Ab initio sampling of transition paths by Conditioned Langevin Dynamics

Marc Delarue, Patrice Koehl, Henri Orland

Ab initio sampling of transition paths by Conditioned Langevin Dynamics

Marc Delarue *
Unité de Dynamique Structurale des Macromolécules,
UMR 3528 du CNRS, Institut Pasteur, 75015 Paris, France
e-mail: delarue@pasteur.fr

Patrice Koehl
Department of Computer Science and Genome Center,
University of California, Davis, CA 95616, USA.
e-mail: koehl@cs.ucdavis.edu

Henri Orland
Institut de Physique Théorique, CEA, URA 2306 du CNRS,
F-91191 Gif-sur-Yvette, France.
and
Beijing Computational Science Research Center,
Building 9, East Zone, ZPark II,
No.10 East Xibeiwang Road, Haidian District, Beijing 100193, China
e-mail: henri.orland@cea.fr

May 15, 2017

*Corresponding author; e-mail: delarue@pasteur.fr

Abstract

We propose a novel stochastic method to generate Brownian paths conditioned to start at an initial point and end at a given final point during a fixed time t_f under a given potential $U(x)$. These paths are sampled with a probability given by the overdamped Langevin dynamics. We show that these paths can be exactly generated by a local *Stochastic Partial Differential Equation (SPDE)*. This equation cannot be solved in general but we present several approximations that are valid either in the low temperature regime or in the presence of barrier crossing. We show that this method warrants the generation of statistically independent transition paths. It is computationally very efficient. We illustrate the method first on two simple potentials, the two dimensional Mueller potential as well as on the Mexican hat potential, and then on the multi-dimensional problem of conformational transitions in proteins using the "Mixed Elastic Network Model" as a benchmark.

Keywords. Langevin dynamics, Stochastic partial differential equations, Transition paths, Conformational transitions.

1 Introduction

Biochemical and biological machines are controlled by their dynamical properties. Indeed, it is the ability of molecules to change conformations that leads to their activity [1]. If one could predict the sequence of events leading from one conformational state to another one, a whole new world would be open to structure-inspired drug design techniques, that could focus not only on end-states but also on intermediates to control or block the reaction [2,3]. Observing experimentally or predicting functional conformational changes is however a very difficult problem. At the core of this problem is the fact that a transition between two conformations of a molecule is a rare event compared to the time scale of the internal dynamics of the molecule. This event is a consequence of random perturbations in the structure of the molecule, drawing its energy from the surrounding heat bath; it is rare whenever the energy barrier that needs to be crossed is high compared to the ambient thermal energy k_BT.
The transition state theory (TST) offers a framework for studying such rare events [4, 5]. The main idea of the TST is that the transition state is a saddle point of the energy surface for the molecule of interest. In many cases, the most probable transition path is then simply the minimum energy path (MEP) along that energy surface. The TST however is limited to situations in which the potential energy surface is rather smooth; it also assumes that every crossing of the energy barrier through the transition state gives rise to a successful reaction. As such, the transition path is often calculated by walking down from the TST to each of the two end states using steepest gradient methods [6]. For systems with a rugged potential energy landscape, or when entropic effects matter, the saddle points do not necessarily play the role of transition states [7].
To alleviate the shortcomings of the TST, Vanden Eijnden and colleagues proposed an alternate view of transitions, the Transition Path Theory (TPT) [7–9]. In principle, TST eliminates the need for sampling the transition path ensemble and 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 [10, 11], gradient descent methods [12–15], the nudged elastic band method [16–18] and the string method [19–24].
Other algorithms are concerned with either finding the Minimum Free Energy Path (MFEP) on the free energy surface for the molecule, [25–28] while others search for paths that minimize a functional, such as Onsager-Machlup functional that is known to reproduce the most probable path for a Langevin equation with constant diffusion coefficient [29], as implemented in the Minimum Action Path (MAP) methods [30–36].
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 [37–40]. Also, a new method called "Milestoning" due to Elber and coll. performs very well [41, 42].
This list is not 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 [43].

Due to the inherent fluctuations underlying the transition phenomenon there are many ways however in which a transition can take place. The methods described above usually generate one path along this transition, the most "probable" one, where probable refers to

minimum energy, free energy, or an action.
Path sampling methods expand upon this view by using this path as a seed to generate a Monte Carlo random walk in the path space of the transition trajectories, and thus generate an ensemble of all possible transition paths [44, 45]. All the relevant kinetic and thermodynamic information related to the transition can then be extracted from the ensemble, such as the reaction mechanism, the transition states, and the rate constants. The main drawback of these methods however is that they are very time consuming and therefore limited to small systems. In addition, they generate highly correlated trajectories because the space of sampled trajectories depends strongly on the initial path.
In parallel to path sampling methods, much effort has recently been dedicated to the development and analysis of Markov State Models (MSMs) [46–48]. 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. Much of the recent work focuses on generating those conformations and the dynamics between them, usually using molecular dynamics simulations. To this day, MSMs remain a computationally intensive method and generate a large number of MD trajectories that are complex to analyze even if recent work by [49] now offers free software to automatically analyze them in a more user-friendly way.

In this paper we are concerned with the problem of path sampling. Following preliminary work by one of us [39], we propose a novel method for generating rapidly completely independent paths using a Stochastic Partial Differential Equation (SPDE). This equation cannot be solved in general. In the original presentation of this equation, dubbed "Langevin Bridges", one of us proposed a simplification based on the symmetric form of the Trotter approximation. This simplification was shown to work well when studying transitions for small systems over short times [39]. It was later found however not to be valid for other conditions; in particular it was not applicable to study transitions between conformations of large molecular systems. In this paper we propose new approximations for solving the SPDE that are valid for different regimes for the dynamics of the system considered, and illustrate their applications on such large bio-molecular systems.

We are especially interested in generating multiple transition paths for conformational transitions in coarse-grained models of proteins. Coarse-graining involves mainly taking one bead per residue, centered on the C-alpha coordinate of each residue. One way to generate a two-well energy landscape is to use a superposition of two coarse-grained elastic potentials, as defined originally by Tirion [50], each one centered on the end points of the trajectory. This usually gives rise to a "cusp" in energy at the transition state, leading to unrealistic values of the activation energy [12, 32, 51], unless ones mixes the two elastic potential with a mixing Temperature, T_m [13, 52]. Previously, we implemented the optimisation of Onsager-Machlup action of a superposition of ENM models, which transform a first-order stochastic equation into a second order deterministic partial derivative equation, leading to a unique trajectory. Here we use the Conditioned Langevin Dynamics (CLD) Equation to generate many plausible trajectories, using the so-called "modulated Mixed ENM" [22, 26, 53], that includes both a energy penalty to avoid steric clashes during the simulation and also non-uniform elastic constants for the pairs of atoms linked by an harmonic potential. Here we use this model to benchmark the CLD method.

The paper is organized as follows. In the next section, we derive the SPDE and describe

the different approximations we have implemented to solve this equation. In the following section, we show applications to two well-studied 2D problems. We next proceed to use the same method with biological macromolecules (proteins in different known conformations), using a simplified energy function, with many more degrees of freedom than in the examples shown in the previous section. Finally, we conclude the paper with a discussion on the extension of the method to study transition pathways for other bio-molecular systems, such as the folding of proteins, and with some caveats on the current limitations of the method.

2 Theory

2.1 Derivation of the bridge equation

We assume that the system is driven by a force $F(x,t)$ and is subject to stochastic dynamics in the form of an overdamped Langevin equation.

For the sake of simplicity, we illustrate the method on a one-dimensional system, the generalization to higher dimensions or larger number of degrees of freedom being straightforward. We follow closely the presentation given in Ref. [39].

The overdamped Langevin equation reads

$$\frac{dx}{dt} = \frac{1}{\gamma}F(x(t),t) + \sqrt{\frac{2k_BT}{\gamma}}\eta(t) \tag{1}$$

where $x(t)$ is the position of the particle at time t, driven by the force $F(x,t)$, γ is the friction coefficient, related to the diffusion coefficient D through the Einstein relation $D = k_BT/\gamma$, where k_B is the Boltzmann constant and T the temperature of the heat bath. In addition, $\eta(t)$ is a Gaussian white noise with moments given by

$$\langle \eta(t) \rangle = 0 \tag{2}$$

$$\langle \eta(t)\eta(t') \rangle = \delta(t-t') \tag{3}$$

The probability distribution $P(x,t)$ for the particle to be at point x at time t satisfies a Fokker-Planck equation [54, 55],

$$\frac{\partial P}{\partial t} = D\frac{\partial}{\partial x}\left(\frac{\partial P}{\partial x} - \beta F P\right) \tag{4}$$

where $\beta = 1/k_BT$ is the inverse temperature. This equation is to be supplemented by the initial condition $P(x,0) = \delta(x-x_0)$, where the particle is assumed to be at x_0 at time $t = 0$. To emphasize this initial condition, we will often use the notation $P(x,t) = P(x,t|x_0,0)$.

