(2)$				
500	0.983 91(3)	0.006 69(4)	0.9727(3)	0.9728(1)	0.9783(2)	0.978 33(7)	0.9783(2)	63 000
5000	0.978 89(2)	0.002 08(4)	0.977 82(8)	0.977 82(5)	0.978 36(4)	0.978 36(3)	0.978 36(5)	9100
50 000	0.978 40(2)	0.000 66(4)	0.978 29(5)	0.978 29(2)	0.978 35(3)	0.978 35(2)	0.978 35(3)	930
S_{HS}	0.983 06(1)	0.004 01(1)	0.9745(5)	0.9791(3)*	0.9788(3)	0.9811(2)	0.9799(1)	176 000
S_{TI}	0.978 358(4)		0.978 358(4)	0.978 358(4)	0.978 358(4)	0.978 358(4)	0.978 358(4)	
S_{series}	0.978 360(1)		0.978 360(1)	0.978 360(1)	0.978 360(1)	0.978 360(1)	0.978 360(1)	
			$N=399$	$S_{\text{scan}}=0.975\,67(4)$				
500	0.981 38(6)	0.005 40(5)	0.9710(5)	0.9697(2)	0.9762(3)	0.9756(1)	0.9759(3)	9500
5000	0.976 25(4)	0.001 70(5)	0.9751(1)	0.975 09(8)	0.975 67(5)	0.975 67(5)	0.975 67(5)	2000
50 000	0.975 68(4)	0.000 53(5)	0.975 57(7)	0.975 57(5)	0.975 63(4)	0.975 63(4)	0.975 63(5)	225
S_{HS}	0.981 41(5)	0.003 35(5)	0.9743(5)	0.9769(3)	0.9779(3)	0.9792(2)	0.9782(2)	5500
S_{TI}	0.975 655(8)		0.975 655(8)	0.975 655(8)	0.975 655(8)	0.975 655(8)	0.975 655(8)	
S_{series}	0.975 652(1)		0.975 652(1)	0.975 652(1)	0.975 652(1)	0.975 652(1)	0.975 652(1)	
			$N=599$	$S_{\text{scan}}=0.973\,95(5)$				
500	0.980 03(8)	0.004 45(7)	0.9706(8)	0.9682(4)	0.9753(4)	0.9741(2)	0.9748(5)	3000
5000	0.974 66(7)	0.001 39(7)	0.9736(2)	0.9735(1)	0.9741(1)	0.974 08(9)	0.9741(1)	450
50 000	0.974 13(5)	0.000 36(7)	0.9741(1)	0.974 05(6)	0.974 09(6)	0.974 09(5)	0.974 09(5)	45
S_{TI}	0.974 04(1)		0.974 04(1)	0.974 04(1)	0.974 04(1)	0.974 04(1)	0.974 04(1)	
S_{series}	0.974 025(1)		0.974 025(1)	0.974 025(1)	0.974 025(1)	0.974 025(1)	0.974 025(1)	

cess, $k-1$ directions (bonds), ν will have already been constructed (these bond directions at each step are denoted by $\nu_1, \ldots, \nu_{k-1}$, where $\nu=1,4$). To determine the direction ν_k(out of three possible directions ν) one enumerates all the possible continuations $Z_k^\nu(f)$ of the chain in a limited number of f future steps (typically less than the remaining bonds) that start from ν of step k, where $Z_k^\nu(f)$ is a partial future partition function and f is the scanning parameter. $Z_k^\nu(f)$ enables one to define TPs for ν,

$$p(\nu|\nu_{(k-1)},\ldots,\nu_1,f) = Z_k^\nu(f)/\sum_{\nu=1}^{4} Z_k^\nu(f), \tag{4}$$

where because immediate reversal is forbidden, the summation is only over the three allowed directions. Using these TPs, the k^{th} step is determined by a random number and the process continues. The construction probability $P_i^0(f)$ of SAW i is the product of the TPs with which the steps have been chosen,

$$P_i^0(f) = \prod_{k=1}^{N} p(\nu_k|\nu_{(k-1)},\ldots,\nu_1,f). \tag{5}$$

For $f \ll N$ the scanning is incomplete and $P_i^0(f)$ is approximate. Due to this "incomplete" scanning, the chain can get trapped in a dead end during construction, meaning that the number n of completed constructions is smaller than the number n_{start} of those started.

In other words, $P_i^0(f)$ is normalized over a subgroup of the random walks that includes all the SAWs and part of the self-intersecting walks. Also, $P_i^0(f)$ is biased, i.e., unlike P_i^B, it is larger for the compact SAWs than for the open ones. This bias can be decreased *systematically* by increasing f, where for a complete future scanning, i.e., $f_{\max}=N-k+1$, the TPs [Eq. (4)] become exact and no trapping occurs.[16] In practical applications the bias is removed by an *importance sampling* procedure, which leads to an unbiased estimation $\bar{S}$ that is exact within the statistical error,

$$\bar{S}/k_B = \ln \frac{1}{n_{\text{start}}} \sum_{t=1}^{n} \frac{1}{P_t^0(f)}. \tag{6}$$

The scanning method can easily be extended to a chain model with finite interactions; in this case the interaction energy $E_{j(\nu)}^k(f)$ of the future chain j that starts from ν with itself and with the rest of the chain is calculated and the corresponding Boltzmann factor contributes to $Z_k^\nu(f)$, rather than 1,

$$Z_k^\nu(f) = \sum_{j(\nu)} \exp[-E_{j(\nu)}^k(f)/k_B T]. \tag{7}$$

D. The hypothetical scanning (HS) method

The HS method (as well as the local states method) is based on the concept that two samples in equilibrium generated by different simulation methods are equivalent in the sense that they both lead to the same estimates (within the statistical errors) of average properties, such as the entropy, energy, and their fluctuations. Relying on this equivalence, one assumes that a given sample of SAWs constructed by *any* exact procedure [e.g., Metropolis MC (Ref. 10) has instead been generated with the scanning method. Thus, for each of the bonds $[\nu_k(i)]$ of SAW i one calculates the TP [Eq. (4)] as if i had been generated with the scanning method (we call this process the *reconstruction* of i, essentially an analysis procedure for obtaining TPs). The product of these TPs leads to $P_i^0(f)$ [Eq. (5)] and to a functional S^A, which can be shown *rigorously* (using Jensen's inequality) to be an upper bound for S,[19]

$$S^A = -k_B \sum_i P_i^B \ln P_i^0(f), \tag{8}$$

where i runs on the *complete* ensemble of SAWs and S^A is a function of f. Because the sample is generated with an exact simulation procedure, S^A is a statistical average defined with the Boltzmann probability, which is normalized over the ensemble of SAWs. Each SAW i is associated with the variable $\ln P_i^0(f)$, where $P_i^0(f)$ is normalized over a larger ensemble that also consists of self-intersecting walks. The fluctuation σ_A of $\ln P_i^0(f)$,

$$\sigma_A = \left\{ \sum_{\text{SAWs } i} P_i^B [S^A + k_B \ln P_i^0(f)]^2 \right\}^{1/2}, \tag{9}$$

is expected to be larger than zero, decreasing with increasing f (i.e., with improving the approximation), which means that S^A and σ_A are positively correlated. This correlation has been found to exist for good enough approximations.[22]

