


Thermodynamics of hydrophobic-polar model proteins on the face-centered cubic latticeMatthew S. Wilson¹* and David P. Landau¹*Center for Simulational Physics, Department of Physics and Astronomy, The University of Georgia, Athens, Georgia 30602, USA* (Received 30 April 2021; accepted 7 July 2021; published 12 August 2021)

The HP model, a coarse-grained protein representation with only hydrophobic (H) and polar (P) amino acids, has already been extensively studied on the simple cubic (SC) lattice. However, this geometry severely restricts possible bond angles, and a simple improvement is to instead use the face-centered cubic (fcc) lattice. In this paper, the density of states and ground state energies are calculated for several benchmark HP sequences on the fcc lattice using the replica-exchange Wang-Landau algorithm and a powerful set of Monte Carlo trial moves. Results from the fcc lattice proteins are directly compared with those obtained from a previous lattice protein folding study with a similar methodology on the SC lattice. A thermodynamic analysis shows comparable folding behavior between the two lattice geometries, but with a greater rate of hydrophobic-core formation persisting into lower temperatures on the fcc lattice.

DOI: [10.1103/PhysRevE.104.025303](https://doi.org/10.1103/PhysRevE.104.025303)**I. INTRODUCTION**

Protein folding is a complex, macromolecular process which is difficult to study computationally due to the enormous number of available physical states and the rough free-energy landscapes that arise in these systems. Such challenges, along with large timescales and length scales, often motivate the use of simplified models with Monte Carlo (MC) simulations in place of all-atom simulations.

The hydrophobic-polar (HP) lattice model [1,2] has been used to study various aspects of protein folding [3–8], where lattice geometry, amino acid representation, and energy function are chosen to simplify the problem. While simple, the model has a huge number of possible configurations [9,10], and finding the ground state is *NP*-complete [11]. For this reason, the HP model is often used to test optimization algorithms and sampling methods [12–20].

Square and simple cubic (SC) lattices are commonly used to study model proteins for a complete (but only qualitatively representative) analysis of the protein folding process, although higher-coordination lattices can be used. Statistical analyses [21–23] of experimentally measured protein structures show that the face-centered cubic (fcc) lattice mimics the backbone geometry quite well compared to other simple lattices. In this paper, the replica-exchange Wang-Landau (REWL) algorithm [24] is used to examine thermodynamics of biologically motivated HP sequences on the fcc lattice, and a direct comparison is made with results from previous folding simulations on the SC lattice [25,26].

The remainder of the paper is organized as follows: Section II describes the models used; Sec. III describes the Wang-Landau and REWL algorithms and details bond-bridging moves for the fcc lattice; Sec. IV presents the

simulation results and compares with their SC counterparts; finally, Sec. V summarizes the results.

II. SIMULATION MODEL**A. HP lattice protein**

First proposed in the 1980's, the HP model [1,2] greatly reduces the degrees of freedom of the represented protein while preserving the essential physics of folding. The 20 possible amino acid types are classified as either hydrophobic (H) or polar (P) residues that are restricted to lie on sites of a rigid lattice and are connected by unbreakable, nearest-neighbor peptide bonds, as shown in Fig. 1. There is no explicitly modeled solvent surrounding the protein; rather, the hydrophobic effects between an aqueous solution and amino acids are implicitly considered as the “driving force” of the folding process that results in a hydrophobic core. Shown in Eq. (1), the total energy of an HP lattice protein

$$\mathcal{H} = -n_{HH}\varepsilon_{HH} \quad (1)$$

is solely determined by the number of neighboring, non-bonded H residues in a conformation (H-H contacts) n_{HH} , and a chosen coupling strength ε_{HH} .

B. HP model on the fcc lattice

While widely used on the SC lattice, the same HP model constraints and energy function can also be directly mapped onto other rigid lattices. Having 12 possible nearest neighbors per site, the fcc lattice is a high-coordination lattice, with 60° and 120° angles allowed between adjacent edges (bonds), as well as the 90° and 180° angles found on the SC lattice. Figure 2 shows the HP model on the fcc lattice.

