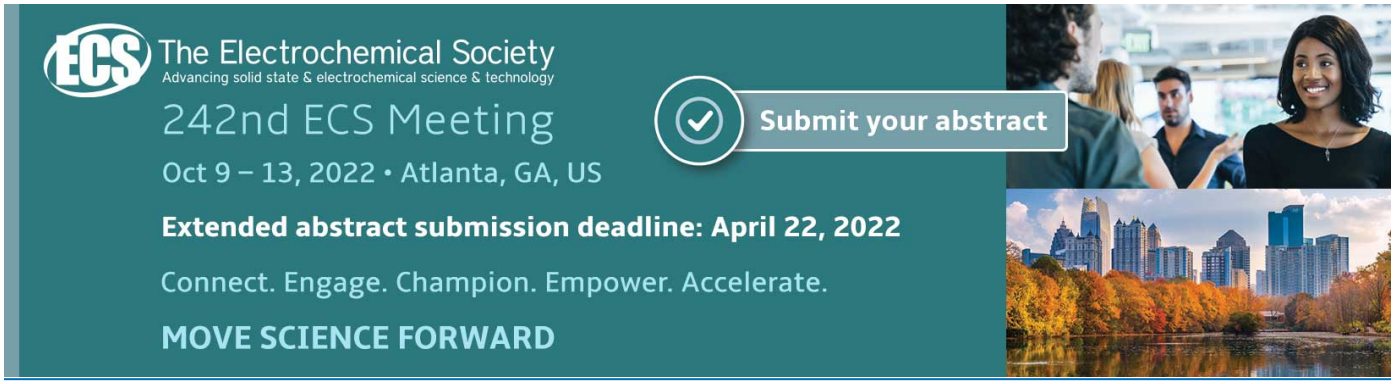
Simulating Met-Enkephalin With Population Annealing Molecular Dynamics



Simulating Met-Enkephalin With Population Annealing Molecular Dynamics

Henrik Christiansen[1], **Martin Weigel**[2], **Wolfhard Janke**[1]

[1]Institut für Theoretische Physik, Universität Leipzig, IPF 231101, 04081 Leipzig, Germany
[2]Centre for Fluid and Complex Systems, Coventry University, Coventry, CV1 5FB, England

E-mail:
henrik.christiansen@itp.uni-leipzig.de
martin.weigel@coventry.ac.uk
wolfhard.janke@itp.uni-leipzig.de

Abstract. Met-enkephalin, one of the smallest opiate peptides and an important neurotransmitter, is a widely used benchmarking problem in the field of molecular simulation. Through its range of possible low-temperature conformations separated by free-energy barriers it was previously found to be hard to thermalize using straight canonical molecular dynamics simulations. Here, we demonstrate how one can use the recently proposed population annealing molecular dynamics scheme to overcome these difficulties. We show how the use of multi-histogram reweighting allows one to accurately estimate the density of states of the system and hence derive estimates such as the potential energy as quasi continuous functions of temperature. We further investigate the free-energy surface as a function of end-to-end distance and radius-of-gyration and observe two distinct basins of attraction.

1. Introduction

The simulation of biomolecular systems with molecular dynamics and related techniques is now a field of vast importance, including applications of paramount fundamental relevance such as protein folding [1] as well as crucial practical challenges such as those encountered, for instance, in drug discovery [2]. The dominant difficulty in this context results from the slow relaxation of such systems in the standard (micro-)canonical dynamics, that typically is a consequence of the presence of several minima in the free-energy landscape [3]. In the past 35 years or so, an array of tools designed to take on this arduous task have been developed. In Monte Carlo, the concept of *generalized ensembles* has led to a number of successful approaches, including simulated tempering [4], parallel tempering [5, 6], and multicanonical simulations [7]. Parallel tempering or replica exchange uses several copies of the system run at different temperatures, and conformation exchanges of copies at nearby temperatures allow for the efficient exploration of the free-energy landscape through the escape to a regime of fast relaxation. While initially proposed for Monte Carlo, this successful meta-algorithm was subsequently also adapted to molecular dynamics [8, 9]. For the latter, a number of alternative strategies for accelerated simulations have been proposed, including so-called "accelerated" simulations [10] as well as metadynamics [11]. These, and related techniques, are based on the idea of adaptively biasing the weight function to gradually overcome energy barriers and enable the sampling of a wide

