Learned mappings for targeted free energy perturbation between peptide conformations ⊘

Soohaeng Yoo Willow ⓘ ; Lulu Kang ⓘ ; David D. L. Minh ✉ ⓘ

Check for updates

View Online   Export Citation

Learned mappings for targeted free energy perturbation between peptide conformations

View Online   Export Citation   CrossMark

Soohaeng Yoo Willow,[1] ID **Lulu Kang,**[2] ID **and David D. L. Minh**[3,a) ID

AFFILIATIONS

[1] Department of Chemistry, Illinois Institute of Technology, Chicago, Illinois 60616, USA
[2] Department of Applied Mathematics, Illinois Institute of Technology, Chicago, Illinois 60616, USA
[3] Department of Chemistry, Department of Biology, and Center for Interdisciplinary Scientific Computation, Illinois Institute of Technology, Chicago, Illinois 60616, USA

[a)]Author to whom correspondence should be addressed: dminh@iit.edu

ABSTRACT

Targeted free energy perturbation uses an invertible mapping to promote configuration space overlap and the convergence of free energy estimates. However, developing suitable mappings can be challenging. Wirnsberger *et al.* [J. Chem. Phys. **153**, 144112 (2020)] demonstrated the use of machine learning to train deep neural networks that map between Boltzmann distributions for different thermodynamic states. Here, we adapt their approach to the free energy differences of a flexible bonded molecule, deca-alanine, with harmonic biases and different spring centers. When the neural network is trained until "early stopping"—when the loss value of the test set increases—we calculate accurate free energy differences between thermodynamic states with spring centers separated by 1 Å and sometimes 2 Å. For more distant thermodynamic states, the mapping does not produce structures representative of the target state, and the method does not reproduce reference calculations.

Published under an exclusive license by AIP Publishing. https://doi.org/10.1063/5.0164662

I. INTRODUCTION

Free energy calculations are a powerful tool that are increasingly used to design materials[1] and drugs.[2] Accurately calculating free energy differences between a pair of thermodynamic states often entails performing simulations of multiple intermediate states along a thermodynamic process connecting the states.[3] Simulating multiple intermediates promotes overlap between the most important configuration spaces in neighboring thermodynamic states, a requirement for the convergence of free energy estimates.[4–6]

Circumventing the simulation of these intermediates could improve simulation accuracy and reduce resource requirements. Simulations of intermediates can have problems that degrade the accuracy of free energy estimates. For example, alchemical processes in which particles are created or destroyed are known to suffer from pathologies such as the end-point catastrophe—there can be poor configuration space overlap between states where particles nearly appear or disappear, leading to spurious discontinuities in free energy estimates along a thermodynamic process—and the

presence of artificial energy minima.[7] Moreover, intermediate states are usually of no particular scientific interest. Avoiding them could lower resource requirements, allowing more scientists to access the tool and reducing the consumption of computing and energy resources.

Targeted free energy perturbation (TFEP)[8] is an approach that can bypass the simulation of intermediate states. Rather than simulating many intermediates, TFEP achieves configuration space overlap via an invertible mapping between the end states. In 2002, Jarzynski[8] formulated TFEP as an extension of the classic free energy perturbation (FEP) identity.[9] In FEP, configurations are sampled from a single state. The free energy is an exponential average of the difference between the potential energy of each sampled configuration in the target and sampled states [Eq. (2)]. On the other hand, instead of using the same configuration for both potential energy evaluations, TFEP uses the potential energy of the *mapped* configuration in the target state. Along with the Jacobian of the mapping, the potential energy of the mapped configuration is incorporated into a generalized work [Eq. (6)] in the exponential average [Eq. (5)].

There were several precedents for the formalism of Jarzynski.[8] In 1985, Voter[10] performed free energy calculations with volume-preserving translations. In 1995, Severance et al.[11] considered transformations of bond lengths and angles based on harmonic force field parameters as well as the rotation of a dihedral angle. Although they did not explicitly mention Jacobians, all of their example mappings had a Jacobian determinant of unity; the term was not needed to reproduce analytical free energies. TFEP can be considered a generalization of the celebrated Jarzynski identity[12,13] that was published in 1997; in the Jarzynski identity, mapping is performed by a nonequilibrium switching process.[14] In 2000, Miller and Reinhardt[15] proposed that such switching processes can incorporate scaling transformations.

There have been several extensions and generalizations to TFEP. Hahn and Then[16] implemented the bidirectional estimator that Jarzynski[8] suggested in the original TFEP paper. Soon after Jarzynski,[8] Meng and Schilling[17] described a multistate generalization of TFEP to arbitrary statistical distributions, which Paliwal and Shirts[18] applied to molecular systems. In addition to free energies, the multistate generalization is capable of estimating arbitrary expectation values.[17,18] Subsequently, TFEP and its extensions have been applied to spin systems and other statistical distributions, but our focus here will be on molecular systems.

Although TFEP is a compelling concept, it has been difficult to apply to molecular systems because of the lack of suitable mapping functions. Mappings have been developed for expanding a cavity in a fluid,[8] inserting a particle into a fluid,[16] the interconversion of water models,[18] and between crystals at different temperatures[19] and volumes.[20] However, the mappings are not broadly generalizable; human intuition and creativity have been required for every distinct application.

While human intelligence is still important, we are now in the age of artificial intelligence. Apropos of the emerging era, a few groups have used deep learning to train maps between molecular systems for TFEP. Wirnsberger et al.,[21] from the Google subsidiary DeepMind, who also developed AlphaFold,[22] learned mappings to ensembles of fluids containing a solute with different radii. Ding and Zhang[23] trained mappings between a Boltzmann and tractable distribution—for which the normalized density is known and independent and identically distributed samples can be easily generated—to compute absolute free energies for different conformations of di-alanine and temperatures of deca-alanine. Subsequently, they applied the same approach to a host–guest system to compute binding free energies.[24] Rizzi et al.[25] trained neural networks for mapping between two levels of quantum theory for a simple chemical reaction in the gas phase. Wirnsberger et al.[26] learned mappings between a lattice model with random perturbations and different phases of Lennard-Jones systems.

In recent years, the use of learned mappings in molecular simulation beyond TFEP has been pioneered by Frank Noé and co-workers. In 2019, Noé et al.[27] introduced Boltzmann generators, which use a reference state and a learned mapping to generate samples that may be reweighed to the Boltzmann distribution. They demonstrated the method for generating samples of a bistable dimer in fluid (initially described in a paper coauthored by the corresponding author[28]) and a small protein. While their publication primarily focused on sampling as opposed to free energies, they did report

computing free energy differences between independent Boltzmann generators using the average (opposed to the exponential average from TFEP[8]) of the generalized work. Subsequently, Sbailò et al.[29] described the use of learned mappings for Monte Carlo moves, as demonstrated in the bistable dimer system. Finally, Invernizzi et al.[30] used learned mappings for replica exchange, performing simulations of di-alanine and tetra-alanine. Besides Noé and co-workers, Mahmoud et al.[31] have developed a hierarchical sampling procedure to generate samples of a medium-sized (106 residue) protein.

