Thermodynamic and Kinetic Characterization of Protein Conformational Dynamics within a Riemannian Diffusion Formalism

Curtis Goolsby,¹ Ashkan Fakharzadeh,² and Mahmoud Moradi^{1,*}

¹*Department of Chemistry and Biochemistry,*

University of Arkansas, Fayetteville, AR 72701, U.S.A.

²*Department of Physics, North Carolina State University, Raleigh, NC 27607, U.S.A.*

(Dated: July 18, 2019)

Abstract

We have formulated a Riemannian framework for describing the geometry of collective variable spaces of biomolecules within the context of collective variable based molecular dynamics simulations. The formalism provides a theoretical framework to develop enhanced sampling techniques, path-finding algorithms, and transition rate estimators consistent with a Riemannian treatment of the collective variable space, where the quantities of interest such as the potential of the mean force, minimum free energy path, the diffusion constant, and the transition rate remain invariant under coordinate transformation due to the Riemannian treatment of the collective variable space. Specific algorithms within this framework are discussed such as the Riemannian umbrella sampling, the Riemannian string method, and a Riemannian-Bayesian estimator of free energy and diffusion constant, which can be used to estimate the transition rate along a minimum free energy path.

* moradi@uark.edu

I. INTRODUCTION

Biomolecular simulations have made enormous progress in recent years. The advent of Molecular dynamics (MD) in order to study biomolecular phenomena [29, 41, 45] has given researchers new insights into previously un-viewable phenomena. MD has overcome [40, 42] the common limitation of experimental techniques which force the user to choose between high-resolution static or low-resolution dynamic pictures of biomolecular systems.

MD does however have a few drawbacks. First, there is a temporal issue in that MD simulations have a shorter time scale than many relevant biological phenomena. In addition, MD has difficulty producing a full description of high-dimensional free energy landscapes. Various enhanced sampling techniques have been developed over the last few decades in order to overcome these limitations [1, 37, 43, 51, 53, 58, 59, 65, 66, 68, 71]. These techniques are often successful for simple toy models (e.g., dialanine peptide [10, 24, 38, 43, 48, 54]); however, practical applications remain quite challenging.

Perhaps the most obvious line of attack for solving these problems is simply to have stronger computing power and more efficient or specialized computer hardware. In the past 40 years, extraordinary advances have been made in this regard. Current state of the art architectures such as Anton [60, 61], peta-scale supercomputers [2, 14, 15], and GPUs [28, 30, 64, 67] have capabilities that allow for simulations with billions of particles over thousands of processors. The size of supercomputers have also given rise to algorithms developed for enhancing the sampling by brute force (rather than statistical mechanics based enhanced sampling) such as particle mesh ewald [5] and dynamic load balancing [22]. The focus of this article, however, is the enhanced sampling techniques that are specifically based on enhancing the sampling within a statistical mechanical framework.

In order to obtain relevant thermodynamic and kinetic [33] properties of a statistical mechanical system, one must integrate over high dimensional spaces, which requires large sets of independent and identically distributed samples. In order to manage the dimensionality of systems the assumption is often made that the vast majority of the high-dimensional space is practically empty (or occupied by microstates of negligible probability) and the occupied space can be approximated by a lower-dimensional manifold which contains all relevant conformations from stable states to important transition states. Ideally, any elementary reaction, can be thought of as a transition between two stable states, characterized

by a transition pathway [11, 23] (or committor function [16]).

Some methods use the ideas of an intrinsic manifold without specifically using a conventional dimensionality reduction technique as in path-optimization techniques [8, 9, 13, 46, 54, 58], which all rely on a localized transition tube. Other methods attempt to identify the intrinsic manifold by using statistical learning methods such as principal component analysis [3], isomap [11], and diffusion map.[23, 25]. Often these techniques are used to analyze MD trajectories [3, 11, 23, 57] or are combined with enhanced sampling as in metadynamics [63, 69] or adaptive biasing force [31].

The use of a coarse/reduced variable space (whether one- or multi-dimensional) in both free energy calculation methods and path-finding algorithms is quite common and various algorithms have been developed to estimate free energies [4, 24, 32, 34, 36, 37, 43, 47, 56, 59, 68] or find transition pathways [8, 9, 13, 46, 49, 54, 58] in such spaces. What is somewhat missing is a theoretical framework that allows for a rigorous treatment of the issues one needs to deal with when working with such methods. For instance, the collective variables used in collective-variable based enhanced sampling methods are often non-linear transformations of atomic coordinates. This complicates their application as noted previously by Johnson and Hummer [39], who show the conventional minimum free energy path (MFEP) obtained from various path-finding algorithms are collective-variable specific and are not invariant under non-linear coordinate transformations. Such difficulties have often been ignored in the past in the majority of the applications of the collective-variable based simulations; however, there has been attempts in addressing them as in the aforementioned work [39] or within the framework of Transition Path Theory [18, 46]. We recently introduced a Riemannian framework for the rigorous treatment of the collective-variable based spaces within the context of enhanced sampling and path-finding algorithms [21], where the distance, free energy, and MFEP are redefined in a way that are all invariant quantities and do not change under coordinate transformations. Our previous work focused on the thermodynamic characterization of protein dynamics within a Riemannian diffusion model [21]. Here we will extend the formalism to kinetics and rate calculation techniques within the same Riemannian framework.

II. FROM EUCLIDEAN TO RIEMANNIAN DIFFUSION

Euclidean geometry was developed by the Greek Euclid circa 300 BC. Riemann's work on generalizing the differential surfaces of $\mathbb{R}^3$ led to progress in many fields of science. Further progress upon his ideas allowed for the formulation of Einstein's General Theory of Relativity [19], progress in group theory [50] and generated differentiable topology.

