

A Rigorous Method for Identifying One-Dimensional Reaction Coordinate in Complex Molecules

Shanshan Wu, Huiyu Li and Ao Ma*

Richard Loan and Hill Department of Biomedical Engineering
The University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607

*correspondence should be addressed to:

Ao Ma
Email: aoma@uic.edu
Tel: (312) 996-7225

Abstract

Understanding the mechanism of functional protein dynamics is critical to understanding protein functions. Reaction coordinate is a central topic in protein dynamics and the grail is to find the one-dimensional reaction coordinate that can fully determine the value of committor (i.e. the reaction probability in configuration space) for any protein configuration. We present a new method that, for the first time, uses a fundamental mechanical operator, the generalized work functional, to identify the rigorous one-dimensional reaction coordinate in complex molecules. For a prototypical biomolecular isomerization reaction, the one-dimensional reaction coordinate identified by the current method can determine committor with an accuracy far exceeding what was achieved by previous methods. This method only requires modest computational cost and can be readily applied to large molecules. Most importantly, the generalized work functional is the physical determinant of the collectivity in functional protein dynamics and provides a tentative roadmap that connects the structure of a protein to its function.

Proteins are the building blocks of biological systems responsible for most biological functions. Understanding the mechanism of protein function is of paramount importance. The central dogma of protein science is that structure determines function. The holy grail is the physical principle that explains how the structure of a protein determines its function. The route to this principle starts with recognizing that most protein functions, such as ligand binding, allostery and effects of mutations, involve significant conformational dynamics because a protein has multiple functional structures and transitions between these structures are required for its function ^{1, 2}. Structure determines function because the specific structure enables the desired functional dynamics. Understanding protein function requires understanding functional dynamics.

Importance of reaction coordinates. Most functional dynamics are activated processes similar to chemical reactions: a protein must cross an activation barrier much higher than thermal energy to move from the reactant (initial) state to the product (final) state. A central concept here is the reaction coordinates (**RCs**): the small number of essential coordinates that fully determine the progress of a reaction ³. In particular, RCs determine the location of the transition state. The reaction progresses in the forward direction when RCs move towards the product state; it regresses when RCs move towards the reactant state; movements of all the other coordinates are irrelevant.

Reaction coordinates are the cornerstone of the standard reaction rate theories. Reaction dynamics is the dynamics of RCs. Indeed, Kramers theory assumes that the dynamics of the RC is governed by a Langevin equation ⁴, which is extended to generalized Langevin equation in Grote-Hynes theory ⁵. Similarly, transition state theory relies on simplifying assumptions on the essential features of the dynamics of the RC ⁶⁻¹⁰. The applicability of all these theories requires full knowledge of the RCs.

Reaction coordinates are also central to enhanced sampling, a major current research direction in computational biophysics. Most algorithms rely on applying biasing potentials on one or a few collective variables (**CVs**) selected by intuitions to increase sampling of the relevant configuration space ¹¹⁻¹⁷. However, these algorithms are effective only if the CVs coincide with RCs. Otherwise, the infamous “hidden barrier” problem will inevitably appear and prevent effective sampling ¹⁸⁻²⁰.

The “hidden barrier” is the actual activation barrier that lies on the true RCs, which is missed by the biasing potentials along CVs that do not overlap with the RCs significantly.

The rigorous concept of RC. Given the central importance of RCs, a rigorous definition is essential. An important concept is committor (p_B): the probability that a dynamic trajectory initiated from a system configuration, with initial momenta drawn from Boltzmann distribution, to reach the product state ^{3, 21-23}. Committor rigorously parameterizes the progress of a reaction (e.g. the transition state has $p_B = 0.5$). Consequently, RCs are the set of coordinates that can fully determine the committor value of any system configuration. Other coordinates in the system are irrelevant. Reaction coordinates that satisfy this definition provide an accurate reduced description of a reaction. In contrast, ad hoc RCs that do not satisfy this definition bear little relevance to reaction dynamics and mechanism ^{24, 25}.

A critical question is: How many RCs are there in a system? The standard reaction rate theories assume that, for a single-channel reaction, there exists a single one-dimensional RC (**1D-RC**) that can fully determine committor. Even though the multi-dimensional extensions of these theories assume that a reaction is controlled by the free energy surface of multiple RCs, it is assumed that there is only one unstable direction in the saddle region ^{7, 26}, which is the direction of the 1D-RC. The seminal work by Berezhkovskii and Szabo ²⁷ showed that it is possible to find a 1D-RC that reproduces the same reaction rate determined by multiple RCs ^{28, 29}.

For single-channel reactions, the concept of multiple RCs is an approximation to the concept of 1D-RC. In the schematic example of Fig. 1a, q_1 and q_2 are considered RCs because they both have significant overlaps with r_c , the 1D-RC. However, they also have components orthogonal to r_c , pointing to their inaccuracy as RCs. The concept of multiple RCs is semi-quantitative and has some leeway; the concept of 1D-RC is quantitative and precise. The former is adopted much more widely ^{3, 30} because, compared to the 1D-RC, they are much easier to identify.

The goal for studying RCs. The rigorously defined RCs offer a chance to bring order to the apparent chaos and complexity of protein dynamics that thwarted understanding of protein function. The small number of RCs compared to the myriad of coordinates in a protein indicates

that protein dynamics has an intrinsic order and governing principles, which is embedded in the RCs as they accurately describe the progress of protein dynamics. The physical cause behind the RCs is the physical determinant of this intrinsic order. The goal for studying RCs is to uncover and, furthermore, understand the principle behind this physical determinant.

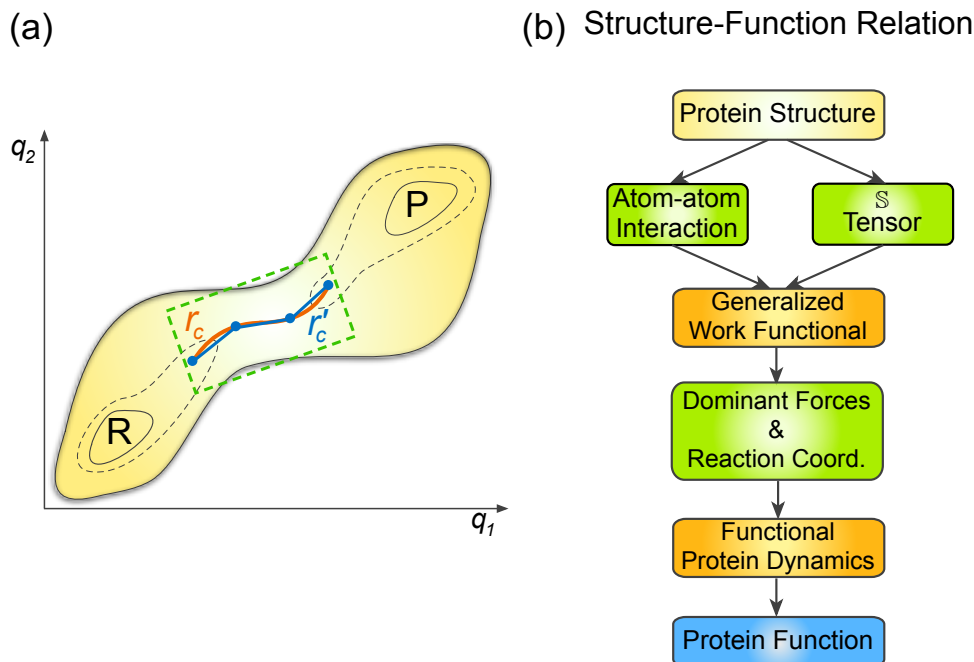


Figure 1: One-dimensional reaction coordinate and structure-function relationship. **(a)** A schematic of the potential energy surface and RCs of a two-dimensional system. The light yellow region indicates the contours of the potential energy surface, with the reactant and product states marked by R and P respectively. The green dashed rectangle marks the transition region. The orange curve r_c indicates the 1D-RC and the blue line segments r'_c is its piecewise linear approximation. **(b)** Flow chart for the tentative roadmap for the physical principle that connects the structure of a protein with its function.

