# Free energy of conformational transition paths in biomolecules: The string method and its application to myosin VI

Victor Ovchinnikov,[1,2] Martin Karplus,[1,3,a)] and Eric Vanden-Eijnden[4,b)]
[1]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138, USA*
[2]*Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA*
[3]*Laboratoire de Chimie Biophysique, ISIS, Université de Strasbourg, 67000 Strasbourg, France*
[4]*Courant Institute of Mathematical Sciences, New York University, New York, New York 10012, USA*

(Received 25 August 2010; accepted 31 December 2010; published online 23 February 2011)

A set of techniques developed under the umbrella of the string method is used in combination with all-atom molecular dynamics simulations to analyze the conformation change between the prepowerstroke (PPS) and rigor (R) structures of the converter domain of myosin VI. The challenges specific to the application of these techniques to such a large and complex biomolecule are addressed in detail. These challenges include (i) identifying a proper set of collective variables to apply the string method, (ii) finding a suitable initial string, (iii) obtaining converged profiles of the free energy along the transition path, (iv) validating and interpreting the free energy profiles, and (v) computing the mean first passage time of the transition. A detailed description of the PPS↔R transition in the converter domain of myosin VI is obtained, including the transition path, the free energy along the path, and the rates of interconversion. The methodology developed here is expected to be useful more generally in studies of conformational transitions in complex biomolecules. © *2011 American Institute of Physics*. [doi:10.1063/1.3544209]

## I. INTRODUCTION

The analysis of large-scale conformational changes in biomolecules is one of the most challenging problems in experimental and computational chemistry. Despite encouraging advances in computer speed, direct observation of such conformational changes in conventional molecular dynamics (MD) simulation is impossible in most cases, because it would require integration times that are orders of magnitude beyond the reach of most available computers. To confront this difficulty, several accelerated sampling techniques have been developed, such as transition path sampling (TPS),[1,2] the string method,[3–5] metadynamics,[6,7] adaptive biasing force,[8] and milestoning.[9,10] In each of these methods, the ultimate goal is to find a detailed and unbiased description of the transition without *a priori* assumptions. Unfortunately, such assumptions are often unavoidable [e.g., one must choose an initial path, collective variables (CVs), or the resolution of the transition path]. In the present paper, we address some of these assumptions in the context of the string method applied to a complex biomolecular system at atomic resolution.[11–13]

Specifically, we use the string method to find the most probable transition path between the prepowerstroke (PPS) and rigor (R) conformers of the converter domain of myosin VI (MVI),[11–13] compute the free energy (FE) profile along the transition path, and estimate the rate of interconversion between the two conformers. The present example illustrates generic challenges that are likely to arise in the application of the string method to large biomolecular systems and provides solutions that can be implemented with present-day computers. We therefore expect that the methodology developed herein will extend the applicability of the string method to a wide class of biomolecules.

The string method provides a representative path for the transition, i.e., the minimum free energy path (MFEP). If certain conditions are met, as described in Sec. II, the MFEP lies at the center of a tube in which the transition is most likely to occur.[14,15] The MFEP is much smoother than a trajectory (such as those found by TPS) and is likely to be more informative about the mechanism of the transition, because it averages out the motions that are unimportant in the transition. The elimination of the unimportant degrees of freedom is achieved because the MFEP is computed in a reduced space of the collective variables that are essential for describing the transition; such collective variables can be center-of-mass (COM) positions of groups of selected atoms, distances between such groups, bond or dihedral angles, etc. The introduction of a reduced collective variable space is necessary to justify the assumptions implicit in the string method and to make the calculation of MFEP affordable. In addition, unlike methods such as metadynamics that also use collective variables but require their number to be rather small, the string method can be used with large sets of collective variables (hundreds or more[16]). Although the string method has been applied to a number of problems,[5,16–18] the present study treats a more complex system, which requires the resolution of issues beyond those encountered in previous applications. For a detailed summary of the methods used in this study and the results obtained by their application to myosin VI, the reader is encouraged to read Secs. IV and V, respectively, before the main body of the paper.

a)Electronic mail: marci@tammy.harvard.edu.
b)Electronic mail: eve2@cims.nyu.edu.

 

Sections II and III describe the methodology and its implementation for the myosin VI converter in detail. In Sec. II we review the general techniques employed in this paper: collective variables are introduced in Sec. II A, the MFEP is discussed in Sec. II B, the string method in collective variables is summarized in Sec. II C, and the calculation of the free energy and the rate of the transition is explained in Sec. II D. Section III addresses the application of the methods to MVI. Section III A discusses the preparation of MD simulation structures. Section III B explains the use of targeted MD simulations to choose collective variables (two sets are chosen in order to ascertain that the free energy profiles and rates calculated are insensitive to the choice of collective variables). Section III C describes the initial conditions used in the string simulations. In Sec. III D we present the transition mechanism together with the corresponding free energy profiles and transition rates. Section IV summarizes the accomplishments and the limitations of the present method, and in Sec. V we highlight key results from the myosin VI simulations.

Additional validation of the present methodology using the simple test case of the alanine dipeptide in vacuum is given in the supplementary materials.[19]

## II. METHODOLOGY

The calculation by the string method of the free energy and the reaction rate associated with an optimized transition path proceed in several steps. First, a set of collective variables that are sufficient to describe the transition are selected. Initial values for the these collective variables (initial "string" of replicas) are determined from a minimum energy path (MEP) calculated between the endpoint structures using the zero-temperature string method in the full space of Cartesian atom positions.[3,20] With the string method in collective variables[5] and starting from the initial string, the replicas are allowed to "relax" in the direction of the negative gradient of the free energy, with the arc-length between adjacent replicas kept approximately constant. The final converged string corresponds to an MFEP between the endpoint states in the collective variable space. Starting from the MFEP, the free energy profile and rate of the reaction can be calculated using the finite-temperature string method[21] and milestoning,[9,18] respectively. Each step of the methodology is outlined below. Because the zero-temperature string method is conceptually similar to the string method in collective variables, it is summarized in the supplementary materials.[19] Appendix A describes some technical calculations. Appendix B provides additional discussion of sources of errors in the finite-temperature string method, and in Appendix C we present a validation of the transition state obtained from one of the present simulations using an ensemble of unbiased trajectories.

### A. Collective variables

An essential aspect of the present method is the selection of a set of collective variables that are appropriate for describing the transition of interest. Collective variables are scalar functions of the atomic coordinates of the system that characterize its state at a coarse-grained level. Examples are the Cartesian positions of representative atoms, the positions of the center of masses of groups of atoms, dihedral angles, and interatomic or atomic-group distances. The starting point for finding a suitable set of collective variables is the assumption that the transition path can be described by specifying the positions of a relatively small number of atoms. Restrained targeted molecular dynamics (RTMD) (Ref. 22) is used to find a small set of atoms, such that applying RTMD forces to the atoms in this set steers the converter structure from one conformation to the other. This set of atoms, denoted the "resolving set (RS)," is then used to define CVs. Letting $\boldsymbol{x}$ denote the Cartesian positions of all the atoms in the system, we identify a set of $K$ CVs, which we denote by $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = (\widehat{\theta}_1(\boldsymbol{x}), \widehat{\theta}_2(\boldsymbol{x}), \ldots, \widehat{\theta}_K(\boldsymbol{x}))$.

The identification of the resolving set and the construction of collective variables for the converter of MVI are described in detail in Sec. III B. The specific CVs sets used are given in Tables II and III.

### B. The minimum free energy path and its interpretation

An MFEP is a path of steepest descent on the free energy surface associated with the collective variables, scaled by a metric tensor that arises from the curvilinear nature of the collective variables, and guarantees that the location of the MFEP is invariant to nonlinear transformations of these variables.[5] More precisely,

> *An MFEP is a path in collective variable space connecting two local minima of* $G(\boldsymbol{\theta})$ *to which the vector* $M(\boldsymbol{\theta})\nabla G(\boldsymbol{\theta})$ *is everywhere tangent.* (1)

In Eq. (1), $M(\boldsymbol{\theta})\nabla G(\boldsymbol{\theta})$ denotes the vector with Cartesian components,

$$\sum_{j=1}^{K} M_{i,j}(\boldsymbol{\theta}) \frac{\partial G(\boldsymbol{\theta})}{\partial \theta_j}, \quad i = 1, 2, \ldots, K, \tag{2}$$

$G(\boldsymbol{\theta})$ is the free energy associated with the collective variables,

$$G(\boldsymbol{\theta}) = -\beta^{-1} \ln \left\langle \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\boldsymbol{x})) \right\rangle, \tag{3}$$

and $M(\boldsymbol{\theta})$ is a tensor given by

$$M_{i,j}(\boldsymbol{\theta}) = \sum_k \frac{1}{m_k} \left\langle \frac{\partial \widehat{\theta}_i(\boldsymbol{x})}{\partial x_k} \frac{\partial \widehat{\theta}_j(\boldsymbol{x})}{\partial x_k} \right\rangle_{\widehat{\boldsymbol{\theta}}(\boldsymbol{x})=\boldsymbol{\theta}}. \tag{4}$$

In Eq. (3), $\beta = 1/k_B T$, where $k_B$ is Boltzmann's constant and $T$ is the temperature, and $\langle\cdot\rangle$ denotes canonical average; in Eq. (4), $m_k$ are the masses of the atoms, the sum is taken over all the coordinates of all the atoms in the system, and $\langle\cdot\rangle_{\widehat{\boldsymbol{\theta}}(\boldsymbol{x})=\boldsymbol{\theta}}$ denotes the conditional average $\langle(\cdot)\delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\boldsymbol{x}))\rangle / \langle\delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\boldsymbol{x}))\rangle$. The estimation of $G(\boldsymbol{\theta})$ and $M(\boldsymbol{\theta})$ in the string method and the calculation of an MFEP are explained in Sec. II C. The significance of the MFEP defined

here is established in transition path theory (TPT).[14,15,23] TPT analyzes the statistical mechanics properties of the reactive trajectories—those by which the transitions from one end-point structure to the other actually happen—and gives expressions for the probability density and current of these trajectories.

Central in the expressions for both the probability density and the current is the committor function $q(\boldsymbol{x}, \boldsymbol{p})$ ($p_{fold}$ in protein folding studies[24]) which gives the probability that, if one initializes the system at position $\boldsymbol{x}$ with momentum $\boldsymbol{p}$, the system will go to one end-point structure (the product) rather than the other (the reactant). In principle, the function $q(\boldsymbol{x}, \boldsymbol{p})$ is an ideal reaction coordinate for describing a transition. However, $q(\boldsymbol{x}, \boldsymbol{p})$, per se, gives limited insight into the transition mechanism because it does not provide direct information on the essential variables that govern the transition.[24–26] Although $q(\boldsymbol{x}, \boldsymbol{p})$ satisfies a closed-form equation of Liouville or Fokker-Planck type, this equation cannot be solved directly in high dimensional systems. The equation for $q(\boldsymbol{x}, \boldsymbol{p})$ can, however, be used as the basis for meaningful approximations. Under the assumptions,

(i) the committor function can be parametrized approximately as a function of the collective variables, i.e., $q(\boldsymbol{x}, \boldsymbol{p}) \approx Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}))$ for some function $Q$, and
(ii) projected in the space of collective variables, most of the probability flux of the reactive trajectories goes through one narrow channel (or a few channels separated by barriers much higher than $k_B T$), referred to as transition tube(s),

it can be shown that, for the MFEPs defined by Eq. (1),[5,14,15,23]

(1) the MFEPs lie at the center of the transition tubes;
(2) locally around the MFEP, the isosurfaces of the committor function (isocommittor surfaces) can be approximated by the surfaces defined by

$$\sum_{i,j=1}^{K} \theta_i'(\alpha) M_{i,j}^{-1}(\boldsymbol{\theta}(\alpha))(\widehat{\theta}_j(\boldsymbol{x}) - \theta_j(\alpha)) = 0; \tag{5}$$

(3) the value of the committor function along the MFEP is

$$Q(\boldsymbol{\theta}(\alpha)) = \frac{\int_0^\alpha m(\alpha^*) e^{\beta F(\alpha^*)} d\alpha^*}{\int_0^1 m(\alpha^*) e^{\beta F(\alpha^*)} d\alpha^*}. \tag{6}$$

In Eqs. (5) and (6), $\boldsymbol{\theta}(\alpha)$ with $\alpha \in [0, 1]$ denotes a parametrization of the MFEP [i.e., for every $\alpha \in [0, 1]$, $\boldsymbol{\theta}(\alpha)$ is a point along the MFEP]; the prime denotes derivative with respect to $\alpha$; $M_{i,j}^{-1}(\boldsymbol{\theta})$ are the elements of the matrix $M^{-1}(\boldsymbol{\theta})$; the scalar $m(\alpha)$ is defined as

$$m(\alpha) = \sum_{i,j=1}^{K} \theta_i'(\alpha) M_{i,j}^{-1}(\boldsymbol{\theta}(\alpha)) \theta_j'(\alpha) \tag{7}$$

and $F(\alpha)$ is the free energy defined as

$$F(\alpha) = -\beta^{-1} \ln \langle \delta(g(\boldsymbol{x}, \alpha)) \rangle, \tag{8}$$

where $g(\boldsymbol{x}, \alpha)$ is a shorthand notation for the left hand side of Eq. (5),

$$g(\boldsymbol{x}, \alpha) = \sum_{i,j=1}^{K} \theta_i'(\alpha) M_{i,j}^{-1}(\boldsymbol{\theta}(\alpha))(\widehat{\theta}_j(\boldsymbol{x}) - \theta_j(\alpha)). \tag{9}$$

The free energy $F(\alpha)$ [not to be confused with the free energy as a function of the collective variables $G(\boldsymbol{\theta})$ defined in Eq. (3)] is the free energy as a function of the committor used as the reaction coordinate, since, by Eq. (5), $g(\boldsymbol{x}, \alpha) = 0$ approximates the isocommittor surface on which $Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{x})) = Q(\boldsymbol{\theta}(\alpha))$. As such, $F(\alpha)$ plotted vs $Q(\boldsymbol{\theta}(\alpha))$ should be insensitive to the choice of collective variables, provided only that these variables are adequate; i.e., that $Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}))$ is a good approximation of the actual committor function $q(\boldsymbol{x}, \boldsymbol{p})$. The free energy $F$ is a one-dimensional function, whereas $G(\boldsymbol{\theta})$ of Eq. (3) is $K$-dimensional, with $K$ equal to the number of collective variables. Profiles of $G$ are evaluated only on the corresponding MFEP and have only local information about the free energy values for points not on the MFEP (see conditions 1 and 2 above). In contrast to $G$, the reaction free energy $F$ maps the entire transition tube onto a single curve. For this reason, we define the function $F(\alpha)$ as the free energy profile of the reaction. In the special case that the transition tube is extremely narrow, or has uniform cross-sectional volume along the path, $G \simeq F$. We will show that this is not the case in the present study (see Sec. III D 5). Additional details on the calculation of free energy profiles and rates of transition can be found in the TPT papers[14,15,23] and in Ref. 5. An alternative method to compute one-dimensional free energy profiles was proposed by Krivov and Karplus.[27,28] The method uses long MD trajectories during which multiple transition events are observed, in combination with the minimum-cut procedure (see Appendix D).

Assumptions (i) and (ii) above formalize the property that the collective variables are "good" variables to describe the transition. In principle, assumption (i) can be checked *a posteriori* by a committor test. This test amounts to launching trajectories from the isosurface on which $Q(\widehat{\boldsymbol{\theta}}(\boldsymbol{x})) = 0.5$ and checking that these trajectories "commit" to the two endpoint structures with equal probability. The committor test can be computationally too costly for diffusive systems because (1) trajectories take a very long time to commit (e.g., micro- to milliseconds) and (2) a very large number of trajectories is needed to ensure a good statistical sample. For these reasons, we are unable to perform a full committor test for the present calculations. Instead, we perform a less stringent test: starting from configurations chosen randomly from a putative transition state ensemble, we launch a collection of relatively short unbiased MD trajectories and examine the distribution of the reaction coordinate values at the end of the simulations. We find that 41% and 59% of the trajectories terminate on the reactant and product sides along the reaction coordinate, respectively, although none of the trajectories reach the reactants or the products. The above "splitting" probability is considered sufficiently close to the optimal 50%/50%, because the committor values are known to change rapidly in the vicinity of a transition state.[29] Details of the calculations can be found in Appendix C.

In addition, we resort to an indirect qualitative strategy to validate assumption (i). We compute the reaction free energy, Eq. (8), using two different sets of collective variables. If the results are similar for the two sets, as we find for the present study (see Sec. III D 4), it is likely that either set is adequate to describe the transition. Note that, in contrast to $F(\alpha)$, the free energy $G$ in Eq. (3) depends on the choice of collective variables and is more difficult to use for cross-validation.

Assumption (ii) can be tested *a posteriori* by examining the transition tubes corresponding to the MFEPs, as discussed in Sec. III D.

### C. String method in collective variables

The string method in collective variables was introduced in Ref. 5. It has been applied to an isomerization reaction in the alanine dipeptide in vacuum[5] and in explicit solvent,[17] the insertion of a coarse-grained model of a protein into a lipid bilayer,[18] and to the collapse of a hydrophobic chain of beads, in which water molecules were represented explicitly and hundreds of thousands of collective variables were used.[16] A number of closely related variants of the method were introduced recently.[30–33] Here we give a brief account of the method we use and refer the interested reader to Refs. 5 and 17 for more details.

The string method in collective variable space is a generalization of the string method in the Cartesian space.[3,20] The essence of the method is to evolve a curve—the string—using $-M(\boldsymbol{\theta})\nabla G(\boldsymbol{\theta})$ as a force while maintaining a prescribed parametrization of the curve. If we parametrize the string as $\boldsymbol{\theta}(\alpha)$ with $\alpha \in [0, 1]$, this evolution can be written as

$$\gamma\dot{\boldsymbol{\theta}} = -M(\boldsymbol{\theta})\nabla G(\boldsymbol{\theta}) + \lambda\boldsymbol{\theta}', \tag{10}$$

where $\gamma$ is an adjustable friction coefficient (discussed below), the dot and the prime denote differentiation with the respect to $t$and $\alpha$, respectively, and $\lambda\boldsymbol{\theta}'$ is a Lagrange multiplier term added to enforce a specific parametrization of the string (in the present case, that the string have uniform arc-length increments; i.e., $|\boldsymbol{\theta}'| =$ constant). Note that the steady-state solution of Eq. (10) satisfies $\lambda\boldsymbol{\theta}' = -M(\boldsymbol{\theta})\nabla G(\boldsymbol{\theta})$, i.e., it is an MFEP according to the definition of Eq. (1).

To integrate Eq. (10), the string is discretized into $N + 1$ representative "images:"

$$\boldsymbol{\theta}_n = \boldsymbol{\theta}(n/N), \qquad n = 0, 1, \ldots, N. \tag{11}$$

At each iteration, these images are evolved in the following two steps:

(1) *Evolution step.* Each image is evolved independently of the others using

$$\boldsymbol{\theta}_n(t+\Delta t) = \boldsymbol{\theta}_n(t) - \gamma^{-1}\Delta t\, M(\boldsymbol{\theta}_n(t))\nabla G(\boldsymbol{\theta}_n(t)). \tag{12}$$

(2) *Reparametrization step.* The images are reparametrized to enforce equal arc-length increments, which requires that

$$|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n| = |\boldsymbol{\theta}_n - \boldsymbol{\theta}_{n-1}|, \quad n = 1, \ldots, N-1, \tag{13}$$

where $|\cdot|$ denotes the Euclidean norm.

