

Realistic Transition Paths for Large Biomolecular Systems: A Langevin Bridge Approach

Patrice Koehl,¹ Marc Delarue,² and Henri Orland^{3,4}

¹*Department of Computer Science and Genome Center, University of California, Davis, CA 95616, USA.*

²*Architecture et Dynamique des Macromolécules Biologiques, UMR 3528 du CNRS, Institut Pasteur, 75015 Paris, France*

³*Department of Physics, School of Sciences, Great Bay University, Dongguan, Guangdong 523000, China*

⁴*Université Paris-Saclay, CNRS, CEA, Institut de Physique Théorique, 91191, Gif-sur-Yvette, France*

(Dated: 2 December 2025)

We introduce a computational framework for generating realistic transition paths between distinct conformations of large biomolecular systems. The method is built on a stochastic integro-differential formulation derived from the Langevin bridge formalism, which constrains molecular trajectories to reach a prescribed final state within a finite time and yields an efficient low-temperature approximation of the exact bridge equation. To obtain physically meaningful protein transitions, we couple this formulation to a new coarse-grained potential combining a Gō-like term that preserves native backbone geometry with a Rouse-type elastic energy term from polymer physics; we refer to the resulting approach as SIDE. We evaluate SIDE on several proteins undergoing large-scale conformational changes and compare its performance with established methods such as MinActionPath and EBDIMS. SIDE generates smooth, low-energy trajectories that maintain molecular geometry and frequently recover experimentally supported intermediate states. Although challenges remain for highly complex motions—largely due to the simplified coarse-grained potential—our results demonstrate that SIDE offers a powerful and computationally efficient strategy for modeling biomolecular conformational transitions.

I. INTRODUCTION

Biomolecules serve as the fundamental workhorses of cells, carrying out nearly all essential biological functions. These functions are governed both by molecular shape—specifically, the three-dimensional (3D) geometries that define biomolecular conformations—and by the inherent change of shape that occurs as they perform their functions. Significant progress has been made in characterizing biomolecular structures, aided by a combination of experimental and computational approaches. Experimental methods such as X-ray crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, and cryo-Electron Microscopy (cryo-EM) have been instrumental in this effort. A particularly notable achievement in this field is the long-sought ability to accurately predict protein 3D structures directly from their amino acid sequences—a challenge that is now considered largely solved. At the forefront of this breakthrough is AlphaFold¹, developed by Google DeepMind, whose creators, Demis Hassabis and John Jumper, received the 2024 Nobel Prize in Chemistry for this accomplishment². Comparable advances have also been achieved by other methods, including RoseTTAFold, developed in David Baker’s laboratory^{3,4}, and ESMFold from Meta AI⁵. However, biomolecular function arises not merely from static structures but from structural dynamics—the continuous motions and conformational transitions that occur over a broad range of temporal and spatial scales. Understanding these dynamics remains one of the central challenges in structural molecular biology. Inspired by the success of deep learning in predicting static protein structures, researchers are now actively exploring other deep learning algorithms

aimed at predicting the conformational changes of proteins (see for example Refs^{6–10}). Currently, a major challenge in the development of such models (and in the development of computational studies of protein dynamics in general) lies in the limited experimental training data characterizing dynamics. Experimentally, only a few techniques can capture time-resolved structural data, and those that do typically probe restricted temporal windows. Computational studies of protein dynamics, especially those aimed at simulating transitions between distinct molecular conformations, have met with limited success. The core difficulty lies in the fact that such conformational transitions represent rare events relative to the molecule’s internal dynamical time scales. These transitions arise from stochastic fluctuations in molecular structure, driven by thermal energy from the surrounding heat bath, and are infrequent whenever the energy barrier separating two conformations is large compared to the ambient thermal energy, $k_B T$ (where T is the temperature and k_B the Boltzmann constant).

There is actually a lot of information on protein dynamics in the PDB itself. Indeed, there are many proteins for which multiple conformations have been characterized. Those conformations arise because of the formation of a complex (for example, the binding of a ligand) or changes in the environment (temperature, pH, salt conditions) that influence the stable conformation observed. What we actually need are techniques that can either capture or predict the changes between those conformations, what is referred to as a transition. Vanden Eijnden and colleagues proposed a general approach to studying those transitions, the Transition Path Theory (TPT)^{11–13}. TPT provides a framework for finding the shortest, or most probable transition path between two conformations of a molecule.

At zero temperature, the TPT is deemed exact. As such, it has served as a touchstone for the development of many path-finding algorithms. Some of those were developed for finding the Minimum Energy Path (MEP) on the energy surface for a molecule, such as morphing techniques^{14,15}, gradient descent methods^{16–19}, the nudged elastic band method^{20–22} and the string method^{23–28}. Other algorithms are concerned with either finding the Minimum Free Energy Path (MFEP) on the free energy surface for the molecule,^{29–32} while others search for paths that minimize a functional, such as Onsager-Machlup functional^{33,34}, as implemented in the Minimum Action Path (MAP) methods^{35–43}.

In some other methods, the Langevin equation is modified to include a bridge between the two end states and enforces the trajectory to join them^{44–49}. In parallel to those methods, effort has been dedicated to the development and analysis of Markov State Models (MSMs)^{50–52}. MSMs aim at coarse-graining the dynamics of the molecular system via mapping it onto a continuous-time Markov jump process, that is, a process whose evolution involves jumps between discretized states representing typical conformations of the original system. A similar concept referred to as “Milestoning” has been proposed by Elber and coll. performs well^{53,54}. Note that this list is not meant to be a comprehensive coverage of all existing techniques for finding transition paths, as this is a very active area of research with new techniques proposed every year.

In preliminary studies by two of the authors^{37,43}, we proposed MinActionPath, a method for computing the MAP between two conformations of a protein on a simplified, two-well free energy surface derived from the ENMs of the two conformations. The energy at any conformation is defined as the minimum of the energy functions derived from the two wells. Using this energy functional, the equations of motion corresponding to the MAP are solved analytically in each well, and continuity conditions at the crossing between the two wells define the transition point. MinActionPath has proved useful for example in characterizing the structural reaction path of a tryptophanyl-tRNA synthetase that involves three conformationally distinct states^{55,56}. There are, however two main drawbacks with MinActionPath that often lead to non-physical paths. First, the ENM potential does not account for steric hindrance, and second, defining the energy surface as the lower envelope of the quadratic wells at the minima leads to problems of smoothness at the crossing between the wells, i.e. to the presence of a cusp in energy at the transition.

This paper addresses the shortcomings of MinActionPath for finding a transition path between two protein conformations in two different directions. First, we propose a different equation of motion that does not assume an energy function that combines quadratic wells at the two conformations. This equation comes from our recent work on Brownian and Langevin bridge equations^{46,48,49}. Starting from an overdamped Langevin equation modified with a propagator to force a transition to the target conformation, we have proposed a non-linear stochastic integro-differential equation (which we will refer to as SIDE), whose approximation at low temperature can be solved efficiently and leads to realistic transitions

both on simple 2D potentials and for biomolecules⁴⁸. Second, in contrast to both MinActionPath and our early work on CLD, we define a new coarse-grained potential to represent the protein conformation that does not mix information from the start and target conformations of the protein of interest. Instead, it considers a Rouse elastic network for each intermediate conformation during the transition, adding collision terms and pseudo-bonded terms to avoid steric clashes and maintain the correct local geometry of the protein.

The paper is organized as follows. In the next section, we describe our equation of motion for finding a transition path between two conformations of a biomolecule and its approximation at low temperature, referred to as SIDE. We then describe two coarse-grained potentials used to quantify the energy along the transition, a mixed elastic potential described in a preliminary study⁴⁸, and a new mixed Gō potential and elastic potential introduced in this study. The next section shows applications to a few biomolecular systems. Finally, we conclude the paper with a discussion on extensions of the method, highlighting some of its current limitations.

II. A BRIDGE EQUATION FOR GENERATING TRAJECTORIES

The path generation strategy we follow is based on a stochastic integro-differential equation, initially referred to as a Langevin bridge (LB) equation⁴⁶ and later amended as the Conditioned Langevin Dynamics (CLD) Equation⁴⁸. As it is core to our strategy to derive transition paths, we briefly describe the derivation. Full details are available in Refs.^{48,49}.

A. The Langevin bridge equation

Consider a system of N particles, each represented by a position vector $\mathbf{r}_i \in \mathbb{R}^3$, $i \in \{1, \dots, N\}$. We are given two conformations for this system, an initial conformation I and a final conformation F . We want to build a trajectory for the system over a given time t_f , such that the system is in state I at $t = 0$ and in state F at $t = t_f$. We will use \mathbf{r}_i^I , \mathbf{r}_i^F , and \mathbf{r}_i to indicate the position of a particle at time $t = 0$, $t = t_f$, and t , respectively. The particles of the system interact through a conservative force derived from a potential U . The system is evolved using overdamped Langevin dynamics

$$\dot{\mathbf{r}}_i = -\frac{1}{\gamma} \nabla_i U + \boldsymbol{\eta}_i(t), \quad (1)$$

where $\mathbf{F}_i = -\nabla_i U$ is the force acting on particle i , $\boldsymbol{\eta}_i$ is the Gaussian random force, and γ is the friction coefficient. The potential U may be the sum of a one-body b and two-body v potential

$$U(\mathbf{r}_1, \dots, \mathbf{r}_N) = \sum_{i=1}^N b(\mathbf{r}_i) + \frac{1}{2} \sum_{1 \leq i \neq j \leq N} v(\mathbf{r}_i - \mathbf{r}_j), \quad (2)$$

and of more complicated interaction terms for the realistic atomic description of molecular systems. The friction coefficient

cient is related to the diffusion constant D and the temperature T through the Einstein relation

$$\gamma = \frac{k_B T}{D} = \frac{1}{D\beta}, \quad (3)$$

where $\beta = 1/k_B T$. The friction is usually taken to be independent of T , so that the diffusion coefficient D is proportional to the temperature T .

The moments of the Gaussian white noise are given by

$$\begin{aligned} \langle \eta_i^k(t) \rangle &= 0, \\ \langle \eta_i^k(t) \eta_j^l(t') \rangle &= 2D \delta_{ij} \delta_{kl} \delta(t - t'), \end{aligned} \quad (4)$$

where the indices k and l denote components of the vector $\eta_i(t)$. As the diffusion constant D is proportional to T , the random force $\eta_i(t)$ is of order \sqrt{T} .

The probability distribution function $P(\{\mathbf{r}_i\}, t | \{\mathbf{r}_i^f\}, 0) = P(\{\mathbf{r}_i\}, t)$ for the system to be at positions $\{\mathbf{r}_i\}$ at time t given that it was at position $\{\mathbf{r}_i^f\}$ at time 0, satisfies the Fokker-Planck (FP) equation

$$\frac{\partial P}{\partial t} = D \sum_i \nabla_i (\nabla_i P + \beta \nabla_i U P). \quad (5)$$

Among all the paths generated by the Langevin equation (1), we are only interested in those that are conditioned to end at a given configuration $\{\mathbf{r}_i^f\}$ at time t_f . To this end, we use the method of Brownian bridges introduced through the Doob transform⁵⁷. We denote by $\mathcal{P}(\{\mathbf{r}_i\}, t)$ the probability that the conditioned system is at $\{\mathbf{r}_i\}$ at time t . We have

$$\mathcal{P}(\{\mathbf{r}_i\}, t) = \frac{P(\{\mathbf{r}_i^f\}, t_f | \{\mathbf{r}_i\}, t) P(\{\mathbf{r}_i\}, t | \{\mathbf{r}_i^f\}, 0)}{P(\{\mathbf{r}_i^f\}, t_f | \{\mathbf{r}_i^f\}, 0)}. \quad (6)$$

The probability $P(\{\mathbf{r}_i\}, t | \{\mathbf{r}_i^f\}, 0)$ satisfies equation (5) whereas the function $Q_1(\{\mathbf{r}_i\}, t) = P(\{\mathbf{r}_i^f\}, t_f | \{\mathbf{r}_i\}, t)$ above satisfies the following reverse or adjoint Fokker-Planck (FP) equation⁵⁸.

$$\frac{\partial Q_1}{\partial t} = -D \sum_i \nabla_i^2 Q_1 + D\beta \sum_i \nabla_i U \cdot \nabla_i Q_1. \quad (7)$$

Using eq.(5) and (7), one can easily see that $\mathcal{P}(\{\mathbf{r}_i\}, t)$ satisfies the modified FP equation

$$\frac{\partial \mathcal{P}}{\partial t} = D \sum_i \nabla_i (\nabla_i \mathcal{P} + \nabla_i (\beta U - 2 \ln Q_1) \mathcal{P}) \quad (8)$$

and that the positions $\{\mathbf{r}_i(t)\}$ of the conditioned system satisfy a modified Langevin equation given by

$$\dot{\mathbf{r}}_i = -\frac{1}{\gamma} \nabla_i U + 2D \nabla_i \ln Q_1 + \eta_i(t). \quad (9)$$

This equation is called a bridge equation⁵⁷. The additional force term $2D \nabla_i \ln Q_1$ conditions the paths and guarantees that they will end at $(\{\mathbf{r}_i^f\}, t_f)$. We can rewrite Q_1 as:

$$Q_1(\{\mathbf{r}_i\}, t) = e^{-\beta(U(\{\mathbf{r}_i^f\}) - U(\{\mathbf{r}_i\}))} \langle \{\mathbf{r}_i^f\} | e^{-(t_f - t)H} | \{\mathbf{r}_i\} \rangle. \quad (10)$$

In equation (10), we have used standard quantum mechanical notation⁵⁹ for the matrix element of the evolution operator e^{-Ht} , where the Hamiltonian H is given by

$$H = -D \sum_i \nabla_i^2 + D\beta^2 W(\{\mathbf{r}_i\}). \quad (11)$$

In equation 11, the potential W is given by

$$W(\{\mathbf{r}_i\}) = \frac{1}{4} \sum_i (\nabla_i U)^2 - \frac{k_B T}{2} \nabla_i^2 U. \quad (12)$$

The driving term $Q_1(\{\mathbf{r}_i\}, t)$ is a sum over all paths joining $(\{\mathbf{r}_i\}, t)$ to $(\{\mathbf{r}_i^f\}, t_f)$, properly weighted by the so-called Onsager-Machlup action⁶⁰, $\exp\left(-\frac{1}{4D} \int_t^{t_f} d\tau \left(\dot{\mathbf{r}} + \frac{1}{\gamma} \nabla U\right)^2\right)$.

