Unfolded protein ensembles, folding trajectories, and refolding rate prediction

A. Das, B. K. Sin, A. R. Mohazab, and S. S. Plotkin

Citation: J. Chem. Phys. **139**, 121925 (2013); doi: 10.1063/1.4817215
View online: http://dx.doi.org/10.1063/1.4817215
View Table of Contents: http://jcp.aip.org/resource/1/JCPSA6/v139/i12
Published by the AIP Publishing LLC.

Unfolded protein ensembles, folding trajectories, and refolding rate prediction

A. Das, B. K. Sin, A. R. Mohazab,[a] and S. S. Plotkin[b]
Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, British Columbia V6T 1Z1, Canada

(Received 29 May 2013; accepted 17 July 2013; published online 23 August 2013)

Computer simulations can provide critical information on the unfolded ensemble of proteins under physiological conditions, by explicitly characterizing the geometrical properties of the diverse conformations that are sampled in the unfolded state. A general computational analysis across many proteins has not been implemented however. Here, we develop a method for generating a diverse conformational ensemble, to characterize properties of the unfolded states of intrinsically disordered or intrinsically folded proteins. The method allows unfolded proteins to retain disulfide bonds. We examined physical properties of the unfolded ensembles of several proteins, including chemical shifts, clustering properties, and scaling exponents for the radius of gyration with polymer length. A problem relating simulated and experimental residual dipolar couplings is discussed. We apply our generated ensembles to the problem of folding kinetics, by examining whether the ensembles of some proteins are closer geometrically to their folded structures than others. We find that for a randomly selected dataset of 15 non-homologous 2- and 3-state proteins, quantities such as the average root mean squared deviation between the folded structure and unfolded ensemble correlate with folding rates as strongly as absolute contact order. We introduce a new order parameter that measures the distance travelled per residue, which naturally partitions into a smooth "laminar" and subsequent "turbulent" part of the trajectory. This latter conceptually simple measure with no fitting parameters predicts folding rates in 0 M denaturant with remarkable accuracy ($r = -0.95$, $p = 1 \times 10^{-7}$). The high correlation between folding times and sterically modulated, reconfigurational motion supports the rapid collapse of proteins prior to the transition state as a generic feature in the folding of both two-state and multi-state proteins. This method for generating unfolded ensembles provides a powerful approach to address various questions in protein evolution, misfolding and aggregation, transient structures, and molten globule and disordered protein phases. © *2013 AIP Publishing LLC*. [http://dx.doi.org/10.1063/1.4817215]

I. INTRODUCTION

Structural biology has historically been grounded by several landmark studies wherein the atomic coordinates of several large molecules have been experimentally determined, giving insight into the mechanisms of their biological function. Principle relatively recent examples include the photosynthetic reaction center,[1] potassium channels[2] and aquaporins,[3] the ribosome,[4–6] the RNA polymerase II transcription complex,[7] and G protein-coupled receptors.[8,9] Despite the triumphs of the structure-function paradigm, there has been emergent evidence of the biological importance of intrinsically disordered proteins[10–15] for which atomic coordinates significantly fluctuate so that a three-dimensional structure is poorly defined, at least in the absence of binding partners[16] or osmotic stabilizing agents.[17]

While current nuclear magnetic resonance (NMR) measurements of chemical shifts, residual dipolar couplings (RDCs), and ^3J couplings can provide information on structural preferences, computer simulations can provide critical information of disordered proteins through an explicit geometrical knowledge of the conformational ensemble. There have thus been recent efforts towards a computational characterization of both intrinsically disordered protein (IDP) ensembles[18–23] and chemically denatured ensembles of natively folded proteins.[24,25] One difficulty however is the time-scales necessary to sample a sufficiently large set of conformations to represent microscopic equilibrium properties of the unfolded ensemble.

In what follows, we first describe our method for generating a diverse, representative ensemble for the unfolded state of a protein. We apply this method here to several proteins, including 4 IDPs and 17 α, β, or mixed natively folded proteins.

As an application of unfolded ensembles, we investigate transformations between unfolded and folded structures, for natively ordered proteins. We ask whether the distance covered during such transformations can predict folding kinetics, and we find several geometrical transformation measures that indeed correlate with folding kinetics for both 2-state and 3-state kinetic folders.

Numerous experiments have pointed to a kinetic molten globule – a semi-collapsed but hydrated state which may have significant secondary structure – as a generic feature of many

[a]Present address: Recon Instruments #100, 1050 Homer Street, Vancouver, British Columbia V6B 3W9, Canada.
[b]Electronic mail: steve@phas.ubc.ca

unfolded proteins in the absence of denaturant (reviewed in Ref. 26). A collapsed ensemble prior to the transition state implies that reconfigurations in the presence of strong bonding interactions and significant steric constraints would be relevant to the folding barrier. Indeed such reconfiguration would be expected to be more substantial if more long range interactions were present, providing a potential explanation for the success of contact order in determining folding rates[27] that is somewhat distinct from explanations involving Flory entropy in loop closure. It is possible, however, that more accurately characterizating such motions could lead to even stronger correlates with folding barriers. We explore this notion and find such a characterization in the distance on average that the protein must travel to adopt the native structure. Taking a portion of the transformation distance where the trajectories become "turbulent" gives a remarkably strong correlation with folding rates in 0 M denaturant.

We organize this paper by first describing our method for generating a conformationally diverse unfolded ensemble, which applies to both intrinsically folded and intrinsically unfolded proteins. An extension of the method to proteins containing disulfide bonds is described and applied. Properties of the unfolded ensemble, including chemical shifts, residual dipolar couplings, structural diversity, clustering properties, and scaling of the radius of gyration with polymer length are described. We then explore minimal transformations between the unfolded ensemble and the native structure for intrinsically foldable proteins, and investigate the correlation with folding rates of such minimal transformations.

II. METHODS

In this section, we first describe our method for generating unfolded ensembles. We then describe our method for generating collections of optimized pathways from these unfolded configurations to the folded structure. The proteins analyzed in this paper are given in Table I. They consist of 8 2-state folders, 9 3-state folders, 6 α-helix proteins,

TABLE I. Proteins and their properties used in this study.[a]

PDB	kin[b]	SS[c]	log (k_f)[d]	log (k_u)[d]	log (k_{mp})[e]	N[f]	ν[g]	ACO	⟨RMSD⟩[h]	⟨TM⟩	⟨GDT⟩	$\langle\mathcal{D}_{NC}\rangle$	$\langle\mathcal{D}\rangle$	$\langle\mathcal{D}^{(lam)}\rangle$	$\langle\mathcal{D}^{(turb)}\rangle$
1L2Y	2	α	12.5	11.5	13	20	0.48	3.7	6.38	0.173	0.462	…	…	…	…
1ENH	2	β	10.5	7.6	8.1	54	0.57	7.4	14.81	0.147	0.218	14.1	27.0	15.3	11.7
1SHG	2	α	1.1	−4.8	−3.7	57	0.6	10.9	16.68	0.133	0.259	16.7	33.5	11	22.5
2CRO	3	β	3.7	−0.5	0.3	65	0.66	7.3	14.27	0.144	0.218	14.6	30.2	11.7	18.5
1CSP	2	β	6.5	2.3	2.7	67	0.57	11	18.01	0.129	0.160	18.3	31.5	16.8	14.7
1VII	2	α	9.4	5.3	10.6	36	0.57	4	9.68	0.149	0.347	8.2	19.2	13.1	6.1
2PDD	2	α	9.8	5.4	9.8	43	0.62	4.8	11.63	0.155	0.302	11	21.3	13.9	7.4
1BNI	3	$\alpha\beta$	2.6	−9.1	−4.3	108	0.57	12.3	21.5	0.127	0.132	22.8	39.8	18.4	21.4
1APS	2	$\alpha\beta$	−1.6	−9	−3.3	98	0.58	21.8	24.36	0.124	0.140	24.8	44.2	17.6	26.6
1A6N	3	α	1.1	−3.8	−1.4	151	0.53	14	28.01	0.131	0.113	28.1	45.7	20.9	24.8
1CBI	3	β	−3.2	−9.8	−6.7	136	0.62	18.8	27.09	0.112	0.112	27.2	45.5	20.1	25.4
1TIT	3	β	3.6	−7.6	−6.9	89	0.49	15.8	20.39	0.126	0.135	20.4	36.5	19.6	17.0
1IMQ	2	α	7.3	−1.9	−1.4	86	0.68	10.4	18.3	0.137	0.203	17	32.7	17.9	14.8
1PSF	3	β	3.2	…	…	69	0.61	11.7	18.43	0.131	0.254	17.8	…	…	…
2A5E	3	$\alpha\beta$	3.5	0.2	0.4	156	0.52	8.3	23.26	0.127	0.120	24.5	41.1	18.2	22.9
2RN2	3	$\alpha\beta$	0.1	−12	−4.6	155	0.54	19.3	28.69	0.129	0.106	31	43.0	18.5	24.4
1RA9	3	$\alpha\beta$	−2.5	−6.1	−5.2	159	0.52	22.3	24.73	0.140	0.110	27.9	47.8	19.1	28.7
IN[i]	…	…	…	…	…	60	0.57	2.7	12.7	0.143	0.233	…	…	…	…
1IYT	…	…	…	…	…	42	0.58	2.5	10.4	0.143	0.446	…	…	…	…
1XQ8	…	…	…	…	…	140	0.67	2.9	22.1	0.121	0.098	…	…	…	…
proTα	…	…	…	…	…	129	0.63	2.5	21.9	0.129	0.130	…	…	…	…

[a]Proteins include Trp-cage miniprotein (1L2Y), Engrailed homeodomain (1ENH), src-homology 3 (SH3) domain (1SHG), phage 434 cro protein (2CRO), cold shock protein (1CSP), chicken villin headpiece (1VII), peripheral subunit-binding domain of dihydrolipoamide acetyltransferase (2PDD), barnase (1BNI), acylphosphatase (1APS), deoxy-myoglobin (1A6N), apo-cellular retinoic acid binding protein I (1CBI), titin, IG repeat 27 (1TIT), colicin E9 immunity protein IM9 (1IMQ), photosystem I protein (1PSF), tumor suppressor P16INK4A (2A5E), ribonuclease H (2RN2), dihydrofolate reductase (1RA9), N-terminal domain of HIV Integrase (IN; IUP structure generated from sequence[28]), amyloid beta-peptide (1-42) (1IYT), alpha-synuclein (1XQ8), Prothymosin alpha (proTα; IUP structure generated from sequence[28]). 1L2Y and 1PSF were used only for unfolded ensemble generation; 1PSF does not have published unfolding or transition midpoint rates, and 1L2Y is temperature-denatured rather than chemically denatured.

[b]Indicates 2-state or 3-state kinetics.

[c]Secondary structure content.

[d]Natural logarithm of experimentally determined refolding and unfolding rates in 0 M denaturant.[29]

[e]Midpoint experimental relaxation rate.[29]

[f]Chain length.

[g]Scaling exponent of the radius of gyration with chain length N.

[h]All distance and alignment metrics are averaged over the equilibrium unfolded ensemble, as indicated by angle brackets. RMSD = root mean squared deviation in Å, TM = score from template modeling alignment, GDT = global distance test- total score, $\mathcal{D}_{NC}$ = Distance between an unfolded conformation and the native, accounting for polymer non-crossing,[30] $\mathcal{D}$ = Geometrical pathways (GP) generated distance,[31] $\mathcal{D}^{(lam)}$ = Laminar component of the GP distance, $\mathcal{D}^{(turb)}$ = Turbulent component of the GP distance. All distance metrics are in units of Å and are an average *per residue*.

[i]For IUP proteins, numbers for ACO indicate an average over all structures in the unfolded ensemble. Numbers for ⟨RMSD⟩, ⟨TM⟩, and ⟨GDT⟩ indicate an average over all pairs in the unfolded ensemble.

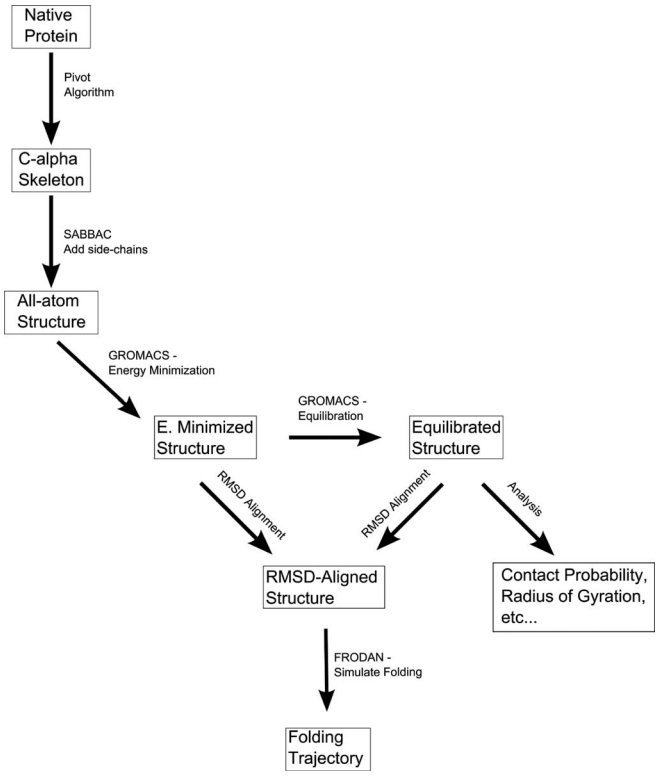

FIG. 1. Overview of algorithm.