Here, we apply learned mappings to free energy differences between different conformations of deca-alanine. Our approach is similar to that of Wirnsberger et al.,[21] but the system is qualitatively different. Wirnsberger et al.[21] modeled a cavity in a fluid of neutral monatomic molecules, which has a simpler potential energy function due to the lack of bonded (bond length, bond angle, and torsion) and Coulomb interactions. Our system is similar to Ding and Zhang's,[23] but the approach is qualitatively different. While Ding and Zhang[23] learned mappings between Boltzmann and tractable distributions, we learn mappings between Boltzmann distributions corresponding to different thermodynamic states (different harmonic biases). Mapping to a tractable distribution has the advantage that the tractable distribution can be selected to have significant configuration space overlap with the unmapped molecular distribution. Moreover, it is generally faster to evaluate the probability density and its derivatives for a tractable distribution than the potential energy of a molecular system. On the other hand, directly mapping between pairs of molecular distributions can permit a subset of coordinates to be perturbed. For processes in which changes are limited to a subset of coordinates, such as those in the focused confinement method,[32,33] such mappings may be particularly fruitful. Hence, both approaches to mapping may find separate niches.

The remainder of the paper is as follows: In Sec. II, we review the theory of TFEP and the loss function. We then describe computational methods (Sec. III), including the deca-alanine model system, the neural network architecture, the training procedure, and free energy estimators. We then report the results along with a discussion of their implications (Sec. IV) and finally describe our conclusions (Sec. V).

II. THEORY

A. Targeted free energy perturbation

Let us first define notation. For a thermodynamic state A, a configuration x has the equilibrium probability density $\rho_A(x) = \exp[-\beta U_A(x)]/Z_A$. In this expression, β is the inverse of the temperature and Boltzmann's constant k_B, $\beta = (k_B T)^{-1}$, $U_A(x)$ is the potential energy of the configuration in state A, and $Z_A = \int \exp[-\beta U_A(x)]dx$ is the configurational integral of the state, an integral over all space. Analogous definitions apply to a second thermodynamic state B. The objective of our calculation is the Helmholtz free energy difference between states A and B, defined as

$$\Delta F = F_B - F_A = -\beta^{-1} \ln\left(\frac{Z_B}{Z_A}\right). \tag{1}$$
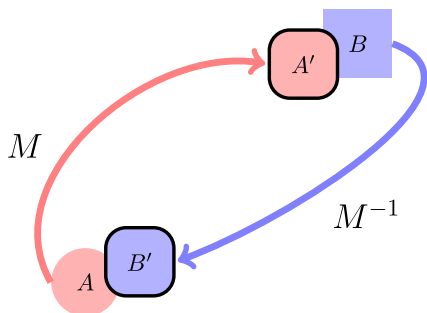
FIG. 1. Schematic showing how mapping can improve configuration space overlap. Each shape represents the highest-probability regions of the respective states.

In principle, a free energy difference ΔF may be calculated based on the classic free energy perturbation identity,[9]

$$\mathbb{E}_A\left[e^{-\beta \Delta U}\right] = e^{-\beta \Delta F}, \qquad (2)$$

where $\mathbb{E}_A[\cdot]$ represents an expectation value in the state A, and $\Delta U(\boldsymbol{x}) = U_B(\boldsymbol{x}) - U_A(\boldsymbol{x})$ is the potential energy difference of the *same* configuration between the two states. (For notational simplicity, the configuration dependence is implicit within the expectation value.) Specifically, a free energy difference may be estimated using the sample mean of the observable $\exp(-\beta \Delta U(\boldsymbol{x}))$ over configurations drawn from state A. As exponential averages suffer from finite-sample bias,[34] better estimation performance may be achieved using samples from both A and B[35] or from multiple states (including A and B).[36] Regardless of the estimator, free energy calculations do not reliably converge to the true values unless there is overlap between the highest-probability regions of configuration space for the pair (or series) of states.[4–6]

Configuration space overlap may be improved by bijective (one-to-one) and invertible mapping (Fig. 1). A mapping is simply a function that changes a configuration $\boldsymbol{x}$ into a new configuration $\boldsymbol{x}' = M(\boldsymbol{x})$. For an invertible mapping, the inverse function transforms $\boldsymbol{x}'$ into $\boldsymbol{x}$, $\boldsymbol{x} = M^{-1}(\boldsymbol{x}')$. Applying a mapping to state A creates a new state A'. The probability of the new state is given by the change of variables formul

$$\rho_{A'}(M(\boldsymbol{x})) = \frac{\rho_A(\boldsymbol{x})}{|J_M(\boldsymbol{x})|}, \qquad (3)$$

where $|J_M(\boldsymbol{x})|$ is the determinant of the Jacobian of the mapping. Similarly, the inverse mapping can be applied to the state B to create a new state B' with density,

$$\rho_{B'}(M^{-1}(\boldsymbol{x})) = \frac{\rho(\boldsymbol{x})}{|J_{M^{-1}}(\boldsymbol{x})|}. \qquad (4)$$

A mapping can be defined such that A' is close to B, and conversely, B' is close to A (Fig. 1).

Mappings may be incorporated into free energy calculations via a generalization of Eq. (2), TFEP,[8]

$$\mathbb{E}_A\left[e^{-\beta \Phi_F}\right] = e^{-\beta \Delta F}. \qquad (5)$$

Equation (5) replaces the ΔU in Eq. (2) with a generalized work

$$\Phi_F(\boldsymbol{x}) = U_B(M(\boldsymbol{x})) - U_A(\boldsymbol{x}) - \beta^{-1} \log|J_M(\boldsymbol{x})|, \qquad (6)$$

in which the configurations in states A and B may be different. We refer to this term as work because it is a generalization[14] of the nonequilibrium work from Jarzynski's identity.[12,13] The work includes a subscript F to denote a "forward" mapping from A to A'. For the "reverse" mapping from B to B', the work includes the subscript R,

$$\Phi_R(\boldsymbol{x}) = U_A(M^{-1}(\boldsymbol{x})) - U_B(\boldsymbol{x}) - \beta^{-1} \log|J_{M^{-1}}(\boldsymbol{x})|. \qquad (7)$$

The designation of a direction as forward or reverse is arbitrary. In the case of an identity mapping, the work is the conventional potential energy difference, and Eq. (5) reduces to Eq. (2). However, a mapping that increases configuration space overlap can improve the convergence of free energy estimates. Indeed, if a mapping perfectly transforms between A and B,

$$\frac{\rho_{A'}(M(\boldsymbol{x}))}{\rho_B(\boldsymbol{x})} = \frac{\rho_{B'}(M^{-1}(\boldsymbol{x}))}{\rho_A(\boldsymbol{x})} = 1, \qquad (8)$$

then only a single sample is needed to estimate the free energy difference.[8,25] In addition to the sample mean estimator based on Eq. (5), mappings have also been incorporated into bidirectional[16] and multistate[17,18] estimators for free energies and expectation values.