We now study the probability over all paths starting at x_0 at time 0 and conditioned to end at a given point x_f at time t_f, to find the particle at point x at time $t \in [0,t_f]$. This probability can be written as

$$\mathcal{P}(x,t) = \frac{1}{P(x_f,t_f|x_0,0)}Q(x,t)P(x,t)$$

where we use the notation

$$P(x,t) = P(x,t|x_0,0)$$
$$Q(x,t) = P(x_f,t_f|x,t)$$

Indeed, the probability for a path starting from $(x_0,0)$ and ending at (x_f,t_f) to go through x at time t is the product of the probability $P(x,t|x_0,0)$ to start at $(x_0,0)$ and to end at (x,t) by the probability $P(x_f,t_f|x,t)$ to start at (x,t) and to end at (x_f,t_f).

The equation satisfied by P is the Fokker-Planck equation mentioned above (4), whereas that for Q is the so-called reverse or adjoint Fokker-Planck equation [54, 55] given by

$$\frac{\partial Q}{\partial t} = -D\frac{\partial^2 Q}{\partial x^2} - D\beta F\frac{\partial Q}{\partial x} \tag{5}$$

It can be easily checked that the conditional probability $\mathcal{P}(x,t)$ satisfies a new Fokker-Planck equation

$$\frac{\partial \mathcal{P}}{\partial t} = D\frac{\partial}{\partial x}\left(\frac{\partial \mathcal{P}}{\partial x} - \left(\beta F + 2\frac{\partial \ln Q}{\partial x}\right)\mathcal{P}\right)$$

Comparing this equation with the initial Fokker-Planck (4) and Langevin (1) equations, one sees that it can be obtained from a Langevin equation with an additional potential force

$$\frac{dx}{dt} = \frac{1}{\gamma}F + 2D\frac{\partial \ln Q}{\partial x} + \sqrt{\frac{2k_BT}{\gamma}}\eta(t) \tag{6}$$

This equation has been previously obtained using the Doob transform [56, 57] in the probability literature and provides a simple recipe to construct a *generalized bridge*. It generates Brownian paths, starting at $(x_0,0)$ conditioned to end at (x_f,t_f), with unbiased statistics. It is the additional term $2D\frac{\partial \ln Q}{\partial x}$ in the Langevin equation that guarantees that the trajectories starting at $(x_0,0)$ and ending at (x_f,t_f) are statistically unbiased. This equation can be easily generalized to any number of degrees of freedom.

In the following, we will specialize to the case where the force F is derived from a potential $U(x)$. The bridge equation becomes

$$\frac{dx}{dt} = -\frac{1}{\gamma}\frac{\partial U}{\partial x} + 2D\frac{\partial \ln Q}{\partial x} + \sqrt{\frac{2k_BT}{\gamma}}\eta(t) \tag{7}$$

In that case, the Fokker-Planck equation corresponding to this modified Langevin equation [58] can be recast into an imaginary time Schrödinger equation [39], and the probability distribution function P can be written as

$$Q(x,t) = P(x_f,t_f|x,t) = e^{-\beta(U(x_f)-U(x))/2}\langle x_f|e^{-H(t_f-t)}|x\rangle \tag{8}$$

where we used standard bra-ket Dirac notation for a matrix element $M_{ij} =< i|M|j >$, H is a "quantum Hamiltonian" defined by [54]

$$H = -D\frac{\partial^2}{\partial x^2} + D\frac{\beta^2}{4}V(x) \tag{9}$$

and the potential V by

$$V = \left(\frac{\partial U}{\partial x}\right)^2 - 2k_B T \frac{\partial^2 U}{\partial x^2} \tag{10}$$

We denote by M the matrix element of the Euclidian Schrödinger evolution operator

$$M(x,t) = \langle x_f | e^{-H(t_f - t)} | x \rangle \tag{11}$$

Using eq.(8) for Q, one can write equation (7) as

$$\frac{dx}{dt} = 2\frac{k_B T}{\gamma}\frac{\partial}{\partial x} \ln M(x,t) + \sqrt{\frac{2k_B T}{\gamma}}\eta(t) \tag{12}$$

We see on the above form that when $t \to t_f$, the matrix element $M(x,t)$ converges to $\delta(x_f - x)$, and it is this singular attractive potential which drives all the paths to x_f at time t_f.

2.2 Transition paths

The bridge equations (7) or (12) can be solved exactly in a certain number of cases [59]. However in general, for systems with many degrees of freedom, the functions $Q(x,t)$ or $M(x,t)$ cannot be computed exactly and one has to resort to some approximations. In the following, we will be mostly interested in problems of energy or entropy barrier crossing, which are of utmost importance in many chemical, biochemical, or biological reactions.

The matrix element $M(x,t)$ can be written as a Feynman path integral

$$M(x,t) = \int_{x(t)=x}^{x(t_f)=x_f} \mathcal{D}x \, \exp\left(-\int_t^{t_f} d\tau \left(\frac{1}{4D}\left(\frac{dx}{d\tau}\right)^2 + \frac{D\beta^2}{4}V(x(\tau))\right)\right) \tag{13}$$

The free case is defined as

$$\begin{aligned} M_0(x,t) &= P_0(x_f, t_f | x, t) \\ &= \int_{x(t)=x}^{x(t_f)=x_f} \mathcal{D}x \, \exp\left(-\int_t^{t_f} d\tau \left(\frac{1}{4D}\left(\frac{dx}{d\tau}\right)^2\right)\right) \\ &= \left(\frac{1}{4\pi D(t_f - t)}\right)^{1/2} e^{-\frac{\left(x_f - x\right)^2}{4D(t_f - t)}} \end{aligned} \tag{14}$$

where P_0 is the probability distribution for a free particle.

Equation (13) can be rewritten as

$$M(x,t) = M_0(x,t)\langle \exp\left(-\frac{D\beta^2}{4}\int_t^{t_f} d\tau V(x(\tau))\right)\rangle_0$$

where the expression $\langle ... \rangle_0$ denotes the expectation value with the Brownian measure P_0.

The convexity of the exponential function implies the Jensen inequality [60], which states that for any operator A and any probability measure, one has

$$\langle e^{-A} \rangle \geq e^{-\langle A \rangle} \tag{15}$$

Equality occurs when the probability is a $\delta-$function; it is thus a good approximation when the operator A has small fluctuations. Taking $A(t)$ to be

$$A(t) = \frac{D\beta^2}{4} \int_t^{t_f} d\tau V(x(\tau)) \tag{16}$$

we have

$$M(x,t) \simeq M_0(x,t) \exp\left(-\frac{D\beta^2}{4} \int_t^{t_f} d\tau \langle V(x(\tau)) \rangle_0 \right) \tag{17}$$

Using the expression

$$\langle V(x(\tau)) \rangle_0 = \frac{1}{M_0(x,t)} \int_{-\infty}^{+\infty} dz P_0(x_f, t_f | z, \tau) V(z) P_0(z, \tau | x, t),$$

after some calculations, we obtain

$$\langle V(x(\tau)) \rangle_0 = \left(\frac{\theta_1 + \theta_2}{4\pi D \theta_1 \theta_2} \right)^{1/2} \int_{-\infty}^{+\infty} dz \exp\left(-\frac{\theta_1 + \theta_2}{4D\theta_1\theta_2} \left(z - \frac{x_f \theta_2 + x\theta_1}{\theta_1 + \theta_2} \right)^2 \right) V(z)$$

where

$$\theta_1 = t_f - \tau, \ \theta_2 = \tau - t$$

After a change of variable, this expression becomes

$$\langle V(x(\tau)) \rangle_0 = \int_{-\infty}^{+\infty} \frac{dz}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2} \right) V\left(X + \sqrt{\frac{2D\theta_1\theta_2}{\theta_1 + \theta_2}} z \right)$$

where

$$X = \frac{x_f \theta_2 + x\theta_1}{\theta_1 + \theta_2}$$

and the constrained Langevin equation (12) becomes

$$\begin{aligned} \frac{dx}{dt} &= \frac{x_f - x}{t_f - t} \\ &- \frac{(D\beta)^2}{2} \int_t^{t_f} d\tau \left(\frac{t_f - \tau}{t_f - t} \right) \int_{-\infty}^{+\infty} \frac{dz}{(2\pi)^{1/2}} \exp\left(-\frac{z^2}{2} \right) \frac{\partial}{\partial X} V\left(X + \sqrt{\frac{2D(t_f - \tau)(\tau - t)}{(t_f - t)}} z \right) \\ &+ \sqrt{\frac{2k_B T}{\gamma}} \eta(t) \end{aligned} \tag{18}$$

or, after the change of variable $u = \frac{\tau - t}{t_f - t} = \frac{\theta_2}{\theta_1 + \theta_2}$,

$$\begin{aligned}
\frac{dx}{dt} &= \frac{x_f - x}{t_f - t} \\
&- \frac{(D\beta)^2}{2}(t_f - t)\int_0^1 du(1-u)\int_{-\infty}^{+\infty}\frac{dz}{(2\pi)^{1/2}}\exp\left(-\frac{z^2}{2}\right)\frac{\partial}{\partial X}V\left(X + \sqrt{2D(t_f - t)u(1-u)}z\right) \\
&+ \sqrt{\frac{2k_BT}{\gamma}}\eta(t)
\end{aligned} \tag{19}$$

where

$$X = x_f u + x(1-u) \tag{20}$$

Interestingly, integration by part with respect to z yields an equivalent form which does not require the cumbersome evaluation of $\partial V/\partial X$.

