JCTC Journal of Chemical Theory and Computation

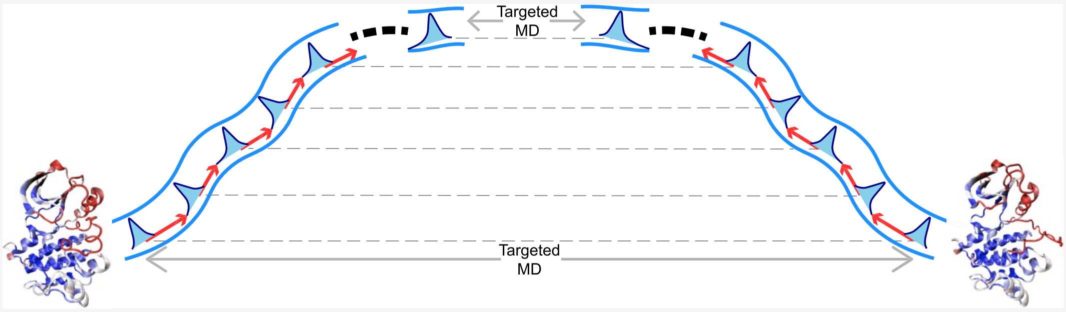
pubs.acs.org/JCTC Article

Identifying a Feasible Transition Pathway between Two Conformational States for a Protein

Yao Li and Haipeng Gong*

Cite This: J. Chem. Theory Comput. 2022, 18, 4529−4543 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information



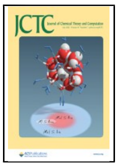
ABSTRACT: Proteins usually need to transit between different conformational states to fulfill their biological functions. In the mechanistic study of such transition processes by molecular dynamics simulations, identification of the minimum free energy path (MFEP) can substantially reduce the sampling space, thus enabling rigorous thermodynamic evaluation of the process. Conventionally, the MFEP is derived by iterative local optimization from an initial path, which is typically generated by simple brute force techniques like the targeted molecular dynamics (tMD). Therefore, the quality of the initial path determines the successfulness of MFEP estimation. In this work, we propose a method to improve derivation of the initial path. Through iterative relaxation-biasing simulations in a bidirectional manner, this method can construct a feasible transition pathway connecting two known states for a protein. Evaluation on small, fast-folding proteins against long equilibrium trajectories supports the good sampling efficiency of our method. When applied to larger proteins including the catalytic domain of human c-Src kinase as well as the converter domain of myosin VI, the paths generated by our method deviate significantly from those computed with the generic tMD approach. More importantly, free energy profiles and intermediate states obtained from our paths exhibit remarkable improvements over those from tMD paths with respect to both physical rationality and consistency with a priori knowledge.

1. INTRODUCTION

The biological functions of proteins are usually realized through the structural transition between different conformational states.[1] In ion channels like Na$_V$ channels, the pore domain switches from a “closed” state to an “open” state during voltage-gated activation to allow the transmembrane permeation of Na$^+$ ions.[2] Glucose transporters like GLUT3 alternate between inward facing and outward facing states so as to facilitate the transport of D-glucose across the membrane.[3] Kinases like ADK[4] and c-Src[5] interconvert between inactive and active conformations during phosphorylation and dephosphorylation processes, which enables the cellular regulation of enzyme activity. Despite the great success, experimental structural determination methods such as nuclear magnetic resonance (NMR), X-ray crystallography, and cryo-electron microscopy (Cyro-EM) can only obtain the high-resolution structures of certain highly stable states for a given protein, thereby missing the dynamic characteristics harbored in other states because of their transient nature or the limitation of in vitro conditions. Small-angle X-ray scattering (SAXS) and single-molecule techniques can capture the structural dynamics to a certain extent but fail to provide molecular details at the atomistic level.

As a powerful computational tool to facilitate the understanding of protein functions as well as the structure-based drug design, molecular dynamics (MD) simulations aim to provide the spatial and temporal information of protein structures simultaneously,[1] although in practice, they often suffer from the lack of convergence because of the long time scales of many biological processes that exceed the limit of in silico simulations. To improve convergence, a variety of strategies have been developed to maximally explore the protein conformational space starting from an experimentally determined structural state. Besides simple elongated simu-

Received: April 18, 2022
Published: June 20, 2022



https://doi.org/10.1021/acs.jctc.2c00390

lations using dedicated hardware such as Anton[6−9] and graphics processing unit (GPU),[10−12] enhanced sampling methods are designed to help the protein escape from local energy basins or overcome barriers. A mainstream class of methods expands the sampling space by modifying the potential energy surface using biasing potentials.[13] For example, energy boosts are exerted on the original potential to reduce the heights of local barriers in accelerated MD (aMD);[14] local minima in the already-sampled region are filled with positive history-dependent Gaussian potentials to avoid redundant sampling in metadynamics[15−17] and its variants;[18−20] deep-learning techniques are engaged for dimensionality reduction in VAE,[21,22] AE,[23,24] and VAMPnet[25] as well as fitting free energy surface or biasing potential in reinforced dynamics,[26] DeepVES[27] and TALOS.[28] Another class of methods enhances the sampling in the less sampled regions by iteratively selecting suitable seed structures as restarting points for new simulations, including but not limited to FEXS,[29] SDS,[30] CoCo MD,[31] and weighted ensemble.[32,33] Besides, novel machine-learning-based methods trying to draw independent samples from the equilibrium states in one shot have emerged, exemplified by variational autoregressive networks[34] and Boltzmann generators.[35]

Albeit powerful, the above methods become less efficient when at least two structural states of the target protein have been determined experimentally. In this scenario, essential interests are focused on how to find the atomistic details of conformational interconversion between these fixed states, instead of expanding the sampling of the overall conformational space. A number of methods have been proposed to fulfill such demands specifically. For instance, path sampling techniques, such as transition path sampling (TPS)[36] and its derivatives, transition interface sampling (TIS),[37−40] milestoning,[41] and forward flux sampling (FFS),[42,43] obtain an ensemble of transition pathways using numerous unbiased sampling trajectories and choose reactive trajectories by weights in the transition-state ensemble for reaction rate estimation. Weighted ensemble methods,[32] on the other hand, sample pathways by spawning child trajectories upon reaching new regions of configuration space and assigning weights to them. TeDA2[44] uses adaptive seed structures to sample from both endpoints in reduced feature spaces. When sufficient amount of sampling is available with these methods, the pathways can be derived by visual inspection or from kinetic network models like Markov state models (MSM).[45−47] However, the sampling itself is usually far from sufficiency because of the ubiquitous high barriers that separate metastable states and the numerous coordinates involved in the process. Additionally, MSMs use crude dimensionality reduction methods like clustering to estimate the probabilities of macrostates and thus cannot provide a rigorous free energy evaluation of the process.

Hence, it is necessary to further reduce the dimensionality by finding the most probable reaction pathway or the minimum free energy path (MFEP), which allows fast and sufficient sampling along it and the subsequent rigorous evaluation of the potential of mean force (PMF). Previous strategies generate an initial path using linear interpolation (morphing methods),[48] targeted MD (tMD),[49] or other enhanced sampling techniques[50,51] and then optimize the guessed path via path searching methods represented by the string method[52] and its variants.[53−55] These path searching methods, however, only allow local optimization around the initial path and frequently fail when the initial path deviates far away from the true MFEP, particularly on the rugged free energy surface where numerous energy peaks located in-between hinder the iterative path refinement. Consequently, a physically unreliable free energy estimation will be produced upon a poor initial path. For instance, in a study of transition between the active and inactive states of c-Src kinase, the free energy profile along the path that was initialized from tMD and then optimized using the string method with swarms of trajectories (SMwST) exhibited an energy barrier of ∼30 kcal/mol,[56] a value that is highly unlikely to be overcome by atomic thermal motions of the system. Therefore, improving the estimation of the initial path is of great importance. The adaptive anisotropic network model (aANM)[57] was an early attempt, which guides the protein to move toward target direction based on the recombination of normal modes. Although fast, such an elastic network model (ENM)[58]-derived method is still naive because the normal modes spanning the space of internal motions are extracted within one native structure,[58] unable to reflect kinetic information of the whole trajectory. Adelman and Grabe combined the string method and weighted ensemble to improve the sampling along the transition path.[59] Unfortunately, when applied on protein targets, the initial path was still generated using a two-state elastic network model that was produced upon two fixed protein states. Sultan and Pande engaged a small number of low-frequency motions estimated from equilibrium MD simulations on the fixed ending structures as collective variables to guide the Metadynamics sampling.[60] Without an adaptive strategy, this kind of method needs substantial sampling of the intermediate states of the transition process to identify a good set of collective variables. Besides, neglecting medium frequency modes may hinder the effectiveness of transition path sampling.

In this work, we developed a collective motion-based bidirectional adaptive sampling (COMBAS) method to construct a physically reasonable transition pathway between a pair of functional states for a protein, particularly when the two states have sufficient structural difference. Starting from both ends, the protein is driven toward the other end along the collective motion that is represented as the optimal linear combination of a large number of motions identified by time−structure-based independent components analysis (tICA)[61,62] in each cycle, followed by sufficient relaxation. Through adaptively updating the ends and the subsequent sampling, the iteratively produced structural frames from relaxation (i.e., equilibrium) MDs in all cycles enable the construction of a feasible transition path, which allows rigorous PMF evaluation after the path optimization using the string method. Here, we first validated the basic principle of this method on three small, fast-folding proteins (chignolin, trp-cage, and villin) and found that COMBAS sampling could successfully capture the major characteristics of the reversible folding processes, nearly to the same extent as the extremely long equilibrium simulations but with tremendously reduced simulation time. Then, we applied this method on more complicated protein systems, where the large-scale conformational changes can hardly be tracked by existent equilibrium simulations. In the investigation on the inactive-active transition for the catalytic domain of human c-Src kinase (denoted as c-Src^{C})[5,63] as well as the interconversion between the pre-power-stroke (PPS) and rigor (R) states for the converter domain of myosin VI (denoted as MVI^{C}),[64,65] the paths generated by COMBAS deviate