The evolution step Eq. (12) uses a forward Euler discretization of Eq. (10) with the Lagrange multiplier term neglected. [The computation of $M(\boldsymbol{\theta}_n(t))\nabla G(\boldsymbol{\theta}_n(t))$ is described below, and the numerical values for $\Delta t$ and $\gamma$ are specified in Sec. III D 1.]

The reparametrization step represents the action of the Lagrange multiplier term in Eq. (10). It is performed in two steps. First, we calculate the piecewise linear function $\ell(\alpha)$, such that $\ell(0) = 0$ and its values at $\alpha = n/N$ with $n = 1, \ldots, N$ are

$$\ell(\frac{n}{N}) = \sum_{m=1}^{n} |\boldsymbol{\theta}_m - \boldsymbol{\theta}_{m-1}|. \tag{14}$$

Second, we compute the images at new parameter values specified by

$$\ell(\alpha_n) = \frac{n}{N}\ell(1) \tag{15}$$

using linear interpolation. These new images satisfy Eq. (13) approximately. Although arbitrary accuracy can be achieved by applying the reparametrization step iteratively, we only used one iteration, after which the constraint was satisfied to greater than 1% accuracy.

For a string that is far from the MFEP (as would be observed in early stages of the string calculation), the reparametrization correction is small compared to the evolution step. On the other hand, after the string has converged to the MFEP, the reparametrization correction exactly cancels the evolution step, so that the string images remain stationary.

The evaluation of the terms $M(\boldsymbol{\theta}_n)$ and $\nabla G(\boldsymbol{\theta}_n)$ is performed using MD sampling with restraints as follows. To each image $\boldsymbol{\theta}_n$, we assign an independent all-atom replica of the system, which we simulate by MD with the restraining potential

$$U(\boldsymbol{x}, \boldsymbol{\theta}_n) = \frac{1}{2}\sum_{i=1}^{K} k_i(\widehat{\theta}_i(\boldsymbol{x}) - \theta_{n,i})^2. \tag{16}$$

The force constants $k_i$ should be chosen such that the constraint $\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}_n$ is approximately satisfied during the MD simulation. In practice, however, it may be advantageous to use low force constants to accelerate equilibration of the replicas during MD. In particular, low force constants permit the use of Hamiltonian replica exchange (REX) (Refs. 34 and 35) to increase conformational sampling, as will be illustrated in Sec. III D 3. The use of low force constants, however, requires correcting the free energy gradients computed from the simulations, e.g., with the umbrella integration (UI) method.[36]

Denoting by $\boldsymbol{x}_n(t)$ the MD trajectory of the replica assigned to image $\boldsymbol{\theta}_n$, the $i$th component of the free energy gradient $\nabla G(\boldsymbol{\theta}_n(t))$ is estimated as

$$-\frac{1}{\Delta t}\int_t^{t+\Delta t} k_i(\widehat{\theta}_i(\boldsymbol{x}_n(t')) - \theta_{n,i}(t))dt' \tag{17}$$

and the $(i, j)$th entry of the tensor $M(\boldsymbol{\theta}_n(t))$, as

$$\frac{1}{\Delta t}\int_t^{t+\Delta t}\sum_{k=1}^{n}\frac{1}{m_k}\frac{\partial\widehat{\theta}_i(\boldsymbol{x}_n(t'))}{\partial x_k}\frac{\partial\widehat{\theta}_j(\boldsymbol{x}_n(t'))}{\partial x_k}dt'. \tag{18}$$

Equations (12), (17), (18) suggest that $\gamma^{-1}\Delta t$ must be small enough so that the solution of Eq. (10) is stable and accurate; yet, $\Delta t$ must be large enough so that the time averages in Eqs. (17) and (18) are converged. This compromise can be achieved by choosing the friction coefficient $\gamma$ sufficiently large. It was shown in Ref. 17 that for large values of $\gamma$, the integration step $\Delta t$ can be made as low as that of a single MD evolution step (1-2fs) without compromising the accuracy of the MFEP. For $\gamma$ large enough, the images $\boldsymbol{\theta}_n$ evolve much more slowly than the corresponding MD replicas and effectively "feel," via Eqs. (12), (17), and (18), the average effect of the MD replica. The averages in Eqs. (17) and (18) need not be converged to attain convergence of Eq. (12) to the MFEP. Furthermore, even for somewhat smaller values of $\gamma$ (with $\Delta t$sufficiently small), the steady-state solution of Eq. (12) will oscillate around the MFEP, but will not diverge from it. The approximate values for $\gamma$ and $\Delta t$ used in the present simulations are 1500 $\text{ps}^{-1}$ and 20 fs, respectively (see Sec. III D 1).

In the present simulations, the averages in Eqs. (17) and (18) are computed using either 10 or 15 MD steps, depending on the simulation (see Sec. III). The images are updated according to Eq. (10), and the string is reparametrized. Evolving and reparametrizing the string once in every several MD iterations only incurs a small additional cost compared to a regular MD simulation. Moreover, the MD replicas can be evolved in parallel using different sets of central processing units (CPUs) for each.

Convergence of the string to the MFEP is assessed by monitoring the quantity $D(t)$ defined as

$$D(t) = \left[\frac{\sum_{n=0}^{N} |\boldsymbol{\theta}_n(t) - \boldsymbol{\theta}_n(0)|^2}{K(N+1)}\right]^{1/2}, \tag{19}$$

which measures the root-mean-square distance (RMSD) by which the images along the string have moved from their initial positions. Convergence is assumed after $D(t)$ reaches a plateau (see Sec. III).

### D. Calculation of free energies and mean first passage times (MFPTs)

#### 1. Free energy of the collective variables along the MFEP

Once the string has converged to the MFEP, the images $\boldsymbol{\theta}_n$ are fixed, and we can compute the approximate gradient of the free energy, Eq. (3), at these images using

$$\frac{\partial G(\boldsymbol{\theta}_n)}{\partial \theta_i} = -\frac{1}{T}\int_0^T k_i(\widehat{\theta}_i(\boldsymbol{x}_n(t)) - \theta_{n,i})dt, \tag{20}$$

where $\boldsymbol{x}_n(t)$ denotes the trajectory of the MD replica evolving in the potential with the restraint term of Eq. (16) and $T$ is taken large enough for the average to converge. The free energy $G(\boldsymbol{\theta})$, relative to the endpoint value $G(\boldsymbol{\theta}_0)$ along the MFEP, can then be obtained by numerical integration. The trapezoidal rule yields

$$G(\boldsymbol{\theta}_n) - G(\boldsymbol{\theta}_0) = \sum_{m=1}^{n} \frac{\nabla G(\boldsymbol{\theta}_m) + \nabla G(\boldsymbol{\theta}_{m-1})}{2}(\boldsymbol{\theta}_m - \boldsymbol{\theta}_{m-1}). \tag{21}$$

It is shown in Ref. 5 that the errors due to the use of a restraint instead of a constraint in the approximation, Eq. (20), are of the order $(\beta k_i)^{-1}$ (see also Sec. III D 3). Thus, if the force constants $k_i$ are chosen sufficiently large, Eq. (21) will yield an accurate free energy profile. In the present application, we found that convergence was difficult to achieve using Eq. (20) as written, and Hamiltonian replica exchange was employed with the MD simulations to improve sampling.

The REX employed herein is based on the replica-exchange umbrella sampling (US) algorithm, in which umbrella potentials corresponding to adjacent "windows" of a progress coordinate are exchanged "on-the-fly" during the simulation.[34] Unlike the implementations in Refs. 34, and 37–39, which implement one-dimensional US, we use the multidimensional biasing potential of Eq. (16). The number of windows equals the number of string replicas, and the windows are centered on the string images $\boldsymbol{\theta}_n$. To use REX, the restrained MD simulations must be performed concurrently for each image, so that exchange moves between adjacent images may be attempted. Specifically, after a prescribed number of iterations, we attempt to switch the atomic coordinates of the MD replicas $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n+1}$ that correspond to the neighboring images $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_{n+1}$, respectively,

$$\{\ldots, (\boldsymbol{x}_n;\boldsymbol{\theta}_n), (\boldsymbol{x}_{n+1};\boldsymbol{\theta}_{n+1}), \ldots\} \rightarrow \{\ldots, (\boldsymbol{x}_{n+1};\boldsymbol{\theta}_n), (\boldsymbol{x}_n;\boldsymbol{\theta}_{n+1}), \ldots\}, \tag{22}$$

where $n$ is chosen randomly from $0, 1, \ldots, N$. The acceptance probability of the move in Eq. (22) is computed according to the Metropolis criterion,

$$p_{\text{acc}}(\boldsymbol{x}_n \leftrightarrow \boldsymbol{x}_{n+1}) = \begin{cases} 1 & \text{if } \Delta \le 0, \\ \exp(-\Delta) & \text{if } \Delta > 0, \end{cases}$$
$$\Delta = \beta[U(\boldsymbol{x}_{n+1},\boldsymbol{\theta}_n) + U(\boldsymbol{x}_n,\boldsymbol{\theta}_{n+1}) - U(\boldsymbol{x}_n,\boldsymbol{\theta}_n) - U(\boldsymbol{x}_{n+1},\boldsymbol{\theta}_{n+1})]. \tag{23}$$

In Eq. (23), $U(\boldsymbol{x},\boldsymbol{\theta})$ is the restraining potential defined in Eq. (16) and $p_{\text{acc}}$ is the probability of accepting the move in Eq. (22). To achieve high-enough acceptance rates in REX, the force constants had to be lowered so that the conformational ensembles corresponding to adjacent images have sufficient overlap. To correct the gradients obtained using lower force constants, we employ the umbrella integration method[36] as follows. For each window centered on $\boldsymbol{\theta}_n$, we have[40]

$$G_n(\boldsymbol{\theta}) = -\frac{1}{\beta}\ln P_n(\boldsymbol{\theta}) - \frac{1}{2}\sum_{i=1}^{K} k_i(\theta_i - \theta_{n,i})^2 + C_n, \tag{24}$$

where $G_n$ is an estimate of $G$ corresponding to the window, $P_n(\boldsymbol{\theta})$ is the probability density of $\boldsymbol{\theta}$, and $C_n$ is an unknown constant. Differentiation of Eq. (24) with respect to $\theta_j$ gives

$$\frac{\partial G_n(\boldsymbol{\theta})}{\partial \theta_j} = -\frac{1}{\beta}\frac{\partial \ln P_n(\boldsymbol{\theta})}{\partial \theta_j} - k_j(\theta_j - \theta_{n,j}). \tag{25}$$

Approximating $P_n(\boldsymbol{\theta})$ by a Gaussian of the form[36]

$$P_n(\boldsymbol{\theta}) = \prod_{i=1}^{K} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\theta_i - \overline{\theta}_{n,i}}{\sigma_{n,i}}\right)^2\right], \tag{26}$$

in which $\overline{\theta}_{n,i}$ and $\sigma_{n,i}$ denote the mean and the standard deviation of the time series $\widehat{\theta}_i(\boldsymbol{x}_n(t))$, respectively, and substituting into Eq. (25) leads to

$$\frac{\partial G_n(\boldsymbol{\theta})}{\partial \theta_j} = -\frac{1}{\beta}\frac{\theta_j - \overline{\theta}_{n,j}}{\sigma^2_{n,j}} - k_j(\theta_j - \theta_{n,j}). \tag{27}$$

The estimates $\partial G_n(\boldsymbol{\theta})/\partial\theta_j$ from all windows are combined according to

$$\frac{\partial G(\boldsymbol{\theta})}{\partial \theta_j} = \sum_{n=0}^{N} \frac{\partial G_n(\boldsymbol{\theta})}{\partial \theta_j} \frac{T_n P_n(\boldsymbol{\theta})}{\sum_{m=0}^{N} T_m P_m(\boldsymbol{\theta})}, \tag{28}$$

in which $T_n$ is the number of integration steps in the window corresponding to $\boldsymbol{\theta}_n$.

The values of $\boldsymbol{\theta}$ at which gradients of $G(\boldsymbol{\theta})$ were computed were obtained by linearly interpolating the original string $\boldsymbol{\theta}_n$ onto a finer string $\boldsymbol{\theta}_m$, $m = 1, \ldots, M = 500$. The free energy profiles were calculated using the trapezoidal rule, as in Eq. (21). The results were insensitive to values of $M$ in the range 200–1000, similar to the conclusions in Ref. 36. The choice of REX parameters is described in Sec. III D 3.

### 2. Free energy as a function of the reaction coordinate

As explained in Sec. II B, the free energy expressed in terms of the collective variables, $G(\boldsymbol{\theta})$, is not the same as the free energy along the parametric reaction coordinate, $F(\alpha)$, defined in Eq. (8). To compute $F(\alpha)$, we use the procedure introduced in Ref. 21. Once the string has converged to the MFEP, we associate a cell $B_n$ with each fixed image $\boldsymbol{\theta}_n$. The cell $B_n$ is defined by

$$B_n = \{\boldsymbol{x} : d_{n,m}(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}_n) < d_{n,m}(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}_m), \quad \text{for all } m \neq n\}. \tag{29}$$

Thus, $B_n$ contains all the points in configuration space that are closest to image $\boldsymbol{\theta}_n$, according to the distances,

$$d_{n,m}(\boldsymbol{\theta}^a, \boldsymbol{\theta}^b) = \left[(\boldsymbol{\theta}^a - \boldsymbol{\theta}^b)^T \left(\frac{M^{-1}(\boldsymbol{\theta}_n) + M^{-1}(\boldsymbol{\theta}_m)}{2}\right)(\boldsymbol{\theta}^a - \boldsymbol{\theta}^b)\right]^{1/2}, \tag{30}$$

where $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^b$ are arbitrary $K$-dimensional vectors. The cells $B_n$ form a tessellation of the entire configuration space. If the tensor $M$ is proportional to the identity, the distances reduce to the standard Euclidean distance.

For two adjacent images on the string, $\boldsymbol{\theta}_n$ and $\boldsymbol{\theta}_{n+1}$, the boundary between the corresponding cells, $B_n$ and $B_{n+1}$, is defined by

$$d_{n,n+1}(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}_n) = d_{n,n+1}(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{\theta}_{n+1})). \tag{31}$$

Using Eq. (30) with $m{=}n+1$ and that $M^{-1} = M^{-T}$, the square of Eq. (31) can be written as

$$0 = (\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n)^T \left(\frac{M^{-1}(\boldsymbol{\theta}_n) + M^{-1}(\boldsymbol{\theta}_{n+1})}{2}\right) \times \left(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) - \frac{\boldsymbol{\theta}_n + \boldsymbol{\theta}_{n+1}}{2}\right). \tag{32}$$

This equation is a second-order finite-difference approximation to Eq. (5) evaluated at $\alpha = (n + 1/2)/N$. Consequently, the boundaries of successive cells along the string are local approximations of the isocommittor surfaces at the MFEP. Equations (8) and (9) imply that, for $n = 1, \ldots, N-1$, up to discretization errors of order $1/N$,

$$\pi_n = Z^{-1}\int_{B_n} e^{-\beta V(\boldsymbol{x})} d\boldsymbol{x} \approx \int_{(n-1/2)/N}^{(n+1/2)/N} e^{-\beta F(\alpha)} d\alpha \approx N^{-1} e^{-\beta F_n}, \tag{33}$$

where $\pi_n$ is the probability to find the system in cell $B_n$ at equilibrium, $V(\boldsymbol{x})$ denotes the MD potential, $Z$ is the configurational partition function, and $F_n = F(\alpha = n/N)$. Equation (33) gives the following estimate for $F_n$ in terms of $\pi_n$:

$$F_n = -\beta^{-1}\ln \pi_n - \beta^{-1}\ln N, \qquad n = 1, \ldots, N-1. \tag{34}$$

As shown in Ref. 21, there is a simple procedure to compute the probabilities $\pi_n$. First, we run independent MD simulations in each of the cells $B_n$ and impose a "reflection" rule at the cell boundaries, in which all particle momenta are reversed to maintain the MD replica in its cell. Specifically, if $(\boldsymbol{x}_n(t), \boldsymbol{p}_n(t))$ denotes the position and momentum at time $t$ of the MD simulation assigned to the replica in cell $B_n$, we set

$$(\boldsymbol{x}_n(t+\delta t), \boldsymbol{p}_n(t+\delta t)) = \begin{cases} (\boldsymbol{x}^*_n(t+\delta t), \boldsymbol{p}^*_n(t+\delta t)) & \text{if } \boldsymbol{x}^*_n(t+\delta t) \in B_n, \\ (\boldsymbol{x}_n(t), -\boldsymbol{p}_n(t)) & \text{if } \boldsymbol{x}^*_n(t+\delta t) \notin B_n, \end{cases} \tag{35}$$

where $(\boldsymbol{x}^*_n(t+\delta t), \boldsymbol{p}^*_n(t+\delta t))$ denotes the time-evolved value of $(\boldsymbol{x}_n(t), \boldsymbol{p}_n(t))$ after one MD step of size $\delta t$. Up to time discretization errors, the trajectories generated in this way sample the canonical distribution for the cell $B_n$. The test involved in Eq. (35) is a distance check, in accord with Eq. (30).

Let $N_{n,m}$ denote the number of collisions the MD replica in cell $B_n$ makes with the boundary of cell $B_m$ during the MD simulation interval $T_n$. For a sufficiently large $T_n$, the quantity

$$\nu_{n,m} = \frac{N_{n,m}}{T_n} \tag{36}$$

gives an estimate of the rate of escape from cell $B_n$ to cell $B_m$. At a statistical steady state, the conservation of probability requires

$$\sum_{\substack{m=0 \\ m\neq n}}^{N} \pi_n \nu_{n,m} = \sum_{\substack{m=0 \\ m\neq n}}^{N} \pi_m \nu_{m,n}, \qquad n = 0, 1, \ldots, N, \tag{37}$$

which can be solved for $\pi_n$ using the normalization condition $\sum_{n=0}^{N}\pi_n = 1$. The free energy $F_n \approx F(\alpha = n/N)$ can then be computed from Eq. (34).

### 3. Transition rate between the initial and final states

To compute the transition rate, the free energy calculation method described in Sec. II D 2 can be combined with the

version of Markovian milestoning proposed in Ref. 18, which builds on the original works in Refs. 9, 10, and 41, and 42.

In Markovian milestoning, one calculates from each MD trajectory the index of the "milestone" that the trajectory crossed most recently; in the string method, the milestones are defined as the boundaries between the cells $B_n$ defined in Eq. (29). The time-evolution of the index is approximated by a continuous-time Markov chain. To calculate the rate of transition, one needs to estimate the rate matrix of this chain. Markovian milestoning is similar to Markov state modeling (MSM) (Refs. 43–50) using master equations. The main difference is that in Markovian milestoning, the system states (milestones) are hypersurfaces, whereas in standard MSMs, the states form a partition of the configuration space. We denote by $a$, $b$, ... the indices of the milestones (not to be confused with the index $n$ of the cells). It was shown in Ref. 18 that the rate of instantaneous transition from milestone $a$ to milestone $b$ ($k_{a,b}$) can be estimated as

$$k_{a,b} = \frac{\sum_{n=0}^{N} \pi_n N_{a,b}^n / T_n}{\sum_{n=0}^{N} \pi_n T_a^n / T_n}, \qquad a \neq b. \tag{38}$$

In Eq. (38), $\pi_n$ is the equilibrium probability to find the system in cell $B_n$ computed from Eq. (37), $N_{a,b}^n$ is the total number of transitions from milestone $a$ to milestone $b$ observed in the MD simulation confined to cell $B_n$, $T_a^n$ is the total time during this simulation during which $a$ was the most recent milestone visited by the system, and $T_n$ is the total duration of the simulation confined to cell $B_n$.