B. Loss functions

While it can be challenging to *design* mappings that improve configuration space overlap and accelerate the convergence of free energy calculations, Wirnsberger *et al.*[21] demonstrated that machine learning may be used to *train* mappings. Inspired by configuration space overlap, they developed loss functions for their machine learning models based on a statistical distance between mapped and targeted distributions, the Kullback–Leibler (KL) divergence.[37] The quality of the forward mapping is quantified by the KL divergence between the mapped density $\rho_{A'}(M(\boldsymbol{x}))$ and the target density $\rho_B(\boldsymbol{x})$

$$D_{\mathrm{KL}}[\rho_{A'} \| \rho_B] = \beta(\mathbb{E}_A[\Phi_F] - \Delta F). \qquad (9)$$

Conversely, the quality of the reverse mapping is described by the KL divergence between $\rho_{B'}(M^{-1}(\boldsymbol{x}))$ and $\rho_A(\boldsymbol{x})$

$$D_{\mathrm{KL}}[\rho_{B'} \| \rho_A] = \beta(\mathbb{E}_B[\Phi_R] + \Delta F). \qquad (10)$$

As β and ΔF are constants with respect to mapping, they do not need to be included in loss functions. Hence, a loss function suitable for training only the forward mapping is

$$\mathcal{L}_F = \bar{\mathbb{E}}_A[\Phi_F], \qquad (11)$$

where $\bar{\mathbb{E}}_A[\cdot]$ is a sample mean estimate of an expectation value in state A. Wirnsberger *et al.*[21] also introduced and recommended a bidirectional loss function,

$$\mathcal{L} = \bar{\mathbb{E}}_A[\Phi_F] + \bar{\mathbb{E}}_B[\Phi_R], \qquad (12)$$

which we use in the present work.

III. COMPUTATIONAL METHODS

A. Model system

Alanine deca-peptide [ACE-(ALA)$_9$-NME] was selected as a model system. The system was modeled in the gas phase at room temperature, where it is a stable α helix. The initial structure was modified by Nguyen and Minh[38] by hand to include the end caps ACE and NME. The system has 102 atoms. Parameters from the AMBER ff14SB force field[39] were used. Harmonic restraints were added to the carbon atom of the first peptide bond and the nitrogen atom of the last peptide bond, leading to the total potential energy,

$$U_\lambda(x) = U_0(x) + k\|\boldsymbol{x}_C - \boldsymbol{x}_C^0\|^2 + k\|\boldsymbol{x}_N - \boldsymbol{x}_N^\lambda\|^2, \qquad (13)$$

where $U_0(x)$ is the standard gas-phase AMBER potential energy. In the harmonic potential, $k = 50$ kJ/mol/Å^2 is the force constant, $\boldsymbol{x}_C$ and $\boldsymbol{x}_N$ are the coordinates of the restrained carbon and nitrogen atoms, respectively, and $\|\cdot\|$ represents the Euclidean norm, the length of the vector. The spring center $\boldsymbol{x}_C^0$ is at the origin, and $\boldsymbol{x}_N^\lambda = (0, 0, \lambda)$, where $\lambda \in \{14, 14.2, 14.4, \ldots, 20.6, 20.8, 21\}$ Å.

For each λ, an ensemble of structures was generated with a molecular dynamics simulation. Isothermal molecular dynamics simulations were performed using the Langevin integrator (LangevinIntegrator) in OpenMM 7.7[40] at a temperature of $T = 300$ K with a time step of 1 fs for a duration of 20 ns. The last 10 000 configurations (every 1 ps for the last 10 ns) were saved.

B. Neural network architecture

Mappings between Cartesian coordinates from pairs of ensembles with different λ were performed based on real-valued non-volume preserving (real NVP) transformations.[41] Real NVP uses a deep neural network composed of a stack of simple bijections, known as affine coupling layers, that include scaling and translation. Because each affine coupling layer only transforms a subset of system coordinates, multiple layers are required to transform all coordinates. Affine coupling layers are especially well-suited for TFEP because they are stably invertible and because the Jacobian of their transformations can be efficiently computed. The initial paper on real NVP uses a mapping to a tractable distribution, but the neural network can be applied to any pair of distributions.

To adapt real NVP to molecular systems in Cartesian coordinates, we tried two sets of affine coupling layers. Initially, we considered coupling layers for x, y, and z dimensions (Fig. 2), which we will refer to as NN 1. Based on reviewer suggestions, we later performed some calculations in which the subsets were backbone heavy atoms, side chain heavy atoms, and hydrogen atoms, which we will refer to as NN 2.

In NN 1, the first coupling layer M_x transforms the x coordinates while leaving the y and z coordinates invariant,

$$M_x(\{x, y, z\}) = \left\{x' = xe^{s(y,z)} + t(y,z), y, z\right\}, \qquad (14)$$

where s and t are functions encoded in the neural network. The inverse of this mapping is $M_x^{-1}(\{x', y, z\})$
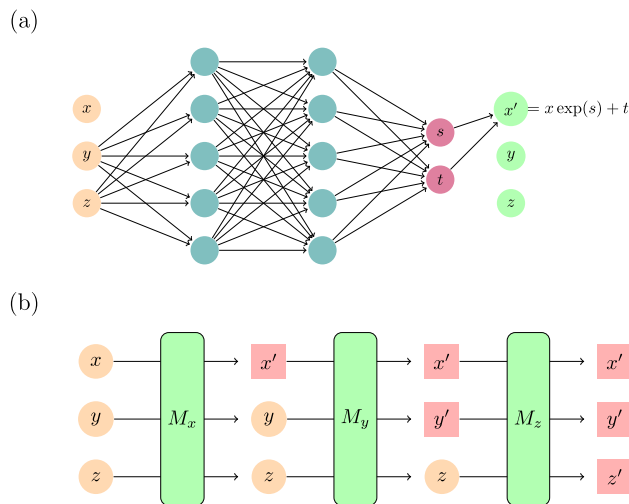

FIG. 2. Schematic of the neural network architecture. (a) In the first coupling layer M_x, y and z are left unchanged and the learnable parameters θ for the s and t functions are trained to transform the x coordinates. (b) The entire real NVP network consists of $M = M_z \circ M_y \circ M_x$.

$= \left\{x = \left[x' - t(y,z)\right]e^{-s(y,z)}, y, z\right\}$. Jacobian determinants of the mappings M_x and M_x^{-1} are

$$\log|J_{M_x}| = \sum_{i=1}^{N_{\text{atoms}}} s_i(y,z), \qquad (15)$$

$$\log|J_{M_x^{-1}}| = -\sum_{i=1}^{N_{\text{atoms}}} s_i(y,z), \qquad (16)$$

where the sum is over the atoms and the limit is the total number of atoms, N_{atoms}. Similarly, the second coupling layer M_y and the third coupling layer M_z transform the y and z coordinates with (x, z) and (x, y) invariants, respectively. Each coupling layer $M_k : \mathbb{T}^{3N} \to \mathbb{T}^{3N}$ used the same architecture but different learnable parameters θ_k. The coupling layers consisted of 10 hidden layers with 128 dimensions (compared to 102 atoms).