$$\begin{aligned}
\frac{dx}{dt} &= \frac{x_f - x}{t_f - t} \\
&- \left(\frac{D}{2}\right)^{3/2}\beta^2\sqrt{t_f - t}\int_0^1 du\sqrt{\frac{1-u}{u}}\int_{-\infty}^{+\infty}\frac{dz}{(2\pi)^{1/2}}\exp\left(-\frac{z^2}{2}\right)zV\left(X + \sqrt{2D(t_f - t)u(1-u)}z\right) \\
&+ \sqrt{\frac{2k_BT}{\gamma}}\eta(t)
\end{aligned} \tag{21}$$

Another change of variable $u = v^2$ allows to get rid of the singularity at $u = 0$.
The integration over the Gaussian variable z can be performed by numerical sampling

$$\int_{-\infty}^{+\infty}\frac{dz}{(2\pi)^{1/2}}\exp\left(-\frac{z^2}{2}\right)F(z) \simeq \frac{1}{N_G}\sum_i F(z_i) \tag{22}$$

where the N_G variables z_i are Gaussian variables (with zero average and unit variance).

However, as we have seen, the approximation (17) is valid if the exponent A does not fluctuate too much over the trajectories relevant to the transition. There are two cases when this approximation can be further simplified and *where the z-integral can be avoided*:

1. Low temperature

   In that case, since $D = k_BT/\gamma$, diffusion is small, thus V can be approximated as $(\frac{\partial U}{\partial x})^2$. In addition, the term $\sqrt{2D(t_f - t)u(1-u)}z$ in (19) is small compared to X and can be neglected. Equation (18) can be simplified to

$$\frac{dx}{dt} = \frac{x_f - x}{t_f - t} - \frac{(D\beta)^2}{2}(t_f - t)\int_0^1 du(1-u)\frac{\partial}{\partial X}\left(\frac{\partial}{\partial X}U(X)\right)^2 + \sqrt{\frac{2k_BT}{\gamma}}\eta(t) \tag{23}$$

2. Barrier crossing

   According to Kramers theory, the total transition time τ_K (waiting + crossing) scales like the exponential of the barrier height ΔE^* while it has been shown that the crossing time (Transition Path Time) τ_c scales like the logarithm of the barrier ΔE^* [61–63]. We have thus $\tau_c << \tau_K$.

   As discussed before, the barrier crossing time is very short compared to the Kramers time. Therefore the transition trajectories are very weakly diffusive, and are thus almost ballistic. Consequently, we have $\sqrt{2Dt_f} \ll |x_f - x_0|$ and again we can neglect the z term in V. Equation (18) becomes

$$\frac{dx}{dt} = \frac{x_f - x}{t_f - t} - \frac{(D\beta)^2}{2}(t_f - t)\int_0^1 du(1-u)\frac{\partial V(X)}{\partial X} + \sqrt{\frac{2k_BT}{\gamma}}\eta(t) \tag{24}$$

All the equations described above are easily generalized to any number of particles in any dimension, interacting with any many-body potentials. They are integro-differential stochastic Markov equations, as the variable X depends only on the stochastic variable $x(t)$. One can generate many independent trajectories by integrating these equations with different noise histories $\eta(t)$.
To test the validity of the main approximation (17), one should compute the variance $(\Delta A)^2 = \langle A^2\rangle - \langle A\rangle^2$ of the random variable $A(t)$ in eq.(16) over all the trajectories generated. Computing the correction to the Jensen inequality, it is easily seen that the approximation is reliable provided

$$R = \frac{(\Delta A(0))^2}{2|\langle A(0)\rangle|} << 1 \tag{25}$$

where we use the value of $A(t)$ at $t = 0$ to define the quality criterion R, because we have observed that this is where it is most indicative of the quality of the trajectory.

2.3 Simulation Time

For barrier crossing, what simulation time t_f should be used? This problem has been approached before for one-dimensional systems [64] but for systems with many degrees-of-freedom there is no theoretical answer to this question. Obviously, for any initial and final state x_i and x_f, there is a set of Langevin trajectories which make the transition in any given time t_f. If the time t_f is very short compared to the typical time scales of large motions of the system, there is a small number of such trajectories, since they require a very specific noise history. As a result, the approximations presented above are reliable and the factor R is much smaller than 1. However, this is not a very interesting regime, as trajectories are driven by the boundary conditions. If we are interested in simulating transition paths, the time t_f should obviously be larger than the typical TPT τ_c. Indeed, if t_f is smaller than τ_c, paths are driven by the final state. On the other hand, if t_f is too large, then we will also simulate part of the waiting time in the wells, where fluctuations are large (except maybe at low temperature). Therefore, in order to simulate transition paths as accurately as possible, one should use a simulation time t_f larger than the typical TPT τ_c, but not much larger. In the following we use the notation $t_f = N_{step}dt$.

3 Results

We now illustrate these concepts on three examples: the Mueller potential, the so-called Mexican hat potential, and then on Elastic Network Models (ENM) of proteins whose end conformations are known.

3.1 The Mueller potential

The Mueller potential is a standard benchmark potential to check the validity of methods for generating transition paths. It is a two dimensional potential given by

$$U(x,y) = \sum_{i=1}^{4} A_i \exp\left(a_i(x-x_i^0)^2 + b_i(x-x_i^0)(y-y_i^0) + c_i(y-y_i^0)^2\right) \tag{26}$$

with

$$A = (-200, -100, -170, 15) \quad a = (-1, -1, -6.5, 0.7) \quad b = (0, 0, 11, 0.6) \tag{27}$$

This potential has 3 local minima denoted by A,B,C, separated by two barriers (Fig.1). The effective potential $V(x,y)$ can be calculated analytically, as well as its gradient. Equations (18), (23) and (24) can easily be solved numerically. We display only the trajectories generated by (24). The simulation time t_f is chosen so that we observe a small waiting time around the initial as well as the final point, namely $t_f = 0.15$. We use 50 points for the integration over u. We display a sample of 500 trajectories obtained from eq. (24) with $t_f = 0.15,\ dt = 10^{-4},\ D = 1$ at temperature $T = 5$. We can compute the average trajectory as well as its variance. These trajectories are displayed on Fig. 1, where we plot the AB, BC and AC trajectories.

To assess the quality of the approximations, we check the criterion (25). For the trajectories AB, we obtain $R \approx 5.310^{-2}$, $R \approx 0.68$ for BC and $R \approx 1.13$ for AC. Therefore, the approximation is quite reliable for the AB trajectories, but less for the others. In fact, it is instructive to study the accuracy of the method when varying t_f. For that matter, in Fig. 2, we plot the factor R as a function of t_f, for the AB transition. We see that for both small and large t_f, the factor R is small, with a maximum at $t_f \approx 0.05$. For small t_f, the trajectories fluctuate around the straight line trajectory joining A to B through high barriers (see Fig. 1A). For large t_f, the trajectories fluctuate around the potential energy valley joining A to B. As t_f increases from small values, the ensemble of trajectories include trajectories going through the high barrier and through the valley, and at $t_f \approx 0.05$, there is a strong mixing of both types of trajectories, giving rise to a large value of R. When t_f increases further, the trajectories going through the barrier disappear from the ensemble, and only valley trajectories remain, yielding a decrease of R.

3.2 The Mexican hat potential

The potential of the Mexican hat is given by

$$U(x,y) = \frac{1}{4}(x^2 + y^2 - 1)^2 \tag{28}$$

and has therefore a circle of minima for $x^2+y^2=1$ with $U=0$ and a maximum at $(0,0)$ with energy $U=1/4$. Again we solve equation (24), using 50 points for the u integral. Given the small barrier of this potential $\Delta U=1/4$, we go to low temperature. On Fig.3A, we plot 100 trajectories, generated at temperature $T=0.1$, starting at $(-1,0)$ and all ending at $(1,0)$. The total time is $t_f=7$ and the time step is $dt=10^{-4}$. The quality criterion (25) gives $R=0.345$. The trajectories divide into three dominant groups, those that take a northern route (30), those that take a southern route (40) along the circle of minima, and those that go directly through the energy barrier (30). The distribution into those three groups was decided based on the mean value Y_{mean} for the y coordinates along the trajectories. If we take a longer duration, the fraction of trajectories that go through the central barrier decreases. For example, for $t_f=10$, there are only 9 of those trajectories, as seen on Fig. 3B. The quality criterion is then $R=0.266$.