6 β-sheet proteins, 5 mixed α/β proteins, and 4 intrinsically disordered proteins. This dataset is not large, but it is diverse, spanning a wide range of rates, size, and structural classes. Native structural homologs as defined through TM-score were not included in the dataset; otherwise, no additional pruning or selection of proteins was made.

A. Generating diverse ensembles of unfolded configurations

We generate unfolded ensembles by employing the following method:

- Generate a diverse coarse-grained (CG) ensemble
- "Foliate" each structure by adding backbone/side-chain degrees of freedom
- Equilibrate each foliated structure for a short time.

The steps in the method are shown schematically in Figure 1. First, if the native crystal or NMR structure existed, it was used as a starting point. If a protein or peptide was intrinsically unfolded or if no pdb file was available, the proteins initial structure was generated *a priori* by submitting the sequence to the I-TASSER structure prediction server,[28] then minimizing and equilibrating the structure. In all cases, the initial structure was used solely as a starting point for the structure generation algorithm.

1. Pivot and crankshaft moves

We coarse grain the initial structure by retaining only the C_α coordinates. This CG structure is then altered by employing a generalization of the pivot algorithm,[19,30,32–34] an efficient algorithm for generating self-avoiding random walk ensembles. We have previously implemented this generalization to generate CG C_α-model unfolded ensembles.[30] For the CG structure, a pivot move selects a particular bond angle and its corresponding dihedral angle at random, and then resamples them from a native-centric Boltzmann distribution[30] (see Figure 2). The corresponding phenomenological energy function contains approximate angle and dihedral stiffness parameters. If after a pivot move the chain sterically interferes with itself, the move is discarded. For a chain of length N, pivot moves are repeatedly attempted until $\mathcal{O}(N)$ successful pivot moves are implemented, such that on average, one pivot move per residue is obtained.

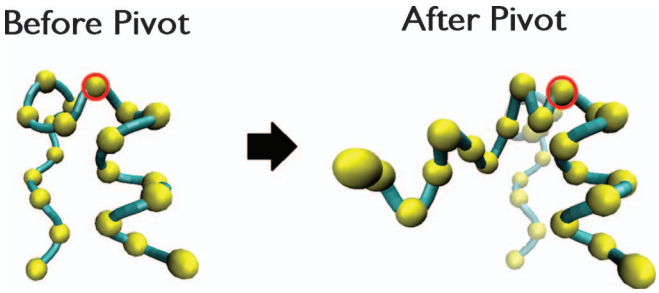

FIG. 2. Illustration of an example pivot move for PDB 1L2Y.

We also considered generated ensembles for proteins with disulfide bonds present. In this case, the pivot algorithm cannot be directly implemented because the disulfide constraint correlates the position of two remote parts of the chain. In this case, we implemented an alternative move for residues bounded by those participating in the disulfide bond. We implemented this procedure for human superoxide dismutase (SOD1), a 153 aa protein that contains a disulfide bond between C57 and C146.

We start by picking a residue at random. If, for a disulfide-bonded protein such as SOD1 (1HL5) (Fig. 3), the residue is before 57 or after 146, a pivot move is implemented for the part of the chain N-terminal or C-terminal to the selected residue. This preserves the coordinates of the disulfide loop. If, however, the selected residue is between 57 and 146, it lies inside the disulfide loop; in this case a non-local "crankshaft" move is implemented between the selected residue and another randomly chosen residue inside the

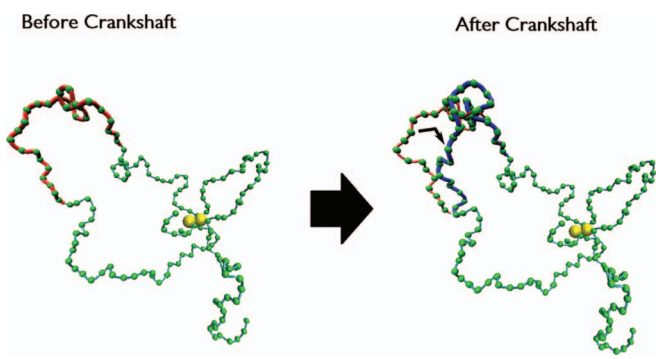

FIG. 3. Crankshaft move for SOD1 (PDB 1HL5), a protein with a long-range disulfide bond between C57 and C146. A minimized, non-equilibrated configuration is shown.

disulfide loop. This "large step" move, which is similar in spirit to the "small-step" biased gaussian steps in torsional space developed by Favrin *et. al.*,[35] randomizes the dihedral angles between the two residues, subject to the constraint that the end point C_α coordinates are fixed. An example crankshaft move is shown in Figure 3. Fixed end-point moves have been studied in detail previously by several authors.[35–39] One advantage of the method we employ here is that the moves result in large global changes of an arbitrarily large collection of contiguous residues. A disadvantage, however, is that the algorithm in its current form is not fast. The trade-off between speed and global reconfigurational change was acceptable for our methodology.

Consider two randomly selected C_α atoms i, j with $j > i$. The deviation in end point position $\delta \boldsymbol{r}_j$ may be Taylor expanded as a function of the $n = j - i - 1$ dihedral angles in the coarse-grained structure between i and j whose rotations would alter the position $\boldsymbol{r}_j$

$$\delta \boldsymbol{r}_j = \sum_{k=1}^{n} \frac{\partial \boldsymbol{r}_j}{\partial \phi_k} \delta \phi_k \,, \tag{1}$$

so that the square of the end point position is given by

$$\delta \boldsymbol{r}_j^2 = \sum_{k,\ell=1}^{n} \delta \phi_k G_{k\ell} \delta \phi_\ell, \tag{2}$$

where the $n \times n$ non-negative, symmetric matrix $\mathbf{G}$ has elements

$$G_{k\ell} = \frac{\partial \boldsymbol{r}_j}{\partial \phi_k} \frac{\partial \boldsymbol{r}_j}{\partial \phi_\ell} \,. \tag{3}$$

The 3 constraints corresponding to $\delta \boldsymbol{r}_j = 0$ means that 4 or more dihedrals are required to have a mode corresponding to an eigenvector of $\mathbf{G}$ with zero eigenvalue, i.e., $j - i = 5$ or greater. However, for $j - i \geq 5$, there will be combinations of dihedrals that leave $\boldsymbol{r}_j$ unchanged. We randomly select one of these eigenvectors and implement a small rotation $\delta \boldsymbol{\phi}$ along it. The end point position generally moves slightly because Eq. (2) is only zero for infinitesimal displacements. We repeatedly implement multiple rotation vectors $\delta \boldsymbol{\phi}$ that leave the end point position $\boldsymbol{r}_j$ nearly unchanged. Occasionally, the deviation in the end point position becomes appreciable (fractions of an Å). We then correct the end point deviation $\delta \boldsymbol{r}_j$ by inverting Eq. (1) to find the vector $\delta \boldsymbol{\phi}$, with components

$$\delta \phi_k = -\left(\frac{\partial \boldsymbol{r}_j}{\partial \phi_k} \right)^{-1} \cdot \delta \boldsymbol{r}_j \,. \tag{4}$$

This gives a set of dihedral rotations that rotates the end point back to its original position. In total, we implement rotations roughly 20 times per bond angle, i.e., $20(j - i)$ times, correcting as needed. The net result is to change the coordinates in a crankshaft fashion, as shown in Figure 3. This constitutes one update of the configuration. As mentioned above, $\sim N$ updates are taken before sampling a new distinct configuration for foliation by adding the remaining side-chain and backbone atoms. We have illustrated non-local crankshaft moves here for the case of a disulfide bonded protein. However, for the remainder of the analysis we consider only non-disulfide-bonded proteins. In principle, one could implement a combination of pivot moves and crankshaft moves for all proteins. Because our unfolded ensembles were not dense, it was more efficient to implement pivot moves; for dense systems however such as polyglutamine repeats for example, it is an interesting future topic to explore combined move sets.

2. Foliation, minimization, and equilibration

Once a distinct CG conformation is obtained by pivot/crankshaft moves, side-chains and backbone are added. We examined two methods of adding atoms, `PULCHRA`[40] and `SABBAC`,[41] which yielded equivalent results; `PULCHRA` is available as an executable program, while calls to `SABBAC` must be uploaded to a server – in an automated way for a large number of configurations.

After side-chain and backbone atoms are added, the protein is energy minimized in GROMACS using a steepest descent algorithm to eliminate steric clashes; Figure 4 shows a rendering of the process for Trp cage (PDB 1L2Y). Though the configuration itself does not change much (RMSD values are approximately 1.4 Å), the energy typically decreases by several orders of magnitude. The radius of gyration typically increases slightly during this process – zero temperature energy minimization tends to favor extended chains when starting from a random initial structure.

The `SABBAC` and `PULCHRA` algorithms occasionally place side chains in sterically clashing positions that are not ameliorated by minimization. Roughly 15% of the initial states are not viable for equilibration and are discarded.

For disulfide bonded proteins, the residues involved in the disulfide bond, along with the adjacent residues in the primary sequence (e.g., residues 56-58 and 145-147 for SOD1), are all held fixed during crankshaft and pivot moves. This allows the side-chains of the disulfide bond (between residues 57 and 146 for SOD1) to be reconstructed from the initial configuration. Essentially, no conformations are lost by this

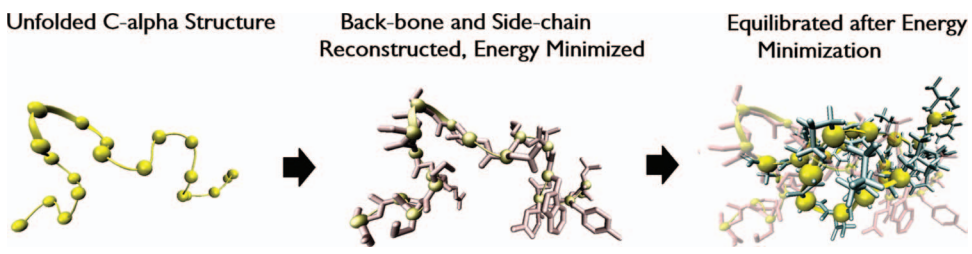

FIG. 4. Schematic figures indicating the processes of backbone and side-chain addition, energy minimization, and 1 ns thermal equilibration.

constraint after equilibration – the arbitrary origin of the coordinates can be thought of as centered around the disulfide bond. PULCHRA and SABBAC may also occasionally protonate histidines on a nitrogen that is likely unprotonated in the native structure. This may be corrected by applying a patch to properly protonate histidines in the unfolded conformational ensemble. We have generally not implemented such a patch, except for the ensemble of SOD1 conformations that are provided in the supplementary material.[47]

Viable all-atom configurations were then thermally equilibrated by molecular dynamics (MD) in explicit SPC water with CHARMM27 force field, using the GROMACS simulation package. All protein atoms are initially at least 20 Å from the faces of a cuboid aligned with the axes of the gyration tensor and having periodic boundary conditions. Simulation conditions were in the *NPT* ensemble at $T = 300$ K using the modified Berendsen (V-rescale) weak coupling thermostat, and 1 bar using the Parrinello-Rahman barostat. The Particle Mesh Ewald method was applied for long-range electrostatics, and a 10 Å cut-off was used for non-bonded electrostatic and van der Waals (VDW) interactions. Covalent bonds lengths to hydrogen atoms were constrained using the LINCS algorithm. The integration time-step was 2 fs, and coordinates were saved every 100 ps. Each initial configuration is simulated generally for 1 ns. For α-synuclein, proTα, and SOD1, configurations were simulated for 5 ns. A sample of 1000 equilibrated conformations from the unfolded ensemble of the disulfide-bonded protein superoxide dismutase (WT SOD1) is provided in the supplementary material.[47]

Several previous approaches have been used to generate unfolded ensembles. Zagrovic and Pande[42] have run thousands of independent MD trajectories in implicit GB/SA solvent for 3 small proteins: Villin, Trp zipper, and BBA5. Each of these simulations all started from the fully extended state ($\phi = -135^\circ$, $\psi = 135^\circ$), mandating individual simulations at least 10 ns long for convergence of quantities such as the radius of gyration. One advantage we found for generating random initial ensembles was that many global properties such as the radius of gyration often came to equilibrium on the 1 ns time scale. Pappu and colleagues[19] have run 50 iterations of 12 ns simulations for 90 randomly generated initial configurations of Gln_5 and Gln_{15}, in explicit TIP4P solvent. Their procedure is similar to the one we employed here, except that there was no coarse-graining step, as we have employed. Other methods of generating unfolded ensembles have involved extracting conformational fragments from databases of high-resolution crystal structures. Sosnick and colleagues[24] find the conformation of an amino acid (accounting for nearest neighbor effects) using a statistical potential for regions outside of helices, sheets, and turns. If there is a steric clash, the conformation is nudged until the clash is resolved. This process certainly perturbs the ensemble from the equilibrium one,[34,43,44] and as well the resulting conformations are not subsequently equilibrated. Still, the model shows excellent agreement with experimental RDCs, for the experimental scenario of chemically denatured proteins in highly anisotropic media. Blackledge and colleagues[18,25] employ a similar approach of extracting conformational fragments from a database. Chains are grown by randomly selecting ϕ/ψ angles from a database of sequence-specific fragments outside of helices or sheets (turns are now included). If in the growth of the chain a steric clash ensues, those angles are rejected and another angle pair is selected until no steric clash is found. This process also perturbs the ensemble from the equilibrium one,[34,43,44] and as above the conformations are not equilibrated. Nevertheless, this model also shows excellent agreement with experimental RDCs in conditions of high denaturant and high anisotropy.