range of the relevant reaction coordinates. In this sense, these techniques are akin to the multicanonical [7] and Wang-Landau [12] methods in the world of Monte Carlo simulations.

While these methods can significantly extend the degree to which the free-energy landscape is being explored and accelerate the convergence to equilibrium, simulations of relevant biopolymers still frequently require very substantial computational resources. The significant growth in the number of cores in available (super-)computers is unfortunately only of limited utility to simulations using these techniques, as exploration of phase space and convergence to equilibrium are intimately linked to the number of time steps, and there is only limited scope for speeding up each step through (parallel) task splitting [13]. A natural way of using the available parallel resources consists of running many short simulations independently and combining the resulting statistics to improve the degree of sampling of the relevant states [14]. While this can work reasonably well, it does not remove the bottleneck of equilibration as each simulation conducted in parallel must at least run for the time required to thermalize. The popular use of Markov state models describing transitions between the valleys [15, 16] can only lead to trustworthy results if a reliable picture of the relevant states has already been deduced from direct simulations before.

In the present contribution, instead, we demonstrate how a recently introduced massively parallel approach to molecular dynamics (MD) simulations, population annealing molecular dynamics (PAMD), can be employed to utilize practically any amount of parallel resources available to improve the sampling in systems that are found hard to thermalize [17, 18]. Population annealing, that was first introduced in the context of Monte Carlo simulations [19, 20, 21, 22], uses a large population of identical system copies (replicas) which are successively cooled down starting from a thermalized population at high temperatures where equilibration is straightforward. At each cooling step, the population is resampled, favoring copies well adapted to the lower temperature. This process is reminiscent of genetic algorithms [23] where a population is evolved according to its fitness. For each new generation mutations occur at a prescribed rate, i.e., random changes in the microscopic variables. Subsequently, the new resulting fitness is calculated. According to some determined threshold, conformations are replicated or pruned. Genetic algorithms are hugely successful as optimization algorithms, i.e., in finding ground-state configurations [24, 25]. However, no thermodynamic information can be extracted. On the other hand, PAMD provides an equilibrium sample at each temperature considered, and it can hence be seen as an equilibrium variant of a combination of simulated annealing [26] and a genetic algorithm.

2. Model and Method

The system studied here is Met-enkephalin, a penta-peptide with amino-acid sequence Tyr-Gly-Gly-Phe-Met, that occurs in many organisms. It inhibits neurotransmitter release upon activation of the appropriate opioid receptor, and so plays a central role in pain regulation [27]. It is said to inhibit tumor growth and metastasis [28]. For the simulation, we cap the ends with a methyl and an acetyl group, respectively, resulting in a total size of 84 atoms. The interactions were modeled by the AMBER force-field ff94 [29].

While previous simulation studies of this system mostly used the replica exchange method for ensuring that equilibration is achieved, we here employ PAMD simulations to study the system [17]. In particular, this involves the following simulation steps:

(i) Set up an equilibrium ensemble of R independent copies of the peptide at some high temperature T_0.

(ii) Choose the next temperature $T_i < T_{i-1}$ from a pre-defined sequence.

