THE JOURNAL OF PHYSICAL CHEMISTRY B

pubs.acs.org/JPCB Article

# Discovering Protein Conformational Flexibility through Artificial-Intelligence-Aided Molecular Dynamics

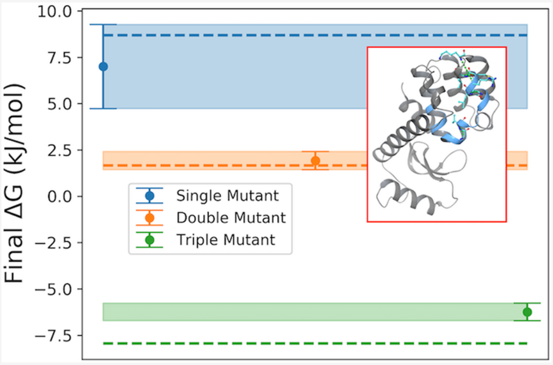
*Published as part of The Journal of Physical Chemistry virtual special issue "Machine Learning in Physical Chemistry".*

Zachary Smith, Pavan Ravindra,[§] Yihang Wang,[§] Rory Cooley, and Pratyush Tiwary*

Cite This: *J. Phys. Chem. B* 2020, 124, 8221−8229 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** Proteins sample a variety of conformations distinct from their crystal structure. These structures, their propensities, and the pathways for moving between them contain an enormous amount of information about protein function that is hidden from a purely structural perspective. Molecular dynamics simulations can uncover these alternative conformations but often at a prohibitively high computational cost. Here we apply our recent statistical mechanics and artificial intelligence-based molecular dynamics framework for enhanced sampling of protein loops. We exemplify the approach through the study of three mutants of the classical test-piece protein T4 lysozyme. We are able to correctly rank these according to the stability of their excited state. By analyzing reaction coordinates, we also obtain crucial insight into why these specific perturbations in sequence space lead to tremendous variations in conformational flexibility. Our framework thus allows an accurate comparison of loop conformation populations with minimal prior human bias and should be directly applicable to a range of macromolecules in biology, chemistry, and beyond.



## ■ INTRODUCTION

Understanding and predicting the relationship among protein sequence, structure, and function has been a long-standing dream in biophysics. One of the many reasons that this is an especially difficult problem is that often a given sequence does not imply one fixed structure. On one hand, the so-called "folding problem" can be considered to be solved through either experiments[1−4] or even artificial intelligence (AI).[5] On the other hand, while it is useful to consider the single most stable protein crystal structure for a given sequence determined through experiments and/or theory, the typical model of proteins has shifted from static objects to fluctuating polymers.[6−8] Considering that the fluctuations of proteins among various structures, even rare ones, have been shown to increase our understanding of the mechanisms underlying protein function,[9−11] these fluctuations are hard to quantify through even state-of-the art experiments.[11−13] One particularly interesting class of conformational rearrangements is the movement of surface-exposed loops which exhibit much greater flexibility than the rest of the protein. This class of problems is relevant both from the perspective of a fundamental understanding of the chemistry of life processes[14,15] as well as discovering druggable targets for inhibiting different diseases that offer both specific and potent action.[16,17]

A particularly well characterized yet puzzling system for studying loop movement as well as the related problem of ligand recognition is the L99A mutant of the protein T4 lysozyme, in which a leucine residue at position 99 is replaced with alanine, opening up a pocket for binding small hydrophobic ligands such as benzene.[18] This L99A mutated protein populates two well-defined states: the ground state (GS) and an excited state (ES), with populations of roughly 97 and 3%, respectively, at 25 °C.[19] The excited state is characterized by the rotation of phenylalanine residue F114 into the binding pocket and the unification of two adjacent helices.[20] Additionally, benzene egress appears to take place during the transition from the ground state to the excited state, as F114 moves into a buried position occupying the cavity previously filled by benzene.[21,22]

**Received:** May 4, 2020
**Revised:** August 21, 2020
**Published:** August 25, 2020

Previous computational studies have attempted to detail underlying mechanisms of this transition, but many aspects are still not fully understood.[23,24] For instance, in addition to the L99A mutant, two other point mutants of the T4 lysozyme have been discovered which very significantly alter the population of the ground and excited states.[20] These two mutants [L99A and G113A (double mutant) and L99A, G113A, and R119P (triple mutant)] have 66:34 and 4:96 ground state/excited state population ratios as opposed to the 97:3 for single mutant L99A.[20] Detailed atomic-level insight into why and how such point mutations (Figure 1) so drastically affect this protein's conformational landscape is still missing. While computational and experimental methods have clarified such structural fluctuations and ligand binding pathways in L99A T4L in recent years, this system still presents a challenge due to the high degree of flexibility and size of the surface-exposed loop near the binding pocket.[23,25] This set of three mutants thus serves as an excellent benchmarking set for methods that seek to sample loop conformations including this current work.

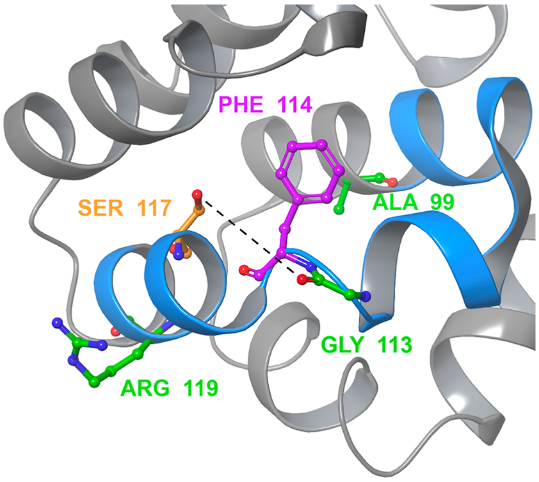


**Figure 1.** Crystal structure of the single mutant L99A with the cavity-adjacent loop (residues 100−120) shown in blue. The mutation sites for the three mutants are shown in green. The phenylalanine residue that blocks the binding pocket in the excited state is shown in purple. The distance between serine 117 (orange) and glycine 113 (green) that represents the formation of the hydrogen bond connecting the phenylalanine-adjacent helices is shown as a dashed line.