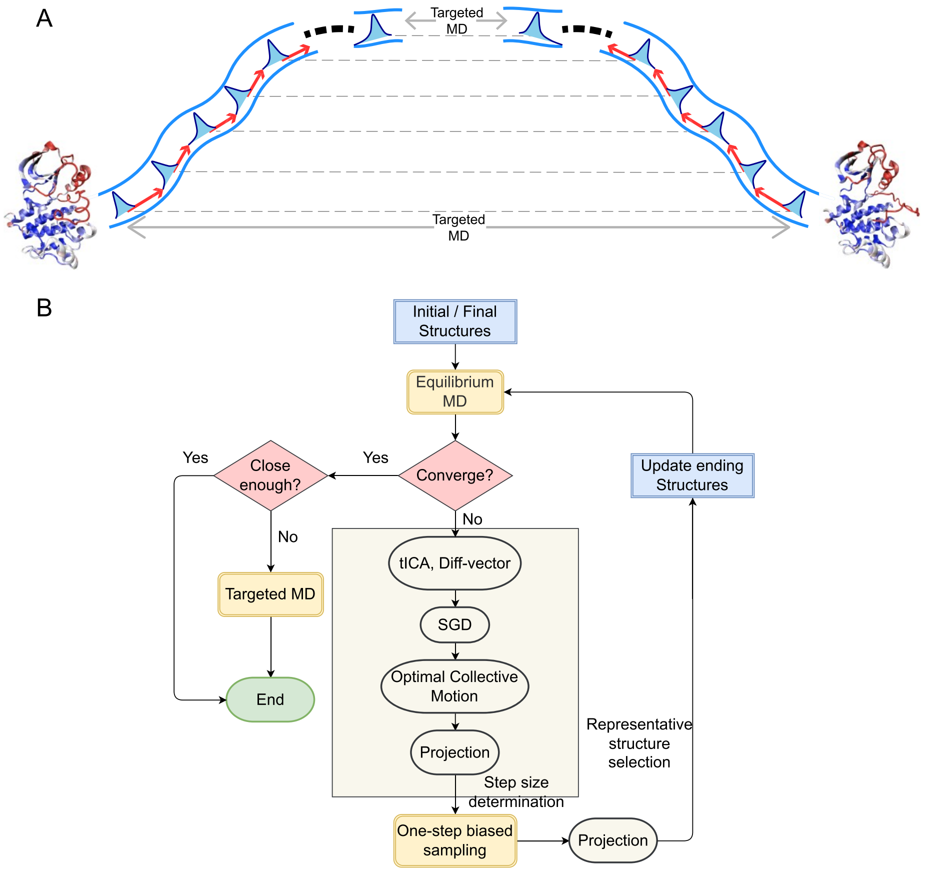


Figure 1. (A) Schematic illustration of the overall COMBAS sampling process. Starting from two known protein conformational states, the algorithm proceeds bidirectionally in an adaptive manner, generating two series of biasing−relaxation iterations. In each cycle, the protein is first forced to move along an ideal direction (i.e., the optimal collective motion, shown as red arrows) calculated from an equilibrium simulation (shown as normal distributions colored in blue) by a small-step biased simulation and then sufficiently relaxed for the next iteration. The dotted lines indicate the directions of Diff-vectors in each cycle. The upper tMD is optional, used only when there exists a gap in the COMBAS sampling region, whereas the lower tMD denotes the traditional way of path initialization for complicated protein systems, which was taken as control in the experiments on c-Src^{C} and MVI^{C}. Structural frames from all equilibrium trajectories finally compose a feasible transition path. (B) Flowchart of COMBAS sampling. Only one computational block starting from one ending structure is shown here. In practice, two blocks are executed simultaneously in each cycle to allow bidirectional sampling.

remarkably from those produced using the traditional tMD approach. More importantly, in comparison to tMD, the path obtained by our method shows substantial improvement in both the free energy estimation and the consistency with a priori knowledge.

2. METHODS

2.1. Generation of the Initial Path by COMBAS.

2.1.1. General Idea and Implementation of the Overall Pipeline. The general purpose of COMBAS is to identify a physically reasonable transition pathway that is spatially close to the true MFEP in principle, at low consumption of computational time and resources. The whole pipeline proceeds in an iterative manner (Figure 1). At the beginning of each cycle, we run conventional MD (cMD) to allow both ending structures to relax locally in the conformational space. For each ending structure, we perform tICA to identify the slowly decorrelating eigenvectors corresponding to low frequency modes, that is, the time−structure-based independent components (tICs), based on the cMD trajectory and then try to find an optimal linear combination of these tICs that is best correlated with the difference vector (Diff-vector) between the two ending structures. The collective motion along this optimal linear combination of slow tICs reflects an ideal movement mode that comprehensively considers physical feasibility and structural guidance. Subsequently, we pick up the cMD snapshot with the largest deviation away from the first structure along the identified collective motion and initialize a short biased MD to moderately amplify the movement in that direction. At the end of each cycle, both ending structures are updated by frames in the biased MD trajectories. The whole scheme is a relaxation-biasing iteration executed bidirectionally from the initial and final structural states of a target protein, with both ending structures and collective motions updated adaptively. The cMD snapshots obtained in the relaxation simulations in all cycles thus compose two sets of intermediates that approach each other gradually in the conformational space. The sampling converges when the smallest Euclidian distance between the two sets of structures within the space of top principal components (PCs) could no longer reduce. After convergence, if the two cMD trajectories in the last cycle do not overlap, we perform tMD to fill the gap. Nevertheless, all sampled structural frames from cMD and tMD if necessary in COMBAS allow the construction of an initial guess of the transition path, which is subsequently optimized using the string method. Finally, a protocol to calculate PMF along the refined path parameterized using the principal curve[66] is established to further reveal thermodynamic properties of the structural transition process. The detailed implementation of COMBAS is described below.

(1) The method is based on the simultaneous generation of pairs of intermediate conformations starting from the known endpoints A and B, two local minima on the potential energy surface. We first align B to A to obtain the 3 N-dimensional representations $\boldsymbol{X}^A$ and $\boldsymbol{X}^B$, that is, the Cartesian coordinates of Cα atoms in A and B, respectively. At the beginning of the ith cycle, where i = 1, 2, 3, ..., we run short cMD simulations to improve the local sampling for the starting structures $\boldsymbol{X}^A_{i-1}$ and $\boldsymbol{X}^B_{i-1}$, respectively. Particularly, in the first cycle, $\boldsymbol{X}^A_0 = \boldsymbol{X}^A$, $\boldsymbol{X}^B_0 = \boldsymbol{X}^B$. The outcome trajectories are superposed to $\boldsymbol{X}^A$ to remove translational and rotational degrees of freedom, resulting in ξ^A_i and ξ^B_i. In this work, 10 ns cMD is performed with 1000 snapshots saved in each trajectory.

(2) We calculate a difference vector d$\vec{v}_i$, named Diff-vector, to define the direction from one ending point to the other in the ith cycle, and normalize it to the unit length:

$$\overrightarrow{\mathrm{d}v_i}^A = \frac{\boldsymbol{X}^B_{i-1} - \boldsymbol{X}^A_{i-1}}{\|\boldsymbol{X}^B_{i-1} - \boldsymbol{X}^A_{i-1}\|}, \tag{1}$$

$$\overrightarrow{\mathrm{d}v_i}^B = -\overrightarrow{\mathrm{d}v_i}^A . \tag{2}$$

(3) Instead of using geometric similarity as a proxy for kinetic similarity, we perform tICA, a dimensionality reduction technique producing slowly decorrelated eigenmodes $\vec{\alpha}^A(\vec{\alpha}^B)$ (i.e., tICs), to extract slow movement modes from the time series ξ_i^A (ξ_i^B). The eigenvalue of the time-lag correlation matrix becomes nearly zero at about the midpoint of d dimensions, and thus, only the eigenvectors with positive eigenvalues are taken into consideration:

$$\vec{\alpha}_j^A(j = 1, 2, 3, ..., [d/2]) , \tag{3}$$

$$\vec{\alpha}_j^B(j = 1, 2, 3, ..., [d/2]). \tag{4}$$

(4) In order to avoid sampling around irrelevant local minima and accelerate convergence, we seek to find the most probable direction, along which the conformational sampling is kinetically accessible. Based on this idea, the Pearson correlation coefficient (PCC) between the linear combination of the top J dominant eigenvectors (with low frequencies) and the Diff-vector is evaluated. Considering that PCC is identical to the correlation cosine between the two vectors under this circumstance, this problem is formulated as an optimization problem with the objective function of

$$f^A(J) = \vec{v}^A_{\mathrm{OCM}}\cdot\overrightarrow{\mathrm{d}v_i}^A \; (J = 1, 2, 3, ..., [d/2]) \tag{5}$$

,

$$\vec{v}^A_{\mathrm{OCM}} = \frac{\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^A}{\left\|\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^A\right\|} \; (J = 1, 2, 3, ..., [d/2]) \tag{6}$$

,

$$f^B(J) = \vec{v}^B_{\mathrm{OCM}}\cdot\overrightarrow{\mathrm{d}v_i}^B (J = 1, 2, 3, ..., [d/2]), \tag{7}$$

$$\vec{v}^B_{\mathrm{OCM}} = \frac{\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^B}{\left\|\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^B\right\|} \; (J = 1, 2, 3, ..., [d/2]). \tag{8}$$

For each J, stochastic gradient descent (SGD) is used to search for the best weights (ω_j, j = 1,2, ..., J) for input tICs, with the loss function set as the opposite of the inner product between the resulting combined vector and Diff-vector under certain J:

$$\mathrm{LOSS}^A(\omega|J) = -\frac{\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^A}{\left\|\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^A\right\|}\cdot\overrightarrow{\mathrm{d}v_i}^A , \tag{9}$$

$$\mathrm{LOSS}^B(\omega|J) = -\frac{\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^B}{\left\|\sum_{j=1}^{J}\omega_j\vec{\alpha}_j^B\right\|}\cdot\overrightarrow{\mathrm{d}v_i}^B . \tag{10}$$

To accelerate calculation, we evaluate the objective functions only for J values at multiples of 10, which means that every 10 tICs are grouped together for the final decision on the optimal J (see the Results for details). This step provides the optimal collective motions $\vec{v}^A_{\mathrm{OCM}}$. and $\vec{v}^B_{\mathrm{OCM}}$

(5) We then calculate projections of all cMD frames (ξ_i) on the optimal collective motion $\vec{v}_{\mathrm{OCM}}$ with respect to the first structure,

$$\mathrm{PROJ}^A = \langle \xi_i{}^A | \vec{v}^A_{\mathrm{OCM}} \rangle, \tag{11}$$

$$\mathrm{PROJ}^B = \langle \xi_i{}^B | \vec{v}^B_{\mathrm{OCM}} \rangle, \tag{12}$$

where the brackets denote the inner product between ξ_i and $\vec{v}_{\mathrm{OCM}}$.

(6) Subsequently, the system is made to move along the optimal collective motion via a one-step biased sampling. "One step" means a short biased simulation (2 ns here) targeted to a putative central structure that is generated by applying a small-step movement along the direction of optimal collective motion. We use a scaling factor k to control the step size. The limit of k → 0 refers to infinitesimally small displacements that are accurate (identical to unbiased sampling) but computationally costly. By selecting a proper k value (0.15 here), we achieve a tradeoff between sampling consumption and negative effects of artificial biases. The step size of one-step biased sampling is determined as:

$$\mathrm{step}^A = k\cdot\mathrm{RMSD}^2(\boldsymbol{X}^A_{i-1}, \boldsymbol{X}^B_{i-1})\cdot\mathrm{gap}^A, \tag{13}$$

$$\mathrm{step}^B = k\cdot\mathrm{RMSD}^2(\boldsymbol{X}^A_{i-1}, \boldsymbol{X}^B_{i-1})\cdot\mathrm{gap}^B, \tag{14}$$

where the gap defined as

$$\mathrm{gap}^A = \mathrm{median}(\mathrm{top}10\%(\mathrm{PROJ}^A)) - \mathrm{median}(\mathrm{bottom}10\%(\mathrm{PROJ}^A)), \tag{15}$$

$$\mathrm{gap}^B = \mathrm{median}(\mathrm{top}10\%(\mathrm{PROJ}^B)) - \mathrm{median}(\mathrm{bottom}10\%(\mathrm{PROJ}^B)), \tag{16}$$

stands for the difference between the median of the top 10% projections and that of the bottom 10% projections and roughly reflects the range of conformational change along $\vec{v}_{\mathrm{OCM}}$ in a cMD trajectory. Here, the root-mean-square-displacement (RMSD) is adopted to dynamically adjust the step size. For example, when the representative structures are separated far away with a large RMSD, the biased sampling should walk by a relatively large step and vice versa. We tested multiple values (0, 1, and 2) for the power of RMSD in the formula and found that using square of RMSD scaled by k and gap in the step could make the calculation converge at a moderate speed and keep k as a nearly constant value simultaneously.