One can define another approximate entropy functional denoted by S^B,[19]

$$S^B = -k_B \sum_{\text{SAWs } i} P_i(f) \ln P_i^0(f), \tag{10}$$

where $P_i(f)=P_i^0(f)/\Sigma P_i^0(f)$. If $P_i^0(f)$ was replaced in Eq. (10) by $P_i(f)$, according to the free-energy minimum principle,[23] S^B would become a lower bound for S; but for SAWs it can only be shown rigorously[19] that $S^B \leq S^A$. However, when reliably estimated for a good enough approximation, S^B has been found in most cases[19,24] to underestimate S, as is also shown in the present calculations. In practice, lower bound behavior can be verified if S^B increases as the approximation improves; one can then assume that this trend would continue for better approximations meaning that S^B would converge to S. S^B can be estimated from a sample of size n by importance sampling,

$$\bar{S}^B = -(k_B/n)\left[\sum_{t=1}^{n} P_{i(t)}^0(f) \ln P_{i(t)}^0(f)\right] / \sum_{t=1}^{n} P_{i(t)}^0(f), \tag{11}$$

where $i(t)$ is SAW i obtained at time t of the correct Boltzmann simulation and the bar above S^B denotes estimation. However, the statistical reliability of this estimation (unlike the estimation of S^A) decreases sharply with increasing chain length, because the overlap between the probability distributions P_i^B and $P_i(f)$ decreases exponentially.

If S^B is a lower bound for S and the deviations of S^A and S^B from S (in the absolute values) are approximately equal, their average S^M becomes a better approximation than either of them individually,

$$S^M = [S^A + S^B]/2. \tag{12}$$

Typically, several approximations for S^A, S^B, and S^M are calculated as a function of f, and their convergence enables one to determine the correct entropy with high accuracy. While application of HS to SAWs has been found to be quite efficient, for structured molecules such as an α helix of a peptide HS has failed because it is impossible to carry out the future scanning within the limited conformational space defined by the local fluctuations of this structure, hence to define appropriate TPs. As discussed below, this problem does not exist with HSMC.

E. The HSMC method

While the TPs defined by HS are deterministic (based on the *entire* conformational space defined by f at step k of the reconstruction process), for a large chain they are always approximate, i.e., $f \ll N$ due to the exponential growth (with f) of the number of future SAWs. The HSMC method overcomes this limitation by seeking to estimate the *exact* TP defined by Eq. (4) with $f_{\max}=N-k+1$, i.e., the whole future is scanned at step k. This is achieved by replacing the exact enumeration of f future steps at k by a MC simulation of the entire future segment of the chain (i.e., steps $k, k+1, \ldots, N$) in the presence of the "frozen past" $[\nu_1, \ldots, \nu_{k-1}]$. The TP denoted by p^{MC} of the actual direction $\nu_k(i)$ in the reconstructed SAW i is calculated from the number of MC steps $n_k^{\nu(i)}$ for which $\nu_k(i)$ was visited during the simulation of total n_{MC} steps at step k,

$$p^{\mathrm{MC}}(\nu_k(i)|\nu_{(k-1)}, \ldots, \nu_1) = n_k^{\nu(i)}/n_{\mathrm{MC}}, \tag{13}$$

and the reconstruction probability of chain i is

$$P_i^{\mathrm{MC}} = \prod_{k=1}^{N} p^{\mathrm{MC}}(\nu_k(i)|\nu_{(k-1)}, \ldots, \nu_1), \tag{14}$$

where, for simplicity, i has been omitted in the TPs. In Eqs. (13) and (14) and in the rest of this paper, for brevity, we denote by MC physical quantities calculated by HSMC; notice, however, that in previous publications these properties were denoted by HS, which in this paper is reserved to denote the results obtained with the HS method. Unlike the deterministic $P_i^0(f)$ [Eq. (5)], P_i^{MC} is defined stochastically. The fact that the entire future is considered is important for systems with strong long-range interactions such as SAWs, proteins, etc; also, unlike $P_i^0(f)$ that is defined over the ensemble of SAWs and part of the ensemble of self-intersecting walks, P_i^{MC} is defined *only* over the ensemble of SAWs. As discussed below, this property of P_i^{MC} distinguishes HSMC from HS in many respects. Still, p^{MC} hence P_i^{MC} are approximate (due to finite simulation lengths), but one can show that as the MC simulation is increased, $p^{\mathrm{MC}} \to p^{\mathrm{exact}}$ and $P_i^{\mathrm{MC}} \to P_i^{\mathrm{B}}$, meaning that S can be estimated by reconstructing a single SAW. In practice, however, P_i^{MC} is approximate, leading to an upper bound for S [compare with Eq. (8)],

$$S^A = -k_B \sum_i P_i^B \ln P_i^{\mathrm{MC}} = \sum_i P_i^B S_i^{\mathrm{MC}}, \tag{15}$$

where $S_i^{\mathrm{MC}}=-k_{\mathrm{B}} \ln P_i^{\mathrm{MC}}$. It can be shown (see Appendix of Ref. 4) that like S^A [Eq. (8)], S^A [Eq. (15)] defined with stochastic probabilities P_i^{MC} is a rigorous upper bound, which is expected to have nonzero fluctuation σ_A [Eq. (9)].

