Computing Absolute Free Energy with Deep Generative Models

Xinqiang Ding and Bin Zhang*
Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

Fast and accurate evaluation of free energy has broad applications from drug design to material engineering. Computing the absolute free energy is of particular interest since it allows the assessment of the relative stability between states without the use of intermediates. In this letter, we introduce a general framework for calculating the absolute free energy of a state. A key step of the calculation is the definition of a reference state with tractable deep generative models using locally sampled configurations. The absolute free energy of this reference state is zero by design. The free energy for the state of interest can then be determined as the difference from the reference. We applied this approach to both discrete and continuous systems and demonstrated its effectiveness. It was found that the Bennett acceptance ratio method provides more accurate and efficient free energy estimations than approximate expressions based on work. We anticipate the method presented here to be a valuable strategy for computing free energy differences.

Free energy is of central importance in both statistical physics and computational chemistry. It has important applications in rational drug design [1] and material property prediction [2]. Therefore, methodology development for efficient free energy calculations has attracted great research interest [3–13]. Many existing algorithms have focused on estimating free energy differences between states and originate from the free energy perturbation (FEP) identity [14]

$$\mathbb{E}_A[e^{-\beta\Delta U}] = e^{-\beta\Delta F}. \tag{1}$$

Here, $\Delta F = F_B - F_A$ is the free energy difference between the two equilibrium states A and B at temperature T and $\beta = 1/k_BT$. $U_A(\boldsymbol{x})$ and $U_B(\boldsymbol{x})$ are the potential energies for a configuration $\boldsymbol{x}$ in state A and B, respectively, and $\Delta U(\boldsymbol{x}) = U_B(\boldsymbol{x}) - U_A(\boldsymbol{x})$. $\mathbb{E}_A$ represents the expectation with respect to the Boltzmann distribution of $\boldsymbol{x}$ in state A,

$$p_A(\boldsymbol{x}) = \frac{e^{-\beta U_A(\boldsymbol{x})}}{Z_A}, \tag{2}$$

where the normalization constant $Z_A = \int e^{-\beta U_A(\boldsymbol{x})} d\boldsymbol{x}$. Computing ΔF with the FEP identity (Eq. 1) only uses samples from state A. It is more efficient to use samples from both states to compute ΔF by solving the Bennett acceptance ratio (BAR) equation [15]

$$\begin{aligned}&\sum_{k=1}^{N_A} f(\beta[\Delta U(\boldsymbol{x}_k^A) - M - \Delta F]) \\ =&\sum_{k=1}^{N_B} f(-\beta[\Delta U(\boldsymbol{x}_k^B) - M - \Delta F])),\end{aligned} \tag{3}$$

where $f(t) = 1/(1+e^t)$ and $M = \ln(N_B/N_A)$. Here, $\{\boldsymbol{x}_k^A, k = 1, ..., N_A\}$ and $\{\boldsymbol{x}_k^B, k = 1, ..., N_B\}$ are samples from the two states. Both the FEP and the BAR method converge poorly when the overlap in the configuration space between state A and B is small. In that case, multiple intermediate states along a path with incremental changes in the configuration space can be introduced to bridge the two states [3]. However, sampling from multiple intermediate states greatly increases the computational cost. It is, therefore, useful to develop techniques that can alleviate the convergence issue without the use of intermediate states [12, 13].

The requirement on a significant overlap between the configuration space of the two states can be circumvented if we compute their free energy difference from the absolute free energy as $\Delta F = F_B - F_A$. The absolute free energy of a state A/B can be obtained from its difference from a reference state A°/B° as $F_{A/B} = F_{A^\circ/B^\circ} - \Delta F_{A/B\to A^\circ/B^\circ}$. For this strategy to be efficient, however, the reference states must bear significant overlap with the states of interest, and their absolute free energy should be available with minimal computational effort. For most systems, designing reference states that satisfy these constraints can be challenging and requires expertise and physical intuition [16–20]. In this letter, we demonstrate that reference states can be constructed with tractable generative models for efficient computation of the absolute free energy [21, 22].