(7) After the biased MD, the trajectory is projected onto the optimal collective motion and the frame with the largest projection is chosen to update the ending structure ($\boldsymbol{X}_i^A$ or $\boldsymbol{X}_i^B$), from which new cMD simulation is initiated in the next cycle. In practice, we also refer to the RMSD between the two ends when making decision, that is, structures with a large projection and a small RMSD are selected.

Steps 1−7 are repeated until the convergence criterion is satisfied.

2.1.2. Convergence Determination of COMBAS. We evaluate convergence after the cMD (step 1, see Figure 1) of each cycle. Specifically, following the principal component analysis (PCA), all historical cMD trajectories are projected into the space of the top principal components (PCs). Here, the number of PCs is determined based on the ratio of total variance accumulatively explained by the selected number of top PCs. Convergence is believed to be reached only when the two sets of cMD frames collected from iterative cycles starting from the two initial structural states can no longer approximate each other in the PC space, or more precisely, the Euclidean distance between the two sets does not drop with the proceeding of iterative cycles. A complete transition path is already available if the two sets overlap in the last cycle. Otherwise, we perform tMD to fill the gap.

2.1.3. Computational Architecture. The overall scheme is a relaxation-biasing iteration, accompanied with sequential or parallel operations. Based on simulation origins (i.e., the initial end points A and B), the calculation can be divided into two blocks, the A started block and B started block. Within each block, the computation is executed sequentially, while two blocks are computed parallelly in each cycle. Notably, the execution of the next iteration must wait until both blocks in the previous cycle are finished so that the ending structures could be updated.

2.2. Path Optimization. For a transition completed within a large sampling space, the location of MFEP is of central interest. Initiating from a predefined or guessed path generated using various sampling techniques, path optimization schemes usually approach the MFEP by expanding the sampling region, including the string method with drifting paths, the traveling-salesman based automated path searching (TAPS)[67,68] that sample conformations perpendicular to the reference path, as well as other path refining methods.

In this study, we perform PCA analysis on the sampled structures (by COMBAS or tMD) to reduce dimensionality of the original Cartesian space. The top five PCs explaining more than 80% of the total variance are selected to construct a reduced feature space. We then adopt the post-hoc string method (PHSM)[69] to generate a putative transition pathway, which is further refined by SMwST.[53] The SMwST improves the sampling and lowers energy barriers by relaxing image conformations on the path iteratively. Here, we use Cartesian coordinates of the Cα atoms as collective variables for SMwST calculation to ensure continuity in the high-dimensional space, as reported in the literature.[70] After 70 iterations of the procedure, the path is well refined to reach convergence.

2.3. Free Energy Calculation. In the work of the finite temperature string (FTS) method,[54,55] the reaction path is parametrized by a principal curve, which simplifies the following PMF calculation.[71] Borrowing this idea, we also adopt the principal curve to parameterize the pathway predetermined in the previous section (Section 2.2) and evaluate the free energy profile along the curve. The principal curve,[66] a nonparametric generalization of principal components, is a smooth one-dimensional curve that passes through the middle of the dataset orthogonally, providing a 1D description of the data. Like in the previous work,[72] we employ umbrella sampling to evaluate the PMF as a function of the curve parameter as described below.

For a sequence of conformations along the path denoted as $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, ..., $\boldsymbol{X}_N$, a principal curve is constructed to fit the dataset of Cartesian coordinates of Cα atoms. Any point $\boldsymbol{X}$ in the configuration space can be projected to the closest point on the curve to obtain the corresponding $\boldsymbol{X}^{\mathrm{cur}}$. We use λ to denote the cumulative curve length such that the arc length between two images is given by

$$\lambda_2 - \lambda_1 = \int \|\mathrm{d}\boldsymbol{X}^{\mathrm{cur}}\|. \tag{17}$$

The umbrella potential is

$$U_i = \frac{k}{2}(\lambda - \lambda_i)^2, \tag{18}$$

$$U_i(\boldsymbol{X}) = \frac{k}{2}(\boldsymbol{X}^{\mathrm{cur}} - \boldsymbol{X}_i^{\mathrm{cur}})^2 \tag{19}$$

where λ_i and $\boldsymbol{X}_i^{\mathrm{cur}}$ are the reference point and the structure for the ith window along the principal curve.

The direction vector from $\boldsymbol{X}_{i-1}^{\mathrm{cur}}$ to $\boldsymbol{X}_i^{\mathrm{cur}}$ is

$$\vec{d}_i = \frac{\boldsymbol{X}_i^{\mathrm{cur}} - \boldsymbol{X}_{i-1}^{\mathrm{cur}}}{\|\boldsymbol{X}_i^{\mathrm{cur}} - \boldsymbol{X}_{i-1}^{\mathrm{cur}}\|}. \tag{20}$$

Hence, for each structure in $\boldsymbol{X}$ that is restrained around $\boldsymbol{X}_i$:

$$\boldsymbol{X}^{\mathrm{cur}} \approx \boldsymbol{X}_i^{\mathrm{cur}} + (\boldsymbol{X} - \boldsymbol{X}_i^{\mathrm{cur}})\cdot\vec{d}_i, \tag{21}$$

$$U_i(\boldsymbol{X}) \approx \frac{k}{2}[(\boldsymbol{X} - \boldsymbol{X}_i^{\mathrm{cur}})\cdot\vec{d}_i]^2. \tag{22}$$

The biasing potential is applied as the summed projection of deviations of Cα atoms coordinates from reference coordinates onto $\vec{d}_i$:

$$p(\boldsymbol{X}_i^t, \boldsymbol{X}_i^{\mathrm{cur}}) = \sum_{j=1}^{n} \vec{d}_i^{\,j}\cdot(U(\boldsymbol{X}_i^t - \boldsymbol{X}_i^t(\mathrm{cog})) - (\boldsymbol{X}_i^{\mathrm{cur}} - \boldsymbol{X}_i^{\mathrm{cur}}(\mathrm{cog}))), \tag{23}$$

where U is the optimal rotation matrix, $\boldsymbol{X}_i^t(\mathrm{cog})$ and $\boldsymbol{X}_i^{\mathrm{cur}}(\mathrm{cog})$ are the centers of the geometry of the current and reference positions, respectively, and $\vec{d}_i^{\,j}$ is the component of the direction vector ($\vec{d}_i$) for the jth atom.

Adjacent umbrella windows are sampled with sufficient overlaps. The final PMF profile is obtained using the weighted

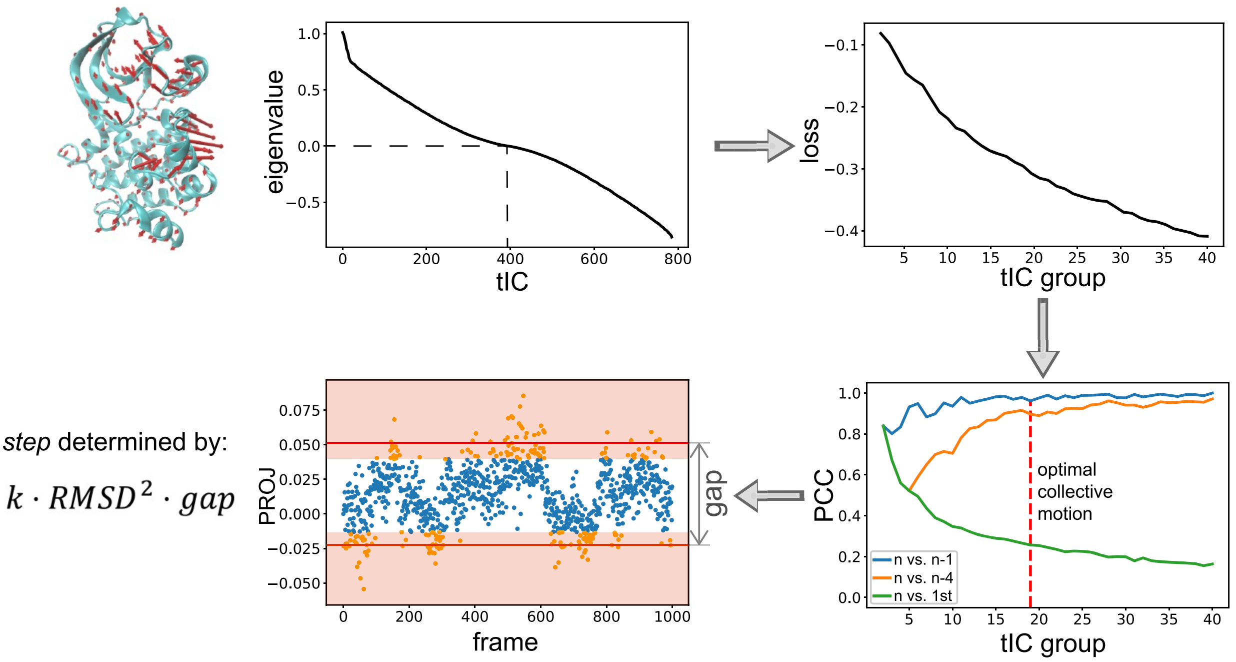


Figure 2. Parameter selection of COMBAS. First, the Diff-vector between ending structures (red arrows in the upper left panel, visualized by the ProDy plugin in VMD) is estimated. Next, tICA is performed on the cMD trajectory, and SGD is conducted to find the best linear combination within each tIC group. After that, the tIC group is chosen based on PCC evaluation, and the final optimal collective motion is thus determined. Finally, the step size for subsequent one-step biased MD is determined based on the projection of cMD frames onto the optimal collective motion (abbreviated as PROJ).

histogram analysis method (WHAM) in combination with the Bayesian block bootstrapping scheme.[73]

2.4. MD Simulation and Computational Details.

Parameterized by the AMBER ff14SB[74] force field, each simulation system was solvated in a periodic water box filled with explicit water molecules and a necessary number of Na^+ and/or Cl^- ions for neutralization. The cMD simulations were performed on the GPU version of OpenMM[10] in the NVT ensemble with temperature held at 300 K. The long-range electrostatic interactions were calculated using the particle Mesh Ewald and the van der Waals cutoff was set as 10 Å. The time step was set as 2 fs with SHAKE turned on to remove the stretching degrees of freedom of hydrogen-involved bonds. We used Amber[75] to perform tMD with the additional energy term defined as

$$E = 0.5 \cdot \mathrm{FRC} \cdot \mathrm{NARMSD} \cdot (\mathrm{CURRMSD} - \mathrm{TGTRMSD})^2, \tag{24}$$

where FRC stands for the force constant and NARMSD denotes the number of atoms involved in RMSD calculation ($C\alpha$ atoms in this work), while CURRMSD and TGTRMSD refer to the current and expected RMSD values, respectively. The one-step biased MD was realized in a similar manner to umbrella sampling, by applying a quadratic potential on $C\alpha$ atoms to penalize the structural deviation from the putative central structure that is estimated from the step size and the optimal collective motion. In this work, both one-step biased MD and umbrella sampling were implemented using NAMD[76] with proper restraining weights. Other parameters in biased MD are similar to those described above for the unbiased MD simulations. All simulations were accelerated by GPU.

We used MDTraj[77] for the postprocessing of MD trajectories, for example, reading, superposing, and coordinate extraction. tICA was conducted using MSMBuilder[45,78] and PCA was implemented with Scikit-learn.[79] Tensorflow[80] was employed for optimizing the loss function by SGD. The workflow of COMBAS was mainly written in Python and NAMD configuration files, with connecting shell scripts to allow automatic running. The codes are freely available at the GitHub site of https://github.com/Gonglab-THU-MD/COMBAS.