Given an arbitrary milestone denoted by $b^*$, the instantaneous rate matrix and the MFPTs to the milestone $b^*$ from the other milestones in the system, denoted by $T_{b,b^*}$ with $b \neq b^*$, satisfy the relationship

$$\sum_{b \neq b^*} k_{a,b} T_{b,b^*} = -1, \qquad a \neq b^*. \tag{39}$$

This equation is a standard result in the theory of Markov chains[51] and its derivation in the context of milestoning can be found in Ref. 18. If the milestones $a^*$ and $b^*$ are chosen as the isocommittor surfaces for the transition between two metastable states of a system (e.g., A and B) with $q_{a^*} \approx 0$ and $q_{b^*} \approx 1$, then $(T_{a^*,b^*})^{-1}$is an estimate of the rate of transition from A to B. Use of an unrealistic low friction with the implicit solvent model (Sec. III D 1), as was done for improved sampling, is expected to result in an overestimate of the transition rate (see Sec. IV).

It was shown in Ref. 52 that the MFPT estimate is exact if the milestones are isocommittor surfaces for the transition. This condition will be satisfied approximately, provided assumptions (i) and (ii) in Sec. II B hold, because in that case the boundaries of the cells $B_n$, which we use as the milestones, are approximations of the isocommittor surfaces.

## III. APPLICATION TO THE MYOSIN VI CONVERTER

### A. Preparation of simulation structures

Crystal structures of MVI in the (R) and (PPS) conformations were obtained from the protein data bank (PDB entries 2BKH and 2V26, respectively). The resolution of the respective structures is 2.4 and 1.75 Å. Only residues 703–788, which correspond to the converter domain, were included. The R and PPS conformers are shown in panels (a) and (b), respectively, in Fig. 1. Atoms that form the basis for the collective variables defined in Sec. III B 2 and listed in Table II

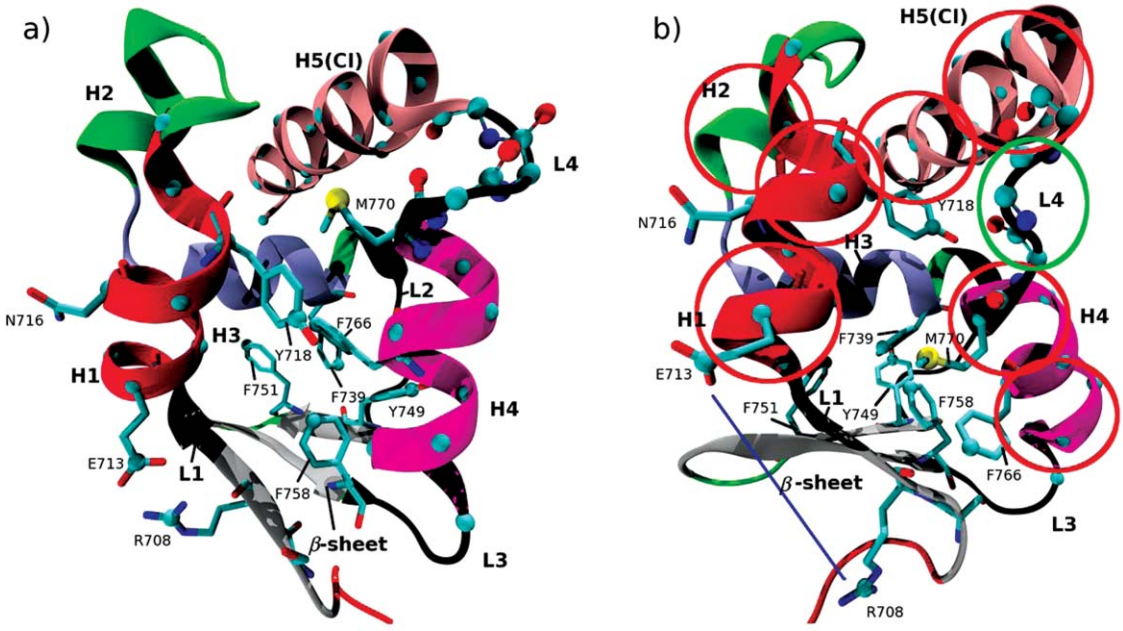


FIG. 1. VI converter in the R (a) and PPS (b) conformation. The secondary structure elements are helices 1–5 (H1–H5), loops 1–4 (L1–L4), and the $\beta$-sheet. We alternatively refer to H5 as the converter insert (CI) to emphasize that it is known to be present only in MVI (Ref. 53). The helices and loops are numbered in the order of increasing residue numbering. Atoms to which forces were applied in the RTMD or string simulations (see Tables I and II) are drawn as spheres. If forces were applied to a sidechain atom, the entire corresponding residue is shown. Atoms that are used to define the collective variables set CVS1 (Table II) are shown as spheres. In (b), red circles are drawn approximately around subdomains that correspond to CV 1–21 in CVS2 (Table III). The green oval is drawn around L4 and corresponds to the collective variables 46–50 in CVS2. The blue line corresponds to the distance CV 51 in CVS2.

are shown as spheres. In the PPS structure, the atoms CD1 and ND2 in residue N716 were interchanged to optimize the local hydrogen bonding patterns. Assignment of histidine protonation states was based on a visual inspection of the local environment for each histidine residue. For both conformers, all histidines were singly protonated on the $\delta$-nitrogen except for H776, which was singly protonated on the $\epsilon$-hydrogen.

The all-atom CHARMM22 topology and parameter files were used in the simulations. To approximate the effect of solvent, we employ the fast analytical continuum treatment of solvation (FACTS) model,[54] which has been shown to yield accurate atomic solvation and pair interaction energies when compared with finite-difference Poisson–Boltzmann data. In applications, FACTS has been shown to maintain stable structures during long simulations ($\approx$ 100 ns) of small proteins and has been implemented in CHARMM.[55,56]

Prior to performing string calculations, each conformer was simulated for 60 ns by equilibrium MD in the canonical ensemble. The average backbone-atom RMSD of the simulation structures from the corresponding crystal structures is 1.6 and 1.25 Å for the R and PPS conformers, respectively, confirming that both converter domain structures are in metastable local minima. The RMSD between the backbone atoms in the x-ray structures is 4.2 Å, while that between the backbone atoms in the average MD structures is 3.9 Å.

## B. Selection of collective variables

As explained in Sec. II B, the quality of a given set of collective variables to describe a complex reaction can only be determined *a posteriori*, either by a committor test or by "cross-validating" the reaction free energy of the reaction rate computed using two or more sets of collective variables. In this section we propose a methodology to generate *a priori* candidate collective variables to describe conformational transitions in proteins.

### 1. Identification of a "resolving set" of atoms using RTMD

RTMD is closely related to the original TMD method.[57,58] The main difference is that targeting forces are applied in the form of a harmonic restraint, rather than a holonomic constraint.[22] We apply RTMD forces only to the atoms in a putative RS starting with their positions in one structure to decrease the best-fit RMSD between this structure and the target structure. Let $\boldsymbol{r}^{rs} \in \mathbb{R}^{3N}$ denote the vector containing the positions of the $N$ atoms in the putative RS to which the RTMD forces are applied. The driving forces are derived from the potential

$$U_{\mathrm{RTMD}}(\boldsymbol{r}^{rs}) = \frac{k}{2}\left(\mathrm{RMSD}(\boldsymbol{r}^{rs}, \boldsymbol{r}_T^{rs}) - \delta\right)^2, \tag{40}$$

where $k$ is the force constant and $\boldsymbol{r}_T^{rs}$ is the (fixed) value of the positions of the atom in the target structure. We start the simulation with $\boldsymbol{r}^{rs}$ in the initial structure and set $\delta$ to the RMSD between $\boldsymbol{r}^{rs}$ and $\boldsymbol{r}_T^{rs}$. The value of $\delta$ is then decreased linearly to zero over $\sim$2 ns. After the zero value of $\delta$ is reached, the restraining potential $U_{\mathrm{RTMD}}(\boldsymbol{r}^{rs})$ is applied for $\sim$3 ns and then scaled linearly to zero over 1 ns, after which the system is allowed to relax for 1–3 ns. The RMSD between all heavy atoms in the relaxed structure and the target structure is taken as the indicator of the quality of our choice of the RS. The RTMD simulations performed in this study and the corresponding RSs are summarized in Table I.

In the RTMD simulations, the RMSD $\left(\boldsymbol{r}^{rs}, \boldsymbol{r}_T^{rs}\right)$ was computed between the positions $\boldsymbol{r}^{rs}$ and $\boldsymbol{r}_T^{rs}$, after a best-fit orientation using the backbone atoms of helix 3 (H3) (see Fig. 1), which does not include any atoms from the RS. Because this orientation set and the RS are different, there is a net force and a net torque acting on the protein. To prevent the rigid-body motion of the simulation structures, the backbone atoms of H3 were restrained to their original positions with harmonic potentials of the form $(1/2)k|\boldsymbol{r}_i - \boldsymbol{r}_i^0|^2$, where $\boldsymbol{r}_i$ denotes the position of atom $i$, $k$ equals 1 kcal/mol/Å$^2$, and the superscript 0 refers to the initial position. These restraints are very unlikely to affect the path of the transition because (i) the restrained helix is relatively far away from the region in which the two endpoint structures differ appreciably and (ii) the RMSD between the helix backbone in the two conformations is only 0.24 Å. These two facts indicate that the helix is not involved in the conformational change and remains intact during the transition.

The first trial RS was chosen based on a visual inspection of the endpoint structures shown in Fig. 1. Two prominent qualitative differences between the structures are (i) the orientation of helix 4 (H4), which is vertical and approximately perpendicular to the $\beta$-sheet in the R structure, and inclined at $\simeq$45° to the $\beta$-sheet in the PPS structure, and (ii) the conformation of loop 4 (L4), which is $\alpha$-helical in the R structure, and unwound in the PPS conformation. In addition, going from the R structure to the PPS structure, H4 appears to rotate about its axis, moving the sidechains of M770 and F766 from the interior to the outside of the converter domain (Fig. 1). Based on these observations, we assumed that much of the conformational transition could be accounted for by the movement of H4, L4, M770, and F766. Therefore, the first RS includes the $C_\alpha$ atoms of H4 and L4, the $S_\delta$ atom of M770, and the $C_\gamma$ atom of F766. Because we expect that the movement of M770 would involve significant motion of Y718 due to steric clashes between the two residues, we also included the $C_\gamma$ atom of Y718 in the RS.

Starting from the first trial RS, which contained 28 atoms (listed as RTMD 1 in Table I), additional atoms or domains were added until the heavy-atom RMSD between the final RTMD simulation structures and the corresponding the target structures was below 2.0 Å in both directions. The smallest tested RS that satisfies these criteria was used in RTMD simulation 5 in Table I. It is composed of 59 atoms, which are drawn as spheres in Fig. 1 and listed in Table I. This RS includes atoms from helices H1 and H5 as well as from several aromatic residues in the core of the converter domain.

We note that the RS found by the above procedure is unlikely to be unique. Furthermore, the composition of the RS will depend on the RMSD criterion above, such that a lower RMSD would probably require a larger RS. Therefore, although the search for the RS is systematic, the final choice of RS is somewhat subjective.

TABLE I. Summary of RTMD simulations. Residue numbers followed by an atom type indicate that only that atom was included. Residue numbers followed by "BB" indicate that the backbone atoms (N C O CA) were included. CHARMM atom nomenclature is used. "n/a" indicates that the corresponding simulations were not performed. Important residues in the RS are shown in Figs. 1(a) and 1(b). "RMSD ($\delta = 0$)" is the minimum RMSD value observed after the target RMSD, $\delta$, reaches 0, but before the forces are switched off. "RMSD (relaxed)" is the average RMSD value observed after the forces are switched off. RMSD values are computed using all heavy atoms. The force constant $\kappa$ is given in units of kcal/mol/Å$^2$. Simulation time is quoted in ns. For brevity, "P" denotes the "PPS" state. RMSD values are quoted in Angstroms.

| RTMD | Resolving set (number of atoms) | $\kappa$ | RMSD ($\delta = 0$) R→P/P→R | RMSD (relaxed) R→P/P→R | Simulation time R→P/P→R |
|---|---|---|---|---|---|
| 1 | 761–769 CA; 770–773 BB; 770 SD; 718 766 CG (28) | 0.5 | 4.2/3.9 | (n/a) | 2/2 |
| 2 | 762–769 BB; 771–773 BB (44) | 0.5 | 3.8/(n/a) | (n/a) | 2 |
| 3 | 761–769 711–720 774–788 CA; 770–773 BB; 770 SD; 716 766 CG (38) | 1.0 | 3.0/3.4 | (n/a) | 2 |
| 4 | 761–769 711–720 774–788 CA; 770–773 BB; 770 SD; 718 766 758 749 708 CZ (56) | 1.0 | 1.75/1.5 | 2.3/1.65 | 6/6 |
| 5 | 761–769 711–720 774–788 CA; 770–773 BB; 770 SD; 713 CD; 718 766 758 749 708 739 751 CZ (59) | 1.0 | 1.25/1.26 | 1.60/1.66 | 6/10 |
| 6 | 761–769 711–720 774–788 CA; 770–773 BB; 770 SD; 713 722 CD; 718 766 758 749 708 739 751 CZ (60) | 1.0 | 1.26/1.26 | 1.75/1.45 | 6/6 |

The evolution of the heavy-atom RMSD in RTMD 5 is shown in Fig. 2. The equilibration (with $\delta$=0) and relaxation (forces turned off) phases were each 2 ns longer for the PPS→R transition to determine whether the final RMSD could dip below the R→PPS value (≃1.6 Å). A comparison of the final RMSD values from the PPS→R and R→PPS simulations shows that the extra 4 ns made no difference.

The fact that the RMSD values (≃ 1.6 Å) between the final relaxed RTMD structures and the target structures are lower than the maximum RMSD values from the NVT simulations of the R and PPS structures (2.6 and 2.0 Å, respectively, see Fig. 2) strongly suggests that each simulation structure has reached a stable conformation near the corresponding target structure.

Because applying targeting forces only to the atoms in the RS used in the RTMD listed as number 5 in Table I is sufficient to enforce a complete conformational change, we assume that the mechanism of the transition can be understood by considering only the positions of these atoms. Two sets of collective variables used for the string simulations in this study were chosen on the basis of the RS used in RTMD 5.

### 2. First set of collective variables (CVS1)

A simple set of collective variables that can be constructed from the RS of RTMD 5 consists of just the positions of the atoms in the RS. Specifically, given the set of 59 atomic position triplets $\boldsymbol{r}_i^{rs} = (x_i^{rs}, y_i^{rs}, z_i^{rs})$ with $i = 1, 2, \ldots, 59$, one can define a set of $3 \times 59 = 177$ CVs,

$$\begin{aligned} \widehat{\theta}_{3(i-1)+1}(\boldsymbol{x}) &= x_i^{rs}, \\ \widehat{\theta}_{3(i-1)+2}(\boldsymbol{x}) &= y_i^{rs}, \\ \widehat{\theta}_{3(i-1)+3}(\boldsymbol{x}) &= z_i^{rs}. \end{aligned} \tag{41}$$

A serious drawback of this definition, however, is that the variables are not invariant under rigid-body translation or rotation of the simulation system. To guarantee rigid-body

TABLE II. Atoms used to define the set of collective variables CVS1. For each atom, three CV are defined, which correspond to the Cartesian $x$, $y$, and $z$ positions. Atoms are specified by their residue ID and the atom type. Atom positions on the converter structure are shown in Fig. 1. The total number of position CV is 177.

| CV description | Residue ID(s) | Atom type(s) | CV indices (number of CV) |
|---|---|---|---|
| Helix 1 position | 711–720 | CA | 1–30 (30) |
| Helix 4 position | 761–769 | CA | 31–57 (27) |
| Converter insert position | 774–788 | CA | 58–102 (45) |
| Loop 4 conformation | 770–773 | N, C, CA, O | 103–150 (48) |
| Sidechain position | R708 | CZ | 151–153 (3) |
| Sidechain position | E713 | CD | 154–156 (3) |
| Sidechain position | Y718 | CZ | 157–159 (3) |
| Sidechain position | F739 | CZ | 160–162 (3) |
| Sidechain position | Y749 | CZ | 163–165 (3) |
| Sidechain position | F751 | CZ | 166–168 (3) |
| Sidechain position | F758 | CZ | 169–171 (3) |
| Sidechain position | F766 | CZ | 172–174 (3) |
| Sidechain position | M770 | SD | 175–177 (3) |

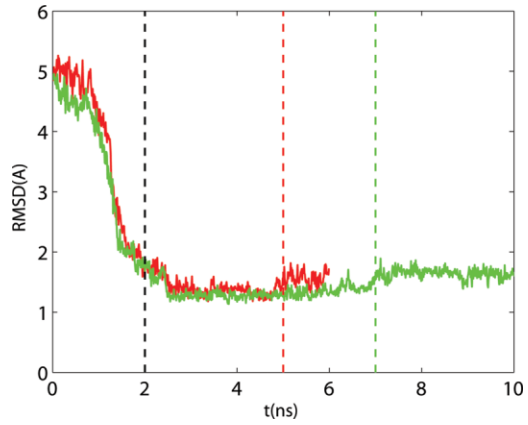


FIG. 2. Evolution of the RMSD of the heavy atoms between the RTMD simulation 5 structures and the corresponding target structures. Red solid line R structure (PPS target); green solid line PPS structure (R target); dashed vertical lines mark the times at which $\delta$ reaches 0.0 (black dashed line), the RTMD forces are switched off for the R→PPS simulation (red dashed line) and the PPS→R simulation (green dashed line).

invariance of the CV defined above, the atomic positions in Eq. (41) were expressed in local coordinates as follows. A local frame of reference was constructed based on the positions of all atoms in H3 (this helix was restrained in the RTMD simulations). First, a mass-weighted $3 \times 3$ correlation tensor was computed as

$$C_{i,j} = \sum_{n=1}^{N} (r_{n,i} - \hat{r}_i)(r_{n,j} - \hat{r}_j) m_n. \tag{42}$$

In this equation, $N$ is the number of atoms that comprise H3, $r_{n,i}$ is the $i$th coordinate of the atom indexed by $n$ (i.e.,$r_{n,1} = x_n, r_{n,2} = y_n, r_{n,3} = z_n$, if $[x_n, y_n, z_n]$ are the $x$, $y$, and $z$ positions of atom $n$), $\hat{r}_i = \sum_{n=1}^{N} m_n r_{n,i} / \sum_{n=1}^{N} m_n$ is the COM position vector of the $N$ atoms, and $m_n$ is the mass of atom $n$. $C_{i,j}$ is symmetric positive definite and therefore has three real eigenvalues, $\lambda_i$, $i = 1, 2, 3$, with three corresponding eigenvectors, $\boldsymbol{v}_i$, that are orthonormal ($\boldsymbol{v}_i \cdot \boldsymbol{v}_j = \delta_{ij}$). These eigenvectors are taken to be the basis vectors of a moving coordinate frame in which the positions CV are computed. To ensure right-handedness of the coordinate frame, one of the eigenvectors occasionally needed to be inverted ($\boldsymbol{v}_i \to -\boldsymbol{v}_i$) during the MD simulations. Additional constraints that guarantee the uniqueness of the computed coordinate frame and the details of computing the derivatives of the position CV expressed in a local frame of reference are given in Appendix A. The set of positions CV defined above is denoted as CVS1.