Our aim in this work is to demonstrate how our recent enhanced molecular dynamics (MD) simulation algorithms grounded in statistical mechanics and AI[26−30] can be used to obtain such insight in this classic test-piece system, thereby opening the possibility to answer such questions in the future in generic proteins with related unanswered questions of practical and fundamental relevance.[31,32] In principle, MD simulations can give such insights into the thermodynamics and biophysical mechanisms behind these protein conformational changes by providing data at all-atom and femtosecond resolution, which can be expensive to obtain in experiments.[33] However, MD simulations have a crippling time scale problem: reaching even the millisecond threshold on even the most specialized supercomputers is very difficult due to the sheer number of interactions that needs to be computed per time step and the short time scale of bond vibrations restricting each time step to 1 or 2 fs at best.[34] Often, these metastable protein conformations are separated by high-energy barriers that make exchanges between these conformations much slower than hundreds of microseconds and therefore practically inaccessible to classical MD simulations.[28]

To overcome this time scale problem, enhanced sampling methods have been designed to accelerate the sampling of protein configurations.[35−37] One such method is metadynamics.[28,38−41] If given a carefully constructed reaction coordinate (RC) that has sufficient overlap with the relevant slow degrees of freedom, metadynamics can gradually enhance fluctuations along this RC to efficiently move the system between known and unknown metastable states in a controlled, reweightable manner.[28] The major limitation of metadynamics and arguably many other enhanced sampling methods as well is the dependence of its performance on the selection of the RC whose fluctuations are enhanced through biasing. Traditionally, this RC has been chosen by hand using chemical intuition and results from previous studies, but in recent years, methods have been developed which aim to learn RCs with minimal human intuition.[26,27,42−51]

One such AI-inspired method is reweighted autoencoded variational bayes for enhanced sampling (RAVE), which is based on the principle of a predictive information bottleneck (PIB), originally proposed to model neuronal behavior in retinal cells[52] and more recently as an explanation of the generic success of deep learning and AI.[53] In a nutshell, the PIB is the minimally complex yet maximally predictive model for describing the evolution of a given dynamical system.[27,54] RAVE learns such a PIB in an iterative procedure, interpreting it as the RC to be used as a biasing variable in enhanced sampling.[26,27] Here, we construct this RC as a linear combination of selected basis functions or order parameters (OPs). These OPs can number in the hundreds to thousands and are generic features such as protein−ligand or protein−protein contacts. Here we first use our automatic mutual information noise omission (AMINO)[30] method that screens for redundancies among this very large and generic set of OPs, determining, for example, if two protein−ligand contacts carry the same information. The output from AMINO is fed to RAVE,[26,27] which then uses the PIB to construct an RC as a linear combination of the AMINO OPs. The RC is then used as a biasing variable in metadynamics, and the biased metadynamics trajectory itself is fed back to RAVE to further optimize the RC. The iteration between deep learning and sampling continues until multiple transitions between different metastable states are sampled. While this study uses metadynamics, any enhanced sampling method that uses a biasing RC can be used with AMINO and RAVE. Through this automated protocol, we are able to compare the changes in ground- and excited-state populations due to point mutations and gain insights into the underlying mechanisms causing these changes. We are also able to analyze the RC and its constituent basis functions for different sequences and gain mechanistic insight into how small perturbations in sequence space lead to enormous effects in the structure space. We expect that the procedure using AMINO, RAVE, and enhanced sampling illustrated in this work can be generalized to sample a wide array of processes undergone by macromolecules, especially those which have not been studied extensively enough to have standard enhanced sampling parameters.

## METHODS

**AMINO.** In this work, we constructed our RC as a linear combination of a set of basis functions, which we call order parameters (OPs). AMINO starts with a large dictionary of OPs and uses an information theoretic approach to identify redundancies in the full dictionary to return a reduced set of OPs for use in RC construction.[30] It is important to note that the output of AMINO is not an RC but a more efficient set of OPs to be used in RC calculations through a method such as SGOOP or TICA.

AMINO runs a k-medoids clustering procedure on the set of OPs based on the following distance metric:

$$D(X;\,Y) = 1 - \frac{I(X;\,Y)}{H(X,\,Y)} \tag{1}$$

$$= 2 - \frac{\sum_{x\in X}\sum_{y\in Y} P(x,\,y)^{*}\log(P(x)^{*}P(y))}{\sum_{x\in X}\sum_{y\in Y} P(x,\,y)^{*}\log(P(x,\,y))} \tag{2}$$

As input to AMINO for each system, we used a combination of a 50 ns unbiased trajectory beginning in ES and a 50 ns unbiased trajectory beginning in GS to form a combined trajectory of 100 ns. Since AMINO needs only estimates of the stationary probability density, preserving the temporal ordering of data points is not needed, and this kind of mixing is acceptable. From these trajectories, the contact points used to construct the OPs for each system were the following:

1. Category I: The C$\alpha$ of every amino acid in the loop, defined to be amino acids in the range of [110, 120].
2. Category II: Every third C$\alpha$ that is not in the above loop range.

In the second category, in principle we could take every amino acid and not just the third, but the latter helps with computational efficiency. In future versions of AMINO, we hope to make it possible to consider every amino acid in a computationally efficient manner. The input OPs to AMINO were the distances between every pair of contact points $p$, $q$, where $p$ is a contact point from category I, $q$ is a contact point from either category I or II, and $p \neq q$.

From this input data, AMINO runs the aforementioned clustering procedure on these OPs for a range of $k$ values, where $k$ is the number of clusters used in k-medoids clustering. AMINO decides the optimal value of $k$ automatically as discussed in ref 30. This forms the procedure that allows for the automated selection of OPs with minimal prior knowledge of the system.