3.3 The Mixed ENM Energy Model for Proteins

We now use the CLD method to explore conformational transitions in proteins using the"Mixed ENM" energy model.

3.3.1 The Energy Model

The energy function is the combination of the mixed elastic model and a collision term

$$U_{tot}=U_{Mix-ENM}+U_{collision} \tag{29}$$

where

$$U_{Mix-ENM}=-\frac{1}{\beta_m}log(e^{-\beta_m U_A}+e^{-\beta_m U_B}) \tag{30}$$

where U_A is the ENM Energy centered on conformation A (initial) and U_B the ENM Energy centered on conformation B (final), as defined originally by Tirion [50], and β_m is the inverse of the mixing Temperature T_m [13].

$$U_A=\sum_{ij}k_{ij}C_{ij}(d_{ij}-d_{ij}^A)^2 \tag{31}$$

$$U_B=\sum_{ij}k_{ij}C_{ij}(d_{ij}-d_{ij}^B)^2+\Delta U \tag{32}$$

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 is the energy difference between the two states and d_{ij}^A and d_ij^B their interatomic distances at rest in conformations A and B, respectively.
The collision energy term, taken as the repulsive part of a Van der Waals potential, reads

$$U_{collision}=\epsilon\sum_{i,j}(\frac{\sigma}{d_{ij}})^{12}=\sum_{i,j}U_{ij} \tag{33}$$

with $\sigma = 2.5$ Angstroms and $\epsilon = 1.0$ kcal/Mole.
One new feature of the U_{ENM} model is that the elastic constants k_{ij} are modulated by the difference in the resting d_{ij} distances of the two states.

$$k_{ij} = \min(\frac{\epsilon_k}{(d_{ij}^A - d_{ij}^B)^2}, k_{max}) \tag{34}$$

In Appendix A we explain this choice of the k_{ij} that essentially conserves the height of the energy barriers for pairs of atoms separated by different distances.
Here we take, as recommended [53], $\epsilon_k = 0.5$ kcal/Mole and $k_{max} = 0.2$ kcal/Mol/Angstrom^2.
The mixed ENM itself is characterized by $R_c = 11.5$ Angstroms and $\Delta U = 0$.

The generalisation of V to multi-dimensional problems is straightforward. For instance,

$$V_{ENM} = \sum_{i=1,N} \sum_{\alpha=1,3} (\nabla_{i,\alpha} U_{ENM})^2 - 2k_B T \sum_{i=1,N} \Delta_i U_{ENM} \tag{35}$$

and

$$V_{collision} = \sum_{i=1,N} \sum_{\alpha=1,3} (\nabla_{i,\alpha} U_{collision})^2 - 2k_B T \sum_{i=1,N} \Delta_i U_{collision} \tag{36}$$

If $U_{tot} = U_{ENM} + U_{collision}$ then there is a cross-term in the V_{tot} term:

$$V_{tot} = V_{ENM} + V_{collision} + 2 \sum_{i,\alpha} \nabla_{i,\alpha} U_{ENM} \nabla_{i,\alpha} U_{collision} = V_{ENM} + V_{collision} + 2V_{cross} \tag{37}$$

All the algebra and derivatives needed to implement this Energy in this method are described in an Appendix in [65]. We also implemented the following Mixing potential, with very similar results.

$$U_{Mix-ENM} = \frac{U_A + U_B - \sqrt{(U_A - U_B)^2 + 4\epsilon^2}}{2} \tag{38}$$

3.3.2 PDB coordinates and parameters

For Adenylate Kinase we used CA-coordinates from files 1AKE.pdb and 4AKE.pdb (213 CA atoms) for the initial and final states of the simulation, respectively. The RMSD between the two forms is 7.2 Angstroms.
For the mixed-ENM we used the weighting scheme described in Eq. 30, $R_c = 11.5\mathring{A}$, and a mixing Temperature $T_m = 1500T$, where T is taken as either $T = 5$ or $T = 10$.
For the Langevin dynamics we took in all cases $dt = 0.001$ and $\gamma = 1$. Hence $dt/\gamma = 1E-3$.
We stress that the approximation used in the new method makes it valid mainly at low temperature.

3.3.3 Comparison of the different variants of the method

In Figure 4 we compare the results obtained for Adenylate Kinase with different versions of the program that use either
i) V (Eq. 21), with $N_G = 100$ points for the Gaussian integral, or
ii) ∇V (Eq. 19) with either $N_G = 1$ (i.e. $z = 0$) or $N_G = 100$ integration points for the Gaussian integral.
In both cases we used $N_U = 50$ points for the evaluation of the 1-D integral in u, using Simpson rule. We see that the results are almost super-imposable with both $N_G = 1$ or $N_G = 100$ if one uses ∇V, confirming the validity of Eq. 23, or using V as in Eq. 21.
We also show the result of an older method (without any 1-D integral) described earlier by one of us [39] and based on a simpler approximation:

$$\frac{d\vec{x}}{dt} = \frac{\vec{x_f} - \vec{x}}{t_f - t} - \frac{1}{2\gamma^2}(t_f - t)\vec{\nabla}V(\vec{x}) + \vec{\eta}(t) \tag{39}$$

where the "driving force" (second term in the r.h.s) results from a more drastic approximation and we see that the energy barrier is higher than in the method presented here.

3.3.4 Scanning the length of the simulation

We next simulate transitions for different lengths of the total transition time. We used $N_{step} = 500, 1000, 2500, 5000, 7500, 10000$. We see different activation energies but the system reaches a plateau about $N_{step} = 5000$. Clearly we see in Figure 5 that for $N_{step} = 7500$ and $N_{step} = 10000$ the system spends more time in the two basins before beginning the transition, as predicted, but without changing the height of the Energy barrier.

3.3.5 Evaluation of the trajectory

We have calculated the criterion R mentioned above for 100 Adenylate Kinase trajectories of different t_f lengths ($N_{step} = 1000, 2500, 5000, 7500, 10000$), at $T = 5, T_m = 1500T$. In all cases we find $R = 0.00456 \pm 0.00002$. This justifies the approximation made to obtain the equations described in the section "Methods".

To further understand the nature of the conformation transition simulated in our trajectories, we monitored $RMSD_1$ and $RMSD_2$ during the course if the simulation, where these quantities refer to the RMSD of the current model with the first form and the second form, respectively. If one calls $RMSD_{1-2}$ the total RMSD between the two extreme forms one can define an order parameter RC [13]:

$$RC(t) = \frac{1}{2}(1 + \frac{RMSD_1(t) - RMSD_2(t)}{RMSD_{1-2}}) \tag{40}$$

We found that the increase of the order parameter RC from 0 to 1 is linear as a function of time, indicating that the trajectories are essentially ballistic.
However, we can see that the trajectories are different from purely linear ones in that i) they are self-avoiding and ii) their Q1-vs-Q2 plots are not linear, as shown in Figure 6A (Q1 is the

Name	Length (aa)	PDB codes	RMSD (Å)	CPU (s)
Ntr	124	1DC7 1DC8	4.5	127
AK	213	1AKE 4AKE	7.2	175
RBP	272	1URP 2DRI	6.2	265
LeuT	492	3TT1 3TT3	3.6	484
5-NT	526	1OID 1HPU	9.3	519

Table 1: CPU time needed to calculate the CLD trajectory for proteins of various lengths

percentage of native contacts in the initial form and Q2 the percentage of native contacts in the final form; contacts are counted for $d_{ij} < R_c$). In Figure 6B we show the same trajectory, but with different realisations of the noise history.

3.3.6 Efficiency of the algorithm

We have tested the program on a number of test cases for proteins of different lengths: Ribonuclease III (RNase), Ca^{++} ATPase (ATPase), Ribose Binding Protein (RBP) and 5' Nucleotidase were studied in [11]; Nitrogen Regulatory protein C (Ntr) was studied in [53], while Adenylate Kinase (AK), Leucine Transporter (LeuT) and also Ca^{++} ATPase were studied in [51].

In Table I, we report the CPU time needed to complete the calculations in the following conditions $N_U = 50$, $N_G = 1$, $N_{step} = 5000$, $T = 5$ and $T_m = 1500T$, the PDB codes of the initial and final conformations, as well as the cRMSD between them, calculated on the C-alpha coordinates. The program was compiled with gfortran and executed on a single CPU of a Linux workstation (Intel Xeon CPU E5-1650v3 at 3.50GHz).

We see than we can generate hundreds of trajectories on a single CPU overnight. We note that recompiling the program using the Intel Fortran compiler further reduces the computing time, by at least a factor of 2.

