MARCH 01 2022

Quantifying biomolecular diffusion with a “spherical cow” model

Frederico Campos Freitas; Sandra Byju; Asem Hassan; Ronaldo Junio de Oliveira; Paul C. Whitford

American Journal of Physics 90, 225–238 (2022)
https://doi.org/10.1119/5.0075952
CHORUS

The Computational Physics Section publishes articles that help students and instructors learn about the computational tools used in contemporary research. Interested authors are encouraged to send a proposal to the editors of the Section, Jan Tobochnik (jant@kzoo.edu) or Harvey Gould (hgould@clarku.edu). Summarize the physics and the algorithm you wish to discuss and how the material would be accessible to advanced undergraduates or beginning graduate students.

Quantifying biomolecular diffusion with a "spherical cow" model

Frederico Campos Freitas,[1,a)] Sandra Byju,[2,b)] Asem Hassan,[2]
Ronaldo Junio de Oliveira,[1,c)] and Paul C. Whitford[2,d)]
[1]*Laboratório de Biofísica Teórica, Departamento de Física, Instituto de Ciências Exatas, Naturais e Educação, Universidade Federal do Triângulo Mineiro, Uberaba, MG, Brazil*
[2]*Department of Physics and Center for Theoretical Biological Physics, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115*

(Received 20 October 2021; accepted 20 October 2021)

The dynamics of biological polymers, including proteins, RNA, and DNA, occur in very high-dimensional spaces. Many naturally occurring polymers can navigate a vast phase space and rapidly find their lowest free energy (folded) state. Thus, although the search process is stochastic, it is not completely random. Instead, it is best described in terms of diffusion along a downhill energy landscape. In this context, there have been many efforts to use simplified representations of the energetics, for which the potential energy is chosen to be a relatively smooth function with a global minimum that corresponds to the folded state. That is, instead of including every type of physical interaction, the broad characteristics of the landscape are encoded in approximate energy functions. We describe a particular class of models, called structure-based models, that can be used to explore the diffusive properties of biomolecular folding and conformational rearrangements. These energy functions may be regarded as the spherical cow for modeling molecular biophysics. We discuss the physical principles underlying these models and provide an entry-level tutorial, which may be adapted for use in curricula for physics and non-physics majors.
https://doi.org/10.1119/5.0075952

I. INTRODUCTION

When studying a complex system, physicists will typically begin by proposing a highly simplified model that includes a few relevant properties of the system. The broad utilization of this strategy inspired the well-known joke in the physics community regarding a spherical cow, with several examples of this approach immortalized by a book entitled by the joke.[1] That is, when studying a cow, a physicist's first approximation is to represent the cow by a sphere of uniform mass and charge density. Starting with this spherical cow, physicists will then investigate the properties of the simplified system before considering additional details. By iteratively introducing new features, complex physical systems can be understood at ever-increasing levels of detail. In contrast to this approach, traditional biological studies aim to provide broad characterizations (e.g., structures and rates) of detailed systems (e.g., molecules in a cell). Thus, at first glance, it may not be obvious how physicists can effectively apply the spherical cow philosophy to biology.

In the following, we will discuss a spherical cow approach to studying molecular biophysics. Specifically, we will explain the ideas behind a class of potential energy functions called structure-based models.[2–4] These models exploit the phenomenological features of biomolecules to provide a simplified version of the energetics. To understand the value of these models, it is necessary to recognize that molecular biology techniques can provide only atomic-resolution descriptions of long-lived stable structures of biomolecules. Accordingly, these configurations must correspond to deep (at least several k_BT) free energy minima. Inspired by this simple observation, structure-based models explicitly define experimental configurations to be stable. That is, the baseline versions of these models do not aim to identify the factors that impart stability. Rather, interactions formed in the native (ground state) configurations are defined to be stabilizing, and all other interactions are treated as repulsive, which ensures that the spatial arrangements are preserved. Given the crude character of the models, it may be surprising that these simplified representations have been able to model a broad range of biomolecular processes, ranging from protein folding[2,5–7] to the dynamics of protein synthesis by the ribosome.[8–10]

We first provide a brief introduction to molecular biology for physics students, followed by a description of simulation techniques and structure-based models. We additionally discuss example calculations that can be adopted and integrated in advanced undergraduate or graduate-level physics courses. Our intent is to provide students (and instructors) with a basic understanding of the biological context and physical principles. To facilitate the adoption of this material, we provide a repository with step-by-step instructions on how to apply the models to simulations.

II. MOLECULAR BIOLOGY: A PRIMER FOR PHYSICS STUDENTS

Because most physics students have limited exposure to biological systems, we first provide some biochemical and physical–chemical background that is necessary for understanding the physical principles that govern biology. We will focus on biological polymers, including proteins and nucleic acids. Due to the intimate relationship of structure and function, we will discuss both chemical composition and empirically determined structural properties. Although our discussion can be found in standard biochemistry texts, this overview allows for a more focused entry into the simulation of biopolymers.

A. Protein structure

Proteins have many roles in the cell, including providing structural integrity, executing chemical reactions, signaling, and regulating gene expression. A protein is a polymer that is formed by a sequence of amino acid residues. Each amino acid or residue [see Fig. 1(a)] is composed of a common amino group (NH_2), carbon (C_α), and carboxyl group (CO_2), while the "side chain" (usually denoted by the letter R) differs for each type of residue. There are 20 naturally occurring amino acids, each defined by the composition of the R group. Each amino acid is linked to the next residue by the formation of a peptide bond, such that a single protein chain is called a "polypeptide." When describing protein structures, the N-terminal end is considered the "beginning" of the chain, and the end of the chain is the C-terminal tail. If we exclude the side chain atoms, we can define the "backbone" of the chain by the repeating set of common atoms. Because the backbone atoms are common to each residue, amino acids are generally classified based on the side chain (R) composition, where they can be acidic, basic, uncharged polar, and non-polar (hydrophobic).

There are several dominant classical forces that describe the energetics of proteins. Along with the covalently linked backbone atoms, there are weaker non-covalent interactions that can be formed between all atoms in the chain. The four types of non-covalent interactions that are relevant in biomolecules are electrostatic, hydrogen bond, van der Waals, and hydrophobic interactions. Charged species interact via long-range Coulomb interactions. Hydrogen bonds are directional dipole–dipole interactions that can be formed between H atoms bound to highly electronegative atoms (donors), such as N, O, or F, with other electronegative atoms (acceptor atoms). The van der Waals interaction accounts for the excluded volume due to the exclusion principle as well as a net attractive force due to instantaneous dipole–dipole interactions. Hydrophobic interactions describe the way by which non-polar hydrophobic (water-repelling) residues favor aggregation to minimize exposure to the polar solvent environment.

To systematically describe a protein, it is necessary to decompose its structure into multiple tiers. At the most basic level, the primary structure is used to define the sequence of amino acids that are present in a single chain, such as the four-amino acid polypeptide in Fig. 1(a). Local structure formation in short segments (typically 10–20 residues) is called the secondary structure, where the two major structural motifs are the α helix and β sheet [see Figs. 1(b) and 1(c)]. In an α helix, the polypeptide twists to form a right-handed helical structure that is stabilized by hydrogen bonds formed along the protein backbone. To aid the inspection of the structure, graphical software will typically display these regions as helical ribbons [Fig. 1(b)]. The second common structural motif is the β sheet, which is formed when two or more spatially adjacent segments of the polypeptide chain align in a parallel or anti-parallel arrangement. These elements are often shown as aligned arrows [Fig. 1(c)]. At a higher level of organization is the tertiary structure of a protein, which typically involves the spatial organization of

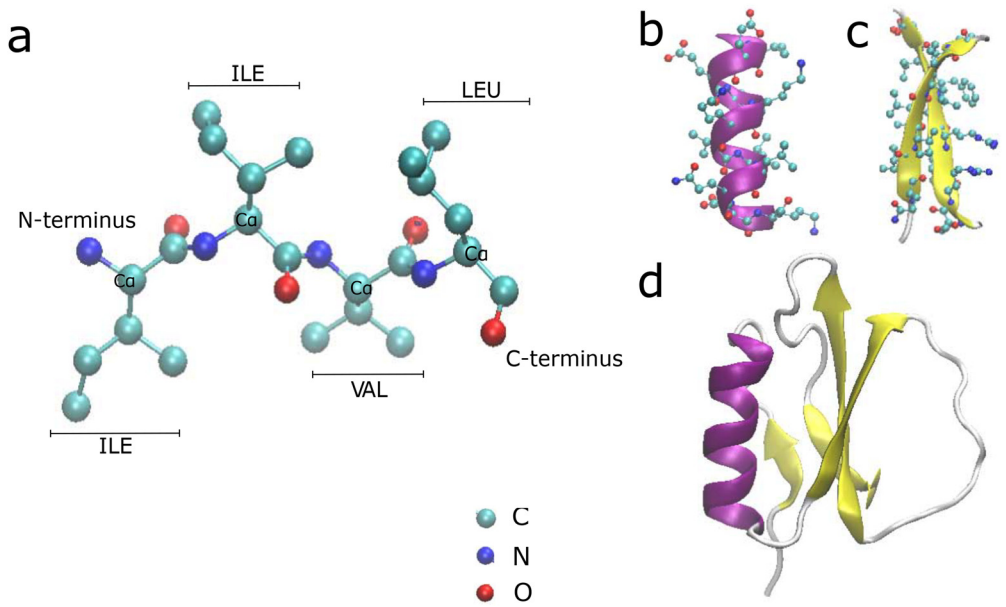

Fig. 1. Protein structure: (a) A polypeptide composed of four amino acid residues; the N- and C-terminal ends are marked. Shown is a ball and stick representation of all non-hydrogen atoms (color coded), with bonds shown as thin cylinders. (b) An α helix of protein chymotrypsin inhibitor (CI2) with helical ribbons representing the polypeptide backbone and side chains shown explicitly using a ball and stick representation. (c) A β sheet of CI2 with aligned arrows representing the polypeptide backbone and side chains shown in a ball and stick representation. (d) Tertiary structure of CI2. Cartoon representation of CI2 in its folded conformation, with α helix (purple) and β sheets (yellow) highlighted.

numerous secondary structure elements [Fig. 1(d)]. When a portion of the chain assembles into an autonomous unit, these elements are often called "domains." The final level of organization is the quaternary structure, which refers to the assembly and association of multiple peptide chains.