**RAVE.** RAVE took the output OPs from AMINO and used an information-bottleneck-based protocol to learn the RC, which was then used in metadynamics to enhance the sampling. The sampling from metadynamics was then fed back into RAVE to learn a better RC, which in turn is used to perform a newer round of metadynamics. The iteration between metadynamics and RAVE is continued until sufficient back-and-forth movement between metastable states is obtained during metadynamics. In RAVE, the RC is defined as the predictive information bottleneck which encodes high-dimensional input $X$ as low-dimensional representation $\chi$, which we interpret as the RC. The optimally encoded RC should be the low-dimensional representation that maximally compresses the input yet also has maximal predictive power for the future state of the input, given by $X_{\Delta t}$. This trade-off between maximal compression and maximal prediction can be quantified using two pieces of mutual information, $I(\chi, X)$ and $I(\chi, X_{\Delta t})$ or, more specifically, the difference between these two mutual pieces of information.[27,29] This has been shown to be equivalent to training a neural network with an encoder–decoder structure to optimize the following objective function[55]

$$L = \int P(X_{\Delta t},\,\chi)\,\ln Q(X_{\Delta t}|\chi)\;\mathrm{d}X_{\Delta t}\;\mathrm{d}\chi \tag{3}$$

Here, $\chi$ is a linear combination of OPs determined by the encoder and $Q(X_{\Delta t}|\chi)$ is the probability learned by the deep neural network to approximate the distribution $P(X_{\Delta t}|\chi)$ from data. In this study, we needed to iterate between metadynamics and RAVE, so we needed to reweight the effect of the biasing potential on the static probability as well as the dynamical propagator. The correction introduced in ref 29 is used to give a better estimation of $P(X_{\Delta t}, \chi)$ from biased trajectories.

**Molecular Dynamics.** All molecular dynamics simulations were run using the AMBER99SB force field through the GROMACS package patched with PLUMED version 2.4.2[56,57] with a 2 fs time step. Temperature and pressure were kept at 300 K and 1 bar using the velocity rescale thermostat[58] and Parrinello–Rahman barostat.[59] The nonbonded interactions were calculated with a 10 Å cutoff, and long-range electrostatics were calculated using the particle-mesh Ewald method.[60] The starting structures were PDB https://www.rcsb.org/structure/1L90 for the single mutant, PDB https://www.rcsb.org/structure/3DMV for the double mutant, and PDB https://www.rcsb.org/structure/2LC9 for the triple mutant. The triple mutant was prepared by mutating residue G113 to A113 using MAESTRO on the starting structure.

The AMINO and RAVE inputs were 50 ns of simulation from the starting structure and 50 ns of simulation from the less-stable state (ES for single and double mutants and GS for the triple mutant). The starting structures for the less stable states were set by finding the minimum RMSD between a reference structure for that state and a metadynamics trajectory along a trial inter-residue contact RC that visited both states. This RMSD was calculated by aligning all C$\alpha$ atoms except 100–120 and measuring the displacement of C$\alpha$ 100–120. The excited-state reference structures were PDB https://www.rcsb.org/structure/2LCB for single and double mutants and PDB https://www.rcsb.org/structure/1L90 for the triple mutant. Mixing these two trajectories for each system allowed us to consider the fluctuations in both states when constructing a basis set of OPs and an RC.

**Metadynamics.** The metadynamics simulations were run using the PLUMED implementation of well-tempered metadynamics with a bias factor of 10, an initial hill height of 1.5 kJ/mol, and bias deposited every picosecond.[39,56,57] The sigma value of the Gaussian bias kernel was set to the standard deviation of the biasing RC from 50 ns of unbiased simulation for each system.

Once an initial 15 ns metadynamics run was completed for each system, the biased trajectory was reweighted and used to run a second round of RAVE.[61,62] The system in these 15 ns explored more conformational space than the initial 100 ns of unbiased simulation, which enabled RAVE to determine a more informative RC. The RC from the second round of RAVE was used for the longer production runs of metadynamics.

The production runs also included a quadratic bias for the single and double mutants to prevent the loop's helix from

Table 1. Reaction Coordinate Definitions for Each Mutant[a]

| single mutant distances | $\chi_1$ weight | $\chi_2$ weight | double mutant distances | $\chi_1$ weight | $\chi_2$ weight | triple mutant distances | $\chi_1$ weight | $\chi_2$ weight |
|---|---|---|---|---|---|---|---|---|
| *G110− L118* | −0.249 | 0.895 | *G110−V111* | 0.820 | 1.000 | *V111−A112* | −0.047 | −0.083 |
| *V111−A112* | −0.498 | 0.052 | *A112−A113* | 0.660 | 0.110 | *N116−S117* | −0.248 | −0.103 |
| *A112−G113* | 0.221 | −0.017 | *A113−S117* | 0.307 | 0.326 | *S117−L118* | 0.087 | 0.559 |
| *T115−N116* | −0.197 | −0.075 | *N116−M120* | −0.262 | −0.228 | I27−T115 | 1.000 | −0.547 |
| Y25−T115 | −0.056 | −0.406 | *S117−L118* | −0.715 | −0.430 | F114−T152 | 0.476 | 1.000 |
| G28−G110 | 0.083 | −0.411 | *L118−R119* | −0.651 | −0.298 | | | |
| T34−F114 | −0.409 | −0.006 | K43−V111 | 0.256 | −0.099 | | | |
| T34−L118 | 0.669 | 0.385 | R76−V111 | 0.558 | −0.181 | | | |
| K43−F114 | 0.120 | −0.425 | R119−K124 | −0.133 | 0.435 | | | |
| A49−N116 | 0.062 | −0.162 | M120−L121 | 1.000 | 0.256 | | | |
| F67−N116 | 0.237 | 0.767 | | | | | | |
| K85−A112 | −0.016 | 0.297 | | | | | | |
| A97−T115 | 1.000 | −0.701 | | | | | | |
| T142−R119 | 0.206 | 1.000 | | | | | | |
| R154−G113 | 0.082 | 0.393 | | | | | | |
| T157- N116 | −0.056 | 0.975 | | | | | | |

[a] $\chi_1$ and $\chi_2$ refer to the first and second components of a two-dimensional RC. The order parameters are defined as the distance between two C$\alpha$ atoms. Order parameters using two cavity-adjacent loop residues are shown before other order parameters and are in italics.

breaking as bias accumulated. The quadratic bias (kJ/mol) had the form of eq 4 for the single mutant and eq 5 for the double mutant, where the RMSD (nm) is calculated using the protocol from the AMINO inputs. These metadynamics runs were performed in duplicate with randomized starting velocities for 500 ns for each of the three mutants.