Defining

$$Q(\{\mathbf{r}_i\}, t) = \langle \{\mathbf{r}_i^f\} | e^{-(t_f - t)H} | \{\mathbf{r}_i\} \rangle, \quad (13)$$

the bridge equation (9) becomes

$$\dot{\mathbf{r}}_i = 2D \nabla_i \ln Q + \eta_i(t) \quad (14)$$

In ref.⁴⁹, we have shown that this equation can be exactly recast in the following form

$$\dot{\mathbf{r}}_i = \frac{\mathbf{r}_i^f - \mathbf{r}_i(t)}{t_f - t} - \frac{2}{\gamma^2} \int_t^{t_f} d\tau \left(\frac{t_f - \tau}{t_f - t} \right) \langle \nabla_i W(\{\mathbf{r}_i(\tau)\}) \rangle_Q + \eta_i(t), \quad (15)$$

where the bracket $\langle \dots \rangle_Q$ denotes the average over all paths joining $(\{\mathbf{r}_i\}, t)$ to $(\{\mathbf{r}_i^f\}, t_f)$, weighted by the action of equation (13)

$$\langle \nabla_i W(\mathbf{r}(\tau)) \rangle_Q = \frac{\langle \{\mathbf{r}_i^f\} | e^{-(t_f - \tau)H} \nabla W(\{\mathbf{r}_i\}) e^{-(\tau - t)H} | \{\mathbf{r}_i\} \rangle}{\langle \{\mathbf{r}_i^f\} | e^{-(t_f - t)H} | \{\mathbf{r}_i\} \rangle} \quad (16)$$

and the Gaussian noise is defined by equation (4). Note that the first term in the right-hand side of equation (15) guarantees that the constraint $\mathbf{r}_i(t_f) = \mathbf{r}_i^f$ is satisfied. It is the only term that is singular at time t_f , since the integral term does not have any singularity at any time. In fact, in the case of a free Brownian particle, the potential W vanishes, and we recover the standard equation for free Brownian bridges

$$\dot{\mathbf{r}}_i = \frac{\mathbf{r}_i^f - \mathbf{r}_i(t)}{t_f - t} + \eta_i(t) \quad (17)$$

Equation (15) is the fundamental equation of motion whose solution defines a transition path between the initial and final conformations. This equation is a nonlinear stochastic equation. It is Markovian, in the sense that the right-hand side of equation (15) depends only on $\mathbf{r}(t)$. However, the presence of the average over all future paths makes it difficult to use.

B. An efficient approximation for weakly dispersed trajectories

We are interested in problems of free energy barrier crossing, which are of importance in many chemical, biochemical, or biological reactions. In this specific situation of barrier

crossing, according to Kramers theory, the total transition time τ_K (waiting + crossing) scales like the exponential of the barrier height $\exp(\beta\Delta E^*)$. In contrast, the crossing time (Transition Path Time) τ_c scales like the logarithm of the barrier $\log\beta\Delta E^{*61-63}$. We have thus $\tau_c \ll \tau_K$. In this case, the transition trajectories are very weakly diffusive and are thus almost ballistic.

In reference Ref. ⁽⁴⁸⁾, by using a cumulant expansion, we obtained a simple approximation for the bridge equations in the transition path time regime mentioned above. Using the exact bridge equation (15), it is easy to recover this approximate equation. The approximation consists of considering that the paths that connect $\{\mathbf{r}_i(t)\}$ to $\{\mathbf{r}_i^F(t)\}$ in (15) reduce to a straight line connecting the two points. The equation for this straight line is

$$\mathbf{r}_i(\tau) = \frac{(\mathbf{r}_i^F - \mathbf{r}_i)\tau + t_f\mathbf{r}_i - t\mathbf{r}_i^F}{t_f - t}. \quad (18)$$

Consequently, eq.(15) becomes

$$\dot{\mathbf{r}}_i = \frac{\mathbf{r}_i^F - \mathbf{r}_i(t)}{t_f - t} - \frac{2}{\gamma^2} \int_t^{t_f} d\tau \left(\frac{t_f - \tau}{t_f - t} \right) \nabla_i W(\{\mathbf{r}_i(\tau)\}) + \boldsymbol{\eta}_i(t) \quad (19)$$

where $\mathbf{r}_i(\tau)$ is given by eq.(18).

Making the change of variable

$$u = \frac{\tau - t}{t_f - t}, \quad (20)$$

we have

$$\mathbf{r}_i(u) = u\mathbf{r}_i^F + (1 - u)\mathbf{r}_i. \quad (21)$$

Equation (19) becomes

$$\dot{\mathbf{r}}_i = \frac{\mathbf{r}_i^F - \mathbf{r}_i(t)}{t_f - t} - \frac{2}{\gamma^2} (t_f - t) \int_0^1 du (1 - u) \nabla_i W(\{\mathbf{r}_i(u)\}) + \boldsymbol{\eta}_i(t). \quad (22)$$

This equation is an integro-differential stochastic Markov equation, as the variable $\dot{\mathbf{r}}_i(t)$ depends only on the stochastic variable $\mathbf{r}_i(t)$ at time t . One can generate many independent trajectories by integrating this equation with different noise histories $\boldsymbol{\eta}_i(t)$. It is the basis of the transition path method presented in this paper.

C. Solving the bridge equation

We found that a simple method to solve the equation is to use a Euler-Maruyama⁶⁴ discretization scheme for the equation, dividing the time t_f in N intervals of size dt , so that $t_f = Ndt$. In this scheme, the state of the molecule $\{\mathbf{r}(k, i)\}$ at time $(k + 1)dt$ is computed from the state at time kdt :

$$\mathbf{r}(k + 1, i) = \mathbf{r}(k, i)dt + \frac{\mathbf{r}^F(i) - \mathbf{r}(k, i)}{t_f - kdt}dt - \frac{2}{\gamma^2} (t_f - kdt)I(k, i) + \sqrt{2Ddt}\boldsymbol{\xi}(k, i), \quad (23)$$

where $\boldsymbol{\xi}(k, i)$ are normalized Gaussian variables:

$$\langle \boldsymbol{\xi}^{(a)}(k, i) \rangle = 0; \quad \langle (\boldsymbol{\xi}^{(a)}(k, i))^2 \rangle = 1. \quad (24)$$

The integral $I(k, i)$ is computed numerically over M steps within the interval $[0, 1]$,

$$I(k, i) = \frac{1}{M} \sum_{l=0}^{M-1} \left(1 - \frac{l}{M}\right) \nabla_i W(\{\mathbf{r}(l, i)\}),$$

where

$$\mathbf{r}(l, i) = \frac{l}{M}\mathbf{r}^F(i) + \left(1 - \frac{l}{M}\right)\mathbf{r}(k, i).$$

III. COARSE-GRAINED POTENTIALS FOR PROTEIN TRANSITION PATHS

The previous section introduced an integro-differential equation for computing the trajectory between 2 conformations of a biomolecule. This equation is valid for any potential U that describes the stability and geometry of the biomolecule considered. However, this potential is essential for generating realistic trajectories. We describe two such potentials, both coarse-grained, i.e., based on a simplified representation that includes only one atom per residue, the C_{α} for a protein.

A. A mixing potential for transition paths for proteins driven by Langevin bridge: the CLD framework

We are concerned with the study of conformational changes between two states of a protein. We use a coarse-grained representation of the protein structure, namely, we only consider one atom per residue, its C_{α} . In our previous study for deriving protein transition paths, we defined an energy function that is the combination of a mixed elastic model and a collision term

$$U = U_{\text{Mix-ENM}} + U_{\text{collision}}. \quad (25)$$

In this equation the mixed elastic potential is defined as

$$U_{\text{Mix-ENM}} = -\frac{1}{\beta_m} \log(e^{-\beta_m U_I} + e^{-\beta_m U_F}), \quad (26)$$

where β_m is the inverse of the mixing Temperature T_m , U_I and U_F are the elastic potential centered on conformation I (initial) and conformation F (final), respectively¹⁷. The two elastic potentials follow the original definition of Tirion⁶⁵:

$$\begin{aligned} U_I &= \sum_{ij} k_{ij} C_{ij} (r_{ij} - r_{ij}^I)^2 \\ U_F &= \sum_{ij} k_{ij} C_{ij} (r_{ij} - r_{ij}^F)^2 + \Delta U_0 \end{aligned} \quad (27)$$

where C_{ij} is a contact matrix that is set to 1 if $d_{ij} < R_c$ and 0 otherwise and k_{ij} is its associated elastic constant. If a pair (i, j) is present in both forms, we take the same elastic constant k_{ij}

for both (see below). ΔU_0 is the (preset) energy difference between the two states and r_{ij}^I and r_{ij}^F their interatomic distances at rest in conformations A and B, respectively. The elastic constants k_{ij} are modulated by the difference in the resting r_{ij} distances of the two states.

$$k_{ij} = \min \left(\frac{\epsilon_k}{(r_{ij}^I - r_{ij}^F)^2}, k_{max} \right) \quad (28)$$

The collision energy term is taken as the repulsive part of a Lennard-Jones potential,

$$U_{collision} = \epsilon \sum_{i,j} \left(\frac{\sigma}{d_{ij}} \right)^{12} \quad (29)$$

The potential W is given by,

$$W = \frac{1}{4} \nabla U^2 - 0.5 k_B T \Delta U \quad (30)$$

Note that W is not a linear combination of an effective energy associated with the elastic term and an effective energy for the collision term, due to the presence of the norm squared of ∇U : there is a cross-term between the two types of potential. All the algebra needed to implement this energy U and its associated effective energy was described in detail in an Appendix in⁴².

It is important to notice that U , and consequently W depends on many parameters. Those include the mixing temperature T_m , the cutoff R_c for the elastic networks, ϵ_k and k_{max} for the mixing elastic constants, ϵ and σ for the collision term, and ΔU_0 defining the energy difference between states A and B. In addition, this energy function is heavily conditioned by the conformations A and B through their elastic potentials. This leads to some ambiguities for some terms. For example, U_I and U_F may include very different numbers of pairs of atom, leading to possible biases in the mixing energy. We have seen that parameterizing this potential is not easy, leading to cases in which we could not build a transition path based on this energy (the ATPase and RNase cases mentioned in Ref.⁴⁸). This lead us to propose a different type of potential, presented below.

B. A Gō like potential for transition paths for proteins driven by Langevin bridge: the SIDE framework

When designing a new potential for building transition paths between two conformations of a protein, our specifications were double: minimize the number of parameters, and reduce the dependencies to the initial and final conformations. We chose a Gō-like potential as it verifies these two requirements. The Gō-like potential only considers the C_α of all residues in the molecule of interest. In addition to the notations already introduced above, we consider also θ_i , the virtual bond angle formed by the C_α s of the consecutive residues i , $i+1$, and $i+2$. Our Gō-like potential at a conformation \mathbf{X} is defined as (see⁶⁶ as well as⁶⁷ for equivalent potentials):

$$U(\mathbf{X}) = U_b(\mathbf{X}) + U_\theta(\mathbf{X}) + U_{vdW}(\mathbf{X}) + U_{el}(\mathbf{X}) \quad (31)$$

with:

$$U_b(\mathbf{X}) = \frac{100\epsilon_G}{2} \sum_{i=1}^{N-1} (r_{i,i+1} - r_{i,i+1}^I)^2,$$

$$U_\theta(\mathbf{X}) = \frac{40\epsilon_G}{2} \sum_{i=1}^{N-2} (\theta_i - \theta_i^I)^2,$$

$$U_{vdW}(\mathbf{X}) = \epsilon_G \sum_{(i < j-3)} \left[\left(\frac{r_{ij}^I}{r_{ij}} \right)^{12} - \left(\frac{r_{ij}^I}{r_{ij}} \right)^6 \right],$$

$$U_{el}(\mathbf{X}) = \frac{\epsilon_G}{2Np} \sum_i \sum_j g(r_{ij}) r_{ij}^2.$$

I stands for the initial conformation, N is the total number of C_α in the protein, Np is the number of pairs considered, ϵ_G is a constant, and $g(r_{ij})$ is a Fermi-like function that provides a smooth cutoff:

$$g(r) = \frac{1}{1 + \exp \left(\frac{r-d_0}{a_0} \right)} \quad (32)$$

with d_0 and a_0 constants (see Figure 1).

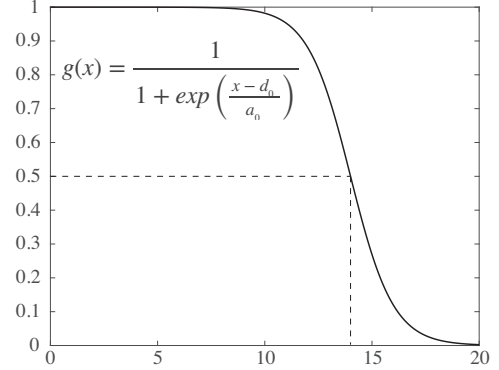


FIG. 1. The Fermi like function $g(x)$ (shown for $d_0 = 14$ and $a_0 = 1$).

The potential defined in equation 31 contains three types of terms:

- a) *Bonded interactions.* Both U_b (a bond potential associated with consecutive C_α s) and U_θ (an angle potential associated with 3 consecutive C_α s) are local potentials meant to maintain the geometry of the protein chain. Two consecutive C_α s are at 3.8 Å or 2.9 Å apart, depending on the peptide bond being trans or cis, respectively. We capture this information from the initial conformation. The angles associated with 3 consecutive C_α s is loosely conserved in protein, but we keep it close to their values in the initial conformation.

- b) A *vdW term*. U_{vdW} is a 12-6 Lennard-Jones potential set to avoid steric clashes during the transition.
- c) An *elastic potential*. The elastic model is different from a standard Tirion-like elastic model in that it does not include a reference structure. It is replaced by the Rouse Model, a coarse-grained model well studied in polymer physics with springs between beads^{68,69}.