The overall mapping M of the real NVP transformation is given as

$$M = M_z \circ M_y \circ M_x, \qquad (17)$$

where $\circ$ means that the output of the right side is the input for the left side. Its inverse is

$$M^{-1} = M_x^{-1} \circ M_y^{-1} \circ M_z^{-1}. \qquad (18)$$

The Jacobian determinants of the mapping M and the inverse mapping M^{-1} are

$$\log|J_M| = \sum_{v\in\{x,y,z\}} \log|J_{M_v}|, \quad \log|J_{M^{-1}}| = \sum_{v\in\{x,y,z\}} \log|J_{M_v^{-1}}|. \quad (19)$$

NN 2 had a similar architecture except that M_x, M_y, and M_z were replaced by M_b for backbone heavy atoms, M_{sc} for

side chain heavy atoms, and M_h for hydrogen atoms, respectively. The number of hidden dimensions was 3N dimensions, where N is the number of atoms. The three mappings form a block such that $M_n = M_b \circ M_{sc} \circ M_h$. The total mapping was composed of five blocks of complete mappings, $M = M_1 \circ M_2 \circ M_3 \circ M_4 \circ M_5$. Calculations were performed with NN 1 unless explicitly mentioned otherwise.

Mappings were implemented in JAX.[42]

C. Training

The real NVP map was initialized as an identity map. This was achieved by initializing the transformation t and scaling s factors to zero, such that $x' = x \exp(0) + 0 = x$. Randomness in the initial conditions was introduced by drawing all other initial parameters from the standard normal distribution.

The real NVP neural network was trained by minimizing the value of the loss function [Eq. (12)]. As the loss function [Eq. (12)] is based on the molecular mechanics energies of mapped configurations, we implemented the AMBER force field in JAX. This implementation is freely available at https://github.com/swillow/jax_amber. To ensure that mapped x_N is similar to x_N^λ, we used $k = 200$ kJ/mol/Å^2 instead of $k = 50$ kJ/mol/Å^2 in the loss function. Data were divided into a training set (80%) and a test set (20%). Out of the last 10 ns of simulation, 2000 configurations (from 10 to 12 ns) were designated as test sets, while the remaining 8000 configurations (from 12 to 20 ns) were designated as training sets. The training loss was minimized using the adaptive moment estimation (Adam) optimizer[43] with a learning rate of 1.0×10^{-4}.

We performed different amounts of training in the initial and later calculations. Initially, training was performed for 150 000 steps. We compared results from mappings obtained after different amounts of training:

1. Early stopping, with the minimum loss of the test set.
2. L ~ 0, where the loss of the training set was closest to zero.
3. Complete, after all 150 000 training steps.

Subsequently, after observing superior free energy estimation from early stopping, we stopped training once the loss of the test set in the last 200 steps was over $3.0k_BT$ greater than the minimum.

D. Free energy estimation

Reference free energies as a function of λ were calculated using the multistate Bennett acceptance ratio (MBAR) estimator[36] without mapping. They made us aware of the entire ensemble of equilibrated structures and energies. Mapped free energies were estimated based on data from pairs of λ using the generalization of the Bennett acceptance ratio[35] described by Hahn and Then.[16] Thus, the mapped calculations were based on a small subset of the structures and energies used in the reference calculations.

IV. RESULTS AND DISCUSSION

A. The free energy landscape of deca-alanine has multiple barriers

The free energy of deca-alanine as a function of the spring position generally increases between $\lambda = 14$ and $\lambda = 21$ Å (Fig. 3). There

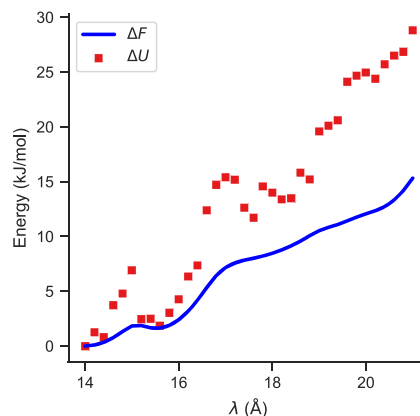**FIG. 3.** The reference free energy surface and average potential energy difference compared to the state with $\lambda = 14$ Å.

are local minima near 14, 16, and 18 Å that are separated by barriers. The barriers correspond to local peaks in the potential energy, suggesting that crossing them corresponds to breaking intramolecular interactions such as hydrogen bonds. As the potential energy barriers are larger than the free energy barriers, they are partially compensated for by entropic increases.

B. Overlap decreases as λ separation increases

While there may be some overlap between thermodynamic states with λ separated by 1 Å, there does not appear to be any overlap between states where it is separated by 2 Å or more ($\Delta\lambda \geq 2$ Å). The overlap between thermodynamic states may be evaluated based on the histograms of Φ_F and $-\Phi_R$. Without mapping (or an identity map), the work is simply the potential energy difference of the same configuration in the pair of states. For both separations, the distributions of the potential energy difference are unimodal. For $\Delta\lambda = 1$ Å,

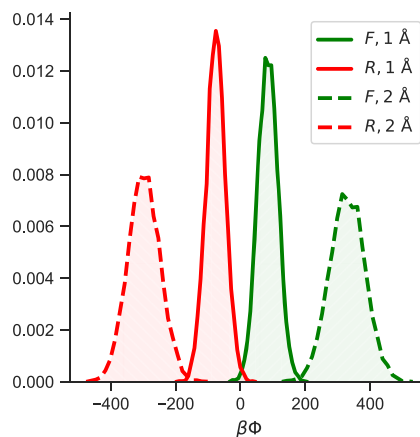**FIG. 4.** Normalized histograms of work without mapping for $\lambda_A = 20$ Å and $\lambda_B = 21$ Å (solid line) and for $\lambda_A = 19$ Å and $\lambda_B = 21$ Å (dashed line). For both pairs, $\beta\Phi_F$ is colored green with forward diagonal hatches, and $-\beta\Phi_R$ is colored red with backward diagonal hatches.
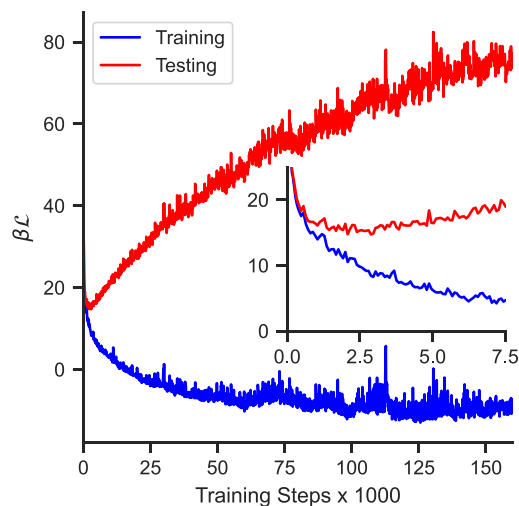
FIG. 5. Loss values for the mapping between states with $\lambda_A = 20$ Å and $\lambda_B = 21$ Å for training (blue) and test (red) data. The inset is a subset of the range emphasizing the minimum of the test loss.

there is some overlap between the tails of the distributions of Φ_F and $-\Phi_R$. For $\Delta\lambda = 2$ Å, the distributions of potential energy differences are broader and further separated, such that there is no evident overlap (Fig. 4). Due to the lack of overlap between these distributions, the free energy difference cannot be accurately calculated (without a mapping that improves overlap).