The results for the triple mutant could be improved by adding a quadratic wall similar to those used for the other two mutants. The triple mutant spent a large quantity of time in high-RMSD conformations that were not considered in the GS or ES which can be alleviated by adding such a barrier. We have chosen to leave one mutant without a barrier in this work to clearly show the difference between trajectories with and without a barrier. The use of these RMSD barriers makes the procedure more efficient but less automated than using AMINO and RAVE alone. This can be improved by the use of a metric that does not rely on crystal structures but describes secondary structure such as the $\alpha$ helix and $\beta$ sheet RMSDs developed by Pietrucci and Laio.[63]

$$V(\text{RMSD}) = 1000(\text{RMSD} - 0.5)^4 \text{ (kJ/mol)} \quad (4)$$

$$V(\text{RMSD}) = 10\,000(\text{RMSD} - 0.4)^4 \text{ (kJ/mol)} \quad (5)$$

**State Definitions.** To stay consistent with the literature,[23,24] we defined these states in terms of three variables: (i) the F114 $\Psi$ dihedral angle, (ii) the hydrogen bonding distance between G113 and S117, and (iii) the RMSD deviation over C$\alpha$ atoms from the non-native conformation for each mutant (PDB https://www.rcsb.org/structure/2LCB for single and double mutants and PDB https://www.rcsb.org/structure/1L90 for the triple mutant). The distance and dihedral components of each state definition were set using the equilibrium value of each OP for a specific state and mutant (i.e., equilibrium distance in the ground state for the single mutant). These equilibrium values were calculated using the average reweighted probabilities across the mutant's two 500 ns metadynamics runs. The equilibrium values were set to the local maxima in the probabilities closest to the definitions reported in ref 23, and then the range around these values was set by hand to cover these local maxima without including other probability maxima. The RMSD cutoffs were then set to be the RMSD of each mutant's equilibrated structure +0.1 Å.

## ■ RESULTS

**Summary of Methods.** Here we briefly summarize all of the methods used in this work, with full details given in Methods. Short unbiased MD simulations were performed for each of the three mutants starting from their respective X-ray crystal structures. Starting from a much larger list of different inter-residue contacts, AMINO[30] generated a minimal representative set of OPs for each of the three systems. A first round of RAVE was then applied to the unbiased MD trajectory to learn an RC as a linear combination of the OP output from AMINO. Well-tempered metadynamics[28] simulations were performed using this RC to accelerate the sampling, and the resultant biased trajectory was fed back to RAVE to learn an improved RC. This iterative procedure allows us to explore conformational space much faster than traditional MD while returning RCs in terms of human-interpretable variables such as the distances between C$\alpha$ atoms used in this study. The iteration between RC optimization through RAVE and sampling through well-tempered metadynamics was terminated when the latter achieved multiple back-and-forth transitions between the starting state and other metastable states. Once RC optimization was completed, two longer 500 ns productions runs of metadynamics were performed for each mutant.

**Order Parameter and Reaction Coordinate Analysis.** The minimal OP set output from AMINO and the subsequently optimized RCs are shown for the three different mutants in Table 1. Two comments are critical here. First, as can be seen in Table 1, the dimensionality of the minimal OP set differed for the three mutants. Second, the RCs were all two-dimensional. We first attempted to perform the protocol of this work with a one-dimensional RC, but subsequent rounds of metadynamics and RAVE were found to show no improvement in sampling. As such, we used a two-dimensional RC wherein sampling improved with subsequent training rounds. This RC was composed of components $\chi_1$ and $\chi_2$ which themselves were expressed as linear combinations of the input OPs.

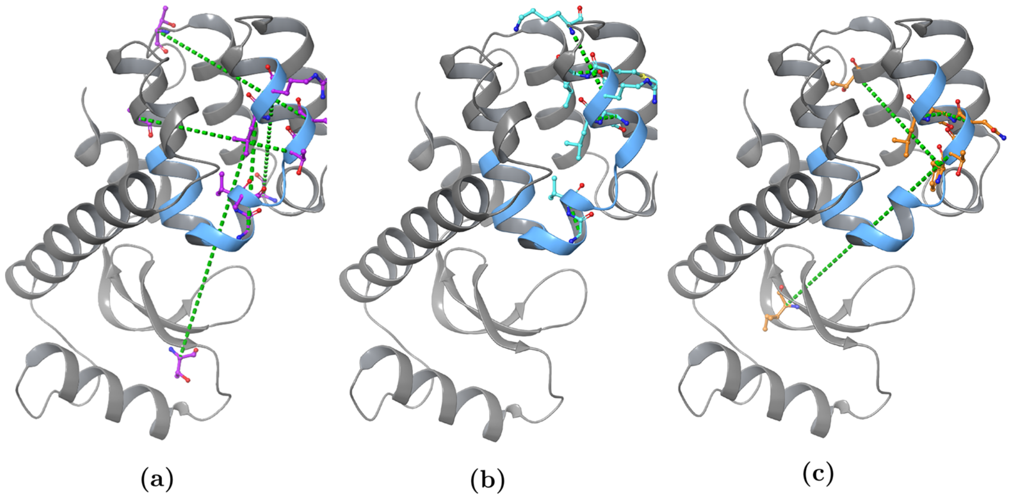


**Figure 2.** Visual representation of the three highest-weighted OPs in each of the two RC components for each mutant. The OPs are shown as dashed green lines, while the cavity-adjacent loop is highlighted in blue. The residues used in OPs are shown in purple for the single mutant (a), cyan for the double mutant (b), and orange for the triple mutant (c). There are fewer than six OPs shown for (b) and (c) because some OPs were among the highest three weights for both RC components.

Looking at the dimensionality of the minimal OP set for each mutant, we can determine how many independent components are needed to describe the loop's dynamics and what type of components is needed to do so. We can see that the single, double, and triple mutants have 16, 10, and 5 independent OPs, respectively. Thus, for the single mutant, there are more possible contacts whose fluctuations can lead to change in the RC or, equivalently, whose fluctuations can cause back-and-forth movement between metastable states. Moreover, each subsequent mutation appears to lower the number of OPs or equivalently lowers the global flexibility of the protein. While the single mutant needs an overall larger number of contacts to describe the RC for conformational change, even more interestingly, the double mutant has a larger number of these contacts solely between residues from the loop region. This means that the loop in the double mutant is more flexible than the other two mutants while the single mutant exhibits greater global flexibility.[64] Though these assertions about the flexibility of the mutants are loosely correlated with the RMSF calculations shown in Figure S5, they are drawn from relatively short unbiased trajectories and could be improved with better sampling.