3.3.7 Comparison with experiments and with other methods

In the test cases studied in [11], there is a known intermediate structure. In Figure 7 we show the RMSD of this structure with all the intermediate conformations along the Conditioned Langevin Dynamics Path, for 5'-nucleotidase (5'-NT) and for Ribose Binding Protein (RBP). The PDB codes for 5'-NT are 1OID, 1OI8 and 1HPU for the initial, putative intermediate and final conformations, respectively. For RBP they are 1URP, 1BA2 and 2DRI.
One clearly sees (red and green) that intermediate conformations along the generated transition path get reasonably close to the proposed experimental intermediate. The blue line shows the average of the CA-CA consecutive distances for all intermediate conformations, which remain close to 3.8 $\mathring{A}$, indicating that the models are not distorted in the trajectory that is generated. We note that the new method (RelaxPath) of one of us [65], that refines the trajectory obtained by MinActionPath [32] using a potential similar to the one described here (including the $V_{collision}$ term) lead to very similar results, as illustrated in Figure 8.
The trajectories generated by MinActionPath, RelaxPath, and CLD are found to be very similar for AK, RBP, and 5'-NT with some minor differences. First, we notice that the CLD

and RelaxPath trajectories are the most similar ones, on all three test cases. The trajectories generated with MinActionPath are very similar to the RelaxPath and CLD trajectories near their end points, but start showing differences near the transition states. This is especially true for 5'-NT. These differences are likely due to the differences in the potentials considered by the three methods. MinActionPath only considers a superposition of the elastic potentials but does not "mix" them, while in both RelaxPath and CLD a collision term is added. We have observed that the absence of this collision term can lead to distortion of the structures along the MinActionPath trajectories [65].

3.3.8 Current limitations of the method

In the case of ATPase (994 residues, PDB codes 1SU4 and 1T5S) and RNase (437 residues, PDB codes 1YY0 and 1YYW), also tested in [11], we could not find a window of the parameters T, T_m, t_f to generate trajectories. We suspect that this is due to a very large RMSD between the two forms (16.1$\mathring{A}$ and 13.4$\mathring{A}$, respectively). In general, we have observed that tuning these parameters is very system-dependent. Indeed, we found that it is not always possible to clearly apprehend the energy landscape generated by the "Mixed ENM" model, which depends crucially on k_1, k_2, T, T_m. In most cases, it is a rather smooth landscape and the transition can be found readily with the Conditioned Langevin Dynamics Equation. In some cases, however, we could not find the right combination of the parameters of the model to make the energy landscape smooth enough.

In this respect it may be worth mentioning the large differences in T_m that were applied in the two cases studied in [13], which also points to some necessary tuning of the model parameters with this kind of "Mixed-ENM" models.

Another limitation is that the energy is purely geometric, apart from the van der Waals repulsion term. However, there are other energy terms, such as the Go-like Energy Model [66], that could be added to the current version of the program. This should be especially useful for studying the folding transition, not just conformational transitions.

Finally, our current implementation mainly use coarse-grained models and this might appear as a severe limitation of the method. However, we have developed methods that allow to reconstruct quickly all atoms from a CA-trace representation of a protein (P. Koehl, unpublished). An immediate application of the method would then be to generate quickly initial trajectories to be refined by more sophisticated methods such as the (all atoms) string method, that does need as an input an initial guess of the trajectory.

4 Conclusion and Perspective

In summary, we show here a novel method to generate ab initio transition path trajectories using a formalism called Conditioned Langevin dynamics. The most crucial parameter of the theory is the total length of the simulation, which must be carefully chosen so as to allow short waiting times around both the initial and final states. We define a quality criterion R

that can be calculated *a posteriori* to see if the approximation made to solve the underlying PDE is justified or not. This criterion is necessary but not sufficient. We have tested the method on simple one- or two-dimensional potentials, as well as in more biologically relevant situations, such as conformational transitions in proteins, with simple coarse-grained potentials.

Compared to our earlier work [32] we have succeeded in improving the generation of trajectories between two known forms of the same macromolecule in two ways: i) first we impose self-avoidance during the trajectory and ii) we generate hundreds of plausible trajectories in a matter of hours on a single CPU desktop workstation.

Proteins usually continuously visit one of several states (e.g. native or denatured, open or closed, apo or holo in the allosteric picture), by making stochastic transitions between them. The picture which emerges is that the system is staying for a long time in one of the minima and then making stochastically rapid transitions to the other minimum. It follows that for most of the time, the system performs harmonic oscillations in one of the wells, which can be described by normal mode analysis. Rarely, there is a very short but interesting physical phenomenon, where the system makes a fast transition between minima. This picture has been confirmed by single molecule experiments, where the waiting time in one state can be measured, although the time for crossing is so short that it cannot be resolved [67]. Note however that the whole transition path trajectory distribution has been measured in the case of DNA hairpin formation [68]. This scenario has also been confirmed recently by very long millisecond molecular dynamics simulations which for the first time show spontaneous thermal folding-unfolding events [69].
Our next goal is to try to simulate by this method the folding transition of proteins and compare it with these folding-unfolding events. To this end, we will implement a more realistic energy potential (Go-model) and also use many different initial states using randomly generated models of unfolded states. If the results compare well, we might be in a position to simulate folding transitions on much larger proteins that can currently be handled with the most powerful available MD methods [70].

5 Acknowledgements

MD acknowledges financial support from a grant of the Agence Nationale de la Recherche (ANR) through the program 10-BINF-03-11. PK acknowledges support from the Ministry of Education from Singapore through grant Number: MOE2012-T3-1-008. We thank Frederic Poitevin for helpful discussions at initial stages of the project.

A Appendix A

When two harmonic potentials with the same curvature $k_{left} = k_{right}$ are mixed the height of the activation barrier varies a lot when the minima are separated by different distances, if they have a common curvature $k_{ij} = k_0$ (Left panel).

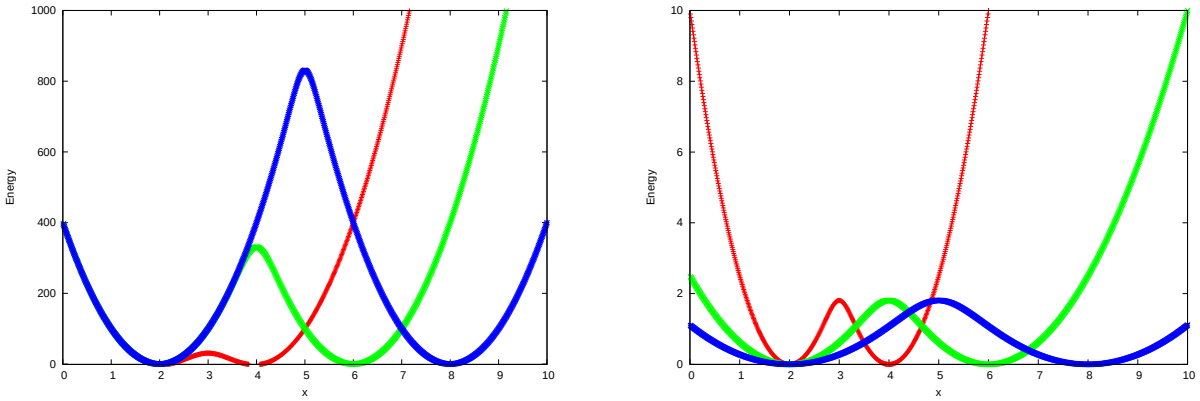

Figure 1: Left: Elastic potential of pairs of atoms with different distances and a constant elastic constant. Right: Modulated elastic constant to conserve the same energy barriers for pairs of atoms separated by different distances

Hence the idea of the modulation of k_{ij} as a function of the distance between the minima, as in Eq. (34) in the main text. Indeed, if we use a modulated elastic constant of the form $k_{ij} = k_0(\frac{1}{d_0^A - d_0^B})^2$ we get the desired result, an energy barrier independent of the position of the two minima, at the expense of different k_{ij} (Right panel).
Also, care must be taken in the case where $\|d_0^A - d_0^B\|$ becomes very small (i.e. less than 2 A), because in this case the stiffness of the elastic potential becomes unreasonably high and one gets stuck into numerical problems during derivative's evaluations.
Therefore, one needs to take a threshold value for the elastic constant, as in the mixed ENM model presented in the main text. At high temperature, the minima of the mixed-ENM are no longer exactly at the two extreme conformations, but they are slightly displaced.

References

[1] D. Zuckerman, *Statistical Physics of Biomolecules: an Introduction*, CRC Press, Boca Raton, Florida, USA, 2010.

[2] E. Laine, C. Goncalves, J. Karst, A. Lenard, A. Rault, W. Tang, T. Malliavin, D. Ladant, and A. Blondel, *Proc. Natl. Acad. Sci. (USA)* **107**, 11277 (2010).

[3] E. Laine, D. Martinze, D. Latant, T. Malliavin, and A. Blondel, *Toxins* **4**, 580 (2012).

[4] H. Eyring, *Chem. Rev.* **17**, 65 (1935).

[5] E. Wigner, *Trans. Faraday Soc.* **34**, 29 (1938).

[6] S. Fischer, K. Olsen, K. Nam, and M. Karplus, *Proc. Natl. Acad. Sci. (USA)* **108**, 5608 (2011).

[7] W. E and E. Vanden-Eijnden, *Annu Rev Phys Chem* **61**, 391 (2010).