B. Structural properties of nucleic acids (RNA and DNA)

The second class of biomolecules that we will discuss are nucleic acids, which are formed by a string of nucleotide monomers. A nucleotide is composed of a backbone pentose sugar, where the carbon atom in the 5′ position is linked to a phosphate group, and a nitrogenous base is covalently bonded to the 1′-carbon atom through a N-glycosidic linkage [Figs. 2(a) and 2(b)]. Nucleic acid carbons belonging to sugar rings have the prime symbol added to their numbers to differentiate them from nucleobase ring carbons. In contrast to proteins, which carry their +1 or −1 charge on the side chain, the nucleic acid backbone has a negative charge that arises from the phosphate group (PO_4^-). As for proteins, the sequence of nucleic acid residues defines the polymer. On each residue, there is a nitrogenous base, which is either a purine (adenine, guanine [Figs. 2(c) and 2(d)]), or a pyrimidine (cytosine, uracil, thymine [Figs. 2(e)–2(g)]). The two major classes of nucleic acids (RNA and DNA) are classified by the presence/absence of a single oxygen atom on the 2′ carbon in the backbone of each residue [Figs. 2(a) and 2(b)].

DNA typically forms the well-known double-stranded helix, where hydrogen bonds are formed between complementary bases in the two chains. This stabilizing energy imparted by Watson–Crick base pairing is in addition to that associated with the stacking of adjacent bases (hydrophobic and van der Waals forces). The most common DNA structure is the right-handed double helix (B form).

In contrast to DNA, RNA molecules adopt a wide range of stable conformations. With this extended versatility, they can contribute to functional conformational dynamics in the cell by serving as biomolecular machines (e.g., the ribosome) or performing enzyme activity (e.g., ribozymes). Although all RNA molecules have similar components, their functional roles have led to the introduction of many different names.

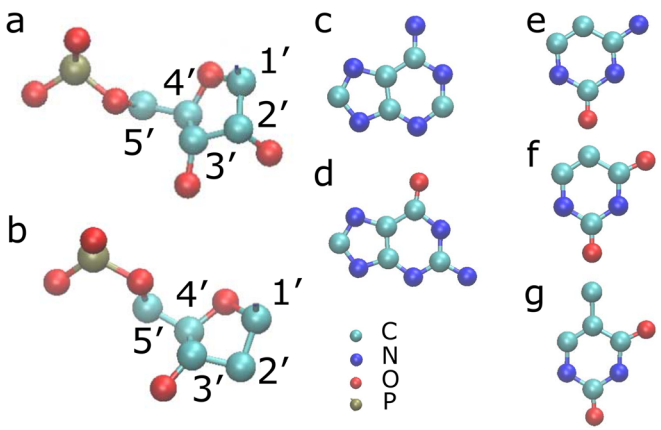

Fig. 2. Nucleic acids. Each nucleotide is composed of a pentose sugar with phosphate group attached to 5′ carbon and nucleobase attached to the 1′ carbon: (a) RNA nucleoside with ribose sugar and a phosphate group connected to the 5′ carbon. (b) DNA nucleoside with deoxyribose sugar and a phosphate group connected to the 5′ carbon. Purine nucleobases are adenine (c) and guanine (d). Pyrimidine nucleobases are cytosine (e), uracil (f), and thymine (g); uracil is unique to RNA as thymine is to DNA.

For example, RNA associated with gene expression is often called mRNA, and RNA present in the ribosome is called rRNA. Regardless, RNA can exist in isolation (i.e., individual chains), where hydrogen bonds and Coulomb forces stabilize a range of structural motifs, including the RNA double helix (A form), hairpin loops, internal loops, bulges, and junctions. Together, these structural elements are referred to as the secondary structure of an RNA molecule. These secondary structure elements arrange to form functional tertiary conformations. In addition to RNA–RNA interactions, the large negative charge of RNA chains leads to a strong dependence of RNA structure on metal ions, including Mg^{2+} and K^+.

C. Protein synthesis, folding, and assembly

DNA, RNA, and proteins are the key biomolecular components that define molecular biology. That is, genetic information is coded in DNA, which is then transcribed to mRNA sequences. Although the transcription of mature mRNA in prokaryotes (single-cell organisms) is performed by RNA polymerase, in eukaryotes (multi-cellular organisms), the mRNA can be further modified through a range of processes, such as splicing. In both cases, the mature mRNA sequence is read and translated into a protein sequence by the ribosome. After the ribosome produces a new protein, the protein must then fold to carry out any given biological purpose.

The most common way to begin to think about protein folding is to begin with what is known as Levinthal's paradox.[11] According to this "paradox," a random search to find the folded conformation would require the age of the universe. For example, for a polypeptide of 100 amino acid residues, with each residue having two allowed conformations, there would be a total of 2^{100} possible conformations. If the random sampling of conformations occurred every picosecond, it would require roughly 10^{10} years for a single protein to fold. However, protein folding occurs many orders of magnitude faster, generally between microseconds and seconds.

To reconcile the apparent paradox, it was recognized that the energy landscapes of proteins must not be random or flat. Instead, the energy landscapes may be thought of as being funnel-shaped,[12] where the principle of minimal frustration[13–16] indicates that there is a lack of large-scale energetic traps. In this framework, protein folding may be described as a diffusive process, where the protein moves along the landscape in the direction of the global minimum. Due to the presence of a large energy gap between the native and unfolded conformations relative to the scale of the energetic roughness, the dynamics in these funneled landscapes yields time scales that are consistent with the dynamics in the cell.

III. THEORETICAL MODELS AND COMPUTATIONAL METHODS

Although experimental molecular biology techniques can determine the atomic structures of stable structures, describing their dynamics requires an understanding of energetics. From a theoretical/computational perspective, we need to specify how atoms within a biomolecule interact, typically by constructing a suitable potential energy function. Once a potential energy function (force field) is defined, we use numerical techniques to evaluate the associated kinetic and thermodynamic properties of the system. The two most

widely used computational techniques are Monte Carlo and molecular dynamics (MD) simulations. For an introduction, see Ref. 17. We will focus on MD techniques, because they are more widely used to study the kinetics and equilibrium distributions in proteins. We first describe the physical principles behind MD simulations, followed by a discussion of the types of force fields that may be applied in simulations of biomolecules.

A. Molecular dynamics simulations

Molecular dynamics simulations involve numerically integrating Newton's equations of motion, using a discrete time step τ (generally femtosecond scale). At each step, the molecular forces are calculated, the coordinates and velocities are incremented in time, and the process is then repeated.[17] In its simplest form, we can use the positions and velocities at time t, along with the force ($\mathbf{F} = -\nabla U$) to determine the positions and velocities at time $t+\tau$ by the relations,

$$\mathbf{X}(t+\tau) = \mathbf{X}(t) + \mathbf{V}(t)\tau \tag{1}$$

$$\mathbf{V}(t+\tau) = \mathbf{V}(t) + \frac{\mathbf{F}(t)}{m}\tau. \tag{2}$$

Although higher-order approximations, such as the Verlet algorithm,[18] are typically applied in research settings, all MD simulations share this approach for determining the time sequence of configurations. By using this simple approach, a trajectory of the motion can be obtained to probe functional conformational transitions,[19] ligand binding,[20] subunit association,[21] and folding transitions.[22,23] We will focus on protein folding/unfolding transitions to illustrate the methods and physical analyses that are available.

To integrate Newton's equations of motion, the potential energy function must be defined. The potential energy function contains terms that specify the nature of the interactions between bonded and non-bonded atoms. Bonded interactions are typically associated with pairs of atoms, triplets that form angles, or quartets that form dihedral angles. These bonded terms define the covalent bond geometry of the biomolecule and approximate the vibrational properties that arise from covalent interactions.

Interactions between non-covalently bonded atoms can be divided into several broad classes. Steric repulsion arises from atomic exclusion, where two atoms are not allowed to overlap their electronic densities. In classical MD simulations, this effect can be described by many possible functional forms, although a $1/r^{12}$ relation is most commonly applied. Another major contributor is electrostatic interactions. Because classical MD simulations do not explicitly describe electrons, partial charges are typically assigned to each atom, and the precise values are intended to reflect the associated (average) electronic densities. These charges interact via Coulomb potentials, screened-Coulomb potentials, or other implicit-solvent representations. In addition to electrostatics, dispersion forces can also lead to attraction between atoms. These effects are commonly approximated by a $1/r^6$ dependence. As we have described, hydrogen bonds can also occur, where proton-mediated interactions are formed between two highly electronegative atoms. Finally, base–stacking interactions are generally attractive in RNA and DNA and provide a critical contribution to overall molecular stability.

A particularly important factor that influences the structure of biomolecules is the nature of solvent (water) interactions. Sometimes, structural water molecules can bind to proteins and form stabilizing hydrogen bonds with acceptor groups. In addition, the solvent can mediate hydrophobic interactions, which are entropic in nature. The hydrophobic effect is due to hydrophobic residues being sequestered from the solvent, which leads to an increase in the configurational entropy of the solvent. We will perform simulations with an implicit-solvent model for which the impact of the solvent is included by an effective representation.

After defining the functional form of different terms in the potential energy function, parameters have to be specified. The parameters can be determined using quantum mechanical calculations, semi-empirical comparisons, effective parameterization strategies, or phenomenologically based arguments. In the latter cases, the energetic terms represent effective interactions between atoms inside the biomolecule. These effective interactions typically account for solvent effects when the solvent is not explicitly modeled, and any other energetic interactions that are not represented explicitly, which can include hydrogen bonds, salt bridges, and structural water molecules.