The positions CV expressed in the local reference frame are invariant with respect to rigid-body motion of the simulation system. Although such motions change the position vector relative to the absolute (simulation) frame, the position vector remains the same in the local frame, because both the frame and the rest of the molecule undergo the same rigid-body motion. We note that other methods of avoiding the change in CV due to rigid-body motion exist, such as those based on simple harmonic restraints,[59] Eckart constraints,[60–62] or moment of inertia tensor constraints.[63] Each method, including ours, comes with disadvantages, such as additional complexity of implementation, or approximation in the force calculations. In the present simulations the eigenvectors of $C_{i,j}$ were always unique, which ensured an unambiguous definition of the local frame. However, this will not be the case for some molecular geometries, such as those in which the atoms that define the coordinate frame have internal symmetry, such as backbones of $\alpha$-helices. (In this case, one principal vector will be directed along the axis of the $\alpha$-helix, but the two remaining vectors are unique only up to a rotation in the plane perpendicular to the first vector). In addition, coordinate frame vectors defined on the basis of very flexible bodies may fluctuate strongly during MD simulation, which could render any function of these vectors (such as positions) effectively ill-defined, and lead to instabilities if forces are applied based on these functions. Thus, the choice of local frame must be made with caution and may require experimentation.

We note that $C_{i,j}$ can be related to the moment of inertia tensor,

$$I_{i,j} = \sum_{n=1}^{N} (|r_n - \hat{r}|^2 \delta_{i,j} - (r_{n,i} - \hat{r}_i)(r_{n,j} - \hat{r}_j)) m_n, \tag{43}$$

by $C_{i,j} = I^* \delta_{i,j} - I_{i,j}$, where $I^* = \sum_{n=1}^{N} |r_n - \hat{r}|^2 m_n$. Since the vectors $\boldsymbol{v}_i$ are orthonormal, they also diagonalize $I^* \delta_{i,j}$ and, consequently, $I_{i,j}$. Thus, the vectors $\boldsymbol{v}_i$ are also eigenvectors of the moment of inertia tensor $I_{i,j}$ (although the corresponding eigenvalues will be different).

The collective variables CVS1 defined in this subsection are listed in Table II and shown in Fig. 1.

#### 3. Second set of collective variables (CVS2)

To cross-validate the results using the strategy explained in Sec. II B, we constructed an additional set of collective variables, hereafter referred to as CVS2. In this set, the number of collective variables was reduced from 177 to 51 by first representing several groups of atomic positions used in CVS1 by the position of their COM expressed in local coordinates as explained above for CVS1 (see CV 1–21 in Table III). In addition, the positions of several atoms in CV1 were replaced by five dihedral angles and one distance between the COMs of two sets of atoms (see CV 46–51 in Table III).

### C. Generation of initial conditions

The string method in collective variables described in Sec. II C requires an initial string [i.e., a value for each $\boldsymbol{\theta}_n(0)$] and $n$ all-atom configurations of the converter for the estimation of $M(\boldsymbol{\theta}_n(t))$ and $\nabla G(\boldsymbol{\theta}_n(t))$ [see Eq. (12)].

These initial conditions were obtained from MEPs generated using the zero-temperature string method in Cartesian coordinates (ZTS), which was implemented in CHARMM following Ref. 20. Because the ZTS method is conceptually and algorithmically very similar to the string method in collective variables described in Sec. II C, with the mean force $\nabla G(\boldsymbol{\theta})$ and the tensor $M(\boldsymbol{\theta})$ replaced by the atomic force, $\nabla V(\boldsymbol{x})$, and the inverse of the mass matrix, respectively, we refer the reader to the supplementary materials for full details.

ZTS was used to generate two MEP from which the initial conditions, described above, were computed. The purpose

TABLE III. CV in set 2. Each COM-position entry corresponds to three Cartesian positions. Residue numbers followed by an asterisk indicate that only the sidechain atoms were included in the CV definition (the backbone atoms C, N, CA, and O were excluded). CVs are indicated on the converter structure in Fig. 1(b). The total number of CV is 51.

| CV description | CV type | Atoms involved | CV indices (No. of CV) | Equivalent CVs in CVS1 |
|---|---|---|---|---|
| H1 position | COM-position | Residue 711–715 | 1–3 (3) | 1–30 |
| | COM-position | Residue 716–720 | 4–6 (3) | |
| H2 position | COM-position | Residue 723–728 | 7–9 (3) | None |
| H4 position | COM-position | Residue 761–765 | 10–12 (3) | 31–57 |
| | COM-position | Residue 766–770 | 13–15 (3) | |
| CI position | COM-position | Residue 774–781 | 16–18 (3) | 58–102 |
| | COM-position | Residue 782–788 | 19–21 (3) | |
| Sidechain position | COM-position | Residue 718* | 22–24 (3) | 157–159 |
| Sidechain position | COM-position | Residue 722* | 25–27 (3) | None |
| Sidechain position | COM-position | Residue 739* | 28–30 (3) | 160–162 |
| Sidechain position | COM-position | Residue 749* | 31–33 (3) | 163–165 |
| Sidechain position | COM-position | Residue 751* | 34–36 (3) | 166–168 |
| Sidechain position | COM-position | Residue 758* | 37–39 (3) | 169–171 |
| Sidechain position | COM-position | Residue 766* | 40–42 (3) | 172–174 |
| Sidechain position | COM-position | Residue 770* | 43–45 (3) | 175–177 |
| L4 conformation | Dihedral ($\phi$) | M770C/K771N/K771CA/K771C | 46 (1) | 103–150 |
| | Dihedral ($\psi$) | K771N/K771CA/K771C/S772N | 47 (1) | |
| | Dihedral ($\phi$) | K771C/S772N/S772CA/S772C | 48 (1) | |
| | Dihedral ($\psi$) | S772N/K772CA/K772C/D773N | 49 (1) | |
| | Dihedral ($\phi$) | S772C/D773N/D773CA/D773C | 50 (1) | |
| Sidechain position | COM distance | Residues 708*, 713* | 51 (1) | 151–156 |

of generating two initial paths was twofold. First, it allowed us to investigate the extent to which the choice of the initial condition affects the computed free energy profile. We found a difference of several kilocalories per mole in the free energy barriers that correspond to the two initial paths, indicating a significant effect (see Sec. III D 1). Second, computing the free energy along two different paths provides a test of the accuracy of the calculation: since the endpoints of both paths correspond to the same metastable states, the free energy difference between the endpoints should be independent of the path. In practice, errors arising from the discretization of the string or insufficient sampling may lead to different free energy values. The magnitude of the difference is a measure of the accuracy of the computed free energy. The first MEP (referred to as MEP1), resolved using 256 replicas, was generated in seven cycles, starting from the two endpoint (R and PPS) structures. In each cycle, a linear interpolation in Cartesian space that doubles the number of replicas is performed, followed by 100 iterations of the ZTS method. Each iteration of the method consists of 20 steps of minimization using the steepest descent minimizer in CHARMM, followed by a reparametrization step, to enforce Eq. (13) with $\boldsymbol{\theta}$ replaced by $\boldsymbol{x}$ (see supplementary materials).

The second MEP (referred to as MEP2) was constructed by changing the direction of rotation of the dihedral angles in a flexible loop (see Fig. 3). In MEP1, the backbone oxygen of residue K771 (indicated by a green arrow) rotates to the "outside," while in MEP2, it passes "under" the upper part of L4 (residue S772). A four-replica path was taken at the end of the first cycle in the generation of MEP1. The two intermediate structures were modified manually to change the direction of loop torsion. This path was then refined to 256 replicas, as described for MEP1.

To quantify the difference between MEP1 and MEP2, we computed the RMSD between corresponding structures along the two paths, using all atoms and using only the atoms that belong to L4. The RMSD plots are shown in Fig. 4(a). The difference in the configurations of L4 is more pronounced

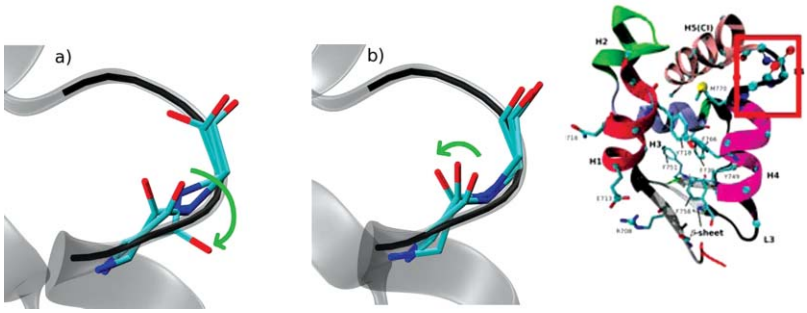


FIG. 3. Conformational change in L4 in zero-temperature path MEP1 (a) and MEP2 (b). The converter in the R conformation is shown in transparent gray. For each case, the R, PPS, and one intermediate conformation of loop 4 are shown. The directions of loop torsion are shown by green arrows. The inset indicates the location of L4 in the converter domain.

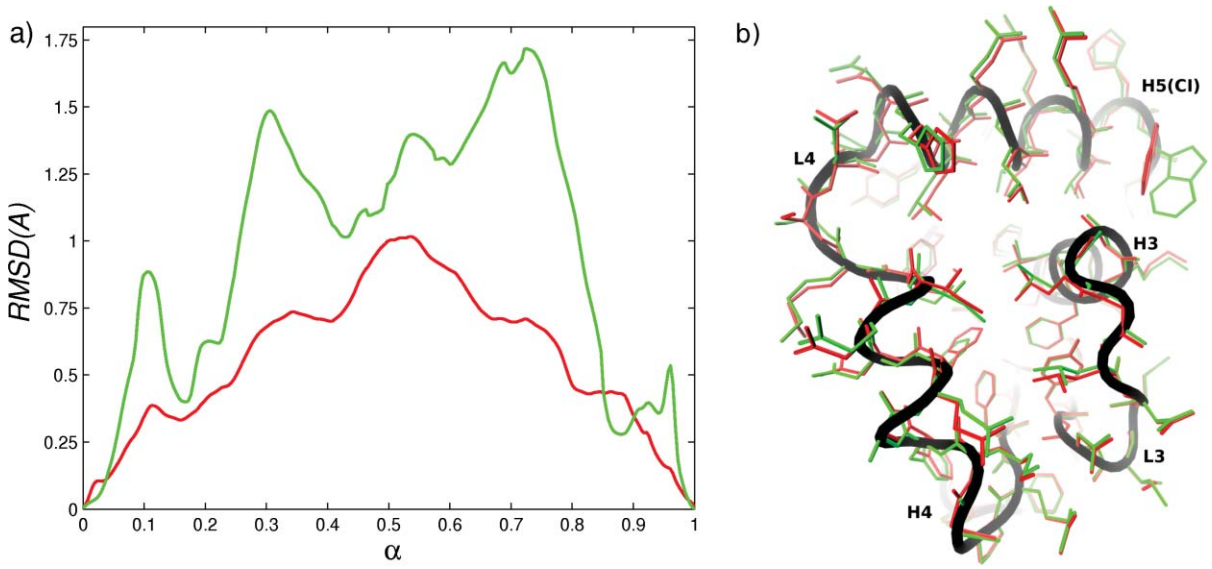


FIG. 4. (a) RMSD between corresponding replicas from MEP1 and MEP2: red solid line computed using all atoms; green solid line computed using atoms of L4. (b) Superposition of structures from MEP1 and MEP2 at $\alpha = 0.66$ (replica 170 of 256); red: MEP1; green: MEP2; black: backbone from MEP2.

than the overall (all-atom) difference between the structures. This is expected because only the coordinates of L4 were modified directly in the construction of MEP2. However, the all-atom RMSD between MEP1 and MEP2 (maximum of 1 Å at $\alpha \simeq 0.5$) suggests that the changes made to L4 propagate through the entire converter domain. In Fig. 4(b) we have superposed structures from MEP1 and MEP2 at $\alpha = 0.65$ after a best-fit alignment.[62] The figure makes clear that differences between the MEPs are not limited to L4, but are present throughout the entire converter structure. The potential energy along MEP1 and MEP2 is shown in Fig. 5. The figure indicates that at $\alpha \simeq 0.35$ and $\alpha \simeq 0.65$ MEP1 is higher in energy than MEP2 by $\simeq$40 kcal/mol. By examining the individual contributions to the total effective potential energy [i.e., bond, angle, linear and improper dihedral, electrostatic, van der Waals (vdW) and solvation energies], we found that most of the difference arises from differences in the dihedral and vdW energy terms. Figure 5 shows that if the dihedral and vdW energy terms are omitted, the profiles corresponding to MEP1 and MEP2 are very similar (the contributions of the two terms are approximately of the same magnitude). To see whether the differences in the energies arise primarily from differences in the conformations of L4, we computed the matrix of interaction energies between all pairs of residues at $\alpha = 0.65$. We found the difference in the interaction energies between L4 and the rest of the converter to be $\simeq$6 kcal/mol (compared with the total energy difference of $\simeq$40 kcal/mol). Furthermore, the difference in the interaction energies was significant for many residue pairs throughout the converter domain. These findings indicate that MEP1 is a higher-energy path with the higher interaction energies not localized in a specific region of the converter (e.g., L4). Despite the $\simeq$40 kcal/mol differences in the potential energies between the MEPs at $\alpha = 0.65$, Fig. 4(b) shows that most of the residue sidechains occupy similar positions and suggests that the origin of the differences in the energies between MEP1 and MEP2 is rather subtle (excluding L4).

Prior to running simulations using the string method in collective variables, the resolution of MEP1 and MEP2 was decreased from 256 to 32 replicas to reduce the computer cost of simulation. The resulting coordinate sets were used to calculate the initial values of the collective variables in CVS1 and CVS2 and to initialize the restrained MD simulations required for the estimation of $M(\boldsymbol{\theta}_n(t))$ and $\nabla G(\boldsymbol{\theta}_n(t))$. The reduction in resolution is justified because the number of dimensions of the CV spaces (177 and 51 for CVS1 and CVS2, respectively), which contain the respective MFEPs, is much lower than the dimensionality of the full Cartesian space (4326). Consequently, the free energy landscape in the space of CV will be much smoother, requiring fewer discretization points. Furthermore, after the string has converged to the MFEP, the resolution can be increased by interpolation to improve the accuracy of discretization, as performed in Sec. III D 3. Animations of MEP1 and MEP2 can be found in the supplementary materials.[19]

## D. Minimum free energy paths, free energies, and rates

### 1. Calculation of the MFEPs

To compute MFEPs, three simulations were performed using the string method in collective variables, as summarized in Table IV (S1–S3). Each simulation was performed using a string discretized into 32 images, with one MD replica assigned to each image, using 1–4 processors per replica, so that the total CPU requirement was between 32 and 128 CPUs. Each replica was simulated in the NVT ensemble at 300 K using the Langevin dynamics thermostat coupled to heavy atoms using a friction constant of 1 $ps^{-1}$. With this value of the friction constant, the temperature computed from the MD simulations fluctuated around 300K with a standard deviation of 6 K (computed over 20 ns of simulation). Covalent bonds to hydrogen atoms were constrained with SHAKE. The FACTS model was used to approximate the effects of

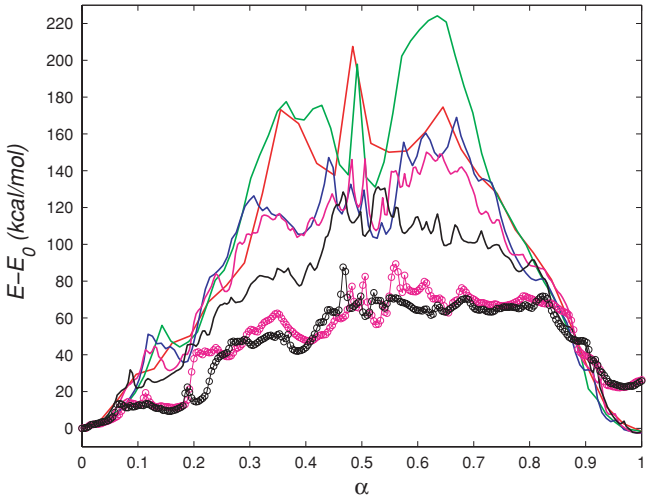


FIG. 5. (a) Potential energy along the initial paths calculated using the zero-temperature string method (ZTS). MEP1: Red solid line iteration 400, 32 replicas; green solid line iteration 500, 64 replicas; blue solid line iteration 600, 128 replicas; magenta solid line iteration 700, 256 replicas; (magenta ○—) iteration 700, 256 replicas; energy from dihedral angles and van der Waals interactions excluded. MEP2: black solid line iteration 700, 256 replicas; (black ○—) iteration 700, 256 replicas; energy from dihedral angles and van der Waals interactions excluded.

solvent.[54] The approximate time required to integrate 1 ns of MD was 15 CPU-core-hours per replica using a multiprocessor cluster equipped with quad-core Intel Xeon CPUs with the InfiniBand interconnect. Thus, simulation S1 (see Table IV) running on 32 quad-core processors of comparable speed (128 cores in total) would take 6 days.

For the string simulations using CVS1, the force constant in the restrained dynamics Eq. (16) was set to 1.0 kcal/mol/Å$^2$ for all CVs. For each replica on the string, ten MD steps were performed to compute the average force and tensor $M$, which corresponds to $\Delta t$ =20 fs in Sec. II C. The CV values were updated using Eqs. (12) and (13). The friction coefficient $\gamma$ was set to 1673 ps$^{-1}$. This value corresponds to five times the minimum value of $\gamma$ for which $\gamma^{-1}\Delta t$ results in stable integration of Eq. (12) (determined by trial and error). In view of the Einstein relation for the Brownian motion ($D = k_B T/m\gamma$),[64] in which $m$ represents particle mass and $D$ is the coefficient of diffusion of the particle, the need to use a large value for $\gamma$ in the string simulation indicates a low rate of diffusion of the string on the landscape of the free energy $G(\boldsymbol{\theta})$ defined by Eq. (3), as would be expected for the evolution of a coarse-grained representation of the system.

For the string simulations with CVS2, the force constants in Eq. (16) were set to 1.0 kcal/mol/Å$^2$ for position CVs and distance CVs and to 10.0 kcal/mol/rad$^2$ for dihedral angle CVs. The average force on each CV was computed after every 15 MD steps, which corresponds to $\Delta t = 30$ fs. $\gamma$ was set to 1255 ps$^{-1}$, determined by trial and error as for simulations with CVS1.

In each simulation, the string was evolved from the initial condition until convergence to the MFEP was obtained, according to Eq. (12). Figure 6 shows the evolution of $D(t)$ in simulations S1–S3 (Table IV). Simulation S3 was run longer (49 ns) than either S1 (36 ns) or S2 (40 ns) to ascertain that $D(t)$ was not continuing to increase. Since $D(t)$ is a measure of the distance that a path has traveled from the initial path, the different plateau values in Fig. 6 corresponding to the simulations S1–S3 reflect the different distances between the MFEPs and the respective MEPs. Although simulations S1 and S2 were initialized from the same initial path, they employ different collective variables (different both in type and number), which is likely to be the reason for the difference in the respective plateau values. In addition, the different plateau distances between the MFEPs and the MEPs may be related to the different widths of the corresponding transition tubes (discussed in Sec. III D 4), with MFEP S1 having the widest transition tube and MFEP S2, the narrowest.