Riemannian geometry provides a robust mathematical framework to develop a formalism for the geometry and dynamics of collective/coarse variable spaces, defined to reduce the dimensionality of the atomic models of macromolecular systems. For instance, consider a transmembrane protein whose transmembrane helices rotate under certain conditions to allow opening or closing of a gate and transporting materials across the membrane. A reasonable collective variable for such a system would be the orientation of the transmembrane helices that can be determined using principal axes of the helices or their orientation quaternions [26, 52]. In order to work in these spaces, we must first have a formalism that allows us to answer common questions which are posed in a typical collective-variable based simulation. Examples of such questions include: what is the distance between two points in the collective variable space? How is the potential of the mean force (PMF) defined at a given point in the collective variable space? How does a biasing potential affect the distribution? How can one find the MFEP? How can we estimate the PMF from biased simulations along the MFEP? What is the diffusion constant along a transition pathway? How can we estimate the rates of a reaction from the PMF and diffusion constants? The questions have previously been answered within a Euclidean framework but with our Riemannian treatment of the collective variable space, some of these concepts and quantities need to be revisited. We will begin to answer these questions one at a time.

Imagine a system containing N atoms which are represented based upon their coordinates in three dimensional space, $\mathbf{x}$, their momenta $\mathbf{p}$, and a potential energy surface $V(\mathbf{x})$. Given that they lie in a canonical equilibrium distribution their motion can be described with $m_i \dot{x}_i = p_i$ and Langevin Dynamics [44] written here as:

$$dp_i = -\left(\frac{\partial}{\partial x_i} V(x) + \gamma_i m_i^{-1} p_i\right) dt + \sqrt{2\gamma_i \beta^{-1}} dB_i. \quad (2.1)$$

Given the usual assumption for brownian motion, i.e. that the system is overdamped and thus the left side of the equation is zero, we can write:

$$dx_i = -\beta D_i \frac{\partial}{\partial x_i} V(x) dt + \sqrt{2D_i} dB_i, \quad (2.2)$$

for which dB_i is a Weiner Process and $D_i = \gamma_i^{-1} \beta^{-1}$ is the diffusion constant for atom i . In order to reduce the variable space to a coarser space we make the relation $\xi : \mathbb{R}^{3N} \rightarrow M$ such that ξ is a multi-dimensional variable which reduces the system such that $(M, \mathbf{g})$ is a complete Riemannian Manifold of dimension n with a position dependent metric $\mathbf{g}(x)$. Thus we can describe an effective potential as:

$$A(\zeta) = -\beta^{-1} \log \langle \delta_\zeta(\xi) \rangle. \quad (2.3)$$

in which ζ is any point on M and $A(\zeta)$ is the Riemannian effective potential which is related to Euclidean PMF, $\tilde{A}(\zeta)$, by:

$$\tilde{A}(\zeta) = -\beta^{-1} \log \int d^{3N} \mathbf{x} d^{3N} \mathbf{p} \prod_i \delta(\xi_i(\mathbf{x}) - \zeta_i) \rho(\mathbf{x}, \mathbf{p}), \quad (2.4)$$

$$A(\zeta) = \tilde{A}(\zeta) + \frac{1}{2} k_B T \log(g(\zeta)) = -k_B T \log \langle \delta(\zeta - \zeta_0) \rangle + \frac{1}{2} k_B T \log(g(\zeta)), \quad (2.5)$$

and δ is the dirac delta function which in Euclidean and then Riemannian formalisms satisfy (6) and (7).

$$\int d^n \zeta \delta(\zeta - \zeta_0) f(\zeta) = f(\zeta_0), \quad (2.6)$$

$$\int d\Omega_\xi \delta_\zeta(\xi) f(\xi) = f(\zeta), \quad (2.7)$$

where Ω is the volume element given by $d\Omega_\xi = \sqrt{g} d^n \xi$. Noticing that both the Euclidean formalism and Riemannian formalism share the same ξ we can now note that their PMF, is the same. We rewrite (3) as:

$$A(\zeta) = -\beta^{-1} \log \int d^{3N} \mathbf{x} d^{3N} \mathbf{p} \rho(\mathbf{x}, \mathbf{p}) \delta_\zeta(\xi(\mathbf{x})). \quad (2.8)$$

From this we can write, in a similar fashion to (2):

$$d\xi = -(\beta D \nabla A(\xi) + \mathbf{b}) dt + \sqrt{2D} d\mathbf{W}, \quad (2.9)$$

where we have removed the position dependence of the diffusion constant, which may now be represented by an arbitrary constant. Instead the position dependence is coupled into the metric, g , as seen below, and dW is once again a Wiener process. In order to continue our formulation we make the assumption that our model is diffusive. This allows us to write our model in terms of the Fokker-Planck (or Smulchowsky) equation [20]:

$$\frac{\partial}{\partial t}u(\boldsymbol{\xi}, t) = \beta D \nabla \cdot (u(\boldsymbol{\xi}, t) \nabla A(\boldsymbol{\xi})) + D \Delta u(\boldsymbol{\xi}, t), \quad (2.10)$$

which contains a gradient defined by $\nabla^i = g^{ij} \frac{\partial}{\partial j}$, the Laplace-Beltrami operator, $\Delta = \nabla \cdot \nabla = \frac{1}{\sqrt{g}} \frac{\partial}{\partial i} \sqrt{g} g^{ij} \frac{\partial}{\partial j}$, and $u(\boldsymbol{\xi}, t)$ is the probability of finding the system at $\boldsymbol{\xi}$ at time t with a boundary condition of $u(\boldsymbol{\xi}, 0) = u_0(\boldsymbol{\xi})$.