The vdW term is limited to pairs of atoms that are within a cutoff distance R_c from each other. The elastic potential handles the cutoff differently, introducing the Fermi-like function $g(r)$ defined in Equation 32. The reason is simple: the vdW potential naturally approaches 0 at large distances because of the 12 and 6 powers of the Lennard Jones potential. In contrast, the elastic potential is proportional to r^2 , i.e., it becomes large for large distances. The presence of the Fermi function imposes a potential of 0 close to $d_0 + a_0$. There is another reason, however, for the Fermi potential. In its absence, the elastic potential would be proportional to the radius of gyration, squared. As such, optimizing the elastic potential would lead to proteins that are as compact as possible, which is not the desired outcome in most cases.

As written, the Gō like potential we consider has 4 parameters: the cutoff R_c and its equivalent d_0 for the Fermi-like function, the width of the latter, a_0 , as well as the factor ϵ common to all terms. The accepted range of values for R_c in the context of a C_α -only elastic model is 12 to 14. In all experiments described below, we use $R_c = d_0 = 14$, $a = 1$, and $\epsilon = 1$.

All the algebra needed to implement the Gō-like energy U and its associated effective energy is described in detail in Appendices A (pairwise interactions), B (angular term), and C (effective energy).

C. Principal component analyses of an ensemble of protein structures

A transition path between two structures of a protein is a sampling of the ensemble of conformations that are accessible to the protein along that path. It is possible to extract information from that ensemble using statistical techniques. In particular, principal component analysis (PCA) is well suited for that task as it enables describing the ensemble through a decomposition process that filters observed motions from the most significant to the least significant scales (see for example Ref.⁷⁰).

PCA has been widely used to describe the essential motions of proteins from MD simulations^{71,72}. It starts from a coordinate matrix, X of size $3N_a \times N_s$ where N_a is the number of atoms considered, and N_s the number of conformations in the ensemble (in our case, the number of images along the trajectory). The first step is to remove overall translations and rotations by aligning each conformation to a reference structure (usually the first structure). Then a new matrix X' is built through centering each row of X . A covariance matrix $C = XX'$ is constructed and diagonalized:

$$C = X'X'^T = EVE^T$$

The column \mathbf{E}_i of E correspond to the i -th principal components (PC), with contribution $\frac{v_i}{\sum_k v_k}$ (the variance of the principal component). Usually the top M (i.e. with the largest variance) two PCs are chosen to capture a low-dimensional representation of the structure space associated with the N_s conformation in the ensemble.

Note that the matrix C has size $3N_a \times 3N_a$. For a large molecular system, it becomes too large to fit on standard computer memory. However, there is no need to compute C explicitly. Indeed, we are only interested in a few of the largest eigenvalues of C that can be computed using an iterative method such as the power method with explicit deflation (see Ref.⁷³). The idea is the following: after finding the largest eigen pair (v_1, \mathbf{E}_1) of C , deflate the matrix:

$$C_1 = C - v_1 \mathbf{E}_1 \mathbf{E}_1^T.$$

Then apply the power method to C_1 to find v_2 , and repeat until the M top eigen pairs of C have been found. The power method requires computing the products of C with vectors V (that ultimately converge to an eigenvector of C). This is done by computing $X'X'^T V$, removing the need to compute C explicitly. Finally, the variance σ_i by an eigenvalue v_i is:

$$\sigma_i = \frac{v_i}{\sum_k v_k} = \frac{v_i}{\text{tr}(C)} = \frac{v_i}{\|X'\|_F^2}$$

where $\text{tr}(C)$ is the trace of C , and $\|X'\|_F$ is the Frobenius norm of X' .

The PCA procedure described above can be used to compare trajectories as follows. First, we collect all conformations along all the trajectories to analyze, defining an ensemble. A coordinate matrix X is then built from this ensemble, and analyzed using the PCA procedure described. The two main principal components, PC1 and PC2, with their eigenvectors \mathbf{E}_1 and \mathbf{E}_2 define a space for the ensemble with reduced dimensionality. The coordinates $c(\mathbf{X})_i$ with $i \in [1, 2]$, of any conformation \mathbf{X} in that space are then computed as

$$c(\mathbf{X})_i = \langle (\mathbf{X} - \mathbf{X}_0), \mathbf{E}_i \rangle$$

IV. RESULTS AND DISCUSSION: STRUCTURAL TRANSITIONS FOR LARGE MOLECULAR SYSTEMS

We test three different strategies for generating a trajectory between two conformations of a protein.

The first strategy relies on solving the Langevin bridge equation discussed in the method section. We apply it with two different potentials, a mixed elastic potential (the CLD method) and the new potential introduced in this study (the SIDE method).

The second strategy, which we will refer as MAP, assumes a harmonic potential at the end states and solves the action minimization problem by integrating the equations of motion at each end state and finding the crossing point of these two solutions, referred to as a transition state. This method, initially referred to as MinactionPath³⁷, is available through a web server⁴³. We used instead a standalone version that is equivalent to the web version.

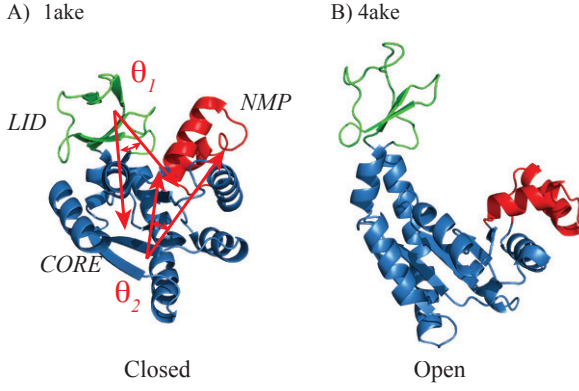


FIG. 2. **Conformational transitions in Adenylate kinase (AKE).** (A) the open state (PDB code: 4AKE); and (B) the closed state (PDB code: 1AKE). AKE consists of three well-defined domains, the rigid CORE (blue, residues 1–29, 60–121, and 160–214) the nucleotide triphosphate binding domain, LID (green, residues 122–159), and the nucleotide monophosphate binding domain NMP (red, residues 30–59). The angle LID-CORE θ_1 is formed by the centers of mass of the backbone of residues of LID (residues 123–155 (LID), hinge (residues 161–165), and CORE (residues 1–8, 79–85, 104–110, and 190–198), whereas the angle NMP-CORE θ_2 is formed by the centers of mass of the backbone of residues of NMP (residues 50–59), CORE (residues 1–8, 79–85, 104–110, and 190–198), and hinge (residues 161–165).

The third strategy, EBDIMS⁷⁴, starts with a stochastic simulation following a Langevin equation,

$$m_i \ddot{\mathbf{r}}_i = -\nabla_i U - \gamma \dot{\mathbf{r}}_i + \boldsymbol{\eta}(t),$$

where U is an elastic potential,

$$U = \frac{1}{2} \sum_i \sum_j (r_{ij} - r_{ij}^0)^2,$$

namely, a Tirion potential⁶⁵ (see equation 27 for details), and $\boldsymbol{\eta}$ is a Gaussian random force. The Langevin dynamics starts at one of the two conformations of the protein under study. It is biased in the direction of the other conformation by computing every k steps a progress variable that compares the interatomic distances in the current conformation with the same distances in the target conformation, and accepting a move only if it reduces this progress variable⁷⁴. EBDIMS is available through a web server⁷⁵, and as a standalone program. We used the latter.

The four different programs, CLD, SIDE, PATH, and EBDIMS were run with the following parameters.

- For CLD, we used the parameters described in Ref.⁴⁸. For its mixed-ENM potential, we set $R_c = 11.5$, and a mixing Temperature $T_m = 1500T$, where T is taken to be $T = 5$. For the Langevin dynamics equation, we set $dt = 0.001$ and $\gamma = 1$.
- For the SIDE potential, we set $\varepsilon = 1$ and $R_c = 14$. The Fermi function is set with $a_0 = 1$ and $d_0 = R_c = 14$. For

the Langevin dynamics, we set $dt = 0.001$, $\gamma = 1$, and $T = 1$.

- MinActionPath has two options for defining elastic network, one based on a cutoff distance and the second that derives the elastic network from the Delaunay complex over the C_α of the protein. We used the former, setting $R_c = 14$ to match with the network defined with the three other methods. Similarly, MinActionPath includes a Tirion potential and a Gō potential; we use the former. This specific version of MinActionPath is referred to as MAP in the following. The trajectories are computed with $t_f = 50$ and $T = 1$.
- EBDIMS has two main accessible parameters, a cutoff value, and the number of unbiased step of Langevin dynamics, k . We have set the cutoff to 6 and k to 1, as recommended by the EBDIMS web server.

A. An illustrative example: Adenylate kinase (AKE)

Adenylate kinase, or AKE in short, is an ubiquitous enzyme that catalyzes the reversible phosphoryl transfer of AMP and ATP into two ADP. AKE comprises of three main domains, the ATP-binding domain also referred to as LID, the AMP binding domain (NMP) and the remainder, referred to as the CORE domain (see Figure 2). AKE undergoes a large-scale conformational change between open and closed states through hinge-like motions. Interestingly, it will also undergo a conformational change in the absence of substrates⁷⁶ which makes the study of its transition particularly amenable to numerical simulation because it removes the complexity of accurately simulating the protein–ligand interactions. As such, AKE’s transition between its open and closed states has been widely studied both by computational simulations (see for example Ref.⁷⁷ and references therein), making it a perfect illustrative example for any method generating transition paths.

Figure 3 illustrates a trajectory generated by SIDE between the closed state (PDB entry 1Ake) and the open state (PDB entry 4ake) of AKE. As expected, the main motion is a hinge motion that leads the LID to separate from the CORE and the NMP. During the transition, the elastic network associated with the protein decreases in size. Initially, the protein is very compact and edges are observed within, and in between the three domains of the protein. As we get closer to the open conformation, the LID maintains only limited interactions with the two other domains; this is captured by the elastic network.

We compare in Figure 4 the four trajectories generated by SIDE, CLD, MAP, and EBDIMS, using both geometry (panel A), and a projection onto the principal components of the space of conformations for AKE (panel B). Figure 4A shows a significant transition in the angle LID-CORE, θ_1 . It changes from 95° to 60° , corresponding to the open and the closed state of LID domain, respectively. In contrast, the angle of NMP-CORE, θ_2 , varies from 65° to 35° , corresponding to the open and semi-open state of the NMP domain, respec-

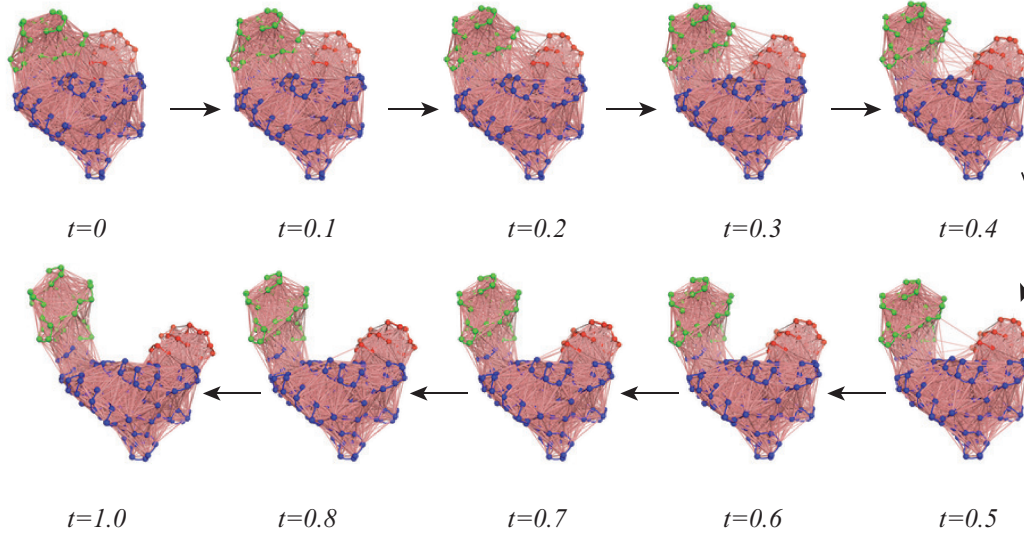


FIG. 3. **SIDE trajectory between the open state and closed state of AKE.** The protein is represented as a string of beads, corresponding to the C_{α} atoms along its main chain, and colored based on its domain definition (see Figure 2 for details). The elastic network at each pose is illustrated with edges colored in salmon pink. The timing for each pose is given as a fraction of the total time given for the construction of the path.

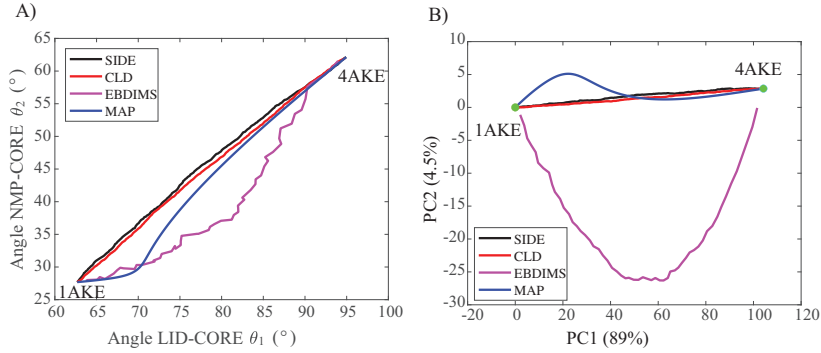


FIG. 4. **Four different trajectories between the open state and closed state of AKE.** We computed trajectories between the open state (1ake) and the closed state (4ake) using SIDE (black), CLD (red), EBDIMS (magenta) and MAP (blue). (A) The angle between the AMP binding domain (NMP) and the CORE is plotted against the angle between the LID and the CORE. (B) Projections of the 4 trajectories along the 2 principal components of their conformational spaces.