B. Coupling to a heat bath

Biomolecules are constantly bombarded by collisions with water and other molecules inside the cell. To describe the exchange of energy between the molecule of interest and the local environment, it is common for MD simulations to couple the dynamics to an external constant temperature heat reservoir. If the number of particles and volume are also held constant, then the canonical NVT ensemble is described. One way to account for coupling to a heat bath is to apply Langevin dynamics. In these applications, we integrate Newton's equations of motion, and the effect of the solvent is taken into account by introducing a drag term and a random force term,

$$m\ddot{\mathbf{X}} = -\nabla U - \gamma\dot{\mathbf{X}} + \sqrt{2\gamma k_B T}\mathbf{R}(t), \tag{3}$$

where $\mathbf{X}$ is the position of an atom (out of N atoms), U is the potential energy function, and γ is an effective drag coefficient. The second term represents dissipative momentum exchange of the protein with the solvent molecules, and the third term represents the random force imparted by collisions with the solvent atoms. This random force is represented as Gaussian white noise with zero mean, and the distribution of values is defined according to the fluctuation-dissipation theorem.[24]

C. Structure-based models: The spherical cow

As discussed in Sec. I, physicists usually begin their study of a complex system by proposing a simplified model (think of how many times you have reduced a system to a simple harmonic oscillator). In the context of understanding the physics of molecular biology, an extremely simplified version of a biomolecule is a structure-based model, which we now describe.

We will employ an all-atom structure-based (SMOG) model[3,4] to demonstrate how simple force fields can be used

in conjunction with MD simulations to study the diffusive aspects of protein folding. In this model, all non-hydrogen atoms are represented as beads of unit mass. The parameters of the force field are defined to stabilize a known biomolecular structure, which is usually obtained from experiment. The force field is intended to represent the effective energetics of the system, after averaging over electrostatics, van der Waals, and solvent effects. With this approach, the potential energy landscape is funneled toward the native conformation. Accordingly, at sufficiently low temperatures, the free energy landscape will also exhibit funnel-like characteristics. In this representation, the absence of non-native attractive interactions is in accordance with the principle of minimal frustration.[13–15] That is, the principle of minimal frustration implies a smooth energy landscape, in which native interactions provide the dominant contribution to the energetics. Although the energy landscape is minimally frustrated, free energy barriers can still occur as a result of steric or entropic factors.

In this model, the potential energy function is given by

$$\begin{aligned} U = & \sum_{\text{bonds}} \frac{\epsilon_r}{2}(r - r_0)^2 + \sum_{\text{angles}} \frac{\epsilon_\theta}{2}(\theta - \theta_0)^2 \\ & + \sum_{\text{impropers}} \frac{\epsilon_\chi}{2}(\chi - \chi_0)^2 + \sum_{\text{backbone-dihedrals}} \epsilon_{\text{bb}} F(\phi - \phi_0) \\ & + \sum_{\text{sidechain-dihedrals}} \epsilon_{\text{sc}} F(\phi - \phi_0) \\ & + \sum_{\text{contacts}} \epsilon_c \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \\ & + \sum_{\text{non-contacts}} \epsilon_{\text{nc}} \left(\frac{\sigma_{\text{nc}}}{r} \right)^{12}, \end{aligned} \tag{4}$$

where

$$F(\phi) = [1 - \cos\phi] + \frac{1}{2}[1 - \cos 3\phi]. \tag{5}$$

The bond, angle, and improper dihedral terms define the covalent bond structure of the protein. The values of the parameters r_0 and θ_0 are taken from the experimentally obtained structure. Non-planar improper dihedrals χ_0 are also given the values adopted in the structure. The energy scale is set by ϵ, with $\epsilon_r = 100\epsilon/\text{Å}^2$, $\epsilon_\theta = 80\epsilon/\text{rad}^2$, $\epsilon_\chi = 40\epsilon/\text{rad}^2$ for planar dihedrals, and $\epsilon_\chi = 10\epsilon/\text{rad}^2$ for other improper dihedrals.

The dihedral angle terms influence the local secondary structure of the protein. The values of ϕ_0 are set to those in the experimentally obtained structure. This model provides more stabilizing energy to backbone dihedrals, where the values of ϵ_{bb} and ϵ_{sc} are assigned values such that $\epsilon_{\text{bb}}/\epsilon_{\text{sc}} = 2$.

Non-bonded interactions include stabilizing interatomic contacts as well as steric repulsion between atom pairs that are not in contact in the native structure. In this simplest form of the model, non-bonded interactions are all described by isotropic pairwise interactions. Contact interactions are defined between any two atoms that are in contact in the native structure. Native contacts are defined based on the shadow algorithm.[25] In this algorithm, any two atoms are considered "in contact" in the experimental structure if they are separated by less than 6 Å, not connected by less than four bonds, and not occluded by any other atoms. All native contacts are given Lennard-Jones-like interactions, with minima set to σ_{ij}, the interatomic distances found in the experimentally obtained structure. Consistent with theoretical analyses,[16,22,26] ϵ_{c} is defined such that the amount of energy in the contacts is twice the energy in the dihedrals,

$$\frac{\sum \epsilon_{\text{c}}}{\sum \epsilon_{\text{bb}} + \sum \epsilon_{\text{sc}}} = 2 \tag{6}$$

and

$$\sum \epsilon_{\text{c}} + \sum \epsilon_{\text{bb}} + \sum \epsilon_{\text{sc}} = N\epsilon. \tag{7}$$

The inverse 12th power steric repulsion terms between all non-contact atoms have $\sigma_{\text{nc}} = 2.5$ Å and $\epsilon_{\text{nc}} = 0.1\epsilon$.

D. Protein folding thermodynamics

For many protein sequences, there is a well-defined folded state that is typically associated with a globular structure. In this structure/state, there are many interactions that contribute to the stability of the folded protein. Interestingly, many proteins reversibly fold and unfold under constant temperature conditions. When the protein is folded, excluded volume interactions and topological constraints strongly limit the configurations that may be adopted, and the folded state may be described as a small ensemble (with low configurational entropy) of structurally similar configurations. In contrast, the unfolded state is characterized by an extended peptide chain, where there is only a small number of native contacts that transiently form and break. As a result, the unfolded protein can explore a vast range of configurations. Accordingly, the folded state has low enthalpy and a low entropy, while the unfolded state has a higher enthalpy and higher entropy.

The relative balance between folded and unfolded states depends on the thermodynamic stability of each state. In the canonical ensemble, systems are driven toward macrostates that minimize the free energy. We will consider the Helmholtz free energy $F = E - TS$, where E is the energy, S is the entropy, and T is the temperature. At high temperatures, there is a greater weight given to entropy, such that systems are driven toward high entropy states, thus lowering the free energy. At low temperatures, the influence of entropy is minimal, and systems are driven toward low energy states. In the case of protein folding, there exists a temperature for many proteins at which the folded and unfolded states have equal free energies, and are therefore equally probable. At this temperature, called the folding temperature T_f, the protein will reversibly interconvert between folded and unfolded states indefinitely.

There are several ways to identify the folding temperature of a protein. First, we can define a reaction coordinate (or order parameter), such as the fraction of native contacts that distinguishes between the two states. We simulate the system at many temperatures and determine the temperature at which the reaction coordinate values associated with the folded state occurs with the same probability as the values associated with the unfolded state. Another way to identify T_f is by the behavior of the specific heat at constant volume C_V. For temperatures less (greater) than T_f, the protein is predominantly in the folded (unfolded) state. For temperatures below or above T_f, the macrostate (folded or unfolded) does

not depend strongly on the temperature, and the specific heat will be small. As the temperature approaches the folding temperature, the protein abruptly undergoes a pseudo-first-order phase transition between the folded and unfolded states. This rapid shift is accompanied by a very large change in the energy, which will manifest itself as a peak in the specific heat (Fig. 3). The peak in C_V indicates the pseudo-phase transition temperature, in this case the folding temperature.

IV. DIFFUSIVE MOTION AND BIOMOLECULAR RATES

To introduce how the concept of diffusion is used to study the dynamics of protein folding, we first discuss some principles of diffusive dynamics.

A. Fundamentals of diffusive dynamics

Diffusion is ubiquitous in fields ranging from physics and chemistry, to economics and other social sciences. From a macroscopic view, we can treat diffusion in terms of the net movement of a quantity due to a gradient in concentration, which drives particles to lower-concentration regions. Fick's first law formalizes the relation between the diffusion coefficient D, concentration ρ, and flux as[27]

$$\mathbf{J} = -D\nabla\rho, \tag{8}$$

where $\mathbf{J}$ is the diffusion flux vector. By applying the principle of mass conservation in a closed system, and assuming a constant diffusion coefficient, we can derive the second Fick's law[27]

$$\frac{\partial\rho}{\partial t} = D\nabla^2\rho. \tag{9}$$

Although Eqs. (8) and (9) provide a deterministic description of diffusion, the motion of an individual particle is largely random and controlled by noise. Historically, the erratic movement of pollen particles in a motionless water droplet first caught the attention of Robert Brown, and the random motion that underlies diffusion is now known as Brownian motion. Brownian motion was subsequently analyzed by Einstein in one of his annus mirabilis' papers.[28,29] The relation between the random thermal motion and diffusion has been discussion by Einstein, Smoluchowski, Langevin, Fokker, among many others.