This summarizes our diffusion model, which was previously introduced in Ref. [21]. In the following we derive a new relation that allows the estimation of the PMF, metric, diffusion constant, and transition rate from unbiased simulations performed along an approximate MFEP.

III. TRANSITION RATE ESTIMATION

We will later discuss how one can find a MFEP using path-finding algorithms. For now, let us assume we have found such a pathway and we have placed a system on a given point close to this pathway at time t . One can write:

$$\begin{aligned} \frac{\partial}{\partial t}u(\boldsymbol{\xi}, t) = & \beta D \frac{\partial}{\partial s} (u(\boldsymbol{\xi}, t) \frac{\partial}{\partial s} A(\boldsymbol{\xi}(s))) + D \frac{\partial^2}{\partial s^2} u(\boldsymbol{\xi}(s), t) \\ & + \beta D \nabla_{\Sigma_s} \cdot (u(\boldsymbol{\xi}(s), t) \nabla_{\Sigma_s} A(\boldsymbol{\xi}(s))) + D \Delta_{\Sigma_s} u(\boldsymbol{\xi}(s), t), \end{aligned} \quad (3.1)$$

where the variable s is the geodesic distance along the MFEP $\boldsymbol{\xi}(s)$, where it parametrizes $\boldsymbol{\xi}(s)$. $\boldsymbol{\kappa}$ is defined on the $n - 1$ dimensional subspace of $\boldsymbol{\xi}$ perpendicular to $\boldsymbol{\xi}(s)$ (with its origin placed on $\boldsymbol{\xi}(s)$). ∇_{Σ_s} denotes the ∇ operator in the subspace of $\boldsymbol{\kappa}$. We have thus projected $\boldsymbol{\xi}(s)$ onto two components including the one-dimensional s and the $n - 1$ dimensional $\boldsymbol{\kappa}$.

Let us now define $U(s, t) \equiv \int_{\Sigma_s} u(\boldsymbol{\kappa}, t) d\Omega_{\boldsymbol{\kappa}}$, where Σ_s is an arbitrary portion of the subspace of $\boldsymbol{\kappa}$ around its origin. On the MFEP, let us also define a univariate PMF $G(s) =$

$A(\xi(s))$. We note that on a converged pathway, $\langle \kappa \rangle = 0$, gives a zero drift pathway in any high dimensional space.

Continuing, we can integrate over individual terms in Relation (3.1), where the LHS of the equation becomes $\frac{\partial}{\partial t}U(s, t)$ and the RHS terms can be approximated as follows. Assuming $u(\xi, t)$ is much larger within a relatively narrow “tube” around the MFEP, we can integrate over the cross section of this tube and the areas around the tube. In other words, we choose Σ_s to be a portion of the κ space that covers the cross section of transition tube that falls within this space as well as some low-probability areas around it. The first term of the RHS of Relation (3.1) will thus be:

$$\beta D \frac{\partial}{\partial s} \int_{\Sigma_s} (u(\xi, t) \frac{\partial}{\partial s} A(\xi(s))) d\Omega_{\kappa} = \beta D \frac{\partial}{\partial s} (U(s, t) \frac{\partial}{\partial s} G(s)), \quad (3.2)$$

where we define a univariate PMF $G(s) = A(\xi(s))$ and assume $A(\xi(s))$ is more or less constant on any cross section of transition tube perpendicular to the MFEP. The next term becomes $D \frac{\partial^2}{\partial s^2} U(s, t)$ after integration and the third term will vanish assuming $\nabla_{\Sigma_s} A(\xi(s))$ stays close to zero within the tube. To be more precise, we assume the following integral is negligible as $\nabla_{\Sigma_s} A(\xi)$ is negligible close the MFEP:

$$\beta D \int_{\Sigma_s} \nabla_{\Sigma_s} \cdot (u(\xi, t) \nabla_{\Sigma_s} A(\xi)) d\Omega_{\kappa} = \beta D \oint_{\partial \Sigma_s} u(s) \nabla_{\Sigma_s} G(s) \cdot \mathbf{n} d\sigma_{\kappa}, \quad (3.3)$$

where the divergence theorem is used. Similarly, we can use the divergence theorem to reduce the last term to:

$$D \int \Delta_{\Sigma_s} u(\xi, t) d\Omega_{\Sigma_s} = D \oint_{\partial \Sigma_s} \nabla_{\Sigma_s} u(s) \cdot \mathbf{n} d\sigma_{\kappa}. \quad (3.4)$$

We assume this term is also negligible since the $\nabla_{\Sigma_s} u(\xi) \cdot \mathbf{n}$ term is only evaluated outside the transition tube and is expected to be close to zero on average. These approximations reduce Relation (3.1) to a one-dimensional diffusion equation in terms of probability density $U(s, t)$ and the univariate PMF (or potential energy) $G(s)$:

$$\frac{\partial}{\partial t} U(s, t) = \beta D \frac{\partial}{\partial s} (U(s, t) \frac{\partial}{\partial s} G(s)) + D \frac{\partial^2}{\partial s^2} U(s, t). \quad (3.5)$$

The thin transition tube, used to simplify Relation (3.1) is a common assumption made for path-finding algorithms such as string method [17]. At this point, we have reduced our atomic model, first to a coarse variable space, and then we have focused upon a single dimension, namely the transition pathway of interest.