TABLE I. Predicting intermediates by constructing transition paths

Name	PDB ID			cRMS (Å)			SIDE		CLD		EBDIMS		MAP	
	Start (A)	Intermediate (I)	Final (C)	(AI)	(IC)	(AC)	R_{best}	IS ^{b)}	R_{best}	IS	R_{best}	IS	R_{best}	IS
5'-NT	1oidA	1oi8B	1hpuD	5.4	4.7	9.3	2.01	58.0	1.61	66.0	2.47	48.0	2.14	55.0
RBP	1ba2A	1urpD	2driA	2.2	4.20	6.2	0.82	63	0.77	65	1.76	20.0	0.70	68
RNase III	1yyoAB	1yz9AB	1yywAB	7.3	13.2	17.5	5.0	30.6	NA ^{c)}	NA	5.4	25.5	6.8	6.2
CA ²⁺ -ATPase	1su4A	1vfpA	liwoA	13.7	10.1	114.0	9.3	7.4	NA	NA	8.9	11.9	9.4	7.6
Myosin	1qviA	1kk7A	1kk8A	16.7	12.0	27.3	4.2	65.0	NA	NA	5.9	50.6	4.2	65.0

^{a)} Minimal cRMS (over CA atoms, in Angstroms) between the trajectory and the intermediate state, computed using Equation 33.

^{b)} Improvement score, in percent, computed using Equation 34. The higher the number, the better. The best score for each protein is highlighted in bold.

^{c)} Trajectory not found (see text)

tively. Interestingly, the two angular transitions appear to occur simultaneously in all 4 trajectories.

Figure 4B is generated by projecting the four trajectories on the first two principal components (PC) of the structural en-

semble generated by combining all snapshots along those trajectories. The two trajectories SIDE and CLD are very close to each other. Those two trajectories are based on the same equation of motions, but with two very distinct potentials. Interestingly, those two trajectories are “ballistic” in the sense that they are relatively straight in structure space.

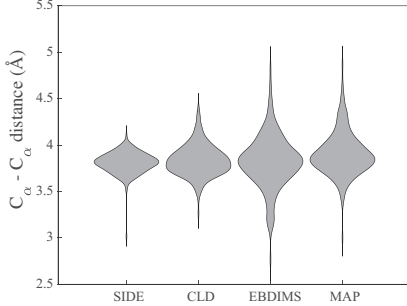


FIG. 5. The distributions of the distances between neighboring C_α s for one snapshot along the trajectory between the open state and closed state of AKE are shown as violin plot. For each method, we picked the snapshot whose corresponding distribution has the largest variance. Note the presence of a peak in those distribution at the C_α - C_α distance of 2.9 corresponding to the cis Proline 87.

The four methods we have tested rely on a coarse-grained representation of the protein structure that only considers the position of the C_α atoms of the molecule. As such, it is legitimate to ask how well the geometry of the molecule is maintained. For each method considered here, we computed the distributions of the distances between neighboring C_α s for all snapshots of the trajectory it generated. In figure 5, we plot the distributions at the snapshot whose corresponding variance is the largest, for all four methods considered. Of the four trajectories, SIDE exhibits the smallest variance in those distributions. This is by far not unexpected, as the potential used by SIDE explicitly constrains the distances between neighboring C_α s. The three other trajectories keep those distances close to the expected value of 3.8 Å, with the largest variance for the EBDIMS trajectory.

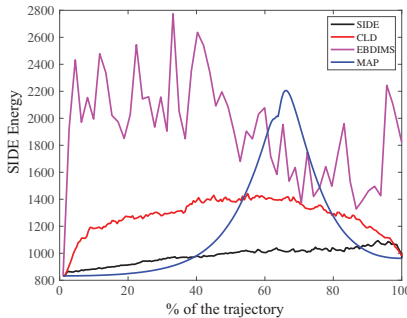


FIG. 6. The Gō-like energy along the four trajectories between the open state and closed state of AKE, SIDE (black), CLD (red), EBDIMS (magenta) and MAP (blue).

In figure 6, we plot the energy used by the SIDE strategy for all four trajectories between the open and closed states

of AKE. This energy has three main components: a Gō like potential that maintains the geometry of the backbone of the protein by restraining the distances and angles between C_α atoms, a vdW term to prevent collision, and an elastic potential to drive the transition. SIDE exhibits the lowest energy along the whole trajectory: this is not surprising, as it is explicitly the energy that drives its equation of motion. Interestingly, however, this energy is relatively constant, hinting that the method was able to find a path that remains in low energy regions of the conformation space. In contrast, the energy of the MAP trajectory shows a significant increase at a transition point. This is inherent to the method behind MAP that computes two quadratic trajectories from the two end points of the path and finds a transition point between those trajectories. The energies along the EBDIMS trajectory are the largest, likely a consequence of the fact that the protein chain along this trajectory exhibits the largest deviation for standard geometry (see figure 5).

B. Detecting intermediate structures along transition paths

Many methods have been developed to generate a path in conformational space between two structural states of a molecular system. As highlighted by Weiss and Levitt¹⁵, however, there is no fully satisfying objective methods to test the biological relevance of such paths. Here we follow the approximate method that they had proposed. Namely, we start from a set of proteins for which there are at least three distinct conformations whose (experimental) structures are known and available in the Protein Data Bank⁷⁸. Two of those conformations, A and C serve as end points for the transition, while the third is set to be the intermediate state, I ; the distinction between the three comes from biology. Trajectories between the two end points are generated, with no knowledge of the intermediate state. We then follow how close the trajectory comes to the intermediate structure,

$$R_{best}(I) = \min_{k \in [1, N]} [cRMS(C_k, I)], \quad (33)$$

where the min is taken over all conformations C_k along the trajectory, and by computing an improvement score IS ^{15,18}:

$$IS = 100 \times \left(1 - \frac{R_{best}(I)}{\min(cRMS(A, I), cRMS(C, I))} \right), \quad (34)$$

As defined, IS is a measure of how close the trajectory comes to the structure of the intermediate, computed as a fraction of how close the start and end points are to this intermediate. In both equations 33 and 34, $cRMS$ stands for the coordinate root mean square deviation computed over all C_α s of the protein.

Results for five proteins included in the original paper from Weiss and Levitt and the four methods for generating transition paths considered in this study are shown in Figures 7, 8, and Table I.

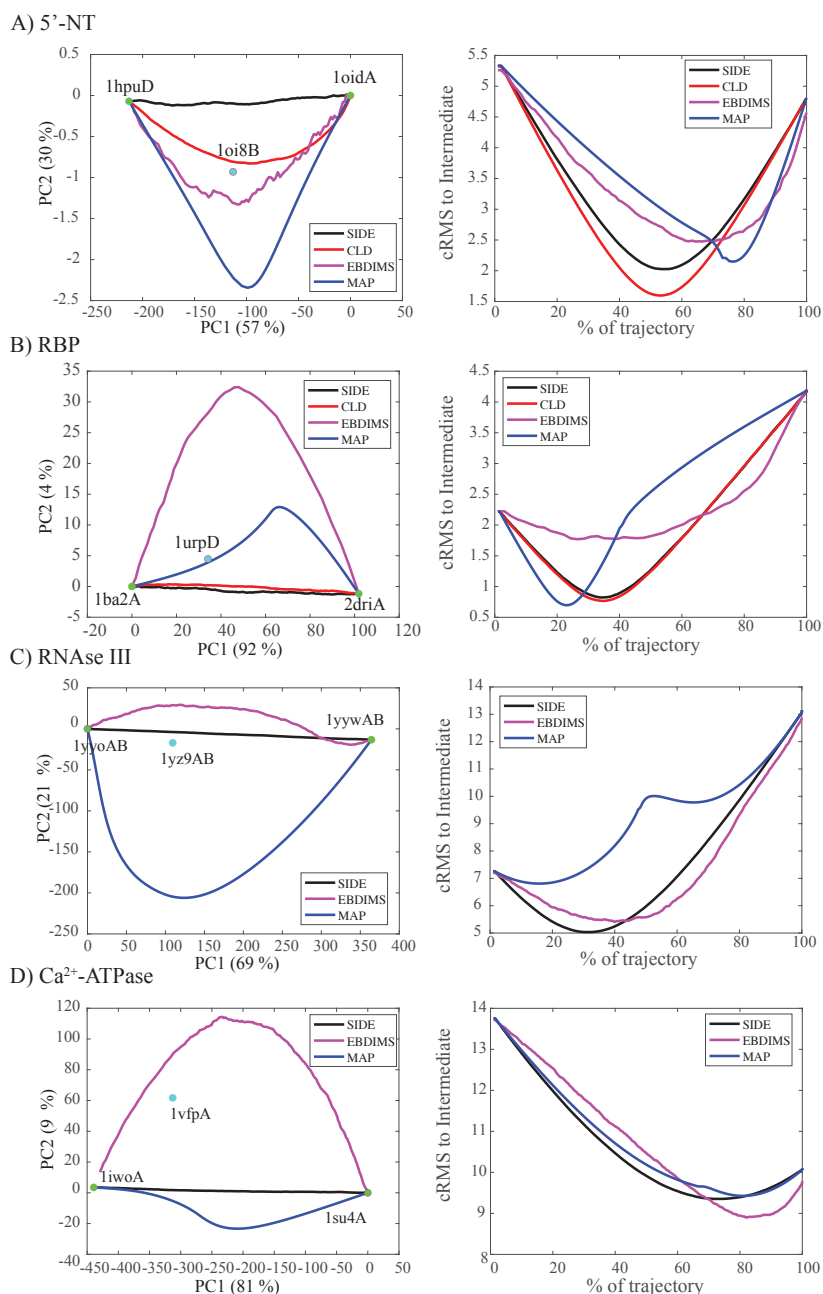


FIG. 7. We computed trajectories between two states A and C for four proteins (5'-NT, RBP, RNase, and Ca²⁺-ATPase) using SIDE (black), CLD (red), EBDIMS (magenta) and MAP (blue). Right panels, projections of the 4 trajectories along the 2 principal components of their conformational spaces, and, left panels, distance of a putative intermediate I to successive snapshots of the trajectory. CLD trajectories for RNase III and Ca²⁺-ATPase and missing as we could not find parameters to get them to converge.

1. Four "simple" transitions

The RBP protein corresponds to the easiest of the four cases considered here: the motion between the start and final states is a simple hinging motion. SIDE, CLD, and MAP perform very similarly in terms of IS score. All three methods get within one Angstrom of the target intermediate structure (1urpD). Interestingly, in the PC space, the intermediate structure is closer to the MAP trajectory, while the IS score

would indicate that the CLD trajectory gets closer. EBDIMS does not seem to identify a trajectory that gets close to the intermediate. This could be just a question of tuning its parameters.

The 5'-NT protein undergoes a large domain rotation. All four methods are able to capture that motion, with the resulting trajectories getting close to the intermediate conformation. Based on IS, CLD performs better (as indicated by the Improvement Score and minimal cRMS), with its trajectory get-

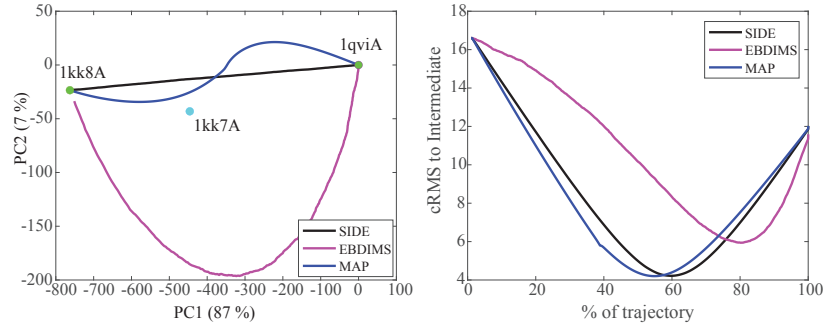


FIG. 8. We computed trajectories between two states of scallop myosin II (1qviA and 1kk8A) using SIDE (black), EBDIMS (magenta) and MAP (blue). Right panel, projections of the 4 trajectories along the 2 principal components of their conformational spaces, and, left panels, distance of a putative intermediate 1kk7A to successive snapshots of the trajectory.

ting closer than 1.65 Å from the intermediate state, for an improvement score of 66%. In contrast to the RBP test case, this observation is confirmed in the PC space.

The RNase III protein undergoes much larger conformational changes than both RBP and 5'-NT as the protein switches from the catalytic form to the non-catalytic form. In this switch, the orientations of the two domains of each of the two chains of RNase are changed drastically. CLD fails to generate such a trajectory; this was already noted in our previous study⁴⁸. Both SIDE and EBDIMS perform well on this protein, generating smooth transitions that get within 5 Å of the intermediate conformation, with improvement scores of 30 % and 26%, respectively. In comparison, MAP fails to get near the intermediate conformation (improvement score of 6.2 %). Note that in PC space, the SIDE and EBDIMS trajectories are very close to each other.

The Ca^{2+} ATPase is the most complicated of the four cases illustrated in Figure 7. The transition between its start and final conformations (apo and holo conformations, respectively) involve a significant structural rearrangement (14 Å). CLD again failed to generate a trajectory. None of the other three methods capture correctly this transition, as none get significantly closer to the intermediate conformation (they do get closer, but stay below 9 Å of the intermediate). This failure should be considered as relative, as the transition involve a complicated series of conformational changes⁷⁹. We note that Mixed-ENM seems to perform best, with an improvement score of 8.5 %, but its trajectory remains more than 9 Å away from the intermediate conformation.

2. A more difficult test case: Myosin

Weiss and Levitt considered a fifth test case, scallop myosin, that is often not included in subsequent studies, as it is the most challenging of the test cases they considered, with the largest structural distance between the two conformations considered (1qviA and 1kk8A), and between those two end points and the structure of the putative intermediate, 1kk7A (see table I). We analyze it in more detail here.

Myosin is a motor protein, hydrolyzing ATP to drive muscle contraction. Weiss and Levitt built a trajectory between pre-

stroke and post-stroke structures, using the near-rigor structure as the intermediate; this corresponds to the order in which they are found in the power cycle⁸⁰.