Challenges in identifying RCs. The first step towards this goal is to identify the less strictly defined multiple RCs, since it is easier than identifying the 1D-RC. However, it turned out to be a daunting challenge. Despite much efforts, over the past two decades there have been only a few successful cases where RCs that can determine committor were identified³⁰⁻³³. In each case, two or three RCs were successfully identified. The most important lesson from these examples is that RCs are often counter-intuitive^{30,31}. This counter-intuitiveness demonstrates why it is challenging to identify the few RCs out of practically infinitely many potential candidates in a complex molecule. For example, over 6,000 candidates were tested in ref.³¹ for the $C_{7eq} \rightarrow \alpha_R$ isomerization of an alanine dipeptide in solution to uncover the three RCs.

As a result, systematic methods that go beyond intuition-based trial-and-error experimentation are required, as it is not feasible to enumerate the enormous number of potential candidates. The first systematic method used neural network and successfully identified the RCs for the $C_{7eq} \rightarrow \alpha_R$ isomerization of an alanine dipeptide in solution that eluded human intuitions³¹. This early success is followed by many machine-learning methods along similar lines, a research direction that is currently attracting intensive attention but also facing significant challenges^{25, 34-38}. More importantly, machine-learning methods cannot answer the question central to the goal of studying RCs: Why do RCs exist and how do they control functional dynamics? Answering this question requires the correct mechanical model of functional dynamics, which machine-learning cannot provide because it can only optimize parameters based on a pre-formulated model but cannot construct a model de novo. An approach based on fundamental physics is required.

Physics-based approach for understanding RCs. The recently developed energy-flow theory offered some insights into the physical principle behind RCs^{32, 39}. It showed that energy flows from the fast to the slow coordinates during a reaction³². Because RCs are the slowest, they carry the highest energy flows. This picture is intuitively appealing and explains why RCs are so important: energy is the currency of dynamics and movements of RCs incur the highest cost, thus they control reaction dynamics by dictating the overall cost. However, this picture is still semi-quantitative and does not answer a critical question: What is the relationship between different RCs?

This question is automatically answered if we can identify the 1D-RC, because it is a function of the multiple RCs, which defines the precise relationship between them. Moreover, 1D-RC answers a key conceptual question: What is the physical origin of collectivity in functional dynamics? It is generally assumed that functional dynamics are controlled by one or a few CVs, the reason that CVs are widely used in enhanced sampling. However, the critical questions have never been answered: Why functional dynamics are collective and what determine the CVs? By definition, 1D-RC is the optimal CV, thus the physical determinant of 1D-RC is the physical origin of collectivity in functional dynamics.

In this paper, we introduced a fundamental mechanical operator rooted in Newton’s law, the generalized work functional (**GWF**), which determines the correct 1D-RC. Applying this new approach to a prototypical biomolecular isomerization process, the $C_{7eq} \rightarrow C_{7ax}$ transition of an alanine dipeptide in vacuum, we obtained, as an accurate approximation, the piecewise linearization of the 1D-RC (e.g. r'_c in the example of Fig. 1a) throughout the entire transition region (i.e. $p_B \in (0, 1)$). Each linear segment is a combination of four component RCs: two dihedrals and two improper dihedrals. This 1D-RC can predict the committor value throughout the entire range of p_B with an accuracy that approaches the limit of the numerically evaluated committor values. This accuracy far exceeds what was achieved by previous methods^{30, 31}. Among the four component RCs, the two improper dihedrals were only identified by the current method; their inclusion significantly improved the accuracy in predicting the committor value.

Most importantly, the GWF is a generic physical property universal to all protein molecules, allowing us to answer fundamental questions concerning RCs and functional protein dynamics. 1) Why do RCs exist and how do they control protein dynamics? 2) What is the relationship between different RCs? 3) What is the origin of collectivity in protein dynamics?

The GWF summarizes the mechanical effects of the couplings between different coordinates. It shows that, due to the special structural features of a protein, a small number of essential forces have high impacts on the dynamics of the entire system, while the other forces have only minor effects. The directions of these essential forces are the RCs. Motions of RCs control these essential forces, thus RCs control the reaction dynamics. The GWF is a tensor; its inherent structure can be revealed by singular value decomposition (**SVD**). The leading singular vector from SVD identifies the single dominant force; its direction is the 1D-RC, which is a function of all the RCs. The motion of 1D-RC is a synergized motion of all the RCs. This synergy is the origin of collectivity in functional dynamics, as dynamics of RCs is the crux of protein dynamics.

From these results emerges a tentative roadmap that connects the structure of a protein to its function (Fig. 1b). Protein structure determines both atom-atom interactions and the structural coupling tensor responsible for the mechanical couplings between different coordinates. Together, they determine the GWF, which determines the directions of the strong forces that appear as the

RCs. Finally, RCs control the functional dynamics, which determines function. In this roadmap, the GWF plays a pivotal role: it is the underlying “mechanical structure” that determines the behaviors of the geometric structure we observe.

Results

Energy-Flow Theory. A major theoretical tool we use for analyzing reaction dynamics is the energy flow theory^{32, 39}. It was motivated by a simple physical intuition: motions of the most important coordinates in a dynamic process require high energy cost. In the Lagrange-Hamiltonian formulation of classical mechanics, energy is the cost function of motion. If the motion of a coordinate requires high energy cost, then it has a slow time scale and controls the rate of a reaction process. To identify these important coordinates, we need a rigorous definition of the energy cost of the motion of a coordinate. This is challenging because the energy function of a system is dominated by complex coupling terms between different coordinates; there is no well-defined energy per coordinate. Therefore, we can only find the energy cost of the motion of a coordinate by projection. This is achieved through integration over partial differentials.

The total differential of a multi-variable function $f(x_1, \dots, x_n)$ is $df = \sum_{i=1}^n \frac{\partial f}{\partial x_i} dx_i$, thus each partial differential $\frac{\partial f}{\partial x_i} dx_i$ is the exact contribution of the change in x_i to the total change in $f(x_1, \dots, x_n)$. The coupling terms in $f(x_1, \dots, x_n)$ are precisely partitioned among all the variables through partial derivatives $\frac{\partial f}{\partial x_i}$. Consequently, $\int \frac{\partial f}{\partial x_i} dx_i$ rigorously defines the contribution to the accumulated change in $f(x_1, \dots, x_n)$ from changes in x_i .

Applying this idea to the potential energy of the system, we define the potential energy flow through a coordinate q_i as the work on q_i ³²:

$$\Delta W_i(t_1, t_2) = - \int_{q_i(t_1)}^{q_i(t_2)} \frac{\partial U(\vec{q})}{\partial q_i} dq_i \quad (1),$$

where $U(\vec{q})$ is the potential energy of the system, $\vec{q}$ is the position vector of the system in the configuration space. According to Eq. (1), $\Delta W_i(t_1, t_2)$ is the change in $U(\vec{q})$ caused by the motion of q_i alone along a dynamic trajectory in the time interval $[t_1, t_2]$. It is a projection of the change

in the total potential energy onto the motion of q_i and a measure of its cost. Kinetic energy flow is not used in the current analysis and is not discussed here ³⁹.