Relation (3.5) is quite similar to the conventional Smoluchovsky equation in a one-dimensional Euclidean space. However, this is due to the fact that we assumed s to represent the geodesic distance along the MFEP path. A slightly more general relation can be derived based on (3.5) for an arbitrary parameter r , parametrizing the MFEP path. The r space will then be associated with a 1D metric $h(r)$. We have $ds^2 = h(r)dr^2$ or $\sqrt{h(r)} = \frac{ds}{dr}$. The Riemannian Smoluchowsky equation is then locally written in terms of r as:

$$\frac{\partial}{\partial t}U(r, t) = \beta D \frac{1}{\sqrt{h(r)}} \frac{\partial}{\partial r} \left(U(r, t) \frac{1}{\sqrt{h(r)}} \frac{\partial}{\partial r} G(r) \right) + D \frac{1}{\sqrt{h(r)}} \frac{\partial}{\partial r} \frac{1}{\sqrt{h(r)}} \frac{\partial}{\partial r} U(r, t), \quad (3.6)$$

which is slightly different from a conventional 1D position-dependent diffusion equation. Here $D/h(r)$ is equivalent to the conventional 1D position-dependent diffusion constant and $G(r)$ is the same as the conventional 1D PMF in terms of r with an extra term $\frac{1}{2}k_B T \log h(r)$, which makes the Riemannian PMF independent of the choice of r .

Finally, we can rewrite Relation (3.6) as:

$$\frac{\partial}{\partial t}U(r, t) = D \frac{1}{\sqrt{h(r)}} \frac{\partial}{\partial r} \left(e^{-\beta G(r)} \frac{1}{\sqrt{h(r)}} \frac{\partial}{\partial r} e^{\beta G(r)} U(r, t) \right) \quad (3.7)$$

Now suppose that one has been able to identify the MFEP without necessarily quantifying the metric. Relation (3.7) provides a framework to determine the dynamics of the system as long as it stays close to the MFEP transition tube. The PMF and metric along r ($G(r)$ and $h(r)$) fully describe the diffusive dynamics and the arbitrary diffusion constant D simply determines the unit of $h(r)$. In Section V, we will discuss in more detail how the estimation of $G(r)$ and/or $h(r)$ is possible using unbiased or biased simulations; however, none of these discussions would be relevant without identifying the MFEP.

IV. PATH FINDING ALGORITHMS

In Sections II and III, we developed the Riemannian framework of path finding algorithms. The PMF defined above is invariant under coordinate transformation as is the MFEP. The MFEP and the formalism of the String Method with Swarms of Trajectories, SMwST [55]. We will also in this section describe Post-Hoc String Method which allows a user to define certain parameters to move from state a to state b while finding a minimum free energy pathway, MFEP [27, 46], by defining a string based on previously performed sampling.

First we present PHSM. In order to explore the reaction coordinates we need to find a MFEP. We can find the MFEP by using a predetermined sample set to obtain a reaction

pathway. Before using PHSM the user must supply a large set of conformations and weights for said conformations. PHSM then requires inputs which give (i) a metric/distance norm $\|\cdot\|$ to measure the closeness of conformations; (ii) an initial pathway of n conformations; (iii) transition tube thickness parameter ϵ ; and (iv) convergence criteria including a threshold δ and a persistence number P : The particular algorithms used below for finding an initial path and re-parametrization have their own parameters (i.e., $\Delta_{\min}/\max$ and δ' , respectively). In order to completely clarify this process, we will describe the entire process now.

The collective variable space and metric are defined as above. Next convergence criteria must be selected, $\Delta_{\min}/\max$, δ , δ' , and ϵ . The $\Delta_{\min}/\max$ determine the spacing of your images. The drift, which can be alternatively defined depending on your variable space must be less than δ' , and finally the transition tube has a thickness of ϵ and must move from its original center by no more than δ which has a usage we will see in a moment. The $\Delta_{\min}/\max$ are then used to generate an initial string which varies from $\zeta(0)$ to $\zeta(1)$ over the coarse variable space with distance between images defined by $\Delta_{\min}/\max$. We note that the values chosen are dependent on the distance metric used. This serves as a seed for the string method that follows. This initial string and subsequent strings are updated with new centers found by Voronoi Tessellation [70]. These centers are rejected if they move more than the ϵ or δ parameters, i.e. if the center strays from the initial center by more than the tube thickness or if the value strays too much from the previous center the value will be discarded. If all the centers are discarded, parameters must be adjusted. The re-parameterization of image centers is performed with de Casteljau's algorithm [12] in order to implement Bezier curves [6, 7] for interpolation. Finally, if the change in the string centers from the previous step is less than δ' , the string is converged.

Now that we have a seed string from PHSM, we will discuss SMwST. SMwST starts from an initial string, defined by N points (or images/windows) denoted x_I , $I=1, \dots, N$. A collective variable ξ is defined as above and the initial image centers are given by the initial string. A swarm of trajectories are created, giving M copies of each point or $N \times M$ systems in total. Each SMwST system then undergoes three iterative steps. First, the system is restrained for a defined number of steps, τ_r , using a harmonic potential, $V_B(x) = U(\xi(x)) = \frac{k}{2}d(\xi(x), \zeta)^2$, which has its center on the current center and the distance metric defined as $d(x, x_0)$. The simulation is then released from their harmonic potential and allowed to drift for a defined number of steps, τ_d . Finally, the system is re-parameterized

to a new center based upon the average $\xi(x(\tau_r + \tau_d))$ values for all M systems associated with an image. Images are kept equidistant and new systems are created until the method reaches convergence to a zero-drift path, i.e. the string centers oscillate. In contrast to a Euclidean SMwST which does not necessarily identify the zero-drift path in the normal coordinates our Riemannian SMwST identifies the zero-drift path in the normal coordinates which provides a coordinate transformation invariant solution.

Our version of SMwST varies in four distinct points from the conventional SMwST:

1. The metric at each image center ξ_I is estimated via:

$$g^{ij} \approx \Sigma_l \frac{m}{m_l} \left\langle \frac{\partial}{\partial x^i} \xi^i \frac{\partial}{\partial x^j} \xi^j \right\rangle_I, \quad (4.1)$$

in which the average over the images is calculated for each iteration by using the $x(\tau_r)$ values observed for all M copies of I.