Unlike the SC lattice, there is no parity problem where residues must be separated by an odd number in the polymer chain in order to be nonbonded nearest neighbors of one another. The HP model has previously been studied

*msw.wilson@uga.edu

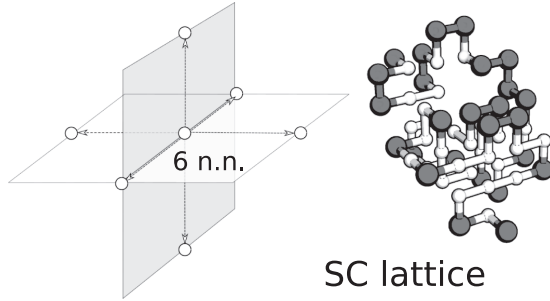


FIG. 1. HP model on the simple cubic lattice: The six nearest neighbors are shown on the left; a typical lattice configuration of the BPTI protein is shown on the right. Black residues are polar (P) and white are hydrophobic (H).

on the fcc lattice to test optimization algorithms with low-energy structure prediction [17,19,20,27], and others have proposed various augmented protein models on the fcc lattice [23,28–31].

Here, we calculate the ground states and full density (multiplicity) of states for the HP model on the fcc lattice to make a direct comparison of the average thermodynamics with previous results on the SC lattice.

C. Benchmark HP sequences

The HP sequences given in Table I are benchmarks from various SC lattice studies (references in the caption of Table I) and have average thermodynamics reported by Wüst and Landau [26] with which to compare. Most of the sequences are mapped from real protein fragments [3,32], with the exception of seq_67 and seq_88 that were specifically designed for the SC lattice, and are known to have threefold degenerate and nondegenerate ground state energies, respectively [13,33].

III. METHODOLOGY

A. Wang-Landau algorithm

The Wang-Landau (WL) algorithm [34,35] estimates the density of states $\hat{g}(E)$ using a modification factor f and flatness criterion p as control parameters, and a histogram $H(E)$ to keep track of visited energies. The algorithm is as follows:

- (1) Start with $f = e$ and $\hat{g}(E) = 1 \forall E$.

- (2) Using MC trial moves, randomly transition between states A and B with probability

$$P(A \rightarrow B) = \min \left(1, \frac{\hat{g}(E_A)}{\hat{g}(E_B)} \right). \quad (2)$$

- (3) After each trial move, update $\hat{g}(E) \leftarrow f \cdot \hat{g}(E)$ and $H(E) \leftarrow H(E) + 1$.

- (4) If $H(E) \geq p \cdot \overline{H(E)} \forall E$ is satisfied, set $H(E) = 0 \forall E$ and let $f \leftarrow \sqrt{f}$.

- (5) Continue to iterate steps (2)–(4) until $f \approx 1$ within some user-defined threshold (e.g., $\ln f_{\text{final}} = 10^{-6}$).

B. Replica-exchange Wang-Landau algorithm

The REWL algorithm [24,36], a parallelized version of the WL algorithm, employs the WL scheme as “replicas” concurrently sampling overlapping energy subspaces called “windows.” Each replica is assigned to a window, and samples within the bounds of this window with its own $\hat{g}(E)$. At a chosen frequency, two replicas, i and j , in the overlapping regions of neighboring windows attempt to exchange states, denoted A and B , according to the exchange probability

$$P_{i,j}(A \rightarrow B) = \min \left(1, \frac{\hat{g}_i(E_A)}{\hat{g}_i(E_B)} \cdot \frac{\hat{g}_j(E_B)}{\hat{g}_j(E_A)} \right). \quad (3)$$

The benefit of not having to converge the whole energy space within a single $\hat{g}(E)$ is a remarkable speed-up and access to larger system sizes [37]. The algorithm can be scaled up in the number of windows, number of walkers per windows, and even the dimensionality of the density of states. After all windows converge, the simulation ends, and the individual $\hat{g}_i(E)$ are shifted together at the point where $\frac{d}{dE} |\ln[\hat{g}_i(E)] - \ln[\hat{g}_{i\pm 1}(E)]|$ is minimal.