3. RESULTS

We first evaluated the sampling efficiency of our COMBAS method on three small proteins, chignolin (PDB id: 5AWL), trp-cage (PDB id: 2JOF; the first NMR model used here), and villin (PDB id: 2F4K), taking the extremely long cMD trajectories (106, 208, and 125 μs, respectively) simulated by D. E. Shaw Research[9] as the reference for comparison. Then, we comprehensively tested COMBAS on two larger human proteins, c-Src tyrosine kinase and myosin VI, where the structural transition between distinct conformational states are impracticable to achieve by pure cMD simulations. For these two proteins, we sampled the transition paths using the COMBAS method and refined the paths using the string method. Because of the lack of available cMD trajectories for comparison, we evaluated against the paths generated following a generic protocol:[56,81,82] initialize the path by pure tMD simulations and refine it using the string method. C-Src is a protein playing key roles in the cell cycle[83] with well-characterized inactive (PDB id: 2SRC) and active (PDB id: 1Y57) states. Here, we only adopted the catalytic domain (residues 260−521) with ATP included and denoted the system as c-$\mathrm{Src}^{\mathrm{C}}$. During activation, the catalytic domain experiences a large conformational change, where the overall RMSD reaches ∼9 Å and certain residues in the A-loop move as far as ∼25 Å. Myosin VI is the only known member of the myosin superfamily that walks in the opposite direction to all

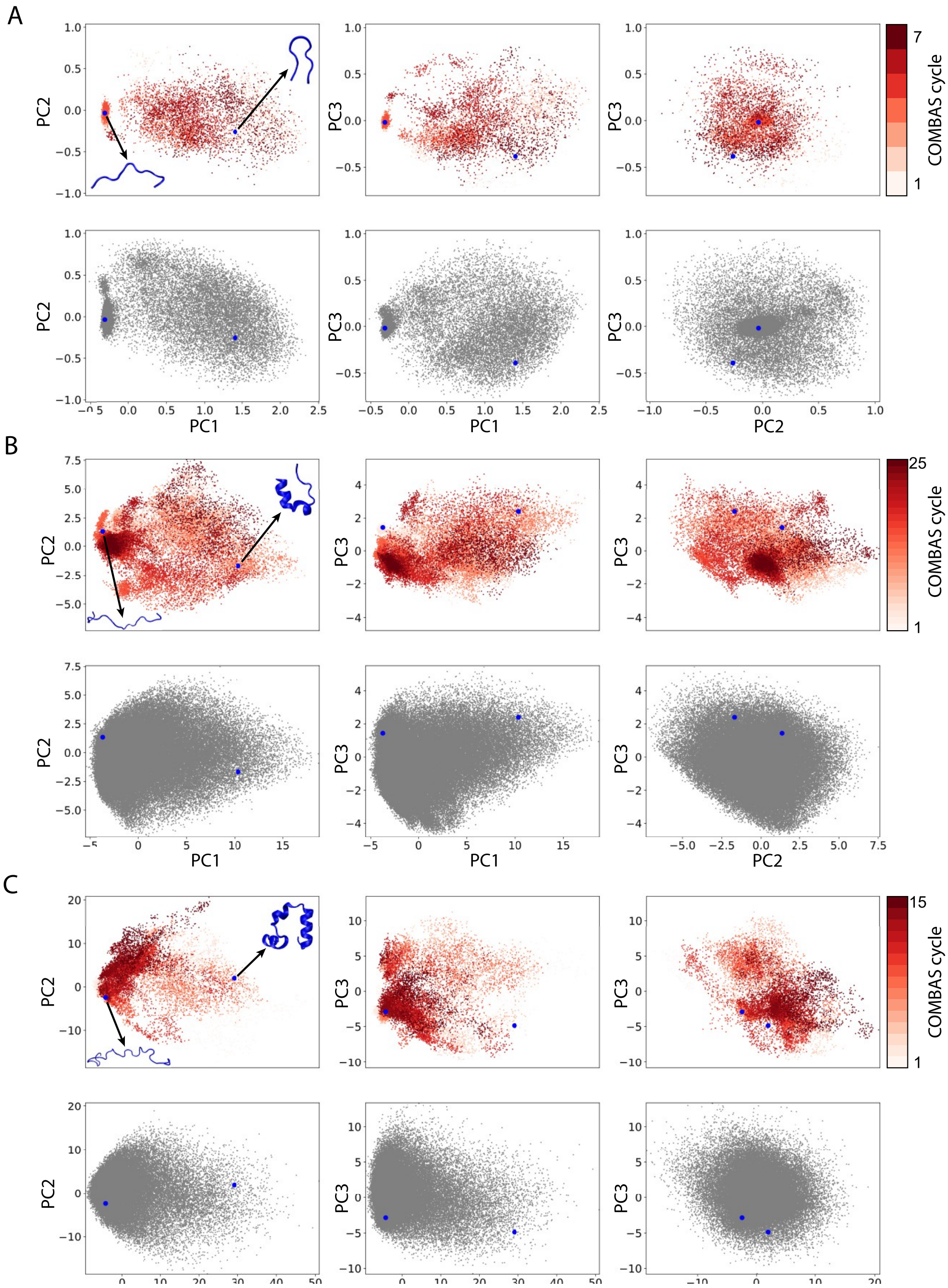


Figure 3. COMBAS sampling was validated via comparison with the cMD reference trajectory on the reversible folding of chignolin (A), trp-cage (B), and villin (C). For each protein, COMBAS cMD trajectories (upper panel) and the reference cMD trajectory (lower panel) are projected into the same top 3 PC space derived from the reference trajectory. For chignolin, the reference trajectory is superposed to its folded structure before performing PCA in the space of Cartesian coordinates of Cα atoms, whereas dimensionality reduction of the other two systems is conducted in the space of pairwise distances between Cα atoms. All the reference trajectories are diluted by 10 folds, and the snapshots sampled by COMBAS are colored from light to dark red corresponding to successive COMBAS cycles. The initial unfolded and folded states are denoted as big blue dots, with the corresponding structures displayed aside in the upper left figures in panels A, B, and C, respectively.

of the other myosins on the actin filament.[84] Here, we isolated the converter domain of myosin VI (residues 703−788, denoted as MVI^C) to study the power stroke process from the pre-power-stroke (PPS) conformer (PDB id: 2 V26) to the rigor (R) conformer (PDB id: 2BKH), during which the overall RMSD is ∼4 Å. Additionally, we also applied COMBAS to a relatively small system, the receiver domain of nitrogen regulatory protein C (denoted as $NtrC^R$), which belongs to the family of two-component systems in bacteria[85] and exhibits an RMSD of ∼3 Å between its inactive state (PDB id: 1 DC7) and active state (PDB id: 1 DC8). Results of $NtrC^R$ can be found in the Supporting Information.

Table 1. Analysis of the Relative Coverage along the Top PCs

	chignolin		trp-cage		villin	
	explained variance ratio	relative coverage (%)	explained variance ratio	relative coverage (%)	explained variance ratio	relative coverage (%)
PC1	0.72	92 ± 2	0.52	85 ± 3	0.43	84 ± 6
PC2	0.06	80 ± 5	0.14	84 ± 3	0.13	89 ± 3
PC3	0.04	87 ± 2	0.09	67 ± 4	0.08	79 ± 3

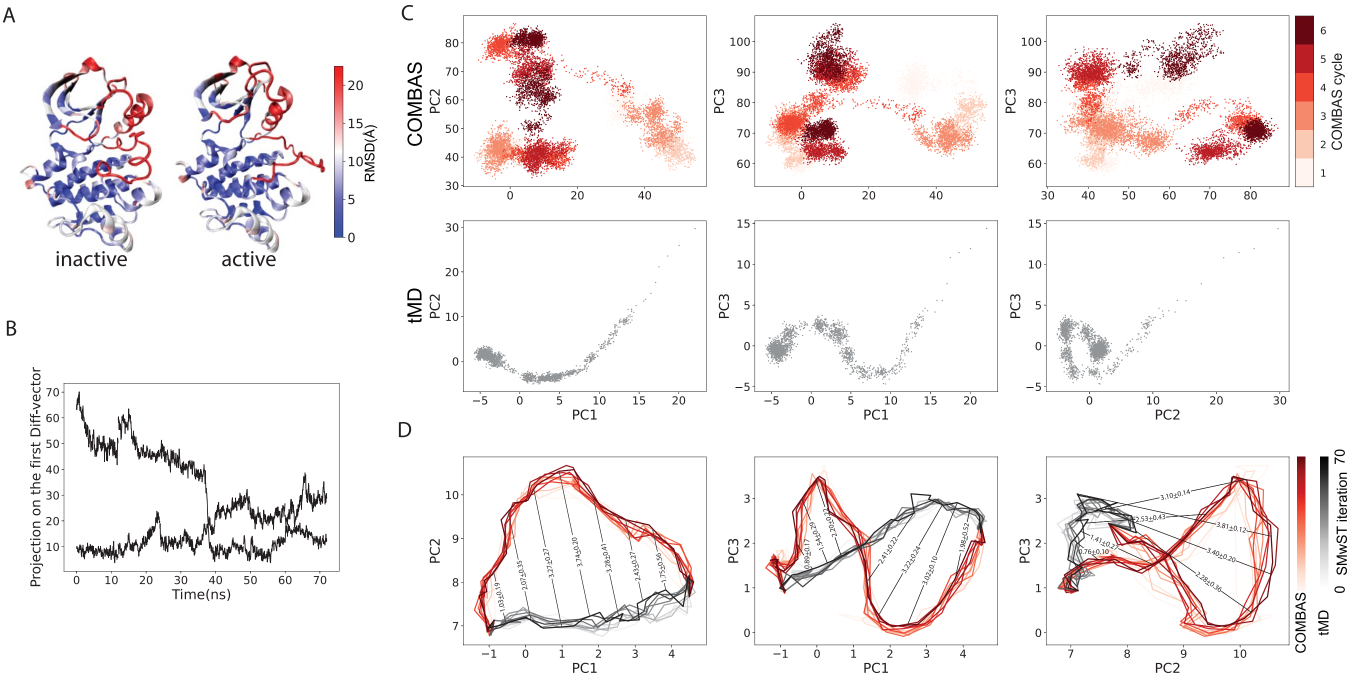


Figure 4. Activation pathways of c-Src^C. (A) Structural illustration of c-Src^C. The residues are colored according to their pairwise RMSD values between the inactive and active states. (B) Projection of historical cMD trajectories onto the Diff-vector between the inactive and active states. Simulations starting from the two structural states cross at ∼40 ns. (C) Projection of COMBAS (upper panel) and tMD (lower panel) samplings in the space of top three PCs. The cMD snapshots sampled in the COMBAS scheme are colored from light to dark red corresponding to COMBAS cycles. (D) Refinement of COMBAS and tMD paths using the string method. Here, only paths at iterations of 0, 7, 14, 21, 28, 35, 42, 49, 56, 63, and 70 are shown, and pairwise distances between images on the two sets of paths in the space of top three PCs are labeled with the mean and standard deviation.