2. The biasing potential is approximated by using $g(\xi_I)$ to calculate the inner product of $U_I(\xi) \approx \frac{k}{2}(\xi - \xi_I) \cdot (\xi - \xi_I)$.
3. As opposed to using Euclidean distance to perform the interpolation necessary to update images, the distance between two consecutive images is defined by $s_I - s_{I-1} = \sqrt{(\xi - \xi_{I-1}) \cdot (\xi - \xi_{I-1})}$, in which $g(\xi_{I-1})$ is used to evaluate the inner product.
4. In the re-parametrization step, the mean drifted centers $\langle \xi(\tau_D) \rangle$, used in conventional SMwST, are shifted by a biasing term $-\mathbf{b}(\xi_I)\tau_d\Delta t$ for each image I. $b_i = Dg^{jk}\Gamma_{jk}^i$ is used to evaluate the geometric drift, where the Christoffel symbols are defined as $\Gamma_{jk}^i = \frac{1}{2}g^{im}(\partial_j g_{mk} + \partial_k g_{mj} - \partial_m g_{jk})$. The metric derivatives can be approximated by finite difference, e.g., $\partial_l g_{mk}(\xi_I) \approx \frac{(g_{mk}(\xi_I) - g_{mk}(\xi_{I-1}))}{(\xi_I^l - \xi_{I-1}^l)}$.

V. PERTURBED FREE ENERGY VS PMF

As in the conventional case, the string centers can then be used as the harmonic centers for US simulations. From this the US trajectories can be used to estimate the perturbed free energies by self-consistently solving:

$$e^{-\beta F_i} = \sum_t \frac{e^{-\beta U_i(\xi^t)}}{\sum_j S_j e^{-\beta(U_j(\xi^t) - F_j)}} \quad (5.1)$$

For which the sums are over all coarse variables and the images. F is an estimator for perturbed free energy and $U_\xi = \frac{k}{2}d(\xi^t, \xi_0)$ are the harmonic potentials calculated, around the point ξ_0 with a distance metric $d(\xi, \xi_0)$, as described in the following five equations.

$$e^{-\beta F(s)} = \int d^{3N} \mathbf{x} d^{3N} \mathbf{p} e^{-\beta U_s(\xi(\mathbf{x})) \rho(\mathbf{x}, \mathbf{p})} \quad (5.2)$$

Or

$$e^{-\beta F(s)} = \int d\Omega_\xi e^{-\beta(A(\xi) + U_s(\xi))} \quad (5.3)$$

$$f(\zeta) = -\beta^{-1} \log \int d\Omega_\xi e^{-\beta(A(\xi) + \frac{k}{2}d(\xi, \zeta)^2)} \quad (5.4)$$

Unlike $F(s)$, which is defined on the $\zeta(s)$ curve only, $f(\xi)$ is defined everywhere on the manifold; however, the two functions have the same values at any point along the curve, i.e., $f(\zeta(s)) = F(s)$. Following a similar approach as in the 1-D case we can employ a Riemannian adaptation of the Weirstrauss transform to derive a meaningful relation between the Riemannian PMF and the perturbed free energy.

$$e^{-\beta A(\zeta)} = \left(\frac{2\pi}{\beta k}\right)^{-\frac{\pi}{2}} e^{-\frac{1}{2\beta k}\Delta} e^{-\beta f(\xi)}|_{\xi=\zeta} \quad (5.5)$$

Given a large force constant, k , equation 24 can be simplified to

$$A(\zeta) \approx f(\zeta) + \frac{1}{2\beta k}(\beta \nabla f(\xi) - \Delta f(\xi))|_{\xi=\zeta} + Constant \quad (5.6)$$

VI. DETERMINING THE RATE OF TRANSITION

Our final formulation mathematically follows closely the work of Hummer [35], except we are working within a Riemannian geometry. We define $h(\mathbf{r}) = \sqrt{D}m^{-\frac{1}{2}}(\mathbf{r})$ and $H(\mathbf{r}, t) = e^{\beta G(\mathbf{r})}U(\mathbf{r}, t)$ such that (14) can be rewritten in an easily discretized form:

$$\mathbf{H}(t + \delta t) = (\mathbf{1} + \mathbf{R}\delta t)\mathbf{H}(t) \quad (6.1)$$

$$\mathbf{H}(\Delta t) = e^{\mathbf{R}\Delta t}\mathbf{H}(0) \quad (6.2)$$

$\mathbf{H}$ is here a discretized representation of the transition probabilities which can be determined empirically and $\mathbf{R}$ is a tridiagonal matrix for which more can be seen in section 4.

We can self consistently [62] solve by maximizing the likelihood between states at t and $t \pm 1$ defined by:

$$L = \prod_{\alpha} e^{\mathbf{R} \Delta t_{\xi_{t \pm 1}, \xi_t}} \quad (6.3)$$

using the tridiagonal $\mathbf{R}$ from (16):

$$R_{i,i} = -R_{i,i+1} - R_{i,i-1} \quad (6.4)$$

$$R_{i,i \pm 1} = \delta r^{-2} h(r_i) h(r_{i \pm \frac{1}{2}}) e^{-\beta(G(r_{i \pm \frac{1}{2}}) - G(r_i))} \quad (6.5)$$

$$R_{i,i \pm 1} = D \delta s_i^{-1} \delta s_{i \pm \frac{1}{2}}^{-1} e^{-\beta(G(r_{i \pm \frac{1}{2}}) - G(r_i))} \quad (6.6)$$

using the relation that $m(r_i) = \frac{\delta s_i}{\delta r_i}$ and r values chosen such that $\delta r_i = \delta r_{i+1}$. Now it can be seen that working in a Riemannian collective variable space is valid for the determination of quantities of interest such as rate, free energy calculations, diffusion, and reaction pathways.