(iii) Resample the population of systems to the new temperature T_i by replicating each copy a number of times proportional to the relative Boltzmann weight $\tau_j \propto e^{-(1/k_B T_i - 1/k_B T_{i-1})E_j}$,

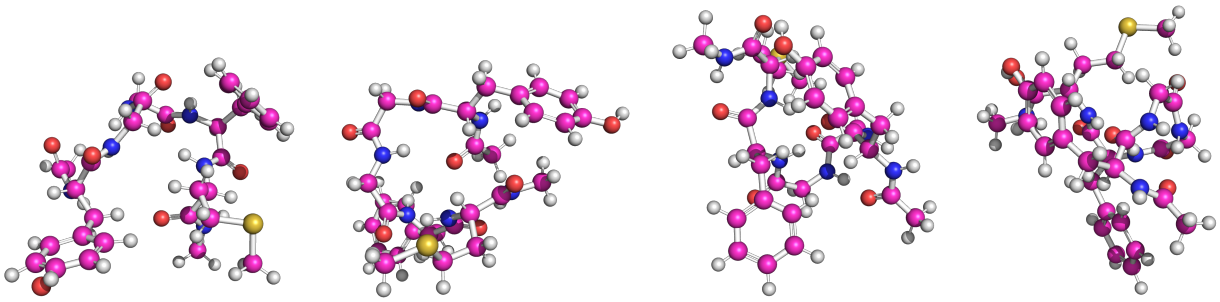


Figure 1. Snapshots of Met-enkephalin at different temperatures, decreasing from left to right. The following color-code applies: pink – carbon, white – hydrogen, red – oxygen, blue – nitrogen, yellow – sulfur.

where E_j is the potential energy of the jth replica. The momenta are simply adjusted to the new temperature by rescaling $p_k \to \sqrt{\frac{T_i}{T_{i-1}}} p_k$.

(iv) Update each replica with θ simulation steps of the underlying MD algorithm.

(v) Calculate observables $\mathcal{O}$ at temperature T_i as population averages.

(vi) Goto step (ii) until T_i reaches or falls below the target temperature T_N.

In parallel tempering, the choice of a suitable temperature protocol is often a challenging task [30, 31]. For PAMD, on the other hand, we have shown in Ref. [18] that an appropriate temperature set may be found "on the fly" by demanding a constant energy histogram overlap without the need of conducting preliminary simulations. For the present study we adopt the previously used temperature scheme consisting of simulations at $T_i = 700$, 585, 489, 409, 342, 286, 239, 200 K. This corresponds to a constant energy histogram overlap of about 30% [18]. Due to the resampling that involves making copies of some population members, it is crucial to employ a stochastic thermostat in PAMD simulations, as otherwise the copies would eternally follow the same trajectory, thus compromising the quality of the statistical sampling [17]. Here we use a Langevin thermostat with 0.5 fs integration step and friction coefficient $\gamma = 1/\text{ps}$. For our simulation we ran 10^4 replica with a total simulation time of 200 ns. MD steps amounting to 25 ns were used to initialize the population at the highest temperature, followed by 21.875 ps of evolution (or 4375 updates) performed at each temperature step. For the molecular dynamics part, which dominates the computational effort, we relied on the package OpenMM [32]. Overall, our simulations show an excellent parallel efficiency of over 85% using 500 cores, far superior to what normally can be achieved in parallel tempering.

3. Results

To convey an overall impression of the behavior of Met-enkephalin, we present typical snapshots of the peptide in Fig. 1. It is clearly seen that compared to the initial configuration at the highest temperature (left), the conformations obtained by population annealing at lower temperatures (to the right) are significantly more compact.