3.1. Parameter Selection of COMBAS. Take c-Src^C (containing 262 residues) as an example. The dimension of this system is 786, and the lag time of tICA performed on the 10 ns equilibrium trajectory is 0.1 ns. As shown in Figure 2, eigenvalues of the first half of all 786 dimensions are greater than 0, indicating that the corresponding tICs represent the dominant slow degrees of freedom in the dynamical process. Discarding tICs with negative eigenvalues, we ranked the remaining tICs by their eigenvalues and grouped them cumulatively, with the number of tICs in each group set as multiples of 10. For instance, group 1 refers to the first 10 tICs, group 2 refers to the first 20 tICs, and group n refers to the first 10 × n tICs. Next, we tried to find the optimal linear combination of tICs within each group by minimizing the loss function (i.e., negative value of PCC between the recombined vector and Diff-vector, see the Methods for details) using SGD. Aiming at selecting global modes but excluding local modes as far as possible, we set up a convergence criterion for identifying a small group of top tICs. In practice, we computed the PCC values between the optimal vectors of the current group (group n) and the first group as well as two previous groups (groups $n - 1$ and $n - 4$). As shown in Figure 2, PCC values converge at a certain group size, meaning that the inclusion of more tICs introduces marginal improvement. Specifically, the optimal tIC group is chosen when the values of two PCC curves (n vs $n - 1$ and n vs $n - 4$, see the blue and orange lines in Figure 2) satisfy $|\text{value}_{i+4} - \text{value}_i|/4 \leq 0.01$ for the current group (i) and two preceding groups ($i - 1$ and $i - 2$). Hence, we chose the best linear combination of this group as the optimal collective motion. By projecting cMD trajectory onto the optimal collective motion, we calculated the median values of the top 10% and the bottom 10% projections and then used them to estimate the step size for the subsequent one-step biased sampling. The parameter selection in the remaining rounds of c-Src^C and MVI^C can be found in Figures S1 and S2, respectively. By projecting the biased trajectory onto the optimal collective motion, we selected representative structures from the late stage of biased sampling that have large projections and small RMSDs to start new cMD simulations for the next cycle.

3.2. Proof of Principle. Here, we validated the sampling efficiency of COMBAS sampling on three small, fast-folding proteins: chignolin, trp-cage, and villin. Specifically, we heated the three-folded systems and then equilibrated them around their melting temperatures (340 K for chignolin, 290 K for trp-cage, and 360 K for villin, the same with the corresponding simulations by D. E. Shaw Research) for 5 ns to obtain relatively stable unfolded structures. Starting from both the

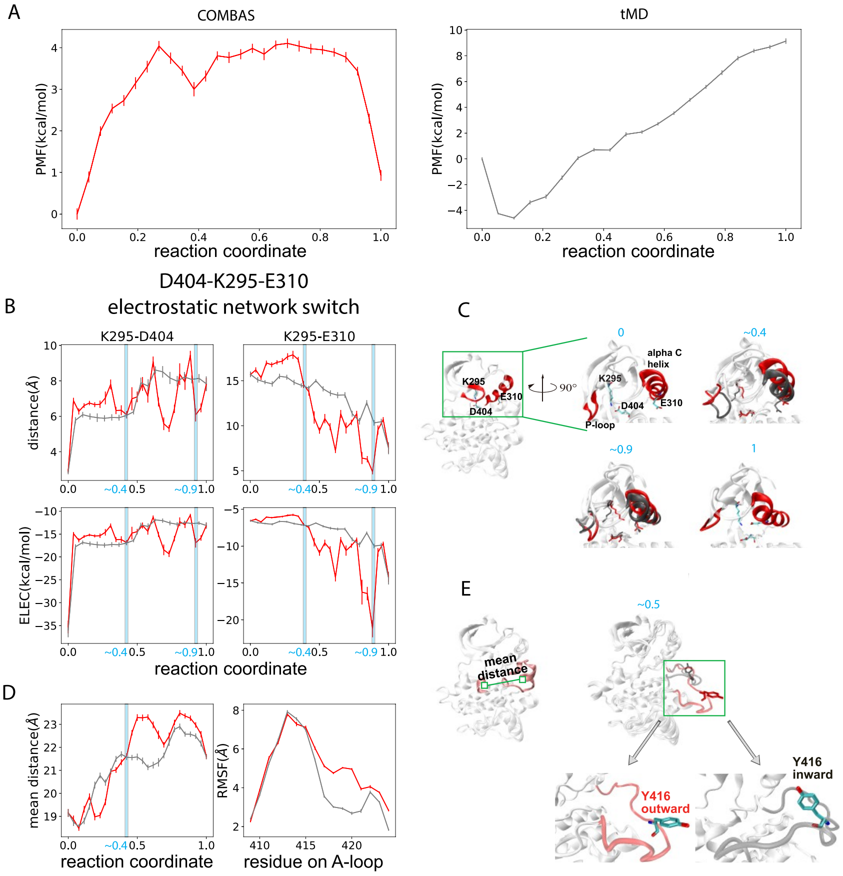


Figure 5. PMF evaluation and structural details for the structural transition of c-Src^C. (A) PMF profiles of the COMBAS (red) and tMD (gray) paths. (B) Switch of the D404-K295-E310 electrostatic network on COMBAS (red) and tMD (gray) paths. Electrostatic interaction is evaluated by the distance between carboxylate O (for D404 and E310) and guanidine N (for K295) and by the magnitude of electrostatic energy (abbreviated as ELEC) in the upper and lower panels, respectively. (C) Representative structural frames at various reaction coordinates (blue numbers on the top of the frames) are selected to highlight the changes of the D404-K295-E310 electrostatic network as well as the associated movement of the αC helix and the P-loop on the COMBAS (red) path vs tMD (dark gray) path. (D) Movement of the fragment between R409 and R424 in the A-loop. The left panel shows the mean distance of this fragment to the center of rigid part of the protein (taking the positions of residues 389−391 as the reference), while the right panel presents the RMSF profile of this fragment. (E) Structural details for the A-loop movement along the COMBAS path. The left panel presents a schematic illustration of the mean distance as evaluated in panel (D). The right panel shows the Y416 outward intermediate sampled by the COMBAS path (red). The corresponding residue remains inward facing on the tMD path (gray).

unfolded and folded states, we then conducted iterative cycles of COMBAS sampling, each consisting of 5 ns cMD and 1 ns biased MD, to resolve the folding processes. To evaluate the sampling efficiency, for each system, we collected all cMD trajectories in the COMBAS cycles and compared with the long cMD reference trajectory (by D. E. Shaw Research) by PCA analysis (Figure 3). Notably, all structural frames were projected into the same PC space, where the top 3 PCs were computed purely based on the cMD reference trajectory. A tMD was also conducted to simulate the folding process and projected into this space (Figure S3). Clearly, the structural frames sampled by COMBAS cMD trajectories exhibit a similar level of coverages on the top 3 PCs when compared to the reference trajectory (Figure 3). As shown in Table 1, the relative coverage, that is, the percentage of the range (along a PC) covered by the reference cMD, which is also explored by COMBAS, exceeds 80% for most PCs, implying that global motions in the folding process are largely tracked by COMBAS sampling. On the other hand, the total simulation time of COMBAS sampling (70, 250, and 150 ns for chignolin, trp-cage, and villin, respectively) is shortened by nearly three orders of magnitude when compared with the cMD reference trajectories (106, 208, and 125 μs, respectively). Hence, the COMBAS method is capable of sampling the feature space to a similar extent as extremely long equilibrium simulations but with remarkably reduced time. These results indicate that COMBAS can capture the structural transition process with a relatively high efficiency.

For each PC, the relative coverage is calculated as the ratio between the absolute coverage range of COMBAS sampling and that of the reference cMD trajectory and is represented in percentage. Bootstrap resampling was performed by 1000 times to estimate the mean values as well as standard errors. Explained variance ratio is the ratio within the total variance explained by a specific PC during PCA analysis.

3.3. Activation of c-Src^C.

3.3.1. Convergence and Path Optimization. The active and inactive states of c-Src^C exhibit a remarkable structure difference (Figure 4A). However, with the proceeding of COMBAS sampling, intermediate structures sampled by iterative cMD trajectories starting from the two distinct conformational states become approximate and even cross each other, when projected on the 1D Diff-vector between the active and inactive states (Figure 4B). In Figure 4C, all cMD snapshots are projected into the space of the top 3

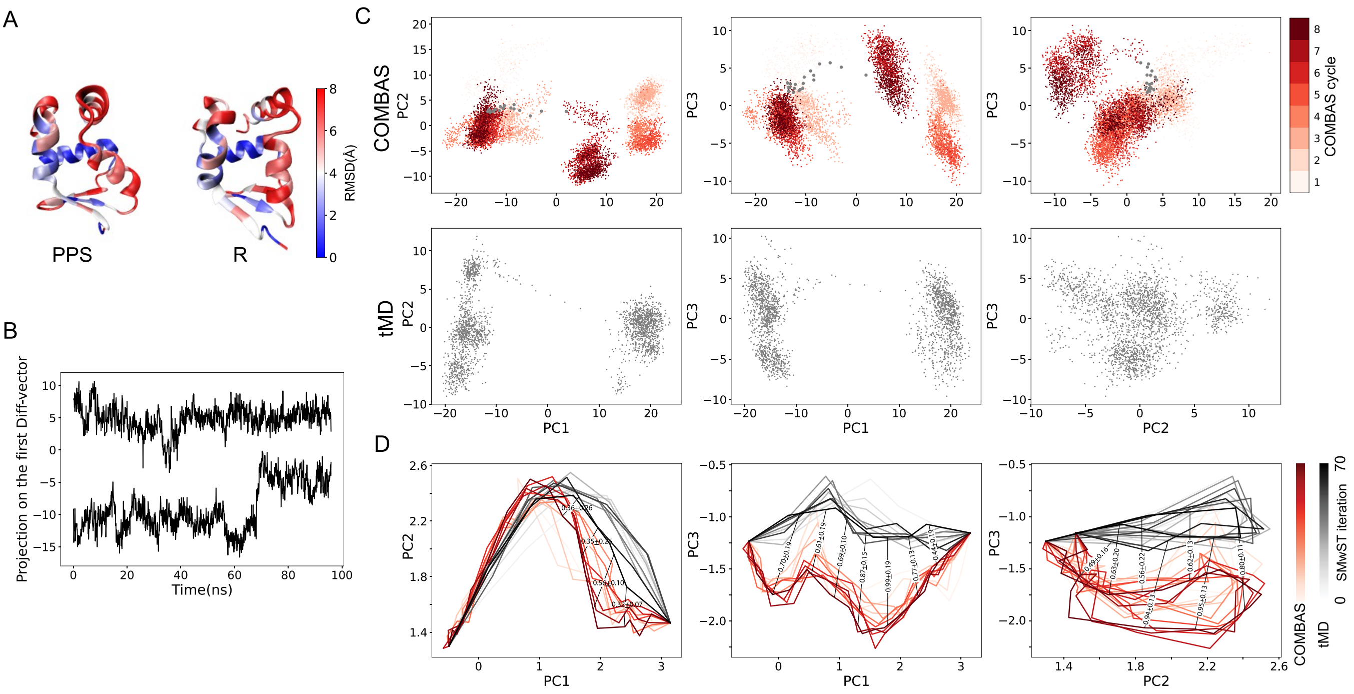


Figure 6. Power stroke pathways of MVI^C. (A) Structural illustration of MVI^C. The residues are colored according to pairwise RMSD values between the PPS and R states. (B) Projection of historical cMD trajectories onto the Diff-vector between the PPS and R states. Simulations starting from two structural states becomes approximate without crossing. (C) Projection of COMBAS (upper panel) and tMD (lower panel) samplings in the space of top three PCs. The cMD snapshots sampled in the COMBAS scheme are colored from light to dark red corresponding to COMBAS cycles. Frames from the extra tMD to fill the gap in COMBAS sampling are highlighted as big gray dots. (D) Refinement of COMBAS and tMD paths using the string method. Here, only paths at iterations of 0, 7, 14, 21, 28, 35, 42, 49, 56, 63, and 70 are shown and pairwise distances between images on the two sets of paths in the space of top 3 PCs are labeled with the mean and standard deviation.