### 2. Description of the MFEP

In this section, we describe the mechanism of the transition corresponding to the MFEP from simulation S3. S3 is

TABLE IV. Summary of string simulations performed in this study. Simulations G2fr and G3fr were performed with REX (see text). For simulations G2f10 and G3f10, all force constants in the restraining potential in Eq. (16) were increased by a factor of 10, as described in the text. Simulation durations correspond to the different procedures associated with the particular type of simulation (see text for details).

| Simulation index | CV set | Initial path | Number of images/cells | Simulation duration (ns) |
|---|---|---|---|---|
| S1 | CVS1 | 1 | 32 | 36 |
| G1c | CVS1 | 1 | 32 | 20 |
| F1 | CVS1 | 1 | 32 | 30 |
| S2 | CVS2 | 1 | 32 | 40 |
| G2c | CVS2 | 1 | 32 | 20 |
| G2f | CVS2 | 1 | 128 | 15 |
| G2fr | CVS2 | 1 | 128 | 20 |
| G2f10 | CVS2 | 1 | 128 | 15 |
| F2 | CVS2 | 1 | 32 | 40 |
| S3 | CVS2 | 2 | 32 | 49 |
| G3c | CVS2 | 2 | 32 | 20 |
| G3f | CVS2 | 2 | 128 | 15 |
| G3fr | CVS2 | 2 | 128 | 20 |
| G3f10 | CVS2 | 2 | 128 | 15 |
| F3 | CVS2 | 2 | 32 | 30 |

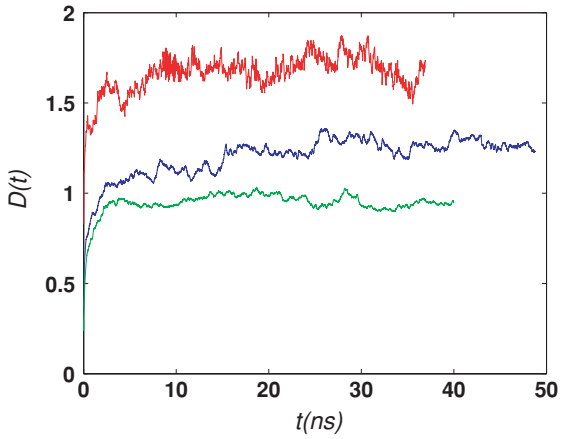


FIG. 6. Convergence of string simulations (see Table IV) monitored by the evolution of $D(t)$ [Eq. (19)]. Red solid line, S1; green solid line, S2; blue solid line, S3.

chosen in preference to S2 because the corresponding free energy profile associated with the committor is probably more accurate (discussed in Sec. III D 4). In addition, the profiles of the free energies $G$ and $F$ associated with simulation S3 exhibit lower barriers (see Sec. III D 3). This transition path, therefore, is more likely to be relevant for the ensemble of all transition paths. A brief discussion of the MFEP from S2 is available in the supplementary materials.[19] S3 is chosen in preference to S1 because the calculated endpoint free energy difference is closer to that of MFEP S2 (discussed in Sec. III D 3).

Six snapshots from the MFEP ordered in the R→PPS direction are shown in Fig. 7 to illustrate the transition mechanism; important residues and secondary structure elements are labeled in Fig. 1. The main differences between the R and PPS conformations of the converter were discussed in Sec. III B 1. Transition in the R→PPS direction begins with a downward motion of H4, which positions M770 in the middle of the converter interior and causes H5 to rotate (R to I5); (the numbers in parentheses correspond to image indices and R and P correspond to the rigor and prepowerstroke states); the hydrogen bonds between R708 and E713 break (I5); H4 continues its downward motion and M770 moves between the aromatic rings of Y718 and F766 (I15 to I22). Aromatic residues in the converter core (F739, Y749, F751, F758, F766) move into their PPS positions (I22 to I25). L4 begins to twist toward its PPS position (I22 to I25), and H4 begins to tilt and twist, accompanied by the motion of M771 and F766 to the outside of the converter interior (I22 to I25). H4 continues to tilt, and CI rotates into its PPS position (I25 to P). An animation of this MFEP can be found in the supplementary materials.[19]

Structures I15 and I25 on the MFEP (Fig. 7) correspond to peaks in the free energies $G(\boldsymbol{\theta})$ and $F(\alpha)$ (not to be confused with the energy along the MEP shown in Fig. 5). Free

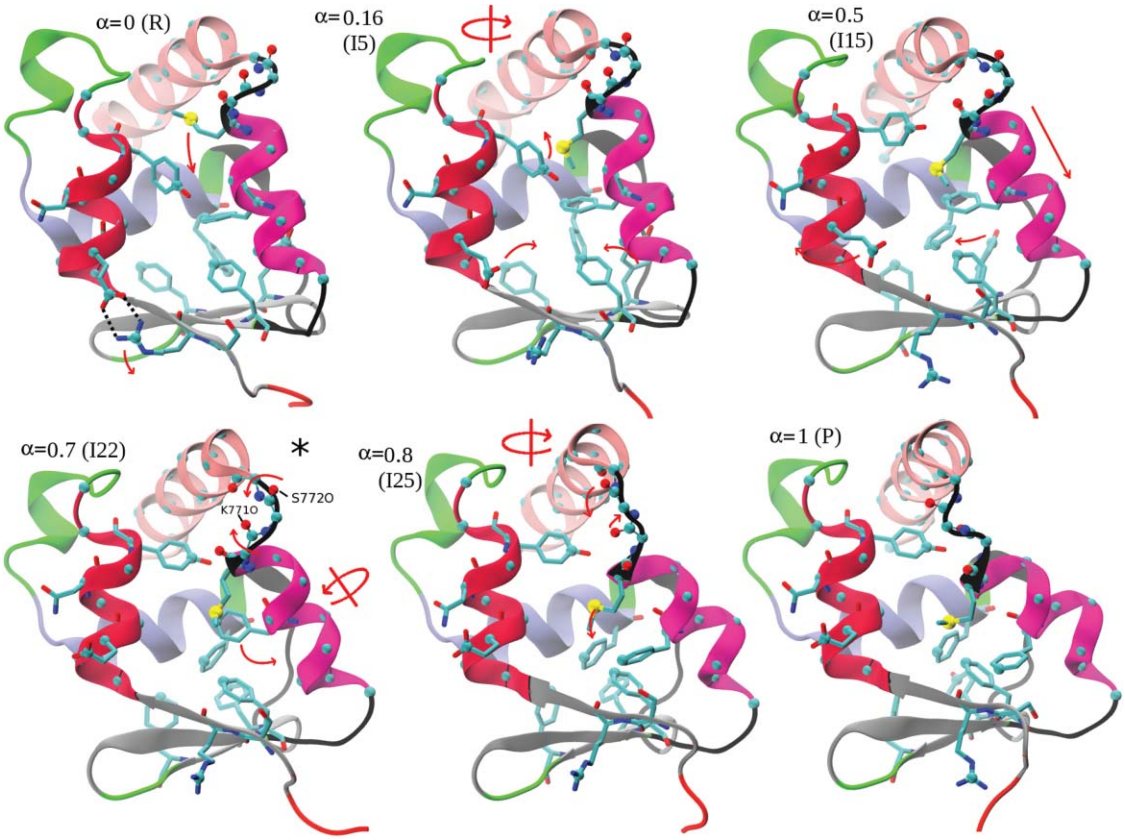


FIG. 7. Snapshots from simulation S3 (Table IV) that illustrate the transition mechanism. The converter structures are shown, as in Fig. 1. Red arrows indicate the conformational change associated with the snapshot. The location of the snapshot on the transition path ($\alpha$ =Ix/I31), and the corresponding image index (I0–I31) are shown for each snapshot. Images I0 and I31 correspond to the rigor (R) and pre-powerstroke (P) states, respectively. Hydrogen bonds between R708 and E713 are indicated by dotted black lines in the R snapshot. The metastable state is indicated by an asterisk. Residues and secondary structure elements are labeled in Fig. 1.

energy profiles are discussed in Secs. III D 3 and III D 4 below. In I15, M770 is involved in repulsive interactions with Y718 and F766 (while passing from the interior of the converter to the outside), and in I25, loop L4 is in a strained configuration (while twisting from the R into the PPS conformation). The global maximum of the committor free energy $F$ corresponds to I15.

Structure I22 on the MFEP (located between I15 and I25) is at a local free energy minimum. In this case, M770 is in the PPS position, but L4 is still in the R position, i.e., M770 has completed crossing its barrier, but L4 has not yet begun. As a qualitative check of whether the snapshot I22 corresponds to a metastable state, a 40 ns equilibrium MD simulation was performed, starting from the structure I22. The heavy-atom RMSD from the initial configuration remained at $\simeq$1.7 Å during the simulation, M770 and L4 both remained close to their starting positions, and no major structural changes in the conformation were observed, consistent with metastability.

### 3. Free energy as a function of the collective variables

In this subsection, we describe the calculation of $G(\boldsymbol{\theta})$, the free energy as a function of the collective variables defined in Eq. (3). We find that considerable care is required to compute accurate and converged profiles of $G$. The free energy of the committor ($F$) is discussed in Subsection III D 4.

After the simulations S1–S3 converged to the corresponding MFEPs, the images $\boldsymbol{\theta}_n$ were held fixed, and restrained MD simulations were performed for approximately 20 ns in each case to compute the gradients $\nabla G(\boldsymbol{\theta}_n)$. These simulations are denoted G1c–G3c in Table IV. Free energies were then computed using the trapezoidal rule [Eq. (21)], as described in Sec. II D 1 (not using umbrella integration).

The resulting FE profiles are shown in Fig. 8(a). The profiles corresponding to the 32-image calculations show that the PPS state has higher free energy than the R state. It was puzzling, however, that the FE difference between the two end-states was ∼12 kcal/mol for G2c, but ∼5 kcal/mol for G3c, because both the collective variables and the endpoint images along the initial string (computed from equilibrated x-ray structures) were the same for these simulations; only the initial paths connecting the two endpoints were different (see Table IV). The difference in paths, however, should have no impact on the FE difference between end states, since free energy is a function of state.

Three potential causes for the discrepancy were considered: (I) integration error of the trapezoidal rule, (II) movement of the endpoints during the string simulations that could position them in different locations on the FE landscape, and (III) inadequate sampling in the estimation of the FE gradient $\nabla G(\boldsymbol{\theta}_n)$.

To rule out (I), the MFEPs from S2 and S3 were interpolated onto 128-image strings using linear interpolation, and the FE was calculated from restrained MD simulations for ∼15 ns, followed by trapezoidal rule integration, as before. Fig. 8(a) shows that the resulting profiles did not change significantly. These simulations are denoted G2f and G3f in Table IV and Figs. 8(a) and 8(b).

To rule out (II), we constructed four-replica strings between the corresponding endpoints of the MFEPs in S2 and S3 using linear interpolation (the resulting path was acceptable because the endpoints were very close, e.g., within ∼1 Å RMSD of one another). Restrained simulations were performed for 20 ns, and the FE profiles was computed as before. The FE difference between the endpoint images was found to be ∼0.5 kcal/mol for each pair, which cannot account for the 7 kcal/mol difference in the FE change described above.

To test (III) efficiently, rather than running additional simulations beyond the original 20 ns (which could require much longer integration times), we combined restrained MD simulations with Hamiltonian replica exchange (see Sec. II D 1). The algorithm is a generalization of umbrella sampling replica exchange,[34] in which restraining potentials are exchanged between neighboring windows, according to the Metropolis criterion.

Twenty-nanosecond simulations with REX were performed using 128-image strings, interpolated from the MFEPs. These simulations are denoted G2fr and G3fr in Table IV and Figs. 8(a) and 8(b). Trial moves for all replicas were attempted simultaneously (i.e., $0 \leftrightarrow 1$, $2 \leftrightarrow 3$, …, or $1 \leftrightarrow 2$, $3 \leftrightarrow 4$, …, by a random decision) once in every 100 MD iterations. The average acceptance probability $p_{\text{acc}}$ was ∼70%, although the minimum $p_{\text{acc}}$ was ∼10%, corresponding to the replica from MFEP S3 located at $\alpha \simeq 0.5$ [see Fig. 8(a)].

The results of the REX simulations shown in Fig. 8(a) demonstrate that 7 kcal/mol discrepancy in the endpoint free energy difference between simulations G2c/G2f and G3c/G3f was caused by insufficient sampling in the evaluation of the gradient $\nabla G(\boldsymbol{\theta}_n)$. The enhanced sampling of REX is most significant for the replicas near $\alpha \simeq 0.5$ of simulations G3. At this location on the path, the free energy gradients computed without REX are underestimated, lowering the free energy by several kilocalories per mole, relative to the simulations with REX. As noted above, $\alpha \simeq 0.5$ is also the location of minimum $p_{\text{acc}}$ values, which means that the MD replicas that correspond to adjacent images near $\alpha \simeq 0.5$ sample somewhat different regions of space and probably have to overcome greater energy barriers in order to exchange. This explanation is consistent with the fact that the free energy profile in the vicinity of $\alpha \simeq 0.5$ has sharp variations. For simulation G2fr, the minimum of $p_{\text{acc}}$ (∼15%) occurs near $\alpha \simeq 0.8$, which is also the location of a sharp free energy peak [Fig. 8(a)].

In addition to the 128-image REX simulations, 32-replica REX simulations were performed. However, the acceptance probabilities were very low in the vicinity of $\alpha \simeq 0.5$ (MFEP from S3) (<1%), and therefore we did not see substantial improvement in the FE profile over that from G2c (results not shown).

As discussed in Sec. II B, since the free energy $G(\boldsymbol{\theta})$ does not correspond to the free energy as a function of the committor reaction coordinate, it is less informative than $F(\alpha)$ and more difficult to interpret. While REX simulations G2fr/G3fr were necessary to establish the accuracy of the thermodynamic integration Eq. (21) for computing $G(\boldsymbol{\theta})$, we decided that the additional computational cost of performing a 128-replica REX simulation for MFEP S1 was not justified, and

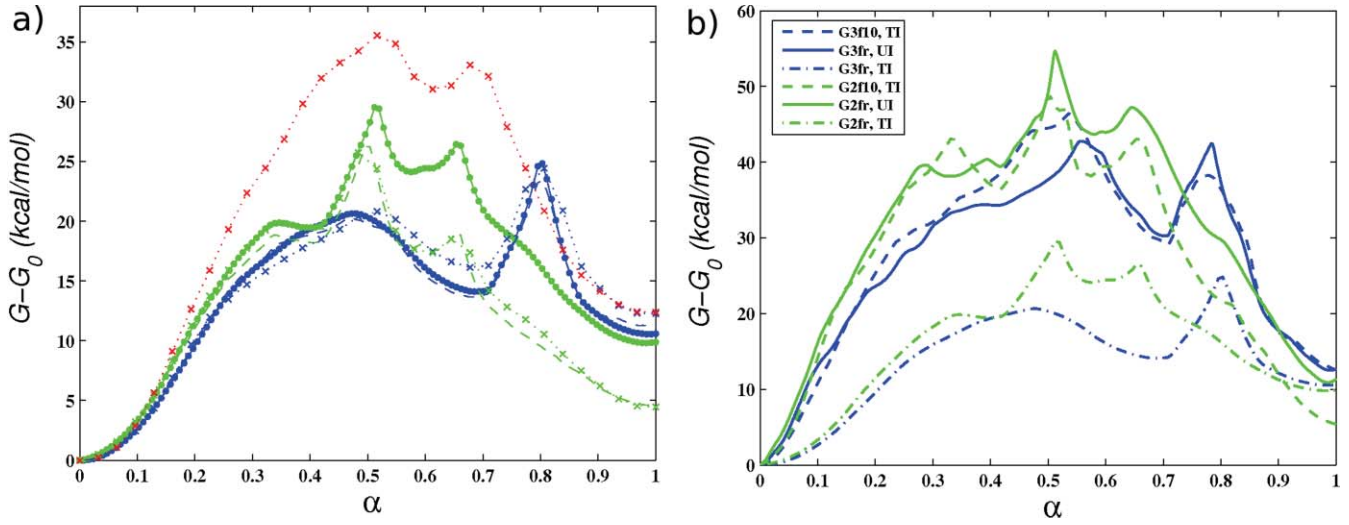


FIG. 8. (a) Free energy of the collective variables ($G$) along the MFEPs (red ×·····), G1c; (green ×·····), G2c; (blue ×·····), G3c; (green - - -), G2f; (blue - - -), G3f; (green •—), G2fr; (blue •—), G3fr. (b) Comparison of the FE $G$ for simulations G2f/G2fr/G2f10 and G3f/G3fr/G3f10 computed using different methods (see text). (Green —), G2fr; (blue —), G3fr. Umbrella integration (corrected gradients with low force constants). (Green - - -), G2f10; (blue - - -), G3f10. Simple trapezoidal integration (uncorrected gradients with high force constants). (Green – · –), G2fr; (Blue – · –), G3fr. Simple trapezoidal integration (uncorrected gradients with low force constants).

this simulation was not performed. The corresponding free energy profile in Fig. 8(a) is therefore less accurate than that for the other simulations.

As described in Sec. II D 1, the errors in the estimation of the gradients $\nabla G(\boldsymbol{\theta}_n)$ due to the use of a restraint instead of a constraint in Eq. (20) are small if the force constants used in the restraint potential are sufficiently large. On the other hand, in order to use REX efficiently in the calculation of FE gradients, the force constants have to be sufficiently low to ensure high acceptance probabilities. To assess the accuracy of the FE profiles from simulations G2 and G3, shown in Fig. 8(a), we corrected the gradients computed from the 128-image REX simulations using UI (Ref. 36) and repeated the integration, as discussed in Sec. II D 1. In addition, we increased all force constants by a factor of 10 and repeated the 128-image simulations (REX was not used because of the high constants), following by integration. These simulations are denoted G2f10 and G3f10 in Table IV.

The resulting FE profiles are compared in Fig. 8(b). In the profiles obtained by integrating the gradients computed from the REX simulations directly (without using UI), the peaks are underestimated almost by a factor of 2, compared to the other two sets of profiles. This indicates that the force constants used in the REX simulations (1.0 kcal/mol/Å$^2$ for position and distance CV to 10.0 kcal/mol/rad$^2$ for dihedral angle CV) are too low to obtain an accurate free energy profile. On the other hand, if UI is used to correct the gradients obtained from REX simulations, the resulting FE profiles agree well with those computed directly from simulations that employ the larger force constants. The good overall agreement also suggests that the higher force constants are sufficiently large to estimate the magnitude of the FE barriers. Unfortunately, in this case, REX cannot be used efficiently, and the gradients are affected by sampling errors, as before (note e.g., the ∼7 kcal/mol endpoint FE difference between the plots corresponding to the high force constants).