We built three trajectories between the pre-stroke (1qviA) and post-stroke (1kk8A) of scallop myosin, using SIDE, EBDIMS, and MAP. Just like for RNase III and Ca^{2+} ATPase, we could not find parameters that would allow us to generate a trajectory with CLD. In Figure 8, we illustrate both the projections of the three trajectories in the PC space of the structural ensemble obtained by combining them, as well as the distances of the intermediate 1kk7A to successive snapshots of the trajectories. SIDE and MAP perform well on this protein, generating smooth transitions that get within 4.2 Å of the intermediate conformation, with improvement scores of 65 % (see table I). The trajectory generated by EBDIMS is significantly different (left panel of Figure 8); it also get closer to the intermediate conformation, albeit to a lesser extent (IS of 50%).

The results illustrated in Figure 8, however, could be misleading as each data point on the figures summarizes a whole protein structure. Both 1qviA and 1kk8A have a few missing residues within loops that may impact the quality of the trajectories designed by the three methods considered. In Figure 9, we show the distributions of the distances between neighboring C_α s for all snapshots of the trajectories (panels A, B, and C), as well as specific distributions at the snapshots corresponding to the distributions with the largest variances (panel D). The SIDE trajectory is very well constrained, maintaining CA-CA distances within 0.1 Å of the ideal value of 3.8 Å. The EBDIMS trajectory shows more variations, with some CA-CA distances close to 5 Å. The most striking behavior, however, is observed for the MAP trajectory, exhibiting large variations near its transition point. Some CA-CA distances are found close to 1 Å, while others are close to 7 Å. As such, even though the trajectory seems reasonable from Figure 8, it should be difficult to use it to generate full atom models for its snapshots.

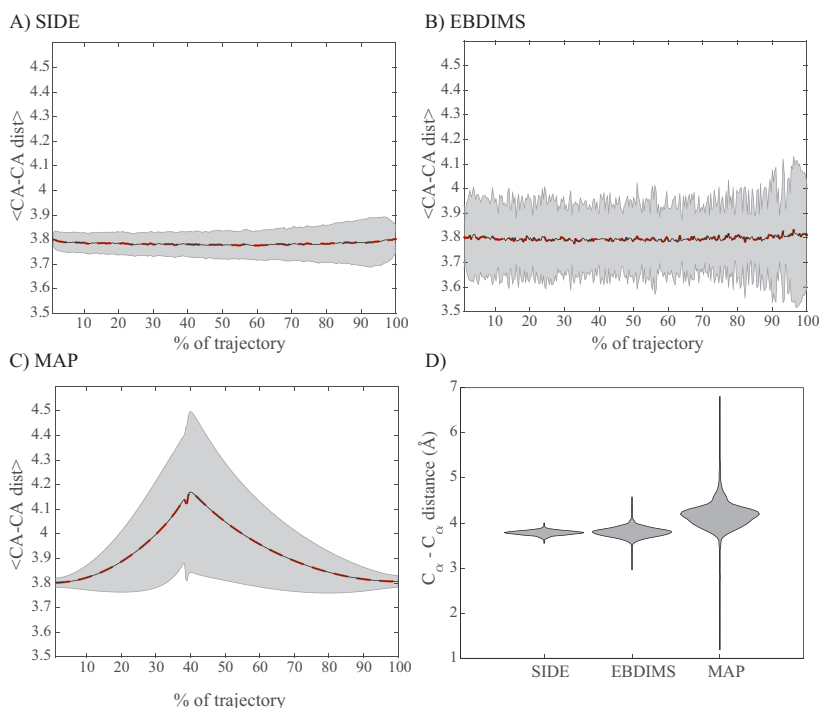


FIG. 9. The mean (dotted line) and standard deviation (grey shaded region) of the distributions of the distances between neighboring C α s for all snapshots along the SIDE (panel A), EBDIMS (panel B), and MAP (panel C) between the pre-stroke and post-stroke structures of scallop myosin. In panel (D) the full distribution is shown as a violin plot for one snapshot along each trajectory. For each method, we picked the snapshot whose corresponding distribution has the largest variance.

3. Overall assessment

The trajectories generated by SIDE, CLD, EBDIMS, and MAP all have merits in their abilities to identify motions responsible to changes of conformation as well as in their capacity to retrieve intermediate structures. There are, however, differences that are worth summarizing. CLD and MAP work well for systems with small cRMS between the two conformations at the end points of the trajectories. When the cRMS becomes large (for example, for RNase III or myosin), CLD fails (more exactly, we could not find parameters to make it work), while MAP “explodes”, i.e., generates trajectories with severe geometric problems. This was already observed before⁴². EBDIMS is more consistent, generating trajectories in all 5 test cases we considered. It does show some distortions along the main chain of the protein. In contrast, SIDE performs well on all 5 cases, leading to trajectories whose snapshots have correct stereochemistry. The difference between EBDIMS and SIDE is likely due to the difference in their potential, as the latter defines the protein main chain geometry explicitly.

C. Limitation of path sampling methods: the VATP case

The examples presented above are standard test cases that have been used in many studies. While useful in highlighting strengths and weaknesses of path sampling methods, they fail

to report on a critical issue for coarse-grained approaches, that is illustrated in the following.

ATP synthase enzymes are “splendid molecular machines”^{81,82} that catalyze the synthesis of ATP. These enzymes exist in two main families: F-type ATP synthases, found in mitochondria, chloroplasts, and bacterial membranes, and V-type ATPases, predominantly located in vacuolar membranes and specialized cellular compartments. Both types share a common architectural principle featuring two coupled rotary motors. The catalytic mechanism fundamentally depends on a rotary motion where proton (or ion) flow through a membrane-embedded rotor domain drives the physical rotation of a central stalk. This mechanical rotation is then transmitted to the catalytic domain, inducing conformational changes in the active sites that enable ATP synthesis in F-type enzymes or ATP hydrolysis to drive proton pumping in V-type enzymes. Understanding this rotary motion and its coupling to the catalysis/hydrolysis reactions is therefore essential to the comprehension of how this enzyme works⁸³.

Here we consider the V-type ATPase (VATP) from *Thermus Thermophilus*. Cryo-EM data revealed three main conformations for the enzyme, corresponding to three states along the full rotation of the rotor at 120° apart⁸⁴. Figure 10 illustrates the architecture of one of those states. The three conformations for VATP are available in the PDB with ids 6qum (state 1), 6r0w (state 2), and 6r0y (state 3). We built three trajectories, S12 (state 1 → state 2), S23 (state 2 → state 3), and S31 (state 3 → state 1), using SIDE, that would ultimately recon-

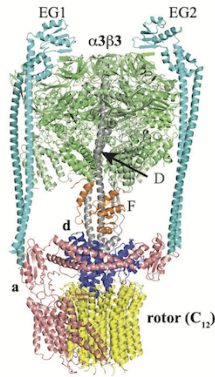


FIG. 10. **The V-ATPase of *Thermus Thermophilus* in the low energy rotational state 1**, PDB code 6qum, color-coded as subunits α and β in green, subunit D is grey, F in orange, the two peripheral stalks EG1 and EG2 is teal and cyan, respectively, domain d in blue, a in salmon, and the rotor formed by 12 helices in yellow.

stitute the whole rotation of the rotor. Results illustrating this complete rotation at the level of the rotor are shown in Figure 11.

Each of the 3 trajectories, S12, S13, S31, captures the corresponding rotation of the rotor. The geometry of the individual helices is well preserved. However, we observe a breathing motion, leading to a reduction in size of the whole rotor, along each of the trajectories. This breathing motion is not expected to be natural. For example, it is unlikely due to the presence of side chains in the lumen of the rotor, as illustrated in Figure 12. The problem is that the potential used by SIDE knows nothing about those side chains. It is based on a coarse-grained representation of the protein in which only the C_α atoms are considered. It includes terms to maintain the geometry of the backbone. Its collision term can't avoid the breathing motion we observe, as atoms within the side chains are not there. This is a general problem. We ran similar simulations with CLD (failed) and with MAP and EBDIMS. The corresponding trajectories exhibit similar non-natural motions, even distortions of the helices of the rotor.

V. CONCLUSIONS AND PERSPECTIVE

In this paper, we addressed the problem of generating paths for a bio-molecular system that start at a given initial configuration and that are conditioned to end at a given final configuration. Our approach follows the ideas of Langevin overdamped dynamics, as expressed with the bridge equation^{46,48,85}. We first revisited this concept of bridge in the context of low temperature, leading to conditioned Langevin dynamics, already described in earlier work^{48,49}. We introduced a new coarse-grained potential that describes the stability of the system and illustrated how the combination of conditioned Langevin dynamics with this new potential in a framework we call SIDE can generate realistic transition pathways for proteins. We tested SIDE against previous iterations of our efforts to generate transition paths as well as against

publicly available tools to compute such tasks. On a test set of 5 proteins originally introduced by Weiss and Levitt¹⁵, we showed that SIDE performs as well as all the other methods tested, with improved conservation of the protein backbone geometry along the trajectory. Finally, we highlight a limitation that is not inherent to our method, but associated with the use of coarse-grained representations of proteins.

Methods that generate trajectories between two conformations of the same protein are defined with two main components: their equation of motion, and the potential they use to represent the energetics of the protein. We tried to disentangle the two by comparing SIDE, our new framework, with CLD, our previous⁴⁸ path sampling technique, as those two methods rely on exactly the same equation of motion, a conditioned Langevin bridge, with the same approximation of low temperature. CLD relies on a mixed-ENM potential that combines information from the two conformations (start and end), while SIDE relies on a pseudo-elastic potential with no (or zero) reference state. CLD adds a repulsive collision term, while SIDE adds both a vdW collision term and a geometric potential to maintain correct main chain geometry along the trajectory. We found that applications of CLD are limited to systems whose transitions involve relatively small changes or simple motion, such as a hinge motion, as it is difficult to parameterize the mixed-ENM potential (this was already observed before⁴⁸). SIDE is much easier to parameterize, at least when it comes to defining the potential, and was successful in (nearly) all the test cases considered (where "nearly" is explained below).

We have provided evidence that SIDE is able to predict intermediate structures in the transition between two conformations of a protein, based on the knowledge of those two conformations only, in addition to generating geometrically realistic trajectories. While promising, those successes should still be considered with caution. The most successful applications described here correspond to cases in which the motions involved in the transition are relatively simple, such as hinge motions or domain rotations. For more complicated motions, such as those involved in the test case Ca^{2+} ATPase or VATP, SIDE (and all other methods tested), perform poorly. Of greater concern, SIDE can generate trajectories with unrealistic motions, such as the breathing found for the rotor of VATP. The latter is clearly associated with limitations of a coarse-grained potential that only considers one atom per residue. While such potentials are useful to reduce computational costs, enabling simulations of large systems, the limitation we illustrate is of concern. One solution to address this problem is to add constraints to the potential. For example, in the case of VATP, we could have introduced constraints to keep the helices of the rotors distributed around a barrel of (nearly) fixed size. This type of ad-hoc solution, however, is difficult to implement as it is not always clear what those constraints should be. Another approach is to refine the description of the protein. We are currently investigating the use of more realistic coarse-grained representations of protein than the C_α only representation, such as the MARTINI force field^{86,87}, more specifically the promising GōMartini3 model⁸⁸ that combines Gō models⁸⁹ with the Martini3 force

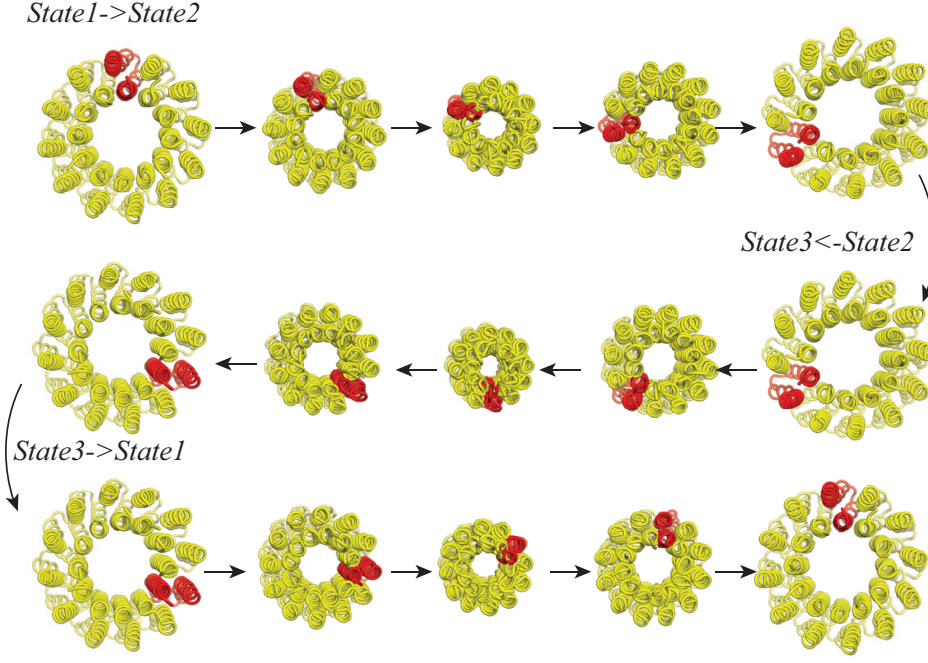


FIG. 11. **Rotation of the rotor of The V-ATPase of *Thermus Thermophilus*.** We plot bottom views (i.e. from the cytoplasm) of the rotor of VATP along the full trajectory reconstituted from the three trajectories state 1 (PDB 6qum) \rightarrow state 2 (PDB 6r0w), state 2 \rightarrow state 3 (PDB 6r0y), and state 3 \rightarrow state 1. One of the 12 helices of the rotor is colored in red to illustrate the rotational motion. All images are at the same scale: note the unusual breathing motion of the whole rotor between each end state of the trajectories.

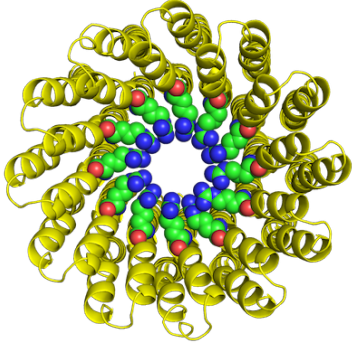


FIG. 12. **The rotor of V-ATPase of *Thermus Thermophilus* in the low energy rotational state 1,** (PDB code 6qum). The 12 helices forming the barrel of the rotor are colored yellow. We highlight the corresponding 12 arginine at position 36 that extends within the lumen of the rotor.

field⁹⁰.