For simplicity, we will consider a one-dimensional representation of a diffusive system and write the Brownian motion in terms of the Langevin equation as

$$m\frac{d^2x}{dt^2} = -\alpha\frac{dx}{dt} + F_r(t), \tag{10}$$

where α is the friction coefficient and $F_r(t)$ is a random force. This random term arises due to collisions with the environment and has the properties that there is no directional bias, $\langle F_r(t)\rangle = 0$, and sequential collisions are random in direction and amplitude, $\langle F_r(t)F_r(t')\rangle = B\delta(t-t')$. We divide both sides of Eq. (10) by m and rewrite it in terms of the particle velocity to find

$$\frac{dv}{dt} = -\gamma v + \zeta(t), \tag{11}$$

where $\gamma = \alpha/m$ and $\zeta(t) = F_r(t)/m$. Equation (11) obeys the same conditions as Eq. (10), and thus $\langle\zeta(t)\rangle = 0$ and $\langle\zeta(t)\zeta(t')\rangle = \Gamma\delta(t-t')$, with $\Gamma = B/m^2$. By applying these conditions to the solution of Eq. (11), we find that the solution of Eq. (11) is

$$\langle v^2\rangle - \langle v\rangle^2 = \frac{\Gamma}{2\gamma}(1 - e^{-2\gamma t}), \tag{12}$$

which at long times approaches $\langle v^2\rangle - \langle v\rangle^2 = \Gamma/(2\gamma)$. By applying the equipartition theorem to the first term, we obtain the explicit temperature dependence,

$$\Gamma = \frac{2\gamma k_B T}{m}, \tag{13}$$

and consequently $B = 2\alpha k_B T$. By using the same strategy, we can determine the mean-square displacement of a particle that is undergoing Brownian motion,

$$\langle x^2\rangle - \langle x\rangle^2 = \frac{\Gamma}{\gamma^2}\left[t - \frac{2}{\gamma}(1 - e^{-\gamma t}) + \frac{1}{2\gamma}(1 - e^{-2\gamma t})\right], \tag{14}$$

where the left-hand side is the mean-squared displacement as a function of time. In the long-time limit, the first term of Eq. (14) is dominant and the mean-square displacement reduces to[30,31]

$$\langle x^2\rangle - \langle x\rangle^2 = 2Dt, \tag{15}$$

where $D = \Gamma/(2\gamma^2)$ and $D = B/(2\alpha^2)$. From Eq. (13), we can write the relation between the diffusion coefficient and the temperature as

$$D = \frac{k_B T}{\alpha}. \tag{16}$$

Equation (16) is also applicable for two and three dimensions. If we substitute the friction coefficient for a spherical particle of radius r moving in a liquid of viscosity μ, Eq. (16) becomes

$$D = \frac{k_B T}{6\pi\mu r}. \tag{17}$$

B. Applying diffusive concepts to biomolecules

Although the early studies of diffusion provided fundamental insights into the principles that govern stochastic processes, it was Hendrik A. Kramers who extended these concepts to analyze the rates of chemical reactions.[32–34] By building on Einstein's results, he derived general diffusion equations that describe the dynamics in the low and high viscosity regimes. The latter, overdamped regime is most relevant to the study of biomolecular dynamics and will be used here to describe the dynamics of large-scale collective processes.

Protein folding is a self-organizing process that is well described in terms of chemical reaction concepts.[35] It is useful to categorize each accessible structure of a biomolecule as belonging to the unfolded, partially folded, or fully folded ensemble. We use the number of native contacts as the reaction coordinate, and an ensemble of configurations is associated with a common number of contacts. If these ensembles

are ordered in terms of their similarity to the folded state, the number of conformations that can be accessed decreases dramatically as the folded state is approached.[12,14,16] The distribution of configurations then resembles a funnel-like entity, where the global minimum corresponds to the folded structure. Based on the principle of minimal frustration,[12,14,16] the energetic roughness on the funnel must be small relative to the difference in energy between the folded (the bottom of the funnel) and unfolded states (the top of the funnel). In this interpretation, the kinetics of folding is governed by the overall slope of the funnel, as well as its roughness, where the latter controls the diffusive properties of the molecule.

Although protein energy landscapes have many dimensions, it is often sufficient to consider the free energy as a function of only a small number of coordinates. These low-dimensional free energy landscapes have two dominant minima that correspond to the unfolded and folded states, similar to the descriptions of products and reactants when discussing chemical kinetics. Chemical kinetics are often described using single atomic distances or angles. In contrast, folding is a collective process that typically requires more elaborate metrics to quantify. A widely used coordinate for describing folding is the fraction of native contacts that are formed as a function of time.[2–4] As the protein folds, more of the native (folded) contacts are formed, and the coordinate will adopt larger values. In many proteins, this reaction coordinate captures the basic diffusive properties of the folding process,[36] which allows the system to be described in terms of the Fokker–Planck equation. This equation describes the time evolution of a stochastic process subject to a deterministic drift,[37]

$$\frac{\partial}{\partial t}P(x,t) = \left[-\frac{\partial}{\partial x}v + \frac{\partial^2}{\partial x^2}D\right]P(x,t), \tag{18}$$

where $P(x, t)$ is the probability density of x, v is the drift velocity (associated with the external force), and D is the diffusion coefficient. Although Eq. (18) is given in terms of the spatial coordinate x, this relation can be used to describe other stochastic processes, and x can represent either a spatial coordinate or a generalized reaction coordinate. In addition, the diffusion coefficient D can depend on the value of x.

For short time scales, the solution of Eq. (18) is given by[37]

$$P(x,t) = -\frac{1}{\sqrt{4\pi D(x_c)t}}\exp\left[-\frac{(x-x_c-v(x_c)t)^2}{4D(x_c)t}\right], \tag{19}$$

for the initial condition $P(x, t = 0) = \delta(x_c)$. Equation (19) represents a Gaussian distribution, initially centered at x_c, moving with velocity $v(x_c)$, where the width of the Gaussian σ increases as the square root of t ($\sigma(t) = \sqrt{2D(x_c)t}$). The drift and diffusion coefficients can be expressed as

$$v(x_c) = \frac{\langle x(t_2)\rangle - \langle x(t_1)\rangle}{\Delta t} \tag{20}$$

and

$$D(x_c) = \frac{\sigma^2(t_2) - \sigma^2(t_1)}{2\Delta t}, \tag{21}$$

where $\Delta t = t_2 - t_1$. In principle, Eqs. (20) and (21) should be evaluated in the limit $\Delta t \to 0$ to obtain the drift and diffusion coefficients from a given dataset. However, in real-world applications, Δt only has to be small enough to ensure convergence of both quantities.

By using Eqs. (20) and (21), it is possible to numerically extract the diffusion and drift coefficients directly from a time series of values for a specific reaction coordinate $x(t)$.[38] We will provide an example of how to use these relations to quantify diffusive dynamics, which we call the drift–diffusion (DrDiff) approach.[39–41] To use this approach, one discretizes the reaction coordinate values into bins, where each bin is centered around x_c with a width of δx_c. By using these binned time values, time-dependent distributions are calculated over the interval $[t_{\text{initial}}, t_{\text{final}}]$. The functional form given by Eq. (19) is then fit to each distribution to provide an estimate of the position of the Gaussian center and standard deviation for each time Δt. Linear regressions for $\sigma^2(t)$ and $x_c(t)$ are then evaluated using all the values obtained from the time interval considered.

For systems that are described well in terms of diffusion on a one-dimensional landscape, we can obtain several relations between diffusion, drift, free energies and rates. For example, the free energy profile can be extracted from the drift velocity and diffusion coefficient using[42]

$$F(x)/k_BT = -\int_{x_{\text{ref}}}^{x}\frac{v(x')}{D(x')}dx' + \ln D(x) + \text{constant}, \tag{22}$$

where the additive constant is related to the arbitrary free energy of reference state x_{ref}. Equation (22) can be derived by assuming the equilibrium probability density P_{eq} is a solution of the steady-state Fokker–Planck equation, Eq. (18) and that $P_{\text{eq}}(x) \propto \exp[-F(x)/k_BT]$, where $F(x)$ is the free energy.[42] In addition, the mean first-passage time τ_f between two points on the profile, which is inversely related to the rate, is given by[43]

$$\tau_f = \int_{x_{\text{unf}}}^{x_{\text{fold}}}dx\int_0^x dx'\frac{e^{\beta[F(x)-F(x\prime)]}}{D(x)}, \tag{23}$$

where $F(x)$ is the coordinate-dependent free energy profile, $D(x)$ is the coordinate-dependent diffusion coefficient, and $\beta = 1/k_BT$. In this context, it is assumed that the coordinate increases as a function of the reaction (folding) and that there is a lower bound of zero.

Equation (23) can be used to obtain folding time scales if x_{unf} and x_{fold} define the unfolded and folded state minima on the free energy profile. The intervening barrier defines the transition state ensemble, which is the collection of configurations that have values of the coordinate for which the free energy is maximal. In the following, we will use the fraction of native contacts as the coordinate for folding, because it has been shown to exhibit diffusive properties and captures the rate-limiting barrier for many systems.[36]

C. WHAM—weighted histogram analysis method

The calculation of the thermodynamic properties of a system can be reduced to determining the density of states. We will employ the weighted histogram analysis method (WHAM) to combine histograms from multiple simulations performed with different thermodynamic parameters (temperature). Instead of providing a derivation, we provide a

brief introduction to WHAM, so that readers may appreciate this statistical mechanics tool.

The potential of mean force[44,45] is widely used as a measure of the free energy change during biomolecular processes. In terms of the probability distribution $p(\xi)$, we may obtain the potential of mean force from the relation

$$W(\xi) = W(\xi^*) - k_B T \ln\left[\frac{p(\xi)}{p(\xi^*)}\right], \tag{24}$$

where ξ^* and $W(\xi^*)$ are arbitrary constants, ξ is the system reaction coordinate, and W is the potential of mean force. This approach is useful when there is sufficient Boltzmann sampling, although the accuracy of the calculation is limited by the quality of the data. WHAM is a powerful technique that allows one to extract unbiased distributions from datasets that were obtained with or without biasing forces. As suggested by the name, WHAM estimates the relative weights of the sampled histograms to generate a set of unbiased free energy profiles. The algorithm applies the following arguments. For a set of Z histograms, the average distribution function is given by[46–48]

$$\langle p(\xi)\rangle = \frac{\sum_{i=1}^{Z} n_i \langle p(\xi)\rangle_i}{\sum_{j=1}^{Z} n_j e^{-\beta\left[w_j(\xi)-f_j\right]}}, \tag{25}$$

where n_i is the number of snapshots used to calculate a histogram/distribution from the ith simulation: $\langle p(\xi)\rangle_i$. From Eq. (25), the free energy of the ith simulation can be calculated from

$$\beta f_i = -\ln \int \langle p(\xi)\rangle e^{-\beta w_i(\xi)} d\xi. \tag{26}$$

Equations (25) and (26) were derived by minimizing the sampling errors in the overlapping regions of the probability distributions.[48] In the WHAM algorithm, Eqs. (25) and (26) are iteratively evaluated until a self-consistent solution is obtained, at which point $p(\xi)$ is approximated by $\langle p(\xi)\rangle$. This solution yields temperature-dependent free energy profiles as well as the specific heat as a function of temperature. We apply a version of WHAM that is distributed with the SMOG2 software package;[4] details can be found in the SMOG2 documentation.