One can define the entropy functional S^B [Eq. (10)] and thus S^M [Eq. (12)] also for HSMC, where S^B becomes a rigorous lower bound of S due to the fact that P_i^{MC} is defined only over the ensemble of SAWs. We express S^B as

$$\begin{aligned} S^B &= -k_B \sum_i P_i^{\mathrm{MC}} \ln P_i^{\mathrm{MC}} = -k_B \frac{\Sigma_i P_i^B [P_i^{\mathrm{MC}} \ln P_i^{\mathrm{MC}}]}{\Sigma_i P_i^B P_i^{\mathrm{MC}}} \\ &= \frac{\Sigma_i P_i^B \exp[-S_i^{\mathrm{MC}}/k_B][S_i^{\mathrm{MC}}]}{\Sigma_i P_i^B \exp[-S_i^{\mathrm{MC}}/k_B]}, \end{aligned} \tag{16}$$

where it is estimated by Eq. (11). Equation (16) emphasizes an *explicit* dependence of S^B on the variable $S_i^{\mathrm{MC}} = -k_B \ln P_i^{\mathrm{MC}}$, that is directly related to the average S^A [Eq. (15)] and its fluctuation σ_A [defined in the same manner as in Eq. (9)]. Because of the stochastic nature of S_i^{MC} it is plausible to assume that when configurations (i) are sampled from the Boltzmann distribution (i.e., with P_i^B), their corresponding S_i^{MC} values occur with a Gaussian probability centered around S^A with standard deviation σ_A. Indeed, such Gaussian behavior has been observed in models for liquid argon and TIP3P water, which has led (see details in Ref. 4) to a Gaussian approximation S_G^B for S^B,

$$S_G^B = -\frac{(\sigma_A)^2}{k_B} + S^A, \tag{17}$$

and to the corresponding S_G^M [see Eq. (12)],

$$S_G^M = (S^A + S_G^B)/2 = S^A - \frac{1}{2}\frac{(\sigma_A)^2}{k_B}. \tag{18}$$

The fact that S_G^B depends only on S^A and σ^A is an advantage because these quantities are typically easier to estimate than S^B (directly) from Eqs. (10), (11), or (16), meaning that S_G^B is expected to be statistically more reliable than S^B. Previous results have shown that this Gaussian distribution is a very good approximation as there is excellent agreement of F_G^B with F^B for cases where F^B is well converged (when finite interactions are defined F replaces S). Again, several approximations for S^A, S_G^B, and S_G^M can be calculated, and their convergence leads to highly accurate free-energy determination. It should be pointed out that formally one can calculate S_G^B also for S_i^{HS} defined by $P_i^0(f)$ of the HS method. However, S_i^{HS} (unlike S_i^{MC}) is not stochastic and thus deviates from a Gaussian distribution, where this deviation increases as the approximation worsens, i.e., with increasing chain length N.

The entropy can be expressed exactly by S^D (see Ref. 4), which can also be estimated from a sample generated with P_i^B. One obtains

$$\begin{aligned} S^D &= -k_B \ln \sum_i P_i^B P_i^{\mathrm{MC}} \\ &= -k_B \ln \left[\sum_i P_i^B [\exp(-S_i^{\mathrm{MC}}/k_B)] \right]. \end{aligned} \tag{19}$$

In practice, the efficiency of estimating S by S^D depends on the fluctuation of this statistical average, which is determined by the fluctuation of S_i^{MC} exponentiated. That is, if the fluctuations in S_i^{MC} are small, then the values for $\exp(-S_i^{\mathrm{MC}}/k_B)$ do not vary drastically, and the averages for S^D (and S^B) can be estimated reliably from a relatively small sample. Still (as for S^B), the direct calculation of S through S^D will not be as statistically reliable as estimating S^A. Obviously, as $S_i^{\mathrm{MC}} \to S$ (i.e., $P_i^{\mathrm{MC}} \to P_i^B$) all fluctuations become zero and S can be obtained from a single configuration. We note additionally that to improve convergence, S^D (like S^B) can be approximated by the Gaussian distribution [for the S_i^{MC} values in Eq. (19)]; applying this approximation leads to S_G^M defined in Eq. (18).

As for S_G^B, one can formally calculate S^D also for $P_i^0(f)$ defined by HS. However, because $P_i^0(f)$ is not defined only on the ensemble of SAWs, $S^D(\mathrm{HS})$ [unlike $S^D(\mathrm{MC})$] will not converge to the correct S even for a very large sample. Convergence could occur for $P_i(f)=P_i^0(f)/\Sigma P_i^0(f)$ which is normalized over the SAWs alone [i.e., $\Sigma P_i^0(f)<1$]; however, calculation of $\Sigma P_i^0(f)$ by HS is impossible. This suggests that S^D calculated by HS for a large chain will always be an upper bound for S.

While the theory above has been introduced for the entire ensemble of SAWs, it also applies to a set of reconstructions of a single chain conformation (see Appendix, Ref. 4). That is, the required averages can be obtained from a set of n independent reconstructions of the same chain, where each reconstruction contributes an estimation for S_i^{MC}. For S^A, for example, these estimations are arithmetically averaged; for S^D the arithmetic average of $\exp(-S_i^{\mathrm{MC}}/k_B)$ is used, etc.

F. Calculation of the entropy by series expansion

For comparison we also present the results obtained with a formula based on series-expansion (exact enumeration) data.[21] The entropy S_{series} is obtained from the total number of SAWs c_N,

$$S_{\mathrm{series}}/k_B = \ln c_N \cong \ln\{\mu^N[a_1 N^{11/32} + a_2 N^{-21/32} + b_1 N^{-37/32} + (-1)^N d_1 N^{-3/2} + (-1)^N d_2 N^{-2}]\}, \tag{20}$$

where $a_1=1.1771(2)$, $a_2=0.554(2)$, $b_1=-0.19(2)$, $d_1=-0.19(2)$, $d_2=0.034(2)$, and $\mu=2.638\,158\,5(10)$ (the error of the last digits appears in parenthesis).

G. Calculation of the entropy by thermodynamic integration

In order to calculate the partition function of a SAW via TI,[25] the system must be linked with a calculable reference state, which in this case is the ideal chain. Samples of chains are generated where monomers are allowed to overlap each other. To effect this, a unitless energy function E is defined where

$$E = \sum_j \varphi_j . \tag{21}$$

φ_j is the "overlap value" at lattice site j, and the summation is carried out over all sites. The overlap value is defined as follows. A lattice site that is occupied by only a single monomer (or is unoccupied) contributes nothing to the energy ($\varphi_j=0$). A doubly occupied site (i.e., a single overlap) contributes $\varphi_j=1$; a triply occupied site (a double overlap) contributes $\varphi_j=2$, and so on. The value of E is thus always an integer. For a SAW, E must be zero.

The above-defined energy function is used to describe a *general* chain which can exist at any arbitrary finite temperature. The partition function for the general chain ensemble is given by

$$Z = \sum_{\mathrm{id}} \exp[-E_i/T], \tag{22}$$

where the sum is carried out over all *ideal chain* configurations i (the total configuration space), and where we have introduced a unitless temperature T. We note that at high (infinite) T, the Boltzmann factor $\exp[-E_i/T]$ is unity and the partition function approaches that of the ideal chain reference state (i.e., $Z_{\mathrm{id}}=4\times 3^{N-1}$, where immediate reversal is forbidden). At low $T(T=0)$, only zero energy configurations will contribute to the summation and the partition function becomes that of the SAW, Z_{SAW}.