B. Chemical shifts and residual dipolar couplings

Chemical shifts were obtained using the program CAMSHIFT.[45] Residual dipolar couplings were obtained using the program PALES.[46] Chemical shifts and RDCs for the proteins in this study are tabulated in the supplementary material.[47]

C. Investigating the correlation between geometrical folding pathways and folding kinetics

We seek direct transformations between an unfolded configuration and the native configuration, to give a measure of average distance between the unfolded ensemble and the folded structure. We considered two generated ensembles: one energy minimized and equilibrated for 1 ns, and another only energy minimized. We quantified the unfolded-folded distance in several ways:

- RMSD to the native structure: calculated for C_α atoms, and then averaged over the unfolded ensemble, which we denote by $\langle RMSD \rangle$.[48]
- Ensemble-averaged TM-score: $\langle$TM-score$\rangle$
- Ensemble-averaged global distance test – total score: $\langle$GDT-TS$\rangle$
- The distance accounting for polymer non-crossing constraints $\langle \mathcal{D}_{NC} \rangle$, as calculated by the algorithm developed by Mohazab and Plotkin.[30]
- The distance corresponding to the RMSD-minimized trajectories generated by the Geometrical Pathways (GP) algorithm of Farrell *et al.*[31]
- A variation of the GP distance where only the smooth "laminar" part of the trajectories are taken, as described below.
- A variation of the GP distance where only the fluctuating "turbulent" part of the trajectories are taken, as described below.
- A variation of the GP distance, wherein trajectories are smoothed by applying a weighting function given by $(\frac{1}{3}, \frac{2}{3}, 1, \frac{2}{3}, \frac{1}{3})$ to consecutive sets of 5 points along the trajectory. This procedure eliminates jagged edges along the trajectory.

1. Order parameters for unfolded structures

The ensemble-averaged order parameters described in this section are all given in Table I for the proteins in this study.

To examine RMSD, structures in the unfolded ensemble were RMSD-aligned to the native structure, and the average (residual) RMSD was calculated, using the software program VMD.[49]

TM-score was calculated with TM-ALIGN,[50] and is given by

$$\text{TM-score} = \max \frac{1}{L_{targ}} \sum_{i=1}^{L_a} \left[1 + \left(\frac{d_i}{d_o(L_{targ})} \right)^2 \right]^{-1},$$

where L_{targ} is the length of the target protein that another protein structure is aligned to, L_a is the number of template-aligned C_α pairs, d_i is the distance between the ith pair of aligned pairs, and $d_o(L_{targ}) = 1.24(L_{targ} - 15)^{1/3} - 1.8$ is a distance parameter that ensures that the average TM-score is not dependent on protein length. max indicates the maximization of this quantity by alignment.

GDT-TS[51] was calculated by file upload to the KoBaMIN web server,[52] and is calculated by

$$\text{GDT-TS} = \max \frac{1}{4N} \left(C_{d/4} + C_{d/2} + C_d + C_{2d} \right),$$

where N is the chain length, $d = 4$ Å is a distance threshold, and $C_{d/4}$, for example, is the number of residues superposed below a threshold of $d/4$ after alignment.

Distance $\mathcal{D}_{NC}$ accounting for polymer non-crossing was calculated using the method in Ref. 30, which is based on the calculation of minimal distance trajectories that we have developed previously.[53–55] This method calculates the approximate distance undertaken by all C_α atoms in transforming between two structures, e.g., an unfolded structure and the native structure, while accounting for polymer non-crossing constraints. The method involves a depth first tree search algorithm to find the shortest distance trajectories between two conformations for a linear self-avoiding polymer. We apply this method here between the native structure and 200 coarse-grained unfolded structures. After coarse-graining (smoothing) conformations by sampling every other bead, each structure was transformed to the folded state by the algorithm discussed in Ref. 30, and the minimal distance cost was found.

The geometrical pathways[31] distance $\mathcal{D}$ was found using the program FRODAN, which calculates a stereochemically acceptable transformation between two all-atom structures, by following a steepest descent pathway that minimizes RMSD. From such transformations, we calculate the distance that C_α atoms have moved. Sample folding trajectories of 5 C_α atoms are shown in Figure 5. The total arc length that all C_α residues have travelled is accumulated, to obtain the total distance for one conformation pair. Similarly to TM-score and GDT-TS, this total is then divided by chain length, which yields the mean distance travelled per residue, for one conformation pair,

$$\mathcal{D} = \frac{1}{N} \sum_{i=1}^{N} \int_{\boldsymbol{r}_i^{(\alpha)}}^{\boldsymbol{r}_i^{(N)}} |d\boldsymbol{r}_i| \ . \tag{5}$$

Here, the sum is over the N C_α atoms, and the integral sums up arc-length increments from initial to final position for each C_α atom (see Figure 5). For the proteins in Table I, the mean distance is about 36 Å.

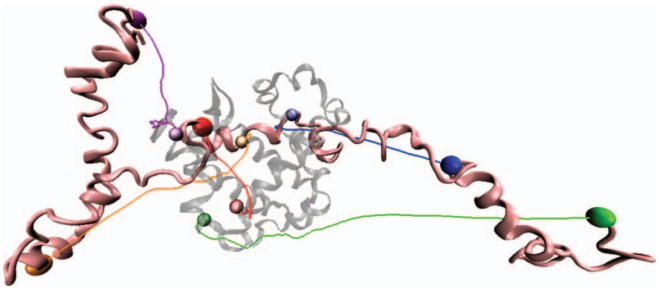

FIG. 5. Example optimal folding trajectories for 5 C_α atoms in apo-myoglobin (1A6N). Unfolded and folded structures are also shown.

Particle trajectories obtained from the GP method tended to be delineated by two parts, an early smooth "laminar" segment, and a late rugged "turbulent" segment (Figs. 6(a) and 6(b)). Figure 6(c) plots the distance travelled per step as a function of step index along a C_α trajectory. The turbulent

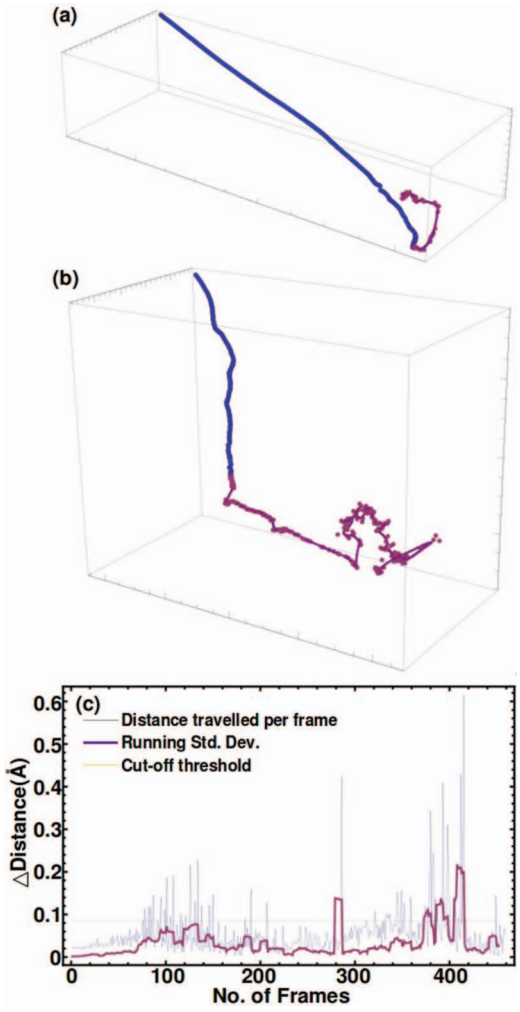

FIG. 6. Each C_α trajectory is divided into a smooth "laminar" and rugged "turbulent" part. Panels (a) and (b) show sample trajectories for $C_\alpha(4)$ and $C_\alpha(75)$ of apo-myoglobin. Panel (a) is predominantly laminar – the corresponding distances are $\mathcal{D}^{(lam)} = 51$ Å, $\mathcal{D}^{(turb)} = 7$ Å. Panel (b) is predominantly turbulent – the corresponding distances are $\mathcal{D}^{(lam)} = 4.6$ Å, $\mathcal{D}^{(turb)} = 12$ Å. (c) Criterion for determining the transition from laminar to turbulent trajectories. When the root variance in the distance travelled per step jumps above a threshold given by 7 times the baseline value, the trajectory from then on is defined as turbulent.

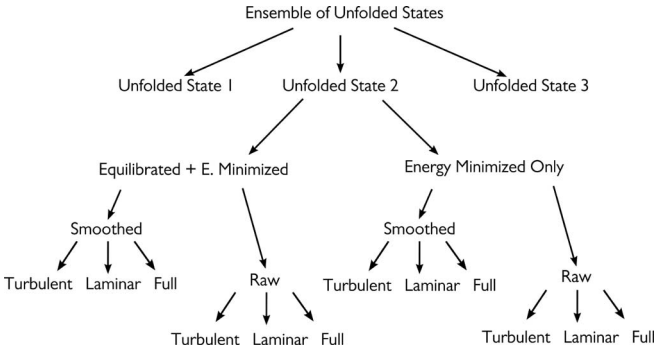

FIG. 7. Different ensembles considered in this study to compare with protein folding kinetics.

segment is characterized by heterogeneous jumps of variable distance. The root variance (standard deviation) in step length, averaged over 10 frames, is also plotted. We determine a baseline standard deviation by averaging the first 50 frames, then when the standard deviation exceeds a threshold which we set to 7 times the baseline, the trajectory is deemed turbulent from then on. We thus partition each trajectory up into laminar and turbulent parts, and define the corresponding accumulated distances from the unfolded structure (α) to the native structure (N)

$$\mathcal{D}^{(lam)} = \frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{r}_i^{(\alpha)}}^{\boldsymbol{r}_i^{(cut)}} |d\boldsymbol{r}_i|\,,$$

$$\mathcal{D}^{(turb)} = \frac{1}{N}\sum_{i=1}^{N}\int_{\boldsymbol{r}_i^{(cut)}}^{\boldsymbol{r}_i^{(N)}} |d\boldsymbol{r}_i|\,. \qquad (6)$$

The transition to turbulence can also be signalled by an abrupt increase in the curvature of the trajectory as shown in Figure S1 of the supplementary material.[47]

Choosing either the minimized or equilibrated ensembles, the smoothed or raw trajectories, and the laminar, turbulent, or full trajectories, gives 12 different measures of distance, as depicted in Figure 7.

Both 2 and 3 state proteins were selected from proteins with known kinetics using the webserver KineticDB.[29] Proteins were ensured to be non-homologous by TM-score, as compared to a non-redundant (NR) protein database[56] (see Fig. 8).

III. RESULTS AND DISCUSSION

A. Chemical shifts and residual dipolar couplings

Chemical shift values for C_α atoms are obtained using the program CAMSHIFT.[45] These values are plotted for Aβ_{1-42} in Figure 9, which show good agreement with the experimental values of Hou *et al.*[57] Chemical shift values were obtained from a generated ensemble of 773 structures. These numbers agree with those from 1 μs explicit water simulations.[22] That said, even better agreement with experimental values is obtained from chemical shifts compiled from databases of loop regions in protein structures (e.g., CAMCOIL,[58] $r = 0.99$).

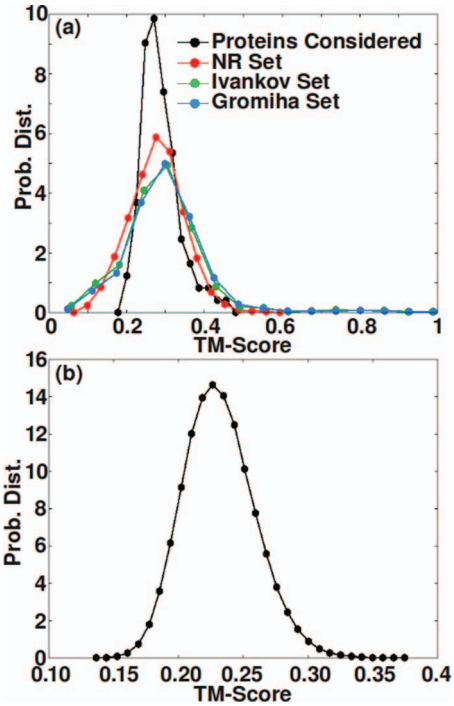

FIG. 8. (Panel (a)) TM-score distributions between *native* structures, showing homology of our dataset compared to a NR dataset,[56] and other datasets used for protein folding kinetics analysis.[84,118] One can see some homologous protein pairs in other datasets. (Panel (b)) TM-score distribution between 1299 unfolded states for α-synuclein. Similar distributions are obtained for other proteins.