V. RESULTS

We discuss an introductory example of how we can quantify diffusive aspects of protein folding using a simplified model. First, we will discuss how to identify the folding temperature for a given protein and model. We then discuss the calculation of free energy barriers and the calculation of diffusion coefficients from folding trajectories. We close with a brief discussion of convergence considerations when applying molecular simulations to study dynamics. For our discussion, we will simulate and analyze the dynamics of the protein chymotrypsin inhibitor 2 (CI2; PDB code 2CI2).[49] CI2 was chosen because it has been widely utilized as a model protein for the study of folding dynamics. In addition, it is a relatively small protein that can be easily simulated, making it an excellent system for demonstrating the techniques and ideas associated with the analysis of diffusion in proteins.

All simulation results presented here are available at the SMOG2_tutorial repository.[50] The repository also provides a file with instructions on how to replicate the simulated trajectories and analysis. Due to the stochastic properties of biomolecular dynamics, individual time traces are not exactly reproducible. However, the statistical properties should be consistent with our discussion.

A. Finding the folding temperature

To study protein folding/unfolding under constant temperature conditions, it is necessary to first identify the folding temperature. To do so, we typically start by performing several constant temperature simulations that span a wide range of temperatures. All simulations were performed with the GROMACS software package.[51,52] GROMACS is a molecular dynamics simulation package that is widely used for the simulation of proteins, nucleic acids, and lipids. This package integrates a given potential energy function, as in Eq. (4), to determine the time evolution of a system. To visualize the simulated trajectory, we use software, such as VMD (visual molecular dynamics). Because GROMACS does not allow for the use of reduced units (the Boltzmann constant is hard coded), a reduced temperature of 1 corresponds to a GROMACS temperature of 120. That is, in reduced units, the Boltzmann constant is set equal to 1. However, because GROMACS uses a value of $0.008\,31 \approx 1/120$ for the Boltzmann constant, a numerical value of $\approx$120 in GROMACS is equivalent to $k_B T = 1$. In structure-based models, room temperature corresponds to approximately 0.5

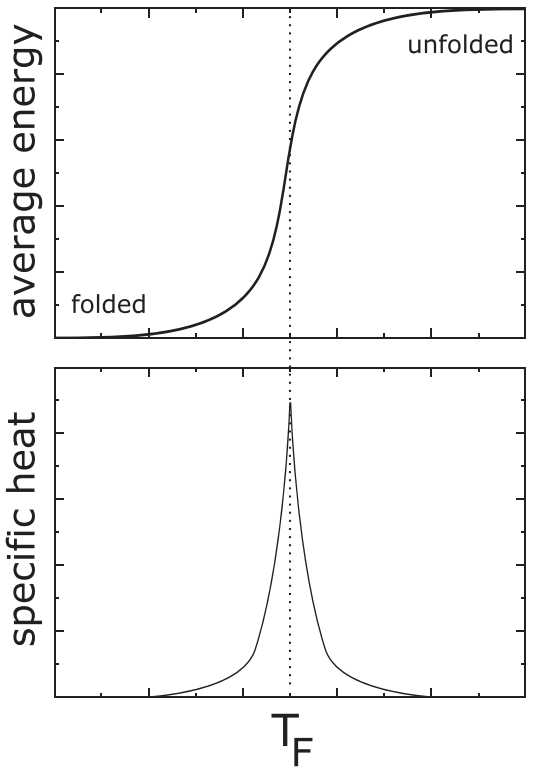

Fig. 3. Schematic of the temperature dependence of the average energy and specific heat for a protein. The specific heat exhibits a large peak around the folding temperature T_f.

reduced energy units. For the first iteration (iteration 1 in Fig. 4), the simulations were performed at eight different temperatures ranging from 0.67 to 1.25 (in reduced temperature units). WHAM was then used to combine the data from all temperatures and generate a specific heat curve (Fig. 4; right-hand curve). This initial set of simulations gives a pronounced peak at $T \approx 1.015$. Because it is possible that poor (non-Boltzmann) sampling can lead to artificial peaks in C_V, we next performed additional simulations near the candidate folding temperature. In iteration 2, six additional simulations were performed for temperatures ranging from 0.99 to 1.03, and the simulations from both iterations were then combined using WHAM. Perhaps surprisingly, we see a clear shift in the C_V peak to lower temperatures. Based on this result, we performed five additional simulations (iteration 3) near the new candidate folding temperature of ≈1.0. In this example, we repeated the process for a total of four iterations. Between the third and fourth iterations, there were minimal changes in C_V, which implies that $T_f \approx 0.994$.

After identifying the folding temperature, we typically perform a much longer simulation at the folding temperature to obtain a large number of spontaneous folding and unfolding events. We performed a single simulation at the folding temperature of 10^9 time steps, which is 100 times longer in duration than the initial simulations. We will refer to this long simulation as the "trajectory." Figure 5 shows the number of native contacts $Q(t)$ as a function of time. A zoomed-in view of the simulation indicates there is a clear separation between the folded (high Q) and unfolded (low Q) states where abrupt transitions occur. In total, ≈80 transitions between folded and unfolded states were found. The plot on the top right in Fig. 5 shows two well defined peaks in the probability distribution. The peak at $Q \approx 450$ is associated with the folded state, and the one centered at $Q = 50$ corresponds to the unfolded state. For reference, the native structure has 597 native contacts. The fact that the ensemble of folded configurations only has ≈80% of the native contacts formed may be surprising. This difference between the theoretical maximum and the most probable value can be understood in terms of simple thermodynamic considerations. That is, when all contacts are formed the protein adopts a highly compact form, which is associated with strong steric interactions that confine the chain. In other words, forming all contacts simultaneously is entropically disfavored.

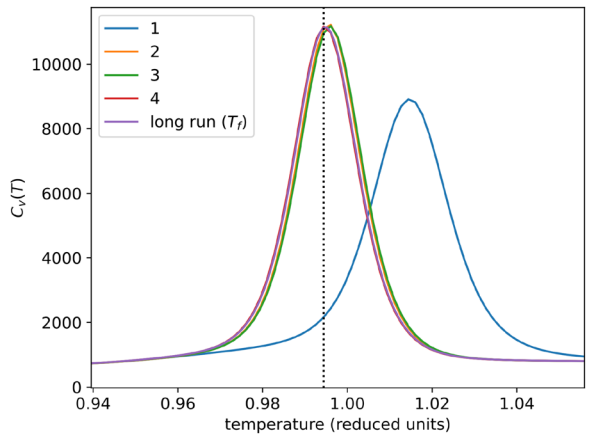

Fig. 4. The specific heat as a function of temperature (in reduced units) for a structure-based model of the protein CI2. To find an initial estimate of the folding temperature, we evaluate the temperature where C_V reaches a peak value. As the sampling is enhanced (iterations 1–4), the peak maximum shifts and moves closer to the correct folding temperature value of $T_f = 0.994$.

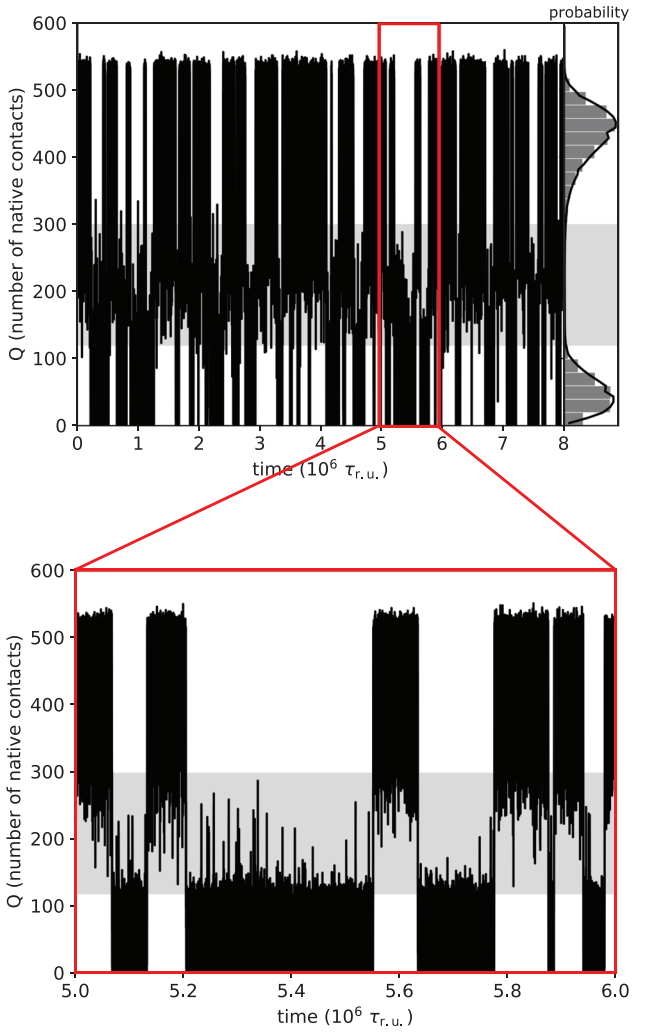

Fig. 5. Number of native contacts Q as a function of time (top). In this trajectory, there are ≈80 transitions between the folded (high Q) and unfolded (low Q) states. The inset on the right shows the corresponding probability density. The bottom plot shows a zoomed-in perspective, which highlights the transition events and relative stabilities of the folded and unfolded states. The transition state ensemble in the gray region corresponds to configurations that successfully cross the underlying free energy barrier.