The difference $\ln Z_{\mathrm{SAW}} - \ln Z_{\mathrm{id}}$ can be evaluated by integration over T or over $1/T$, using the derivative relations,

$$\frac{d\ln Z}{dT} = \left(\frac{1}{Z}\right)\sum_{\mathrm{id}} \frac{E_i}{T^2}\exp[-E_i/T] = \left\langle \frac{E}{T^2} \right\rangle, \tag{23}$$

and

$$\frac{d\ln Z}{d(1/T)} = \left(\frac{1}{Z}\right)\sum_{\mathrm{id}} -E_i \exp[-E_i/T] = -\langle E\rangle . \tag{24}$$

The corresponding integrals are, respectively,

$$\ln\left[\frac{Z_{\mathrm{SAW}}}{Z_{\mathrm{id}}}\right] = \int_{\infty}^{0} \frac{d\ln Z}{dT} dT \quad \text{and} \quad \ln\left[\frac{Z_{\mathrm{SAW}}}{Z_{\mathrm{id}}}\right] = \int_{0}^{\infty} \frac{d\ln Z}{d(1/T)} d(1/T). \tag{25}$$

We have chosen to use both of these relations where we conduct the integration in two stages as

$$\ln\left[\frac{Z_{\mathrm{SAW}}}{Z_{\mathrm{id}}}\right] = \int_{0}^{1/T^*} \frac{d\ln Z}{d(1/T)} d(1/T) + \int_{T^*}^{0} \frac{d\ln Z}{dT} dT, \tag{26}$$

where T^* is an intermediate temperature. The left-hand term thus quantifies the change in $\ln Z$ for going from an ideal chain to the general chain at T^*, and the right-hand term is the change from this point to the SAW.

In our implementation, the generalized chain was simulated at a total of 199 temperatures. The relevant result in these simulations is the average energy $\langle E\rangle$; these values are used for derivative points [Eqs. (23) and (24)] in the numerical evaluation of Eq. (26). In the first stage/series [corresponding to the left-hand term in Eq. (26)], 100 simulation temperatures were spaced evenly in $1/T$, ranging from $1/T=0$ (the ideal chain) to $1/T^*$ where $T^*=0.757\,575\,75$ ($1/T^*=1.32$). In the second stage [for the right-hand term in Eq. (26)], 100 simulation temperatures were spaced evenly in T, ranging from T^* to $T=0$ (the SAW). With the finite limits in Eq. (26), a simple trapezium integration was adequate. It should also be noted that the number of simulation points employed in this work was actually well more than was necessary. We note further that the results are insensitive to the choice of T^* as long as there are enough points. One could drastically reduce the number of simulation points (thereby increasing the efficiency) by careful (optimized) choice of T^* and/or by implementing less simple-minded quadrature techniques; however, the present performance is sufficient for our purposes. Details about the MC simulations appear below in Sec. III A.

III. RESULTS AND DISCUSSION

We have calculated the entropy of SAWs consisting of $N=29$, 49, 99, 149, 249, 399, and 599 bonds. The results of Table I were obtained by reconstructing a single chain conformation (see Appendix, Ref. 4), i.e., by n replicate reconstructions (based on different sets of random numbers) of a *straight* SAW of N bonds, while the results in Table II were obtained by reconstructing a sample of SAWs.

A. MC simulations and the HSMC reconstruction procedure

The efficiency of HSMC is affected considerably by the MC procedure employed in the reconstruction process. On a square lattice, "crankshaft" moves are in most cases rejected due to the strong excluded volume interactions while corner moves have somewhat higher acceptance rate.[8] Therefore, for the reconstruction process we have used a MC procedure based on 50% corner moves (that provide local conformational changes) and 50% "pivot" moves that have been shown to effectively induce global changes.[26] This procedure has been employed not only in the reconstruction process, but also for generating samples of SAWs (to be reconstructed by HSMC and HS) and for the TI simulations.

The HSMC calculations are based on the sample size n, the number of reconstructed SAWs and n_{future}, which is related to the number of future MC steps per bond applied during the reconstruction process as defined below. First we note that the first bond of the chain is not reconstructed; its probability is always 1/4. The number of MC steps n_{MC} for bond k is scaled as $n_{\mathrm{MC}}=(N-k+1)n_{\mathrm{future}}$, meaning that the maximal number of future MC steps is applied for the reconstruction of the second bond (to which corresponds the largest future segment of $N-1$ bonds), while the last bond (N) is allotted the minimal number of MC steps. Because each simulation at step k always starts from the structure of the reconstructed chain it is important to let the future SAW equilibrate, otherwise p^{MC} [Eq. (13)] would (on average) be too high; therefore, 300 MC steps per future bond are used for equilibration. As discussed earlier, the larger is n_{future} the better (i.e., smaller) is S^A [Eq. (15)], the larger is S^B [Eq. (16)] [and S_G^B, Eq. (17)] and the smaller is the fluctuation σ_A [Eq. (9)]. To demonstrate this effect, the results for each chain length are presented in the tables for $n_{\mathrm{future}}=500$, 5000, and 50 000, where the corresponding sample size n is decreased, which results in approximately the same computer time for each calculation. Notice that for a single chain (Table I), n is the number of reconstructions applied to the same straight chain, while for a sample of chains (Table II), n is the number of different configurations reconstructed, one reconstruction is performed for each configuration.

For the TI process the chains were simulated as described above, by the 50/50 ratio of pivot and corner moves, where in this case the entire chain is moveable (except of the first bond). The total simulation length was the same at each temperature, however, it varied depending on chain size. 6×10^7 MC moves were carried out at each temperature for $N=29$, where run lengths of $10^8, 10^8, 6\times10^7, 5\times10^7, 3.2\times10^7$, and 2.4×10^7 steps were used for $N=49$, 99, 149, 249, 399, and 599, respectively. All of these runs were replicated nine times (i.e., nine independent simulations were performed), thus yielding nine independent integration results (trials) for each chain size. Our final reported result is the average of these trials, with the standard deviation of the mean being used as the uncertainty estimate.

B. Results by TI, series expansion, and the scanning method

To a large extent, we judge the performance of HSMC by comparing its results to those obtained by other techniques, such as the scanning method [Eq. (6), Ref. 20], series expansion [Eq. (20)], TI [Eq. (26)], and HS (using $f=8$); therefore, we start by discussing the results of these methods which appear in both tables.