[8] W. E and E. Vanden-Eijnden, *J. Stat. Phys.* **123**, 503 (2006).

[9] E. Vanden-Eijnden, *Adv. Exp. Med. Biol.* **797**, 91 (2014).

[10] M. Kim, R. Jernigan, and G. Chirikjian, *Biophys. J.* **83**, 1620 (2002).

[11] D. R. Weiss and M. Levitt, *J. Mol. Biol.* **385**, 665 (2009).

[12] P. Maragakis and M. Karplus, *J. Mol. Biol.* **352**, 807 (2005).

[13] W. Zheng, B. Brooks, and G. Hummer, *Proteins: Struct. Func. Bioinfo.* **69**, 43 (2007).

[14] M. Tekpinar and W. Zheng, *Proteins: Struct. Func. Bioinfo.* **78**, 2469 (2010).

[15] F. Pinski and A. Stuart, *Journal of Chemical Physics* **132**, 184104 (2010).

[16] H. 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, chapter 16, pp. 385–404, World Scientific, Singapore, 1998.

[17] G. Henkelman, B. Uberuaga, and H. Jonsson, *J. Chem. Phys.* **113**, 9901 (2000).

[18] D. Sheppard, R. Terrell, and G. Henkelman, *J. Chem. Phys.* **128**, 134106 (2008).

[19] W. E, W. Ren, and E. Vanden-Eijnden, *Phys. Rev. B* **66**, 052301 (2002).

[20] W. Ren, E. Vanden-Eijnden, P. Maragakis, and W. E, *J. Chem. Phys.* **123**, 134109 (2005).

[21] W. E, W. Ren, and E. Vanden-Eijnden, *J. Chem. Phys.* **126**, 164103 (2007).

[22] E. Vanden-Eijnden and M. Venturoli, *J. Chem. Phys.* **130**, 194103 (2009).

[23] W. Ren and E. Vanden-Eijnden, *J. Chem. Phys.* **138**, 134105 (2013).

[24] L. Maragliano, B. Roux, and E. Vanden-Eijnden, *J. Chem. Theory Comput.* **10**, 524 (2014).

[25] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, *J. Chem. Phys.* **125**, 24106 (2006).

[26] A. Pan, D. Sezer, and B. Roux, *J. Phys. Chem. B.* **112**, 3432 (2008).

[27] Y. Matsunaga, H. Fujisaki, T. Terada, T. Furuta, K. Moritsugu, and A. Kidera, *PLoS Comput. Biol.* **8**, e1002555 (2012).

[28] D. Branduardi and J. D. Faraldo-Gomez, *J. Chem. Theory Comput.* **9**, 4140 (2013).

[29] D. Dürr and A. Bach, *Commun. Math. Phys.* **60**, 153 (1978).

[30] R. Olender and R. Elber, *J. Chem. Phys.* **105**, 9299 (1996).

[31] P. Eastman, N. Gronbech-Jensen, and S. Doniach, *J. Chem. Phys.* **114**, 3823 (2001).

[32] J. Franklin, P. Koehl, S. Doniach, and M. Delarue, *Nucl. Acids. Res.* **35**, W477 (2007).

[33] P. Faccioli, M. Sega, F. Pederiva, and H. Orland, *Phys. Rev. Lett.* **97**, 108101 (2006).

[34] E. Vanden-Eijnden and M. Heymann, *J. Chem. Phys.* **128**, 061103 (2008).

[35] X. Zhou, W. Ren, and W. E, *J. Chem. Phys.* **128**, 104111 (2008).

[36] S. Chandrasekaran, J. Dhas, N. Dokholyan, and C. Carter Jr, *Struct. Dyn.* **3**, 012101 (2016).

[37] A. Stuart, P. Wiberg, and J. Voss, *Commun .Math. Sci* **2**, 685 (2004).

[38] M. Hairer, A. Stuart, and J. Voss, *Ann. Appli. Proba* **17**, 1657 (2007).

[39] H. Orland, *J. Chem. Phys.* **134**, 174114 (2011).

[40] J. Mattingly, N. Pillai, and A. Stuart, *Ann. Appli. Proba* **22**, 881 (2012).

[41] A. Faradjian and R. Elber, *J. Chem. Phys.* **120**, 10880 (2004).

[42] J. Bello-Rivas and R. Elber, *J. Chem. Phys.* **142**, 094102 (2015).

[43] L. Chong, A. Saglam, and D. Zuckerman, *Curr. Opin. Struct. Biol.* **43**, 88 (2017).

[44] L. Pratt, *J. Chem. Phys.* **85**, 5045 (1986).

[45] P. Bolhuis, C. Dellago, P. L. Geissler, and D. Chandler, *Annu. Rev. Phys. Chem.* **53**, 291 (2002).

[46] J. Chodera, N. Singhal, V. Pande, K. Dill, and W. Swope, *J. Chem. Phys.* **126**, 155101 (2007).

[47] G. Bowman, K. Beauchamp, G. Boxer, and V. Pande, *J. Chem. Phys.* , 124101 (2009).

[48] V. Pande, K. Beauchamp, and G. Bowman, *Methods* **52**, 99 (2010).

[49] M. Harrigan, M. Sultan, C. Hernandez, B. Husic, P. Eastman, C. Schwantes, K. Beauchamp, R. McGibbon, and V. Pande, *Biophys. J.* **112**, 10 (2017).

[50] M. Tirion, *Phys. Rev. Lett.* **77**, 1905 (1996).

[51] A. Das, M. Gur, M. H. Cheng, S. Jo, I. Bahar, and B. Roux, *PLoS Comput. Biol.* **10**, e1003521 (2014).

[52] R. Best, Y. Chen, and G. Hummer, *Structure* **12**, 1755 (2005).

[53] J. Adelman and M. Grabe, *J. Chem. Phys.* **138**, 044105 (2013).

[54] N. van Kampen, *Stochastic Processes in Physics and Chemistry*, North–Holland, Amsterdam, The Netherlands, 2007.

[55] R. Zwanzig, *Nonequilibrium Statistical Mechanics*, Oxford University Press, Oxford, United Kingdom, 2001.

[56] J. Doob, *Bull. Soc. Math. France* **85**, 431 (1957).

[57] P. Fitzsimmons, J. Pitman, and M. Yor, Markovian bridges: construction, Palm interpretation, and splicing, in *Seminar on Stochastic Processes, 1992*, Birkhaeuser, Boston, MA, USA, 1992.

[58] P. Chaikin and T. Lubentsky, *Principles of Condensed Matter Physics*, Cambridge University Press, Cambridge, United Kingdom, 2000.

[59] S. Majumdar and H. Orland, *J. Stat. Mech. Theor. Exp.* **2015**, P06039 (2015).

[60] J. Jensen, *Acta Math.* **30**, 175 (1906).

[61] I. Gopich and A. Szabo, *J. Chem. Phys.* **124**, 154712 (2006).

[62] W. Kim and R. Netz, *J. Chem. Phys.* **143**, 224108 (2015).

[63] E. Carlon and H. Orland, *submitted* .

[64] B. Zhang, D. Jasnow, and D. Zuckerman, *J. Chem. Phys.* **126**, 074504 (2007).

[65] P. Koehl, *J. Chem. Phys.* **145**, 184111 (2016).

[66] H. Taketomi, Y. Ueda, and N. Go, *Int. J. Pept. Prot. Res.* **7**, 445 (1975).

[67] H. Chung, K. McHale, J. Louis, and W. Eaton, *Science* **335**, 981 (2012).

[68] K. Neupane, D. Forster, D. Dee, H. Yu, F. Wang, and M. Woodside, *Science* **352**, 239 (2016).

[69] K. Lindorff-Larsen, S. Piana, R. Dror, and D. Shaw, *Science* **334**, 517 (2011).

[70] S. Piana, K. Lindorff-Larsen, and D. Shaw, *Proc. Natl. Acad. Sci. (USA)* **110**, 5915 (2013).