To gain mechanistic insights, we need to look at how a mechanical quantity $A(\Gamma)$ (e.g. potential energy flow) change as a reaction progresses. We first project $A(\Gamma)$ onto a quantity $\xi(\Gamma)$ that parameterizes the progress of a reaction, which we call the projector, then average over the ensemble of reactive trajectories:

$$\langle \delta A(\xi^*) \rangle = \frac{\int d\Gamma \rho(\Gamma) \delta A(\xi(\Gamma) \rightarrow \xi(\Gamma) + d\xi) \delta(\xi(\Gamma) - \xi^*)}{\int d\Gamma \rho(\Gamma) \delta(\xi(\Gamma) - \xi^*)}$$

$$\langle \Delta A(\xi_1 \rightarrow \xi_2) \rangle = \int_{\xi_1}^{\xi_2} \langle \delta A(\xi) \rangle \quad (2)$$

Here, $\rho(\Gamma)d\Gamma$ is the probability of finding the system in an infinitesimal volume $d\Gamma$ around a phase-space point Γ in the reactive trajectory ensemble; $\delta(x)$ is the Dirac δ -function; $\delta A(\xi(\Gamma) \rightarrow \xi(\Gamma) + d\xi)$ is the change in A_i in a differential interval $[\xi(\Gamma), \xi(\Gamma) + d\xi]$; $\langle \Delta A(\xi_1 \rightarrow \xi_2) \rangle$ is the change in A in a finite interval $[\xi_1, \xi_2]$; ΔA can be ΔW_i ^{32, 39}. The ensemble of reactive trajectories consists of trajectories that cover the transition period of a reaction process but exclude the waiting period ³.

Generalized Work Functional. The source of complexity of protein dynamics is the highly entangled mechanical couplings between different coordinates. The potential energy flow through a coordinate q_i provides an exact account of the energetic cost of its motion but does not account for the effects of the mechanical couplings between different coordinates. These mechanical couplings are a consequence of the mathematical structure of Lagrange's equation.

The Lagrange's equation for a generalized coordinate q_i is: $\frac{d}{dt} \left(\frac{\partial L}{\partial \dot{q}_i} \right) = \frac{\partial L}{\partial q_i}$, where L is the system Lagrangian $L = K - U$, defined as the difference between the kinetic energy $K = \frac{1}{2} \sum_{j,k=1}^N \mathbb{S}_{ij} \dot{q}_j \dot{q}_k$ and the potential energy U . Here, $\mathbb{S}$ is the structural coupling tensor: $\mathbb{S}_{ik} = \sum_{\alpha=1}^N m_{\alpha} \frac{\partial x_{\alpha}}{\partial q_i} \frac{\partial x_{\alpha}}{\partial q_k}$, where the sum is over all the N coordinates in the system; m_{α} and x_{α} are the mass and a Cartesian coordinate of atom α . After expanding all the terms, we obtain:

$$F_i = -\frac{\partial U}{\partial q_i} = \sum_{k=1}^N (\dot{S}_{ik}\dot{q}_k + S_{ik}\ddot{q}_k) + \frac{1}{2} \sum_{j,k=1}^N \frac{\partial S_{jk}}{\partial q_i} \dot{q}_j \dot{q}_k \quad (3).$$

On the left-hand side is the force induced by dq_i , the infinitesimal movement of q_i . The right-hand side involves the velocities ($\dot{q}_k, k \in [1, N]$) and accelerations ($\ddot{q}_k$) of all the coordinates in the system. This means F_i directly affects the movements of all the coordinates in the system. The effects of F_i on q_k and the effects of F_k on q_i are determined by the values of S_{ik} and $\dot{S}_{ik}$, which are determined by the time-dependent system configuration. In contrast, the Lagrange's equation for a Cartesian coordinate x_α is the same as the Newton's equation: $F_\alpha = m_\alpha \ddot{x}_\alpha$; F_α directly affects only the motion of x_α , not any other coordinate. This is because $S_{\alpha\beta} = m_\alpha \delta_{\alpha\beta}$ in Cartesian coordinates, where $\delta_{\alpha\beta}$ is the Kronecker- δ .

Because F_i directly affects motions of all coordinates, it is important to have a measure of the impact of F_i on the motion of q_j ($j \neq i$). One choice is $\langle F_i dq_j \rangle$ ($i \neq j$); $\langle \dots \rangle$ is the average over the reactive trajectory ensemble. The physical motivation is that, if F_i strongly affects the motion of q_j , $\langle F_i dq_j \rangle$ should have a high value. Otherwise, it will be small. Accordingly, we define the generalized work of F_i on q_j as:

$$\mathbb{W}_{ij}(\xi; \xi_0 \rightarrow \xi_1) = \int_{\xi_0}^{\xi_1} \langle F_i(\xi) dq_j(\xi) \rangle \quad (4),$$

which accumulates the impact of F_i on the motion of q_k as the reaction progresses from ξ_0 to ξ_1 . The collection of all the $\mathbb{W}_{ij}(\xi)$ is a tensorial functional $\mathbb{W}(\xi)$ in the configuration space, which we call the generalized work functional because $F_i dq_j$ ($i \neq j$) is a generalization of the concept of mechanical work. For a specific value ξ_1 , $\mathbb{W}(\xi)$ is a tensor $\mathbb{W}(\xi; \xi_0 \rightarrow \xi_1)$, which we call the generalized work tensor.

With this definition, $\mathbb{W}(\xi)$ summarizes the accumulated effects of the pairwise mechanical couplings between different coordinates as a function of the progress of a reaction. Because $\mathbb{W}_{ij} \neq \mathbb{W}_{ji}$, SVD is the proper tool to extract its inherent structure: $\mathbb{W}(\xi; \xi_0 \rightarrow \xi_1) = \sum_{i=0}^{N-1} \sigma_i \vec{u}_i \cdot \vec{v}_i^T$, where σ_i is the i -th singular value, $\vec{u}_i$ is the i -th left and $\vec{v}_i^T$ is the transpose of the i -th right singular vector. The leading term in this decomposition, $\sigma_0 \vec{u}_0 \cdot \vec{v}_0^T$, is the dominant component of

$\mathbb{W}(\xi; \xi_0 \rightarrow \xi_1)$. Accordingly, $\vec{F}_0 = -\partial U / \partial \vec{u}_0$ is the force in the direction of $\vec{u}_0$. This force has the largest overall effects on the motions of individual coordinates in the system. Because the 1D-RC is generally considered the “driving force” of a reaction process³, it is intuitively enticing to contemplate that it may coincide with $\vec{u}_0$. If this is the case, we have also found in GWF the physical determinant of the 1D-RC.

An Example of Biomolecular Conformational Change. To test if the GWF can determine the 1D-RC, we applied it to the $C_{7eq} \rightarrow C_{7ax}$ isomerization of an alanine dipeptide in vacuum. This process is a prototype of conformational dynamics of proteins because alanine dipeptide is the smallest example of complex molecules. Here, we define a complex molecule as molecule whose non-RCs form a heat bath that can provide the RCs with enough energy to cross the activation barrier^{21, 27, 31, 32, 37, 40}. In contrast, small molecules need an external energy source, such as buffer gas in gas-phase and solvents in solution-phase reactions. The isomerization of an alanine dipeptide in vacuum carries some fundamental features that are unique to reactions in protein-like molecules but are absent in small molecules. Consequently, it has served as a standard system for testing methods for identifying RCs^{3, 30, 31}. This isomerization process is mainly a rotation around the ϕ dihedral (Fig. 2a,b). In previous studies, two backbone dihedrals ϕ and θ_1 were identified as the RCs, as they can determine the committor value with adequate accuracy^{31, 41}.