Thermodynamic properties at the temperatures included in the simulation protocol can be extracted from regular population averages. Beyond that, effectively continuous estimates of observables can be derived from the density of states $\Omega(E)$ that can be obtained using the weighted-histogram analysis method (WHAM) [33, 34] applied to the potential energy E. Details of the application to population annealing can be found in Ref. [35]. The result of this analysis for Met-enkephalin is shown in Fig. 2(a). Note that due to the continuous nature of the energy

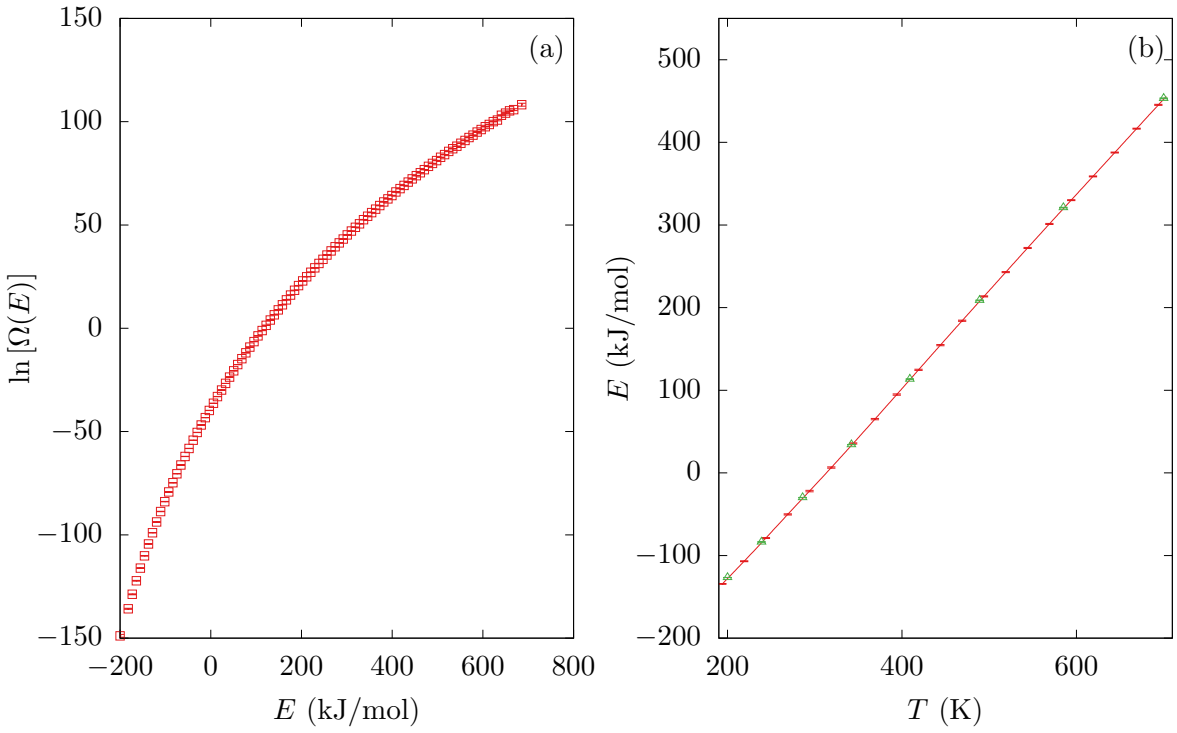


Figure 2. (a) Logarithm of the density of states for Met-enkephalin obtained from our PAMD simulation using WHAM [33, 34, 35]. (b) (Potential) energy E as derived from the density of states of (a) as a function of temperature T. The green triangles indicate population averages at the actual simulation temperatures.

functional one needs to use binning to apply WHAM. Here, we used 200 equally spaced bins in the range from the lowest energy observed to the highest. (The binning is only applied after the simulation, in the post-processing of data.) Using this result for the density of states, it is then possible to extract observable estimates at any temperature such as

$$\langle E\rangle(T) = \frac{\sum_E E\Omega(E)e^{-E/k_BT}}{\sum_E \Omega(E)e^{-E/k_BT}}. \tag{1}$$

The resulting plot is shown in Fig. 2(b) for temperatures in the simulation range of 200 K $< T <$ 700 K.