PCs (explaining >75% of the variance). Again, cMD trajectories starting from the inactive and active states gradually approach each other with the proceeding of iterative cycles. The convergence criterion becomes satisfied within six cycles, which cost a total of 72 ns of simulation time for this system. In contrast to the tMD simulation that simply connects the two ending states, COMBAS sampling is diverse, likely taking a tortuous contour to achieve the large-scale conformational change. We then adopted PHSM to pick up putative pathways from COMBAS and tMD samplings, respectively, and engaged the SMwST (referred as the string method in the rest of this work) to further refine them. As shown in Figure 4D, because the string method can only provide local optimization, the final paths are still very close to the initial paths, and therefore, the significant difference between COMBAS and tMD paths still persists after iterative refinement using the string method (see Figure S4 for the convergence check). To quantitatively describe this difference, we calculated the distances between corresponding images on the two sets of paths in the PC space (labeled as mean ± std. in Figure 4D). Based on the one-way analysis of variance (ANOVA) analysis, the COMBAS paths and tMD paths obtained from all iterations in the string method are significantly different, with the p-value < 0.001. This observation reinforces the importance of initial path selection in the derivation of MFEP.

3.3.2. PMF Calculation and Conformational Change along the Path. We parametrized the refined paths of COMBAS and tMD using the principal curve and employed umbrella sampling to evaluate the PMF profiles along both paths (Figure 5A). The COMBAS and tMD paths are divided into 27 and 20 windows, respectively, with 36 ns simulation in each window (see Figure S5 for the convergence of PMF calculation, as justified by the sufficient overlaps between the distributions of neighboring windows along the principal curve). Along both paths, c-Src^C is more stable in the inactive state (reaction coordinate $\lambda = 0$) than in the active state ($\lambda = 1$), as reported in previous studies.[56,86] However, the magnitude of free energy difference exhibits significant distinction: ~10 kcal/mol for tMD path vs ~1 kcal/mol for COMBAS path. The small free energy difference obtained using our method is in excellent agreement with the work of Sultan et al., in which the free energy differences between active and inactive ensembles were found to fall within 1−2 kcal/mol across all sequences of seven Src family kinase (SFK) members.[87] A more recent study[88] also supports the negligible free energy difference between the two states of c-Src^C. Moreover, the PMF curve only presents a small barrier of 4 kcal/mol on the COMBAS path ($\lambda \approx 0.27$), in sharp contrast to the ~14 kcal/mol barrier on the tMD path. As a kinase, the inactive c-Src should be able to sample the active state occasionally, exposing its Y416 for phosphorylation in order to finalize the activation process. In this perspective, PMF of the COMBAS path is clearly more consistent with the functional requirement of c-Src because the inactive−active conformational transition could be overcome by thermal motion. Notably, the free energy barrier of 4 kcal/mol on the COMBAS path agrees well with the value measured by Shukla et al.[86] (4−5 kcal/mol) using a large set of unbiased MD trajectories and is remarkably lower than other previous estimations[56,88] (~30 and ~12 kcal/mol, for example).

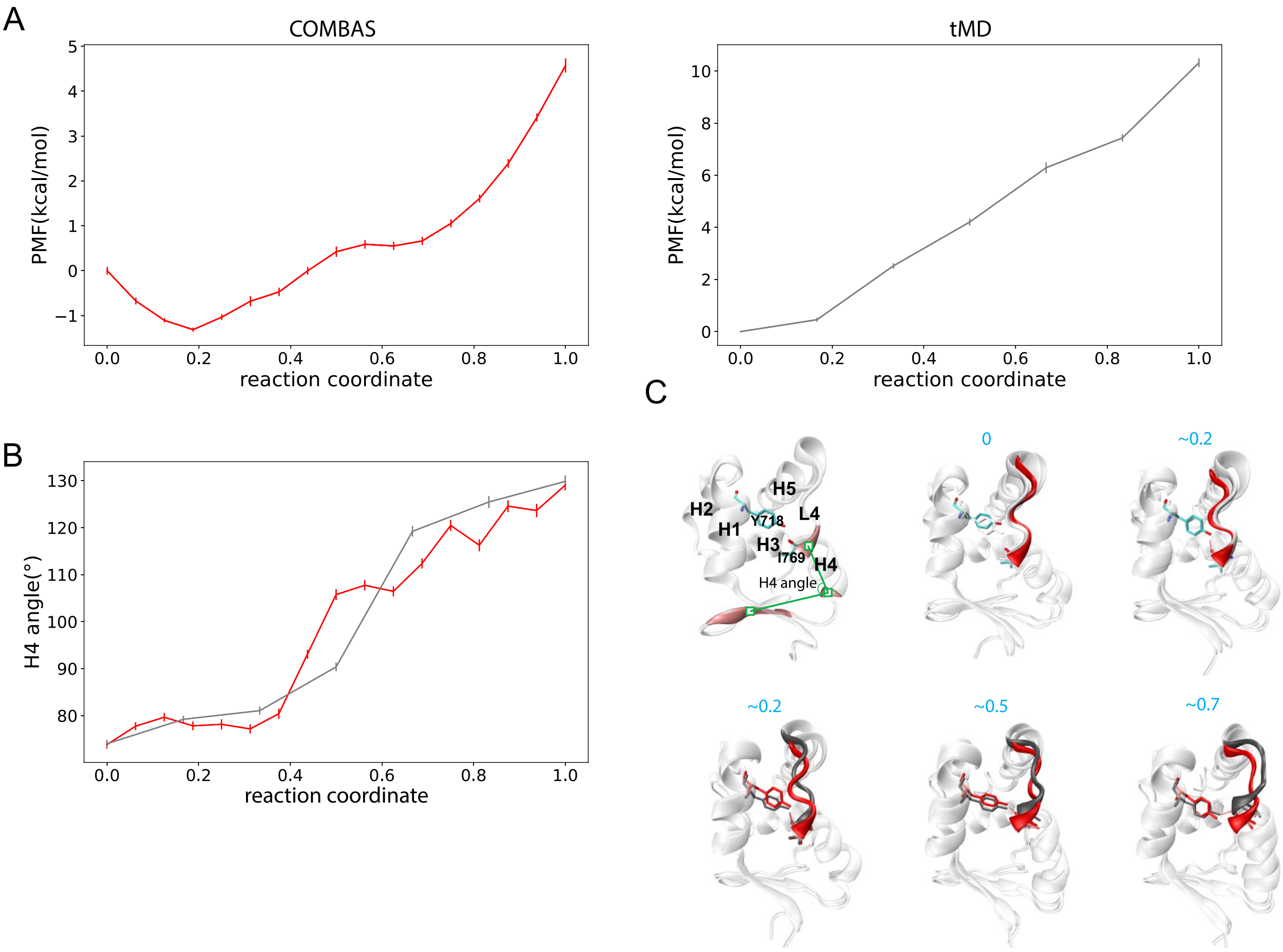


Figure 7. PMF evaluation and structural details for the structural transition of MVI^C. (A) PMF profiles of the COMBAS (red) and tMD (gray) paths. (B) Rotation of the H4 helix along the COMBAS (red) and tMD (gray) paths. (C) L4 loop and the interhelix interactions between Y718 and I769. *Upper panel*: COMBAS snapshots (red) vs PPS state (light gray). *Lower panel*: COMBAS (red) images vs tMD (dark gray) ones. The blue numbers indicate the approximate reaction coordinates for the corresponding structural frames. Hydrogen bonds are highlighted by dashed lines.

We then analyzed the structural details for the paths of COMBAS and tMD. During activation, c-Src^C experiences considerable conformational change, particularly the unfolding of the A-loop (residues 404−424) from short helix-like pieces to an extended loop, which exposes Y416 for phosphorylation, and the accompanying inward rotation of the αC helix (residues 300−317). We first explored the αC helix rotation. On the COMBAS path, K295 drifts away from D404, its hydrogen bonding patterner in the inactive state, at the very beginning (Figure 5B, left panel) and then moves toward E310 on the αC helix to form favorable interactions (λ = 0.8−0.9, Figure 5B, right panel), which may facilitate the inward rotation of the αC helix and the associated uplift of the glycine-rich P loop (Figure 5C). On the tMD path, however, K295 seldom forms interactions with E310 (Figure 5B, right panel). Moreover, unlike the tMD path, the K295-D404 interaction on the COMBAS path is restored temporarily at the late stage of activation ($\lambda \approx$ 0.9, Figure 5B, left panel). This breaking−reformation behavior agrees with the switch of the D404-K295-E310 electrostatic network reported by Ozkirimli and Post.[89]

As for the A-loop, aside from the initiated rearrangement of the hydrophobic core constituted by residues F307, M314, L407, and L410 as well as the opening of the α-helix-like piece as reported before,[56] the E310-R409 hydrogen bond remains intact for a long time before breaking at a very late stage ($\lambda \approx$ 0.9) on the COMBAS path, different from a previous report that E310-R409 hydrogen bond only exists at local minima of the free energy profile.[56] The sustained E310-R409 interaction might be helpful for stabilizing the N-terminal part of the A-loop (residues 404−408). The most important difference between the paths of COMBAS and tMD resides on the fragment between R409 and R424. Quantified by a mean distance from the rigid part of the protein (the center of residues 389−391, see Figure 5E, left panel), this fragment of the A loop is far more extended showing a significantly larger mean distance (for λ > 0.4, see Figure 5D, left panel) and is also more flexible with a larger root-mean-square-fluctuation (RMSF) (Figure 5D, right panel) on the COMBAS path than on the tMD path. In addition, the phosphorylation site Y416 experiences a novel inward−outward−inward transition along the COMBAS path (Figure 5E), an observation that is not found on the tMD path nor reported before. Previous single-molecule and other experimental investigation on Vav1[90] and the complex of p27 and Cdk2/cyclin-A[91] have suggested that protein dynamics may lead to transient exposure of the phosphorylation sites, allowing those residues to interact with the kinase. More directly, Henriques and Lindorff−Larsen[92] captured the transient exposure of phosphorylation sites of the p27 complex (Y88 and Y74) by enhanced sampling simulations and proposed that this motion may exist for buried residues of many other proteins. These previous studies may support the rationality of the transient Y416 outward intermediates sampled by COMBAS.

3.4. Power Stroke of MVI^C. *3.4.1. Convergence and Path Optimization.* The PPS and R states of MVI^C have moderate structural differences (Figure 6A). Again, during COMBAS sampling, the two sets of structural intermediates sampled by iterative cMD trajectories starting from the PPS and R conformers get close to each other, as projected on the Diff-vector (Figure 6B) and in the space of the top three PCs (Figure 6C). However, unlike the c-Src^C case, even when COMBAS sampling becomes convergent at the eighth cycle (with a total simulation time of 96 ns), there still exists a gap between the two parts, although the gap size is smaller than that in the pure tMD sampling. Therefore, we had to conduct an extra brute-force tMD to fill the gap (see the big gray dots in Figure 6C, upper panel). Nevertheless, the combination of all sampled structures allowed the construction of the putative path, which was then optimized using the string method iteratively (convergence check is shown in Figure S4). After path refinement using the string method, the COMBAS path is still significantly different (p-value < 0.001 by ANOVA analysis) from the one generated by pure tMD simulations, particularly in the direction of PC3 (Figure 6D).