VII. ORIENTATION QUATERNION

Among the different routes to describe rotation and translation, perhaps using quaternions are more convenient one. In our approach, quaternions are 4-vector (q_0, q_1, q_2, q_3) which are normalized, $\sum_i q_i^2 = 1$. By definition the quaternion can be represented by $(\cos(\frac{\theta}{2}), \sin(\frac{\theta}{2}) \hat{u})$, where θ is the optimum rotation angle and $\hat{u}$ is a unit vector associated with the optimum axis of rotation.

To estimate the distance between two orientation quaternion collective variables, $(q - q')^2 = g^{ij}(q - q')_i (q - q')_j$, we use our approximation for the metric in (4.1). Following Dill (cite Dill's paper), the orientation quaternions should satisfy the eigenvalue equation $Sq = \lambda q$, where S is a 4×4 symmetric, traceless matrix, made of the correlation matrix (Euclidean

coordinates). The λ are the eigenvalues S , and corresponding to eigenvectors (q). Differentiating both sides of the above eigenvalue equation we have:

$$\frac{\partial S}{\partial x^l} q^{(1)} + S \frac{\partial q^{(1)}}{\partial x^l} = \frac{\partial \lambda^{(1)}}{\partial x^l} q^{(1)} + \lambda^{(1)} \frac{\partial q^{(1)}}{\partial x^l} \quad (7.1)$$

here $q^{(1)}$ is the corresponding eigenvector to the leading eigenvalue $\lambda^{(1)} = \lambda_{max}$. Using the orthogonality property between eigenvectors we obtain:

$$\begin{aligned} \frac{\partial \lambda^{(1)}}{\partial x^l} &= q^{(1)} \cdot \frac{\partial S}{\partial x^l} q^{(1)} \\ \frac{\partial q^{(1)}}{\partial x^l} &= \sum_{\alpha=2}^4 \frac{1}{\lambda_{\alpha} - \lambda^{(1)}} (q^{(\alpha)} \cdot \frac{\partial S}{\partial x^l}) q^{(\alpha)} \end{aligned} \quad (7.2)$$

We can express the derivative of quaternions as

$$\frac{\partial q}{\partial x^l} = \sum_{\alpha=2}^4 A^{(\alpha)} q^{(\alpha)} \quad (7.3)$$

where we drop the index (1). So we can approximate our metric as

$$g^{ij} \approx \sum_l \frac{m}{m_l} \left\langle \frac{\partial}{\partial x^l} q^i \frac{\partial}{\partial x^l} q^j \right\rangle_I = \sum_l \frac{m}{m_l} \left\langle \sum_{\alpha=2}^4 \sum_{\beta=2}^4 A^{(\alpha)} A^{(\beta)} q^{(\alpha)i} q^{(\beta)j} \right\rangle_I \quad (7.4)$$

-
- [1] C. Abrams and G. Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2013.
- [2] F. Allen, G. Almasi, W. Andreoni, D. Beece, B. J. Berne, A. Bright, J. Brunheroto, C. Cascaval, J. Castanos, P. Coteus, P. Crumley, A. Curioni, M. Denneau, W. Donath, M. Eleftheriou, B. Fitch, B. Fleischer, C. J. Georgiou, R. Germain, M. Giampapa, D. Gresh, M. Gupta, R. Haring, H. Ho, P. Hochschild, S. Hummel, T. Jonas, D. Lieber, G. Martyna, K. Maturu, J. Moreira, D. Newns, M. Newton, R. Philhower, T. Picunko, J. Pitera, M. Pitman, R. Rand, A. Royyuru, V. Salapura, A. Sanomiya, R. Shah, Y. Sham, S. Singh, M. Snir, F. Suits, R. Swetz, W. C. Swope, N. Vishnumurthy, T. J. C. Ward, H. Warren, and R. Zhou. Blue Gene: A vision for protein science using a petaflop supercomputer. *IBM Systems Journal*, 40:310–327, 2001.
- [3] A. Amadei, A. B. M. Linnsen, and H. J. C. Berendsen. Essential dynamics of proteins. *PROTEINS: Structure, Function, and Genetics*, 17:412–425, 1993.
- [4] V. Babin, C. Roland, and C. Sagui. Adaptively biased molecular dynamics for free energy calculations. *Journal of Chemical Physics*, 128(13):134101, 2008.