After running REWL, the normalized $\hat{g}(E)$ can be used as fixed weights in a multicanonical [38] production run that samples $\propto 1/g(E)$. From this run, the $\hat{g}(E)$ can be improved through reweighting, or additional averages can be calculated for the system, as described in Eq. (5) at the end of the section. We used a large production run to improve estimates of $\ln[g(E)]$ from REWL for seq_124, which was the most challenging sequence to adequately sample.

C. Monte Carlo trial move set

The following MC trial moves are reversible and ergodic, and are implemented in an unbiased fashion that does not need to count forward (backward) transitions between states:

Single-site pull move [39,40]: Displace one residue and “pull” the rest of the chain along occupied positions until reconnected.

Bond-rebridging moves [41], Hamiltonian path [42]: Cut and form new pairs of backbone bonds without changing residue locations.

Pivot move [17,43]: Rotate a random portion of the chain with random axis and angle.

Diagonal (kink flip) move: Displace one residue if it remains connected to its bonded neighbors.

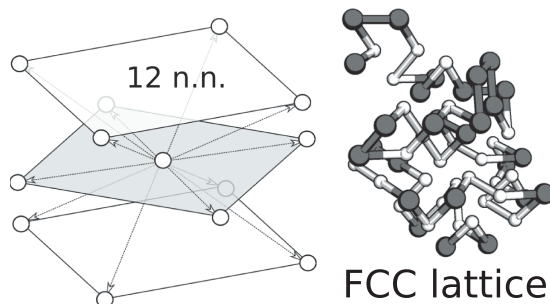


FIG. 2. HP model on the face-centered cubic lattice: The 12 nearest neighbors are shown on the left; a typical lattice configuration of the BPTI protein is shown on the right. Black residues are polar (P) and white are hydrophobic (H).

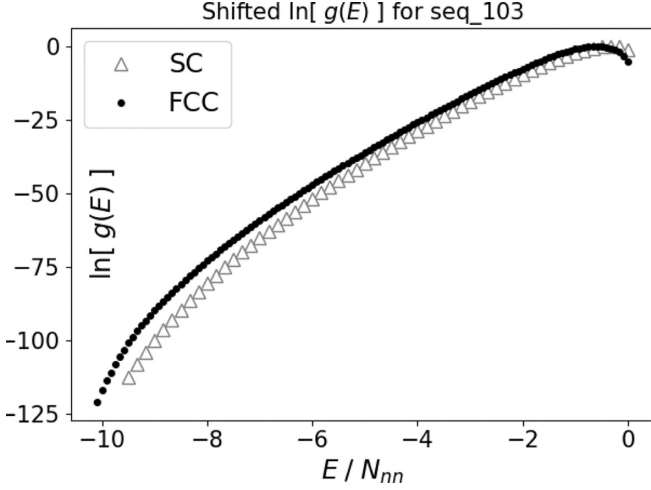


FIG. 4. Natural log of $g(E)$ [shifted so that $\max\{g(E)\} = 1$] for seq_103 on the SC (gray triangles) and fcc (black circles) lattices. The x axis is normalized by N_{nn} , and results for fcc have approximately twice as many points as SC. Error bars are smaller than the symbol sizes.

Note that $\beta \equiv (k_B T)^{-1}$ and \mathcal{Z} can be calculated for all temperatures from a single WL or REWL simulation. An example of a shifted $\ln[g(E)]$ spanning over 50 orders of magnitude is plotted in Fig. 4 for seq_103 on both the SC (triangles) and fcc (circles) lattices.