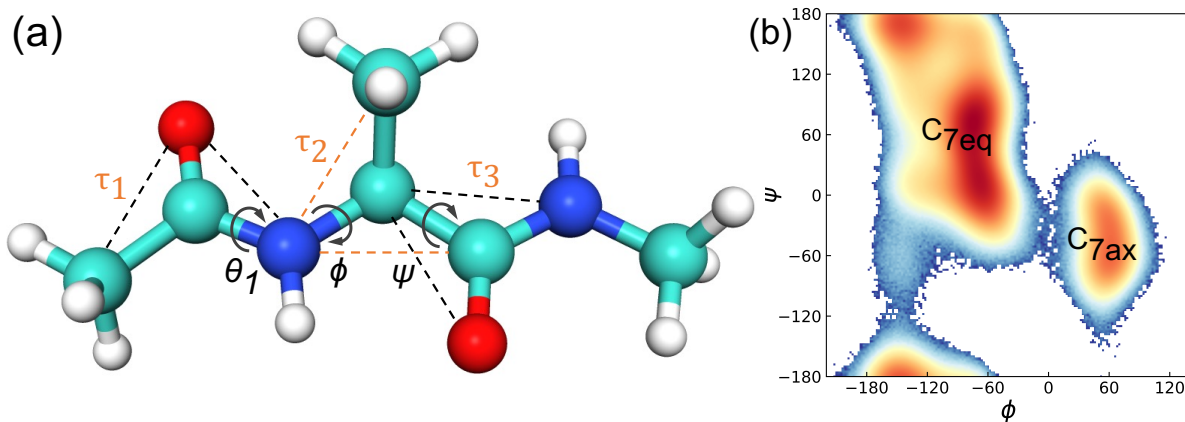


Figure 2: Schematics for the $C_{7eq} \rightarrow C_{7ax}$ isomerization of alanine dipeptide. (a) The molecular structure of alanine dipeptide. For a dihedral, its bond of rotation is marked by a curved arrow. For an improper dihedral, each of the two planes that define it is spanned by 3 atoms, with the central atom chemically bonded to the other 2 atoms. We mark each plane by connecting the 2 atoms that are not bonded to each other with a dashed line. The edge shared by the 2 planes is the bond of rotation for the improper dihedral. (b) The definition of the C_{7eq} and C_{7ax} basins in the (ϕ, ψ) -plane. The heat map is the logarithm of the joint probability $p(\phi, \psi)$ obtained from a 6 μ s equilibrium molecular dynamics simulation.

For analyses using energy flow and GWF, internal coordinates are the proper choice because they are the natural coordinates for describing protein motions and automatically satisfy all the

restraints from the bonded interactions in a protein (Fig. 3a,b). In contrast, Cartesian coordinates cannot provide useful mechanistic information because their movements are dominated by strong restraint forces from bonded interactions that bear little relevance to the mechanism of protein dynamics.

1D-RC is Determined by the GWF. Figure 3a shows the GWF for all the proper and improper dihedrals in the system computed with committor as the projector. The motion of the major RC, ϕ , is strongly affected by the forces from four coordinates: ϕ itself, the other known RC θ_1 , and two improper dihedrals τ_1 and τ_2 (Fig. 2a). Interestingly, both τ_1 and τ_2 are important players in transferring kinetic energy into ϕ based on our previous analysis of kinetic energy flows in this system³⁹. In contrast, $\mathbb{W}_{\phi\psi}(p_B)$ is significant but $\mathbb{W}_{\psi\phi}(p_B)$ is small, suggesting that F_ϕ has significant impact on the motion of ψ but F_ψ has little influence on the motion of ϕ . This suggests that the motion of ψ is slaved to the motion of ϕ because the latter is the dominant factor for determining F_ϕ ^{42, 43}. Consequently, ψ , though important for distinguishing the reactant and product states, is not an RC.

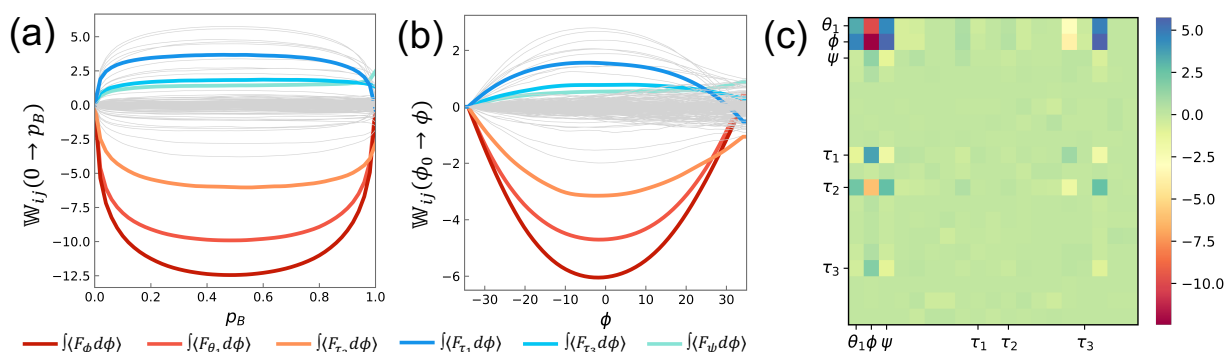


Figure 3: Generalized work functional and generalized work tensor. **(a)** The GWF with p_B as the projector for all the dihedrals and improper dihedrals in the system. There are $19 \times 19 = 361$ lines in the plot. The thick colored lines are $\int_{\phi(0)}^{\phi(p_B)} \langle F_i d\phi \rangle$ for q_i that has significant magnitude, corresponding to the blocks with distinct colors in the second column of the generalized work tensor in panel (c). The rest are colored light gray. **(b)** The GWF with ϕ as the projector. **(c)** The 19×19 generalized work tensor $\mathbb{W}(p_B; 0 \rightarrow 0.5)$ calculated for all the dihedrals and improper dihedrals in the system. Each tensor element is colored based on its value.

Figure 3c shows $\mathbb{W}(p_B; 0 \rightarrow 0.5)$, the generalized work tensor for $p_B = 0$ to 0.5 . We included all the dihedrals and improper dihedrals in constructing $\mathbb{W}(p_B; 0 \rightarrow 0.5)$, consisting of 19 coordinates in total. We did not include bonds or bond angles in $\mathbb{W}(p_B; 0 \rightarrow 0.5)$ because terms involving forces from them have high noise. This is because these forces are of high magnitude and oscillate very fast compared to forces from dihedrals.

The leading left singular vector of $\mathbb{W}(p_B; 0 \rightarrow 0.5)$ is: $\mathcal{R}_c = 0.57\theta_1 + 0.72\phi + 0.34\tau_2 - 0.21\tau_1$. To test the accuracy of $\mathcal{R}_c$, we carried out the standard “committor test” and compared the performance of $\mathcal{R}_c$ with RCs identified in previous studies^{3, 30, 31, 44}. Because RCs must be sufficient for determining p_B , configurations that share the same values of the RCs but differ in other coordinates should all have the same p_B value. To test the quality of specific RCs, we harvest an ensemble of configurations that all have the values of the RCs corresponding to the transition state, with the other coordinates randomly sampled. The distribution of p_B values for this ensemble should peak around 0.5 if the RCs can determine p_B ; a narrower distribution indicates higher accuracy of the RCs.

Figure 4a shows the results of the committor test. The orange line is the distribution of p_B for configurations with $\phi = 0^\circ$ and $\psi = 40^\circ$, corresponding to the value of ϕ at the transition state. This distribution peaks at $p_B = 0$ and $p_B = 1$, indicating that ϕ alone cannot determine committor. The green line is the distribution of p_B for configurations with $\phi = 0^\circ$ and $\theta_1 = 3.6^\circ$. This combination was identified in previous studies as sufficient for determining committor^{30, 31}. Indeed, this distribution peaks at $p_B = 0.5$, though it is rather broad, indicating either missing RCs or the uncertainty inherent in the numerical calculation of p_B ⁴⁴. This is the limit reached by previous methods, both intuition-guided trial-and-error and machine-learning with neural network^{30, 31}.