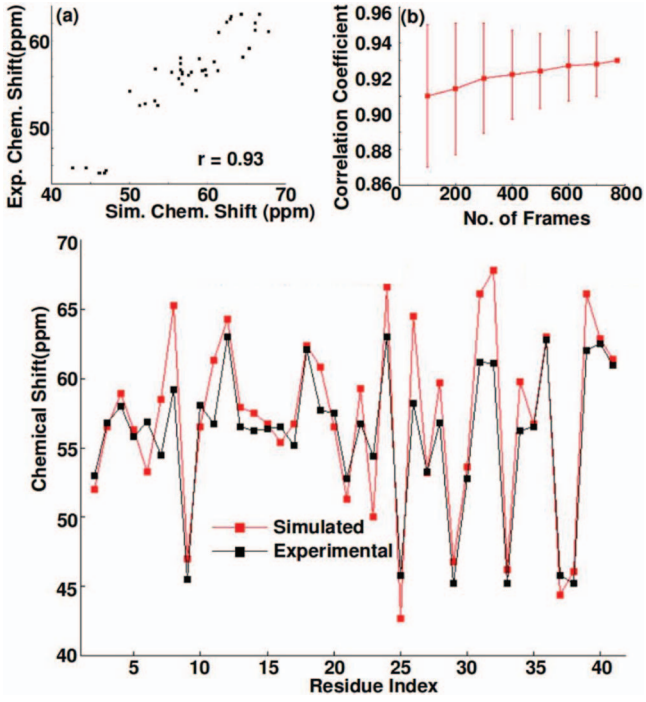

FIG. 9. Comparison between experimental and simulated $^{13}C_\alpha$ chemical shift values, for Aβ_{1-42}. (Main panel) Black data points are experimental values from Ref. 57, red data points are those from the simulated ensemble of 773 conformations, using CAMSHIFT. (Inset (a)) Scatter plot of experimental *vs* simulated chemical shifts ($r = 0.93$). (Panel (b)) Convergence study of the correlation coefficient between experimental and simulated data. Mean correlation coefficient is shown; vertical bars indicate the standard deviation of correlation coefficient values when random subsets with a given number of frames are taken from the total dataset.

RDCs measure the nuclear dipole coupling between two spin 1/2 nuclei such as the amide H and labelled ^{15}N. This coupling between nuclei i and j depends on the angle the internuclear bond makes with the orientation of the whole molecule, and is given in the principle axis frame by

$$D_{ij}(\theta,\phi) = \frac{D_{max}}{2}\left[S_{zz}(3\cos^2\theta - 1) + (S_{xx} - S_{yy})\sin^2\theta\cos 2\phi\right], \quad (7)$$

where S_{zz} is the axial component of the alignment tensor, and $(S_{xx} - S_{yy})$ is the rhombic component of the alignment tensor, and $|S_{zz}| > |S_{yy}| \geq |S_{xx}|$. The angles θ and ϕ give the orientation of the vector in the principle basis, and D_{max} sets the scale of dipolar interactions and is given by $D_{max} = \gamma_i\gamma_j\mu_o h/(8\pi^3\bar{r}_{ij}^3)$, where $\bar{r}_{ij}$ is the effective internuclear distance accounting for libration of internuclear vector,[59] γ_i, γ_j are gyromagnetic ratios of nuclei i, j, μ_o is the magnetic permeability of vacuum, and h is Planck's constant.

RDCs were obtained from simulated ensembles using `Pales`;[46] we have provided tables of amide NH RDC values for our generated ensembles in the supplementary material.[47] While simulated RDCs of native or near-native structures show good correlation with experimental RDCs, simulated RDCs of unfolded ensembles have not shown strong correlation with experimental RDCs in previous studies;[60] as well, we also did not see significant correlation with experimental values.

There are many potential reasons for this. Experimental RDC values change significantly depending on the degree of extension of the orienting liquid crystal (e.g., polyacrylamide), and also vary with denaturant concentration and pH. The chemically denatured ensemble may significantly differ from the unfolded ensemble. Unfolded ensembles may show partial native structure,[61] and at least partial collapse,[62,63] depending on net charge or hydrophobicity. Urea or GuHCl-denatured ensembles on the other hand tend to exhibit relatively simple self-avoiding walk behavior.[64–66] Finally, because steric obstacles represent absorbing boundaries for the probability distribution of a disordered polymer, the effect of steric hindrance itself, due to the aligning liquid crystals, significantly modifies the structure of the unfolded ensemble. Thus, the experimental conditions under which RDCs are obtained may result in very different ensembles than those in simulation, which model an isolated protein in the absence of denaturant at pH $\sim$ 7.

In spite of this, several recent models of the unfolded state show remarkable agreement with experimental RDCs in denaturant and in stretched polyacrylamide.[18,24,25] These models reconstruct the unfolded ensemble from unstructured elements in the protein data bank, and evidently are very good models of unfolded proteins in aligning media. It is not clear on the other hand how such models would predict RDCs as external conditions were varied, e.g., as the polyacrylamide were relaxed or compressed, without some phenomenological adjustment. It may be best to think of the RDC values we obtain here as "unperturbed" RDCs arising solely due to correlations between bond vectors and the inherent anisotropy of the polymer, rather than the induced anisotropy.

B. Polymer scaling laws and persistence length

After equilibration, the radius of gyration R_G was obtained for all subsequences with length $\leq N$, for each protein in Table I. The slope of the log-log plot gives the exponent of the scaling law $R_G = r_1 N^{\nu}$, where in 3-dimensions $\nu = 3/5$ for a self-avoiding random walk, 1/2 in the Θ-state, and 1/3 for a compact globule state.[67] Figure 10(a) shows a plot of the radius of gyration *vs* timestep for several trajectories along with the average over trajectories, for the 129 aa intrinsically unfolded protein (IUP) proTα. Figure 10(b) shows the scaling law for the radius of gyration obtained by taking subsequences of the full length protein and averaging R_G for those lengths over the 5 ns equilibrium ensemble.

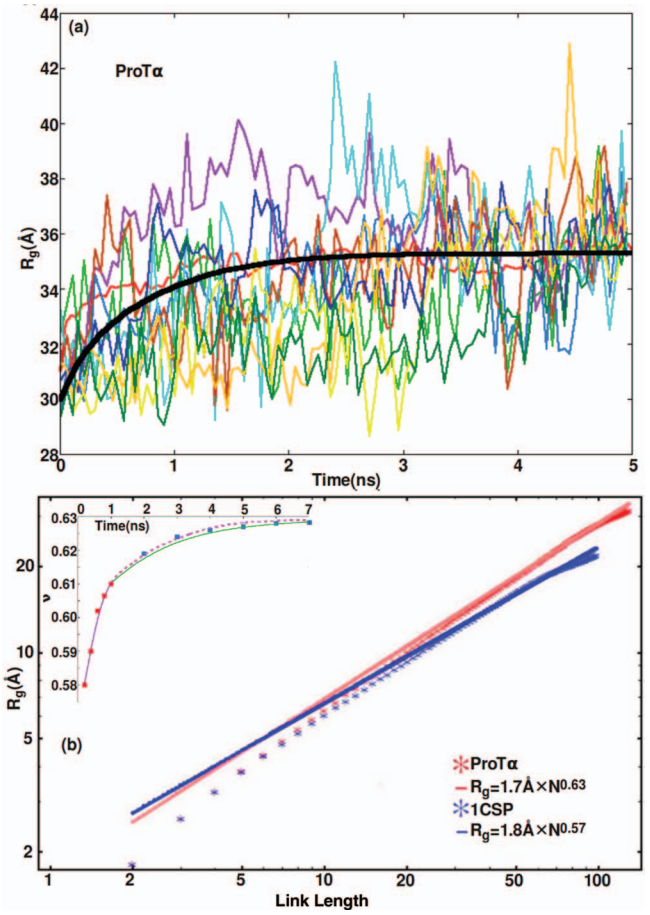

FIG. 10. (a) Radius of gyration vs. time (equilibration process), for proTα: a highly charged, intrinsically disordered protein. The relaxation time is about 0.8 ns, and the asymptotic value of the radius of gyration R_G is about 35.5 Å. (b) Scaling of the radius of gyration R_G with chain length, obtained by taking all subsections of a given length and finding the ensemble averaged radius of gyration. (Inset) Extrapolation procedure to find the asymptotic value of the scaling exponent ν. The value of ν is obtained for ensembles at a given equilibration time. This value converges exponentially to the $t \to \infty$ value. Extrapolation from ensembles with $t \leq 1$ ns gives an asymptotic value of 0.633, while extrapolation from ensembles with $t \leq 5$ ns gives an asymptotic value of 0.631. A similar conclusion was obtained from extrapolation of the data for α-syn. Thus, extrapolation of ν from $t \leq 1$ ns ensembles is likely to be sufficiently accurate in general.

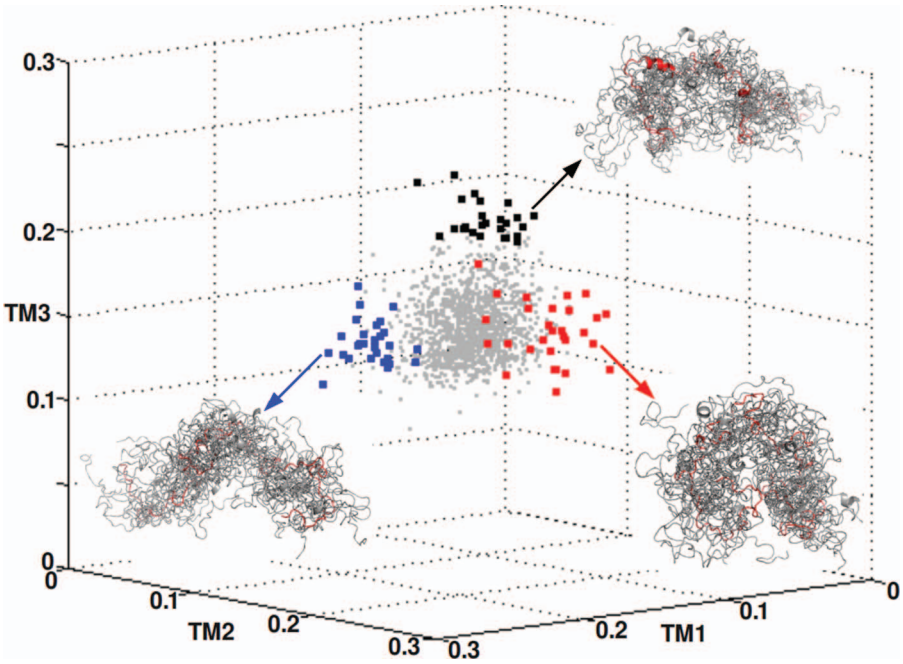

FIG. 11. Nearest neighbor clustering using TM-score of 1299 structures of α-synuclein, projected onto the TM-scores to the centroid structures of the largest three clusters (blue, red, and black, respectively). Representative conformations in each cluster are shown. The lack of distinct clustering indicates diverse sampling of the unfolded ensemble.

We obtain the scaling exponent ν for the ensembles at several times, shown in the inset of Figure 10(b). The exponent converges exponentially to an asymptotic value at $t \to \infty$. Extrapolating from times $\leq$1 ns is sufficient to obtain this asymptotic value to high accuracy. From this procedure, a scaling exponent that is slightly larger than a self-avoiding random walk (SAW) is obtained for proTα, likely because of the high charge density of this protein. We generally see scaling exponents for the unfolded ensemble with values between those in Θ-solvents and those in good solvents (Table I). The IUPs in our study tended to have larger scaling exponents than foldable proteins (0.62 *vs.* 0.57 on average) – some of the IUPs such as proTα are highly charged; none fall into the class of collapsed globule IUPs.[19,68] Our observations on scaling exponents are generally consistent with recent experimental observations by Schuler and colleagues[63] from single-molecule spectroscopy experiments.

An estimate for the persistence length in the unfolded state may be obtained from the prefactor r_1 as $\ell_p = (2\nu + 1)(2\nu + 2)r_1^2/(2b)$, where $b = 3.8$ Å is the C_α–C_α distance;[69] a SAW distribution has been assumed in this estimate. From this estimate, the persistence lengths we measured varied from 2 to 4 Å, averaging around 3 Å; these values were somewhat smaller than those obtained experimentally from force spectroscopy studies: typically around 3.5-4.0 Å (see, e.g., Refs. 70 and 71).

C. Clustering analysis

We undertook a clustering analysis for α-synuclein using the program `maxcluster` with nearest neighbor clustering.[72] The lack of strong clustering would be evidence of the success of our method to generate a diverse unfolded ensemble. Figure 11 shows the distribution of 1299 configurations of α-synuclein, projected along the TM-scores to the centroid structures of the three largest clusters. The elements of the clusters are indeed not well differentiated from the other structures, and the elements of the top three clusters are themselves fairly unrelated to the cluster centroid, with low TM-scores (cluster centroids have TM-scores of unity along their respective axis and are not shown – the centroids barely pull structures from the bulk ensemble).