Ensemble averages are denoted by angled brackets $\langle \rangle$, where Eq. (5) shows the calculation for some arbitrary observable Q using the density of states from REWL and histograms $H(E)$, $H(E, Q)$ accumulated during a multicanonical production run,

$$\langle Q \rangle = \frac{1}{\mathcal{Z}} \sum_E \bar{Q}(E) g(E) e^{-\beta E}, \quad (5)$$

where

$$\bar{Q}(E) = \frac{1}{H(E)} \sum_Q Q H(E, Q).$$

The heat capacity [Eq. (6)] is equivalent to the thermal rate of hydrophobic-core formation [Eq. (7)], and is calculated at a wide range of temperatures.

$$C_V = k_B \beta^2 [\langle E^2 \rangle - \langle E \rangle^2] \quad (6)$$

$$= -\varepsilon_{HH} \frac{d\langle n_{HH} \rangle}{dT}. \quad (7)$$

The peaks and shoulders present in $C_V(T)$ indicate transitions of the sequence's structural properties, and can be analyzed along with additional structural quantities to elucidate the precise behavior, if necessary. Structural observables can be calculated using Eq. (5) in a multicanonical [38] production run with the $g(E)$ from REWL as sampling weights.

When comparing properties between the SC and fcc lattices, the coordination number, or number of nearest neighbors N_{nn} , is used to normalize temperature scales. Shown in

Figs. 1 and 2, this quantity is defined as

$$N_{nn} = \begin{cases} 6, & \text{SC lattice,} \\ 12, & \text{fcc lattice.} \end{cases} \quad (8)$$

IV. RESULTS FOR BENCHMARK HP SEQUENCES

Thermodynamic results from REWL for the six benchmark sequences on the fcc lattice are compared to the results from Wüst and Landau [26], which used serial WL to simulate the same sequences on the SC lattice. Figure 5 shows the comparison of the specific heat C_V/N for the two lattice types, with the reduced temperatures normalized by N_{nn} . Ground state energies and configurations are reported for the sequences at the top left of each plot.

The specific heat curves (C_V/N) in Fig. 5 each show distinct maxima in the reduced temperature region between 0.08 and 0.125, indicating the coil-globule transition, where the model proteins change from an extended conformation to a collapsed, disordered state. This first structural transition in the folding process involves an increase in the number of HH contacts as a disordered hydrophobic core forms. Both lattice geometries show similar coil-globule transition temperatures (when shifted by N_{nn}), but with the fcc lattice having a slightly greater transition temperature and magnitude of C_V/N .

At temperatures below the coil-globule collapse, the formation of an ordered, low-energy H core happens in one or two more structural transitions that are signified by “shoulders” in C_V/N . For the three shortest sequences, the optimal folded state occurs after a signal at $k_B T / (N_{nn} \varepsilon_{HH}) \leq 0.075$. Seq_67 and seq_88 are known to have threefold and unique ground state degeneracies on the SC lattice, respectively, and both have a sharp peak in C_V/N below which the model protein is optimally folded.

The three sequences with length > 100 residues show two additional structural transitions on the fcc lattice below $k_B T / (N_{nn} \varepsilon_{HH}) \leq 0.075$, where a dense H core is first formed but then refolded/rearranged as the energy is minimized. For seq_124 and seq_136 on the SC lattice, the peaks in C_V/N below $k_B T / (N_{nn} \varepsilon_{HH}) \leq 0.025$ are a result of undersampled $g(E)$ for the minimal energy, and should be regarded as spurious. The lowest-temperature signals in C_V/N for the fcc lattice results consistently extend down to $k_B T / (N_{nn} \varepsilon_{HH}) = 0.025$, whereas the SC results show essentially no thermal response at these temperatures.