In general, the use of AI-based frameworks suffer from an interpretability issue due to the black-box nature of AI. In RAVE, the use of a simple linear encoder circumvents this problem and allows us to look directly at the weights of the OPs themselves. When considering the highest three weights for each RC component of each mutant (shown in Figure 2), the single mutant has few interloop contacts while the double mutant has primarily interloop contacts and the triple mutant has a mix of interloop contacts. This again means that the double mutant shows greater local flexibility and this local flexibility plays a key role in the cavity−adjacent loop's rearrangements, while the triple mutant is globally as well as locally less flexible than the other two given the much smaller number of independent OPs. Furthermore, we needed to consider a two-component RC to sample the loop's movement, which implies that there are two processes that must occur in order for the conformation to change. In summary, we see that while conformational change in the single and the triple mutants needs to be triggered by long-distance fluctuations in the protein, changes in the double mutant can be triggered through short-distance fluctuations. This difference could be why these three sequences differ so distinctly in their propensity for taking different conformations.

**Free-Energy Calculation.** With the two-dimensional RCs as described in the previous subsection for different mutants, we then performed 500 ns of metadynamics for each mutant. Two independent copies were performed to reduce noise and ascertain run-to-run variability. The effect of the metadynamics bias was reweighted in order to recover unbiased statistics.[65] These reweighted statistics were then combined with definitions of various states/conformations to calculate the free-energy difference between the ground and excited states. We remind the reader that we defined these using the F114 $\Psi$ dihedral angle, the hydrogen bonding distance between G113 and S117, and the RMSD deviation over C$\alpha$ atoms from the non-native conformation for each mutant. The distance and dihedral angle as shown in Figure 1 have been used in previous studies,[23] but we found them to be insufficient to define the states because they could yield false positives when compared to experimental estimates of the conformational free-energy difference.[24] Specifically, the secondary structure of the cavity-adjacent loop may become distorted over time, especially during metadynamics, which may yield conformations that match the distance and dihedral criteria for a state but no longer correspond to the experimentally reported structures.

These unfolded states must be accounted for in both state definition and metadynamics simulation. Even a biased simulation with an optimized RC, or equivalently a very long unbiased simulation, can get trapped in these unfolded configurations with very high RMSDs from any state of interest. In principle, these trapping states are not a problem because the simulation would eventually return to regions of greater interest. However, they can lead to debilitating computational efficiency. Therefore, we included a wall in terms of the RMSD that prevents the protein from exploring conformations far from the states of interest in our metadynamics simulation. (See the details in Methods.) The state definitions using distance, the dihedral angle, and the RMSD were set using a semiautomatic procedure whose details are provided in Methods.

Table 2. State Definitions for the Ground and Excited States of Each Mutant Using F114 Ψ, the 113−117 Distance, and the RMSD

| mutant | GS distance (Å) | GS Ψ (radians) | ES distance (Å) | ES Ψ (radians) | RMSD cutoff (Å) |
|---|---|---|---|---|---|
| L99A | 7 ± 1.5 | 0.75 ± 0.375 | 2.5 ± 1.5 | −0.75 ± 0.375 | <3.6 |
| L99A, G113A | 6 ± 1.5 | 0.75 ± 0.375 | 2.5 ± 1.5 | −0.30 ± 0.375 | <4.4 |
| L99A, G113A, R119P | 6 ± 1.5 | 0.75 ± 0.375 | 2.5 ± 1.5 | −0.75 ± 0.375 | <5.0 |

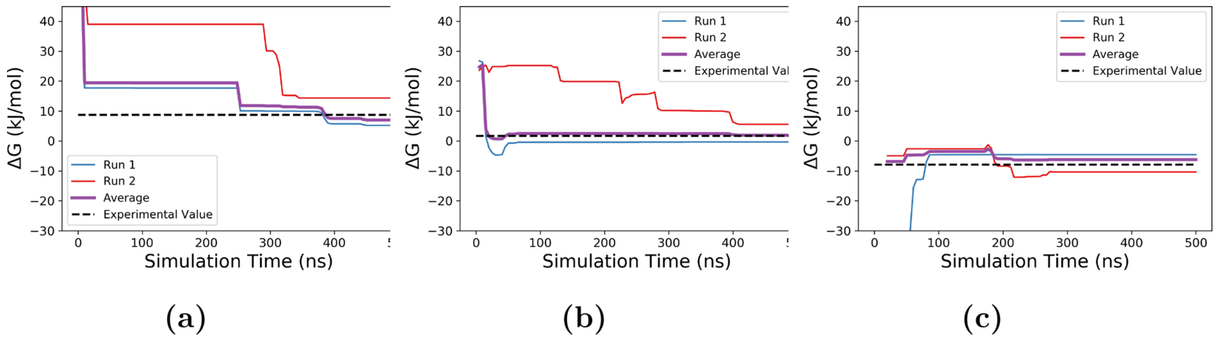


Figure 3. Free-energy differences between the excited and ground states over time for the single mutant (a), double mutant (b), and triple mutant (c) starting from the first time step where both the ground and excited states have been visited. The evolution of Δ$G$ for each replica is shown in red and blue. The Δ$G$ corresponding to the average probabilities of the two replicas is shown in purple. It is important to note that this averaging was done by averaging the Boltzmann probabilities of each state and then calculating the Δ$G$ from these averaged probabilities. The Δ$G$ value reported in ref 20 is shown as a dashed line.