The workflow for calculating the absolute free energy is as follows. State A is used as an example for the discussion, but the same procedure applies to state B. We first draw samples, $\{\boldsymbol{x}_k^A, k = 1, ..., N_A\}$, from the Boltzmann distribution $p_A(\boldsymbol{x})$. We then learn a tractable generative model, $q_\theta(\boldsymbol{x})$, that maximizes the likelihood of observing these samples by fine-tuning the set of parameters θ. Here tractable generative models refer to probabilistic models that have the following two properties: (i) the normalized probability (or probability density), $q_\theta(\boldsymbol{x})$, can be directly evaluated for a given configuration $\boldsymbol{x}$ without the need of sampling or integration; (ii) independent configurations can be efficiently sampled from the probability distribution. Examples of tractable generative models include the neural autoregressive models [21] and the normalizing flow models [22, 23], both of which are used in the following examples. The generative model defines a new equilibrium state A°, which serves as an excellent reference to state A. Because it

is parameterized from samples of state A, most probable configurations from A° should resemble those from A by design, and the overlap between the two states is guaranteed. In addition, since $q_\theta(\boldsymbol{x})$ is normalized, the partition function Z_{A° of state A° is equal to 1, i.e.,

$$Z_{A^\circ} = \int q_\theta(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1. \tag{4}$$

The absolute free energy of the reference state A° is $F_{A^\circ} = -(1/\beta) \ln Z_{A^\circ} = 0$. With the reference state defined, the absolute free energy for state A can be determined by solving a similar BAR equation as Eq. 3. Our use of tractable generative models ensures that sample configurations can be easily produced for the reference state to be combined with those from state A for solving the BAR equation.

We note that a closely related algorithm for computing the absolute free energy has been introduced in variational methods [24–27]. In these prior studies, $q_\theta(\boldsymbol{x})$ was optimized by minimizing the Kullback-Leibler (KL) divergence [28] from $q_\theta(\boldsymbol{x})$ to $p_A(\boldsymbol{x})$

$$\begin{aligned} D_{\mathrm{KL}}(q_\theta||p_A) &= \int q_\theta(\boldsymbol{x}) \ln \frac{q_\theta(\boldsymbol{x})}{p_A(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= \beta(\langle W_{A^\circ \to A} \rangle - F_A), \end{aligned} \tag{5}$$

where $\langle W_{A^\circ \to A} \rangle = \mathbb{E}_{A^\circ}[U_A(\boldsymbol{x}) - U_{A^\circ}(\boldsymbol{x})]$. We define the potential energy of a configuration $\boldsymbol{x}$ in state A° as

$$U_{A^\circ}(\boldsymbol{x}) = -\frac{1}{\beta} \ln q_\theta(\boldsymbol{x}). \tag{6}$$

Because $D_{\mathrm{KL}}(q_\theta||p_A)$ is non-negative, $\langle W_{A^\circ \to A} \rangle$ is an upper bound of F_A. As $D_{\mathrm{KL}}(q_\theta||p_A)$ decreases along the optimization, $\langle W_{A^\circ \to A} \rangle$ is expected to approach closer to the true free energy and was used for its estimation.

Our methodology is different from the variational methods in two aspects. Firstly, instead of $D_{\mathrm{KL}}(q_\theta||p_A)$, we used

$$\begin{aligned} D_{\mathrm{KL}}(p_A||q_\theta) &= \int p_A(\boldsymbol{x}) \ln \frac{p_A(\boldsymbol{x})}{q_\theta(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} \\ &= \beta(\langle W_{A \to A^\circ} \rangle + F_A) \end{aligned} \tag{7}$$