Minimal energy (E_{\min}) states (and all other energy states) are not known *a priori*, but are identified during the REWL simulation. Minimal energies and representative structures are shown in the top left corners of the plots in Fig. 5. Our REWL method shows superior performance over the modified pruned-enriched-Rosenbluth method (nPERM) chain-growth algorithm that was used in previous studies [45] for the fcc HP model, with significantly lower ground state energies reported here (-121 vs -116 for seq_103, -164 vs -154 for seq_124, and -174 vs -168 for seq_136). Ground state energies are identified as optimal using the constraint programming HPstruct tool from the CPSP-tools server [46–49], with the exception of seq_88, that is assured by HPstruct to have an optimal value greater than -143 (for which we found a value of $E_{\min} = -141$).

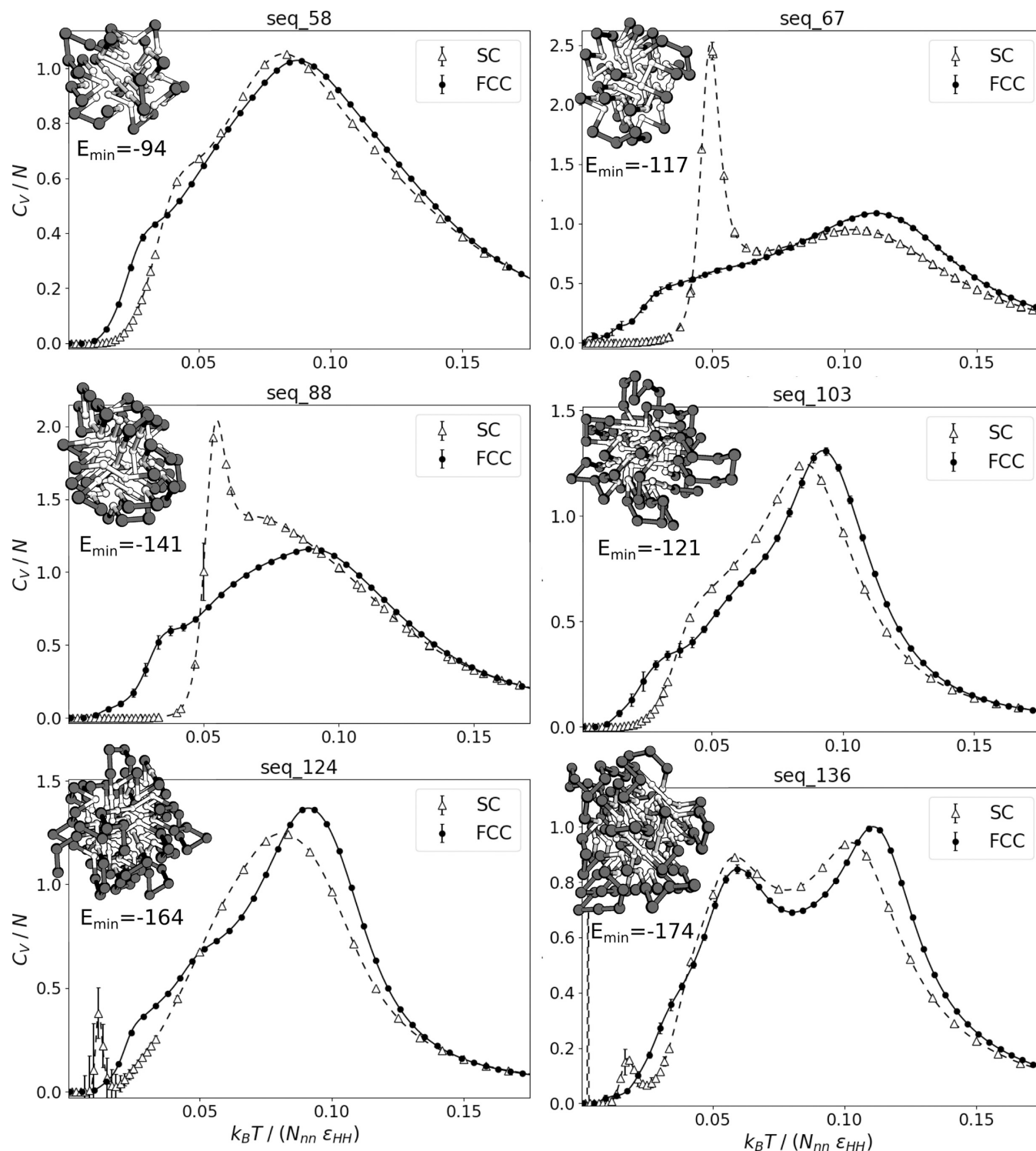


FIG. 5. Thermodynamic comparison of benchmark sequences folding on SC (dashed with triangles) and fcc (solid with black circles) lattices. Reduced temperature is given on the x axis, and normalized by N_{nn} . Ground state structures and energies are shown in the top left of each plot. The black residues are polar (P), and the white are hydrophobic (H). Error bars for the fcc results are calculated from 20 independent trials. Where not shown, the error bars are smaller than the symbol sizes.

V. CONCLUSIONS

This work details the simulation of HP model proteins on the fcc lattice, where the thermodynamics of folding are calculated and directly compared with the SC lattice. Not only

are the ground state energies found for the chosen sequences, but the full density of states is determined using the REWL algorithm with a set of unbiased MC moves chosen from the literature. The implementation of bond-rebridging moves where newly formed bonds need not be parallel is detailed for

the fcc lattice. We study the simple HP model to explicitly observe similar folding transitions between the SC and fcc lattice geometries.