It is not surprising that the FE profiles computed directly from simulations that use low force constants underestimate the magnitude of the FE profiles. Equation (17), which is used to approximate gradients of $G(\boldsymbol{\theta})$, becomes exact (assuming no sampling errors) for a "smoothed" free energy defined by,

$$G^*(\boldsymbol{\theta}) = -\beta^{-1} \ln \left\langle C_1 e^{-\beta U(\boldsymbol{x},\boldsymbol{\theta})} \right\rangle$$
$$= -\beta^{-1} \ln[(e^{-\beta G} * \mathcal{N}(\mathbf{0}, (\beta \mathbf{k})^{-1}))(\boldsymbol{\theta})], \qquad (44)$$

in which $U(\boldsymbol{x}, \boldsymbol{\theta})$ is the potential defined in Eq. (16), $\mathcal{N}(\mathbf{0}, (\beta \mathbf{k})^{-1})$ is the multivariate Gaussian distribution centered at zero with variances $(\beta k_i)^{-1}$ for $i = 1, \ldots, K$, $(*)$ is the convolution operation, and $C_1$ is a normalization constant. Equation (44) follows from the definition of $G(\boldsymbol{\theta})$ in Eq. (3) and the properties of the delta function. Thus, the use of the estimate in Eq. (17) is equivalent to applying a Gaussian filter to $\langle \delta(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(\boldsymbol{x})) \rangle$(the probability density of $\boldsymbol{\theta}$) and computing the gradients of the smoother FE $G^*(\boldsymbol{\theta})$. Using low force constants implies that the variances of the Gaussian will be large, which will lead to greater smoothing of the peaks and valleys of the true FE $G(\boldsymbol{\theta})$. Note, also, that in the limit of zero $k_i$, $G^*(\boldsymbol{\theta})$ is identically zero.

The preceding discussion demonstrates that obtaining accurate profiles of the free energy as a function of the collective variables can be laborious. In the present simulations we used replica exchange to improve sampling, which required reinterpolation of the MFEP from 32 to 128 images to obtain high exchange probabilities, quadrupling the computational cost. Furthermore, because the use of REX requires lowering force constants in the restraining potentials, an additional postprocessing step is needed to compute corrected FE gradients by UI.

The definition of $G(\boldsymbol{\theta})$ given in Eq. (3) depends on the collective variables chosen to describe the transition. Not surprisingly, profiles of $G(\boldsymbol{\theta})$ obtained using different CV sets but the same initial paths (i.e., G1 vs G2) in Fig. 8(a) are quite

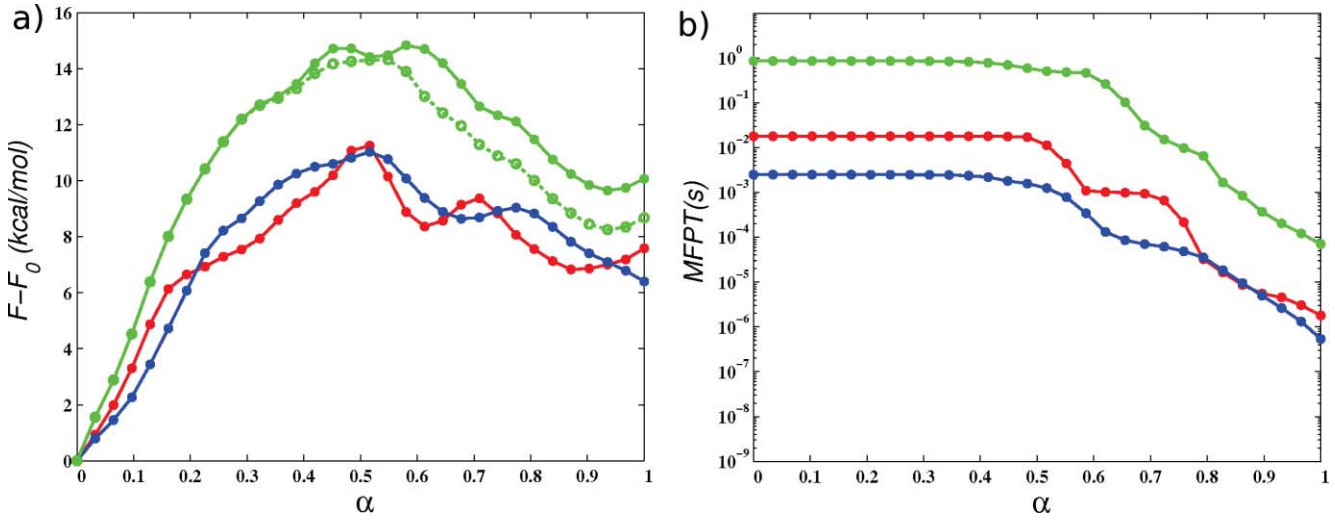


FIG. 9. (a) Free energy of reaction ($F$). (b) MFPT to the milestone nearest to the PPS state. The symbols are as follows: (red •—), F1; (green •—), F2; (green ○ · · ·), corrected F2 (only for a); (blue •—), F3; see text and Table IV for a description of each simulation.

different, although the corresponding MFEPs represent the same transition. We show in Subsection III D 4 that the differences between the different simulations can be accounted for if one considers instead the free energy of the reaction coordinate. We also discuss the correspondence between the structures along the MFEP shown in Fig. 7 and their free energy.

#### 4. One-dimensional free energy profiles and rates of transition

The free energy as a function of the committor reaction coordinate [Eq. (8) in Sec. II B] was computed as described in Sec. II D 2. After the string had converged to the corresponding MFEP, the images $\boldsymbol{\theta}_n$ were fixed, and unrestrained MD simulations were performed concurrently for each cell to estimate the rates of escape $\nu_{m,n}$. Three simulations were carried out, denoted for brevity as free energy simulations F1, F2, and F3, in correspondence to the string simulations S1, S2, and S3. The computational requirements of the free energy simulations were approximately the same as those for the string calculations used to compute the MFEPs and considerably less than those required to compute accurate profiles of $G$ using REX (see Table IV). Convergence of the free energy simulations was assessed by monitoring the total rate of escape from all cells. This quantity reached a plateau after 5–10 ns of MD simulation for simulations F1 and F3 and after ∼15 ns for simulation F2. The apparent reason for the longer time to convergence, required for F2, is a fairly large number of collisions between cells corresponding to nonadjacent images compared to simulations F1 and F3. This issue will be discussed further below because it affects the accuracy of the computed free energy profiles.

After the total rate of escape was stationary, MD integration was continued for 20 ns for simulations F1 and F3 and for 25 ns for simulation F2. Statistics obtained from these trajectory segments were used to compute the free energy and the rate. The FE profiles computed from the entire 20–25 ns trajectory segment were compared to those computed using only the first half of the corresponding segment. The maximum difference was ∼0.4 kcal/mol.

Free energy profiles computed from simulations F1–F3 are shown in Fig. 9(a), and the MFPTs from the milestones $B_i \cap B_{i+1}$ (the "main" milestones, since they approximate isocommittor surfaces, as described in Sec. II D 2) for $i = 1, 2, \ldots, N-2$ to the (last) milestone $B_{N-1} \cap B_N$, are shown in Fig. 9(b). The FE profiles for simulations F1 and F3 are in good agreement, even though the corresponding collective variables sets have different type and size (see Tables II and III). In particular, the FE barriers and the FE difference between the end states are in agreement to within ∼1 kcal/mol. The second FE peak appears to be in slightly different locations ($\alpha \simeq 0.71$ for F1 and $\alpha \simeq 0.78$ for F3). This shift is due to the different parametrizations of the respective MFEPs from S1 and S3. Both are parametrized by arc-length, but the definition of the arc-length involves the collective variables, which are different in type. The actual atomic configurations that correspond to this FE peak are similar for the two paths (compared in the supplementary materials[19]). The profiles of the MFPT for transitions from the main milestones to the last milestone along the paths computed in S1 and S3 are also in fair agreement. The profiles show that the MFPT to the last milestone is approximately constant for the milestones that correspond to $\alpha < 0.5$, indicating that the main FE barrier that occurs near $\alpha = 0.5$ for both paths is the transition "bottleneck."

The FE profile from simulation F2, on the other hand, has consistently higher energies than the profiles from F1 and F3. Examining the rates of escape $\nu_{n,m}$ [from which the free energy is computed via Eqs. (37) and (34)], we observed that in simulation F2, a large fraction (13%) of the total rate of escape from all cells $B_n$ involved pairs of cells that correspond to nonadjacent images (for simulations F1 and F3, this fraction was 0.3% and 0.7%, respectively). Most of the "flux" between nonadjacent cells involved a replica restricted to one of the cells $B_i$ for $i = 14, \ldots, 21$ (93% of the cases), which correspond approximately to $\alpha \in [0.45, 0.67]$ in Fig. 9, and no replicas restricted to an endpoint cell ($i = 0$ or $i = 31$). Although the presence of nonzero fluxes between nonadjacent cells poses no concerns for the estimation of the probabilities

$\pi_n$, it implies that the planar approximation to the isocommittor surface in Eq. (5) is not accurate at certain points away from the MFEP within the transition tube (see Appendix B). The inadequacy of the approximation suggests a problem with assumption (ii) in Sec. II B for simulation S2 (i.e., that the transition tube through which most of the reactive trajectories proceed is narrow).

It should be stressed that this finding does not invalidate the use of CVS2 in general, because the quality of the approximation of isocommittor surfaces by hyperplanes depends not only on the collective variables but also on the curvature of the MFEP and the local free energy landscape (see Appendix B). The fact that the flux between nonadjacent cells in simulation F3 is only 0.3% of the total flux suggests that the present problem does not arise for simulation F3. An approximately corrected free energy profile can be computed by setting the fluxes between nonadjacent cells to zero. This (*ad hoc*) approach corresponds most closely to discarding a portion of the reactive trajectories that are sufficiently far from the MFEP that the hyperplane approximation of the isocommittor surfaces is inaccurate. While accounting for such trajectories properly would produce a more accurate free energy profile, it is impossible to do this without a higher-order approximation to the isocommittor surface.

The corresponding FE profile [included in Fig. 9(a)] shows a significant improvement in the free energy difference between the endpoint states relative to simulations F1 and F3. Since, as mentioned above, 93% of the flux between nonadjacent cells involves MD replicas in cells $B_i$ for $i = 14, \ldots, 21$, and no MD replicas in the endpoint cells, the discarding of the fluxes should affect to the greatest extent the values of the free energy near the barrier ("middle" of the path) and to a lesser extent the FE difference between the endpoints.

One possible reason for the higher FE barrier in simulation F2 relative to simulation F3 is that the FE profile computed in F2 is less accurate due to the spurious transitions between nonadjacent cells described above. In addition, the higher barrier in F2 may be caused by differences in the MFEPs. Recall that calculations S2 and S3 start from initial paths 1 and 2, respectively (see Sec. III C), which specify opposite directions of rotation of the dihedral angles in the loop L4. These dihedrals are explicitly present in the CV set used in simulations S2/G2/F2 and S3/G3/F3 (CVS2), which means that the MFEPs and the transition tubes associated with simulations F2 and F3 cannot intersect. Therefore, paths from F3 may simply be higher in free energy than those from F2. It seems somewhat surprising that the free energy barriers computed from the simulations F1 and F2 differ by $\simeq$3 kcal/mol [Fig. 9(a)]. However, although the corresponding simulations S1 and S2 were both initialized from zero-temperature path 1, some instantaneous MD configurations observed in simulation F1 were found to be more consistent with path 2. This observation suggests that simulation F1 represents a broader ensemble of transition paths, since it contains configurations that are similar to those from both paths 1 and 2. A broader ensemble of paths is likely to result in a lower free energy barrier. As a check of this conjecture, we estimated the width of the transition tube from the free energy simulations (F1–F3) by computing the average,

$$w_n = \langle (\boldsymbol{\theta}_n - \widehat{\boldsymbol{\theta}}(\boldsymbol{x}))^T M^{-1}(\boldsymbol{\theta}_n)(\boldsymbol{\theta}_n - \widehat{\boldsymbol{\theta}}(\boldsymbol{x})) \rangle_{B_n}^{1/2}, \tag{45}$$

for each cell $B_n$.

From the simulations F1, F2, and F3, respectively, we found the values of $w_n$ for cells at $\alpha = 0.5 \pm 0.1$ to be approximately $2w_{\text{ave}}$, $1w_{\text{ave}}$, and $1.5w_{\text{ave}}$, where $w_{\text{ave}}$ is the average width of the transition tube ($w_n < 1w_{\text{ave}}$ at the endpoint states). This is in accord with the suggestion that transition tube associated with MFEP S2 is narrower than those for MFEPs S1 and S3, explaining the higher free energy barrier obtained in simulation F2. The quantity $w_n$, however, is only useful as a qualitative measure of the size of the ensemble of transition paths. It is contaminated by large statistical errors because of the high dimensionality of the CV space (177 and 51 for CVS1 and CVS2). In addition, if the shape of the transition tube is irregular, which is likely to be the case in many dimensions, $w_n$ will not provide an accurate estimate for the volume of the transition tube. We note, also, that that the low apparent width of the transition tube in F2 may be caused by spurious reflections of MD replicas due to collisions with nonadjacent cells (see above).

Having computed profiles of the one-dimensional free energy as a function of the committor, we briefly return to the free energy as a function of the collective variables ($G$). Although computing $G$ accurately requires great care (see Sec. III D 3), the significance of this free energy for the present simulation system appears to be limited, because $G$ cannot be related simply to the reaction free energy $F$. Indeed, although the MFEP lies approximately in the center of the corresponding transition tube, it does not, by itself, contain information on how the volume of this transition tube varies along the path (see Appendix B). From the above estimates of the width of the transition tubes for simulations F1–F3, we can conclude that the transition tube is generally larger at intermediate states along the path compared to the endpoint states. This indicates that the entropic contribution to the reaction FE neglected in $G$ due to the "freezing" of the $K$ collective variables is nonuniform along the MFEP. Consistent with this interpretation, a comparison of Figs. 8(b) and 9(a) shows a dramatic reduction in the free energy values at intermediate locations along the free energy profiles, going from $G$ to $F$. Higher entropic contributions to the FE at intermediate values of the reaction would imply that, in Cartesian space, more configurations are required to characterize these intermediate states.

The curve in Fig. 10 shows the RMSD computed between the structures in each cell $B_n$ and the initial structure (a corresponding replica on the MFEP) for simulation F3. The RMSD is generally higher for intermediate cells than for the endpoint cells, and the highest value occurs at the location of the FE barrier [$\alpha \simeq 0.5$ in Fig. 9(a)]. Surrounding the RMSD curve, we show overlays of four configurations taken from simulation F3 that correspond to $\alpha = 0$, 0.5, 0.8, and 1.0. The intermediate values of $\alpha$ correspond to the two FE peaks in Fig. 9(a) and also to the snapshots I15 and I25 on the MFEP shown in Fig. 7. The larger differences between the overlaid conformations are evident for $\alpha = 0.5$ and 0.8, than for $\alpha = 0$ and

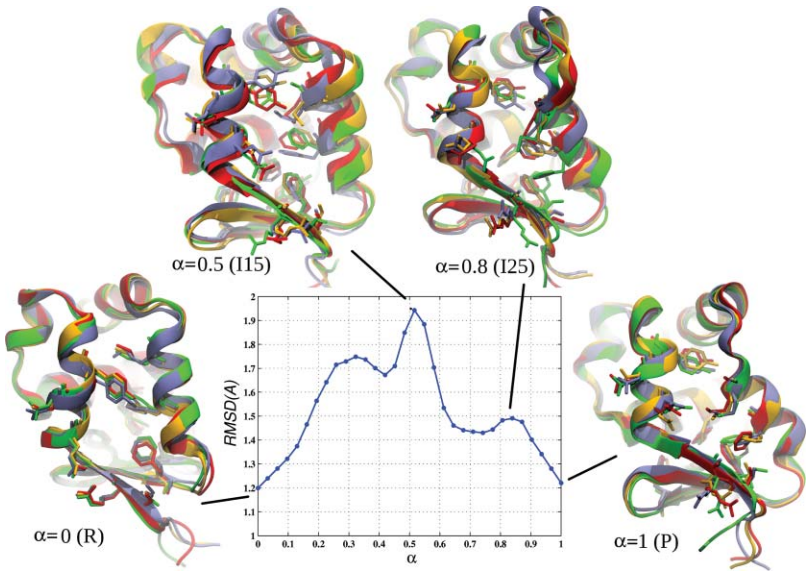


FIG. 10. The curve corresponds to the RMSD computed between the structures in each cell $B_n$ and the initial structure in each cell for simulation F3. The structures correspond to instantaneous MD configurations from simulation F2 in Table IV constrained to cells $B_0$, $B_{15}$, $B_{25}$, and $B_{31}$.

1.0, consistent with the RMSD plot, and suggest that transitions occur via many paths.

Despite the differences in the magnitudes of $G$ and $F$, a qualitative comparison of Figs. 8(b) and 9(a) suggests that profiles of $G$ and $F$ are correlated, so that based on $G$, one can identify correctly not only the relative stability of the endpoint states but also configurations along the MFEPs that correspond to high-energy and metastable states. However, if one aims to compute a one-dimensional free energy profile (versus a parametric curve on a multidimensional landscape) and rates of transition, in addition to the calculation of the MFEP(s), separate free energy simulations are required.

### 5. Estimates of energy and entropy of transition

To estimate the energetic and entropic contributions to the free energy profiles in Fig. 9(a), we computed the average potential energy in each cell $B_n$ from the trajectories in F1–F3. The corresponding energy profiles are shown in Fig. 11. The standard deviation of the energies in each cell was ∼27 kcal/mol, and the standard deviation of the averages shown in the figure is about 9 kcal/mol (based on computing nine block averages ∼2 ns in length). Because of the large error bars, the relative contributions of the energy and entropy discussed herein should be considered qualitative.

We note that that the endpoints have approximately equal energies, which indicates that the FE difference between the R and PPS conformers of the converter (∼8 kcal/mol) is of entropic origin. As a rough check of this conclusion, we estimated the configurational entropy of the R and PPS conformers from quasiharmonic analysis of 60 ns unbiased MD simulations of the two conformers. Using the harmonic oscillator formula in Ref. 65, the value for $-T\Delta S$ for the PPS state was ∼17 kcal/mol higher than that for the R state, qualitatively consistent with the main results.

A second observation from Figs. 11 and 9(a) is that for intermediate values of the progress variable $\alpha$, the entropic contribution to the FE is similar in magnitude to the energetic contribution (∼10 kcal/mol). This finding is qualitatively consistent with the larger widths of the transition tube in the vicinity of $\alpha \simeq 0.5$ than at the endpoints, as discussed in Subsection III D 4, and the somewhat broader conformational ensembles found for MD trajectories restricted to intermediate cells $B_n$ (Fig. 10). This result underscores the important point that a single path connecting two states separated by a barrier is unlikely to provide a full description of the transition. However, this does not impact the utility of the MFEP as an "average" path, i.e., located in the center of a transition tube that may itself be quite wide.