We first would like to point out the surprising accuracy for large N obtained by the series-expansion formula [Eq. (20)] that is based on extrapolating exact enumeration data for relatively short chains. Thus, the results for S_{TI} and S_{series} are equal within the error bars for all N, with comparable errors for $N=49$, 99, and 149. However, for $N=29$ the error in S_{series} is significantly larger than that of S_{TI} and for $N>149$ error(S_{TI})-error(S_{series}) increases constantly with N. For $N=29$ the DMC and TI values are equal within comparable errors.

The results obtained with the scanning method long ago[20] (based on a relatively small scanning parameter $f=6$) are also very good. They are equal to the TI and series results for all N accept for $N=599$, where S_{scan} is smaller due to a bias (for generating compact chains) that was not removed completely by the importance sampling procedure [Eq. (6)]. For $N\geqslant249$ the statistical errors of S_{scan} are significantly larger than those of S_{TI}. In what follows, for comparison we shall consider the TI and series results to be exact.

C. Entropy by reconstructing straight chains

The results obtained by n replicate reconstructions of a straight chain appear in Table I. Part of the data has already been provided in Ref. 6; however, the HS results and those for S_G^B, S_G^M, and for the chain length $N=29$ are new.

The table supports the expectations of the HSMC theory presented in Sec. II. Thus, for all chain lengths, as n_{future} is increased from 500 to 50 000, the fluctuation decreases, S^A decreases and remains an upper bound, and S^B and S_G^B increase remaining lower bounds. For $n_{\mathrm{future}}=500$ the S_G^B re-

TABLE II. HSMC results for the entropy per bond obtained from a sample of chain configurations. For details, see the caption of Table I.

n_{future}	S^A/k_B	σ_A/k_B	S^B/k_B	S_G^B/k_B	S^M/k_B	S_G^M/k_B	S^D/k_B	n
			$N=29$	$S_{\text{DMC}}=1.016\ 147(5)$				
500	1.023 66(2)	0.023 19(2)	1.008 66(5)	1.008 07(3)	1.016 16(3)	1.015 87(2)	1.016 09(3)	1 250 000
5000	1.016 89(2)	0.007 24(2)	1.015 37(3)	1.015 37(2)	1.016 13(2)	1.016 13(2)	1.016 13(2)	125 000
50 000	1.016 21(2)	0.002 37(2)	1.016 04(3)	1.016 05(2)	1.016 13(2)	1.016 13(2)	1.016 13(2)	12 500
S_{TI}	1.016 145(3)		1.016 145(3)	1.016 145(3)	1.016 145(3)	1.016 145(3)	1.016 145(3)	
S_{series}	1.016 15(1)		1.016 15(1)	1.016 15(1)	1.016 15(1)	1.016 15(1)	1.016 15(1)	
			$N=49$	$S_{\text{scan}}=1.000\ 904(4)$				
500	1.009 59(2)	0.019 23(2)	0.992 15(7)	0.991 46(4)	1.000 87(4)	1.000 53(3)	1.000 78(4)	1 248 547
5000	1.001 72(2)	0.006 00(2)	0.999 96(5)	0.999 95(2)	1.000 84(3)	1.000 84(2)	1.000 84(3)	124 763
50 000	1.000 94(2)	0.001 89(2)	1.000 77(4)	1.000 77(2)	1.000 86(2)	1.000 86(2)	1.000 86(3)	12 467
S_{HS}	1.001 49(1)	0.004 34(1)	1.000 26(2)	1.000 57(1)	1.000 88(2)	1.001 03(1)	1.000 94(1)	250 000
S_{TI}	1.000 897(3)		1.000 897(3)	1.000 897(3)	1.000 897(3)	1.000 897(3)	1.000 897(3)	
S_{series}	1.000 899(5)		1.000 899(4)	1.000 899(4)	1.000 899(4)	1.000 899(4)	1.000 899(4)	
			$N=99$	$S_{\text{scan}}=0.987\ 726(5)$				
500	0.998 40(3)	0.015 39(3)	0.9762(2)	0.9750(1)	0.9873(1)	0.98668(5)	0.9873(1)	249 621
5000	0.988 83(3)	0.004 78(3)	0.9866(1)	0.986 57(4)	0.987 70(5)	0.987 70(3)	0.987 70(5)	24 907
50 000	0.987 86(3)	0.001 53(3)	0.987 63(5)	0.987 63(3)	0.987 75(3)	0.987 75(3)	0.987 75(3)	2476
S_{HS}	0.989 94(1)	0.005 07(1)	0.9856(2)	0.9874(1)	0.9878(1)	0.988 7(1)	0.988 17(5)	250 000
S_{TI}	0.987 727(3)		0.987 727(3)	0.987 727(3)	0.987 727(3)	0.987 727(3)	0.987 727(3)	
S_{series}	0.987 730(3)		0.987 730(3)	0.987 730(3)	0.987 730(3)	0.987 730(3)	0.987 730(3)	
			$N=149$	$S_{\text{scan}}=0.982\ 740(3)$				
500	0.994 60(3)	0.013 47(5)	0.9688(5)	0.9676(2)	0.9817(3)	0.9811(1)	0.9818(3)	249 628
5000	0.983 98(3)	0.004 28(5)	0.9813(1)	0.981 26(7)	0.982 63(5)	0.982 62(4)	0.982 64(5)	24 860
50 000	0.982 93(3)	0.001 43(5)	0.982 62(5)	0.982 62(4)	0.982 77(3)	0.982 77(3)	0.982 77(4)	2470
S_{TI}	0.982 742(3)		0.982 742(3)	0.982 742(3)	0.982 742(3)	0.982 742(3)	0.982 742(3)	
S_{series}	0.982 740(2)		0.982 740(2)	0.982 740(2)	0.982 740(2)	0.982 740(2)	0.982 740(2)	
			$N=249$	$S_{\text{scan}}=0.978\ 36(2)$				
500	0.991 88(5)	0.011 49(7)	0.961(2)	0.9590(4)	0.976(1)	0.9755(2)	0.976(1)	50 451
5000	0.979 77(4)	0.003 74(7)	0.9760(2)	0.9763(1)	0.9779(1)	0.978 03(8)	0.9780(1)	7261
50 000	0.978 51(4)	0.001 29(7)	0.9781(1)	0.978 09(6)	0.978 30(5)	0.978 30(5)	0.978 30(5)	938
S_{HS}	0.983 06(1)	0.004 01(1)	0.9745(5)	0.9791(3)*	0.9788(3)	0.9811(2)	0.9799(1)	176 000
S_{TI}	0.978 358(4)		0.978 358(4)	0.978 358(4)	0.978 358(4)	0.978 358(4)	0.978 358(4)	
S_{series}	0.978 360(1)		0.978 360(1)	0.978 360(1)	0.978 360(1)	0.978 360(1)	0.978 360(1)	
			$N=399$	$S_{\text{scan}}=0.975\ 67(4)$				
500	0.9908(1)	0.0099(1)	0.955(3)	0.9517(8)	0.973(2)	0.9712(4)	0.971(2)	6670
5000	0.977 29(8)	0.0032(1)	0.9727(5)	0.9733(3)	0.9750(3)	0.9753(2)	0.9752(3)	1577
50 000	0.9759(1)	0.0012(1)	0.9754(2)	0.9754(1)	0.9757(1)	0.9757(1)	0.9757(1)	115
S_{HS}	0.981 41(5)	0.003 35(5)	0.9743(5)	0.9769(3)	0.9779(3)	0.9792(2)	0.9782(2)	5500
S_{TI}	0.975 655(8)		0.975 655(8)	0.975 655(8)	0.975 655(8)	0.975 655(8)	0.975 655(8)	
S_{series}	0.975 652(1)		0.975 652(1)	0.975 652(1)	0.975 652(1)	0.975 652(1)	0.975 652(1)	
			$N=599$	$S_{\text{scan}}=0.973\ 95(5)$				
500	0.9904(2)	0.0087(2)	0.957(5)	0.945(2)	0.974(3)	0.968(1)	0.969(3)	2540
5000	0.9760(2)	0.0030(2)	0.970(2)	0.971(1)	0.973(1)	0.9733(4)	0.973(1)	316
50 000	0.9743(1)	0.0010(2)	0.9738(5)	0.9737(3)	0.9741(3)	0.9740(2)	0.9741(3)	60
S_{TI}	0.974 04(1)		0.974 04(1)	0.974 04(1)	0.974 04(1)	0.974 04(1)	0.974 04(1)	
S_{series}	0.974 025(1)		0.974 025(1)	0.974 025(1)	0.974 025(1)	0.974 025(1)	0.974 025(1)	