We found that, for both the dominant clusters and the unfolded ensemble, the ends of the 140 aa protein tend to be closer on average than 140 aa stretches of other proteins that we had investigated (end to end distances $r_{ee}(\alpha\text{syn}) \approx 89$ Å, whereas $r_{ee}(1A6N_{140}) \approx 96$ Å), consistent with previous experimental NMR data that indicated aggregation inhibiting, long-range tertiary interactions between the N- and C-termini.[73] It is a valid question as to whether longer simulation times would result in enhancement of long-range tertiary interactions. Our configurations for α-syn were equilibrated for 5 ns; taking the equilibrated radius of gyration R_G of α-syn for our ensemble and treating the polymer as a self-avoiding chain in a good solvent, the longest Rouse-Zimm-like relaxation times are[74] $\tau_r \approx \frac{1}{3}\frac{\eta_s}{k_B T} R_G^3 \approx 10$ ns. This number is significantly longer than the relaxation time we observed for the radius of gyration for α-syn: $\approx$0.8 ns, most likely because the slowest Rouse-Zimm modes have already come to equilibrium by construction of the unfolded ensemble. On the other hand, single molecule Förster resonance energy transfer (FRET) measurements in the denatured state of cold shock protein give longer global reconfiguration times, $\approx$50 ns.[75] As well, specific, tertiary contacts are observed to have significantly slower formation rates,[76,77] e.g., for the naphthalene-xanthone labelled 56 aa peptide poly(GS)$_{28}$, the

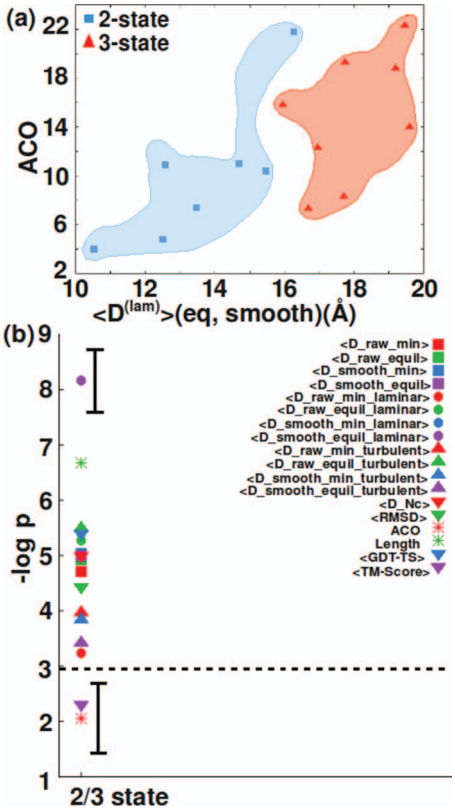

FIG. 12. (a) Scatter plot of the absolute contact order (ACO) and average laminar distance (equilibrium ensemble, with smoothed trajectories), for the 15 natively folded proteins in Table I. 2-state proteins (blue squares) and 3-state proteins (red triangles) are well-clustered by $\mathcal{D}^{(lam)}$, but not by ACO, as can be seen by inspection, i.e., by projecting data onto each order parameter. Closed curves circumscribing each class of protein are a guide to the eye. (b) Statistical significance (p-values) that the various metrics for 2-state and 3-state folders arise from different distributions, as determined by t-test.[30] $-\log(p)$ is plotted, so that a higher number indicates better ability to distinguish between the two classes. The dashed black horizontal line indicates a threshold of 5% for statistical significance. Only ACO and `maxcluster`-determined TM-score fail to distinguish 2-state from 3-state folders. Error bars for ACO and $\mathcal{D}^{(lam)}$ are obtained by removing 1 data point at random from the dataset, recomputing $-\log(p)$, and then calculating the standard deviation for the resulting collection of values. Notation used in this panel is further described in Figure 14.

time constant for the formation of specific end-to-end contacts is ≈170 ns.[76] Exploration of such contact dynamics in α-syn and other proteins is an interesting topic of future research.

We had found previously that a coarse-grained measure of the mean distance $\mathcal{D}$ in Eq. (5), as well as ⟨RMSD⟩ and chain length, significantly discriminates 2- and 3-state folders.[30] Here, we investigated which order parameters in Sec. II C cluster 2-state folders sufficiently separate from 3-state folders, such that they may be discriminated from each other. Figure 12(a) shows a scatter plot of the clustering along ACO and $\mathcal{D}^{(lam)}$ (with smoothed trajectories and for the equilibrium ensemble). Figure 12(b) plots the negative logarithm of the statistical significance, based on a t-test,[30] that each order parameter distinguishes 2-state from 3-state proteins (see Table S3 for listed p-values). More significant distinguishers have larger values on this plot.

D. Correlations between geometrical folding pathways and folding kinetics

Having generated statistically diverse, quasi-equilibrium, unfolded ensembles, we turned to the question as to whether transformations between such an ensemble and the native state could address folding kinetics. We investigated optimal folding trajectories both by our previous non-crossing method,[30] and by the GP method of Thorpe and colleagues;[31] these are described in Sec. II C. We focused primarily on the GP method because it applied to all-atom systems, and because the transformation could be visualized at all intermediate stages.

Information on the folding mechanism is gained from determining which quantity correlates with rate for a given structural or kinetic class of protein. For example, the fact that absolute contact order (ACO)[27,78] or extensions such as long range order[79–82] or total contact distance[83] correlate well with rate for 2-state proteins indicates a dominance of the process of loop closure, through the formation of native contacts, as the rate limiting step in folding. The fact that these quantities do not strongly correlate with rates for 3-state proteins, and that chain length does,[84] indicates that other, perhaps more subtle mechanisms embodying topological complexity may play a role in determining folding barrier heights.

Since the early studies of contact order, many subsequent studies have investigated correlates with protein folding kinetics across protein classes. Extensions of contact order have been developed from polymer theory,[85] which support early polymer physics models predicting rate determining barriers scaling as $\sim N^{1/2}$.[86] Mean field theory had predicted folding barriers increasingly linearly with protein chain length.[87] Rates were found to correlate with thermodynamic properties, including native stability[88] and heterogeneity of contact formation probability and ϕ-values.[89–91] Combinations of contact order and length have been taken.[92] Rates were observed to correlate with contact clustering[93] and to anticorrelate with the number of tightly packed contacts defined through Delaunay edges.[94] Folding rates for 2-state proteins were observed to correlate with helix, turn, and hairpin secondary structure propensity,[95] and inversely with chain length, in a model with 4 weighting coefficients. Consistent with this, both 2-state and 3-state folding rates anticorrelate with residual length after helical segments are renormalized to 3 residues in length,[96] implying significant secondary structure formation in the transition state. Folding times for two-state proteins were also observed to correlate with chain length within structural classes, but not across them;[80] the difficulty for ACO as a predictor across structural classes is also born out in statistical physics models.[97] Other studies have found that rates of 3-state proteins significantly *anticorrelated* with α and β secondary structure length, and that 2-state protein rates anticorrelate with β+loop secondary structure length.[98] Rates of both 2- and 3-state proteins can now be predicted to very high-accuracy from sequence alone,[94,99–101] but such approaches generally involve a large number of parameters – up to 20 for each amino acid type, or up to 49 amino acid properties – so that there is not much prospect of uncovering an underlying mechanism.

FIG. 13. Optimal folding trajectory of $C_\alpha(50)$ in apo-myoglobin (1A6N). The trajectory is curved, due to steric constraints with the remainder of the protein. $C_\alpha(50)$ is shown as blue spheres in the initial and final states. The region of protein N-terminal to $C_\alpha(50)$ in the initial unfolded state is shown in red. This transforms to the short helix N-terminal to $C_\alpha(50)$ in the final position.

Here, we asked the simple, physically based question as to whether the distance covered by optimal trajectories, generated as described in Sec. II C, would anticorrelate with folding rate across proteins (both 2- and multi-state). We also asked if simpler alignment measures such as RMSD, TM-SCORE, and GDT-TS would anticorrelate (for RMSD) or correlate (for TM-SCORE and GDT-TS) with rates, when averaged over the unfolded ensembles generated as described in Sec. II A.

In finding quantities related to distance $\mathcal{D}$ travelled, we found that the GP trajectories often meandered significantly from straightline motion, indicating that even for minimal transformations, steric hindrance induces curvilinear motion for folding trajectories (Fig. 13). We record the laminar and turbulent components of the distance between all unfolded conformations and the native structure, as described in Sec. II C 1.

Figure 14 shows a matrix of the Pearson correlation coefficients (upper triangle) and statistical significance (lower triangle), between the order parameters investigated here, including the 12 variants of distance described in Figure 7, ⟨RMSD⟩, ⟨TM-SCORE⟩ obtained from the program `maxcluster`, ⟨GDT-TS⟩, ACO, and protein length. Also included in the table are the log experimental folding and unfolding rates k_f and k_u, along with the log of the midpoint relaxation rate k_{mp}. A corresponding table with Kendall values is given in the supplementary material.[47]

Smoothing the trajectories did not significantly change an order parameter's correlation with rates. On the other hand, in nearly all cases, equilibration *decreased* an order parameter's correlation with rates. We suspect that this may be an indication of either the fast-mixing experimental protocol often used to dilute an initially high-denaturant state, or of problems with the force fields for the SPC solvent model, in which case more refined solvent models such as TIP4P may yield improved results. On the other hand, it is fortunate that the minimized, non-equilibrated ensemble performs so well – the computationally expensive procedure of equilibration may then not be necessary.

Alignment metrics to the native structure, such as the ensemble-averaged ⟨RMSD⟩, ⟨TM-SCORE⟩, and ⟨GDT-TS⟩, all showed significant correlation with folding rates across 2- and 3-state proteins. It is perhaps surprising that a quantity as simple as RMSD has not been tested in folding rate prediction, but this may be because information on pairs of structures rather than a single (native) structure is needed to calculate it.

The mean distances obtained by our previous polymer noncrossing method for coarse-grained C_α-model polymers[30] correlate very strongly with those obtained from the GP method ($r = 0.99$, $p = 6 \times 10^{-12}$). Moreover, these mean distances $\langle\mathcal{D}\rangle$ correlate remarkably strongly with the mean RMSD of the unfolded ensemble $\langle RMSD\rangle$ ($r = 0.94$ on average), so that this latter quantity may be used as a crude proxy for either distance calculation. This is a fortunate result in the sense that the calculation of $\langle RMSD\rangle$ is less computationally intensive than distance calculations by either method. It also means that the qualitative result is captured by simple polymer models. We will see however that significant quantitative effects are observed that depend on the all-atom steric volume of the protein, and its role in obstructing or guiding folding.

All variants of the distance travelled – minimized/equilibrated and smoothed/raw – showed significant correlation with folding rates. However, the laminar component of the distance travelled does not correlate strongly with rates, and is insignificant for rates in water. We have checked the convergence of the correlation between various distance metrics and folding rates; convergence is achieved fairly quickly, after about 100 unfolded configurations (Fig. S2 in the supplementary material[47]).

Figure 15 gives a synopsis of the correlation between various quantities and folding rates, in terms of minus log base 10 of their statistical significance (e.g., a significance of 10^{-4} would give a value of 4 on the plot). The most striking feature is the degree of correlation shown by the turbulent component of the distance travelled with folding rates in water ($r = -0.95$, $p = 1 \times 10^{-7}$). A scatter plot of folding rate vs. $\mathcal{D}^{(turb)}$ is shown in Figure 16(a). The ensemble-averaged RMSD also correlates significantly with folding rate – comparable in general to ACO (Fig. 16(b)).

The turbulent motion involves nonlinear docking and registering motions between at least partially formed secondary structures. It appears as "late-stage" reconfiguration of structured elements, and may be thought of as measuring the difficulty in fitting secondary structured units together. The observation that these motions appear to govern the barrier that determines rates implies a transition state ensemble with significant native structure present.

Of the quantities we investigated, unfolding rates at 0 M denaturant anticorrelate strongest with ACO ($r = -0.84$, $p = 1 \times 10^{-4}$); the entropy of loop closure governs the unfolding barrier as it does the folding barrier, and implies a mechanism for kinetic stability of native structures:

Pearson Corr. Coeff. p−value	<D_raw _min>	<D_raw _equil>	<D_smooth _min>	<D_smooth _equil>	< D_raw _min _laminar>	<D_raw _equil _laminar>	<D_smooth _min _laminar>	<D_smooth _equil _laminar>	< D_raw _min _turbulent>	<D_raw _equil _turbulent>	<D_smooth _min _turbulent>	<D_smooth _equil _turbulent>	< RMSD_min >	< RMSD_equil >	ACO	Length	<GDT_TS>	<TM_score>	ln_kf	ln_ku	ln_kmp
<D_raw _min>	1.	0.99	1.	0.99	0.75	0.71	0.87	0.9	0.96	0.91	0.97	0.97	0.97	0.96	0.86	0.9	−0.92	−0.84	−0.9	−0.81	−0.84
<D_raw _equil>	11.	1.	0.99	1.	0.76	0.7	0.87	0.91	0.94	0.93	0.95	0.98	0.98	0.97	0.83	0.93	−0.93	−0.89	−0.86	−0.8	−0.83
<D_smooth _min>	17.	11.	1.	0.99	0.76	0.71	0.88	0.92	0.95	0.91	0.96	0.97	0.97	0.96	0.85	0.92	−0.92	−0.84	−0.89	−0.79	−0.82
<D_smooth _equil>	11.	17.	11.	1.	0.78	0.72	0.89	0.92	0.92	0.92	0.94	0.98	0.98	0.98	0.82	0.94	−0.93	−0.89	−0.85	−0.78	−0.8
< D_raw _min _laminar>	2.8	3.	3.	3.2	1.	0.85	0.95	0.76	0.51	0.55	0.57	0.74	0.77	0.81	0.65	0.78	−0.83	−0.74	−0.47	−0.54	−0.55
<D_raw _equil _laminar>	2.5	2.4	2.5	2.6	4.3	1.	0.78	0.56	0.54	0.4	0.6	0.75	0.71	0.77	0.62	0.74	−0.7	−0.7	−0.48	−0.49	−0.48
<D_smooth _min _laminar>	4.5	4.6	4.8	5.	7.	3.2	1.	0.91	0.69	0.73	0.72	0.82	0.89	0.9	0.74	0.87	−0.9	−0.78	−0.65	−0.65	−0.68
<D_smooth _equil _laminar>	5.4	5.7	5.9	6.	3.	1.5	5.6	1.	0.82	0.89	0.83	0.83	0.91	0.88	0.69	0.91	−0.91	−0.81	−0.76	−0.65	−0.72
< D_raw _min _turbulent>	7.5	6.5	7.1	6.	1.3	1.4	2.4	3.8	1.	0.93	0.99	0.92	0.91	0.87	0.82	0.82	−0.81	−0.75	−0.95	−0.8	−0.84
<D_raw _equil _turbulent>	5.7	6.5	5.7	6.1	1.5	0.85	2.7	5.1	6.4	1.	0.92	0.89	0.9	0.87	0.75	0.82	−0.84	−0.79	−0.86	−0.78	−0.82
<D_smooth _min _turbulent>	8.7	7.4	8.1	6.9	1.6	1.8	2.6	3.9	13.	6.	1.	0.94	0.92	0.9	0.83	0.85	−0.83	−0.79	−0.93	−0.79	−0.82
<D_smooth _equil _turbulent>	8.7	10.	8.2	9.8	2.8	2.9	3.8	3.8	5.9	4.9	6.9	1.	0.96	0.97	0.84	0.9	−0.89	−0.88	−0.84	−0.8	−0.8
< RMSD_min >	8.6	9.6	9.1	10.	3.1	2.5	5.	5.7	5.5	5.4	6.	7.8	1.	0.99	0.85	0.94	−0.92	−0.88	−0.86	−0.8	−0.79
< RMSD_equil >	7.7	9.1	7.9	9.9	3.7	3.1	5.3	4.8	4.6	4.5	5.2	9.	11.	1.	0.82	0.92	−0.92	−0.89	−0.81	−0.79	−0.77
ACO	4.5	3.8	4.3	3.8	2.1	1.8	2.8	2.4	3.7	2.9	3.9	4.	4.2	3.8	1.	0.67	−0.77	−0.6	−0.87	−0.84	−0.8
Length	5.4	6.2	5.8	6.6	3.2	2.8	4.6	5.5	3.7	3.7	4.2	5.2	6.6	6.	2.2	1.	−0.87	−0.96	−0.74	−0.64	−0.66
<GDT_TS>	5.8	6.2	5.9	6.4	3.9	2.4	5.4	5.5	3.6	4.1	4.	5.	6.	5.9	3.1	4.5	1.	0.84	0.71	0.7	0.77
<TM_score>	4.	5.	4.1	5.1	2.7	2.5	3.2	3.5	2.9	3.3	3.4	4.8	4.9	4.9	1.8	7.6	4.	1.	0.65	0.63	0.62
ln_kf	5.3	4.4	5.	4.1	1.1	1.1	2.	3.	7.	4.5	6.5	4.1	4.4	3.6	4.5	2.7	2.5	2.1	1.	0.86	0.85
ln_ku	3.5	3.5	3.4	3.3	1.4	1.2	2.	2.1	3.4	3.2	3.4	3.5	3.5	3.3	4.	2.	2.5	1.9	4.4	1.	0.92
ln_kmp	4.1	3.8	3.8	3.5	1.5	1.2	2.3	2.6	4.	3.7	3.7	3.4	3.3	3.1	3.5	2.1	3.1	1.9	4.2	6.	1.

FIG. 14. Correlation matrix for all geometrical parameters, as well as experimental folding rates. The upper triangular elements are Pearson correlation coefficients. The lower triangular elements are the corresponding statistical significance values, which are represented as $-\log_{10}$ so that, e.g., 4.5 corresponds to $p = 10^{-4.5} = 3.2\times\ 10^{-5}$. Red represents strong positive correlation; blue represents strong negative correlation. "_raw" indicates numbers taken from the raw trajectory, while "_smooth" indicates numbers taken from the smoothed trajectory. Trajectories are further divided into "_laminar" and "_turbulent" parts. Initial ensembles are either equilibrated "_equil," or pre-equilibration (energy minimized only or "_min"). Other parameters shown include ACO, protein length, GDT-TS, TM-score, natural log of the folding and unfolding rates in 0 M denaturant, and natural log of relaxation rate at the transition midpoint.

long-range contacts promote more cooperative unfolding barriers. The distance-based metrics are still significant: the correlation coefficient of $\mathcal{D}$ with folding rate is -0.81.

The dominance of turbulent distance as a rate-predictor goes away at the transition midpoint (Fig. 15(c)). The total distance travelled along with the turbulent distance are still the strongest predictors however. Laminar components of the motion, while themselves weak predictors, become more significant in these conditions than they are in water.

This effect is consistent with expansion of the unfolded state as solvent conditions are varied by adding denaturant.[75] In the absence of denaturant, the unfolded state may generally be a significantly collapsed molten-globule like state.[26,102] Early kinetic folding intermediates of apoMb are nearly as collapsed as the native structure.[103] A pre-collapsed, compact unfolded state, driven primarily by hydrophobic interactions, is seen as well in two-state folding reactions for SH3,[104] thermostable variants of cold shock protein,[105] destabilizing

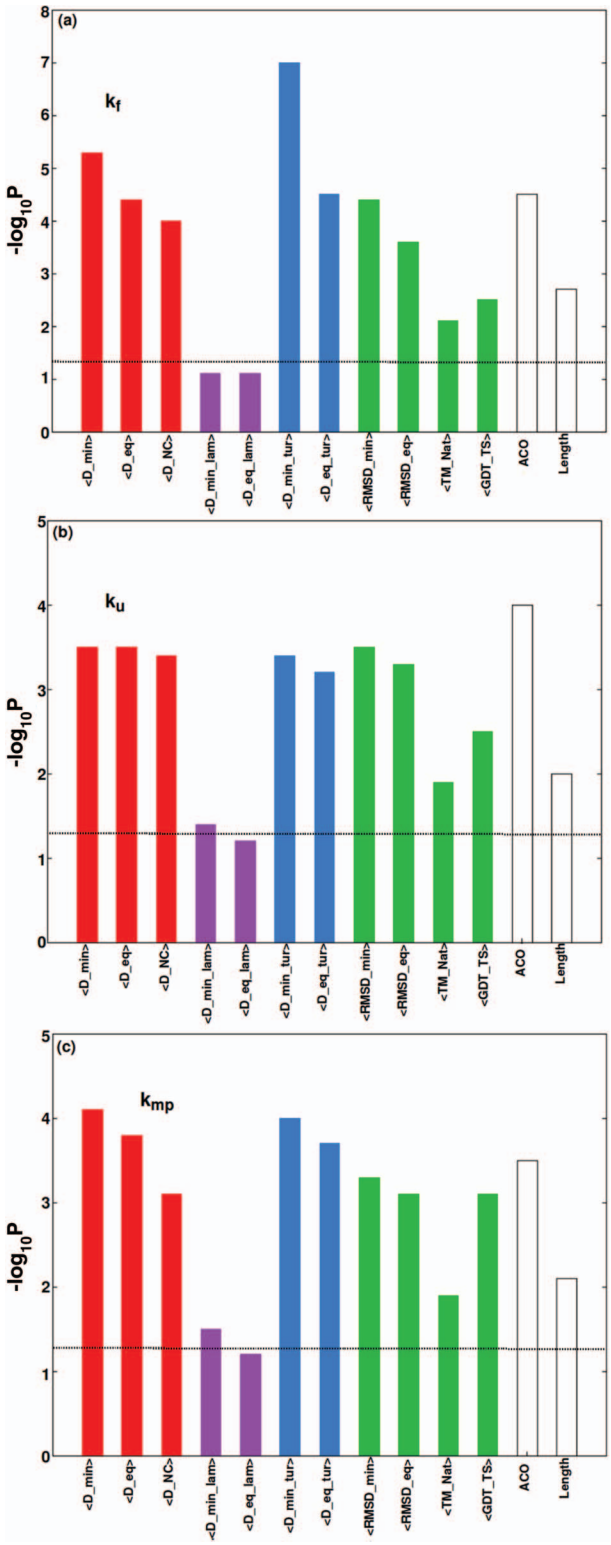

FIG. 15. (Panel (a)) Correlation of various distance metrics with experimental refolding rate in water, for the dataset of proteins listed in Table I. Raw (rather than smoothed) data are taken here. Minus the log base 10 of the statistical significance is plotted, and the horizontal dashed line gives the threshold of statistical significance ($p = 0.05$). The best predictor of folding rates in water, the turbulent distance, has a significance of 10^{-7}. Each integer below this value in the plot corresponds to a decrease in significance by an order of magnitude. (Panel (b)) Same as panel (a) but for experimental unfolding rate in water. Here, ACO emerges as the strongest correlator of unfolding rate. (Panel (c)) Same as panel (a) but for relaxation rate at the transition midpoint. Here, several variants of the distance travelled correlate best with relaxation rate, e.g., both $\mathcal{D}$ and $\mathcal{D}^{(turb)}$ have a correlation coefficient $r = -0.84$.

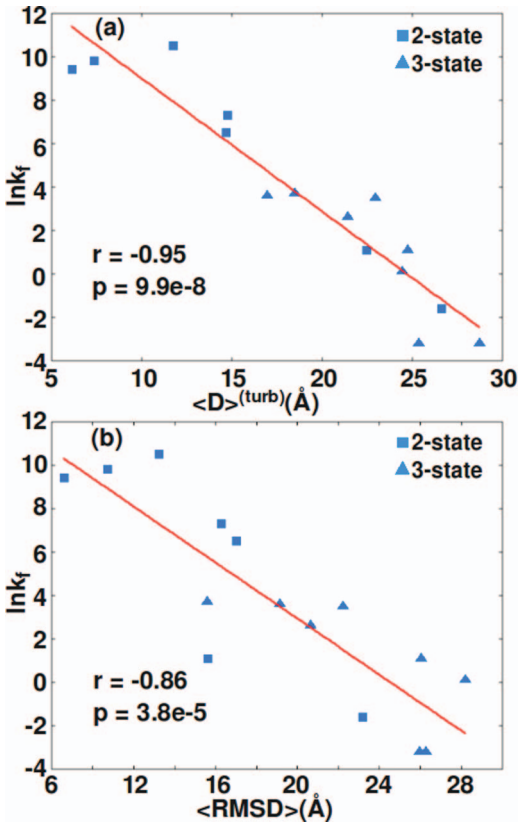

FIG. 16. (Panel (a)) Scatter plot of experimental folding rate at 0 M denaturant with the unfolded ensemble-averaged turbulent distance travelled during folding, corresponding to late-stage protein reconfiguration of structured elements. (Panel (b)) Scatter plot of the folding rate at 0 M denaturant with the ensemble-averaged RMSD between unfolded structures and the native. For both plots, the pre-equilibrated, energy-minimized, ensemble is taken, and raw rather than smoothed data are taken. Data for 2-state proteins are shown as squares, data for 3-state proteins are shown as triangles.

mutants of WT NTL9,[106] and Trp-cage miniprotein.[107] Fast initial collapse on the time-scale of tens of nanoseconds is seen directly in FRET measurements of BBL[108] and simulation of small proteins such as Villin.[42,109] Collapse emerges naturally in statistical field theories of heteropolymer collapse[86,110–113] as well as coarse grained computer simulations of folding.[114–116] Further along in the folding process, the transition state of CI2 has been interpreted as a globally collapsed, condensed nucleus with significant native structure, only slightly expanded, and lacking specific native packing interactions.[117] These experimental observations support the changing nature of the unfolded state with denaturant, which is reflected in which component of the minimal distance serves as the best predictor of folding rates.

IV. CONCLUSIONS

Here, we have connected the problem of generating unfolded ensembles with refolding kinetics, by applying transformations between unfolded conformations and the native structure. We developed a method to generate a diverse, quasi-equilibrium unfolded ensemble by employing coarse-grained sampling, foliation of the coarse-grained structure with side chains and backbone, and short equilibration of each

configuration. Ensembles for proteins with disulfide bonds can be generated as well, by employing non-local crankshaft-like moves.

Chemical shifts showed general agreement with experimental values, while residual dipolar couplings did not correlate with experimental values. We proposed some possible reasons for this discrepancy, including the fact that steric liquid crystal media that would induce rotational anisotropy in experiments would themselves modify the distribution of disordered conformations – a phenomenon that is not a factor for folded proteins.

Distance metrics as applied here between unfolded conformations and the native structure can also be applied to the ensemble of unfolded conformations, to obtain a general measure of the connectivity of the unfolded state. Distance metrics correlated strongly with common metrics of structural similarity, e.g., RMSD. The average RMSD between the folded structure and the unfolded ensemble correlated as strongly with folding rate as absolute contact order did.

The turbulent distance characterizes motion towards the folded structure that involves steric avoidance, jostling, rearrangement, and ultimately docking of highly structured units. The amount of motion this involved correlated most significantly with folding time ($r = 0.95$, $p = 1 \times 10^{-7}$). The dominance of "turbulent" minimal trajectories in predicting folding rates in water, but not at the transition midpoint, is a manifestation of a largely collapsed unfolded state in water or similar conditions, and the importance of reconfigurational motions from this state in finding transition states conducive to rapid progress toward the native structure.

ACKNOWLEDGMENTS

S.S.P acknowledges funding support from PrioNet Canada, NSERC, and computational support from the WestGrid high-performance computing consortium.