For all but the shortest HP sequence, the temperature at which the coil-globule transition signal occurs is slightly increased, presumably due to the availability of many more compact, globular states during the collapse. The rate of hydrophobic-core formation is also significantly larger at low temperatures for all tested sequences on the fcc lattice. Most evident in the results for the longest three sequences, the shoulder in C_V/N below $k_B T / (N_{\text{nn}} \varepsilon_{HH}) \approx 0.05$ suggests that the structural transition associated with an ordered H-core formation may be split into an additional step on the fcc lattice, or is at least more reliable statistically than the results with the SC lattice. This effect is a result of the additional geometric freedom of the fcc lattice, which enables a larger number of accessible low-energy states in the folding process. Results for seq_67 and seq_88 are unsurprisingly different between the two lattice types, as the low degeneracy is not preserved on the fcc lattice. Minimal energy structures identified during each simulation are found to have the optimal H cores, as verified by the HPstruct tool available online.

The presented methodology is effective for identifying and sampling configurations on the fcc lattice and has general utility for simulating polymer and protein models with the fcc lattice geometry. Such models have relevance in current

research topics including *ab initio* protein structure prediction [50,51] and the study of helical and fractal polymers [52–54]. Furthermore, the modification of Eq. (1) to incorporate more realistic secondary structural motifs in protein folding studies is a known challenge [30,31,55,56] and a possible direction for future research that could employ this methodology. Here, we provide thermodynamic data that show the similarity of HP model protein folding simulations on the fcc lattice with the commonly used SC lattice, with the fcc geometry having slightly higher and lower temperatures associated with the coil-globule and ground state transitions, respectively. These results will be a useful benchmark for future thermodynamic inquiries using similar fcc lattice models.