sults are slightly inferior (i.e., lower) than those of S^B. However, for $n_{\text{future}}=5000$ and 50 000, S_G^B and S^B are equal within error bars that are, however, two to three times smaller for S_G^B than for S^B; therefore, the corresponding results for S_G^M are equal to those of S^M but with slightly smaller errors.

In all cases S^M and S_G^M are equal (within the error bars) to S^D, to the TI and series results, and for $N<599$ also to the scanning results. However, the error bars of TI are the smallest. The fact that for each N the S^M (and S_G^M) results for $n_{\text{future}}=5000$ and 50 000 (and in some cases for $n_{\text{future}}=500$) are equal (and they are also equal to the TI values) demonstrates that the absolute values of S^A and $S^B(S_G^B)$ deviate equally from the correct results. Overall the HSMC statistical errors are small (0.002%–0.005%); however, much more computer time has been invested in the simulations of the longer chains.

We also obtained results with HS where the entropy was calculated from a generated sample of chains (see next section) with a limited but systematic scanning of $f=8$. Our main interest has been only to check how much are these results larger than those obtained by HSMC (based on the stochastic MC scanning of the entire future); therefore, the HS results were calculated only for several chain lengths of $N=49$, 99, 249, and 399. Indeed, the S^A(HS) values [Eq. (8)] are always larger than the exact ones, where the deviation increases with N; thus, for $N=49$ the HS value is relatively accurate, comparable to that of HSMC($n_{\text{future}}=5000$), while for $N=399$ S^A(HS) worsens becoming close to S^A(HSMC) for $n_{\text{future}}=500$. Correspondingly, σ_A(HS) is always larger than σ_A(HSMC) obtained for $n_{\text{future}}=5000$ (except for $N=49$). A similar trend is observed for S^B(HS) which is always a lower bound but smaller than S^B(HSMC) obtained for $n_{\text{future}}=5000$.

The results for S^M(HS) are very close to the correct ones for $N=49$ and 99, but overestimate the correct values as chain length increases, where for $N=399$ the error is of $\sim 0.2\%$. As discussed earlier, S_G^B(HS) is not well defined and indeed it constitutes a lower bound only for $N=49$ and 99 [where its values are larger than the corresponding S^B(HSMC) values], becoming larger than the exact value for larger N. The related average S_G^M is always larger than the exact value with the largest deviation of 0.36% occurring for $N=399$. As expected (see the last paragraph of Sec. II E) S^D(HS) is always an upper bound, which is slightly smaller than the corresponding S_G^M. These results demonstrate that the performance of HS is inferior to that of HSMC.

D. Entropy by reconstructing a sample of chains

In practice, however, one would apply HSMC to samples of chains of different conformations, therefore a second set of results has been obtained from thermodynamic samples of SAWs. To generate such samples we have carried out long MC runs (based on the pivot and corner moves described previously) starting from a straight chain, equilibrating for 300 MC steps per bond, where every 2300 MC steps per bond the current conformation (i) was selected for reconstruction as described earlier. (This same prescription was also used to generate samples for the HS method.)

Calculation of the entropy from a sample of chains is more difficult than for a straight chain. Because the frozen past of the chain is not straight, part of the conformational space of the future SAWs might become unreachable with our dynamic pivot/corner MC procedure; this might affect, in particular, the movement of the treated bond k hence the corresponding transition probability. Because the reconstruction starts from configuration i, in an extreme case bond $k(i)$ will be unable to move to another direction leading to $p^{\text{MC}}=1$; in another case it will change direction but may never return to its original direction in chain i leading to $p^{\text{MC}}=0$. Such TPs will affect significantly the probability P_i^{MC} [Eqs. (13) and (14)] of the chain. Notice, however, that these cases do not demonstrate a drawback of the HSMC method but reflect the strong excluded volume interaction of SAWs on a square lattice and the inability inherent in our MC procedure to search the entire conformational space (i.e., the procedure is nonergodic). As discussed below, these problems can be alleviated by generating the future SAWs with more efficient MC techniques. Such problems have not been encountered in application of HSMC to fluid systems (argon and water) and peptides. Obviously, this problem will be weakened significantly for SAWs on a simple cubic lattice, for example, where the excluded volume interactions are less severe than on a square lattice.