as the objective function for learning $q_\theta(\boldsymbol{x})$. $\langle W_{A \to A^\circ} \rangle = \mathbb{E}_A[U_{A^\circ}(\boldsymbol{x}) - U_A(\boldsymbol{x})]$. We note that minimizing the KL divergence from $p_A(\boldsymbol{x})$ to $q_\theta(\boldsymbol{x})$ is equivalent to learning the generative model by maximizing its likelihood on the training data. Moreover, because $D_{\mathrm{KL}}(p_A||q_\theta)$ is also non-negative, $\langle -W_{A \to A^\circ} \rangle$ is a lower bound of F_A. Therefore, minimizing $D_{\mathrm{KL}}(p_A||q_\theta)$ is equivalent to maximizing the lower bound $\langle -W_{A \to A^\circ} \rangle$. At the face value, it may seem that $D_{\mathrm{KL}}(q_\theta||p_A)$ is a better objective function than $D_{\mathrm{KL}}(p_A||q_\theta)$ for model training since its optimization only requires samples from $q_\theta(\boldsymbol{x})$. As aforementioned, sampling from $q_\theta(\boldsymbol{x})$ can be made computationally efficient by the use of tractable generative models. On the other hand, training by $D_{\mathrm{KL}}(p_A||q_\theta)$ requires samples from $p_A(\boldsymbol{x})$, the collection of which often requires costly long timescale simulations with Monte Carlo or molecular dynamics techniques. The caveat is that, unlike $D_{\mathrm{KL}}(p_A||q_\theta)$, $D_{\mathrm{KL}}(q_\theta||p_A)$ is not a convex function of $q_\theta(\boldsymbol{x})$. Therefore, the optimization with $D_{\mathrm{KL}}(q_\theta||p_A)$ is more susceptible to traps from local minima. Converging to the global minimum can be challenging unless simplifications, such as mean field approximations [24], are introduced to factorize the probability distribution $q_\theta(\boldsymbol{x})$. When $p_A(\boldsymbol{x})$ is a high dimensional distribution and the system exhibits multistability, optimizing $D_{\mathrm{KL}}(q_\theta||p_A)$ often leads to solutions that cover only one of the metastable states [26].

Another important difference between our methodology and variational methods is the expression used for estimating F_A. In particular, $\langle W_{A^\circ \to A} \rangle$ is an upper bound of the free energy and only becomes exact when the probability distributions from generative models and the state of interest are identical. On the other hand, our use of the BAR equation (Eq. 3) relaxes this requirement, and F_A can be accurately determined even if the model training is not perfect and there are significant differences between the two distributions. In all but trivial examples, we anticipate the learning process to not converge to the true distribution $p_A(\boldsymbol{x})$ due to its high dimensionality and complexity. The BAR estimation will be crucial to ensure the accuracy of free energy calculations, because the estimate from BAR is asymptotically unbiased and has the lowest variance among asymptotically unbiased estimators [29].

In the following, we apply the introduced methodology to two systems that are of significant interest in statistical physics and computational chemistry. As a benchmark, we first computed the absolute free energy of a 20-spin Sherrington-Kerkpatrick (SK) model [30], the value of which can be determined from complete enumeration as well. The discrete configurations of the SK model will be represented using $\boldsymbol{s}$ instead of $\boldsymbol{x}$. Though we introduced the methodology with continuous variables, all the equations can be trivially extended to $\boldsymbol{s}$ by replacing the integrals with summations over the spin configurations. The potential energy of a configuration $\boldsymbol{s} = (s_1, s_2, ..., s_N)$ is defined as

$$U_A(\boldsymbol{s}) = \frac{1}{\sqrt{N}} \sum_{j>i} J_{ij} s_i s_j, \tag{8}$$

where $s_i \in \{-1, +1\}$ and $N = 20$. J_{ij} were chosen randomly from the standard normal distribution. 5000 samples were drawn from the complete configurational ensemble based on the probability distribution $p(\boldsymbol{s}) = e^{-\beta U_A(\boldsymbol{s})}/Z_A$ with $\beta = 2.0$. These samples were used to train the reference state A° by minimizing $D_{\mathrm{KL}}(p_A||q_\theta)$ (Eq. 7). The reference probability $q_\theta(\boldsymbol{s})$ was defined with a neural autoregressive density estimator (NADE) [21] as