ACKNOWLEDGMENTS

We thank Thomas Wüst for an illuminating discussion, for sharing the SC lattice thermodynamic data [26], and for insights on the SC lattice codes [57]. We thank Alfred Farris for many helpful discussions and suggestions. This study was supported in part by resources and technical expertise from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

-
- [1] K. A. Dill, *Biochemistry* **24**, 1501 (1985).
 - [2] K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
 - [3] K. A. Dill, K. M. Fiebig, and H. S. Chan, *Proc. Natl. Acad. Sci. USA* **90**, 1942 (1993).
 - [4] G. Shi, T. Wüst, and D. P. Landau, *Phys. Rev. E* **94**, 050402(R) (2016).
 - [5] A. C. Farris, G. Shi, T. Wüst, and D. P. Landau, *J. Chem. Phys.* **149**, 125101 (2018).
 - [6] A. D. Swetnam and M. P. Allen, *Phys. Rev. E* **85**, 062901 (2012).
 - [7] Y. W. Li, T. Wüst, and D. P. Landau, *Phys. Rev. E* **87**, 012706 (2013).
 - [8] T. Wüst, D. Reith, and P. Virnau, *Phys. Rev. Lett.* **114**, 028102 (2015).
 - [9] R. Schram, G. Barkema, and R. Bisseling, *J. Stat. Mech.: Theory Exp.* (2011) P06019.
 - [10] R. D. Schram, G. T. Barkema, R. H. Bisseling, and N. Clisby, *J. Stat. Mech.: Theory Exp.* (2017) 083208.
 - [11] B. Berger and T. Leighton, *J. Comput. Biol.* **5**, 27 (1998).
 - [12] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, E. I. Shakhnovich, and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **92**, 325 (1995).
 - [13] T. C. Beutler and K. A. Dill, *Protein Sci.* **1996**, 2037 (1996).
 - [14] H.-P. Hsu, V. Mehra, W. Nadler, and P. Grassberger, *Phys. Rev. E* **68**, 021113 (2003).
 - [15] S. C. Kou, J. Oh, and W. H. Wong, *J. Chem. Phys.* **124**, 244903 (2006).
 - [16] M. Bachmann and W. Janke, *J. Chem. Phys.* **120**, 6779 (2004).
 - [17] J.-J. Tsay and S.-C. Su, *Proteome Sci.* **11**, S19 (2013).
 - [18] J. Liu, G. Li, and J. Yu, *Phys. Rev. E* **84**, 031934 (2011).
 - [19] M. Rashid, M. A. H. Newton, T. Hoque, S. Shatabda, D. N. Pham, and A. Sattar, *BMC Bioinf.* **14**, S16 (2013).
 - [20] I. Dotu, M. Cebrián, P. Van Hentenryck, and P. Clote, in *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence* (AAAI Press, Palo Alto, CA, 2008), Vol. 1, p. 241.
 - [21] G. Raghunathan and R. L. Jernigan, *Protein Sci.* **6**, 2072 (1997).
 - [22] Z. Bagci, R. Jernigan, and I. Bahar, *Polymer* **43**, 451 (2002).
 - [23] M. Mann, R. Saunders, C. Smith, R. Backofen, and C. M. Deane, *Adv. Bioinf.* **2012**, 148045 (2012).
 - [24] T. Vogel, Y. W. Li, T. Wüst, and D. P. Landau, *Phys. Rev. E* **90**, 023302 (2014).
 - [25] T. Wüst and D. P. Landau, *Phys. Rev. Lett.* **102**, 178101 (2009).
 - [26] T. Wüst and D. P. Landau, *J. Chem. Phys.* **137**, 064903 (2012).
 - [27] J. Liu, B. Song, Y. Yao, Y. Xue, W. Liu, and Z. Liu, *Phys. Rev. E* **90**, 042715 (2014).
 - [28] L. Toma and S. Toma, *Protein Sci.* **8**, 196 (1999).
 - [29] T. Hoque, M. Chetty, and A. Sattar, *J. Comput. Biol.* **16**, 85 (2009).
 - [30] P. Pokarowski, A. Kolinski, and J. Skolnick, *Biophys. J.* **84**, 1518 (2003).
 - [31] P. Pokarowski, K. Droste, and A. Kolinski, *J. Chem. Phys.* **122**, 214915 (2005).
 - [32] E. E. Lattman, K. M. Fiebig, and K. A. Dill, *Biochemistry* **33**, 6158 (1994).
 - [33] K. Yue and K. A. Dill, *Proc. Natl. Acad. Sci. USA* **92**, 146 (1995).
 - [34] F. Wang and D. P. Landau, *Phys. Rev. E* **64**, 056101 (2001).
 - [35] F. Wang and D. P. Landau, *Phys. Rev. Lett.* **86**, 2050 (2001).
 - [36] T. Vogel, Y. W. Li, and D. P. Landau, *J. Phys.: Conf. Ser.* **1012**, 012003 (2018).