The spatial structure of the peptide can be characterized using quantitative measures such as the end-to-end distance,

$$R_{\mathrm{ee}} = |\mathbf{R}_N - \mathbf{R}_1|, \tag{2}$$

where $\mathbf{R}_N$ is the position of the last atom of the peptide and $\mathbf{R}_1$ that of the first. For a peptide it is not particularly useful to apply this definition directly and, instead, we rather use the distance between the two outermost carbon atoms. While it is known that this quantity can be misleading, e.g., when "by chance" the start and end monomer in an otherwise extended conformation are very close to each other, it may provide useful insight when directly compared to other physical quantities. A more robust indicator of the average size is the (squared) radius-of-gyration,

$$R_{\mathrm{g}}^2 = \frac{\sum_{i=1}^N m_i(\mathbf{R}_i - \mathbf{R}_{\mathrm{com}})^2}{\sum_{i=1}^N m_i}, \tag{3}$$

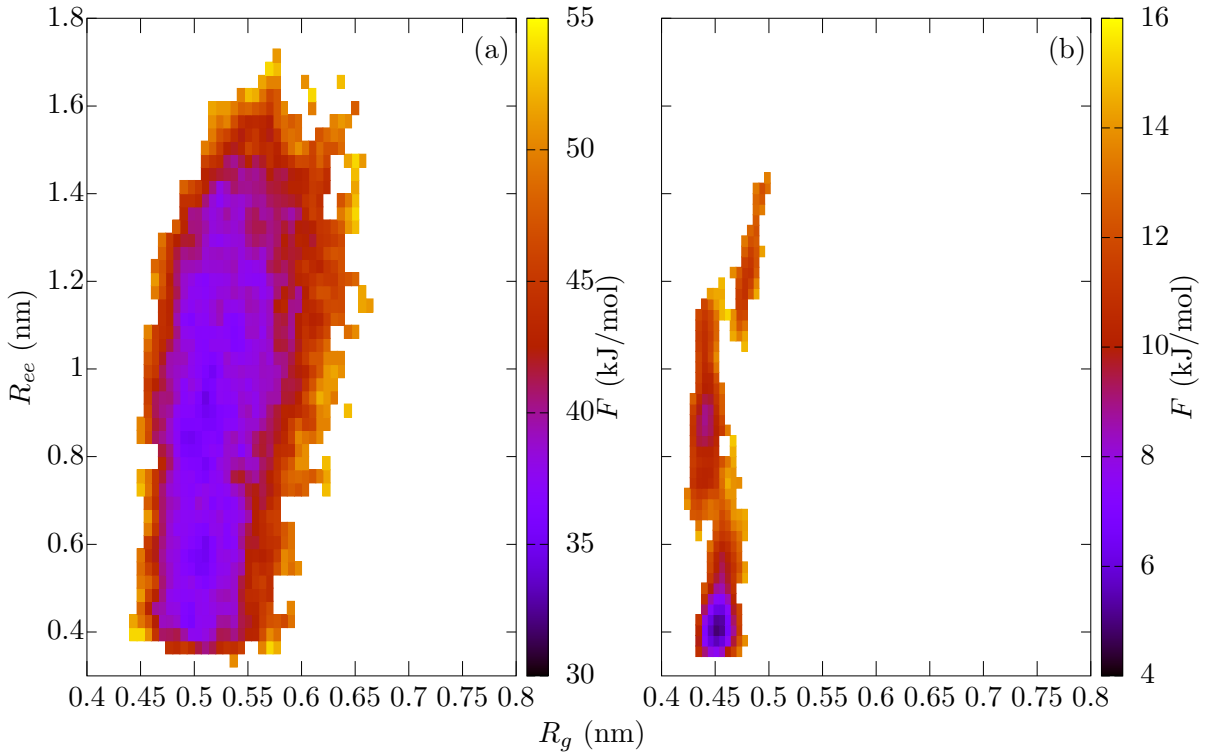


Figure 3. Plots of the end-to-end distance versus the radius-of-gyration for our simulation of Met-enkephalin at (a) $T = 700$ K and (b) $T = 200$ K.

where the sum from $i = 1$ to N runs over all atoms and $\mathbf{R}_{\text{com}} = \sum_i m_i \mathbf{R}_i / \sum_i m_i$ is the center-of-mass position of the peptide. This observable thus averages the squared distance of each atom weighted by its mass to the center-of-mass and is hence a well defined measure for the current extension of the peptide.