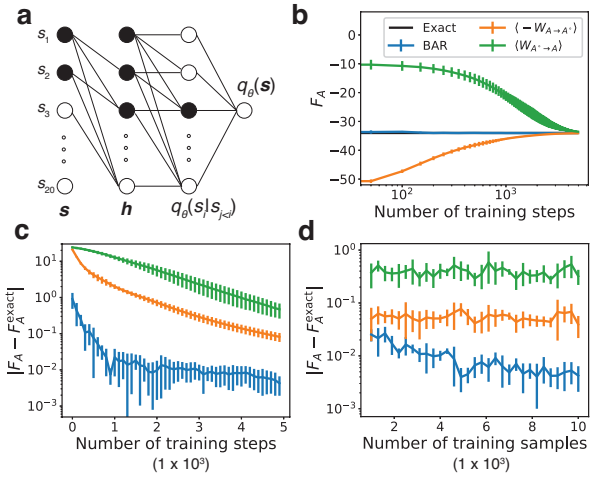

FIG. 1. Performance of different free energy estimators on the Sherrington-Kerkpatrick (SK) model. (a) A schematic representation of the neural autoregressive model used to parameterize q_θ. Any node i in the third layer that represents $q_\theta(s_i|s_{j<i})$ is only connected to nodes (spins) from the input layer ($\boldsymbol{s}$) with indices less than i via the hidden layer ($\boldsymbol{h}$). For example, the 3rd node (black) in the 3rd layer is only connected to the first two input nodes via the three hidden nodes. (b) The absolute free energy of the SK model calculated with three estimators as a function of training steps compared with the exact result obtained from complete enumeration. (c, d) Errors of the estimated absolute free energy versus the number of steps and the number of samples used for training q_θ. The coloring scheme is identical to that in part b.

a product of conditional distributions

$$q_\theta(\boldsymbol{s}) = \prod_{i=1}^{N} q_\theta(s_i|s_1, ..., s_{i-1}). \tag{9}$$

$q_\theta(s_i|s_1, ..., s_{i-1})$ were parameterized using feed forward neural networks with one hidden layer and the neural network's connections are specifically designed such that it maintains the autoregressive property, i.e., $q_\theta(s_i|s_1, ..., s_{i-1})$ only depends on $s_1, ..., s_i$ (Fig. 1a). The autoregressive model is well suited for discrete data and was chosen for its flexibility [31–34]. It is also tractable since, given a configuration $\boldsymbol{s}$, the conditional probabilities can be directly evaluated using the neural network to compute the reference probability $q_\theta(\boldsymbol{s})$ from Eq. 9. In addition, independent samples from $q_\theta(\boldsymbol{s})$ can be drawn with little computational cost via a sequential procedure that samples the configuration for the spin i from $q_\theta(s_i|s_1, ..., s_{i-1})$ after having determined the values for s_1 to s_{i-1}. More details on the implementation of the neural network are provided in the Supplemental Material [35].

After training $q_\theta(\boldsymbol{s})$ for some numbers of steps [36, 37], 5000 configurations were independently drawn from $q_\theta(\boldsymbol{s})$. These configurations, together with the training inputs sampled from $p_A(\boldsymbol{s})$, were used to determine the absolute free energy of the SK model. In Figs. 1b and 1c, we compare results from the three estimators: the lower bound $\langle -W_{A\to A^\circ}\rangle$, the upper bound $\langle W_{A^\circ\to A}\rangle$, and the solution of the BAR equation (Eq. 3), with the exact value.

Consistent with the expressions for the KL divergence (Eqs. 5 and 7), $\langle -W_{A\to A^\circ}\rangle$ and $\langle W_{A^\circ\to A}\rangle$ provide lower and upper bounds of the exact value, respectively. However, at the early stages of model parameterization with small training step numbers, both bounds deviate significantly from the true value. This deviation is expected and is a direct result of the difference between the two probability distributions $q_\theta(\boldsymbol{s})$ and $p_A(\boldsymbol{s})$. As the training proceeds, the agreement between the distributions improves and $\langle -W_{A\to A^\circ}\rangle$ and $\langle W_{A^\circ\to A}\rangle$ gradually converge to the exact result after 5000 steps (Fig. 1b). On the other hand, the BAR estimator converges much faster to the true value with a smaller error (Figs. 1b and 1c). Importantly, even after 100 steps of training, the error of the BAR estimator has reduced to be less than 1.0 k_BT, whereas both the errors of $\langle -W_{A\to A^\circ}\rangle$ and $\langle W_{A^\circ\to A}\rangle$ are large than 10 k_BT. Therefore, the BAR estimator is more efficient for computing the absolute free energy than the bounds, especially when there is a large difference between the generative model and the true distribution. In addition, varying the number of samples used for training $q_\theta(\boldsymbol{s})$ has different effects on the accuracy of converged results for the three approaches (Fig. 1d). For both $\langle -W_{A\to A^\circ}\rangle$ and $\langle W_{A^\circ\to A}\rangle$, increasing the number of training samples from 1×10^3 to 1×10^4 does not significantly change the accuracy of their results. In contrast, using more training samples significantly reduces the error of the BAR estimator. This is because solutions of the BAR equation are asymptotically unbiased for estimating F_A, whereas $\langle -W_{A\to A^\circ}\rangle$ and $\langle W_{A^\circ\to A}\rangle$ are not [29].