## IV. DISCUSSION OF METHODOLOGY

In the study of the prepowerstroke ↔ rigor transition of the converter domain of myosin VI, several challenges were addressed that are likely to arise in the application of the string method to other complex biomolecular systems. The choice of

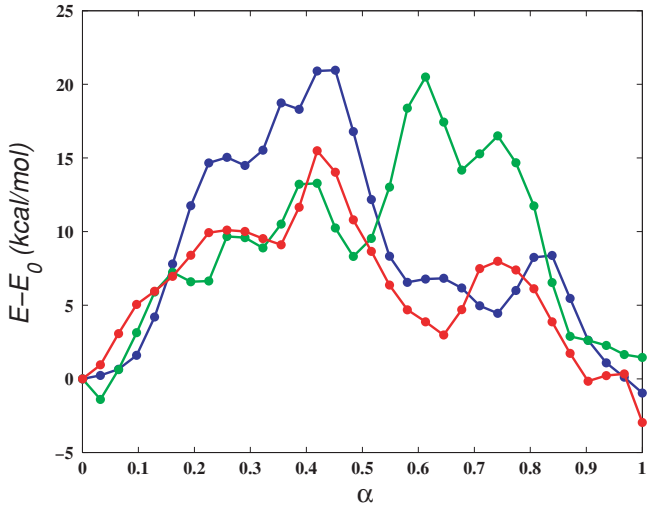


FIG. 11. Average potential energy of the tessellations $B_n$; (red •—), F1; (green •—), F2; (blue •—), F3.

CVs to describe the transition under study is far from obvious. In the present approach, we first found a set of atoms, such that applying forces to just these atoms in a targeting simulation (RTMD) drives either conformation toward the other. To be confident that the final RTMD simulation structure was in a local minimum corresponding to the target, the structure was "relaxed" by gradually removing the restraints and allowing it to equilibrate without external forces; the relaxed structure was within 1.6 Å (heavy-atom RMSD) of the target. Using this atomic set, we constructed two sets of CV. The first CV set is comprised of just the positions of the atoms, expressed in an internal frame of reference. In the second set, the number of CV is reduced by a factor of 3 compared to the first set. Since any CV set that can be used to parametrize the committor function with reasonable accuracy should represent the behavior of the system equally well (and, in particular, lead to the same free energy profiles and rates), using the two different CV sets allows us to "cross-validate" the results. The fact that we have obtained similar paths and reaction free energies for the transition in the MVI converter indicates the robustness of the present approach.

To use the string method, one must have initial values for the CV as well as complete coordinates for the corresponding MD replicas. Starting with a linear interpolant in Cartesian coordinates between the two structures, a minimum energy path was obtained with the zero-temperature string method. Although linear interpolation produces distorted intermediate structures, subsequent energy minimizations restore equilibrium bond lengths and angles, and alleviate bad contacts. We note that other interpolation algorithms can be used for zero-temperature path generation, such as the chain-of-states method,[66] the nudged elastic band method[67] and conjugate peak refinement.[68] The main advantage of the zero-temperature string over these methods is that it can be used "out-of-the-box" with many solvation models in CHARMM, including FACTS. Although initial conditions to the string methods can also be generated from targeted molecular dynamics simulations with constraints[57,58] or restraints[22] (as well as TMD enhanced with Monte-Carlo sampling in trajectory space[69]), we did not employ such methods. Since TMD simulation structures in general are not in equilibrium (i.e., far from an MFEP), the direction in which TMD is performed can introduce directional bias into the transition path (this was also found for the symmetric TMD of Ref. 22). In contrast, in (symmetric) interpolation methods, endpoint structures contribute equally.

Two types of free energies were used in this study [see Eqs. (3) and (8)]. $G$ is the free energy as a function of the collective variables along the MFEP, and $F$ is the free energy of reaction as a function of the committor function $q$ (plotted versus the parameter $\alpha$). Thus, the free energy $G$ is a $K$-dimensional function, with $K$ equal to the number of collective variables, whereas $F$ is one-dimensional. The profiles of $G$ computed in this study are evaluated only along the corresponding MFEP and have no information about the free energy values for points not on the MFEP (other than that they are higher in free energy locally near the MFEP). In contrast to $G$, the reaction free energy $F$ maps the entire transition tube onto a single curve. In the special case that the transition tube is extremely narrow or has uniform cross-sectional volume along the path, $G \simeq F$. If the volume of the tube cross section is variable (which is true for the present system), $G$ does not provide an accurate measure of the free energy of the reaction. A qualitative comparison of the profiles of $G$ and $F$ [Figs. 8(b) and 9(a)] suggests that $G$ can be used to identify high-energy states and metastable intermediates.

The function $F$ can be calculated by taking a Boltzmann average of $G$ along all reactive trajectories that pass through the transition tube with the same value of the committor function $q(\alpha)$. This approach is equivalent to the Boltzmann-weighted integration of $G$ over the isocommittor surfaces. To perform this integration, we make use of assumption (ii) in Sec. II B and approximate the isocommittor surfaces by hyperplanes of $\boldsymbol{\theta}(\boldsymbol{x})$ that are perpendicular to the MFEP (scaled by $M^{-1}$). The free energy $F$ is computed by tessellating the configurational space into cells with boundaries that coincide locally with the hyperplanes. MD is then used to estimate the rate of escape from each cell by recording the number of collisions between the cell boundaries.

If assumption (ii) is satisfied, then the hyperplane approximation to the isocommittor surface is accurate, and only collisions between adjacent cells will be observed. Conversely, a significant number of collisions between nonadjacent cells indicates a failure of assumption (ii). In this case, $F$ defined by Eq. (8) will not be an accurate approximation of the true free energy of the reaction coordinate (i.e., FE of the committor).

In one of the free energy simulations performed, 13% of all collisions were between nonadjacent cells (in the other two cases, this number was less than 1%), and the endpoint free energy difference computed from this simulation differed from the other two. In a corrected free energy profile, which was calculated by setting the fluxes between nonadjacent cells to zero, the free energy difference was in agreement with the other two profiles.

The preceding discussion suggests that optimal sets of CV for studying transitions are those which correspond to relatively narrow transition tubes. It is likely that the constructions of appropriate CV will require specific analysis of each transition.

Although the use of the FACTS model to represent the effects of solvation is unrelated to the string method per se, it was essential for the present calculations. A simulation with explicit water molecules would increase the size of the system to approximately 17 000 atoms, compared to 1442 atoms required with FACTS. The calculation of additional MD forces arising from FACTS was found to slow down the overall speed of the calculation by a factor of 2. Therefore, the computational cost of performing MD simulations with explicit solvation would be roughly six times greater.

Reaction rate calculations with implicit solvent models require inclusion of realistic solvent friction. To improve conformational sampling, in the present study we used a simple Langevin friction model with a very low value (1 $ps^{-1}$) for the friction coefficient (see Sec. III D 1). This choice of friction kernel will most likely lead to an overestimation of the rate. An approximate magnitude of the error can be obtained by considering a friction model based on hydrodynamic interactions. Models in which atoms or atom groups are treated as

point particles interacting via the Oseen or the Rotne–Prager tensors[70,71] have led to good agreement between the computed and experimentally determined translational and rotational diffusion tensors[72] [provided that a hydration shell of thickness 1.1–1.5 Å (Refs. 72 and 73) is modeled around the protein]. Such a model could be incorporated into the present calculations via, e.g., the Langevin equation[74] using a parametrization of the friction constants based on Stokes' law and an accessible surface area model.[75,76] Venable and Pastor[76] found that if the friction coefficient in their hydrodynamic model is reduced by a factor of 10, the computed diffusional and rotational tensors increase by factors of $\simeq$6 and $\simeq$15, respectively. Reference 76 recommends using mass-scaled friction constants of $\simeq$50 $ps^{-1}$ compared to 1 $ps^{-1}$ used in the present study. These considerations suggest that with a realistic solvent friction model the rate of transition would be slowed by an order of magnitude. Another concern with solvation models is the long-time stability of the protein structure. In the present calculations, the two endpoint structures were stable for around 100 ns of simulation (including equilibration and string calculations). FACTS was thus an excellent choice for the myosin VI converter domain, but this may not be true for all systems.

## V. CONCLUDING DISCUSSION

The present results demonstrate the utility of the string method in computing transition paths, free energy profiles and rates of transition in complex biomolecular systems. In particular, the prepowerstroke $\leftrightarrow$ rigor transition of myosin VI considered here is complicated in that it involves rearrangements of residue side chains, large motions of alpha helices, and changes in the backbone structure of a flexible loop (see Sec. III D 1 and Fig. 7). The presence of diffusive motion along the transition path is evident in the behavior of unbiased trajectories launched from the transition state found in one of the simulations: the evolution of the system projected onto the reaction coordinate is slow, with frequent reversals of the direction of motion (see Appendix C). Such behavior makes it difficult to use transition path sampling,[1,2] because the trajectories would require very long integration times to reach the endpoint structures. Since the reaction coordinate for transitions in complex biomolecular systems is usually unknown, the ability of the string method to use a fairly large number of collective variables in the approximation of the reaction coordinate (defined here as the committor function) is an essential aspect of the method. In contrast, other methods based on collective variables, such as metadynamics[6] or adaptive biasing force,[8] require their number to be rather small. Thus, the string method, although not trivial to use (see Sec. II), appears to be well-suited for obtaining transition paths and free energies in complex biomolecular systems.

The present study of the prepowerstroke $\leftrightarrow$ rigor transition of the myosin VI converter is an important step toward the understanding of the powerstroke.[11–13] The mean first passage times of the transition computed from the simulations [$10^{-3}$ –$10^{-2}$ s; see Fig. 9(b)] are consistent with an experimental study that find the rate-limiting step of the powerstroke transition in myosin VI to be $\leq 90$ $s^{-1}$.[77] (However, as mentioned at the end of Sec. II D 3, the rate calculation does not take into account the implicit solvent approximation, and probably overestimates the rate by an order of magnitude.) The computed free energy profiles reveal the structures that have the highest free energies along the transition paths and that correspond to the transition states, as shown by an analysis of unbiased trajectories (see Appendix C). The predicted structures can be tested by experimental mutagenesis studies, e.g., by mutating residues Y718, F766, or M770 (see Fig. 7). In addition, the simulations predict the existence of a metastable state along the transition path at $\alpha \simeq 0.65$ [see Fig. 9(a)], which suggests that the lever arm of myosin VI may occupy a position that is intermediate between the rigor and prepowerstroke states, and could help to explain the variable step size observed for myosin VI dimers.[78,79] Finally, we find that the prepowerstroke converter conformation is higher in free energy than the rigor conformation by $\simeq$8 kcal/mol. This result should be viewed with caution, however, because only the converter domain was simulated in the present study; the rest of the myosin VI molecule that was excluded from the simulations to decrease computational cost may preferentially stabilize one conformation of the converter. Calculations that treat the entire myosin VI head are in progress to examine this question.

## ACKNOWLEDGMENTS

The authors are grateful to Giovanni Ciccotti, Maddalena Venturoli, and Kwangho Nam for stimulating discussions. Supercomputer resources for the calculations performed in this study were provided by the National Energy Resource Supercomputing Center and the FAS Research Computing Group at Harvard. V.O. acknowledges financial support under the NRSA fellowship 1F32GM083422-01. The work done at Harvard was supported in part by a grant from the National Institutes of Health and by a grant from the Human Frontiers Grant Science Program. E.V.-E. was supported by a grant from the National Science Foundation (Grant No. DMS-0708140) and by a grant from the Office of Naval Research (Grant No. N00114-04-1-6046).

## APPENDIX A: DERIVATIVES OF THE POSITION CV

Let $A = [\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$ and let $o = [o_1, o_2, o_3]$ denote the origin of the coordinate frame $A$ ($o$ is the COM of the atom group that defines the frame). Let $p^*$ denote the absolute coordinates of a point $P$. In the frame of $A$, the coordinates of $P$ are $p = A^T (p^* - o)$, and we have used the orthogonality of $A$. Differentiating this expression with respect to an absolute atomic coordinate, $r_i^*$, we obtain

$$\frac{\partial p}{\partial r_i^*} = \frac{\partial A^T}{\partial r_i^*}\left(p^* - o\right) + A^T\left(\frac{\partial p^*}{\partial r_i^*} - \frac{\partial o}{\partial r_i^*}\right). \tag{A1}$$

The only nontrivial derivative in Eq. (A1) is $\partial A^T/\partial r_i^*$. Since $A$ depends on $r_i^*$ through $C_{i,j}$, i.e., $A\left(r_i^*\right) = A\left(C_{i,j}\left(r_i^*\right)\right)$, we first compute the derivative of $A$ with respect to the components of $C_{i,j}$. Let $\boldsymbol{v}$ be an eigenvector and $\lambda$ the corresponding eigenvalue. Differentiating the eigenvalue relation

$(C_{i,j} - \delta_{ij}\lambda)v_j = 0$ with respect to $C_{p,q}$, we have

$$\delta_{ip}\delta_{jq}v_j - \frac{\partial\lambda}{\partial C_{p,q}}v_i + (C_{i,j} - \delta_{ij}\lambda)\frac{\partial v_j}{\partial C_{p,q}} = 0, \tag{A2}$$

where $v_i$ is the $i$th component of $\boldsymbol{v}$. Multiplying Eq. (A2) by $v_i$ and applying the eigenvalue relation yields

$$\delta_{ip}\delta_{jq}v_iv_j - \frac{\partial\lambda}{\partial C_{p,q}}v_iv_i + v_i(C_{i,j} - \delta_{ij}\lambda)\frac{\partial v_j}{\partial C_{p,q}} = 0, \tag{A3}$$

$$v_pv_q - \frac{\partial\lambda}{\partial C_{p,q}} = 0, \tag{A4}$$

$$\frac{\partial\lambda}{\partial C_{p,q}} = v_pv_q, \tag{A5}$$

where we have assumed that the eigenvectors $\boldsymbol{v}_i$ have been normalized to unity and summed on $i$ and $j$. Inserting Eq. (A5) into Eq. (A3) and rearranging, we obtain a following matrix equation that involves the eigenvector derivative:

$$(C_{i,j} - \delta_{ij}\lambda)\frac{\partial v_j}{\partial C_{p,q}} = v_q(v_pv_i - \delta_{ip}). \tag{A6}$$

Equation (A6) cannot be solved because $C_{i,j} - \delta_{ij}\lambda$ is singular. The singularity can be removed by augmenting Eq. (A6) with the relation

$$\frac{1}{2}\frac{\partial(\boldsymbol{v}\cdot\boldsymbol{v})}{\partial C_{p,q}} = v_j\frac{\partial v_j}{\partial C_{p,q}} = 0, \tag{A7}$$

which expresses the fact that the norm of the eigenvector $v$ is constant (which we set to unity). Equations (A6) and (A7) can now be inverted to yield $\partial v_j/\partial C_{p,q}$. The procedure is carried out for each eigenvalue/eigenvector pair. The derivative $\partial C_{i,j}/\partial r^*_{qk}$ (where $r^*_{qk}$ is the $k$-component of the absolute position of atom $q$) is computed by differentiating Eq. (42) (with $r$ replaced by $r^*$),

$$\frac{\partial C_{i,j}}{\partial r^*_{qk}} = m_q\left(\left(r^*_{qj} - \widehat{r^*}_{qj}\right)\delta_{ik} + \left(r^*_{qi} - \widehat{r^*}_{qi}\right)\delta_{jk}\right). \tag{A8}$$

The derivative $\partial A^T_{i,j}/\partial r^*_k$ can now be computed as

$$\frac{\partial A^T_{i,j}}{\partial r^*_k} = \frac{\partial A_{j,i}}{\partial r^*_k} = \sum_{p=1}^{3}\sum_{q=1}^{3}\frac{\partial \boldsymbol{v}_{ji}}{\partial C_{p,q}}\frac{\partial C_{p,q}}{\partial r^*_k}, \tag{A9}$$

where $\boldsymbol{v}_{ji}$ is the $i$th component of $\boldsymbol{v}_j$.

We note that position CVs expressed in a local moving frame are somewhat more costly to calculate than those in an absolute frame, since the tensor $C_{i,j}$ must be diagonalized at every MD iteration, and computing eigenvector derivatives requires that Eqs. (A6) and (A7) to be solved six times for each eigenvector (only six inversions are necessary because of the symmetry in $C_{i,j}$). Furthermore, each derivative in Eq. (A1) requires two matrix multiplications. To increase the speed of matrix diagonalization, a special-purpose routine for $3 \times 3$ symmetric positive semidefinite matrices was implemented in CHARMM, in which the cubic characteristic polynomial is solved using Cardano's formula, and the eigenvectors are determined using Cramer's rule. In addition, the $3 \times 3$ matrix multiplications were programmed in-line to avoid function calls.

Since, for a right-handed coordinate frame, one can invert any two coordinate vectors simultaneously to generate another right-handed coordinate frame, an additional constraint on the eigenvectors is needed to guarantee a unique solution for the coordinate frame. (There are four ways to define a right-handed frame using three orthonormal vectors.) Given the coordinate frame computed at the previous timestep, $\left[\boldsymbol{v}_1^{\text{prev}}, \boldsymbol{v}_2^{\text{prev}}, \boldsymbol{v}_3^{\text{prev}}\right]$, the new coordinate frame $[\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3]$ is chosen from the four possible right-handed frames such that the scalar product

$$\boldsymbol{v}_1^{\text{prev}}\cdot\boldsymbol{v}_1 + \boldsymbol{v}_2^{\text{prev}}\cdot\boldsymbol{v}_2 + \boldsymbol{v}_3^{\text{prev}}\cdot\boldsymbol{v}_3, \tag{A10}$$

is maximal. This criterion ensures that the frame vectors evolve continuously during the simulation. The coordinate frame at the first step of the simulation is chosen randomly from the four possible definitions. To compute the coordinate frame vectors consistently for the images along the string, a similar criterion is used. Given two adjacent images, $i$ and $i+1$, and a best-fit rotation matrix $A$ that aligns the atoms (those used to define the frame) of image $i+1$ with those of image $i$ in the sense of minimal RMSD, the coordinate frame chosen for the image $i+1$ is that which maximizes the product

$$A\boldsymbol{v}_1^{i+1}\cdot\boldsymbol{v}_1^i + A\boldsymbol{v}_2^{i+1}\cdot\boldsymbol{v}_2^i + A\boldsymbol{v}_3^{i+1}\cdot\boldsymbol{v}_3^i. \tag{A11}$$

The constraint in Eq. (A11) is enforced explicitly at the beginning of a string simulation, and is guaranteed to hold at each step of the simulation in view of the constraint in Eq. (A10).