The blue line is the distribution of p_B for configurations with $\mathcal{R}_c = -80.9^\circ$. It sharply peaks at $p_B = 0.5$ and has a narrow width, showing the high accuracy of $\mathcal{R}_c$. This significant improvement in the accuracy in determining committor value compared to previous results is due to two factors: 1) the inclusion of two improper dihedrals τ_1 and τ_2 , 2) the correct coefficients in the analytical expression of $\mathcal{R}_c$. Because the effects of τ_1 and τ_2 on reaction dynamics are much weaker than those of θ_1 and ϕ , shown by their smaller coefficients in $\mathcal{R}_c$, they are much more difficult to detect, explaining why they eluded both human intuitions and machine learning^{30, 31}. The fact that they are uncovered by $\mathbb{W}(p_B; 0 \rightarrow 0.5)$ demonstrates the capability and precision of the GWF in capturing the essence of dynamics.

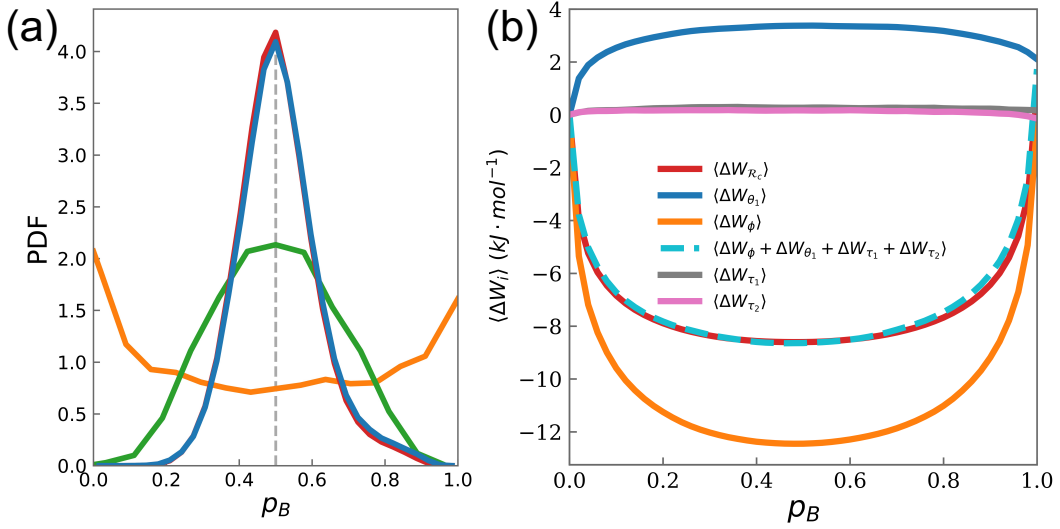


Figure 4: Committor test and potential energy flow. **(a)** Comparison of the results of committor tests for different RCs. **Orange:** PDF of p_B values of the ensemble of configurations ($\phi=0^\circ$ and $\psi=40^\circ$) with ϕ as the RC, the same as used in ref. [30]. **Green:** PDF of p_B values of the ensemble of configurations ($\phi=0^\circ$ and $\theta_1=3.6^\circ$) with both ϕ and θ_1 as the RCs, the same choice as in refs. [30, 31]. **Blue:** PDF of the ensemble of configurations with $R_c=-80.9^\circ$. **Red:** PDF of the ensemble of configurations with $R_c^*=-99.5^\circ$. **(b)** Potential energy flows through ϕ (orange), θ_1 (blue), τ_1 (pink), τ_2 (grey), and R_c (red). The cyan dashed line represents $\langle \Delta W_{\phi} + \Delta W_{\theta_1} + \Delta W_{\tau_1} + \Delta W_{\tau_2} \rangle$.

The red line is the distribution of p_B for configurations with $R_c^* = 0.57\theta_1 + 0.71\phi + 0.34\tau_1 - 0.21\tau_2 - 0.1\tau_3 = -99.5^\circ$, which is almost identical to the blue line. This suggests that the improper dihedral τ_3 is a very minor RC, consistent with its small coefficient in R_c^* .

Moreover, the matrix elements of $\mathbb{W}(p_B; 0 \rightarrow 0.5)$ have inherent numerical uncertainty, especially because p_B is numerically evaluated^{3, 30, 44}. The reason that R_c^* has minimal improvement over R_c is likely that we have already reached the numerical limit, thus R_c is likely the optimal 1D-RC. Since the GWF determines the rigorous 1D-RC, it is likely the correct physical operator for rigorous analysis of protein dynamics in general.

Figure 4b shows that $\Delta W_{R_c} \approx \Delta W_{\phi} + \Delta W_{\theta_1}$. Because ϕ and θ_1 carry the dominant majority of the total potential energy flow through the system based on previous studies^{32, 39}, this means R_c is responsible for the majority of the total potential energy flow through the system, confirming the 1D-RC as the single dominant channel for energy flows during a reaction.

Is 1D-RC Curved? An important conceptual question in reaction dynamics is whether the 1D-RC is straight or curved. There is no physical reason to expect that the 1D-RC should be straight, but in practice it is often assumed so for simplicity^{27, 28}. This assumption is likely true for small molecules because the barrier top region is very narrow. For complex molecules, this assumption is less warranted due to the extended span of the barrier top. Since alanine dipeptide is the smallest complex molecule, it offers an excellent place to examine this matter.