B. Calculating drift velocities and diffusion coefficients

Unlike applications of diffusion that describe the flux of particles that arise from a concentration gradient, diffusion in protein folding describes the probability flux between the members of a conformational ensemble. To describe this ensemble, it is often suitable to choose reaction coordinates to measure the "folded-ness" of the system. Although there is no guarantee that a given coordinate will be suitable for the analysis of diffusion, we will find that the number of contacts is suitable for describing the kinetics and thermodynamics of this protein.

Although textbooks often describe a diffusion coefficient as a constant, it can also be time and coordinate dependent.[40,53–55] In the context of protein folding, there have been many techniques applied to quantify diffusive properties[56,57] from experiments[58,59] and simulations.[38,60,61]

To introduce the analysis and interpretation of diffusive dynamics, we apply the DrDiff approach described in Sec. IV B to evaluate both the drift and diffusion coefficients as well as the free energy profile from a simulated dataset.

We analyze a long constant-temperature simulation (see Fig. 6) to estimate the coordinate dependent diffusion and drift coefficients. As mentioned in Sec. V A, this simulation includes more than 80 folding/unfolding events where the protein spontaneously samples fully folded and unfolded configurations. Figure 6(a) shows the diffusion coefficient calculated using Eq. (21). The diffusion coefficient increases with the number of native contacts and then decreases as the folded state is reached. We additionally calculate the drift velocity [see Fig. 6(b)], which is zero at distinct values of Q. Intuitively, these states should correspond to free energy minima and maxima. The diffusion and drift coefficients were then used to estimate the free-energy profile using Eq. (22). Based on the time traces, there are two deep free energy minima that are separated by a clear barrier [Fig. 6(c)]. By comparing the probability density peaks with the free energy curve, it is clear that the two minima are more highly sampled than the transition state, which is the region where the free energy is no less than 1 k_BT below the peak value.

The increase in the diffusion coefficient from the unfolded state to the transition state [Fig. 6(a)] reveals how entropic factors can influence diffusive properties. That is, during folding, a protein must initially collapse, where it becomes increasingly difficult to form additional contacts due to a high level of residual disorder in the chain. However, once a critical set of contacts is formed, the protein is sufficiently confined that the rearrangements that result in formation of additional contacts are essentially the only allowable motions.

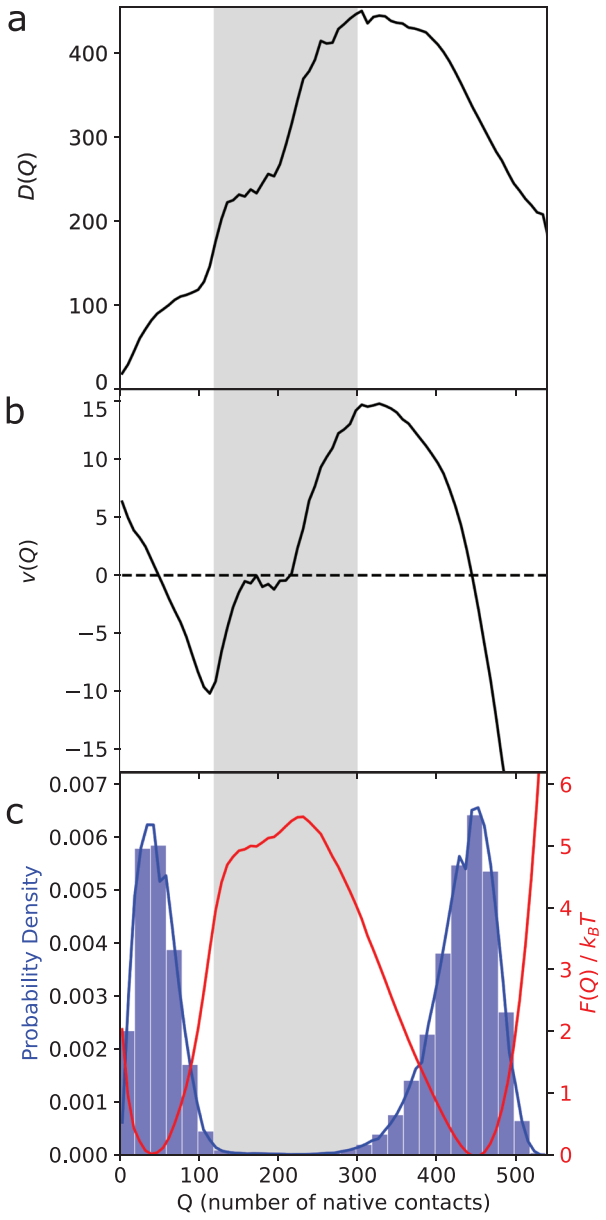

Fig. 6. (a) Coordinate-dependent diffusion coefficient obtained using DrDiff, Eq. (21). The diffusion coefficient increases from the unfolded state until the end of the transition state (gray shaded area) and then decreases as the protein approaches the folded state. (b) Coordinate-dependent drift coefficient extracted from the trajectory data. The gray shaded area shows the transition state, defined as the region where the free energy is within k_BT of its peak value at the barrier. (c) Probability density as a function of the number of native contacts (blue, peaks near $Q = 50$ and 450), alongside the free energy profile integrated from Eq. (22) (red, peak near $Q = 250$).

C. Calculating the free energy barrier

There are several techniques to determine the free energy from a simulation. The most direct method is to evaluate the potential of mean force from the probability density, using

$$F_{\text{eq}} = -K_{\text{B}}T \ln [P(Q)] + \mathbf{C}, \tag{27}$$

where $P(Q)$ is the probability density for the chosen reaction coordinate Q, and $\mathbf{C}$ is an arbitrary constant. The green curve in Fig. 7 represents the free energy calculated from Eq. (27). Figure 7 also displays the free energy profile calculated from multiple simulations at different temperatures that are combined using WHAM. In addition, the free energy was estimated based on Eq. (22), where the drift and diffusion coefficients were obtained from the DrDiff approach. The three techniques were applied to demonstrate the suitability of the reaction coordinate for describing the diffusive process.

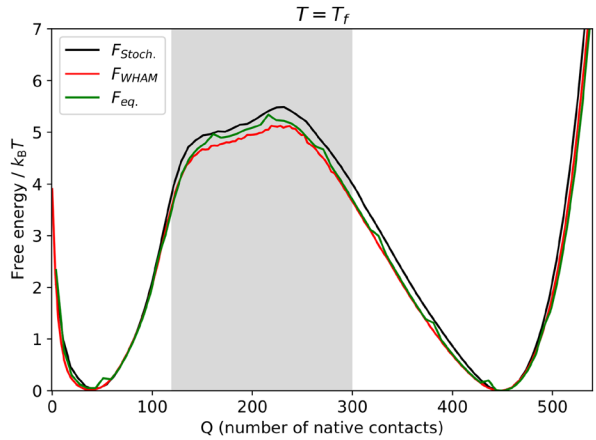

Fig. 7. Comparison of the free energy profiles obtained by three different techniques. The green curve presents the free energy obtained from the trajectory probability density, using Eq. (27). In black is the free energy integrated from the coordinate-dependent drift and diffusion coefficients using Eq. (22). Both are compared with the result from WHAM (red). The gray shaded area is the transition state ensemble.

We find that the main features of the free energy profile at T_f are the same using the three techniques we considered. The positions of the minima and barrier are the same, although there are small differences in the barrier heights. The agreement suggests that the kinetics are well described in terms of one-dimensional diffusion along our chosen reaction coordinate. We stress that agreement between the methods is not guaranteed for an arbitrary dynamical system. It is possible that some systems when described by certain coordinates may appear to exhibit subdiffusive dynamics, even if the underlying dynamics is not truly subdiffusive.[62]

D. Convergence and sampling considerations

We now discuss the importance of both duration and frequency of data storage/analysis when estimating kinetic and thermodynamic quantities from simulations. Because these considerations are not unique to a particular form of analysis, the general strategies can be applied to other areas as well.

To illustrate a simple method for verifying convergence, we analyze fragments of the long trajectory used for the previous calculations. First, we analyzed the initial 10^6 time steps of the simulation using DrDiff. The results will be referred to as 1M and contain the first ten percent of the trajectory (file Q-119.5.segment1.dat in the tutorial repository). Figure 8(a) shows the free energy profile (dashed blue line) estimated from these dataset. As expected, the free energy is undefined for small values of Q because no unfolding transitions occurred in these initial frames, and hence, the unfolded conformations were not represented. The diffusion coefficient is also very noisy in the transition state region, and no values are available for the unfolded ensemble. Next, the first 5×10^6 trajectory steps, referred to as 5M, were analyzed and the results are displayed in Fig. 8 (dashed orange lines). This part of the simulation included only four folding/unfolding transitions. Nonetheless, the diffusion coefficients and free energies from the 5M dataset are comparable with those obtained from the complete trajectory (80 transitions).

Another challenge when calculating diffusive properties is the choice of the proper frequency for saving/analyzing simulated frames. For the DrDiff approach, it is necessary that the reaction coordinate changes by small increments between saved configurations. To illustrate the influence of this point, the results from the complete saved trajectory were compared to the results obtained when only every wth frame is considered, where w is known as the stride value. A stride value of $w = 1$ means that all configurations were saved, keeping the original trajectory intact. A stride value of $w = 10$ means that every tenth configuration is saved, resulting in a trajectory ten times smaller with a time step ten times longer. Figure 9 shows how the different stride values impact the convergence of the diffusion coefficient values. As expected, larger deviations are observed as w is increased. Interestingly, the positions of the free energy minima and barrier were insensitive to the stride value [Fig. 9(a)]. In contrast, the free energy barrier height changed by nearly $2k_BT$ as w was varied. This dependence on w highlights how, even when describing the same system, various thermodynamic and kinetic properties may converge at different rates. Accordingly, it is always necessary to verify the convergence and robustness of each quantity analyzed. In research studies, a more extended and iterative process than the one we discussed here is typically required to ensure that there is a satisfactory level of convergence.