- [5] P. F. Batcho, D. A. Case, and T. Schlick. Optimized particle-mesh Ewald/multiple-time step integration for molecular dynamics simulations. *J. Chem. Phys.*, 115(9):4003–4018, 2001.
- [6] P. Bezier. Definition numerique des courbes et surfaces i. *Automatisme*, 11, 1966.
- [7] P. Bezier. Definition numerique des courbes et surfaces ii. *Automatisme*, 12, 1966.
- [8] L. Cao, C. Lv, and W. Yang. Hidden conformation events in dna base extrusions: A generalized-ensemble path optimization and equilibrium simulation study. *J. Chem. Theory Comput.*, 9(8):3756–3768, 2013.
- [9] M. Chen and W. Yang. On-the-path random walk sampling for efficient optimization of minimum free-energy path. *J. Comput. Chem.*, 30(11):1649–1653, 2009.
- [10] R. Crehuet and M. J. Field. A temperature-dependent nudged-elastic-band algorithm. *Journal of Chemical Physics*, 118(21):9563–9571, 2003.
- [11] P. Das, M. Moll, H. Stamati, L. E. Kavraki, and C. Clementi. Low-dimensional, free-energy landscapes of protein folding reactions by nonlinear dimensionality reduction. *Proceedings of the National Academy of Sciences, USA*, 103(26):9887–9890, 2006.
- [12] P. de Casteljaeu. Courbes et surfaces a poles. *Societe Anonyme Andre Citroen, Internal Report*, 1985.
- [13] G. Díaz Leines and B. Ensing. Path finding on high-dimensional free energy landscapes. *Physical Review Letters*, 109:020601, Jul 2012.
- [14] T. Diede, C. F. Hagenmaier, G. S. Miranker, J. J. Rubinstein, and W. Worley. The titan graphics supercomputer architecture. *Computer*, 21(9):13–30, 1988.
- [15] A. Dubrow. What got done in one year at nsf’s stampede supercomputer. *Computing in Science & Engineering*, 17(2):83–88, 2015.
- [16] W. E, W. Ren, and E. Vanden-Eijnden. Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes. *Chemical Physics Letters*, 413(1-3):242 – 247, 2005.
- [17] W. E. and E. Vanden-Eijnden. Towards a theory of transition paths. *Journal of Statistical Physics*, 123(3):503, May 2006.
- [18] W. E and E. Vanden-Eijnden. Transition-path theory and path-finding algorithms for the study of rare events. *Annual Review of Physical Chemistry*, 61:391–420, 2010.
- [19] A. Einstein. The formal foundation of the general theory of relativity. 1914.

- [20] M. v. S. Einstein, A. *Untersuchungen über die Theorie der Brownschen Bewegung*. Ostwalds Klassiker der exakten Wissenschaften. Deutsch, 1997.
- [21] A. Fakharzadeh and M. Moradi. Effective Riemannian diffusion model for conformational dynamics of biomolecular systems. *Journal of Physical Chemistry Letters*, 7:4980–4987, 2016.
- [22] J.-L. Fattebert, D. F. Richards, and J. N. Glosli. Dynamic load balancing algorithm for molecular dynamics based on voronoi cells domain decompositions. *Computer Physics Communications*, 183(12):2608–2615, 2012.
- [23] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis. Systematic determination of order parameters for chain dynamics using diffusion maps. *Proceedings of the National Academy of Sciences, USA*, 107(31):13597–13602, 2010.
- [24] A. L. Ferguson, A. Z. Panagiotopoulos, P. G. Debenedetti, and I. G. Kevrekidis. Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide. *Journal of Chemical Physics*, 134(13):135103, 2011.
- [25] A. L. Ferguson, A. Z. Panagiotopoulos, I. G. Kevrekidis, and P. G. Debenedetti. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chemical Physics Letters*, 509(1):1–11, 2011.
- [26] G. Fiorin, M. L. Klein, and J. Hénin. Using collective variables to drive molecular dynamics simulations. *Molecular Physics*, 111(22–23):3345–3362, 2013.
- [27] K. Fukui. Formulation of the reaction coordinate. *The Journal of Physical Chemistry*, 74(23):4161–4163, 1970.
- [28] A. W. Gotz, M. J. Williamson, D. Xu, D. Poole, S. Le Grand, and R. C. Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 1. generalized born. *Journal of chemical theory and computation*, 8(5):1542–1555, 2012.
- [29] T. Hansson, C. Oostenbrink, and W. F. van Gunsteren. Molecular dynamics simulations. *Current Opinion in Structural Biology*, 12(2):190–196, 2002.
- [30] M. J. Harvey, G. Giupponi, and G. D. Fabritiis. ACEMD: Accelerating biomolecular dynamics in the microsecond time scale. *Journal of Chemical Theory and Computation*, 5(6):1632–1639, 2009.
- [31] B. Hashemian, D. Millán, and M. Arroyo. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *Journal of Chemical Physics*, 139(21):214101, 2013.

- [32] T. Hayami, K. Kasahara, H. Nakamura, and J. Higo. Molecular dynamics coupled with a virtual system for effective conformational sampling. *Journal of computational chemistry*, 2018.
- [33] M. Held and F. Noé. Calculating kinetics and pathways of protein–ligand association. *European journal of cell biology*, 91(4):357–364, 2012.
- [34] T. Huber, A. E. Torda, and W. F. van Gunsteren. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput.-Aided Mol. Des.*, 8:695–708, 1994.
- [35] G. Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New Journal of Physics*, 7:34, 2005.
- [36] G. Hummer and I. G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *Journal of Chemical Physics*, 118:10762–10773, 2003.
- [37] S. Izrailev, S. Stepaniants, M. Balsera, Y. Oono, and K. Schulten. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophysical Journal*, 72:1568–1581, 1997.
- [38] H. Jang and T. B. Woolf. Multiple pathways in conformational transitions of the alanine dipeptide: An application of dynamic importance sampling. *Journal of Computational Chemistry*, 27(11):1136–1141, 2006.
- [39] M. E. Johnson and G. Hummer. Characterization of a dynamic string method for the construction of transition pathways in molecular reactions. *Journal of Physical Chemistry B*, 116:8573–8583, 2012.
- [40] M. Karplus and J. Kuriyan. Molecular dynamics and protein function. *Proceedings of the National Academy of Sciences, USA*, 102:6679–6685, 2005.
- [41] M. Karplus and J. A. McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 265:654–652, 2002.
- [42] J. L. Klepeis, K. Lindorff-Larsen, R. O. Dror, and D. E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Current Opinion in Structural Biology*, 19:120–127, 2009.
- [43] A. Laio and M. Parrinello. Escaping free energy minima. *Proceedings of the National Academy of Sciences, USA*, 99(20):12562–12566, 2002.
- [44] A. Leach. *Molecular Modelling: Principles and Applications*. Prentice Hall, 1996.