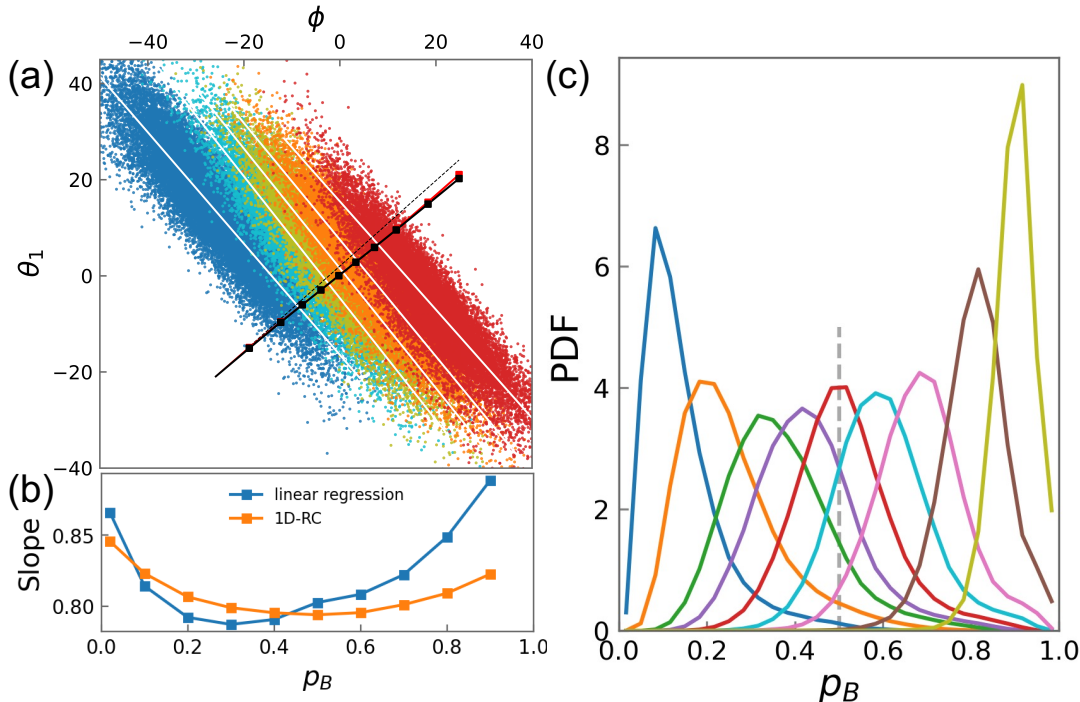


Figure 5: The curvedness of 1D-RC. **(a)** The scattered points represent projection of ensembles of configurations with $p_B = 0.02, 0.3, 0.5, 0.7, 0.9$ onto the (ϕ, θ_1) -plane. The white straight line through a colored “cloud” is its linear regressions, thus each one represents an approximation to the intersection of the corresponding iso-committor surface with the (ϕ, θ_1) -plane. The solid black line is $\mathcal{R}_c(p_B)$ constructed as a concatenation of 10 short line segments; each segment is the $\mathcal{R}_c$ for an interval of p_B . The $\mathcal{R}_c(p_B)$ for a specific p_B^* is the leading left singular vector of the tensor $\mathbb{W}(p_B; 0 \rightarrow p_B^*)$. To construct the short line segments corresponding to different values of p_B , we first prepare 11 ensembles of configurations; each one contains configurations of a specific p_B value: $p_B = 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.98$. Each ensemble of configurations are then projected onto the (ϕ, θ_1) -plane and the center of mass, c_i ($i = 1, \dots, 11$), of each ensemble is computed. The ϕ coordinate of this series of c_i defines a sequence of positions along the ϕ -direction: $\phi_i = -18.9^\circ, -12.3^\circ, -7.8^\circ, -3.9^\circ, -0.2^\circ, 3.3^\circ, 7.2^\circ, 11.7^\circ, 18.4^\circ, 24.8^\circ$. We start from c_1 as the starting point of the first line segment, with its slope as the slope of $\mathcal{R}_c(0.02)$ projected onto (ϕ, θ_1) -plane. The first segment ends at c_2 , with its ϕ -coordinate equal to ϕ_2 , and its θ_1 -coordinate equal to: $\text{slope}(\mathcal{R}_c(0.02)) \cdot (\phi_2 - \phi_1)$. The other segments are constructed in a similar way. The red solid line is constructed in the same manner as the black solid line, but the direction for each line segment is chosen as the direction normal to the linear regression of each p_B ensemble. The black dashed line is a straight line in the direction normal to the linear regression of the ensemble of $p_B = 0.9$. Both solid lines deviate from this straight dashed line, demonstrating the curvedness of the 1D-RC. The black and red solid lines are essentially coincidental, showing that $\mathcal{R}_c$ is perpendicular to the isocommittor surfaces throughout the transition region. **(b)** The slopes of the line segments in panel (a) as function of committor. **(c)** Results of committor tests for ensembles of configurations with $\mathcal{R}_c(p_B) = -89.0^\circ, -86.3^\circ, -83.9^\circ, -82.6^\circ, -80.9^\circ, -79.2^\circ, -77.5^\circ, -74.8^\circ, -71.8^\circ$, corresponding to $p_B = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$, respectively.

Because the GWF in Fig. 3a is nonlinear, the direction of the 1D-RC determined by $\mathbb{W}(p_B; 0 \rightarrow p_B^*)$ likely changes with p_B^* , which means the direction of the 1D-RC in different intervals of committor value will be different. Figure 5a (solid black line) shows the projection of the concatenation of 1D-RCs for all the intervals of p_B^* onto the (ϕ, θ_1) -plane, amounts to a piecewise linearization of the true 1D-RC. We choose this projection because it is easier to visualize in a plane, and ϕ and θ_1 are the dominant RCs. Indeed, the 1D-RC is slightly curved. This curvilinearity is quantified in Fig. 5b (orange line), which shows the slope of the 1D-RC as a function of p_B . Figure 5c shows the results of the committor test on the curved 1D-RC. Indeed, it can predict committor value with high accuracy over the entire range of p_B .

To better understand this curvilinearity of the 1D-RC, we projected onto the (ϕ, θ_1) -plane five ensembles of configurations. Each ensemble consists of configurations on the iso-committor surface of a specific $p_B^* \pm \delta$; the five ensembles correspond to $p_B^* = 0.02, 0.3, 0.5, 0.7, 0.9$; $\delta = 0.05$. It is often assumed that an iso-committor surface is a hyperplane in the configuration space^{27, 28, 35}. If this is true, each ensemble of configurations should form a straight strip with a width determined by δ in the (ϕ, θ_1) -plane. Instead, each ensemble of configurations clusters into an ellipse. To find out the intersections of each iso-committor surface with the (ϕ, θ_1) -plane, we obtain a straight line through each ellipse by linear regression. Figure 5a (white lines) shows that the direction of these iso-committor lines changes slowly with p_B^* . The red solid curve is the concatenation of the normal directions of all the iso-committor lines; it is slightly bent. Intriguingly, it essentially coincides with the 1D-RC, suggesting that the 1D-RC, though curved, is always perpendicular to the iso-committor surfaces.

The New Method is Robust and Cost-Effective. In the results above, the GWF was computed using committor as the projector. Since numerical evaluation of committor is computationally expensive⁴⁴, it could limit the use of the current method for identifying the 1D-RC in large systems. Therefore, we computed the GWF using ϕ as the projector, which is a good order parameter that can distinguish the reactant and product basins³, but insufficient for parameterizing the progression of the reaction (Fig. 4a). Since RCs control the dynamics of the transition period, not the dynamics within the reactant basin, we consider the range of $\phi \in [-35^\circ, 35^\circ]$, corresponding to the region of $p_B \in (0, 1)$.

Figure 3b shows $\mathbb{W}(\phi)$, the GWF calculated with ϕ as the projector, which differ from $\mathbb{W}(p_B)$ in Fig. 3a. This is expected because projector significantly affects the ensemble average. Importantly, the 1D-RC from $\mathbb{W}(\phi; -35^\circ \rightarrow 0^\circ)$, $\mathcal{R}_c^\phi = 0.57\theta_1 + 0.71\phi + 0.37\tau_1 - 0.19\tau_2$, is essentially identical to $\mathcal{R}_c$. Therefore, we can use an order parameter as the projector instead of the committor. This tremendously decreases the computational cost of the current method compared to previous methods^{31, 36}, as now the dominant cost is to harvest sufficient number of reactive trajectories, which is low compared to cost of computing the committor^{31, 36}. Most importantly, substituting p_B with an order parameter does not affect the accuracy of the resulting 1D-RC.

Another important factor is that the number of reactive trajectories required for obtaining the correct 1D-RC is much lower than what is required for smooth GWF curves. The results in Fig. 3 are averages over 2,000 trajectories, but the 1D-RC obtained from averaging over 200 trajectories, $\mathcal{R}_c(200 \text{ traj}) = 0.59\theta_1 + 0.70\phi + 0.32\tau_1 - 0.24\tau_2$, is effectively the same as $\mathcal{R}_c$ obtained from 2,000 trajectories. Moreover, the 1D-RC obtained from 100 trajectories is: $\mathcal{R}_c(100 \text{ traj}) = 0.47\theta_1 + 0.56\phi + 0.64\tau_1 - 0.23\tau_2$. Although it does not have the correct coefficients, it does correctly distinguish the four important coordinates, $\phi, \theta_1, \tau_1, \tau_2$, from the other coordinates in the system, which is what previous methods aimed at and provides highly valuable mechanistic insights^{31, 35, 36}. The reason behind this robustness is probably two-fold. 1) The GWF tensor is an integration of the GWF curves, thus noise in the GWF curves is canceled out by integration, leading to improved accuracy in the 1D-RC. 2) The SVD results are likely constrained by the structure of the GWF tensor and robust to numerical errors. This makes the current approach a robust and cost-effective method for identifying the 1D-RC in complex molecules.