VI. REMARKS

There are many opportunities for physicists to identify novel problems in the biological sciences. We have discussed how a spherical-cow-like model can be used to obtain insights into the diffusive dynamics of protein folding. Our discussion represents only the beginning of what may be explored with these models. Although protein folding is a

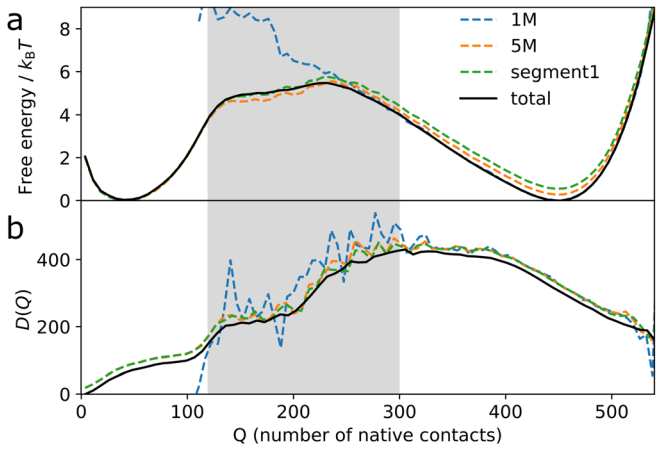

Fig. 8. Results from a complete trajectory at the folding temperature (total) compared with results from three subsets: (a) Free energy integrated from Eq. (22) for different trajectory segments (dashed lines) alongside the full trajectory (solid black line). The numbers are related to the file numbers available in the repository. (b) Corresponding coordinate-dependent diffusion coefficients obtained using DrDiff via Eq. (21). With no transitions, the 1M trajectory segment (first 10^6 time steps) shows discrepancies in the transition state and low Q values. All other analyzed segments show free energy profiles and diffusion coefficient curves that agree with those generated from analysis of the complete trajectory. The gray shaded area is the transition state region.

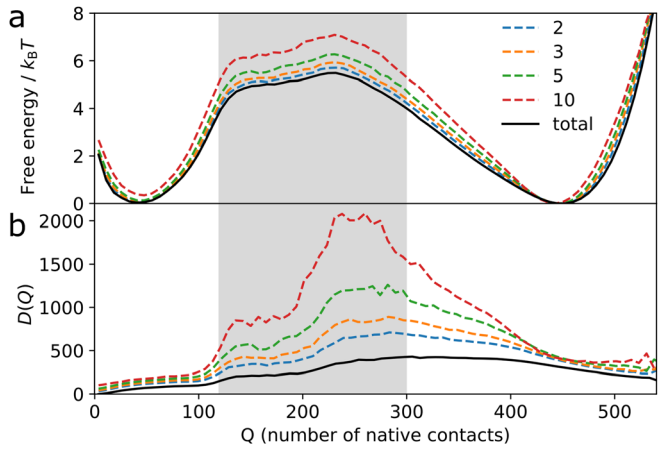

Fig. 9. Results from a complete trajectory (total) compared with the results of trajectory subsets saved after skipping every w time steps, with w varying from 2 to 10: (a) Free energy integrated from Eq. (22) for five values of w applied to the trajectory at T_f. The positions of the free energy minima are the same, despite the different values of w, while the barrier is reduced by $\approx 2k_BT$ when w is increased. (b) The corresponding coordinate-dependent diffusion coefficient is obtained via Eq. (21). As w increases, the deviations from the total trajectory (black) increase because there are fewer sampled Q values and because the drift-diffusion analysis is based on short-time dynamics, some of which is missed as w is increased. The gray shaded area is the transition state region.

mature field, the ideas developed in this context can be applied to a broad range of biological processes. For example, the study of reaction coordinates and diffusion is providing insights into the relation between structure, energetics, and dynamics in very large molecular assemblies, such as bacterial[9] and eukaryotic[63] ribosomes. We expect that as the physics community continues to investigate areas of biology, new questions will be posed and answered, which will reveal the organizing principles of molecular biological processes.

ACKNOWLEDGMENTS

P.C.W. was supported by NSF Grant No. MCB-1915843. Work at the Center for Theoretical Biological Physics was also supported by the NSF (Grant No. PHY-2019745). F.C.F. was financed by the Coodenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (Capes)-Finance Code 001. Financial support for R.J.O. was provided by Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, No. APQ-00941-14) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Nos. 438316/2018-5 and 312328/2019-2).

APPENDIX: SIMULATION VISUALIZATION TUTORIAL

Visual molecular dynamics (VMD) is a powerful three-dimensional (3D) molecular visualization and analysis software package[64] that stands out in its capability to visualize extremely large biomolecules and arbitrary graphics objects along with its long molecular dynamics trajectories. In addition, VMD includes a powerful scripting interface that supports TCL-based commands[64,65] for manipulating visualizations and performing analyses.

Platform-specific VMD can be downloaded.[66] Once VMD is installed, it starts with three default windows: VMD Main, VMD OpenGL Display, and VMD console. We explore the basic capabilities of VMD by going through the steps required to generate various representations of the chymotrypsin inhibitor 2 protein.

(1) Download the PDB formatted file of the chymotrypsin inhibitor 2 protein (ID:2CI2) from the RCSB Protein Data Bank.[67] This protein can be found using the search option on the PDB webserver.
(2) To load the atomic structure file (.pdb) of the molecule, select File → New Molecule in the VMD Main window and use the Browse option to choose the.pdb file in the Molecule File Browser window; then press Load. You will be able to see the 2CI2 molecule in the OpenGL Display, and the.pdb file will be listed in the Main window.
(3) To modify the 3D visualization of the molecule, there are three modes by which the mouse may be used to alter the perspective/view: rotation, translation, and scaling. The mouse mode can be chosen from VMD Main → Mouse → R, T, or S.
(4) VMD also provides options to choose the mode of depth perception when viewing the molecule. VMD Main → Display → Perspective (strong depth perception)/ Orthographic (low depth perception). We find that larger molecules are much easier to view with Orthographic, though Perspective is usually sufficient for small systems. Although the default background color is black, it is often desirable to use other colors such as white. To change the background color in the OpenGL Display to white, go to VMD Main → Graphics → Colors → Display (under Categories)→ Background (under Names)→ white.
(5) VMD has many options available for graphical representations of the molecule and atom selections. The graphical representations window can be accessed from VMD Main → Graphics → Representations. This will provide options for Atom Selection, Drawing Method, Coloring Method, etc. The default representation is "Lines" for Style, "Name" for Color and "all" for Selection. It is possible to either edit the default representation or create additional representations using Create Rep option.
(6) The different options available for Drawing Methods can be found under Draw style. We have found that the most useful representations are VDW (sphere for each atom), Tube (representing only backbone traces) and NewCartoon [protein secondary structure; similar to Fig. 1(b)].
(7) Similarly, there are several options available to color the molecule or the selection based on for example, Name (atom type), ResType (residue type), Secondary Structure, and Backbone.
(8) The selection tab provides the keywords or single words you can string together using Boolean operators (e.g., not protein) to generate atom selections of your interest.
(9) VMD also supports loading trajectories from molecular simulations into the loaded molecule. To load a trajectory of a molecule, first select the molecule in VMD Main window, followed by File → Load Data into Molecule. The Molecule File Browser window pops up and Load Files for option will have the molecule selected. At this point, use the Browse option to select the trajectory file and press load. When following the accompanying tutorial,[41] the trajectory files will have the suffix ".xtc". The loaded trajectory can then be viewed using the animation tools at the bottom of the Main window.

[a]ORCID: 0000-0001-9195-9900.
[b]ORCID: 0000-0003-0220-195X.
[c]ORCID: 0000-0003-4860-309X.
[d]Author to whom correspondence should be addressed: p.whitford@northeastern.edu, ORCID: 0000-0001-7104-2265.

[1]J. Harte, *Consider a Spherical Cow: A Course in Environmental Problem Solving* (University Science Books, Sausalito, 1988).
[2]C. Clementi, H. Nymeyer, and J. Onuchic, "Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins," J. Mol. Biol. **298**(5), 937–953 (2000).
[3]P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, "An all-atom structure-based potential for proteins: Bridging minimal models with all-atom empirical forcefields," Proteins **75**(2), 430–441 (2009).
[4]J. K. Noel, M. Levi, M. Raghunathan, H. Lammert, R. L. Hayes, J. N. Onuchic, and P. C. Whitford, "SMOG 2: A versatile software package for generating structure-based models," PLoS Comput. Biol. **12**(3), e1004794 (2016).
[5]M. Cheung, J. Finke, B. Callahan, and J. Onuchic, "Exploring the interplay between topology and secondary structural formation in the protein folding problem," J. Phys. Chem. B **107**(40), 11193–11200 (2003).
[6]L. Chavez, J. Onuchic, and C. Clementi, "Quantifying the roughness on the free energy landscape: Entropic bottlenecks and protein folding rates," J. Am. Chem. Soc. **126**(27), 8426–8432 (2004).