To alleviate these problems we have taken several measures. First, before carrying out the future sampling at step k the program checks the nearest neighbor sites of monomer k (located at the end of bond $k-1$); if all four of them are already occupied by chain monomers (i.e., step k has only one choice) the future sampling is avoided, the TP(k) is defined as 1, and the next step $(k+1)$ is treated. When $p^{\text{MC}}=0$ or 1 occurs, p^{MC} is calculated by the (systematic) HS method, i.e., by an exact enumeration of the future SAWs of $f=8$ bonds and this value is considered in the calculation of P_i^{MC}. Still, the reconstruction probability of some chains might be affected significantly by similar problems (i.e., p^{MC} values that are incorrectly very small or close to 1). Because the Boltzmann probability of all chains is the same, one can ignore the contribution of such chains to the average entropy. In practice, a SAW i with $-k_B \ln P_i^{\text{MC}}$ beyond four standard deviations of the average is not considered in the averaging of the entropy.

Comparing the results in Tables I and II demonstrates the increase in sampling difficulty and decrease in accuracy involved in reconstructing a sample of chains. Thus, while S^A (sample) in Table II (as expected) is an upper bound that decreases as n_{future} increases, it is always larger (i.e., worse) than the corresponding S^A (straight) in Table I; for $n_{\text{future}}=50\,000$ the deviations are small for $N\leqslant 149$ but increase for larger N. A similar trend is observed for σ_A (sample) that always decreases (as expected) with increasing n_{future} but it is larger than the corresponding σ_A (straight). Notice that for $n_{\text{future}}=50\,000$ S^A (sample) and σ_A (sample) are always better (i.e., smaller) than S^A(HS) and σ_A(HS), which again reflects the superior accuracy of HSMC. S^B and S_G^B always increase with n_{future} and for $n_{\text{future}}=5000$ and 50 000 they are in most cases equal with slightly lower errors for S_G^B. Again, these values are always smaller (i.e., worse) than those in Table I, where the deviations are small for $N \leqslant 149$ and increase for larger N. For $N \leqslant 99$ the (six) results for S^M, S_G^M, and S^D for $n_{\text{future}}=5000$ and 50 000 are all equal (within the error bars), whereas for larger N these functionals have slightly better values at 50 000 than at 5000. For $N \leqslant 249$ the best results for S^M, S_G^M, and S^D (i.e., for $n_{\text{future}}=50\,000$) are equal to those of Table I, while for $N=399$ and 599 the results for the straight chains are more accurate with errors that are ~six times smaller.

The conclusion from the above comparison is that for any model studied it is more efficient to carry out a relatively large number of reconstructions (replicates) of a small number of "good" chain configurations than to reconstruct a thermodynamic sample of chains.

E. Discussion

The results of the two tables show that for a given amount of computer time it is preferable to increase n_{future} using relatively small values of sample size n; this leads to improved (smaller) S^A, larger S^B, hence better estimates S^M and S^D (in particular, for large N). This effect is significant, in particular, for a sample of chains, where, for $N=49$, for example, $S^A(n_{\text{future}}=50\,000)$ is equal in both tables, while for $n_{\text{future}}=5000$ and 500 the results in Table II are always worse (larger) than the corresponding results of Table I. Thus, the best (lowest) S^A will be obtained in the extreme case, where only a single (good) chain is reconstructed with the maximal n_{future} for a given amount of computer time. However, this would come with a price that the information provided by the other functionals would be lost because $S^A=S^B=S^M=S^D$.

An inherent inefficiency of HSMC lies in the need to carry out $N-1$ simulations for an N-bond SAW. Still, the performance of HSMC for a sample of SAWs can be improved by changing the scaling function discussed in Sec. II C, which controls the extent of simulation applied to each bond in the reconstruction process. However, the most significant factor affecting efficiency is the simulation method used for the chain reconstruction. Thus, our preliminary simulations based on corner moves alone have converged extremely slowly, and adding the pivot moves improved performance dramatically. In three dimensions, where the excluded volume effect is weaker, one can add crankshaft moves (and other moves, see Ref. 8) that are expected to increase efficiency further. To improve accuracy one can increase the scanning parameter used in the HS parts of the processes from $f=8$ to12 (and even to 14).

The pivot moves are very important for an open chain, but they become unsuitable for a SAW enclosed in a small volume or for a highly compact SAW with attractive interactions at low temperature, where only local MC moves are applicable. Notice, however, that simulating these restricted models (on a square or a simple cubic lattice) with dynamic MC procedures based on local moves is generally nonergodic and extremely inefficient, meaning that the corresponding HSMC reconstructions will be inefficient as well. On the other hand, restricted SAW models are better handled by step-by-step construction procedures.[14–17] The scanning method, for example, is ergodic and due to its "feelers" one can generate chains in restricted environments quite efficiently. Thus, the idea would be to implement within the framework of HSMC a suitable growth procedure, which will lead to exact results, unlike HS. Notice that growth procedures provide the entropy by themselves from their generated sample of chains; however, a suitable HSMC/growth procedure would enable estimating the entropy from a given trajectory.

An interesting test case is a model of multiple SAWs enclosed in a "box," studied previously by the scanning and HS methods,[24] where chains are added successively to an initially empty box. However, with HS only the *partial* future of a reconstructed chain is considered, whereas HSMC can take into account the entire future, including that of the reconstructed chain and the positions and conformations of the as yet unreconstructed chains. If the system is not extremely dense local dynamic MC moves would suffice. In the extreme case where all sites are populated (density=1) one can apply simulation methods as those implemented by Pakula and Reiter.[27] It should be emphasized that HSMC can handle volumes with any shape and boundary conditions, where defining a suitable reference state for TI is not trivial.

Chain models with finite interactions have been defined on enriched lattices (i.e., with a large coordination number, such as the bond fluctuating model) and have been simulated by dynamic MC procedures. All of these models can be treated by HSMC. Such models have been used to study protein folding trajectories, for example, where transitions between different conformational regions (microstates) occur, but their relative populations can be obtained only crudely from the trajectory. However, these populations can be calculated with high accuracy by applying HSMC locally to these microstates, in the same way it has been applied to the helical, extended, and hairpin microstates of polyglycine molecules.[5] It should be noticed that the entropy of microstates (i.e., local fluctuations) can also be obtained *approximately* by the harmonic and quasiharmonic techniques[12] or the local states method,[13] while a similar calculation by TI is a standing problem. Returning to the present model of SAWs, it appears that the most efficient is the scanning method (where a run for generating SAWs of $N=599$ provides results for all intermediate N), followed by TI, where HSMC is the least efficient. For example, the tabulated TI value for a 399-bond SAW required ~100 h CPU, while for HSMC ($n_{\text{future}}=50\,000$), generating the 225 chains in Table I took 945 h CPU. It is stressed, however, that these levels of precision will often not be necessary in novel investigations on related polymer systems. A single reconstruction of a 399-bond SAW for $n_{\text{future}}=50\,000$ requires far less computational investment (~4.2 h CPU), and already gives a result of $S=0.9757(6)$.