Next, we applied the methodology to a molecular system, the alanine dipeptide in the gas phase. This system presents features commonly encountered in biomolecular simulations with continuous phase space over a rugged energy landscape. Its high dimensionality renders a complete enumeration of the configurational space to compute the absolute free energy for benchmarking impractical. Instead, we computed the free energy difference between two metastable states using their absolute free energy to compare against the value determined from umbrella sampling. The two metastable states, A and B, were defined using the backbone dihedral angle ϕ (C-CA-N-CY): $0° < \phi \leq 120°$ for state A and $\phi \leq 0°$ or $\phi > 120°$ for state B (Fig. 2a).

To compute the absolute free energy of the two states, we designed two references using normalizing flow based generative models [23, 31]. Specifically, each $q_\theta(\boldsymbol{x})$ was parameterized with multiple bijective transformations, $T_1, ..., T_K$, to convert a random variable $\boldsymbol{u}$ to a peptide

configuration, i.e.,

$$\boldsymbol{x} = T(\boldsymbol{u}) = T_K \circ \cdots \circ T_1(\boldsymbol{u}). \tag{10}$$

$\boldsymbol{u}$ shares the same dimension as $\boldsymbol{x}$, and we used the standard Gaussian distribution $\mathcal{N}(0, I)$ as its base distribution $p_u(\boldsymbol{u})$. Based on the formula of variable change in probability density functions, we have

$$\ln q_\theta(\boldsymbol{x}) = \ln p_u(\boldsymbol{u}) - \sum_{k=1}^{K} \ln |J_{T_k}(\boldsymbol{u}_k)|, \tag{11}$$

where $\boldsymbol{u}_k = T_k \circ \cdots T_1(\boldsymbol{u})$. J_{T_k} is the Jacobian matrix of the transformation T_k, and $|\cdot|$ denotes the absolute value of the determinant. The generative model defined with normalizing flows is again tractable since for a given configuration $\boldsymbol{x}$, the corresponding probability density $q_\theta(\boldsymbol{x})$ can be evaluated using the right hand side of Eq. 11 with $\boldsymbol{u} = T^{-1}(\boldsymbol{x})$. We can also draw independent peptide configurations by first sampling $\boldsymbol{u}$ from the base distribution $p_u(\boldsymbol{u})$ and then applying the bijective transformations $\boldsymbol{x} = T(\boldsymbol{u})$.

The bijective transformations, $T_1, ..., T_k$, are often designed to be as flexible as possible to ensure the expressive power of $q_\theta(\boldsymbol{x})$. The only constraint enforced here is for the time complexity of calculating the determinant of the Jacobian matrices for both T_k and the inverse transformation to be $\mathcal{O}(D)$, where D is the dimension of $\boldsymbol{u}$. This constraint allows an efficient evaluation of the probability function defined in Eq. 11. Here for the alanine dipeptide, $\boldsymbol{u}$ was first transformed into the internal coordinates $\boldsymbol{z}$, which was then converted to the Cartesian coordinates $\boldsymbol{x}$, i.e., $\boldsymbol{z} = T_{K-1} \circ \cdots T_1(\boldsymbol{u})$ and $\boldsymbol{x} = T_K(\boldsymbol{z})$. The transformation between $\boldsymbol{z}$ and $\boldsymbol{x}$, T_K, is determined by the topology of the alanine dipeptide and fixed. The real-valued non-volume preserving (realNVP) transformations [22], which are one kind of normalizing flows, were used as the bijective transformations between $\boldsymbol{u}$ and $\boldsymbol{z}$. Each realNVP transformation is a neural affine coupling layer that keeps a subset of the input components unchanged and applies an affine transformation to the remaining input components with the slope and intercept terms parameterized by the fixed components. More details about the implementation of the transformations, $T_1, ..., T_K$, are included in the Supplemental Material [35].

The two reference models were separately trained using 2×10^4 configurations collected from each state via molecular dynamics simulations with the CHARMM force field [38]. With the reference states, we computed the absolute free energy for state A and B using the three estimators. As shown in Figs. 2b and 2c, the BAR estimator again converges much faster than the upper and lower bounds. The results calculated using the upper bound are not shown here because they are much larger than that of the lower bound and the BAR estimator (Fig. S4). Unlike the results for the SK model, the two bounds no longer converge to the same value or to the BAR estimator, and their difference can be as large as 25 kcal/mol (Fig. S4). The large gap between the two bounds suggests that the generative models are still quite different from the true distributions $p_A(\boldsymbol{x})$ or $p_B(\boldsymbol{x})$ even after the training of $q_\theta(\boldsymbol{x})$ has converged. We expect the numbers from the BAR estimator to be correct, as it is guaranteed to converge to the exact result and does not require the generative models to precisely reproduce the original distributions. Furthermore, they lie in between the two bounds in both cases (Fig. S4), as expected for the exact values of the absolute free energy. Therefore, for this molecular system, the two bounds cannot be used for reliable estimation of the absolute free energy.