Once these state definitions, shown in Table 2, had been established for each mutant, we calculated the relative thermodynamic stability of these conformations, as quantified by their free-energy difference. In Figure 3 we show the free-energy differences between the excited and ground states over the two independent 500 ns metadynamics runs each for the single, double, and triple mutants respectively. These free-energy differences were calculated using the time evolution of the Boltzmann probabilities of the excited and ground states. For the average of the two runs, we average the Boltzmann probabilities for each state across the runs and calculate the free energy from these averaged probabilities. While there is run-to-run scatter, the averaged values shown in Figure 4 (see Methods for averaging and error estimate protocols) are in excellent qualitative and quantitative agreement with the experimental values reported by Bouvignies et al.[20] It should be noted that the plateaus in free energy in Figure 3 can correspond to either a converged estimate of the free-energy difference or a prolonged exploration of conformations that do not correspond to their ground or excited state. However, as can be seen from the trajectories of different OPs as well as movies provided in the Supporting Information (SI), we have ample movement between different conformations. In the SI we also provide various free-energy profiles (Figures S1−S4) along relevant OPs, including the RMSD and RC components. These profiles clearly show why we needed to use restraints in the RMSD space. In order to ascertain how sensitive our calculations are to the use of such restraints, we did not apply such a restraint for the triple mutant. While we still obtained the relative free energy in good agreement with experiment (Figure 3), corresponding Figure S3 in the SI when compared with Figures S1 and S2, where we did use restraints, shows the usefulness of doing so. Finally, Figure S4 in the SI shows how the estimate from Figure 3 would have changed with different RMSD values used in defining states (Table 2). The values are extremely robust until we reach the range where the restraints were active and the sampling is no longer reliable.

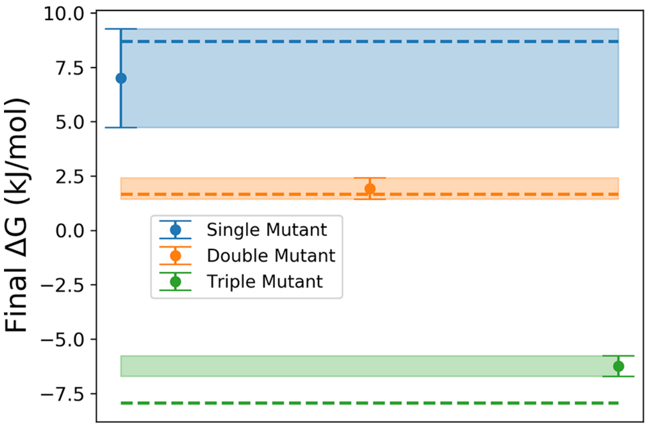


Figure 4. Free-energy differences between the excited and ground states after 500 ns of metadynamics simulation were determined using two replicas for each state. Each mutant is represented by the final Δ$G$ value with error bars obtained by block averaging. For comparison, the Δ$G$ corresponding to the distributions reported in ref 20 are shown as dotted lines matching the color of the corresponding mutant.

## DISCUSSION AND CONCLUSIONS

In this work, we have proposed and demonstrated an all-atom molecular dynamics-based simulation protocol for quantifying flexible loop conformational propensities as a function of the underlying protein sequence. The enhanced sampling procedure of this work represents an automated framework for sampling loop conformations which can be used for both the free-energy ranking of known states and the discovery of new states. Here we have referred to the classic but challenging test piece of mutations in the T4 lysozyme family of proteins. Comparing the final free-energy differences and their error calculated with block averaging (Figure 4), a clear ranking emerges that recovers the correct order of thermodynamic stability. Not only is the ranking correct, but for all three mutants we obtain conformational free-energy differences that are nearly identical to the experimental values, within thermal fluctuation and sampling error margins. In addition to obtaining correct thermodynamic propensities of the different conformations, we have also gained crucial atomic-scale insight into the differences in the dynamics among the three mutants. We see that the total number of OPs needed to describe the dynamics of each mutant decreases as additional mutations occur, suggesting a decrease in global flexibility. Despite the trend in the total number of OPs, the double mutant has the largest number of interloop OPs, suggesting increased local flexibility in the cavity-adjacent loop. We also obtain insight from looking at the weights of the different OPs that build up the RC for the three systems. We see that the increased local flexibility in the double mutant plays a key role in its conformational rearrangements.

While this work presents a step toward a fully automated procedure for sampling loop conformations in all-atom resolution, we would like to highlight that the state definitions for loop conformations are still an open problem requiring a large quantity of expert knowledge. Though the state definition is not yet automated in our work, thereby leading to us not calling it fully automated as of yet, progress has been made in developing generalized methods for the state definition, for example, by using path collective variables.[24] We also point out that the agreement of absolute free-energy differences with experimental values for three systems is not sufficient for the statistical validity of our protocol. That will need more tests across many more systems, which has not really been possible so far with all-atom MD simulations apart from a few recent advanced methods.[66] Methods such as those in ref 66 and in this article should be helpful in facilitating precisely these type of calculations across many systems in a high-throughput manner. In conclusion, the combination of statistical mechanics and AI-based methods AMINO, RAVE, and metadynamics allowed us to enhanced the sampling of the ground and excited states in flexible proteins in a semi-automatic manner with minimal human intervention. Through this we could obtain accurate free energies as well as physically relevant basis functions that give a direct mechanistic understanding of protein conformational plasticity and possibly inspire future experimental or computational studies.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpcb.0c03985.

Additional figures and analysis (PDF)

## ■ AUTHOR INFORMATION

### Corresponding Author

**Pratyush Tiwary** − *Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States;* orcid.org/0000-0002-2412-6922; Email: ptiwary@umd.edu

### Authors

**Zachary Smith** − *Biophysics Program and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States*

**Pavan Ravindra** − *Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States*

**Yihang Wang** − *Biophysics Program and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States*

**Rory Cooley** − *Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.0c03985

### Author Contributions

§P.R. and Y.W. contributed equally to this work.

### Author Contributions

All authors contributed to running the simulations, analyzing the results, and writing the manuscript.

### Notes

The input files necessary for reproducing the metadynamics simulations reported in this article can be found on PLUMED-NEST[57] at plumed-nest.org/eggs/20/007/. The scripts necessary for reproducing the OP and RC calculations can be found on GitHub at github.com/tiwarylab.

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We thank Schrodinger LLC for providing us access to their software suite. We thank the Donors of the American Chemical Society Petroleum Research Fund (no. PRF 60512-DNI6), the NCI-UMD Partnership for Integrative Cancer Research, and the COMBINE fellowship (DGE-1632976) for financial support. We thank Deepthought2, MARCC, Biowulf, and XSEDE (projects CHE180007P and CHE180027P) for the computational resources used in this work.