## APPENDIX B: A DISCUSSION OF FLUXES BETWEEN NON-ADJACENT CELLS

A tessellation of a hypothetical two-dimensional space of collective variables shown in Fig. 12 can be used to understand the implication of fluxes between nonadjacent

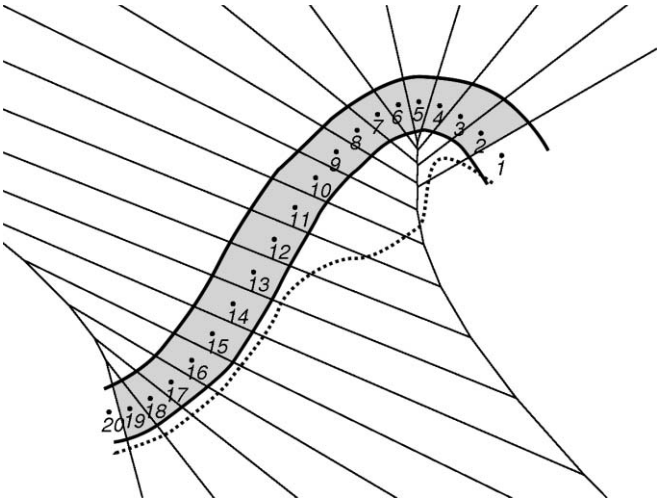


FIG. 12. A tessellation of a two-dimensional space of collective variables based on a hypothetical MFEP discretized into 20 images. The cells $B_n$ are drawn in thin solid lines. Two putative transition tubes are shown with boundaries drawn in thick solid and dotted lines (the upper dotted line coincides with the solid line). The area inside the first transition tube (solid lines) is shaded in gray. The numbers correspond to the images and to the associated cells $B_n$.

images. In the ideal case, trajectories initiated from the points the MFEP will stay within the transition tube bounded by the two thick black lines. Then, for any internal image point $i$, only the fluxes $B_i \leftrightarrow B_{i-1}$ and $B_i \leftrightarrow B_{i+1}$ are possible (these are fluxes between adjacent cells). If, on the other hand, the ensemble of trajectories is bounded by the dotted line, trajectories launched inside cells $B_i$ for $i$=2,3,4,6,7,8,9,10 will collide with boundaries that correspond to nonadjacent cells. For example, trajectories in $B_9$ and $B_{10}$ will collide with the cell $B_1$. In this case, some hyperplanes $B_i \cap B_{i+1}$ provide poor approximations to the corresponding isocommittor surfaces in regions through which reactive trajectories pass with a significant probability. Indeed, if all of the hyperplanes coincide precisely with isocommittor surfaces, then it will be impossible for any trajectory restricted to a cell $B_i$ to cross into any cell other than $B_{i\pm1}$, since this would imply that two isosurfaces that correspond to different values of the committor function must intersect.

The effect of the curvature of the MFEP on the occurrence of fluxes between nonadjacent cells can be understood by referring to Fig. 12. For replicas 10–16, the path is approximately straight, i.e., the curvature is low, and the boundaries of adjacent cells $B_n$ are approximately parallel. The figure suggests that the widest transition tube that is located inside $B_i$ for $i = 10, \ldots, 16$ and that does not include boundaries between nonadjacent cells is much larger than the one shaded in gray. For replicas 2–7, however, the path is curved, and even a slightly larger transition tube contains boundaries between no-adjacent cells (e.g., $B_4$ and $B_6$).

The curvature of the MFEPs computed in this study was approximated using the formula

$$C(n) = \frac{|\boldsymbol{\theta}_{n+1} - 2\boldsymbol{\theta}_n + \boldsymbol{\theta}_{n-1}|}{|\boldsymbol{\theta}_{n+1} - \boldsymbol{\theta}_n|^2}, \quad n = 1, \ldots, N-1, \tag{B1}$$

which uses a second-order finite-difference approximation of second derivatives of the MFEPs and where explicit dependence on time $t$ has been omitted.

For the values given below, $C(n)$ was averaged over the duration of the simulations S2 and S3. In the vicinity of $\alpha = 0.5$, (n $\simeq$ 15), $C \simeq 1.5$ for MFEP S2, and $C \simeq 1.1$ for MFEP S3; at the endpoints ($n = 1$ and $n = 30$), $C \simeq 1$ for both MFEPs. This difference may explain why the flux between nonadjacent cells in simulation F2 is significantly larger than that in simulation F3. We also computed the curvature for the MFEP from simulation S1. We found that $C(n) \simeq 1.1$ at the endpoints of the path and $C(n) \simeq 1.3$ for intermediate values of $n$. However, the curvature from MFEP S1 should not be compared directly to that from MFEP S2 (or MFEP S3) because the CVs are different (see Tables II and III).

## APPENDIX C: ANALYSIS OF THE TRANSITION STATE IN F3 USING UNBIASED TRAJECTORIES

Whether a given region (or surface) of configurational space corresponds to a transition state is typically tested by commitment analysis, which involves launching trajectories from configurations sampled from the region and checking whether the trajectories reach the endpoint structures with equal probability.

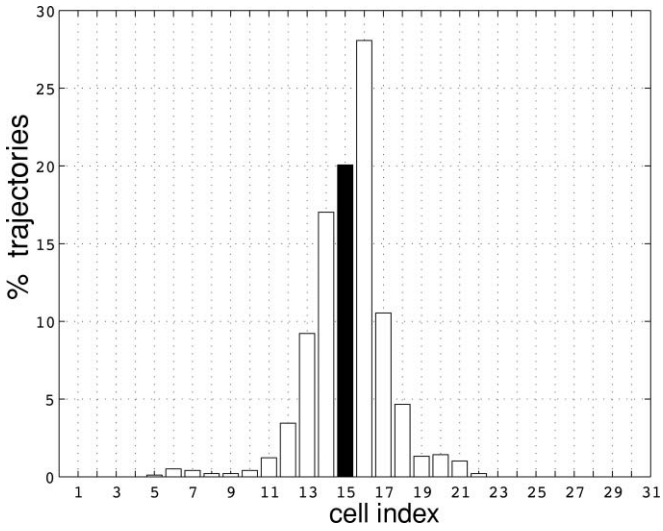


FIG. 13. Percent histogram of the indices of the cells $B_n$ computed for the final structures of the unbiased trajectories launched from configurations in the cell $B_{15}$ (filled in black color) of simulation F3.

Because the full committor test was computationally infeasible for the present system due to the long integration times involved (see below), the following (less demanding) test was performed. Twenty-five instantaneous structures were chosen randomly from the MD trajectory from simulation F3 confined to cell $B_{15}$ ($\alpha \simeq 0.5$). For each structure, 40 sets of momenta were generated according to the Maxwell–Boltzmann distribution using different random seeds, and $25 \times 40 = 1000$ 4-ns trajectories were generated by regular MD at 300 K (with the same thermostat parameters as used for the main calculations). For the final structure in each trajectory, the corresponding cell $B_n$ was determined by using the criterion in Eq. (29). The set of final cell indices is shown in a histogram in Fig. 13. We find that none of the trajectories reach the reactants or the products in the allotted simulation time. In addition, most of the trajectories do not travel far, terminating in the cells $B_{14}$, $B_{15}$, and $B_{16}$. In particular, 20% of the trajectories remain in the initial cell ($B_{15}$), suggesting that long integration times would be needed for these trajectories to commit to an endstate. We also find that 41% and 59% of the trajectories that escape from $B_{15}$ terminate on the reactant side ($B_n$ for $n < 15$) and the product side ($B_n$ for $n > 15$), respectively. Since the committor values are known to change rapidly in the vicinity of a transition state,[29] the 41%/59% ratio should be considered satisfactory.

In Fig. 14 we show the time series of the index of the cells $B_n$ computed for three unbiased trajectories. The two series shown in blue and green were chosen randomly from the entire sample of the trajectories; as expected from Fig. 13, the corresponding trajectories terminate near the starting cell in cells $B_{16}$ and $B_{17}$. The time series drawn in red correspond to a rare trajectory chosen such that it terminates in cell $B_6$ (see Fig. 13). The figure shows that the evolution of the trajectories is rather slow, with frequent reversals of the direction of motion, consistent with diffusive behavior.

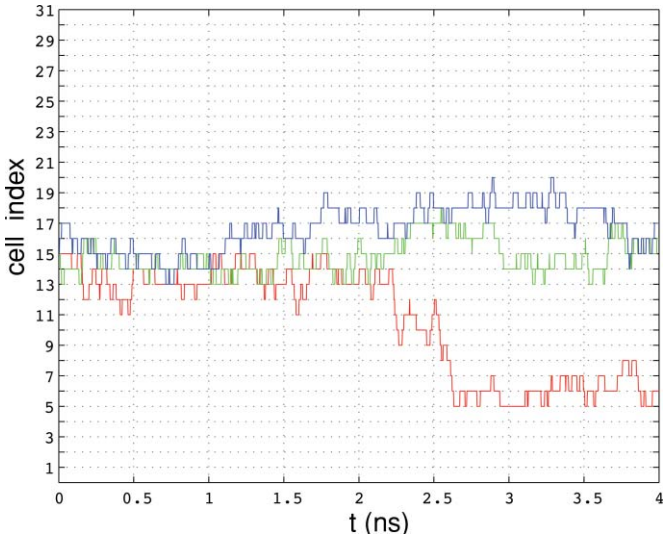


FIG. 14. Time series of the cell index for three unbiased trajectories launched from configurations in the cell $B_{15}$ (see text).

## APPENDIX D: FREE ENERGY PROFILES COMPUTED BY KRIVOV AND KARPLUS

In this appendix, we describe the difference between the one-dimensional free energy profiles computed in the present study and those computed by Krivov and Karplus in Sec. 2.4 of Ref. 27. Krivov and Karplus construct an MSM from long MD simulation trajectories, with the system states characterized by trajectory clustering, and the rates of transitions between the states computed directly from the trajectory. The values of the committor function are evaluated directly from the MSM, and the partition function $Z(q^* - \Delta < q(\boldsymbol{x}, \boldsymbol{p}) < q^* + \Delta)$ of a region that corresponds to a particular value $q^*$ of the committor function is proportional to the number of transitions across the $q(\boldsymbol{x}, \boldsymbol{p}) = q^*$ surface, computed using the minimum-cut[80] and the balanced-cut algorithms.[27] The difference from the present definition of the free energy profile is that the free energy is plotted as a function of the progress coordinate $r(q^*) = Z(q(\boldsymbol{x}, \boldsymbol{p}) < q^*)/Z(q(\boldsymbol{x}, \boldsymbol{p}) \leq 1)$. The use of $r(q^*)$ ensures that the computed free energy profile is invariant with respect to arbitrary invertible transformations of $r(q^*)$.[27,28] The two types of free energy profiles are related simply by a transformation of the abscissa; $r(q^*)$ can be computed from the present results using Eqs. (6)–(8).

[1]C. Dellago, P. Bolhuis, and P. Geissler, Adv. Chem. Phys. **123**, 1 (2002).
[2]P. Bolhuis, D. Chandler, C. Dellago, and P. Geissler, Annu. Rev. Phys. Chem. **53**, 291 (2002).
[3]W. E, W. Ren, and E. Vanden-Eijnden, Phys. Rev. B **66**, 052301 (2002).
[4]W. E, W. Ren, and E. Vanden-Eijnden, J. Phys. Chem. B **109**, 6688 (2005).
[5]L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, J. Chem. Phys. **125**, 024106 (2006).
[6]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U.S.A. **99**, 12562 (2002).
[7]M. Iannuzzi, A. Laio, and M. Parrinello, Phys. Rev. Lett. **90**, 238302 (2003).
[8]J. Hénin, G. Fiorin, C. Chipot, and M. L. Klein, J. Chem. Theory Comput. **6**, 35 (2010).
[9]A. Faradjian and R. Elber, J. Chem. Phys. **120**, 10880 (2004).
[10]D. Shalloway and K. Faradjian, J. Chem. Phys. **124**, 054112 (2006).
[11]J. Ménétrey, A. Bahloul, A. Wells, C. Yengo, C. Morris, H. L. Sweeney, and A. Houdusse, Nature (London) **435**, 779 (2005).
[12]J. Ménétrey, P. Llinas, M. Mukherjea, H. Sweeney, and A. Houdusse, Cell **131**, 300 (2007).
[13]J. Ménétrey, P. Llinas, J. Cicolari, G. Squires, X. Liu, A. Li, H. Sweeney, and A. Houdusse, EMBO J. **27**, 244 (2008).
[14]W. E and E. Vanden-Eijnden, J. Stat. Phys. **123**, 503 (2006).
[15]E. Vanden-Eijnden, *Computer Simulations in Condensed Matter: From Materials to Chemical Biology*, edited by M. Ferrario, G. Ciccotti, and K. Binder (Springer, Berlin, 2006),Vol. 1, pp. 439–478.
[16]T. F. Miller III, E. Vanden-Eijnden, and D. Chandler, Proc. Natl. Acad. Sci. U.S.A. **104**, 14559 (2007).
[17]L. Maragliano and E. Vanden-Eijnden, Chem. Phys. Lett. **446**, 182 (2007).
[18]E. Vanden-Eijnden and M. Venturoli, J. Chem. Phys. **130**, 194101 (2009).
[19]See supplementary material at http://dx.doi.org/10.1063/1.3544209 for a summary of the zero-temperature string method, the alanine dipeptide test case, and for further discussion of MFEPs S2 and S3; for an animation of MEP1; for an animation of MEP2; for an animation of MFEP S3.
[20]W. E, W. Ren and E. Vanden-Eijnden, J. Chem. Phys. **126**, 164103 (2007).
[21]E. Vanden-Eijnden and M. Venturoli, J. Chem. Phys. **130**, 194103 (2009).
[22]J. Apostolakis, P. Ferrara, and A. Caflisch, J. Chem. Phys. **110**, 2099 (1999).
[23]P. Metzner, C. Schütte, and E. Vanden-Eijnden, J. Chem. Phys. **125**, 084110 (2006).
[24]R. Du, V. Pande, A. Grosberg, T. Tanaka, and E. Shakhnovich, J. Chem. Phys. **109**, 334 (1998).
[25]A. Ma and A. Dinner, J. Phys. Chem. B **109**, 6769 (2005).
[26]W. E, W. Ren, and E. Vanden-Eijnden, Chem. Phys. Lett. **413**, 242 (2005).
[27]S. Krivov and M. Karplus, J. Phys. Chem. B **110**, 12689 (2006).
[28]S. Krivov and M. Karplus, Proc. Natl. Acad. Sci. U.S.A. **105**, 13841 (2008).
[29]R. Best and G. Hummer, Proc. Natl. Acad. Sci. U.S.A. **102**, 6732 (2005).
[30]I. Khavrutskii, K. Arora, and C. Brooks III, J. Phys. Chem. **125**, 174108 (2006).
[31]I. Khavrutskii and J. McCammon, J. Chem. Phys. **127**, 124901 (2007).
[32]D. Branduardi, F. Gervasio, and M. Parrinello, J. Chem. Phys. **126**, 054103 (2007).
[33]A. Pan, D. Sezer, and B. Roux, J. Phys. Chem. B **112**, 3432 (2008).
[34]Y. Sugita, A. Kitao, and Y. Okamoto, J. Chem. Phys. **113**, 6042 (2000).
[35]H. Fukunishi, O. Watanabe, and S. Takada, J. Chem. Phys. **116**, 9058 (2002).
[36]J. Kästner and W. Thiel, J. Chem. Phys. **123**, 144104 (2005).
[37]K. Murata, Y. Sugita, and Y. Okomoto, Chem. Phys. Lett. **385**, 1 (2004).
[38]H. Lou and R. Cuckier, J. Phys. Chem. B **110**, 24121 (2006).
[39]M. Wolf, J. Jongejan, J. Laman, and S. de Leeuw, J. Phys. Chem. B **112**, 13493 (2008).
[40]G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
[41]A. West, R. Elber, and D. Shalloway, J. Chem. Phys. **126**, 145104 (2007).
[42]R. Elber, Biophys. J. **92**, L85 (2007).
[43]D. Shalloway, J. Chem. Phys. **105**, 9986 (1996).
[44]G. Hummer and I. Kevrekidis, J. Chem. Phys. **118**, 10762 (2003).
[45]W. Swope, J. Pitera, and F. Suits, J. Phys. Chem. B **108**, 6571 (2004).
[46]W. Swope, J. Pitera, F. Suits, M. Pitman, M. Eleftheriou, B. Fitch, R. Germain, A. Rayshubski, T. Ward, Y. Zhestkov, and R. Zhou, J. Phys. Chem. B **108**, 6582 (2004).
[47]S. Krivov and M. Karplus, J. Chem. Phys. **117**, 10894 (2002).
[48]J. Chodera, N. Singhal, V. Pande, K. Dill, and W. Swope, J. Chem. Phys. **126**, 15501 (2007).
[49]F. Noe, I. Horenko, C. Schütte, and J. Smith, J. Chem. Phys. **126**, 155102 (2007).
[50]N. Buchete and G. Hummer, J. Phys. Chem. B **112**, 6057 (2008).
[51]J. Norris, *Markov Chains*, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge, 2004).
[52]E. Vanden-Eijnden, M. Venturoli, G. Ciccotti, and R. Elber, J. Chem. Phys. **129**, 174102 (2008).
[53]A. Wells, A. Lin, L. Chen, D. Safer, S. Cain, T. Hasson, B. Carragher, R. Milligan, and H. Sweeney, Nature (London) **401**, 505 (1999).
[54]U. Haberthür and A. Caflisch, J. Comput. Chem. **29**, 701 (2008).
[55]B. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).

[56]B. R. Brooks, C. L. Brooks III, A. D. MacKerell, Jr., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, M. Karplus, J. Comput. Chem. **30**, 1545 (2009).
[57]M. Engels, E. Jacoby, P. Krüger, J. Schlitter, and A. Wollmer, Protein Eng. **5**, 669 (1992).
[58]J. Schlitter, M. Engels, P. Krüger, E. Jacoby, and A. Wollmer, Targeted molecular dynamics simulation of conformational change—application to the T → R transition in insulin Mol. Simul. **10**, 291 (1993).
[59]W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, J. Chem. Phys. **123**, 134109 (2005).
[60]C. Eckart, Phys. Rev. **47**, 552 (1935).
[61]R. Czerminski and R. Elber, J. Chem. Phys. **92**, 5580 (1990).
[62]W. Kabsch, Acta Cryst. **A32**, 922 (1976).
[63]L. Maragliano, G. Cottone, G. Ciccotti, and E. Vanden-Eijnden, J. Am. Chem. Soc. **132**, 1010 (2010).
[64]H. Risken, *The Fokker-Planck Equation. Methods of Solution and Applications*, ed. (Springer-Verlag, Berlin, 1989).
[65]I. Andricioaei and M. Karplus, J. Chem. Phys. **115**, 6289 (2001).
[66]R. Elber and M. Karplus, Chem. Phys. Lett. **139**, 375 (1987).
[67]G. Jónsson and K. Jacobsen, "Nudged elastic band method for ending minimum energy paths," in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. Berne, G. Ciccotti, and D. Coker (World Scientific, Singapore, 1998) pp. 385–404.
[68]S. Fisher and M. Karplus, Chem. Phys. Lett. **194**, 252 (1992).
[69]J. Hu, A. Ma, and A. Dinner, J. Chem. Phys. **125**, 114101 (2006).
[70]C. W. Oseen, *Hydrodynamik* (Akademisches, Leipzig, 1927).
[71]J. Rotne and S. Prager, J. Chem. Phys. **50**, 4831 (1969).
[72]J. de la Torre, M. Huertas, and B. Carrasco, Biophys. J. **78**, 719 (2000).
[73]S. Aragon and D. Hahn, Biophys. J. **91**, 1591 (2006).
[74]R. Zwanzig, Adv. Chem. Phys. **15**, 325 (1969).
[75]R. Pastor and M. Karplus, J. Phys. Chem **92**, 2636 (1988).
[76]R. Venable and R. W. Pastor, Biopolymers **27**, 1001 (1988).
[77]E. D. L. Cruz, E. Ostap, and H. Sweeney, J. Biol. Chem. **276**, 32373 (2001).
[78]R. Rock, S. Rice, A. Wells, T. Purcell, J. Spudich, and H. Sweeney, Proc. Natl. Acad. Sci. U.S.A. **98**, 13655 (2001).
[79]H. Park, A. Li, L.-Q. Chen, A. Houdusse, P. Selvin, and L. Sweeney, Proc. Natl. Acad. Sci. U.S.A., **104**, 778 (2007).
[80]R. Gomory and T. Hu, SIAM (Soc. Ind. Appl. Math.) J. Appl. Math. **9**, 551 (1961).