Even for $\Delta\lambda \leq 2$ Å, our unmapped work distributions are more distinct from one another than those reported in Wirnsberger *et al.*[21] Wirnsberger *et al.*[21] focused on a solute in a Lennard-Jones fluid, in which the states A and B were generated by changing the solute radii from $R_A = 2.5974\sigma$ to $R_B = 2.8444\sigma$. For $\sigma = 3.15$ Å, the radius of the solute in state A is 8.18 Å and that of the solute in state B is 8.96 Å, for a difference of 0.78 Å. The distance between unmapped work distributions was about $\beta(\mathbb{E}_A[\Phi_F] + \mathbb{E}_B[\Phi_R]) \sim 40$. In contrast, we observe $\beta(\mathbb{E}_A[\Phi_F] + \mathbb{E}_B[\Phi_R]) \sim 200$ and 600 for 1 and 2 Å, respectively.

C. Loss values for the training and test sets diverge after overfitting

In all of the learned mappings, the loss values for the training and test sets exhibit different behaviors from one another (Fig. 5). At the start of training, both loss values quickly decrease. With additional training steps, the training loss continues to trend downward and even becomes negative. On the other hand, the test loss approaches a minimum and sharply increases after around 2000 steps.

A deviation between the training and test loss is common in machine learning, including the learned mapping in Wirnsberger *et al.*[21] and Rizzi *et al.*[25]—and is an indicator of overfitting to the training set. For this reason, Wirnsberger *et al.*[21] adopted an early stopping criterion. Rizzi *et al.*[25] suggested that overtraining leads to a systematic bias in the free energy estimate; thus, they estimated

free energies based on evaluation sets separate from their training sets.

D. Mapping increases the overlap between forward and reverse work distributions

With early stopping, mapping improves overlap between the work distributions and the feasibility of free energy estimation [Fig. 6(a)]. At this mapping, which provides a minimum loss for the test set, there is considerable overlap between the distributions of Φ_F and $-\Phi_R$. Trends in the average work and free energy difference are consistent with applying Jensen's inequality to the expectation values in the TFEP expression, Eq. (5). We anticipate that $\mathbb{E}_A[\Phi_F] > -\beta^{-1} \ln \mathbb{E}_A[e^{-\beta\Phi_F}] = \Delta F$. Similarly, $\mathbb{E}_B[\Phi_R] > -\Delta F$. Hence, expectation values of the work bound the free energy according to $-\mathbb{E}_B[\Phi_R] < \Delta F < \mathbb{E}_A[\Phi_F]$. While these inequalities are true for expectation values, they do not necessarily hold for estimates based on finite sample sizes, which are not only imprecise but systematically biased.[34] With early stopping, however, the order of expectation values and the free energy follow the anticipated trend. Consequently, the free energy difference, which is at the value of Φ in which the densities are equal,[16] is straightforward to identify. However, further training leads to unexpected trends in the distributions of generalized energies.

When the loss is close to zero, we observe that both estimated averages are less than the free energy difference, $\hat{\mathbb{E}}_B[-\Phi_R] \sim \hat{\mathbb{E}}_A[\Phi_F] < \Delta F$ [Fig. 6(b)]. With perfect mapping, we would anticipate that $\mathbb{E}_B[-\Phi_R] = \mathbb{E}_A[\Phi_F] = \Delta F$. The deviation of the expectation values from the free energy difference demonstrates that perfect mapping is not achieved. The potential energy distributions of the mapped configurations mostly overlap with the targeted (unmapped) potential energy distribution (Fig. S2 of the supplementary material). This overlap between potential energy distributions shows that the source of the observed inequality is from the Jacobian: $\log|J_M| \neq 0$ and $\log|J_{M^{-1}}| \neq 0$.

Finally, at the end of training, when the training loss values are negative ($\mathcal{L} < 0$), the forward work values are actually much smaller than the reverse work values. As shown in Fig. 6(c), the order of averages and free energies is flipped such that $\hat{\mathbb{E}}_A[\Phi_F] \ll \Delta F \ll \hat{\mathbb{E}}_B[-\Phi_R]$. This flipping occurs because the mapped energies $U_B(M(\mathbf{x}_A))$ and $U_A(M^{-1}(\mathbf{x}_B))$ are lower than the targeted energies $U_B(\mathbf{x}_B)$ and $U_A(\mathbf{x}_A)$ (Fig. ?? of the supplementary material). Molecular mechanics potential energies, especially Lennard-Jones repulsion terms, are very sensitive to small changes in coordinates; there are many opportunities for energy optimization relative to samples from the Boltzmann distribution. While these low energies demonstrate that the optimizer is working well, the generated structures are not representative of the target distribution. Moreover, the overlap between work distributions is reduced compared to when the loss is close to zero.

Except with early stopping, trends in the work distributions of the test set do not match those of the training set (Fig. ?? in the supplementary material). With early stopping, the training and test sets have comparable work distributions. When $\mathcal{L} \sim 0$, the work distributions for the test set are similar to the distributions after early stopping; $\rho(\beta\Phi_F)$ and $\rho(-\beta\Phi_R)$ of the test set do not have nearly complete overlap. For complete training, the anticipated order of averages and free energies $\hat{\mathbb{E}}_B[-\Phi_R] < \Delta F$
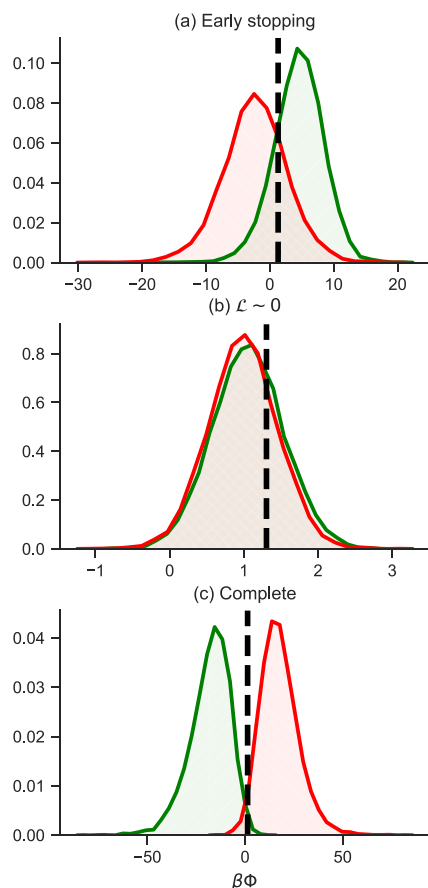
FIG. 6. Effect of stopping criterion on work distributions for training data between $\lambda_A = 20$ Å and $\lambda_B = 21$ Å. Normalized histograms are shown for mapping functions from (a) early stopping, (b) when the loss is closest to zero, and (c) after 150 000 training steps. $\beta\Phi_F$ are colored green with forward diagonal hatches, and $-\beta\Phi_R$ are colored red with backward diagonal hatches. The reference free energy difference ΔF obtained via the MBAR estimator is represented by the vertical dashed line. A similar figure for the testing data is available as Fig. S1 in the supplementary material.