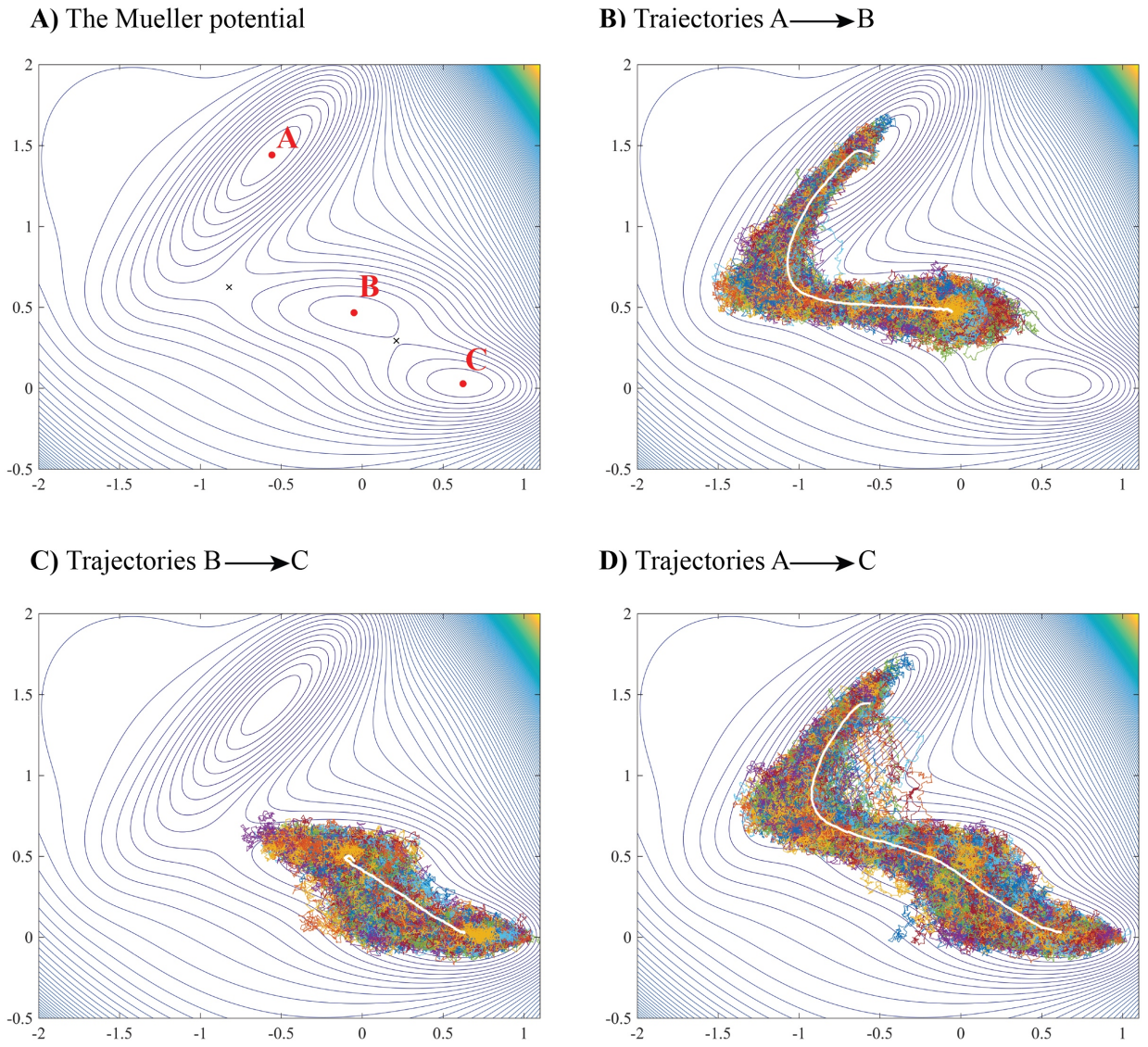

Figure 1: **Conditioned Langevin dynamics (CLD) trajectories on the Mueller potential**. (A) Contour plot of the Mueller potential, with the three minima labeled A, B, and C, and the two saddle points between those minima indicated with an x. 500 converged trajectories between the minima A and B (B), B and C (C), and A and C (D). The unweighted mean trajectories are shown in white.

.

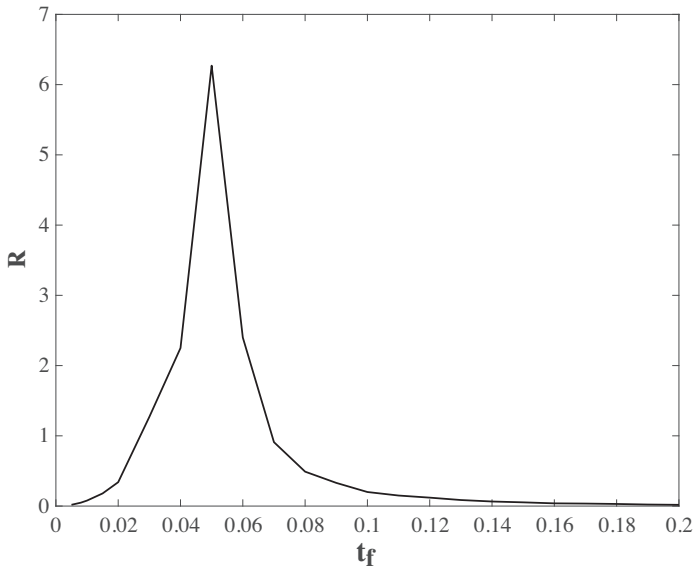

Figure 2: **Evaluation of the CLD trajectories for the Mueller potential**. The quality factor R plotted as a function of the total duration of the transition t_f, for the AB trajectories in the Mueller potential.

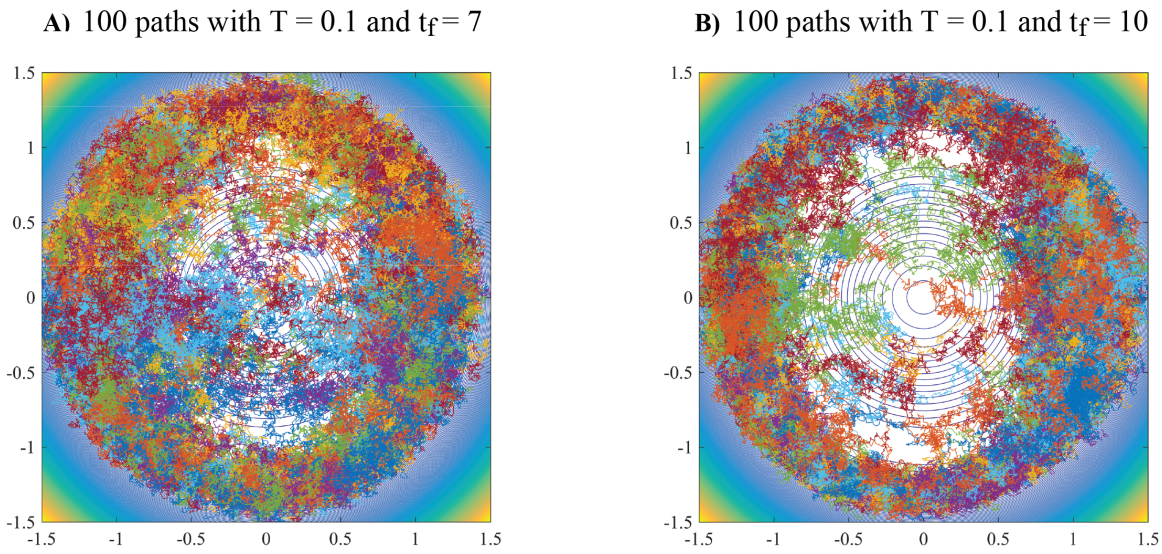

Figure 3: **Langevin bridge trajectories on the Mexican hat potential**. (A) 100 converged trajectories, all starting at $(-1, 0)$ and ending at $(1, 0)$, and generated at temperature $T = 0.1$ with a duration $t_f = 7$. Note that with this short transition time, many trajectories go through the barrier region. (B) Same as in (A), but with t_f now set to 10. Most of the trajectories now follow the circle of minima; those trajectories are nearly equally divided into two groups, those that follow the upper side of the circle (44), and those that follow the lower side (47) (see text for details).

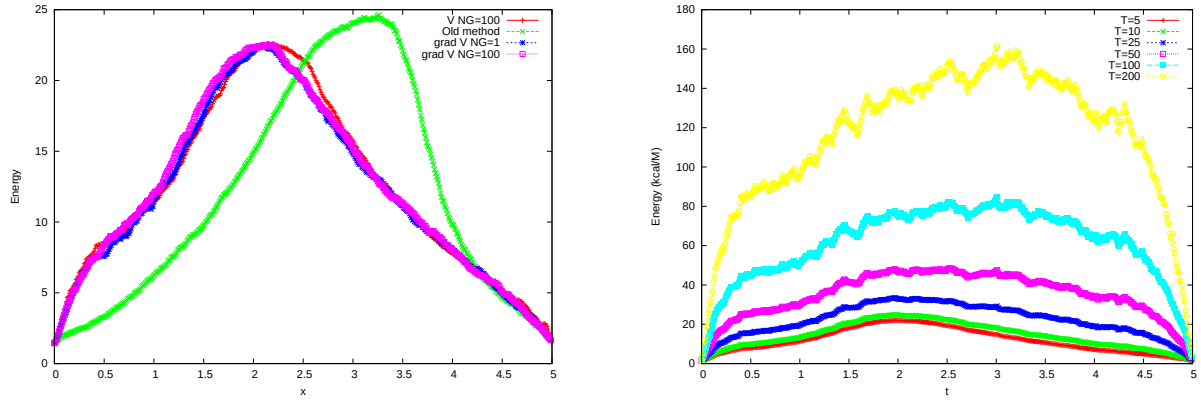

Figure 4: **Adenylate kinase trajectories.** (Left) Different versions of the CLD method were tested: without derivatives of V and $N_G = 100$ (red) for the evaluation of the gaussian integral or with the derivatives of V, with $N_G = 1$, i.e. $z = 0$, (blue) or $N_G = 100$ (magenta). A comparison with an older version of the method [39] is also shown in green. (Right) Runs with different temperatures ranging from $T = 5$ to $T = 200$ using the derivative of V and $N_G = 1$.