## ■ REFERENCES

(1) Dill, K. A.; Ozkan, S. B.; Shell, M. S.; Weikl, T. R. The protein folding problem. *Annu. Rev. Biophys.* **2008**, *37*, 289−316.

(2) Thirumalai, D.; O'Brien, E. P.; Morrison, G.; Hyeon, C. Theoretical perspectives on protein folding. *Annu. Rev. Biophys.* **2010**, *39*, 159−183.

(3) Dill, K. A.; MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **2012**, *338*, 1042−1046.

(4) Schuler, B.; Eaton, W. A. Protein folding studied by single-molecule FRET. *Curr. Opin. Struct. Biol.* **2008**, *18*, 16−26.

(5) AlQuraishi, M. End-to-end differentiable learning of protein structure. *Cell Syst* **2019**, *8*, 292.

(6) James, L. C.; Tawfik, D. S. Conformational diversity and protein evolution − a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **2003**, *28*, 361−368.

(7) Tokuriki, N.; Tawfik, D. S. Protein dynamism and evolvability. *Science* **2009**, *324*, 203−207.

(8) Jensen, R. A. Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **1976**, *30*, 409−425.

(9) Welch, G. R.; Somogyi, B.; Damjanovich, S. The role of protein fluctuations in enzyme action: A review. *Prog. Biophys. Mol. Biol.* **1982**, *39*, 109−146.

(10) Wrabl, J. O.; Gu, J.; Liu, T.; Schrank, T. P.; Whitten, S. T.; Hilser, V. J. The role of protein conformational fluctuations in allostery, function, and evolution. *Biophys. Chem.* **2011**, *159*, 129−141.

(11) Frauenfelder, H.; Chen, G.; Berendzen, J.; Fenimore, P. W.; Jansson, H.; McMahon, B. H.; Stroe, I. R.; Swenson, J.; Young, R. D.

A unified model of protein dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 5129−5134.

(12) Wüthrich, K. The second decade − Into the third millenium. *Nat. Struct. Biol.* **1998**, *5*, 492−495.

(13) Pellecchia, M.; Sem, D. S.; Wüthrich, K. NMR in drug discovery. *Nat. Rev. Drug Discovery* **2002**, *1*, 211−219.

(14) Streaker, E. D.; Beckett, D. Ligand-linked structural changes in the Escherichia coli biotin repressor: The significance of surface loops for binding and allostery. *J. Mol. Biol.* **1999**, *292*, 619−632.

(15) Dagliyan, O.; Tarnawski, M.; Chu, P.-H.; Shirvanyants, D.; Schlichting, I.; Dokholyan, N. V.; Hahn, K. M. Engineering extrinsic disorder to control protein activity in living cells. *Science* **2016**, *354*, 1441−1444.

(16) Hanson, S. M.; Georghiou, G.; Thakur, M. K.; Miller, W. T.; Rest, J. S.; Chodera, J. D.; Seeliger, M. A. What makes a kinase promiscuous for inhibitors? *Cell Chem. Biol.* **2019**, *26*, 390−399.

(17) Xiang, Z.; Steinbach, P. J.; Jacobson, M. P.; Friesner, R. A.; Honig, B. Prediction of side-chain conformations on protein surfaces. *Proteins: Struct., Funct., Genet.* **2007**, *66*, 814−823.

(18) Eriksson, A. E.; Baase, W. A.; Wozniak, J. A.; Matthews, B. W. A cavity-containing mutant of T4 lysozyme is stabilized by buried benzene. *Nature* **1992**, *355*, 371−373.

(19) Mulder, F. A. A.; Mittermaier, A.; Hon, B.; Dahlquist, F. W.; Kay, L. E. Studying excited states of proteins by NMR spectroscopy. *Nat. Struct. Biol.* **2001**, *8*, 932−935.

(20) Bouvignies, G.; Vallurupalli, P.; Hansen, D. F.; Correia, B. E.; Lange, O.; Bah, A.; Vernon, R. M.; Dahlquist, F. W.; Baker, D.; Kay, L. E. Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. *Nature* **2011**, *477*, 111−114.

(21) Feher, V. A.; Schiffer, J. M.; Mermelstein, D. J.; Mih, N.; Pierce, L. C.; McCammon, J. A.; Amaro, R. E. Mechanisms for benzene dissociation through the excited state of T4 lysozyme L99A Mutant. *Biophys. J.* **2019**, *116*, 205−214.

(22) Wang, Y.; Martins, J.; Lindorff-Larsen, K. Biomolecular conformational changes and ligand binding: From kinetics to thermodynamics. *Chem. Sci.* **2017**, *8*, 6466−6473.

(23) Schiffer, J.; Feher, V.; Malmstrom, R.; Sida, R.; Amaro, R. Capturing invisible motions in the transition from ground to rare excited states of T4 lysozyme L99A. *Biophys. J.* **2016**, *111*, 1631−1640.

(24) Wang, Y.; Papaleo, E.; Lindorff-Larsen, K. Mapping transiently formed and sparsely populated conformations on a complex energy landscape. *eLife* **2016**, *5*, No. e17505.

(25) McCammon, J. A.; Gelin, B. R.; Karplus, M.; Wolynes, P. G. The hinge-bending mode in lysozyme. *Nature* **1976**, *262*, 325−326.

(26) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **2018**, *149*, 072301.

(27) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Past-future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat. Commun.* **2019**, *10*, 3573.

(28) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annu. Rev. Phys. Chem.* **2016**, *67*, 159−184.

(29) Wang, Y.; Tiwary, P. *Understanding the role of predictive time delay and biased propagator in RAVE*; arXiv e-prints 2020, arXiv:2002.06099.

(30) Ravindra, P.; Smith, Z.; Tiwary, P. Automatic mutual information noise omission (AMINO): Generating order parameters for molecular systems. *Mol. Syst. Des. Eng.* **2020**, *5*, 339−348.

(31) Huse, M.; Kuriyan, J. The conformational plasticity of protein kinases. *Cell* **2002**, *109*, 275−282.