[1] J. Deisenhofer, O. Epp, K. Miki, R. Huber, and H. Michel, Nature (London) **318**, 618 (1985).
[2] D. A. Doyle, J. M. Cabral, R. A. Pfuetzner, A. Kuo, J. M. Gulbis, S. L. Cohen, B. T. Chait, and R. MacKinnon, Science **280**, 69 (1998).
[3] P. Agre, L. S. King, M. Yasui, W. B. Guggino, O. P. Ottersen, Y. Fujiyoshi, A. Engel, and S. Nielsen, J. Physiol. **542**, 3 (2002).
[4] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, Science **289**, 905 (2000).
[5] A. P. Carter, W. M. Clemons, D. E. Brodersen, R. J. Morgan-Warren, B. T. Wimberly, and V. Ramakrishnan, Nature (London) **407**, 340 (2000).
[6] F. Schluenzen, A. Tocilj, R. Zarivach, J. Harms, M. Gluehmann, D. Janell, A. Bashan, H. Bartels, I. Agmon, F. Franceschi, and A. Yonath, Cell **102**, 615 (2000).
[7] A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell, and R. D. Kornberg, Science **292**, 1876 (2001).
[8] J. A. Pitcher, N. J. Freedman, and R. J. Lefkowitz, Annu. Rev. Biochem. **67**, 653 (1998).
[9] V. Cherezov, D. M. Rosenbaum, M. A. Hanson, S. G. Rasmussen, F. S. Thian, T. S. Kobilka, H.-J. Choi, P. Kuhn, W. I. Weis, B. K. Kobilka, and R. C. Stevens, Science **318**, 1258 (2007).
[10] P. E. Wright and H. J. Dyson, J. Mol. Biol. **293**, 321 (1999).
[11] V. N. Uversky, J. R. Gillespie, and A. L. Fink, Proteins: Struct., Funct., Bioinf. **41**, 415 (2000).
[12] A. K. Dunker, J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and Z. Obradovic, J. Mol. Graphics Modell. **19**, 26 (2001).
[13] P. Tompa, Trends Biochem. Sci. **27**, 527 (2002).
[14] H. J. Dyson and P. E. Wright, Nat. Rev. Mol. Cell Biol. **6**, 197 (2005).
[15] A. K. Dunker, C. Oldfield, J. Meng, P. Romero, J. Yang, J. Chen, V. Vacic, Z. Obradovic, and V. Uversky, BMC Genomics **9**, S1 (2008).
[16] P. E. Wright and H. J. Dyson, Curr. Opin. Struct. Biol. **19**, 31 (2009).
[17] I. Baskakov and D. W. Bolen, J. Biol. Chem. **273**, 4831 (1998).
[18] P. Bernadó, C. W. Bertoncini, C. Griesinger, M. Zweckstetter, and M. Blackledge, J. Am. Chem. Soc. **127**, 17968 (2005).
[19] X. Wang, A. Vitalis, M. A. Wyczalkowski, and R. V. Pappu, Proteins: Struct., Funct., Bioinf. **63**, 297 (2006).
[20] N. G. Sgourakis, Y. Yan, S. A. McCallum, C. Wang, and A. E. Garcia, J. Mol. Biol. **368**, 1448 (2007).
[21] A. Vitalis, X. Wang, and R. V. Pappu, Biophys. J. **93**, 1923 (2007).
[22] O. O. Olubiyi and B. Strodel, J. Phys. Chem. B **116**, 3280 (2012).
[23] S. Æ. Jónsson, S. Mohanty, and A. Irbäck, Proteins: Struct., Funct., Bioinf. **80**, 2169 (2012).
[24] A. K. Jha, A. Colubri, K. F. Freed, and T. R. Sosnick, Proc. Natl. Acad. Sci. U.S.A. **102**, 13099 (2005).
[25] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge, Proc. Natl. Acad. Sci. U.S.A. **102**, 17002 (2005).
[26] J. B. Udgaonkar, Arch. Biochem. Biophys. **531**, 24 (2013).
[27] K. W. Plaxco, K. T. Simons, and D. Baker, J. Mol. Biol. **277**, 985 (1998).
[28] Y. Zhang, BMC Bioinf. **9**, 40 (2008).
[29] N. S. Bogatyreva, A. A. Osypov, and D. N. Ivankov, Nucleic Acids Res. **37**, D342 (2009).
[30] A. R. Mohazab and S. S. Plotkin, PloS ONE **8**, e53642 (2013).
[31] D. W. Farrell, K. Speranskiy, and M. F. Thorpe, Proteins: Struct., Funct., Bioinf. **78**, 2908 (2010).
[32] M. Lal, Mol. Phys. **17**, 57 (1969).
[33] N. Madras and A. D. Sokal, J. Stat. Phys. **50**, 109 (1988).
[34] S. Hadizadeh, A. Linhananta, and S. S. Plotkin, Macromolecules **44**, 6182 (2011).
[35] G. Favrin, A. Irback, and F. Sjunnesson, J. Chem. Phys. **114**, 8154 (2001).
[36] N. Gō and H. A. Scheraga, Macromolecules **3**, 178 (1970).
[37] L. R. Dodd, T. D. Boone, and D. N. Theodorou, Mol. Phys. **78**, 961 (1993).
[38] D. Hoffmann and E.-W. Knapp, Eur. Biophys. J. **24**, 387 (1996), see http://dx.doi.org/10.1007/BF00576711.
[39] M. R. Betancourt, J. Chem. Phys. **123**, 174905 (2005), see http://link.aip.org/link/?JCP/123/174905/1.
[40] P. Rotkiewicz and J. Skolnick, J. Comput. Chem. **29**, 1460 (2008).
[41] J. Maupetit, R. Gautier, and P. Tufféry, Nucleic Acids Res. **34**, W147 (2006).
[42] B. Zagrovic, C. D. Snow, S. Khaliq, M. R. Shirts, and V. S. Pande, J. Mol. Biol. **323**, 153 (2002).
[43] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 356 (1955).
[44] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
[45] K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, and M. Vendruscolo, J. Am. Chem. Soc. **131**, 13894 (2009).
[46] M. Zweckstetter, Nat. Protoc. **3**, 679 (2008).
[47] See supplementary material at http://dx.doi.org/10.1063/1.4817215 for supporting figures, tables, and description.
[48] RMSD values for backbone and all heavy atoms were also examined and gave similar results.
[49] W. Humphrey, A. Dalke, and K. Schulten, J. Mol. Graphics **14**, 33 (1996).
[50] Y. Zhang and J. Skolnick, Nucleic Acids Res. **33**, 2302 (2005).
[51] A. Zemla, Nucleic Acids Res. **31**, 3370 (2003).
[52] J. P. Rodrigues, M. Levitt, and G. Chopra, Nucleic Acids Res. **40**, W323 (2012).
[53] S. S. Plotkin, Proc. Natl. Acad. Sci. U.S.A. **104**, 14899 (2007).
[54] A. R. Mohazab and S. S. Plotkin, J. Phys. Condens. Matter **20**, 244133 (2008).
[55] A. R. Mohazab and S. S. Plotkin, Biophys. J. **95**, 5496 (2008).
[56] B. Thiruv, G. Quon, S. Saldanha, and B. Steipe, BMC Struct. Biol. **5**, 12 (2005).
[57] L. Hou, H. Shao, Y. Zhang, H. Li, N. K. Menon, E. B. Neuhaus, J. M. Brewer, I.-J. L. Byeon, D. G. Ray, M. P. Vitek, T. Iwashita, R. A. Makula, A. B. Przybyla, and M. G. Zagorski, J. Am. Chem. Soc. **126**, 1992 (2004).
[58] A. De Simone, A. Cavalli, S.-T. D. Hsu, W. Vranken, and M. Vendruscolo, J. Am. Chem. Soc. **131**, 16332 (2009).
[59] G. Lipari and A. Szabo, J. Am. Chem. Soc. **104**, 4546 (1982).
[60] N. G. Sgourakis, M. Merced-Serrano, C. Boutsidis, P. Drineas, Z. Du, C. Wang, and A. E. Garcia, J. Mol. Biol. **405**, 570 (2011).

[61] O. Tcherkasskaya and V. N. Uversky, Proteins: Struct., Funct., Bioinf. **44**, 244 (2001).
[62] S. Doniach, Chem. Rev. **101**, 1763 (2001).
[63] H. Hofmann, A. Soranno, A. Borgia, K. Gast, D. Nettels, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **109**, 16155 (2012).
[64] C. Tanford, K. Kawahara, and S. Lapanje, J. Biol. Chem. **241**, 1921 (1966).
[65] D. K. Wilkins, S. B. Grimshaw, V. Receveur, C. M. Dobson, J. A. Jones, and L. J. Smith, Biochemistry **38**, 16424 (1999).
[66] J. E. Kohn, I. S. Millett, J. Jacob, B. Zagrovic, T. M. Dillon, N. Cingel, R. S. Dothager, S. Seifert, P. Thiyagarajan, T. R. Sosnick, M. Z. Hasan, V. S. Pande, I. Ruczinski, S. Doniach, and K. W. Plaxco, Proc. Natl. Acad. Sci. U.S.A. **101**, 12491 (2004).
[67] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, Ithaca, 1953).
[68] V. N. Uversky, Protein Sci. **11**, 739 (2002).
[69] B. Hammouda, Adv. Polymer Sci. **106**, 87 (1993).
[70] R. B. Best, B. Li, A. Steward, V. Daggett, and J. Clarke, Biophys. J. **81**, 2344 (2001).
[71] D. Brockwell, E. Paci, R. Zinober, G. Beddard, P. Olmsted, D. Smith, R. Perham, and S. Radford, Nat. Struct. Biol. **10**, 731 (2003).
[72] D. Shortle, K. T. Simons, and D. Baker, Proc. Natl. Acad. Sci. U.S.A. **95**, 11158 (1998); see also http://www.sbg.bio.ic.ac.uk/maxcluster.
[73] C. W. Bertoncini, Y.-S. Jung, C. O. Fernandez, W. Hoyer, C. Griesinger, T. M. Jovin, and M. Zweckstetter, Proc. Natl. Acad. Sci. U.S.A. **102**, 1430 (2005).
[74] M. Doi and S. F. Edwards, *The Theory of Polymer Dynamics* (Oxford University Press, Oxford, 1986), Chap. 4.
[75] D. Nettels, I. V. Gopich, A. Hoffmann, and B. Schuler, Proc. Natl. Acad. Sci. U.S.A. **104**, 2655 (2007).
[76] F. Krieger, B. Fierz, O. Bieri, M. Drewello, and T. Kiefhaber, J. Mol. Biol. **332**, 265 (2003).
[77] J. Kubelka, J. Hofrichter, and W. A. Eaton, Curr. Opin. Struct. Biol. **14**, 76 (2004).
[78] K. W. Plaxco, K. T. Simons, I. Ruczinski, and D. Baker, Biochemistry **39**, 11177 (2000).
[79] M. M. Gromiha and S. Selvaraj, J. Mol. Biol. **310**, 27 (2001).
[80] A. Y. Istomin, D. J. Jacobs, and D. R. Livesay, Protein Sci. **16**, 2564 (2007).
[81] B. Harihar and S. Selvaraj, Biopolymers **91**, 928 (2009).
[82] T. Zou and S. B. Ozkan, Phys. Biol. **8**, 066011 (2011).
[83] H. Zhou and Y. Zhou, Biophys. J. **82**, 458 (2002).
[84] D. N. Ivankov, S. O. Garbuzynskiy, E. Alm, K. W. Plaxco, D. Baker, and A. V. Finkelstein, Protein Sci. **12**, 2057 (2003).
[85] M. Rustad and K. Ghosh, J. Chem. Phys. **137**, 205104 (2012).
[86] D. Thirumalai, J. Phys. I **5**, 1457 (1995).
[87] S. S. Plotkin, J. Wang, and P. G. Wolynes, J. Chem. Phys. **106**, 2932 (1997).
[88] A. R. Dinner and M. Karplus, Nat. Struct. Biol. **8**, 21 (2001).
[89] S. S. Plotkin and J. N. Onuchic, J. Chem. Phys. **116**, 5263 (2002).
[90] B. Oztop, M. R. Ejtehadi, and S. S. Plotkin, Phys. Rev. Lett. **93**, 208105 (2004).
[91] U. Bastolla, P. Bruscolini, and J. L. Velasco, Proteins: Struct., Funct., Bioinf. **80**, 2287 (2012).
[92] N. Koga and S. Takada, J. Mol. Biol. **313**, 171 (2001).
[93] C. Micheletti, Proteins: Struct., Funct., Bioinf. **51**, 74 (2003).
[94] Z. Ouyang and J. Liang, Protein Sci. **17**, 1256 (2008).
[95] H. Gong, D. G. Isom, R. Srinivasan, and G. D. Rose, J. Mol. Biol. **327**, 1149 (2003).
[96] D. Ivankov and A. Finkelstein, Proc. Natl. Acad. Sci. U.S.A. **101**, 8942 (2004).
[97] P. Bruscolini, A. Pelizzola, and M. Zamparo, Phys. Rev. Lett. **99**, 038103 (2007).
[98] J.-T. Huang, J.-P. Cheng, and H. Chen, Proteins: Struct., Funct., Bioinf. **67**, 12 (2007).
[99] M. M. Gromiha, J. Chem. Inf. Model. **45**, 494 (2005).
[100] H.-B. Shen, J.-N. Song, and K.-C. Chou, J. Biomed. Sci. Eng. **2**, 136 (2009).
[101] L. Chang, J. Wang, and W. Wang, Phys. Rev. E **82**, 051930 (2010).
[102] A. D. Miranker and C. M. Dobson, Curr. Opin. Struct. Biol. **6**, 31 (1996).
[103] D. Eliezer, P. A. Jennings, P. E. Wright, S. Doniach, K. O. Hodgson, and H. Tsuruta, Science **270**, 487 (1995).
[104] A. Dasgupta and J. B. Udgaonkar, J. Mol. Biol. **403**, 430 (2010).
[105] C. Magg and F. X. Schmid, J. Mol. Biol. **335**, 1309 (2004).
[106] B. Anil, Y. Li, J.-H. Cho, and D. P. Raleigh, Biochemistry **45**, 10110 (2006).
[107] K. H. Mok, L. T. Kuhn, M. Goez, I. J. Day, J. C. Lin, N. H. Andersen, and P. Hore, Nature (London) **447**, 106 (2007).
[108] M. Sadqi, L. J. Lapidus, and V. Muñoz, Proc. Natl. Acad. Sci. U.S.A. **100**, 12117 (2003).
[109] Y. Duan and P. A. Kollman, Science **282**, 740 (1998).
[110] J. D. Bryngelson and P. G. Wolynes, Biopolymers **30**, 177 (1990).
[111] S. S. Plotkin and J. N. Onuchic, Q. Rev. Biophys. **35**, 111 (2002).
[112] S. S. Plotkin and J. N. Onuchic, Q. Rev. Biophys. **35**, 205 (2002).
[113] G. Ziv, D. Thirumalai, and G. Haran, Phys. Chem. Chem. Phys. **11**, 83 (2009).
[114] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[115] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Biochemistry **34**, 3066 (1995).
[116] Z. Guo and D. Thirumalai, Biopolymers **36**, 83 (1995).
[117] A. R. Fersht, Curr. Opin. Struct. Biol. **5**, 79 (1995).
[118] M. Gromiha, A. Thangakani, and S. Selvaraj, Nucleic Acids Res. **34**, W70 (2006).