3.4.2. PMF Calculation and Conformational Change along the Path. Similar to the previous case, we evaluated the PMF profiles along the paths of COMBAS and tMD. Both PMF profiles suggest that MVI^C in the PPS conformation (reaction coordinate $\lambda = 0$, also called the lever-up state) is energetically more stable than the R conformation ($\lambda = 1$, also called the lever-down state) (Figure 7A). This observation is generally consistent with a previous study on a different myosin (myosin V) claiming that the power stroke (lever-up to lever-down) is an endergonic process because of the angular energy of the lever.[93,94] Notably, the free energy difference between the two states is about 4−5 kcal/mol for the COMBAS path, close to the minimum energy difference of ∼3.1 kcal/mol measured by Alhadeff and Warshel.[95] In contrast, the value is overestimated as ∼10 kcal/mol when using the tMD path, reinforcing that a proper initial path is essential for reliable thermodynamic evaluation.

The converter domain of MVI exhibits two prominent differences between the PPS and R states: (1) the orientation of helix 4 (H4), which inclines at about 45° to the β-sheet in the PPS state but becomes vertical and approximately perpendicular to the β-sheet in the R state, and (2) the conformation of loop 4 (L4), which is α-helical in the PPS state but becomes extended in the R state. We used the scalar angle between the centers of mass for residue 771, residue 763, and residues 755−758 to quantify the rotation of H4 during the transition (Figure 7C, left panel). Clearly, the H4 reorientation occurs earlier on the COMBAS path than on the tMD path (Figure 7B). Before the H4 rotation, at the reaction coordinate of $\lambda \approx 0.2$ of the COMBAS path, the L4 loop becomes slightly strained, allowing the formation of a hydrogen bond between Y718 and I769 (Figure 7C, upper panel), which coincides with the metastable intermediate state at the corresponding position on the PMF curve (Figure 7A, left panel). Besides the rotation of the H4 helix, major structural differences between images on the COMBAS and tMD paths lie in the L4 loop as well as the inter-helix interaction (mostly hydrogen bond) between Y718 of H1 and I769 of H4 (Figure 7C, lower panel). In Figure S6, we also borrowed the collective variables reported previously[71] (listed in Table S1) to further describe the difference of L4 between the two paths.

4. DISCUSSION

In this work, we developed a COMBAS method to quickly derive a feasible transition pathway between two fixed conformational states for a protein through relaxation-biasing iterations. After optimization using the string method, the path derived from COMBAS sampling enables rigorous free energy evaluation for the structural transition process. The overall pipeline could be conducted automatically, without any requirement of a priori knowledge for order parameter selection. When evaluated in small benchmark systems that have been sufficiently sampled by cMD previously, COMBAS sampling exhibits a comparable level of occupancy in the conformational space to the extremely long simulations executed on Anton. In the investigation of functional structural transition for more complicated protein systems including c-Src^C, MVI^C, and NtrC^R (see the Supporting Information and Figure S7), paths generated by COMBAS sampling are more physically reasonable than those generated by pure tMD simulations. Among the three cases, the superiority of the COMBAS path over the tMD path is highly remarkable for c-Src^C that has the largest structural deviation between its functional states, becomes weakened for MVI^C that has a moderate structural deviation, and diminishes for NtrC^R that has only minor structural difference. Simultaneously, the proportion of extra tMD simulations in COMBAS sampling increases in the same order. The reason is that COMBAS is more suitable for sampling the large-scale structural transition process because the optimal collective motion in each COMBAS cycle is obtained from slow tICs that correspond to global motions rather than local motions.

In this initial implementation of COMBAS, we chose tMD to fill the gap on the transition paths of MVI^C and NtrC^R, which is simple and fast but may result in a large deviation from the true MFEP. To overcome this problem, enhanced sampling techniques like aMD, FEXS, and SDS may be engaged to replace tMD in the future (see Figure S8 for our preliminary trial using aMD). Moreover, the convergence check of COMBAS sampling was conducted in the space of the top three PCs in this work but may need more PCs to describe the global motions for more complex protein systems, which may hinder visualization. As an alternative, we could substitute PCA by nonlinear dimensionality reduction methods like the multidimensional scaling.[96] In addition, instead of decomposing the cMD motions using tICA, we could use nonlinear methods such as kernel tICA[97] or neural network methods such as the recurrent neural network (RNN)[98] and the long short-term memory (LSTM)[99] network to capture the nonlinear transformation from time-series trajectories. This is also one of our future research directions.

Although we only used the Cartesian coordinates of Cα atoms for COMBAS sampling, various side chain conformations were also sampled, for example, the switch of the electrostatic network and the inward−outward−inward movement of Y416 in c-Src^C, accompanying the backbone movement. Thus, COMBAS sampling could provide a comprehensive description of the structural transition process. In a previous study,[86] extensive cMD simulations starting from an initial path of c-Src^C activation sampled by tMD in combination with MSM calculation successfully identified a key intermediate state as the target for drug design. Considering the improved initial path sampling by COMBAS and novel intermediate states (e.g., the Y416 outward conformation of c-Src^C) identified in this work, the construction of transition networks using MSM from our COMBAS path may allow the identification of new drug targets in the future. Although awaiting further validation, the COMBAS sampling proposed in this work may provide novel insights into the mechanistic study and drug design.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jctc.2c00390.

> 8 supporting figures (Figures S1−S8), 1 supporting table (Table S1), and supporting results for NtrC^R (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Haipeng Gong − *MOE Key Laboratory of Bioinformatics, School of Life Sciences and Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, China*; orcid.org/0000-0002-5532-1640; Email: hgong@tsinghua.edu.cn

Author

Yao Li − *MOE Key Laboratory of Bioinformatics, School of Life Sciences and Beijing Advanced Innovation Center for Structural Biology, Tsinghua University, Beijing 100084, China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jctc.2c00390

Author Contributions

The manuscript was written through contributions of both authors. Both authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank Dr. Meng Ke and Dr. Juanrong Zhang for discussion. This work has been supported by the National Natural Science Foundation of China (#32171243) and by the Beijing Advanced Innovation Center for Structural Biology. The authors gratefully thank D. E. Shaw Research for generously providing the reference cMD trajectories.

■ REFERENCES

(1) Karplus, M.; McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **2002**, *9*, 646−652.

(2) Clairfeuille, T.; Xu, H.; Koth, C. M.; Payandeh, J. Voltage-gated sodium channels viewed through a structural biology lens. *Curr. Opin. Struct. Biol.* **2017**, *45*, 74−84.

(3) Deng, D.; Sun, P.; Yan, C.; Ke, M.; Jiang, X.; Xiong, L.; Ren, W.; Hirata, K.; Yamamoto, M.; Fan, S.; Yan, N. Molecular basis of ligand recognition and transport by glucose transporters. *Nature* **2015**, *526*, 391−396.

(4) Kubitzki, M. B.; de Groot, B. L. The atomistic mechanism of conformational transition in adenylate kinase: a TEE-REX molecular dynamics study. *Structure* **2008**, *16*, 1175−1182.

(5) Cowan-Jacob, S. W.; Fendrich, G.; Manley, P. W.; Jahnke, W.; Fabbro, D.; Liebetanz, J.; Meyer, T. The crystal structure of a c-Src complex in an active conformation suggests possible steps in c-Src activation. *Structure* **2005**, *13*, 861−871.

(6) Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossváry, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **2008**, *51*, 91−97.

(7) Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L.-S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y.-H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Schafer, U. B.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE Press: New Orleans, Louisana, 2014; pp 41−53.

(8) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-level characterization of the structural dynamics of proteins. *Science* **2010**, *330*, 341−346.

(9) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517−520.

(10) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* **2017**, *13*, No. e1005659.

(11) Phillips, J. C.; Stone, J. E.; Schultent, K. Adapting a Message-Driven Parallel Application to GPU-Accelerated Clusters. In *International Conference for High Performance Computing, Networking, Storage and Analysis* 2008; p 444-+.

(12) Salomon-Ferrer, R.; Gotz, A. W.; Poole, D.; Le Grand, S.; Walker, R. C., Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. *J. Chem. Theory Comput.* **2013**, *9*, 3878−3888.

(13) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced sampling in molecular dynamics. *J. Chem. Phys.* **2019**, *151*, No. 070902.

(14) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919−11929.

(15) Salvalaglio, M.; Tiwary, P.; Parrinello, M. Assessing the Reliability of the Dynamics Reconstructed from Metadynamics. *J. Chem. Theory Comput.* **2014**, *10*, 1420−1425.

(16) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, No. 020603.

(17) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 12562−12566.

(18) Babin, V.; Roland, C.; Sagui, C. Adaptively biased molecular dynamics for free energy calculations. *J. Chem. Phys.* **2008**, *128*, 134101.

(19) Raiteri, P.; Laio, A.; Gervasio, F. L.; Micheletti, C.; Parrinello, M. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. *J. Phys. Chem. B* **2006**, *110*, 3533−3539.

(20) Darve, E.; Rodriguez-Gomez, D.; Pohorille, A. Adaptive biasing force method for scalar and vector free energy calculations. *J. Chem. Phys.* **2008**, *128*, 144120.

(21) Sultan, M. M.; Wayment-Steele, H. K.; Pande, V. S. Transferable Neural Networks for Enhanced Sampling of Protein Dynamics. *J. Chem. Theory Comput.* **2018**, *14*, 1887−1894.

(22) Hernandez, C. X.; Wayment-Steele, H. K.; Sultan, M. M.; Husic, B. E.; Pande, V. S. Variational encoding of complex dynamics. *Phys. Rev. E* **2018**, *97*, No. 062412.

(23) Lemke, T.; Peter, C. EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J. Chem. Theory Comput.* **2019**, *15*, 1209−1215.

(24) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.* **2018**, *148*, 241703.

(25) Mardt, A.; Pasquali, L.; Wu, H.; Noe, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5.

(26) Zhang, L.; Wang, H. Reinforced dynamics for enhanced sampling in large atomic and molecular systems. *J. Chem. Phys.* **2018**, *148*, 124113.

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

(28) Zhang, J.; Yang, Y. I.; Noe, F. Targeted Adversarial Learning Optimized Sampling. *J. Phys. Chem. Lett.* **2019**, *10*, 5791−5797.

(29) Zhang, J.; Gong, H. Frontier Expansion Sampling: A Method to Accelerate Conformational Search by Identifying Novel Seed Structures for Restart. *J. Chem. Theory Comput.* **2020**, *16*, 4813−4821.

(30) Harada, R.; Shigeta, Y. Efficient Conformational Search Based on Structural Dissimilarity Sampling: Applications for Reproducing Structural Transitions of Proteins. *J. Chem. Theory Comput.* **2017**, *13*, 1411−1423.

(31) Shkurti, A.; Styliari, I. D.; Balasubramanian, V.; Bethune, I.; Pedebos, C.; Jha, S.; Laughton, C. A. CoCo-MD: a simple and effective method for the enhanced sampling of conformational space. *J. Chem. Theory Comput.* **2019**, *15*, 2587−2596.

(32) Zuckerman, D. M. *Overview of Weighted Ensemble Simulation: Path-sampling, Steady States, Equilibrium*; Oregon Health & Science University, 2017.

(33) Donovan, R. M.; Tapia, J.-J.; Sullivan, D. P.; Faeder, J. R.; Murphy, R. F.; Dittrich, M.; Zuckerman, D. M. Unbiased rare event sampling in spatial stochastic systems biology models using a weighted ensemble of trajectories. *PLoS Comput. Biol.* **2016**, *12*, No. e1004611.

(34) Wu, D.; Wang, L.; Zhang, P. Solving statistical mechanics using variational autoregressive networks. *Phys. Rev. Lett.* **2019**, *122*, No. 080602.

(35) Noé, F.; Olsson, S.; Köhler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, No. aaw1147.