Finally, it is worth mentioning that while all the results presented here relate to protein structural transitions, there is nothing in the equation of motion or even the potential that would prevent applications to nucleic acids, both RNAs and DNAs, as well as to protein-nucleic acids complexes.

ACKNOWLEDGMENTS

The work discussed here originated from a visit by P.K. at the Institut de Physique Théorique, CEA Saclay, France, during the fall of 2025. He thanks them for their hospitality and financial support.

Appendix A: Pairwise potential: Gradient, Hessian, and Laplacian

1. Notations

Let P be a protein consisting of N atoms, with atom i characterized by its position \mathbf{r}_i . The whole molecule is described by a $3N$ position vector $\mathbf{X} = (\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N)^T \in \mathbb{R}^{3N}$. We set \mathbf{X}^0 to be the start conformation of the system. For two atoms i and j of P , we set $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$ and $r_{ij}^0 = |\mathbf{r}_i^0 - \mathbf{r}_j^0|$ to be the Euclidean distances between them in a conformation \mathbf{X} and in the ground-state conformation \mathbf{X}^0 , respectively. Similarly, we define $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{r}_{ij}^0 = \mathbf{r}_i^0 - \mathbf{r}_j^0$ the vectors along the edge (i,j) in \mathbf{X} and in \mathbf{X}^0 , respectively. The unit vector along the same edge is referred to as \mathbf{e}_{ij} and is equal to

$$\mathbf{e}_{ij} = \frac{\mathbf{r}_{ij}}{r_{ij}} \quad (\text{A1})$$

We write I_3 for the 3×3 identity matrix.

2. A generic pairwise potential

For a pair atoms (i, j) we define

$$U_{ij} = f(r_{ij}) \quad (\text{A2})$$

f is a function that is C^∞ almost everywhere. This general formulation applies to the bond term, the collision term, and the elastic term of the full potential we have used to study conformational transitions in proteins. Note that in the case of the collision potential, f is not defined for $r_{ij} = 0$.

3. Gradient

It is convenient to define the vector \mathbf{W}_{ij}

$$\mathbf{W}_{ij} = (0, \dots, 0, \mathbf{e}_{ij}, 0, \dots, 0, -\mathbf{e}_{ij}, 0, \dots, 0), \quad (\text{A3})$$

namely, \mathbf{W}_{ij} is a vector in \mathbb{R}^{3N} that is zero everywhere, except at positions i and j where it is equal to the unit vector \mathbf{e}_{ij} and its opposite, respectively. We have,

$$\nabla U_{ij} = f'(r_{ij}) \mathbf{W}_{ij} \quad (\text{A4})$$

4. Hessian

When we take second derivatives with respect to the Cartesian coordinates of particles i and j , the Hessian of the pair contribution U_{ij} is a $3N \times 3N$ matrix which we write in 3×3 -block form as

$$H = \begin{bmatrix} 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & H(i, i) & \cdots & H(i, j) & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & H(j, i) & \cdots & H(j, j) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \quad (\text{A5})$$

Expressions for the different blocks in this matrix are obtained by differentiating the gradient. Notice first that

$$\nabla_{\mathbf{r}_i} \mathbf{e}_i = \frac{1}{r_{ij}} (I_3 - \mathbf{e}_{ij} \mathbf{e}_{ij}^T). \quad (\text{A6})$$

Then,

$$H(i, i) = f''(r) \mathbf{e}_{ij} \mathbf{e}_{ij}^T + \frac{f'(r_{ij})}{r_{ij}} (I_3 - \mathbf{e}_{ij} \mathbf{e}_{ij}^T). \quad (\text{A7})$$

As U_{ij} depends only on relative position, and not absolute position, it is easy to verify that

$$\begin{aligned} H(j, i) &= H(i, j) = -H(i, i) \\ H(j, j) &= H(i, i). \end{aligned} \quad (\text{A8})$$

5. Laplacian and its gradient

As the trace is a linear operator, we get

$$\begin{aligned} \text{tr}(H_{ii}) &= f''(r_{ij}) \text{tr}(\mathbf{e}_{ij} \mathbf{e}_{ij}^T) + \frac{f'(r_{ij})}{r_{ij}} \text{tr}(I_3 - \mathbf{e}_{ij} \mathbf{e}_{ij}^T) \\ &= f''(r_{ij}) + 2 \frac{f'(r_{ij})}{r_{ij}}. \end{aligned} \quad (\text{A9})$$

As the full trace involves both H_{ii} and H_{jj} and that these two blocks are equal, the trace of the Hessian H associated with the pair (i, j) is:

$$\Delta U_{ij} = 2f''(r_{ij}) + 4 \frac{f'(r_{ij})}{r_{ij}}. \quad (\text{A10})$$

Note that the gradient of the Laplacian is then:

$$\nabla(\Delta U_{ij}) = \left(2f'''(r_{ij}) + 4 \frac{f''(r_{ij})}{r_{ij}} - 4 \frac{f'(r_{ij})}{r_{ij}^2} \right) \mathbf{W}_{ij} \quad (\text{A11})$$

6. Complete pairwise potentials: bond, collision, and elastic

A potential associated with the whole protein is the sum of the inter atomic interactions over a preset list of pairs S :

$$U = \sum_{(i,j) \in S} f(r_{ij}). \quad (\text{A12})$$

Linear superposition applies and we can just sum the gradients, Hessian, and Laplacian.

Let us consider a single continuous chain with N C_α atoms. The "bond" potential is designed to constrain the links between consecutive C_α s, i.e., to maintain their lengths as constant as possible. It is defined as:

$$U_b = k_b \sum_{i=1}^{N-1} (r_{ij} - r_{ij}^0)^2, \quad (\text{A13})$$

where we have set $j = i + 1$ and $k_b = 100\epsilon_G$. This pairwise potential maps exactly to the generic case considered above, using $f(r_{ij}) = (r_{ij} - r_{ij}^0)^2$ with r_{ij}^0 being a constant. Its gradient, Hessian, and Laplacian are then derived from the equations above:

$$\begin{aligned} \nabla U_b &= 2k_b \sum_i (r_{ij} - r_{ij}^0) \mathbf{W}_{ij} \\ H_b(i, j) &= -2k_b \mathbf{e}_{ij} \mathbf{e}_{ij}^T \\ &\quad - \frac{2k_b (r_{ij} - r_{ij}^0)}{r_{ij}} (I_3 - \mathbf{e}_{ij} \mathbf{e}_{ij}^T) \\ H_b(i, i) &= - \sum_{j \in \{i-1, i+1\}} H_b(i, j) \\ \Delta(U_b) &= 4k_b \sum_i \sum_{j \in \{i-1, i+1\}} (3 - 2 \frac{r_{ij}^0}{r_{ij}}) \\ \nabla(\Delta U_b) &= -8k_b \sum_i \sum_{j \in \{i-1, i+1\}} \frac{r_{ij}^0}{r_{ij}} \mathbf{W}_{ij} \end{aligned} \quad (\text{A14})$$

Note that special care is needed for the first and last atoms of the chain.

The collision potential is a 12-6 Lennard Jones potential set to reduce the number of collision during the dynamics., It is defined as

$$U_{vdW} = \sum_i \sum_{j \in N(i)} \left(\frac{C_1}{r_{ij}^{12}} - \frac{C_2}{r_{ij}^6} \right), \quad (\text{A15})$$

where $N(i)$ defines the neighborhood of i , i.e., the list of atoms that are within a cutoff distance R_c of i , and $C_1 = C_2 = \epsilon_G$. This pairwise potential maps also to the generic case considered above, using $f(r_{ij}) = \frac{C_1}{r_{ij}^{12}} - \frac{C_2}{r_{ij}^6}$. Its gradient, Hessian, and Laplacian are derived from the equations above:

$$\begin{aligned} \nabla U_{vdW} &= \sum_i \sum_{j \in N(i)} \left(\frac{-12C_1}{r_{ij}^{13}} + \frac{6C_2}{r_{ij}^7} \right) \mathbf{w}_{ij} \\ H_{vdW}(i, j) &= - \left(\frac{156C_1}{r_{ij}^{14}} - \frac{42C_2}{r_{ij}^8} \right) \mathbf{e}_{ij} \mathbf{e}_{ij}^T \\ &\quad - \left(\frac{12C_1}{r_{ij}^{14}} - \frac{6C_2}{r_{ij}^8} \right) (I_3 - \mathbf{e}_{ij} \mathbf{e}_{ij}^T) \\ H_{vdW}(i, i) &= - \sum_{j \in N(i)} H_{col}(i, j) \\ \Delta(U_{vdW}) &= \sum_i \sum_{j \in N(i)} \left(\frac{132C_1}{r_{ij}^{14}} - \frac{30C_2}{r_{ij}^8} \right) \\ \nabla(\Delta U_{vdW}) &= \sum_i \sum_{j \in N(i)} \left(-\frac{1848C_1}{r_{ij}^{15}} + \frac{240C_2}{r_{ij}^9} \right) \mathbf{w}_{ij} \end{aligned} \quad (\text{A16})$$

The elastic potential is a simple quadratic potential:

$$U_{el} = k_e \sum_i \sum_{j \in N(i)} g(r_{ij}) r_{ij}^2, \quad (\text{A17})$$

where $k_e = \frac{10\epsilon_G}{N_{pair}}$, N_{pair} is the number of pairs (i, j) , $N(i)$ is the cutoff-based neighborhood of i already defined for the collision term, $g(r_{ij})$ is a Fermi-like function that defines a smooth cutoff:

$$g(r) = \frac{1}{1 + \exp\left(\frac{r-d_0}{a_0}\right)} \quad (\text{A18})$$

where d_0 and a_0 are constants. We need three levels of derivatives for $g(r)$:

$$\begin{aligned} g'(x) &= -\frac{1}{a_0} g(x) (1 - g(x)) \\ g''(x) &= \frac{1}{a_0^2} g(x) (1 - g(x)) (1 - 2g(x)) \\ g'''(x) &= -\frac{1}{a_0^3} g(x) (1 - g(x)) (1 - 6g(x) + 6g(x)^2) \end{aligned} \quad (\text{A19})$$

The gradient, Hessian, and Laplacian of the elastic potential are then derived as:

$$\begin{aligned} \nabla U_{el} &= k_e \sum_i \sum_{j \in N(i)} (g'(r_{ij}) r_{ij}^2 + 2g(r_{ij}) r_{ij}) \mathbf{w}_{ij} \\ H_{el}(i, j) &= -k_e (g''(r_{ij}) r_{ij}^2 + 4g'(r_{ij}) r_{ij} + 2g(r_{ij})) \mathbf{e}_{ij} \mathbf{e}_{ij}^T \\ &\quad - k_e (g'(r_{ij}) r_{ij} - 2g(r_{ij})) (I_3 - \mathbf{e}_{ij} \mathbf{e}_{ij}^T) \\ H_{el}(i, i) &= - \sum_{j \in N(i)} H_{el}(i, j) \\ \Delta(U_{el}) &= k_e \sum_i \sum_{j \in N(i)} (2g''(r_{ij}) r_{ij}^2 + 12g'(r_{ij}) + 12g(r_{ij})) \\ \nabla(\Delta U_{el}) &= k_e \sum_i \sum_{j \in N(i)} (2g'''(r_{ij}) r_{ij}^2 + 16g''(r_{ij}) r_{ij} \\ &\quad + 24g'(r_{ij})) \mathbf{w}_{ij} \end{aligned} \quad (\text{A20})$$

Appendix B: Angular potential: Gradient, Hessian, and Laplacian

Let us consider the pseudo angle θ_i formed by three consecutive C_α along the backbone of a protein, i , j , and k , and centered at j . As above, we set the positions of these atoms as \mathbf{r}_i , \mathbf{r}_j , and \mathbf{r}_k . The role of the angular potential is to restrain this angle θ_j to match its value θ_j^0 in the starting conformation of the protein:

$$U_\theta(j) = K_a (\theta_j - \theta_j^0)^2. \quad (\text{B1})$$

Compared to the pairwise potentials defined in Appendix A, $U_\theta(j)$ is a three-body potential. Computing its derivatives require care. We start with the simpler problem of computing the derivatives of θ_j with respect to the positions r_i , r_j , and r_k .