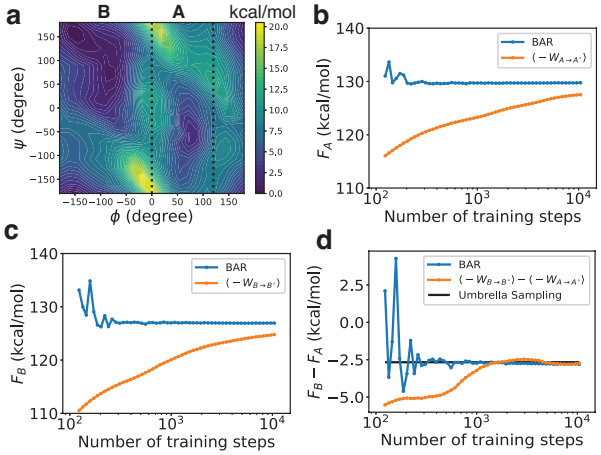

FIG. 2. Performance of different free energy estimators on the alanine dipeptide. (a) The free energy surface of the alanine dipeptide on two collective variables of ϕ and ψ. The two metastable states, A and B, correspond to regions with $0° < \phi \leq 120°$ and $\phi \leq 0°$ or $\phi > 120°$, respectively. (b, c) The absolute free energy of state A and B computed with different estimators. (d) The free energy difference between state A and B computed with different estimators.

We further evaluated the accuracy of the three estimators in computing the free energy differences between state A and B. For comparison, we determined the free energy difference using umbrella sampling [3] as well. Results of estimated free energy differences are shown in Figs. 2d and S4d. Similarly to the results for the absolute free energy, the BAR estimator converges much faster to the result from umbrella sampling than the two bounds. To our surprise, difference between the two lower bounds eventually coincide with the correct result as well. Because the lower bounds converge to biased values for the absolute free energies of state A and state B, we believe its good performance on the free energy difference is due to error cancellation. Therefore, the method introduced here can be used to compute free energy differences between states by calculating the absolute free energy of each state separately. It is computationally more efficient

than umbrella sampling since no sampling from intermediate states are needed.

In summary, we demonstrated that the framework based on deep generative models succeeds at computing the absolute free energy using sample configurations from the state of interest and is applicable for both discrete and continuous systems. Although the two examples tested here are relatively small in size, the free energy calculation method can be readily applied to larger lattice models and more complex biomolecular systems in gas phase or with implicit solvation. It could greatly facilitate the evaluation of protein-ligand binding affinity and protein conformational stability while accounting for entropic contributions. Generalizing the methodology to problems with explicit solvation can be challenging, due to the difficulty in parameterizing the generative models in a configuration space with significantly higher dimensionality. Furthermore, in addition to translational and rotational symmetry, these models must account for the permutation symmetry with respect to individual solvent molecules. It is worth noting that multiple studies [13, 26, 39, 40] have introduced approaches for designing deep generative models that are invariant to permutations. Combining these approaches with our framework to compute the absolute free energy of biomolecular systems with explicit solvation would be an exciting direction for future studies.

This work was supported by the National Institutes of Health (Grant 1R35GM133580-01).

* binz@mit.edu

[1] M. Rami Reddy and M. D. Erion, *Free energy calculations in rational drug design* (Kluwer Academic/Plenum Publishers, 2001) p. 384.

[2] S. Auer and D. Frenkel, Nature **409**, 1020 (2001).

[3] G. M. Torrie and J. P. Valleau, Journal of Computational Physics **23**, 187 (1977).

[4] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, Journal of Computational Chemistry **13**, 1011 (1992).

[5] W. L. Jorgensen and C. Ravimohan, The Journal of Chemical Physics **83**, 3050 (1985).

[6] M. R. Shirts and J. D. Chodera, Journal of Chemical Physics **129**, 124105 (2008).