In summary, calculation of S is a central problem in computer simulation, and HSMC with its unique features constitutes a new tool for obtaining S independent of other methods. With HSMC all interactions are considered, and its accuracy depends only on the amount of MC sampling. Furthermore, a "self-checking" accuracy analysis is inherent in the method, based on verifying the increase and decrease of the rigorous upper and lower bounds, S^B, S_G^B, and S^A, and the decrease of σ_A, as the approximation improves. Finally, unlike other methods, HSMC is of general applicability, covering liquids (argon and water), microstates of polypeptide molecules, and in this work also random coil polymers. HSMC can be applied to any type boundary conditions, which is very difficult to handle by TI, and unlike most methods, enables one to extract the absolute entropy from a given sample, where only a small number of SAWs (and even a single chain) need to be reconstructed; this is important for studying relaxation processes, such as protein folding.

ACKNOWLEDGMENT

This work was supported by NIH Grant Nos. R01 GM61916 and 1R01 GM66090.

[1] D. L Beveridge and F. M. DiCapua, Annu. Rev. Biophys. Biophys. Chem. **18**, 431 (1989); P. A. Kollman, Chem. Rev. (Washington, D.C.) **93**, 2395 (1993); W. L. Jorgensen, Acc. Chem. Res. **22**, 184 (1989).

[2] H. Meirovitch, in *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd (Wiley, New York, 1998), Vol.12, p. 1.

[3] A. Szarecka, R. P. White, and H. Meirovitch, J. Chem. Phys. **119**, 12084 (2003); R. P. White and H. Meirovitch, J. Chem. Phys. **119**, 12096 (2003); R. P. White and H. Meirovitch, Proc. Natl. Acad. Sci. U.S.A. **101**, 9235 (2004).

[4] R. P. White and H. Meirovitch, J. Chem. Phys. **121**, 10889 (2004).

[5] S. Cheluvaraja and H. Meirovitch, Proc. Natl. Acad. Sci. U.S.A. **101**, 9241 (2004); S. Cheluvaraja and H. Meirovitch, J. Chem. Phys. **122**, 054903-1 (2005).

[6] R. P. White, J. Funt, and H. Meirovitch, Chem. Phys. Lett. **410**, 430 (2005).

[7] I. Carmesin and K. Kremer, Macromolecules **21**, 2189 (1988); H.-P. Deutsch and K. Binder, J. Chem. Phys. **94**, 2294 (1991); T. Geisinger, M. Müller, and K. Binder, J. Chem. Phys. **111**, 5251 (1999); M. Müller, K. Binder, and L. Schäfer, Macromolecules **33**, 4568 (2000); G. Xu and W. L. Mattice, J. Chem. Phys. **117**, 3440 (2002); D. Chen and W. L. Mattice, Polymer **45**, 3877 (2004); Y. Termonia, Biomacromolecules **5**, 2404 (2004); Y. Termonia, Text. Res. J. **73**, 74 (2003).

[8] A. Sokal, in *Monte Carlo and Molecular Dynamics Simulations in Polymer Science*, edited by K. Binder (Oxford University Press, New York, 1955), pp. 47–124.

[9] H. Taketomi, Y. Ueda, and N. Gō, Int. J. Pept. Protein Res. **7**, 449 (1975); K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989); D. J. Covell and R. L. Jernigan, Biochemistry **29**, 3287 (1990); D. A. Hinds and M. Levitt, Proc. Natl. Acad. Sci. U.S.A. **89**, 2536 (2004); G. F. Berriz and E. I. Shakhnovich, Curr. Opin. Colloid Interface Sci. **4**, 72 (1999); A. Kolinski, M. Milik, J. Rycombel, and J. Skolnick, J. Chem. Phys. **103**, 4312 (1995); Y. Zhang and J. Skolnick, Biophys. J. **87**, 2647 (2004); D. M. Zuckerman, J. Phys. Chem. B **108**, 5127 (2004).

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

[11] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).

[12] N. Gō and H. A. Scheraga, J. Chem. Phys. **51**, 4751 (1969); A. T. Hagler, P. S. Stern, R. Sharon, J. M. Becker, and F. Naider, J. Am. Chem. Soc. **101**, 6842 (1979); M. Karplus and J. N. Kushick, Macromolecules **14**, 325 (1981).

[13] H. Meirovitch, Chem. Phys. Lett. **45**, 389 (1977); H. Meirovitch, S. C. Koerber, J. Rivier, and A. T. Hagler, Biopolymers **34**, 815 (1994).

[14] F. T. Wall, L. A. Hiller, and D. J. Wheeler, J. Chem. Phys. **22**, 1036 (1954).

[15] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955); F. T. Wall and J. J. Erpenbeck, J. Chem. Phys. **30**, 634 (1959); Z. Alexandrowicz, J. Chem. Phys. **51**, 561 (1969).

[16] H. Meirovitch, J. Phys. A **15**, L735 (1982); H. Meirovitch, J. Chem. Phys. **89**, 2514 (1988).

[17] J. Bascle, T. Garel, H. Orland, and B. Velikson, Biopolymers **33**, 1843 (1993); P. Grassberger and R. Hegger, J. Phys. A **27**, 4069 (1994); P. Grassberger, Phys. Rev. E **56**, 3682 (1997); S. K. Kumar, I. Szleifer, and A. Z. Panagiotopoulos, Phys. Rev. Lett. **66**, 2935 (1991).

[18] H. Meirovitch, J. Phys. A **16**, 839 (1983).

[19] H. Meirovitch, Phys. Rev. A **32**, 3709 (1985).

[20] H. Meirovitch, Macromolecules **18**, 563 (1985).

[21] A. J. Guttmann and I. G. Enting, J. Phys. A **21**, L165 (1988); A. R. Conway, I. G. Enting, and A. J. Guttmann, J. Phys. A **26**, L1519 (1993).

[22] H. Meirovitch and Z. Alexandrowicz, J. Stat. Phys. **15**, 123 (1976); H. Meirovitch, Chem. Phys. **111**, 7215 (1999).

[23] T. L. Hill, *Statistical Mechanics Principles and Selected Applications* (Dover, New York, 1956).

[24] H. Meirovitch, J. Chem. Phys. **97**, 5803 (1992).

[25] M. Muller and W. Paul, J. Chem. Phys. **100**, 719 (1994).

[26] N. Madras and A. D. Sokal, J. Stat. Phys. **47**, 573 (1987).

[27] T. Pakula, Macromolecules **20**, 679 (1987); J. Reiter, Macromolecules **23**, 3811 (1990).