(36) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291−318.

(37) van Erp, T. S.; Bolhuis, P. G. Elaborating transition interface sampling methods. *J. Comput. Phys.* **2005**, *205*, 157−181.

(38) Van Erp, T. S.; Caremans, T. P.; Kirschhock, C. E.; Martens, J. A. Prospects of transition interface sampling simulations for the theoretical study of zeolite synthesis. *Phys. Chem. Chem. Phys.* **2007**, *9*, 1044−1051.

(39) Swenson, D. W.; Bolhuis, P. G. A replica exchange transition interface sampling method with multiple interface sets for investigating networks of rare events. *J. Chem. Phys.* **2014**, *141*, No. 044101.

(40) Cabriolu, R.; Skjelbred Refsnes, K. M.; Bolhuis, P. G.; van Erp, T. S. Foundations and latest advances in replica exchange transition interface sampling. *J. Chem. Phys.* **2017**, *147*, 152722.

(41) Faradjian, A. K.; Elber, R. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **2004**, *120*, 10880−10889.

(42) Allen, R. J.; Frenkel, D.; ten Wolde, P. R. Forward flux sampling-type schemes for simulating rare events: efficiency analysis. *J. Chem. Phys.* **2006**, *124*, 194111.

(43) Hussain, S.; Haji-Akbari, A. Studying rare events using forward-flux sampling: Recent breakthroughs and future outlook. *J. Chem. Phys.* **2020**, *152*, No. 060901.

(44) Yuan, Y.; Zhu, Q.; Song, R.; Ma, J.; Dong, H. A Two-Ended Data-Driven Accelerated Sampling Method for Exploring the Transition Pathways between Two Known States of Protein. *J. Chem. Theory Comput.* **2020**, *16*, 4631−4640.

(45) Bowman, G. R.; Huang, X.; Pande, V. S. Using generalized ensemble simulations and Markov state models to identify conformational states. *Methods* **2009**, *49*, 197−201.

(46) Prinz, J. H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schutte, C.; Noe, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(47) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *J. Am. Chem. Soc.* **2018**, *140*, 2386−2396.

(48) Weiss, D. R.; Levitt, M. Can morphing methods predict intermediate structures? *J. Mol. Biol.* **2009**, *385*, 665−674.

(49) Schlitter, J.; Engels, M.; Kruger, P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graphics* **1994**, *12*, 84−89.

(50) Lee, J.; Lee, I. H.; Joung, I.; Lee, J.; Brooks, B. R. Finding multiple reaction pathways via global optimization of action. *Nat. Commun.* **2017**, *8*, 15443.

(51) Mitsutake, A.; Sugita, Y.; Okamoto, Y. Generalized-ensemble algorithms for molecular simulations of biopolymers. *Biopolymers* **2001**, *60*, 96−123.

(52) Weinan, E.; Ren, W.; Vanden-Eijnden, E. String method for the study of rare events. *Phys. Rev. B* **2002**, *66*, No. 052301.

(53) Pan, A. C.; Sezer, D.; Roux, B. Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B* **2008**, *112*, 3432−3440.

(54) Weinan, E.; Ren, W.; Vanden-Eijnden, E. Finite temperature string method for the study of rare events. *J. Phys. Chem. B* **2005**, *109*, 6688−6693.

(55) Vanden-Eijnden, E.; Venturoli, M. Revisiting the finite temperature string method for the calculation of reaction tubes and free energies. *J. Chem. Phys.* **2009**, *130*, 194103.

(56) Gan, W.; Yang, S.; Roux, B. Atomistic view of the conformational activation of Src kinase using the string method with swarms-of-trajectories. *Biophys. J.* **2009**, *97*, L8−L10.

(57) Yang, Z.; Majek, P.; Bahar, I. Allosteric transitions of supramolecular systems explored by network models: application to chaperonin GroEL. *PLoS Comput. Biol.* **2009**, *5*, No. e1000360.

(58) Bahar, I.; Lezon, T. R.; Yang, L. W.; Eyal, E. Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* **2010**, *39*, 23−42.

(59) Adelman, J. L.; Grabe, M. Simulating rare events using a weighted ensemble-based string method. *J. Chem. Phys.* **2013**, *138*, No. 044105.

(60) Sultan, M. M.; Pande, V. S. tICA-Metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J. Chem. Theory Comput.* **2017**, *13*, 2440−2447.

(61) Schwantes, C. R.; Pande, V. S. Improvements in Markov State Model Construction Reveal Many Non-Native Interactions in the Folding of NTL9. *J. Chem. Theory Comput.* **2013**, *9*, 2000−2009.

(62) Perez-Hernandez, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noe, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, No. 015102.

(63) Xu, W.; Doshi, A.; Lei, M.; Eck, M. J.; Harrison, S. C. Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol. Cell* **1999**, *3*, 629−638.

(64) Ménétrey, J.; Bahloul, A.; Wells, A. L.; Yengo, C. M.; Morris, C. A.; Sweeney, H. L.; Houdusse, A. The structure of the myosin VI motor reveals the mechanism of directionality reversal. *Nature* **2005**, *435*, 779−785.

(65) Ménétrey, J.; Llinas, P.; Mukherjea, M.; Sweeney, H. L.; Houdusse, A. The structural basis for the large powerstroke of myosin VI. *Cell* **2007**, *131*, 300−308.

(66) Hastie, T.; Stuetzle, W. Principal curves. *J. Am. Stat. Assoc.* **1989**, *84*, 502−516.

(67) Zhu, L.; Sheong, F. K.; Cao, S.; Liu, S.; Unarta, I. C.; Huang, X. TAPS: A traveling-salesman based automated path searching method for functional conformational changes of biological macromolecules. *J. Chem. Phys.* **2019**, *150*, 124105.

(68) Xi, K.; Hu, Z.; Wu, Q.; Wei, M.; Qian, R.; Zhu, L. Assessing the Performance of Traveling-salesman based Automated Path Searching (TAPS) on Complex Biomolecular Systems. *J. Chem. Theory Comput.* **2021**, *17*, 5301−5311.

(69) Moradi, M.; Enkavi, G.; Tajkhorshid, E. Atomic-level characterization of transport cycle thermodynamics in the glycerol-3-phosphate:phosphate antiporter. *Nat. Commun.* **2015**, *6*, 8393.

(70) Ke, M.; Yuan, Y.; Jiang, X.; Yan, N.; Gong, H. Molecular determinants for the thermodynamic and functional divergence of uniporter GLUT1 and proton symporter XylE. *PLoS Comput. Biol.* **2017**, *13*, No. e1005603.

(71) Ovchinnikov, V.; Karplus, M.; Vanden-Eijnden, E. Free energy of conformational transition paths in biomolecules: the string method and its application to myosin VI. *J. Chem. Phys.* **2011**, *134*, No. 085103.

(72) Song, H. D.; Zhu, F. Finite Temperature String Method with Umbrella Sampling: Application on a Side Chain Flipping in Mhp1 Transporter. *J. Phys. Chem. B* **2017**, *121*, 3376−3386.

(73) Habeck, M. Bayesian estimation of free energies from equilibrium simulations. *Phys. Rev. Lett.* **2012**, *109*, No. 100601.

(74) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.

(75) Case, D. A.; Cheatham, T. E., 3rd; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M., Jr.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.

(76) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Henin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kale, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, No. 044130.

(77) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernandez, C. X.; Schwantes, C. R.; Wang, L. P.; Lane, T. J.; Pande, V. S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528−1532.

(78) Harrigan, M. P.; Sultan, M. M.; Hernandez, C. X.; Husic, B. E.; Eastman, P.; Schwantes, C. R.; Beauchamp, K. A.; McGibbon, R. T.; Pande, V. S. MSMBuilder: Statistical Models for Biomolecular Dynamics. *Biophys. J.* **2017**, *112*, 10−15.

(79) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(80) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467. 2016.

(81) Ovchinnikov, V.; Karplus, M. Analysis and elimination of a bias in targeted molecular dynamics simulations of conformational transitions: application to calmodulin. *J. Phys. Chem. B* **2012**, *116*, 8584−8603.

(82) Lev, B.; Murail, S.; Poitevin, F.; Cromer, B. A.; Baaden, M.; Delarue, M.; Allen, T. W. String method solution of the gating pathways for a pentameric ligand-gated ion channel. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E4158−E4167.

(83) Parsons, S. J.; Parsons, J. T. Src family kinases, key regulators of signal transduction. *Oncogene* **2004**, *23*, 7906−7909.

(84) Mugnai, M. L.; Thirumalai, D. Kinematics of the lever arm swing in myosin VI. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E4389−E4398.

(85) Stock, A. M.; Robinson, V. L.; Goudreau, P. N. Two-component signal transduction. *Annu. Rev. Biochem.* **2000**, *69*, 183−215.

(86) Shukla, D.; Meng, Y.; Roux, B.; Pande, V. S. Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nat. Commun.* **2014**, *5*, 3397.

(87) Sultan, M. M.; Kiss, G.; Pande, V. S. Towards simple kinetic models of functional dynamics for a kinase subfamily. *Nat. Chem.* **2018**, *10*, 903−909.

(88) Wu, H.; Huang, H.; Post, C. B. All-atom adaptively biased path optimization of Src kinase conformational inactivation: Switched electrostatic network in the concerted motion of alphaC helix and the activation loop. *J. Chem. Phys.* **2020**, *153*, 175101.

(89) Ozkirimli, E.; Post, C. B. Src kinase activation: A switched electrostatic network. *Protein Sci.* **2006**, *15*, 1051−1062.

(90) Li, P.; Martins, I. R.; Amarasinghe, G. K.; Rosen, M. K. Internal dynamics control activation and activity of the autoinhibited Vav DH domain. *Nat. Struct. Mol. Biol.* **2008**, *15*, 613−618.

(91) Tsytlonok, M.; Sanabria, H.; Wang, Y.; Felekyan, S.; Hemmen, K.; Phillips, A. H.; Yun, M. K.; Waddell, M. B.; Park, C. G.; Vaithiyalingam, S.; Iconaru, L.; White, S. W.; Tompa, P.; Seidel, C. A. M.; Kriwacki, R. Dynamic anticipation by Cdk2/Cyclin A-bound p27 mediates signal integration in cell cycle regulation. *Nat. Commun.* **2019**, *10*, 1676.

(92) Henriques, J.; Lindorff-Larsen, K. Protein Dynamics Enables Phosphorylation of Buried Residues in Cdk2/Cyclin-A-Bound p27. *Biophys. J.* **2020**, *119*, 2010−2018.

(93) Mukherjee, S.; Warshel, A. Electrostatic origin of the unidirectionality of walking myosin V motors. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17326−17331.

(94) Mukherjee, S.; Alhadeff, R.; Warshel, A. Simulating the dynamics of the mechanochemical cycle of myosin-V. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 2259−2264.

(95) Alhadeff, R.; Warshel, A. Reexamining the origin of the directionality of myosin V. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 10426−10431.

(96) Mead, A. Review of the Development of Multidimensional Scaling Methods. *J. R. Stat. Soc., Ser. D* **1992**, *41*, 27−39.

(97) Schwantes, C. R.; Pande, V. S. Modeling molecular kinetics with tICA and the kernel trick. *J. Chem. Theory Comput.* **2015**, *11*, 600−608.

(98) Graves, A.; Schmidhuber, J. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2008; Vol. *21*, pp 545−552.

(99) Tealab, A. Time series forecasting using artificial neural networks methodologies: A systematic review. *Fut. Comput. Inf. J.* **2018**, *3*, 334−340.