FIG. 7. Comparison of free energy estimates for a separation of 1 Å. Mapped free energy differences were computed between λ_A and $\lambda_B = \lambda_A + 1$ Å for $\lambda_A \in \{14, 15, \ldots, 20\}$ Å. (a) Free energy surfaces were computed as a cumulative sum. (b) Individually mapped $\Delta\hat{F}$ were compared to reference values. In both panels, estimates were computed with maps from early stopping (magenta squares) and $\mathcal{L} \sim 0$ (green circles).

$< \hat{\mathbb{E}}_A[\Phi_F]$ is preserved, but the distributions are very broad. The corresponding large test loss indicates that there is overfitting in the training set.

E. Mapping enables accurate free estimation for smaller $\delta\lambda$

For most pairs of states where $\Delta\lambda = 1$ Å, TFEP is able to accurately reproduce reference free energy differences. Mapped estimates are most accurate between pairs of states with ΔF less than 2 kJ/mol, less accurate when ΔF is between 2 and 4 kJ/mol, and least accurate for ΔF greater than 4 kJ/mol [Fig. 7(b)].

Among stopping criteria, TFEP is more reliable when using a learned mapping with early stopping. When the loss value is near zero, TFEP performance is sometimes improved over early stopping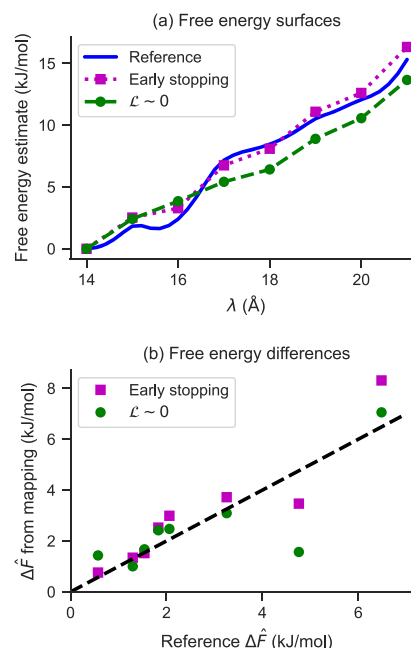, but in several cases with large free energy differences, the estimate significantly deviates from the reference value. When $-\mathbb{E}_B[\Phi_R] = \mathbb{E}_A[\Phi_F] = \Delta F$, we would expect that the free energies are easy to predict. However, we have already shown [Fig. 6(b)] that the learned mappings do not provide perfect mappings and that there is overfitting to the training set.

Differences between stopping criteria are more evident for free energy surface reconstruction [Fig. 7(a)]. With early stopping, the estimated free energy surface reproduces the local minimum near 16 Å and the barrier between that state and at 18 Å. Finer details in the free energy surface are not evident because the mapped estimate was performed at intervals of 1 Å as opposed to 0.2 Å. In contrast to the accurate surface reconstruction using the mapping from early stopping, when the training loss is near zero, the estimated free energy surface has few features.

These results demonstrate the ability to achieve configuration space overlap and accurate free energy estimation through learned mapping, as opposed to sampling of additional intermediate states. Comparable free energy differences are obtained using a fifth of the number of samples between adjacent λ values. While the mapped estimation procedure incurs additional costs in training and is more complex, it may lead to an overall reduction in computational cost.

Using early stopping, reasonable mapped free energy estimates were also obtained for select pairs of states for which $\Delta\lambda = 2$. For $\lambda_A = 18$ Å and $\lambda_B = 20$ Å, the mapped estimate based on early

TABLE I. Comparison of TFEP and MBAR free energy estimates for 4 Å separations. TFEP estimates were based on learned mappings with two different neural network architectures. For each architecture, results from two independent trainings are separated by a slash.

(λ_A, λ_B)	ΔF_{ref}	ΔF_{TFEP} (NN 1)	ΔF_{TFEP} (NN 2)
(14, 18)	8.461	−20.4/−35.2	−25.1/−34.1
(15, 19)	8.691	−25.4/−11.4	−31.8/−19.9
(16, 20)	9.269	11.3/7.8	−10.6/−17.9
(17, 21)	8.154	4.8/17.2	4.8/−1.3

stopping (3.1 kJ/mol) is better than the $\mathcal{L} \sim 0$ estimate (7.4 kJ/mol) at reproducing the reference free energy (3.6 kJ/mol). For $\lambda_A = 19$ Å and $\lambda_B = 21$ Å, the early stopping estimate (5.5 kJ/mol) is also closer to the reference free energy (4.8 kJ/mol) than the $\mathcal{L} \sim 0$ estimate (6.6 kJ/mol). These pairs of states are not separated by potential energy barriers (Fig. 3).

F. There is poor mapping for more distant states

In contrast to the accurate performance for smaller $\Delta\lambda$, TFEP does not provide accurate free energy differences for pairs of states in which the spring center is separated by 4 Å (Table I). Not only are TFEP results significantly different from the reference calculation, but the two independent training replicates diverge from one another. This suggests that the optimization leads to maps that are quite different from one another. Thus, these maps are at local as opposed to global minima of the loss function. The results for NN 2 were worse than for NN 1. To pinpoint the causes and effects of poor mapping, we further analyzed the work, potential energy, and structural distributions of the mapping between $\lambda_A = 14$ Å and $\lambda_B = 18$ Å.

For this pair of states, TFEP fails to reproduce reference free energies because the work distributions are still not overlapping (Fig. 8). Without mapping, there is a gap of over 2000 k_BT between the peaks of the forward and reverse work distributions. Mapping significantly reduces this gap, but there is no significant probability density at the free energy difference, making it difficult to estimate accurately. The mapping based on NN 2 does not reduce the gap as significantly as the mapping based on NN 1. Thus, the rest of our results are based on NN 1.

The failure to achieve work distribution overlap occurs because mapping does not reproduce the targeted potential energy distribution (Fig. 9). Overall, the mapped structures have a higher potential energy than the target distribution. For the forward mapping, the potential energy distribution is shifted higher and broadened. There is significant overlap with the target's potential energy distribution. For the reverse mapping, there is more pronounced shifting and broadening, and the high-energy tail of the distribution is heavier. These suggest that the forward mapping is more successful than the reverse mapping.