(32) Vijayan, R. S. K.; He, P.; Modi, V.; Duong-Ly, K. C.; Ma, H.; Peterson, J. R.; Dunbrack, R. L.; Levy, R. M. Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. *J. Med. Chem.* **2015**, *58*, 466−479.

(33) Adcock, S. A.; McCammon, J. A. Molecular dynamics: Survey of methods for simulating the activity of proteins. *Chem. Rev. (Washington, DC, U. S.)* **2006**, *106*, 1589−1615.

(34) Pan, A. C.; Borhani, D. W.; Dror, R. O.; Shaw, D. E. Molecular determinants of drug−receptor binding kinetics. *Drug Discovery Today* **2013**, *18*, 667−673.

(35) Tiwary, P.; van de Walle, A. In *Multiscale Materials Modeling for Nanomechanics*; Weinberger, C. R., Tucker, G. J., Eds.; Springer International Publishing: Cham, 2016; pp 195−221.

(36) Ribeiro, J. M. L.; Tsai, S.-T.; Pramanik, D.; Wang, Y.; Tiwary, P. Kinetics of ligand−protein dissociation from all-atom simulations: Are we there yet? *Biochemistry* **2019**, *58*, 156−165.

(37) Wang, Y.; Ribeiro, J. M. L.; Tiwary, P. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr. Opin. Struct. Biol.* **2020**, *61*, 139−145.

(38) Bussi, G.; Laio, A.; Tiwary, P. In *Handbook of Materials Modeling: Methods: Theory and Modeling*; Andreoni, W., Yip, S., Eds.; Springer International Publishing: Cham, 2018; pp 1−31.

(39) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(40) Dama, J. F.; Parrinello, M.; Voth, G. A. Well-tempered metadynamics converges asymptotically. *Phys. Rev. Lett.* **2014**, *112*, 240602.

(41) Dama, J. F.; Rotskoff, G.; Parrinello, M.; Voth, G. A. Transition-tempered metadynamics: Robust, convergent metadynamics via on-the-fly transition barrier estimation. *J. Chem. Theory Comput.* **2014**, *10*, 3626−3633.

(42) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, No. eaaw1147.

(43) Noé, F.; De Fabritiis, G.; Clementi, C. Machine learning for protein folding and dynamics. *Curr. Opin. Struct. Biol.* **2020**, *60*, 77−84.

(44) Preto, J.; Clementi, C. Fast recovery of free energy landscapes via diffusion-map-directed molecular dynamics. *Phys. Chem. Chem. Phys.* **2014**, *16*, 19181−19191.

(45) Tiwary, P.; Berne, B. J. Spectral gap optimization of order parameters for sampling complex molecular systems. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 2839−2844.

(46) Smith, Z.; Pramanik, D.; Tsai, S.-T.; Tiwary, P. Multi-dimensional spectral gap optimization of order parameters (SGOOP) through conditional probability factorization. *J. Chem. Phys.* **2018**, *149*, 234105.

(47) Lamim Ribeiro, J. M.; Tiwary, P. Toward achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through RAVE. *J. Chem. Theory Comput.* **2019**, *15*, 708−719.

(48) Bonati, L.; Zhang, Y.-Y.; Parrinello, M. Neural networks-based variationally enhanced sampling. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 17641−17647.

(49) Brotzakis, Z. F.; Parrinello, M. Enhanced sampling of protein conformational transitions via dynamically optimized collective variables. *J. Chem. Theory Comput.* **2019**, *15*, 1393−1398.

(50) Mondal, J.; Tiwary, P.; Berne, B. J. How a kinase inhibitor withstands gatekeeper residue mutations. *J. Am. Chem. Soc.* **2016**, *138*, 4608−4615.

(51) Chiavazzo, E.; Covino, R.; Coifman, R. R.; Gear, C. W.; Georgiou, A. S.; Hummer, G.; Kevrekidis, I. G. Intrinsic map dynamics exploration for uncharted effective free-energy landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E5494−E5503.

(52) Palmer, S. E.; Marre, O.; Berry, M. J.; Bialek, W. Predictive information in a sensory population. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6908−6913.

(53) Shwartz-Ziv, R.; Tishby, N. *Opening the black box of deep neural networks via information*; arXiv preprint 2017, arXiv:1703.00810 2017.

(54) Still, S. Information bottleneck approach to predictive inference. *Entropy* **2014**, *16*, 968−989.

(55) Alemi, A. A.; Fischer, I.; Dillon, J. V.; Murphy, K. *Deep variational information bottleneck*; arXiv preprint 2016, arXiv:1612.00410 2016.

(56) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604−613.

(57) Bonomi, M.; Bussi, G.; Camilloni, C.; Tribello, G. A.; Banáš, P.; Barducci, A.; Bernetti, M.; Bolhuis, P. G.; Bottaro, S.; Branduardi, D.; et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **2019**, *16*, 670−673.

(58) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(59) Parrinello, M.; Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys. (Melville, NY, U. S.)* **1981**, *52*, 7182−7190.

(60) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An Nlog(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089−10092.

(61) Tiwary, P.; Parrinello, M. From metadynamics to dynamics. *Phys. Rev. Lett.* **2013**, *111*, 230602.

(62) Salvalaglio, M.; Tiwary, P.; Parrinello, M. Assessing the reliability of the dynamics reconstructed from metadynamics. *J. Chem. Theory Comput.* **2014**, *10*, 1420−1425.

(63) Pietrucci, F.; Laio, A. A collective variable for the efficient exploration of protein beta-sheet structures: Application to SH3 and GB1. *J. Chem. Theory Comput.* **2009**, *5*, 2197−2201.

(64) Verma, D.; Jacobs, D. J.; Livesay, D. R. Changes in lysozyme flexibility upon mutation are frequent, large and long-ranged. *PLoS Comput. Biol.* **2012**, *8*, No. e1002409.

(65) Tiwary, P.; Parrinello, M. A time-independent free energy estimator for metadynamics. *J. Phys. Chem. B* **2015**, *119*, 736−742.

(66) He, P.; Zhang, B. W.; Arasteh, S.; Wang, L.; Abel, R.; Levy, R. M. Conformational free energy changes via an alchemical path without reaction coordinates. *J. Phys. Chem. Lett.* **2018**, *9*, 4428−4435.