1. Gradient, Hessian, and Laplacian of θ_j :

Let $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $\mathbf{r}_{kj} = \mathbf{r}_k - \mathbf{r}_j$. Let \mathbf{e}_{ij} and \mathbf{e}_{kj} be the corresponding unit vectors. We also set $r_{ij} = |\mathbf{r}_{ij}|$ and $r_{kj} = |\mathbf{r}_{kj}|$. We introduce the cosine and sine of the angle at j :

$$c = \mathbf{e}_{ij} \cdot \mathbf{e}_{kj}, \quad s = \sqrt{1 - c^2}. \quad (\text{B2})$$

The angle at vertex j is then

$$\theta_j = \arccos(c). \quad (\text{B3})$$

The cosine c is a smooth function of \mathbf{r}_{ij} and \mathbf{r}_{kj} :

$$c = \frac{\mathbf{r}_{ij} \cdot \mathbf{r}_{kj}}{r_{ij} r_{kj}} \quad (\text{B4})$$

Its first derivatives are

$$\begin{aligned} \frac{\partial c}{\partial \mathbf{r}_{ij}} &= \frac{1}{r_{ij}} (\mathbf{e}_{kj} - c \mathbf{e}_{ij}), \\ \frac{\partial c}{\partial \mathbf{r}_{kj}} &= \frac{1}{r_{kj}} (\mathbf{e}_{ij} - c \mathbf{e}_{kj}). \end{aligned} \quad (\text{B5})$$

The second derivatives of c with respect to \mathbf{r}_{ij} and \mathbf{r}_{kj} are 3×3 matrices:

$$\begin{aligned}\frac{\partial^2 c}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{ij}} &= \frac{1}{r_{ij}^2} [-(\mathbf{e}_{kj} \mathbf{e}_{ij}^T + \mathbf{e}_{ij} \mathbf{e}_{kj}^T) + 3c \mathbf{e}_{ij} \mathbf{e}_{ij}^T - c I_3], \\ \frac{\partial^2 c}{\partial \mathbf{r}_{kj} \partial \mathbf{r}_{kj}} &= \frac{1}{r_{kj}^2} [-(\mathbf{e}_{ij} \mathbf{e}_{kj}^T + \mathbf{e}_{kj} \mathbf{e}_{ij}^T) + 3c \mathbf{e}_{kj} \mathbf{e}_{kj}^T - c I_3], \\ \frac{\partial^2 c}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}} &= \frac{1}{r_{ij} r_{kj}} [I - \mathbf{e}_{ij} \mathbf{e}_{ij}^T - \mathbf{e}_{kj} \mathbf{e}_{kj}^T + c \mathbf{e}_{ij} \mathbf{e}_{kj}^T].\end{aligned}\quad (\text{B6})$$

Since $\theta_j = \arccos(c)$, its derivatives follow from the chain rule:

$$\begin{aligned}\frac{\partial \theta_j}{\partial \mathbf{r}_{ij}} &= -\frac{1}{s} \frac{\partial c}{\partial \mathbf{r}_{ij}} \\ \frac{\partial \theta_j}{\partial \mathbf{r}_{kj}} &= -\frac{1}{s} \frac{\partial c}{\partial \mathbf{r}_{kj}}\end{aligned}\quad (\text{B7})$$

Differentiating again gives the Hessian blocks for θ_j with respect to \mathbf{r}_{ij} and \mathbf{r}_{kj} (each 3×3):

$$\frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{ij}} = -\frac{1}{s} \frac{\partial^2 c}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{ij}} - \frac{c}{s^3} \left(\frac{\partial c}{\partial \mathbf{r}_{ij}} \frac{\partial c}{\partial \mathbf{r}_{ij}}^T \right), \quad (\text{B8})$$

$$\frac{\partial^2 \theta_j}{\partial \mathbf{r}_{kj} \partial \mathbf{r}_{kj}} = -\frac{1}{s} \frac{\partial^2 c}{\partial \mathbf{r}_{kj} \partial \mathbf{r}_{kj}} - \frac{c}{s^3} \left(\frac{\partial c}{\partial \mathbf{r}_{kj}} \frac{\partial c}{\partial \mathbf{r}_{kj}}^T \right), \quad (\text{B9})$$

$$\frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}} = -\frac{1}{s} \frac{\partial^2 c}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}} - \frac{c}{s^3} \left(\frac{\partial c}{\partial \mathbf{r}_{ij}} \frac{\partial c}{\partial \mathbf{r}_{kj}}^T \right). \quad (\text{B10})$$

Using

$$\frac{\partial \mathbf{r}_{ij}}{\partial \mathbf{r}_{ii}} = I, \quad \frac{\partial \mathbf{r}_{ij}}{\partial \mathbf{r}_{ji}} = I, \quad \frac{\partial \mathbf{r}_{kj}}{\partial \mathbf{r}_{ki}} = I, \quad \frac{\partial \mathbf{r}_{kj}}{\partial \mathbf{r}_{jk}} = I, \quad (\text{B11})$$

we can express the gradient and Hessian of θ_j with respect to the positions of i , j , and k :

$$\begin{aligned}\nabla_{p_i} \theta_j &= \frac{\partial \theta}{\partial \mathbf{r}_{ij}}, \\ \nabla_{p_k} \theta_j &= \frac{\partial \theta}{\partial \mathbf{r}_{kj}}, \\ \nabla_{p_j} \theta_j &= -\frac{\partial \theta}{\partial \mathbf{r}_{ij}} - \frac{\partial \theta}{\partial \mathbf{r}_{kj}}.\end{aligned}\quad (\text{B12})$$

The Hessian of θ_j can be written in block form:

$$H_\theta = \begin{bmatrix} H_{ii} & H_{ij} & H_{ik} \\ H_{ji} & H_{jj} & H_{jk} \\ H_{ki} & H_{kj} & H_{kk} \end{bmatrix}, \quad (\text{B13})$$

where each block is a 3×3 matrix:

$$\begin{aligned}H_{ii} &= \frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{ij}}, \\ H_{ik} &= \frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}}, \\ H_{ij} &= -\frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{ij}} - \frac{\partial^2 \theta}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}}, \\ H_{jk} &= -\frac{\partial^2 \theta_j}{\partial \mathbf{r}_{kj} \partial \mathbf{r}_{kj}} - \left(\frac{\partial^2 \theta}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}} \right)^T, \\ H_{jj} &= \frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{ij}} + \frac{\partial^2 \theta_j}{\partial \mathbf{r}_{kj} \partial \mathbf{r}_{kj}} + \frac{\partial^2 \theta_j}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}} + \left(\frac{\partial^2 \theta}{\partial \mathbf{r}_{ij} \partial \mathbf{r}_{kj}} \right)^T, \\ H_{ki} &= (H_{ik})^T, \\ H_{kj} &= (H_{jk})^T, \\ H_{kk} &= \frac{\partial^2 \theta}{\partial \mathbf{r}_{kj} \partial \mathbf{r}_{kj}}.\end{aligned}\quad (\text{B14})$$

Finally, we give a formula for the Laplacian of θ_j , i.e., the trace of its Hessian:

$$\Delta \theta_j = \text{tr}(H) = \frac{2}{s} \left(\frac{c}{r_{ij}^2} - \frac{1}{r_{ij} r_{kj}} + \frac{c}{r_{kj}^2} \right). \quad (\text{B15})$$

Defining the auxiliary scalar

$$F = \frac{c}{r_{ij}^2} - \frac{1}{r_{ij} r_{kj}} + \frac{c}{r_{kj}^2},$$

the gradients of $\Delta \theta_j$ with respect to the edge vectors \mathbf{r}_{ij} and \mathbf{r}_{kj} are

$$\begin{aligned}\frac{\partial \Delta \theta}{\partial \mathbf{r}_{ij}} &= \frac{2}{s} \left(-\frac{2c}{r_{ij}^3} + \frac{1}{r_{ij}^2 r_{kj}} \right) \mathbf{e}_{ij} \\ &\quad + \frac{2}{s} \left(\frac{1}{r_{ij}^2} + \frac{1}{r_{kj}^2} + \frac{Fc}{s^2} \right) \frac{\mathbf{e}_{kj} - c \mathbf{e}_{ij}}{r_{ij}}, \\ \frac{\partial \Delta \theta}{\partial \mathbf{r}_{kj}} &= \frac{2}{s} \left(-\frac{2c}{r_{kj}^3} + \frac{1}{r_{ij} r_{kj}^2} \right) \mathbf{e}_{kj} \\ &\quad + \frac{2}{s} \left(\frac{1}{r_{ij}^2} + \frac{1}{r_{kj}^2} + \frac{Fc}{s^2} \right) \frac{\mathbf{e}_{ij} - c \mathbf{e}_{kj}}{r_{kj}}.\end{aligned}\quad (\text{B16})$$

Finally, the gradients of $\Delta \theta$ with respect to vertex positions are

$$\begin{aligned}\nabla_{p_i} \Delta \theta &= \frac{\partial \Delta \theta}{\partial \mathbf{r}_{ij}}, \\ \nabla_{p_k} \Delta \theta &= \frac{\partial \Delta \theta}{\partial \mathbf{r}_{kj}}, \\ \nabla_{p_j} \Delta \theta &= -\frac{\partial \Delta \theta}{\partial \mathbf{r}_{ij}} - \frac{\partial \Delta \theta}{\partial \mathbf{r}_{kj}}.\end{aligned}\quad (\text{B17})$$

2. Gradient, Hessian, and Laplacian of $U_\theta(j)$:

We have everything we need to define the derivatives of $U_\theta(j)$. First, the gradient is given by:

$$\nabla U_\theta(j) = 2(\theta_j - \theta_j^0) \nabla \theta_j \quad (\text{B18})$$

where $\nabla \theta_j$ is given in equation B12.

The Hessian of $U_\theta(j)$ is given by:

$$H(U_\theta(j)) = 2(\nabla \theta_j)(\nabla \theta_j)^T + 2(\theta_j - \theta_j^0) H_\theta \quad (\text{B19})$$

with H_θ defined in equations B13 and B14.

The Laplacian of $U_\theta(j)$ is equal to:

$$\Delta U_\theta(j) = 2(\nabla \theta_j)^2 + 2(\theta_j - \theta_j^0) \Delta \theta_j \quad (\text{B20})$$

where $\Delta \theta_j$ is given in equation B17.

Finally, the derivatives of the Laplacian of $U_\theta(j)$ are given by:

$$\begin{aligned} \nabla \Delta U_\theta(j) = & 4H_\theta \nabla \theta + 2\Delta \theta_j \nabla \theta + \\ & 2(\theta_j - \theta_j^0) \nabla \Delta \theta \end{aligned} \quad (\text{B21})$$

Appendix C: The effective potential

Recall that the potential W is defined by:

$$W = \frac{1}{4}(\nabla U)^2 - \frac{kT}{2}\Delta U \quad (\text{C1})$$

where U is the full potential, namely the sum of the bond, angle, collision, and elastic potentials. While U is linear in those potentials, it is not the case of W , because of the term (∇U) . However, we are really interested in the gradient of W , given by:

$$\nabla W = \frac{1}{2}H_U \nabla U - \frac{kT}{2}\nabla \Delta U \quad (\text{C2})$$

We have, by linearity,

$$\begin{aligned} \nabla U &= \nabla U_b + \nabla U_\theta + \nabla U_{vdW} + \nabla U_{el} \\ H_U &= H_b + H_{ang} + H_{col} + H_{el} \\ \Delta E &= \Delta U_b + \Delta U_\theta + \Delta U_{vdW} + \Delta U_{el}, \end{aligned} \quad (\text{C3})$$

with all these terms defined in appendices A and B.

- ¹J. , R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature* **596**, 583–589 (2021).
- ²The Nobel Committee for Chemistry, “The Nobel Prize in Chemistry 2024,” Nobel Prize Press Release (2024).
- ³M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, *et al.*, “Accurate prediction of protein structures and interactions using a three-track neural network,” *Science* **373**, 871–876 (2021).
- ⁴R. Krishna, J. Wang, W. Ahern, P. Sturmfels, P. Venkatesh, I. Kalvet, G. R. Lee, F. S. Morey-Burrows, I. Anishchenko, I. R. Humphreys, *et al.*, “Generalized biomolecular modeling and design with RoseTTAFold All-Atom,” *Science* **384**, eadl2528 (2024).

- ⁵Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, “Evolutionary-scale prediction of atomic-level protein structure with a language model,” *Science* **379**, 1123–1130 (2023).
- ⁶P. Mollaei and A. Barati Farimani, “Activity map and transition pathways of g protein-coupled receptor revealed by machine learning,” *Journal of Chemical Information and Modeling* **63**, 2296–2304 (2023).
- ⁷H. Park and E. Tajkhorshid, “Machine learning guided sampling of protein transition pathways,” *Biophysical journal* **122**, 282a (2023).
- ⁸K. Georgouli, R. R. Stephany, J. O. Tempkin, C. Santiago, F. Aydin, M. A. Heimann, L. Pottier, X. Zhang, T. S. Carpenter, T. Hsu, *et al.*, “Generating protein structures for pathway discovery using deep learning,” *Journal of Chemical Theory and Computation* **20**, 8795–8806 (2024).
- ⁹B. Liu, J. G. Boysen, I. C. Unarta, X. Du, Y. Li, and X. Huang, “Exploring transition states of protein conformational changes via out-of-distribution detection in the hyperspherical latent space,” *Nature Communications* **16**, 349 (2025).
- ¹⁰Y. T. Pang, K. M. Kuo, L. Yang, and J. C. Gumbart, “Deeppath: Overcoming data scarcity for protein transition pathway prediction using physics-based deep learning,” *bioRxiv* (2025).
- ¹¹W. E and E. Vanden-Eijnden, “Towards a theory of transition paths,” *J. Stat. Phys.* **123**, 503–23 (2006).
- ¹²W. E and E. Vanden-Eijnden, “Transition-path theory and path-finding algorithms for the study of rare events,” *Annu Rev Phys Chem* **61**, 391–420 (2010).
- ¹³E. Vanden-Eijnden, “Transition path theory,” *Adv. Exp. Med. Biol.* **797**, 91–100 (2014).
- ¹⁴M. Kim, R. Jernigan, and G. Chirikjian, “Efficient generation of feasible pathways for protein conformational transitions,” *Biophys. J.* **83**, 1620–1630 (2002).
- ¹⁵D. R. Weiss and M. Levitt, “Can morphing methods predict intermediate structures?” *J. Mol. Biol.* **385**, 665–674 (2009).
- ¹⁶P. Maragakis and M. Karplus, “Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase,” *J. Mol. Biol.* **352**, 807–822 (2005).
- ¹⁷W. Zheng, B. Brooks, and G. Hummer, “Protein conformational transitions explored by mixed elastic network models,” *Proteins: Struct. Func. Bioinfo.* **69**, 43–57 (2007).
- ¹⁸M. Tekpinar and W. Zheng, “Predicting order of conformational changes during protein conformational transitions using an interpolated elastic network model,” *Proteins: Struct. Func. Bioinfo.* **78**, 2469–2481 (2010).
- ¹⁹F. Pinski and A. Stuart, “Transition paths in molecules: gradient descent in pathspace,” *J. Chem. Phys.* **132**, 184104 (2010).
- ²⁰H. Jonsson, G. Mills, and K. W. Jacobsen, “Nudged Elastic Band method for finding minimum energy paths of transitions,” in *Classical and Quantum Dynamics in Condensed Phase Simulations*, edited by B. J. Berne, G. Ciccotti, and D. F. Coker (World Scientific, Singapore, 1998) Chap. 16, pp. 385–404.
- ²¹G. Henkelman, B. Uberuaga, and H. Jonsson, “A climbing image nudged elastic band method for finding saddle points and minimum energy paths,” *J. Chem. Phys.* **113**, 9901–9904 (2000).
- ²²D. Sheppard, R. Terrell, and G. Henkelman, “Optimization methods for finding minimum energy paths,” *J. Chem. Phys.* **128**, 134106 (2008).
- ²³W. E, W. Ren, and E. Vanden-Eijnden, “String method for the study of rare events,” *Phys. Rev. B* **66**, 052301 (2002).
- ²⁴W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, “Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide,” *J. Chem. Phys.* **123**, 134109 (2005).
- ²⁵W. E, W. Ren, and E. Vanden-Eijnden, “Simplified and improved string method for computing the minimum energy paths in barrier-crossing events,” *J. Chem. Phys.* **126**, 164103 (2007).
- ²⁶E. Vanden-Eijnden and M. Venturoli, “Revisiting the finite temperature string method for the calculation of reaction tubes and free energies,” *J. Chem. Phys.* **130**, 194103 (2009).
- ²⁷W. Ren and E. Vanden-Eijnden, “A climbing string method for saddle point search,” *J. Chem. Phys.* **138**, 134105 (2013).
- ²⁸L. Maragliano, B. Roux, and E. Vanden-Eijnden, “Comparison between Mean Forces and Swarms-of-Trajectories String Methods,” *J. Chem. Theory Comput.* **10**, 524–533 (2014).
- ²⁹L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, “String method in collective variables: minimum free energy paths and isocommit-