It is also possible to apply the current method to analyzing existing long-time scale MD simulations of complex conformational dynamics, such as the datasets by D. E. Shaw research. The major requirement is that the progress of the conformational transition of interest can be characterized by an order parameter selected by the user. This transition can involve more than two stable/metastable states and the user does not have to correctly identify all the stable states from the beginning. The basic idea can be explained by a schematic example of a two-dimensional system (Fig. 6a). In this example, q_1 and q_2 are two RCs and there are three stable states, S_1, S_2

and S_3 , in the q_1, q_2 -space. The order parameter α can identify S_1 and S_3 but not S_2 from the simulation data and is the projector in GWF calculations. The key is that both the GWF $\mathbb{W}(\alpha)$ and the 1D-RC are functions of α .

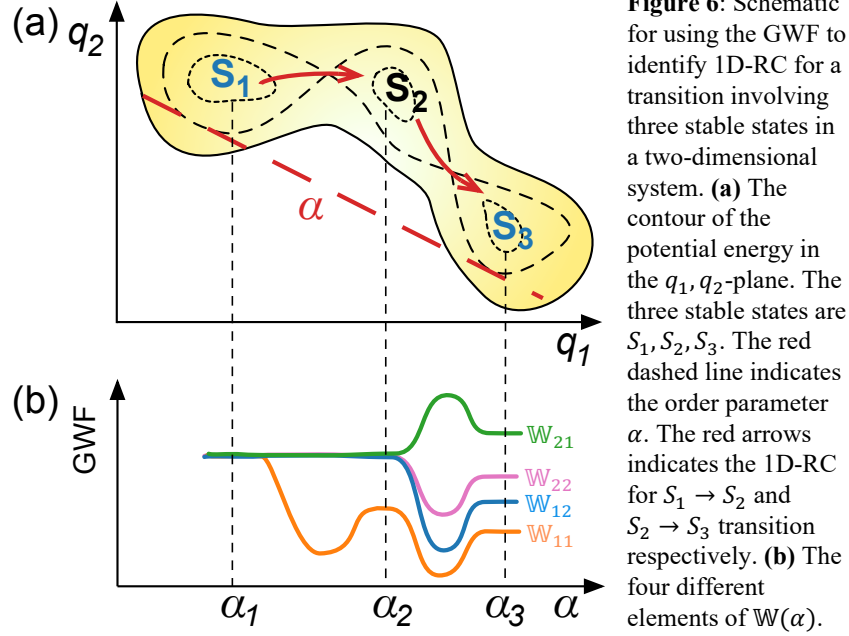


Figure 6b shows the expected features of the four elements of $\mathbb{W}(\alpha)$. Because q_1 is the only RC for $S_1 \rightarrow S_2$ transition, only $\mathbb{W}_{11}(\alpha)$ has significant magnitude. All the other terms of $\mathbb{W}(\alpha)$ are vanishingly small because q_1 and q_2 are not coupled in this region, leaving $\mathbb{W}_{12}$ and $\mathbb{W}_{21}$ essentially zero, and q_2 only undergoes thermal fluctuations, leading to complete cancelations in $\mathbb{W}_{22}$. The resulting 1D-RC will be just q_1 . In the $S_2 \rightarrow S_3$ transition, both q_1 and q_2 are RCs, thus all the elements of $\mathbb{W}(\alpha)$ are significant and the resulting 1D-RC is a linear combination of q_1 and q_2 . Consequently, we expect to observe a significant change in the direction of the 1D-RCs determined in the intervals (α_1, α_2) and (α_2, α_3) respectively, corresponding to the two red arrows in Fig. 6a. This kind of change in the 1D-RC determined at different values of the projector could be due to hidden stable states in the initial analysis, or the intrinsic nonlinearity of 1D-RC in complex systems. The actual situation could then be determined by further analysis. In the discussions above, we assume that the simulation data contain sufficient number of instances of the transition of interest so that the GWF analysis converges. Otherwise, the existing trajectory segments that contain the transition of interest can be used as seeds for TPS to proliferate more reactive trajectories.

Discussions

In this paper, we developed a general mechanical operator, the GWF, that can help to answer some fundamental questions in protein dynamics.

Collectivity in Functional Protein Dynamics. Functional protein dynamics often involve significant changes in the global protein structure, which are assumed to be controlled by collective modes. However, it remains a puzzle what these collective modes are and what their physical origin is.

A global conformational change results from aggregation of many local movements. Each local movement is described by an internal coordinate. Individual local movements are mechanically coupled together by the $\mathbb{S}$ tensor, thus they are synergized with each other, forming collective modes. Understanding the collectivity in conformational dynamics requires understanding what this synergy is and what determines it.

The GWF encapsulates the effects of the couplings induced by the $\mathbb{S}$ tensor. Its left singular vectors are the collective mode responsible for the global conformational change. The coefficients of these singular vectors define the synergy between different internal coordinates. For example, in $\mathcal{R}_c^*$, ϕ , θ_1 and τ_1 correlate with each other but anti-correlate with τ_2 and τ_3 . The GWF is the physical determinant of this synergy and the collectivity of protein dynamics.

A Tentative Roadmap for the Structure-Function Relationship of Proteins. A critical link between the structure and the function of a protein is its functional dynamics, which is controlled by the RCs. The RCs are determined by the GWF, which summarizes how the forces in the protein impact the motions of the coordinates that are extensively coupled together by the $\mathbb{S}$ tensor. Both the forces and the $\mathbb{S}$ tensor are determined by the protein structure.

Summing these factors up leads to a tentative roadmap for the structure-function relationship (Fig. 1b). The structure of a protein determines both atom-atom interactions and the $\mathbb{S}$ tensor. Together,

they determine the GWF, which determines the dominant forces in the protein and the RCs. The RCs control the functional dynamics, which determines the protein function.

Finally, the GWF provides a rigorous and computationally efficient method for identifying the 1D-RC for a single-channel reaction. For the $C_{7eq} \rightarrow C_{7ax}$ isomerization of an alanine dipeptide in vacuum, we obtained the piecewise linearization of the true 1D-RC over the entire transition region, which showed that the 1D-RC is slightly curved. The curved 1D-RC can predict the committor value with an accuracy far exceeding what was achieved before by both human intuitions and machine learning, attesting the value of rigorous physics-based method. Most importantly, the 1D-RC is determined by a general mechanical operator, without resorting to any informatics approach or human intuitions.

Methods

All simulations were performed using the molecular dynamics software suite GROMACS⁴⁵ with transition path sampling implemented. Amber 94 force field was used to facilitate comparison with previous results^{32, 39, 41, 46, 47}. The structure of the alanine dipeptide was minimized using steepest descent algorithm and heated to 300 K using velocity rescaling with a coupling constant of 0.2 ps. The system was then equilibrated for 200 ps and no constraints were applied. The time step of integration was 1 fs. Basin C_{7eq} is defined as $-190^\circ < \phi < -55^\circ$ and $-60^\circ < \psi < 190^\circ$; basin C_{7ax} is defined as $50^\circ < \phi < 100^\circ$ and $-80^\circ < \psi < 0^\circ$. We used transition path sampling method to generate the ensemble of reactive trajectories between these two basins that are used in all the analyses discussed here⁴⁰. All the averaged quantities discussed in the text were averaged over 2,000 trajectories.

Acknowledgement

We thank the National Science Foundation for financial support (award: CHE-1665104) and the anonymous reviewer for the excellent suggestions that help to improve the manuscript.