From a structural perspective, the failure of reverse mapping is related to the formation of hydrogen bonds. While state A is characterized by a stable α-helix, state B is comparatively unfolded. Near the ends of the helix, structures mapped from state B to state A are similar to structures from state A (Fig. 10). However, the middle of the helix is compressed and lacks hydrogen bonds. The
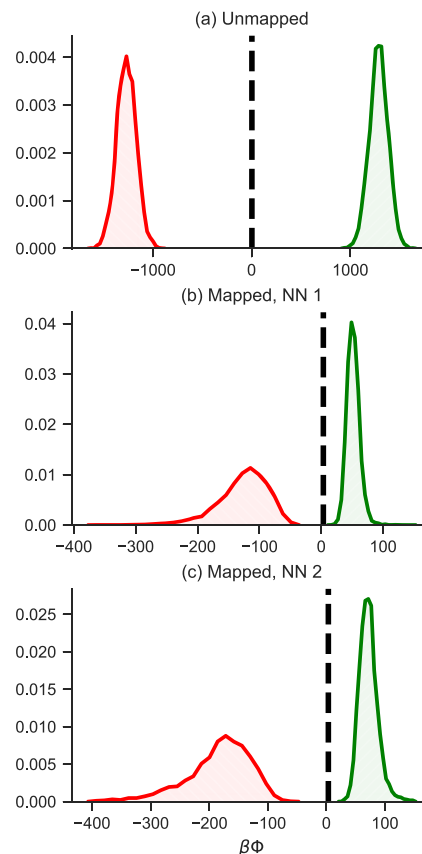

FIG. 8. Normalized histograms of work (a) without and (b and c) with mapping between $\lambda_A = 14$ Å and $\lambda_B = 18$ Å. Mapping was performed with the first (b) and second (c) neural network architectures. $\beta\Phi_F$ are colored green with forward diagonal hatches, and $-\beta\Phi_R$ are colored red with backward diagonal hatches. In (b) and (c), the histogram range was limited to between −400 and 150.

mapped states are in a metastable state, as opposed to the most stable structure of the peptide.

The shortcomings of our more distant mappings suggest that improved neural network architectures and training procedures may be required to map between distant thermodynamic states across rugged potential energy surfaces. Reviewers suggested that mapping could be improved by a more expressive neural network that includes additional coupling layers, e.g., multiple repetitions of Fig. 2(b). Another research area is the development of improved training algorithms. A possible approach to better training of mappings may be to borrow ideas from sampling. Sbailò *et al.*[29] demonstrated that learned mappings may be used to propose Monte Carlo moves between metastable states of a dimer. Mappings capable of generating reasonable structures in the target distribution will probably work well for TFEP.

As mentioned in the introduction, an alternative to mapping between distant molecular distributions is to map each molecular distribution to a comparable tractable distribution. Using this strategy, Ding and Zhang[23] were able to reproduce free energy differences between deca-alanine at very different temperatures, 300 and
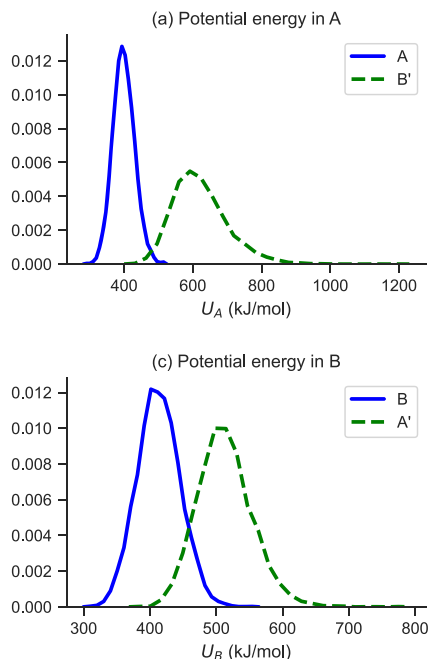
FIG. 9. Normalized histograms of potential energies U for (a) $\lambda_A = 14$ and (b) $\lambda_B = 18$ Å, based on samples from the distribution (line) or mapped samples (dashed line).
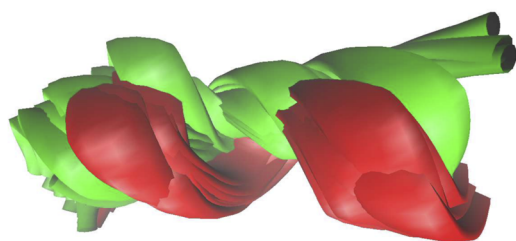

FIG. 10. Comparison of the targeted structures (red color) with the mapped structures (green color) in the reverse process from $\lambda_B = 18$ Å to $\lambda_A = 14$ Å.

500 K. Subsequently, Ding and Zhang[24] applied the same strategy to obtaining binding free energies for a host–guest complex. Reference calculations were performed with an attach–pull–release procedure with multiple intermediate states, suggesting that the end states occupy quite different configuration spaces. However, it is unclear whether their methods will work well on larger and more flexible protein-ligand systems. Moreover, as mentioned in the introduction, mapping between molecular distributions may be the only way to map subsets of coordinates.

V. CONCLUSIONS

We have successfully used learned mappings in TFEP to reproduce reference free energy differences between conformations of a flexible bonded molecule with a rugged free energy surface, deca-alanine. The learned mappings were most successful when the unmapped states were similar. On the other hand, our procedure was unable to learn a mapping from the unfolded to the alpha helical structure of the model peptide.

SUPPLEMENTARY MATERIAL

Two figures. Fig S1 shows the effect of stopping criterion on estimated probability densities of $\beta\Phi$ for test data between $\lambda_A = 20$ Å and $\lambda_B = 21$ Å. Fig S2 shows estimated probability densities of potential energies and terminal nitrogen positions for simulated and mapped samples for $\lambda_A = 20$ Å and $\lambda_B = 21$ Å.

ACKNOWLEDGMENTS

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Soohaeng Yoo Willow: Conceptualization (supporting); Investigation (lead); Methodology (lead); Software (lead); Visualization (lead); Writing – original draft (lead). **Lulu Kang**: Conceptualization (supporting); Methodology (supporting); Project administration (supporting); Supervision (supporting); Writing – review & editing (supporting). **David D. L. Minh**: Conceptualization (lead); Funding acquisition (lead); Methodology (equal); Project administration (lead); Supervision (lead); Visualization (supporting); Writing – review & editing (lead).

DATA AVAILABILITY

Code to perform mapped free energy calculations and sample data are available at https://github.com/swillow/jax_amber.

REFERENCES

[1] K. Tolborg, J. Klarbring, A. M. Ganose, and A. Walsh, "Free energy predictions for crystal stability and synthesis ability," Digital Discovery **1**, 586 (2022).

[2] K. A. Armacost, S. Riniker, and Z. Cournia, "Novel directions in free energy methods and applications," J. Chem. Inf. Model. **60**, 1 (2020).

[3] A. Pohorille, C. Jarzynski, and C. Chipot, "Good practices in free-energy calculations," J. Phys. Chem. B **114**, 10235 (2010).

[4] D. A. Pearlman and P. A. Kollman, "The lag between the Hamiltonian and the system configuration in free energy perturbation calculations," J. Chem. Phys. **91**, 7831 (1989).

[5] D. Wu and D. A. Kofke, "Phase-space overlap measures. II. Design and implementation of staging methods for free-energy calculations," J. Chem. Phys. **123**, 084109 (2005).

[6] D. Wu and D. A. Kofke, "Phase-space overlap measures. I. Fail-safe bias detection in free energies calculated by molecular simulation," J. Chem. Phys. **123**, 054103 (2005).