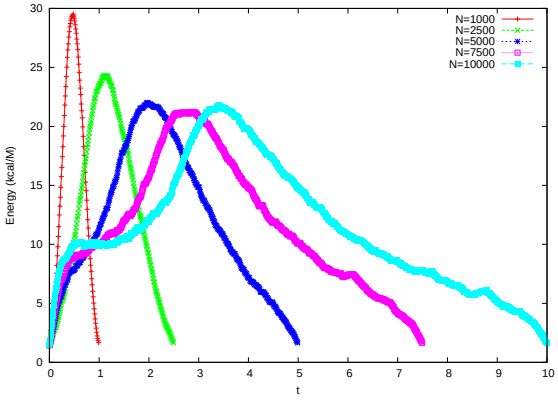

Figure 5: **Influence of the length of the simulation $t_f = N_{step}dt$ on the energy profile for Adenylate kinase**. All trajectories were generated with $T = 5$, $T_m = 1500T$, $N_u = 50$ and $N_G = 1$, for $N_{step} = 1000$ to 10000 and $dt = 0.001$.

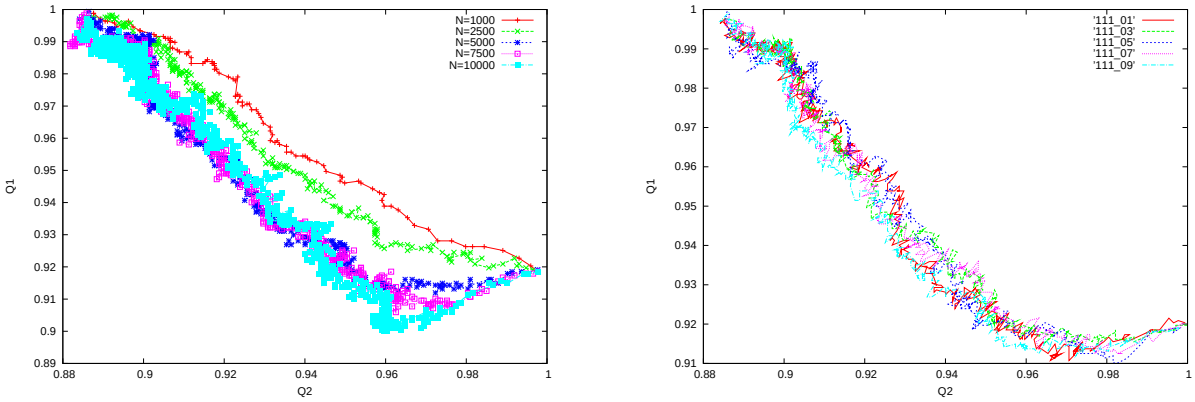

Figure 6: **Evaluation of the (non-) linearity of Adenylate kinase CLD trajectory**. (Left) Percentage of native contacts for the initial form (Q_1) as a function of the percentage of native contacts for the final form (Q_2) for $N_{step} = 1000$ to 10000. (Right) Different realisations of the adenylate kinase trajectory for different noise histories.

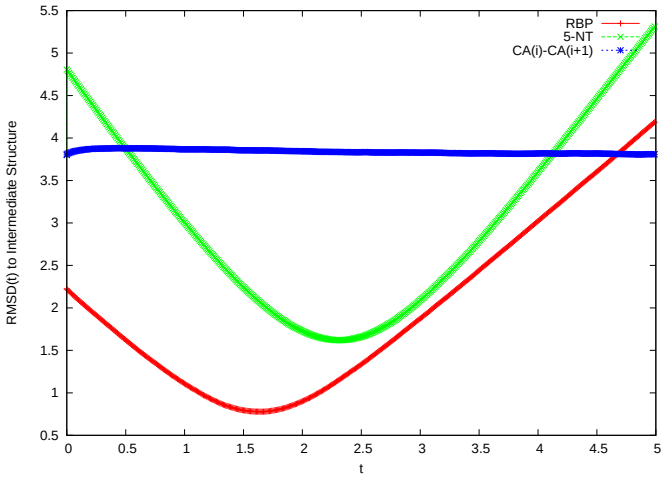

Figure 7: **Distance of a putative intermediate to successive snapshots of the trajectory**. (Green) 5'-nucleotidase (5-NT), (Red) Ribose Binding Protein (RBP), (Blue) The mean distance between successive CA atoms (in Angstroms).

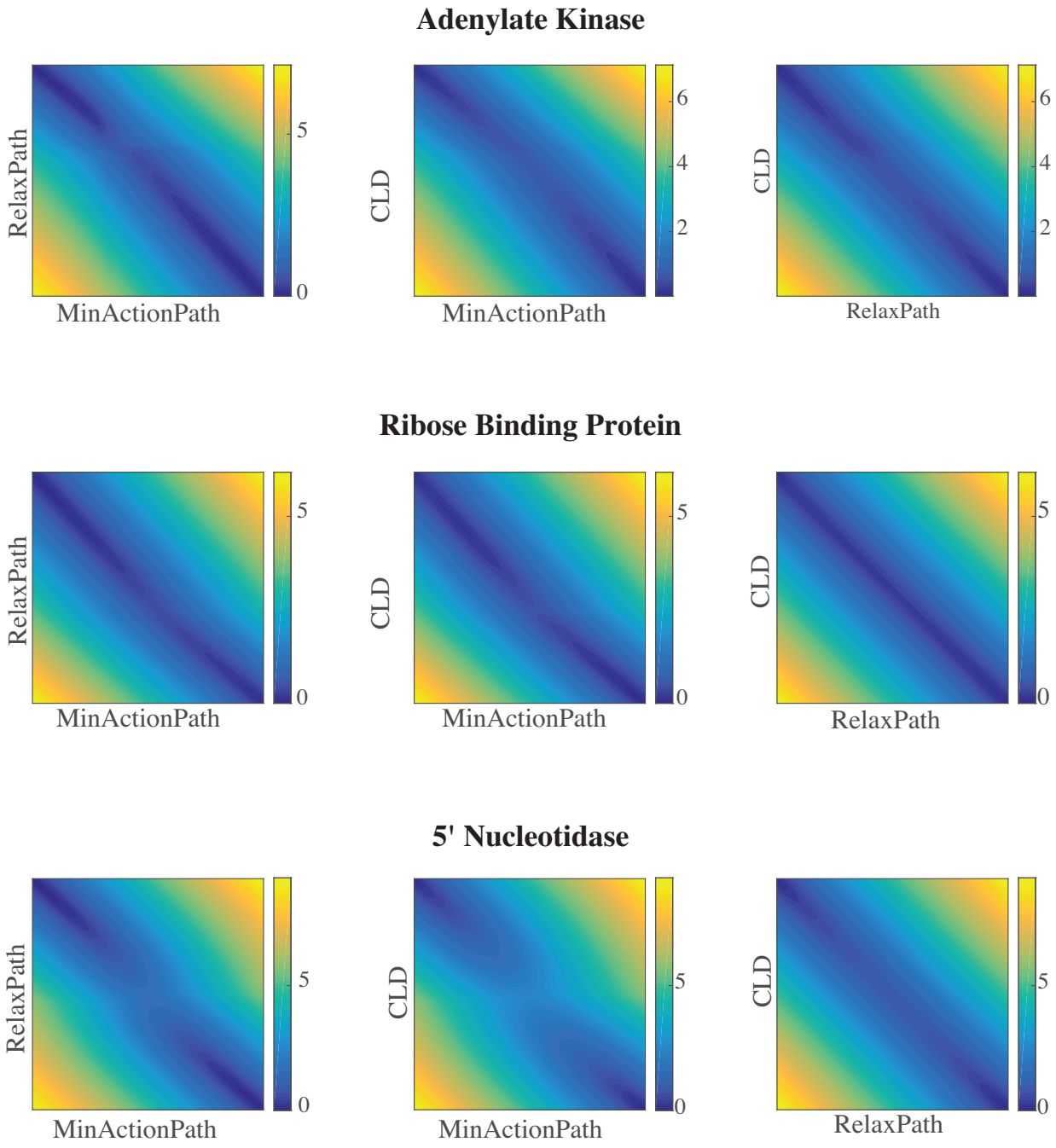

Figure 8: **A comparison of computed trajectories for three test cases**. Adenylate Kinase, Ribose Binding Protein, and 5'-Nucleotidase. The trajectories generated by CLD (this work), MinActionPath [32] and RelaxPath [65] by generating the contour plot of the between-structure cRMS values along the trajectories. The cRMS are computed using the Cα atoms only.