- tor surfaces," *J. Chem. Phys.* **125**, 24106 (2006).
- ³⁰A. Pan, D. Sezer, and B. Roux, "Finding transition pathways using the string method with swarms of trajectories," *J. Phys. Chem. B*, **112**, 3432–3440 (2008).
 - ³¹Y. Matsunaga, H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu, and A. Kidera, "Minimum free energy path of ligand-induced transition in adenylate kinase," *PLoS Comput. Biol.* **8**, e1002555 (2012).
 - ³²D. Branduardi and J. D. Faraldo-Gomez, "String method for calculation of minimum free-energy paths in Cartesian space in freely-tumbling systems," *J. Chem. Theory Comput.* **9**, 4140–4154 (2013).
 - ³³D. Dürr and A. Bach, "The onsager-machlup function as lagrangian for the most probable path of a diffusion process," *Commun. Math. Phys.* **60**, 153–170 (1978).
 - ³⁴S. Raja, M. Šípka, M. Psenka, T. Kreiman, M. Pavelka, and A. S. Krishnapriyan, "Action-minimization meets generative modeling: Efficient transition path sampling with the onsager-machlup functional," *arXiv preprint arXiv:2504.18506* (2025).
 - ³⁵R. Olender and R. Elber, "Calculation of classical trajectories with a very large time step: formalism and numerical examples," *J. Chem. Phys.* **105**, 9299–9315 (1996).
 - ³⁶P. Eastman, N. Gronbech-Jensen, and S. Doniach, "Simulation of protein folding by reaction path annealing," *J. Chem. Phys.* **114**, 3823 (2001).
 - ³⁷J. Franklin, P. Koehl, S. Doniach, and M. Delarue, "Minactionpath: maximum likelihood trajectory for large-scale structural transitions in a coarse grained locally harmonic energy landscape," *Nucl. Acids. Res.* **35**, W477–W482 (2007).
 - ³⁸P. Faccioli, M. Sega, F. Pederiva, and H. Orland, "Dominant pathways in protein folding," *Phys. Rev. Lett.* **97**, 108101 (2006).
 - ³⁹E. Vanden-Eijnden and M. Heymann, "The geometric minimum action method for computing minimum energy paths," *J. Chem. Phys.* **128**, 061103 (2008).
 - ⁴⁰X. Zhou, W. Ren, and W. E, "Adaptive minimum action method for the study of rare events," *J. Chem. Phys.* **128**, 104111 (2008).
 - ⁴¹S. Chandrasekaran, J. Dhas, N. Dokholyan, and C. Carter Jr, "A modified path algorithm rapidly generates transition states comparable to those found by other well established algorithms," *Struct. Dyn.* **3**, 012101 (2016).
 - ⁴²P. Koehl, "Minimum action transition paths connecting minima on an energy surface," *J. Chem. Phys.* **145**, 184111 (2016).
 - ⁴³P. Koehl, R. Navaza, M. Tekpinar, and M. Delarue, "Minactionpath2: path generation between different conformations of large macromolecular assemblies by action minimization," *Nucleic Acids Research* **52**, W256–W263 (2024).
 - ⁴⁴A. Stuart, P. Wiberg, and J. Voss, "Conditional path sampling of sdes and the langevin mcmc method," *Commun. Math. Sci.* **2**, 685–697 (2004).
 - ⁴⁵M. Hairer, A. Stuart, and J. Voss, "Analysis of SPDEs arising in path sampling part ii: the nonlinear case," *Ann. Appl. Probab.* **17**, 1657–1706 (2007).
 - ⁴⁶H. Orland, "Generating transition paths by langevin bridges," *J. Chem. Phys.* **134**, 174114 (2011).
 - ⁴⁷J. Mattingly, N. Pillai, and A. Stuart, "Diffusion limits of the random walk metropolis algorithm in high dimensions," *Ann. Appl. Probab.* **22**, 881–930 (2012).
 - ⁴⁸M. Delarue, P. Koehl, and H. Orland, "Ab initio sampling of transition paths by conditioned langevin dynamics," *J. Chem. Phys.* **147**, 152703 (2017).
 - ⁴⁹P. Koehl and H. Orland, "Sampling constrained stochastic trajectories using brownian bridges," *The Journal of Chemical Physics* **157**, 054105 (2022).
 - ⁵⁰J. Chodera, N. Singhal, V. Pande, K. Dill, and W. Swope, "Automatic discovery of metastable states for the construction of markov models of macromolecular conformational dynamics," *J. Chem. Phys.* **126**, 155101 (2007).
 - ⁵¹G. Bowman, K. Beauchamp, G. Boxer, and V. Pande, "Progress and challenges in the automated construction of markov state models for full protein systems," *J. Chem. Phys.* **124**, 124101 (2009).
 - ⁵²V. Pande, K. Beauchamp, and G. Bowman, "Everything you wanted to know about markov state models but were afraid to ask," *Methods* **52**, 99–105 (2010).
 - ⁵³A. Faradjian and R. Elber, "Computing time scales from reaction coordinates by milestoneing," *J. Chem. Phys.* **120**, 10880–10890 (2004).
 - ⁵⁴J. Bello-Rivas and R. Elber, "Exact milestoneing," *J. Chem. Phys.* **142**, 094102 (2015).
 - ⁵⁵P. Laowanapiban, M. Kapustina, C. Vonnrhein, M. Delarue, P. Koehl, and C. W. Carter, "Independent saturation of three trprs subsites generates a partially assembled state similar to those observed in molecular simulations," *Proc. Natl. Acad. Sci. (USA)* **106**, 1790–1795 (2009).
 - ⁵⁶V. Weinreb, L. Li, S. N. Chandrasekaran, P. Koehl, M. Delarue, and C. W. Carter, *J. Biol. Chem.* **289**, 4367–4376 (2014).
 - ⁵⁷J. Doob, "Conditional brownian motion and the boundary limits of harmonic functions," *Bull. Soc. Math. France* **85**, 431–458 (1957).
 - ⁵⁸N. V. Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, The Netherlands, 1992).
 - ⁵⁹R. Feynman and A. Hibbs, *Quantum Mechanics and Path Integrals* (McGraw-Hill, New York, NY, 1965).
 - ⁶⁰L. Onsager and S. Machlup, "Fluctuations and irreversible processes," *Phys. Rev.* **91**, 1505–1512 (1953).
 - ⁶¹I. Gopich and A. Szabo, "Theory of the statistics of kinetic transitions with application to single-molecule enzyme catalysis," *J. Chem. Phys.* **124**, 154712 (2006).
 - ⁶²W. Kim and R. Netz, "The mean shape of transition and first-passage paths," *J. Chem. Phys.* **143**, 224108 (2015).
 - ⁶³M. Laleman, E. Carlon, and H. Orland, "Transition path time distributions," *J. Chem. Phys.* **147**, 214103 (2017).
 - ⁶⁴P. Kloeden and E. Platen, *Numerical solution of stochastic differential equations* (Springer, Berlin, Germany, 1992).
 - ⁶⁵M. Tirion, "Large amplitude elastic motions in proteins from a single parameter, atomic analysis," *Phys. Rev. Lett.* **77**, 1905–1908 (1996).
 - ⁶⁶C. Clementi, H. Nymeyer, and J. Onuchic, "Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? an investigation for small globular proteins," *J. Mol. Biol.* **298**, 937–953 (2000).
 - ⁶⁷H. Na, T.-L. Lin, and G. Song, "Generalized spring tensor models for protein fluctuation dynamics and conformation changes," in *Protein Conformational Dynamics*, edited by K.-I. Han, X. Zhang, and M.-j. Yang (Springer International Publishing, Cham, 2014) pp. 107–135.
 - ⁶⁸P. E. Rouse, "A theory of the linear viscoelastic properties of dilute solutions of coiling polymers," *J. Chem. Phys.* **21**, 1272–1280 (1953).
 - ⁶⁹P.-G. de Gennes, "Dynamics of ideal chains," in *Scaling Concepts in Polymer Physics* (Cornell University Press, Ithaca, NY, 1979) Chap. 6, pp. 208–243.
 - ⁷⁰H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics* **2**, 433–459 (2010).
 - ⁷¹A. Amadei, A. B. Linssen, and H. J. Berendsen, "Essential dynamics of proteins," *Proteins: Structure, Function, and Bioinformatics* **17**, 412–425 (1993).
 - ⁷²H. J. Berendsen and S. Hayward, "Collective protein dynamics in relation to function," *Current opinion in structural biology* **10**, 165–169 (2000).
 - ⁷³Y. Saad, *Numerical Methods for Large Eigenvalue Problems* (SIAM, Philadelphia, PA, 2011).
 - ⁷⁴L. Orellana, O. Yoluk, O. Carrillo, M. Orozco, and E. Lindahl, "Prediction and validation of protein intermediate states from structurally rich ensembles and coarse-grained simulations," *Nature Communications* **7**, 12575 (2016).
 - ⁷⁵L. Orellana, J. Gustavsson, C. Bergh, O. Yoluk, and E. Lindahl, "ebdms server: protein transition pathways with ensemble analysis in 2d-motion spaces," *Bioinformatics* **35**, 3505–3507 (2019).
 - ⁷⁶K. A. Henzler-Wildman, M. Lei, V. Thai, S. J. Kerns, M. Karplus, and D. Kern, "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis," *Nature* **450**, 913–916 (2007).
 - ⁷⁷S. L. Seyler and O. Beckstein, "Sampling large conformational transitions: adenylate kinase as a testing ground," *Mol. Simul.* **40**, 855–877 (2014).
 - ⁷⁸H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucl. Acids. Res.* **28**, 235–242 (2000).
 - ⁷⁹C. Toyoshima and H. Nomura, "Structural changes in the calcium pump accompanying the dissociation of calcium," *Nature* **418**, 605–611 (2002).
 - ⁸⁰D. M. Himmel, S. Gourinath, L. Reshetnikova, Y. Shen, A. G. Szent-Györgyi, and C. Cohen, "Crystallographic findings on the internally uncoupled and near-rigor states of myosin: further insights into the mechanics of the motor," *Proc. Natl. Acad. Sci. (USA)* **99**, 12645–12650 (2002).
 - ⁸¹P. D. Boyer, "The ATP synthase—a splendid molecular machine," *Annual review of biochemistry* **66**, 717–749 (1997).

- ⁸²W. Junge and N. Nelson, “ATP synthase,” *Annu. Rev. Biochem.* **84**, 631–657 (2015).
- ⁸³R. K. Nakamoto, J. A. Baylis Scanlon, and M. K. Al-Shawi, “The rotary mechanism of the ATP synthase,” *Arch. Biochem. Biophys.* **476**, 43–50 (2008).
- ⁸⁴L. Zhou and L. A. Sazanov, “Structure and conformational plasticity of the intact thermus thermophilus V/A-type ATPase,” *Science* **365**, eaaw9144 (2019).
- ⁸⁵S. Majumdar and H. Orland, “Effective langevin equations for constrained stochastic processes,” *J. Stat. Mech. Theory Exp.* **2015**, P06039 (2015).
- ⁸⁶S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, and A. H. De Vries, “The MARTINI force field: coarse grained model for biomolecular simulations,” *J. Phys. Chem. B.* **111**, 7812–7824 (2007).
- ⁸⁷L. Monticelli, S. K. Kandasamy, X. Periole, R. G. Larson, D. P. Tieleman, and S.-J. Marrink, “The MARTINI coarse-grained force field: Extension to proteins,” *J. Chem. Theory Comput.* **4**, 819–834 (2008).
- ⁸⁸P. C. T. Souza, L. Borges-Araújo, C. Brasnett, R. A. Moreira, F. Grünewald, P. Park, L. Wang, H. Razmazma, A. C. Borges-Araújo, L. F. Cofas-Vargas, L. Monticelli, R. Mera-Adasme, M. N. Melo, S. Wu, S. J. Marrink, A. B. Poma, and S. Thallmair, “GōMartini 3: From large conformational changes in proteins to environmental bias corrections,” *Nature Communications* **16**, 4051 (2025).
- ⁸⁹H. Taketomi, Y. Ueda, and N. Gō, “Studies on protein folding, unfolding and fluctuations by computer simulation: I. the effect of specific amino acid sequence represented by specific inter-unit interactions,” *International journal of peptide and protein research* **7**, 445–459 (1975).
- ⁹⁰P. C. Souza, R. Alessandri, J. Barnoud, S. Thallmair, I. Faustino, F. Grünewald, I. Patmanidis, H. Abdizadeh, B. M. Bruininks, T. A. Wassenaar, *et al.*, “Martini 3: a general purpose force field for coarse-grained molecular dynamics,” *Nature methods* **18**, 382–388 (2021).