- [37] Y. W. Li, T. Vogel, T. Wüst, and D. P. Landau, *J. Phys.: Conf. Ser.* **510**, 012012 (2014).
- [38] B. A. Berg and T. Neuhaus, *Phys. Lett. B* **267**, 249 (1991).
- [39] N. Lesh, M. Mitzenmacher, and S. Whitesides, in *Proceedings of the 7th Annual International Conference RECOMB* (ACM, New York, 2003), p. 188.
- [40] H. J. Böckenhauer, A. Z. M. D. Ullah, L. Kapsokalivas, and K. Steinhöfel, in *WABI, Lecture Notes in Computer Science Vol. 5251* (Springer, Berlin, 2008), p. 369.
- [41] J. M. Deutsch, *J. Chem. Phys.* **106**, 8849 (1997).
- [42] R. Oberdorf, A. Ferguson, J. L. Jacobsen, and J. Kondev, *Phys. Rev. E* **74**, 051801 (2006).
- [43] N. Madras and A. D. Sokal, *J. Stat. Phys.* **50**, 109 (1988).
- [44] D. Reith and P. Virnau, *Comput. Phys. Commun.* **181**, 800 (2010).
- [45] T. Vogel, Diploma thesis, Leipzig University, 2004.
- [46] M. Mann, S. Will, and R. Backofen, *BMC Bioinf.* **9**, 230 (2008).
- [47] M. Mann, C. Smith, M. Rabbath, M. Edwards, S. Will, and R. Backofen, *Bioinformatics* **25**, 676 (2009).
- [48] M. Mann, R. Backofen, and S. Will, in *Proceedings of the 5th WCB*, Vol. WCB09 (2009), [arXiv:0910.3848](https://arxiv.org/abs/0910.3848).
- [49] R. Backofen and S. Will, *Constraints* **11**, 5 (2006).
- [50] S. R. D. Torres, D. C. B. Romero, L. F. N. Vasquez, and Y. J. P. Ardila, in *GECCO '07: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation* (ACM, New York, 2007), p. 393.
- [51] M. A. Rashid, S. Iqbal, F. Khatib, M. T. Hoque, and A. Sattar, *Comput. Biol. Chem.* **61**, 162 (2016).
- [52] M. Baiesi, G. Barkema, E. Carlon, and D. Panja, *J. Chem. Phys.* **133**, 154907 (2010).
- [53] J.-C. Walter, M. Baiesi, G. T. Barkema, and E. Carlon, *Phys. Rev. Lett.* **110**, 068301 (2013).
- [54] R. Schram, G. Barkema, and H. Schiessel, *J. Chem. Phys.* **138**, 224901 (2013).
- [55] A. Dal Palù, A. Dovier, and E. Pontelli, in *Logic for Programming, Artificial Intelligence, and Reasoning: Proceedings of the 12th International Conference, LPAR 2005*, edited by G. Sutcliffe and A. Voronkov, *Lecture Notes in Computer Science Vol. 3835* (Springer, Berlin, 2005), p. 48.
- [56] J.-J. Tsay, S.-C. Su, and C.-S. Yu, *Int. J. Mol. Sci.* **16**, 15149 (2015).
- [57] A. C. Farris, T. Wüst, and D. P. Landau, *Am. J. Phys.* **87**, 310 (2019).