[7] E. Schneider, L. Dai, R. Q. Topper, C. Drechsel-Grau, and M. E. Tuckerman, Physical Review Letters **119**, 150601 (2017).

[8] A. Pohorille, C. Jarzynski, and C. Chipot, Journal of Physical Chemistry B (2010), 10.1021/jp102971x.

[9] P. V. Klimovich, M. R. Shirts, and D. L. Mobley, Journal of Computer-Aided Molecular Design (2015), 10.1007/s10822-015-9840-9.

[10] P. Kollman, Chemical Reviews (1993), 10.1021/cr00023a004.

[11] A. M. Hahn and H. Then, Physical Review E - Statistical, Nonlinear, and Soft Matter Physics **79**, 011113 (2009).

[12] C. Jarzynski, Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics **65**, 5 (2002).

[13] P. Wirnsberger, A. J. Ballard, G. Papamakarios, S. Abercrombie, S. Racanière, A. Pritzel, D. J. Rezende, and C. Blundell, (2020), arXiv:2002.04913.

[14] R. W. Zwanzig, Journal of Chemical Physics **22**, 1420 (1954).

[15] C. H. Bennett, Journal of Computational Physics **22**, 245 (1976).

[16] W. G. Hoover, S. G. Gray, and K. W. Johnson, The Journal of Chemical Physics **55**, 1128 (1971).

[17] D. Frenkel and A. J. Ladd, The Journal of Chemical Physics **81**, 3188 (1984).

[18] W. G. Hoover and F. H. Ree, The Journal of Chemical Physics **47**, 4873 (1967).

[19] L. M. Amon and W. P. Reinhardt, Journal of Chemical Physics **113**, 3573 (2000).

[20] F. M. Ytreberg and D. M. Zuckerman, Journal of Chemical Physics **124**, 104105 (2006).

[21] B. Uria, M.-A. Côté, K. Gregor, I. Murray, and H. Larochelle, Journal of Machine Learning Research **17**, 1 (2016).

[22] L. Dinh, J. Sohl-Dickstein, and S. Bengio, (2016), arXiv:1605.08803.

[23] D. J. Rezende and S. Mohamed, 32nd International Conference on Machine Learning, ICML 2015 **2**, 1530 (2015).

[24] M. Opper and D. Saad, *Advanced mean field methods : theory and practice* (MIT Press, 2001) p. 273.

[25] D. Wu, L. Wang, and P. Zhang, Physical Review Letters **122**, 080602 (2019).

[26] F. Noé, S. Olsson, J. Köhler, and H. Wu, Science **365**, eaaw1147 (2019).

[27] S. H. Li and L. Wang, Physical Review Letters (2018), 10.1103/PhysRevLett.121.260601.

[28] S. Kullback and R. A. Leibler, The Annals of Mathematical Statistics **22**, 79 (1951).

[29] M. R. Shirts, E. Bair, G. Hooker, and V. S. Pande, Physical Review Letters **91**, 140601 (2003).

[30] D. Sherrington and S. Kirkpatrick, Physical Review Letters **35**, 1792 (1975).

[31] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, (2019), arXiv:1912.02762.

[32] D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling, in *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Curran Associates, Inc., 2016) pp. 4743–4751.

[33] G. Papamakarios, T. Pavlakou, and I. Murray, in *Advances in Neural Information Processing Systems* (2017).

[34] C. W. Huang, D. Krueger, A. Lacoste, and A. Courville, in *35th International Conference on Machine Learning, ICML 2018* (2018).

[35] "See Supplemental Material at [URL will be inserted by publisher] for detailed decription on the autoregressive density estimator and normalizing flow models and free energy results on alanine dipeptide.".

[36] D. P. Kingma and J. L. Ba, in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).

[37] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. D. Facebook, A. I. Research, Z. Lin, A. Desmaison, L. Antiga, O. Srl, and A. Lerer, in *Advances in Neural*

Information Processing Systems 32 (2019).

[38] B. R. Brooks, C. L. Brooks III, A. D. Mackerell Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, and others, Journal of computational chemistry **30**, 1545 (2009).

[39] J. Köhler, L. Klein, and F. Noé, (2019), arXiv:1910.00753.

[40] C. M. Bender, J. J. Garcia, K. O'Connor, and J. Oliva, (2019), arXiv:1902.01967.