[7] T.-S. Lee, B. K. Allen, T. J. Giese, Z. Guo, P. Li, C. Lin, T. D. McGee, D. A. Pearlman, B. K. Radak, Y. Tao, H.-C. Tsai, H. Xu, W. Sherman, and D. M. York, "Alchemical binding free energy calculations in AMBER20: Advances and best practices for drug discovery," J. Chem. Inf. Model. **60**, 5595 (2020).

[8] C. Jarzynski, "Targeted free energy perturbation," Phys. Rev. E **65**, 046122 (2002).

[9] R. Zwanzig, "High-temperature equation of state by a perturbation method. I. Nonpolar gases," J. Chem. Phys. **22**, 1420 (1954).

[10] A. F. Voter, "A Monte Carlo method for determining free-energy differences and transition state theory rate constants," J. Chem. Phys. **82**, 1890 (1985).

[11] D. L. Severance, J. W. Essex, and W. L. Jorgensen, "Generalized alteration of structure and parameters: A new method for free-energy perturbations in systems containing flexible degrees of freedom," J. Comput. Chem. **16**, 311 (1995).

[12] C. Jarzynski, "Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach," Phys. Rev. E **56**, 5018 (1997).

[13] C. Jarzynski, "Nonequilibrium equality for free energy differences," Phys. Rev. Lett. **78**, 2690 (1997).

[14] H. Oberhofer, C. Dellago, and P. L. Geissler, "Biased sampling of nonequilibrium trajectories: Can fast switching simulations outperform conventional free energy calculation methods?," J. Phys. Chem. B **109**, 6902 (2005).

[15] M. A. Miller and W. P. Reinhardt, "Efficient free energy calculations by variationally optimized metric scaling: Concepts and applications to the volume dependence of cluster free energies and to solid–solid phase transitions," J. Chem. Phys. **113**, 7035 (2000).

[16] A. M. Hahn and H. Then, "Using bijective maps to improve free-energy estimates," Phys. Rev. E **79**, 011113 (2009).

[17] X.-L. Meng and S. Schilling, "Warp bridge sampling," J. Comput. Graphical Stat.," **11**, 552 (2002).

[18] H. Paliwal and M. R. Shirts, "Multistate reweighting and configuration mapping together accelerate the efficiency of thermodynamic calculations as a function of molecular geometry by orders of magnitude," J. Chem. Phys. **138**, 154108 (2013).

[19] T. B. Tan, A. J. Schultz, and D. A. Kofke, "Efficient calculation of temperature dependence of solid-phase free energies by overlap sampling coupled with harmonically targeted perturbation," J. Chem. Phys. **133**, 134104 (2010).

[20] S. G. Moustafa, A. J. Schultz, and D. A. Kofke, "Very fast averaging of thermal properties of crystals by molecular simulation," Phys. Rev. E **92**, 043303 (2015).

[21] P. Wirnsberger, A. J. Ballard, G. Papamakarios, S. Abercrombie, S. Racanière, A. Pritzel, D. Jimenez Rezende, and C. Blundell, "Targeted free energy estimation via learned mappings," J. Chem. Phys. **153**, 144112 (2020).

[22] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," Nature **596**, 583 (2021).

[23] X. Ding and B. Zhang, "Computing absolute free energy with deep generative models," J. Phys. Chem. B **124**, 10166 (2020).

[24] X. Ding and B. Zhang, "DeepBAR: A fast and exact method for binding free energy computation," J. Phys. Chem. Lett. **12**, 2509 (2021).

[25] A. Rizzi, P. Carloni, and M. Parrinello, "Targeted free energy perturbation revisited: Accurate free energies from mapped reference potentials," J. Phys. Chem. Lett. **12**, 9449 (2021).

[26] P. Wirnsberger, G. Papamakarios, B. Ibarz, S. Racanière, A. J. Ballard, A. Pritzel, and C. Blundell, "Normalizing flows for atomic solids," Mach. Learn.: Sci. Technol. **3**, 025009 (2022).

[27] F. Noé, S. Olsson, J. Köhler, and H. Wu, "Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning," Science **365**, eaaw1147 (2019).

[28] J. P. Nilmeier, G. E. Crooks, D. D. L. Minh, and J. D. Chodera, "PNAS Plus: Nonequilibrium candidate Monte Carlo is an efficient tool for equilibrium simulation," Proc. Natl. Acad. Sci. U. S. A. **108**, E1009 (2011); arXiv:1105.2278.

[29] L. Sbailò, M. Dibak, and F. Noé, "Neural mode jump Monte Carlo," J. Chem. Phys. **154**, 074101 (2021).

[30] M. Invernizzi, A. Krämer, C. Clementi, and F. Noé, "Skipping the replica exchange ladder with normalizing flows," J. Phys. Chem. Lett. **13**, 11643 (2022).

[31] A. H. Mahmoud, M. Masters, S. J. Lee, and M. A. Lill, "Accurate sampling of macromolecular conformations using adaptive deep learning and coarse-grained representation," J. Chem. Inf. Model. **62**, 1602 (2022).

[32] A. van der Vaart, P. B. Orndorff, and S. T. Le Phan, "Calculation of conformational free energies with the focused confinement method," J. Chem. Theory Comput. **15**, 6760 (2019).

[33] P. B. Orndorff, S. T. Le Phan, K. H. Li, and A. van der Vaart, "Conformational free-energy differences of large solvated systems with the focused confinement method," J. Chem. Theory Comput. **16**, 5163 (2020).

[34] D. M. Zuckerman and T. B. Woolf, "Systematic finite-sampling inaccuracy in free energy differences and other nonlinear quantities," J. Stat. Phys. **114**, 1303 (2004).

[35] C. H. Bennett, "Efficient estimation of free energy differences from Monte Carlo data," J. Comput. Phys. **22**, 245 (1976).

[36] M. R. Shirts and J. D. Chodera, "Statistically optimal analysis of samples from multiple equilibrium states," J. Chem. Phys. **129**, 124105 (2008).

[37] S. Kullback, *Information Theory and Statistics* (John Wiley, 1959).

[38] T. H. Nguyen and D. D. L. Minh, "Intermediate thermodynamic states contribute equally to free energy convergence: A demonstration with replica exchange," J. Chem. Theory Comput. **12**, 2154 (2016).

[39] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling, "ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB," J. Chem. Thory Comput. **11**, 3696 (2015).

[40] P. Eastman, J. Swails, J. D. Chodera, R. T. McGibbon, Y. Zhao, K. A. Beauchamp, L.-P. Wang, A. C. Simmonett, M. P. Harrigan, C. D. Stern, R. P. Wiewiora, B. R. Brooks, and V. S. Pande, "OpenMM 7: Rapid development of high performance algorithms for molecular dynamics," PLoS Comput. Biol. **13**, e1005659 (2017).

[41] L. Dinh, J. Sohl-Dickstein, and S. Bengio, "Density estimation using Real NVP," arXiv:1605.08803 [cs.LG] (2017).

[42] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang (2018). "JAX: Composable transformations of Python + NumPy programs," Github, V. 0.3.13, Software. http://github.com/google/jax

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 [cs] (2017).

15 July 2024 21:28:23