[7]S. Yang, S. Cho, Y. Levy, M. Cheung, H. Levine, P. Wolynes, and J. Onuchic, "Domain swapping is a consequence of minimal frustration," Proc. Natl. Acad. Sci. U. S. A. **101**(38), 13786–13791 (2004).
[8]J. K. Noel and P. C. Whitford, "How EF-Tu can contribute to efficient proofreading of aa-tRNA by the ribosome," Nat. Commun. **7**, 13314 (2016).
[9]M. Levi, J. K. Noel, and P. C. Whitford, "Studying ribosome dynamics with simplified models," Methods **162–163**, 128–140 (2019).
[10]M. Levi, K. Walak, A. Wang, U. Mohanty, and P. C. Whitford, "A steric gate controls P/E hybrid-state formation of tRNA on the ribosome," Nat. Commun. **11**, 5706 (2020).
[11]C. Levinthal, "How to fold graciously," in *Mossbauer Spectroscopy in Biological Systems* (University of Illinois, Urbana, 1969), Vol. 67, pp. 22–24.
[12]P. E. Leopold, M. Montal, and J. N. Onuchic, "Protein folding funnels: A kinetic approach to the sequence-structure relationship," Proc. Natl. Acad. Sci. U. S. A. **89**(18), 8721–8725 (1992).
[13]J. Bryngelson and P. Wolynes, "Spin glasses and the statistical mechanics of protein folding," Proc. Natl. Acad. Sci. U. S. A. **84**(21), 7524–7528 (1987).
[14]J. D. Bryngelson and P. G. Wolynes, "Intermediates and barrier crossing in a random energy-model (with applications to protein folding," J. Phys. Chem. **93**(19), 6902–6915 (1989).
[15]J. Bryngelson and P. Wolynes, "A simple statistical field-theory of heteropolymer collapse with application to protein folding," Biopolymers **30**(1–2), 177–188 (1990).
[16]J. D. Bryngelson, J. N. Onuchic, N. D. Socci, and P. G. Wolynes, "Funnels, pathways, and the energy landscape of protein-folding: A synthesis," Proteins **21**(3), 167–195 (1995).
[17]H. Gould, J. Tobochnik, and W. Christian, *An Introduction to Computer Simulation Methods: Applications to Physical Systems* (CreateSpace Independent Publishing Platform, Scotts Valley, 2017).
[18]L. Verlet, "Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules," Phys. Rev. **159**, 98–103 (1967).
[19]M. Levi and P. C. Whitford, "Dissecting the energetics of subunit rotation in the ribosome," J. Phys. Chem. B **123**, 2812–2923 (2019).
[20]D. L. Mobley, J. D. Chodera, and K. A. Dill, "Confine-and-release method: Obtaining correct binding free energies in the presence of protein conformational change," J. Chem. Theory Comput. **3**(4), 1231–1235 (2007).
[21]H. Kim, S. C. Abeysirigunawarden, K. Chen, M. Mayerle, K. Ragunathan, Z. Luthey-Schulten, T. Ha, and S. A. Woodson, "Protein-guided RNA dynamics during early ribosome assembly," Nature **506**(7488), 334–338 (2014).
[22]M. Eastwood and P. Wolynes, "Role of explicitly cooperative interactions in protein folding funnels: A simulation study," J. Chem. Phys. **114**(10), 4702–4716 (2001).
[23]K. Lindorff-Larsen, S. Piana, R. Dror, and D. Shaw, "How fast-folding proteins fold," Science **334**(6055), 517–520 (2011).
[24]R. Kubo, "The fluctuation-dissipation theorem," Rep. Prog. Phys. **29**(1), 255–284 (1966).
[25]J. K. Noel, P. C. Whitford, and J. N. Onuchic, "The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function," J. Phys. Chem. B **116**(29), 8692–8702 (2012).
[26]H. Lammert, P. G. Wolynes, and J. N. Onuchic, "The role of atomic level steric effects and attractive forces in protein folding," Proteins **80**, 362–373 (2012).
[27]R. Pathria and P. D. Beale, "Fluctuations and nonequilibrium statistical mechanics," in *Statistical Mechanics*, 3rd ed., edited by R. Pathria and P. D. Beale (Academic, Boston, 2011), pp. 583–635.
[28]A. Einstein, "On the motion of small particles suspended in liquids at rest required by the molecular-kinetic theory of heat," Ann. Phys. **17**, 549–560 (1905).
[29]A. Einstein, "Investigations on the theory of the Brownian movement," Ann. Phys. **19**, 371–381 (1906).
[30]T. Tomé and M. J. de Oliveira, *Stochastic Dynamics and Irreversibility* (Springer-Verlag GmbH, New York, 2015).
[31]M. A. Islam, "Einstein–Smoluchowski diffusion equation: A discussion," Phys. Scr. **70**(2–3), 120–125 (2004).
[32]H. Kramers, "Brownian motion in a field of force and the diffusion model of chemical reactions," Physica **7**, 284–304 (1940).
[33]H. Brinkman, "Brownian motion in a field of force and the diffusion theory of chemical reactions," Physica **22**(1–5), 29–34 (1956).
[34]H. Brinkman, "Brownian motion in a field of force and the diffusion theory of chemical reactions. II," Physica **22**(1–5), 149–155 (1956).
[35]N. D. Socci, J. N. Onuchic, and P. G. Wolynes, "Diffusive dynamics of the reaction coordinate for protein folding funnels," J. Chem. Phys. **104**(15), 5860–5868 (1996).
[36]S. S. Cho, Y. Levy, and P. G. Wolynes, "P versus Q: Structural reaction coordinates capture protein folding on smooth landscapes," Proc. Natl. Acad. Sci. U. S. A. **103**(3), 586–591 (2006).
[37]H. Risken, *The Fokker-Planck Equation* (Springer Berlin Heidelberg, Berlin, Heidelberg, 1996).
[38]S. Yang, J. N. Onuchic, and H. Levine, "Effective stochastic dynamics on a protein folding energy landscape," J. Chem. Phys. **125**(5), 054910 (2006).
[39]R. J. de Oliveira, "Stochastic diffusion framework determines the free-energy landscape and rate from single-molecule trajectory," J. Chem. Phys. **149**(23), 234107 (2018).
[40]F. C. Freitas, A. N. Lima, V. D. G. Contessoto, P. C. Whitford, and R. J. D. Oliveira, "Drift-diffusion (DrDiff) framework determines kinetics and thermodynamics of two-state folding trajectory and tunes diffusion models," J. Chem. Phys. **151**(11), 114106 (2019).
[41]The associated computational tools are available for download at <https://github.com/ronaldolab/DrDiff>.
[42]D. I. Kopelevich, A. Z. Panagiotopoulos, and I. G. Kevrekidis, "Coarse-grained kinetic computations for rare events: Application to micelle formation," J. Chem. Phys. **122**(4), 044908 (2005).
[43]A. Szabo, K. Schulten, and Z. Schulten, "First passage time approach to diffusion controlled reactions," J. Chem. Phys. **72**(8), 4350–4357 (1980).
[44]J. G. Kirkwood, "Statistical mechanics of fluid mixtures," J. Chem. Phys. **3**(5), 300–313 (1935).
[45]B. Roux, "The calculation of the potential of mean force using computer-simulations," Comput. Phys. Commun. **91**(1–3), 275–282 (1995).
[46]A. Ferrenberg and R. Swendsen, "New Monte-Carlo technique for studying phase-transitions," Phys. Rev. Lett. **61**(23), 2635–2638 (1988).
[47]A. Ferrenberg and R. Swendsen, "Optimized Monte-Carlo data-analysis," Phys. Rev. Lett. **63**(12), 1195–1198 (1989).
[48]S. Kumar, D. Bouzida, R. Swendsen, P. Kollman, and J. Rosenberg, "The weighted histogram analysis method for free-energy calculations on biomolecules. 1. The method," J. Comput. Chem. **13**(8), 1011–1021 (1992).
[49]C. A. McPhalen and M. N. G. James, "Crystal and molecular structure of the serine proteinase inhibitor CI-2 from barley seeds," Biochemistry **26**(1), 261–269 (1987).
[50]See <https://github.com/smog-server/SMOG2_tutorial> for the repository with instructions on how to make the simulations and all the simulated data analyzed in this work.
[51]E. Lindahl, B. Hess, and D. van der Spoel, "GROMACS 3.0: A package for molecular simulation and trajectory analysis," J. Mol. Model. **7**(8), 306–317 (2001).
[52]M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," SoftwareX **1–2**, 19–25 (2015).
[53]J. Chahine, R. J. Oliveira, V. B. P. Leite, and J. Wang, "Configuration-dependent diffusion can shift the kinetic transition state and barrier height of protein folding," Proc. Natl. Acad. Sci. U. S. A. **104**(37), 14646–14651 (2007).
[54]R. J. Oliveira, P. C. Whitford, J. Chahine, V. B. P. Leite, and J. Wang, "Coordinate and time-dependent diffusion dynamics in protein folding," Methods **52**, 91–98 (2010).
[55]R. B. Best and G. Hummer, "Coordinate-dependent diffusion in protein folding," Proc. Natl. Acad. Sci. U. S. A. **107**(3), 1088–1093 (2010).
[56]K. Schulten, Z. Schulten, and A. Szabo, "Dynamics of reactions involving diffusive barrier crossing," J. Chem. Phys. **74**(8), 4426–4432 (1981).
[57]M. Gruebele, "The fast protein folding problem," Annu. Rev. Phys. Chem. **50**(1), 485–516 (1999).
[58]V. Munoz and W. A. Eaton, "A simple model for calculating the kinetics of protein folding from three-dimensional structures," Proc. Natl. Acad. Sci. U. S. A. **96**(20), 11311–11316 (1999).
[59]J. Kubelka, J. Hofrichter, and W. A. Eaton, "The protein folding 'speed limit'," Curr. Opin. Struct. Biol. **14**(1), 76–88 (2004).
[60]G. Hummer, "From transition paths to transition states and rate coefficients," J. Chem. Phys. **120**(2), 516–523 (2004).

[61]S. V. Krivov and M. Karplus, "Diffusive reaction dynamics on invariant free energy profiles," Proc. Natl. Acad. Sci. U. S. A. **105**(37), 13841–13846 (2008).
[62]S. V. Krivov, "Is protein folding sub-diffusive?," PLoS Comput. Biol. **6**(9), e1000921 (2010).
[63]F. C. Freitas, G. Fuchs, R. J. de Oliveira, and P. C. Whitford, "The dynamics of subunit rotation in a eukaryotic ribosome," Biophysica **1**(2), 204–221 (2021).
[64]W. Humphrey, A. Dalke, and K. Schulten, "VMD: Visual molecular dynamics," J. Mol. Graph. **14**(1), 33–38 (1996).
[65]J. Ousterhout, *TCL and the TK Toolkit* (Addison-Wesley, Reading, MA, 1994).
[66]See <https://www.ks.uiuc.edu/Research/vmd/> for more information about the VMD software.
[67]The PDB Protein Data Bank is located at <https://www.rcsb.org/>.

Proteins

Check it out—when I tug the C-terminal tail, the binding tunnel squeezes! (Source: https://xkcd.com/1430/)