- [45] K. Lindorff-Larsen, S. Piana, R. O. Dror, and D. E. Shaw. How fast-folding proteins fold. *Science*, 334:517–520, 2011.
- [46] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *Journal of Chemical Physics*, 125(2):024106, 2006.
- [47] P. R. L. Markwick, L. C. T. Pierce, D. B. Goodin, and J. A. McCammon. Adaptive accelerated molecular dynamics (Ad-AMD) revealing the molecular plasticity of P450cam. *Journal of Physical Chemistry Letters*, 2(3):158–164, 2011.
- [48] M. Mezei. Evaluation of the adaptive umbrella sampling method. *Molecular Simulation*, 3:301–313, 1989.
- [49] Y. Miao, F. Feixas, C. Eun, and J. A. McCammon. Accelerated molecular dynamics simulations of protein folding. *Journal of computational chemistry*, 36(20):1536–1549, 2015.
- [50] J. Milnor. On fundamental groups of complete affinely flat manifolds. *Advances in Mathematics*, 25(2), 1977.
- [51] A. Mitsutake, Y. Mori, and Y. Okamoto. Enhanced sampling algorithms. In L. Monticelli and E. Salonen, editors, *Biomolecular Simulations*, volume 924 of *Methods in Molecular Biology*, pages 153–195. Humana Press, New York, 2013.
- [52] M. Moradi, V. Babin, C. Roland, and C. Sagui. Reaction path ensemble of the B–Z-DNA transition: a comprehensive atomistic study. *Nucleic Acids Research*, 41(1):33–43, 2013.
- [53] M. Moradi, V. Babin, C. Sagui, and C. Roland. Recipes for free energy calculations in biomolecular systems. In L. Monticelli and E. Salonen, editors, *Biomolecular Simulations*, volume 924 of *Methods in Molecular Biology*, pages 313–337. Humana Press, New York, 2013.
- [54] A. C. Pan, D. Sezer, and B. Roux. Finding transition pathways using the string method with swarm of trajectories. *Journal of Physical Chemistry B*, 112(11):3432–3440, 2008.
- [55] J. Pan, S. Tristram-Nagle, N. Kucerka, and J. F. Nagle. Temperature dependence of structure, bending rigidity, and bilayer interactions of dioleoylphosphatidylcholine bilayers. *Biophysical Journal*, 94:117–124, 2008.
- [56] F. Paul. *Markov state modeling of binding and conformational changes of proteins*. PhD thesis, Universität Potsdam Potsdam, 2017.
- [57] J. R. Perilla and T. B. Woolf. Towards the prediction of order parameters from molecular dynamics simulations in proteins. *Journal of Chemical Physics*, 136:164101, 2012.

- [58] W. 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. *Journal of Chemical Physics*, 123(13):134109, 2005.
- [59] J. Schlitter, M. Engels, P. Krüger, E. Jacoby, and A. Wollmer. Targeted molecular dynamics simulation of conformational change — application to the T $\leftrightarrow$ R transition in insulin. *Molecular Simulation*, 10(2–6):291–308, 1993.
- [60] D. E. Shaw, R. O. Dror, J. K. Salmon, J. P. Grossman, K. M. Mackenzie, J. A. Bank, C. Young, M. M. Deneroff, B. Batson, K. J. Bowers, E. Chow, M. P. Eastwood, D. J. Ierardi, J. L. Klepeis, J. S. Kuskin, R. H. Larson, K. Lindorff-Larsen, P. Maragakis, M. A. Moraes, S. Piana, Y. Shan, and B. Towles. Millisecond-scale molecular dynamics simulations on anton. In *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*, SC ’09, pages 39:1–39:11, New York, NY, USA, 2009. ACM.
- [61] D. E. Shaw, J. Grossman, J. A. Bank, B. Batson, J. A. Butts, J. C. Chao, M. M. Deneroff, R. O. Dror, A. Even, C. H. Fenton, et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 41–53. IEEE Press, 2014.
- [62] A. Singharoy, C. Chipot, M. Moradi, and K. Schulten. Chemomechanical coupling in hexameric protein–protein interfaces harnesses energy within v-type atpases. *Journal of the American Chemical Society*, 139(1):293–310, 2016.
- [63] V. Spiwok, P. Lipovová, and B. Králová. Metadynamics in essential coordinates: Free energy simulation of conformational changes. *Journal of Physical Chemistry B*, 111(12):3073–3076, 2007.
- [64] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten. Accelerating molecular modeling applications with graphics processors. *Journal of Computational Chemistry*, 28:2618–2640, 2007.
- [65] Y. Sugita, A. Kitao, and Y. Okamoto. Multidimensional replica-exchange method for free-energy calculations. *Journal of Chemical Physics*, 113(15):6042–6051, 2000.
- [66] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314:141–151, 1999.

- [67] D. E. Tanner, J. C. Phillips, and K. Schulten. GPU/CPU algorithm for generalized Born / solvent-accessible surface area implicit solvent calculations. *Journal of Chemical Theory and Computation*, 8:2521–2530, 2012.
- [68] G. M. Torrie and J. P. Valleau. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella Sampling. *Journal of Chemical Physics*, 23(2):187–199, 1977.
- [69] G. A. Tribello, M. Ceriotti, and M. Parrinello. Using sketch-map coordinates to analyze and bias molecular dynamics simulations. *Proceedings of the National Academy of Sciences, USA*, 109(14):5196–5201, 2012.
- [70] G. Voronoi. Nouvelles applications des parametres continus a la theorie des formes quadratiques. *J. reine angew*, 133, 1907.
- [71] D. M. Zuckerman and T. B. Woolf. Efficient dynamic importance sampling of rare events in one dimension. *Physical Review E*, 63(1):016702, 2000.