References

1. Frauenfelder, H.; Sligar, S. G.; Wolynes, P. G., The energy landscapes and motions of proteins. *Science* **1991**, *254* (5038), 1598-603.
2. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007**, *450* (7172), 964-972.
3. Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L., Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* **2002**, *53*, 291-318.
4. Kramers, H. A., Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* **1940**, *VII* (4), 284-304.
5. Grote, R. F.; Hynes, J. T., The Stable States Picture of Chemical-Reactions .2. Rate Constants for Condensed and Gas-Phase Reaction Models. *J Chem Phys* **1980**, *73* (6), 2715-2732.
6. Wigner, E., The transition state method. *Trans. Faraday Soc.* **1938**, *34*, 29-41.
7. Hanggi, P.; Talkner, P.; Borkovec, M., Reaction-Rate Theory - 50 Years after Kramers. *Rev Mod Phys* **1990**, *62* (2), 251-341.
8. Berne, B. J.; Borkovec, M.; Straub, J. E., Classical and Modern Methods in Reaction-Rate Theory. *J Phys Chem-US* **1988**, *92* (13), 3711-3725.
9. Chandler, D., Statistical-Mechanics of Isomerization Dynamics in Liquids and Transition-State Approximation. *J Chem Phys* **1978**, *68* (6), 2959-2970.
10. Pechukas, P., Statistical Approximations in Collision Theory. In *Dynamics of Molecular Collisions Part B*, Miller, W. H., Ed. Plenum: New York, 1976; Vol. 2, p 269.
11. Barducci, A.; Bussi, G.; Parrinello, M., Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys Rev Lett* **2008**, *100* (2), 020603.
12. Darve, E.; Rodriguez-Gomez, D.; Pohorille, A., Adaptive biasing force method for scalar and vector free energy calculations. *J Chem Phys* **2008**, *128* (14), 144120.
13. Hamelberg, D.; Mongan, J.; McCammon, J. A., Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. *J Chem Phys* **2004**, *120* (24), 11919-29.
14. Invernizzi, M.; Piaggi, P. M.; Parrinello, M., Unified Approach to Enhanced Sampling. *Phys Rev X* **2020**, *10* (4).
15. Kong, X. J.; Brooks, C. L., lambda-Dynamics: A new approach to free energy calculations. *Journal of Chemical Physics* **1996**, *105* (6), 2414-2423.
16. Laio, A.; Parrinello, M., Escaping free-energy minima. *Proc Natl Acad Sci U S A* **2002**, *99* (20), 12562-6.
17. Valsson, O.; Tiwary, P.; Parrinello, M., Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annu Rev Phys Chem* **2016**, *67*, 159-84.
18. Zheng, L. Q.; Chen, M. G.; Yang, W., Random walk in orthogonal space to achieve efficient free-energy simulation of complex systems. *P Natl Acad Sci USA* **2008**, *105* (51), 20227-20232.
19. Zheng, L. Q.; Chen, M. G.; Yang, W., Simultaneous escaping of explicit and hidden free energy barriers: Application of the orthogonal space random walk strategy in generalized ensemble based conformational sampling. *J Chem Phys* **2009**, *130* (23).
20. Zheng, L. Q.; Yang, W., Practically Efficient and Robust Free Energy Calculations: Double-Integration Orthogonal Space Tempering. *J Chem Theory Comput* **2012**, *8* (3), 810-823.

21. Du, R.; Pande, V. S.; Grosberg, A. Y.; Tanaka, T.; Shakhnovich, E. S., On the transition coordinate for protein folding. *J Chem Phys* **1998**, *108* (1), 334-350.
22. Ryter, D., On the Eigenfunctions of the Fokker-Planck Operator and of Its Adjoint. *Physica A* **1987**, *142* (1-3), 103-121.
23. Onsager, L., Initial recombination of ions. *Phys. Rev.* **1938**, *54* (8), 554-557.
24. Wang, Y.; Ribeiro, J. M. L.; Tiwary, P., Past-future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nat Commun* **2019**, *10* (1), 3573.
25. Mardt, A.; Pasquali, L.; Wu, H.; Noe, F., VAMPnets for deep learning of molecular kinetics. *Nat Commun* **2018**, *9* (1), 5.
26. Langer, J. S., Statistical theory of the decay of metastable states. *Ann. Phys.* **1969**, *54* (2), 258-275.
27. Berezhkovskii, A.; Szabo, A., One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *J Chem Phys* **2005**, *122* (1).
28. Ma, A.; Nag, A.; Dinner, A. R., Dynamic coupling between coordinates in a model for biomolecular isomerization. *J Chem Phys* **2006**, *124* (14), 144911.
29. Berezhkovskii, A. M.; Szabo, A., Diffusion along the splitting/commitment probability reaction coordinate. *J Phys Chem B* **2013**, *117* (42), 13115-9.
30. Bolhuis, P. G.; Dellago, C.; Chandler, D., Reaction coordinates of biomolecular isomerization. *Proc Natl Acad Sci U S A* **2000**, *97* (11), 5877-82.
31. Ma, A.; Dinner, A. R., Automatic method for identifying reaction coordinates in complex systems. *J Phys Chem B* **2005**, *109* (14), 6769-79.
32. Li, W.; Ma, A., Reaction mechanism and reaction coordinates from the viewpoint of energy flow. *J Chem Phys* **2016**, *144* (11), 114103.
33. Quaytman, S. L.; Schwartz, S. D., Reaction coordinate of an enzymatic reaction revealed by transition path sampling. *P Natl Acad Sci USA* **2007**, *104* (30), 12253-12258.
34. Frassek, M.; Arjun, A.; Bolhuis, P. G., An extended autoencoder model for reaction coordinate discovery in rare event molecular dynamics datasets. *J Chem Phys* **2021**, *155* (6), 064103.
35. Antoniou, D.; Schwartz, S. D., The stochastic separatrix and the reaction coordinate for complex systems. *J Chem Phys* **2009**, *130* (15), 151103.
36. Peters, B.; Trout, B. L., Obtaining reaction coordinates by likelihood maximization. *J Chem Phys* **2006**, *125* (5), 054108.
37. Li, W.; Ma, A., Recent developments in methods for identifying reaction coordinates. *Mol Simul* **2014**, *40* (10-11), 784-793.
38. Noe, F.; Olsson, S.; Kohler, J.; Wu, H., Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365* (6457).
39. Li, H.; Ma, A., Kinetic energy flows in activated dynamics of biomolecules. *J. Chem. Phys.* **2020**, *153* (9), 094109.
40. Bolhuis, P. G.; Chandler, D.; Dellago, C. and Geissler, P. L., Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Ann. Rev. Phys. Chem.* **2002**, *53*, 291-318.
41. Bolhuis, P. G.; Dellago, C.; Chandler, D., Reaction coordinates of biomolecular isomerization. *P Natl Acad Sci USA* **2000**, *97* (11), 5877-5882.

42. Frauenfelder, H.; Chen, G.; Berendzen, J.; Fenimore, P. W.; Jansson, H.; McMahon, B. H.; Strope, I. R.; Swenson, J.; Young, R. D., A unified model of protein dynamics. *P Natl Acad Sci USA* **2009**, *106* (13), 5129-5134.
43. Frauenfelder, H.; Fenimore, P. W.; Chen, G.; McMahon, B. H., Protein folding is slaved to solvent motions. *Proc Natl Acad Sci U S A* **2006**, *103* (42), 15469-72.
44. Dellago, C.; Bolhuis, P. G.; Geissler, P. L., Transition path sampling. *Adv Chem Phys* **2002**, *123*, 1-78.
45. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J Chem Theory Comput* **2008**, *4* (3), 435-447.
46. Case, D. A.; Pearlman, D. A.; Caldwell, J. W.; Cheatham III, T. E.; Ross, W. S.; Simmerling, C. L.; Darden, T. A.; Merz, K. M.; Stanton, R. V.; Cheng, A. L., Amber 5.0. *University of California, San Francisco* **1997**.
47. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P., Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput Phys Commun* **1995**, *91* (1-3), 1-41.

Structure-Function Relation