While it can already be useful to consider averages and potentially higher moments of quantities such as R_{ee} and R_{g}, it is even more instructive to consider their full probability distributions. In general, define the probability density of an observable $\mathcal{O}$ as

$$P(\mathcal{O}) \sim \iint \mathrm{d}\mathbf{q}\mathrm{d}\mathbf{p}\, P(\mathbf{q},\mathbf{p})\delta\left[\mathcal{O}(\mathbf{q},\mathbf{p}) - \mathcal{O}\right], \tag{4}$$

where $\mathbf{q}$ are the positions and $\mathbf{p}$ the momenta of the atoms. It is then sometimes useful to analyze the *free-energy surface*

$$F(\mathcal{O}) = -k_B T \ln\left[P(\mathcal{O})\right] + c. \tag{5}$$

The constant c is arbitrary and chosen to be zero here. This signifies the observation that only free-energy differences are significant, as the normalization is unknown from simulations.

As an illustration, we plot in Fig. 3 the free-energy surfaces of the end-to-end distance versus the radius-of-gyration for (a) the highest temperature of 700 K and (b) the lowest temperature of 200 K. As expected, both the end-to-end distance and radius-of-gyration are significantly smaller for lower temperatures, where the peptide is in a more compact conformation. At $T = 200$ K the free-energy landscape splits in two minima separated by a region in which no conformations are observed. This signifies two different folding states, which using canonical simulations would not have been observable in a single run.

4. Conclusion

We have described population annealing for molecular dynamics and performed simulations for the penta-peptide Met-enkephalin, illustrating the utility of this approach for efficiently thermalizing the system and estimating thermodynamic quantities. With the help of the weighted-histogram analysis method it is possible to get reliable estimates of the density of states that in turn enable the derivation of quasi continuous estimates of observables as a function of temperature. This peptide has several free-energy minima, which we present by plotting the free-energy surface of the end-to-end distance versus the radius-of-gyration. Here, at least two basins of attractions are observed, indicating the presence of a free-energy barrier at $T = 200$ K. In the population simulation method employed here, at low temperatures different basins are occupied by replicas according to their statistical weight, thus ensuring a fair sampling of the full free-energy landscape while allowing the efficient use of massively parallel computational resources.

Acknowledgments

This project was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Grant No. 189 853 844 – SFB/TRR 102 (project B04), and further supported by the Deutsch-Französische Hochschule (DFH-UFA) through the Doctoral College "$\mathbb{L}^4$" under Grant No. CDFA-02-07, the EU Marie Curie IRSES network DIONICOS under Grant No. PIRSES-GA-2013-612707, and the Leipzig Graduate School of Natural Sciences "BuildMoNa".

References

[1] Scheraga H A, Khalili M and Liwo A 2007 *Annu. Rev. Phys. Chem.* **58** 57–83
[2] Kitchen D B, Decornez H, Furr J R and Bajorath J 2004 *Nat. Rev. Drug Discov.* **3** 935–949
[3] Janke W (ed) 2007 *Rugged Free Energy Landscapes — Common Computational Approaches to Spin Glasses, Structural Glasses and Biological Macromolecules* (*Lect. Notes Phys.* vol 736) (Berlin: Springer)
[4] Marinari E and Parisi G 1992 *Europhys. Lett.* **19** 451–458
[5] Geyer C J 1991 Markov chain Monte Carlo maximum likelihood *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (New York: American Statistical Association) pp 156–163
[6] Hukushima K and Nemoto K 1996 *J. Phys. Soc. Jpn.* **65** 1604–1608
[7] Berg B A and Neuhaus T 1992 *Phys. Rev. Lett.* **68** 9–12
[8] Hansmann U H E 1997 *Chem. Phys. Lett.* **281** 140–150
[9] Sugita Y and Okamoto Y 1999 *Chem. Phys. Lett.* **314** 141–151
[10] Hamelberg D, Mongan J and McCammon J A 2004 *J. Chem. Phys.* **120** 11919
[11] Laio A and Parrinello M 2002 *PNAS* **99** 12562–12566
[12] Wang F and Landau D P 2001 *Phys. Rev. Lett.* **86** 2050–2053
[13] Weigel M 2018 Monte Carlo methods for massively parallel computers *Order, Disorder and Criticality* vol 5 ed Holovatch Y (Singapore: World Scientific) pp 271–340
[14] Gross J, Zierenberg J, Weigel M and Janke W 2018 *Comput. Phys. Commun.* **224** 387–395
[15] Chodera J D, Singhal N, Pande V S, Dill K A and Swope W C 2007 *J. Chem. Phys.* **126** 155101
[16] Pande V S, Beauchamp K and Bowman G R 2010 *Methods* **52** 99–105
[17] Christiansen H, Weigel M and Janke W 2019 *Phys. Rev. Lett.* **122** 060602
[18] Christiansen H, Weigel M and Janke W 2019 *J. Phys. Conf. Ser.* **1163** 012074
[19] Iba Y 2001 *Trans. Jpn. Soc. Artif. Intell.* **16** 279–286
[20] Hukushima K and Iba Y 2003 *AIP Conf. Proc.* **690** 200–206
[21] Machta J 2010 *Phys. Rev. E* **82** 026704
[22] Barash L Y, Weigel M, Borovský M, Janke W and Shchur L N 2017 *Comput. Phys. Commun.* **220** 341–350
[23] Holland J H *et al.* 1992 *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (MIT Press, Cambridge, USA)
[24] Deaven D M and Ho K M 1995 *Phys. Rev. Lett.* **75** 288–291
[25] Beasley J E and Chu P C 1996 *Eur. J. Oper. Res.* **94** 392–404
[26] Kirkpatrick S, Gelatt C D and Vecchi M P 1983 *Science* **220** 671–680
[27] Hughes J, Smith T, Kosterlitz H, Fothergill L A, Morgan B and Morris H 1975 *Nature* **258** 577–579

[28] Kuniyasu H, Luo Y, Fujii K, Sasahira T, Moriwaka Y, Tatsumoto N, Sasaki T, Yamashita Y and Ohmori H 2010 *Gut* **59** 348–356
[29] Cornell W D, Cieplak P, Bayly C I, Gould I R, Merz K M, Ferguson D M, Spellmeyer D C, Fox T, Caldwell J W and Kollman P A 1995 *J. Am. Chem. Soc.* **117** 5179–5197
[30] Katzgraber H G, Trebst S, Huse D A and Troyer M 2006 *J. Stat. Mech.: Theory and Exp.* **2006** P03018
[31] Bittner E, Nussbaumer A and Janke W 2008 *Phys. Rev. Lett.* **101** 130603
[32] Eastman P, Swails J, Chodera J D, McGibbon R T, Zhao Y, Beauchamp K A, Wang L P, Simmonett A C, Harrigan M P, Stern C D *et al.* 2017 *PLOS Comput. Biol.* **13** e1005659
[33] Ferrenberg A M and Swendsen R H 1989 *Phys. Rev. Lett.* **63** 1195–1198
[34] Kumar S, Rosenberg J M, Bouzida D, Swendsen R H and Kollman P A 1992 *J. Comput. Chem.* **13** 1011–1021
[35] Barash L, Marshall J, Weigel M and Hen I 2019 *New. J. Phys.* **21** 073065