Unfolded protein ensembles, folding trajectories, and refolding rate prediction: Supplementary Material

A. Das, B. K. Sin, A. R. Mohazab, and S. S. Plotkin
Department of Physics and Astronomy, University of British Columbia, 6224 Agricultural Road, Vancouver, B.C. V6T 1Z1, Canada
steve@phas.ubc.ca

(Dated: 15 July 2013)

I. UNFOLDED ENSEMBLE FOR SUPEROXIDE DISMUTASE

A sample of 1000 5ns-equilibrated conformations from the unfolded ensemble of the disulfide-bonded protein superoxide dismutase (WT SOD1) is provided in two formats:

- A concatenated pdb file `sod1_all.pdb` (169 MB)
- An xtc trajectory file `sod1_all.xtc` (10.8 MB) along with a single pdb file `0001.pdb` (170K).

To view these files in `VMD`, issue the following commands from the terminal:

```
vmd -pdb sod1_all.pdb
```

or:

```
vmd -pdb 0001.pdb -xtc sod1_all.xtc
```

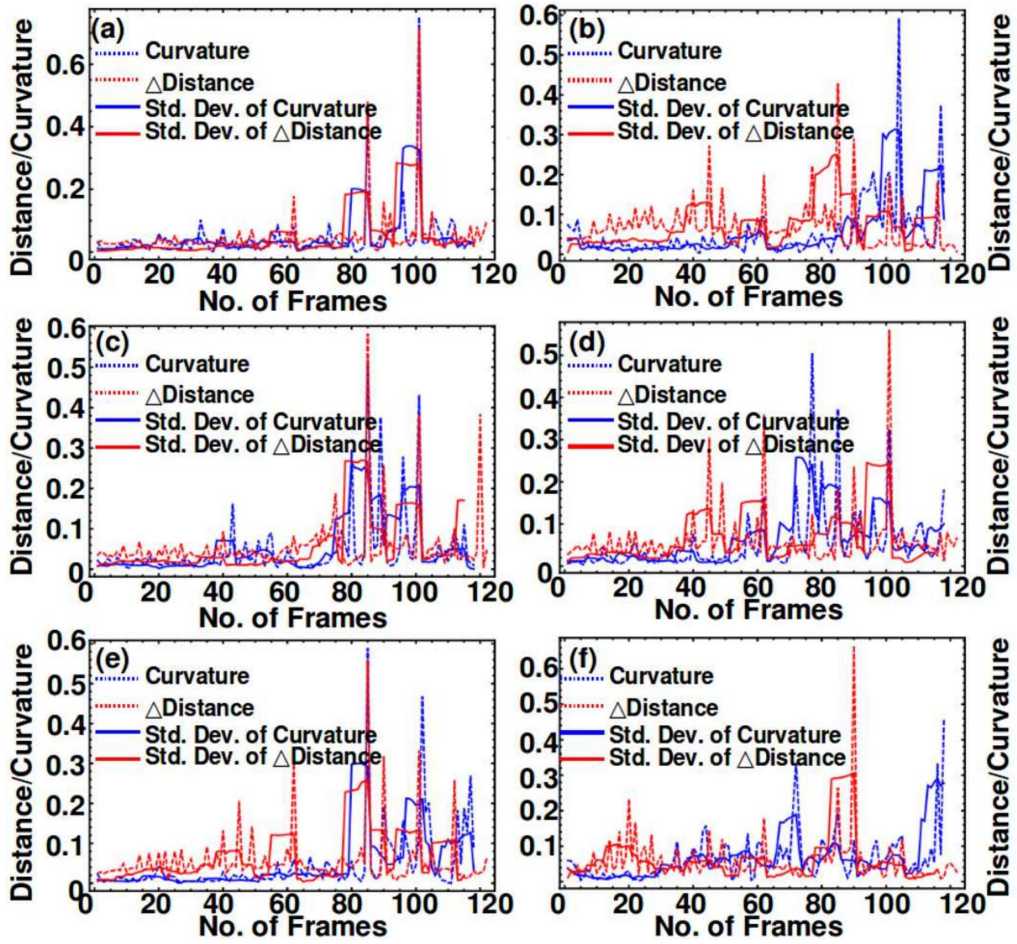

FIG. S1. **Transition from laminar to turbulent trajectories.** Each panel shows properties of the folding trajectory for selected C_α atoms. (Red dashed) Distance $\Delta\mathcal{D}$ traversed per frame, (Blue dashed) Curvature of the trajectory, (Red solid) running standard deviation of $\Delta\mathcal{D}$ over 10 frames, (Blue solid) running standard deviation of the curvature over 10 frames. Panels (a), (c), and (e) show good correlation between the two methods, while panels (b), (d), and (f) show some discrepancies between the two methods.

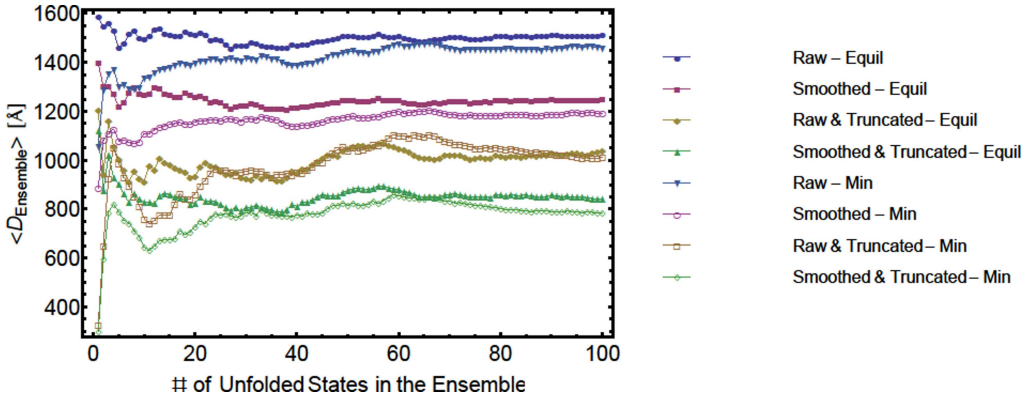

FIG. S2. **Convergence of mean distance travelled with ensemble size.** Convergence of various metrics for the total distance, as number of configurations in the unfolded ensemble is increased, for engrailed homeodomain 1ENH. Total distances converge after approximately 100 configurations.

Kendall Corr. Coeff. / p−value	<D_raw _min>	<D_raw _equil>	<D_smooth _min>	<D_smooth _equil>	< D_raw_min _laminar>	<D_raw _equil _laminar>	<D_smooth _min _laminar>	<D_smooth _equil _laminar>	< D_raw_min _turbulent>	<D_raw _equil _turbulent>	<D_smooth _min _turbulent>	<D_smooth _equil _turbulent>	< RMSD_min >	< RMSD_equil >	ACO	Length	<GDT_TS>	<TM_score>	ln_kf	ln_ku	ln_kmp
<D_raw _min>	1.	0.89	1.	0.9	0.6	0.79	0.71	0.77	0.85	0.71	0.87	0.81	0.87	0.83	0.73	0.79	−0.75	−0.73	−0.74	−0.56	−0.59
<D_raw _equil>	5.4	1.	0.89	0.98	0.6	0.75	0.71	0.81	0.77	0.79	0.79	0.85	0.9	0.9	0.62	0.79	−0.78	−0.77	−0.66	−0.6	−0.56
<D_smooth _min>	6.7	5.4	1.	0.9	0.6	0.79	0.71	0.77	0.85	0.71	0.87	0.81	0.87	0.83	0.73	0.79	−0.75	−0.73	−0.74	−0.56	−0.59
<D_smooth _equil>	5.6	6.5	5.6	1.	0.62	0.77	0.73	0.79	0.75	0.77	0.77	0.87	0.92	0.92	0.64	0.81	−0.8	−0.79	−0.64	−0.58	−0.54
< D_raw_min _laminar>	2.7	2.7	2.7	2.9	1.	0.77	0.89	0.68	0.45	0.47	0.47	0.56	0.58	0.66	0.52	0.62	−0.67	−0.56	−0.39	−0.39	−0.54
<D_raw _equil _laminar>	4.4	4.	4.4	4.2	4.2	1.	0.81	0.79	0.64	0.54	0.66	0.68	0.77	0.77	0.6	0.81	−0.82	−0.75	−0.53	−0.43	−0.5
<D_smooth _min _laminar>	3.7	3.7	3.7	3.9	5.4	4.6	1.	0.79	0.56	0.58	0.58	0.68	0.7	0.77	0.64	0.73	−0.78	−0.68	−0.51	−0.47	−0.57
<D_smooth _equil _laminar>	4.2	4.6	4.2	4.4	3.4	4.4	4.4	1.	0.73	0.71	0.71	0.66	0.75	0.75	0.54	0.79	−0.78	−0.73	−0.59	−0.45	−0.48
< D_raw_min _turbulent>	5.	4.2	5.	4.	1.7	3.	2.5	3.9	1.	0.75	0.94	0.73	0.71	0.68	0.7	0.64	−0.63	−0.58	−0.82	−0.52	−0.56
<D_raw _equil _turbulent>	3.7	4.4	3.7	4.2	1.8	2.3	2.6	3.7	4.	1.	0.73	0.87	0.73	0.77	0.56	0.66	−0.65	−0.64	−0.63	−0.62	−0.54
<D_smooth _min _turbulent>	5.2	4.4	5.2	4.2	1.8	3.2	2.6	3.7	6.	3.9	1.	0.75	0.77	0.7	0.75	0.66	−0.69	−0.6	−0.84	−0.54	−0.57
<D_smooth _equil _turbulent>	4.6	5.	4.6	5.2	2.5	3.4	3.4	3.2	3.9	5.2	4.	1.	0.83	0.9	0.7	0.71	−0.73	−0.7	−0.64	−0.64	−0.56
< RMSD_min >	5.2	5.6	5.2	5.8	2.6	4.2	3.5	4.	3.7	3.9	4.2	4.8	1.	0.92	0.71	0.85	−0.88	−0.83	−0.68	−0.58	−0.54
< RMSD_equil >	4.8	5.6	4.8	5.8	3.2	4.2	4.2	4.	3.4	4.2	3.5	5.6	5.8	1.	0.68	0.81	−0.82	−0.79	−0.63	−0.58	−0.54
ACO	3.9	2.9	3.9	3.	2.2	2.7	3.	2.3	3.5	2.5	4.	3.5	3.7	3.4	1.	0.6	−0.65	−0.54	−0.7	−0.6	−0.67
Length	4.4	4.4	4.4	4.6	2.9	4.6	3.9	4.4	3.	3.2	3.2	3.7	5.	4.6	2.7	1.	−0.84	−0.94	−0.57	−0.5	−0.5
<GDT_TS>	4.	4.3	4.	4.5	3.3	4.7	4.3	4.3	3.	3.1	3.4	3.8	5.3	4.7	3.1	4.9	1.	0.82	0.6	0.54	0.53
<TM_score>	3.9	4.2	3.9	4.4	2.5	4.	3.4	3.9	2.6	3.	2.7	3.5	4.8	4.4	2.3	6.	4.7	1.	0.55	0.52	0.44
ln_kf	3.9	3.2	3.9	3.1	1.4	2.2	2.1	2.6	4.6	2.9	4.8	3.1	3.4	2.9	3.5	2.5	2.7	2.3	1.	0.68	0.6
ln_ku	2.5	2.7	2.5	2.6	1.4	1.6	1.8	1.7	2.2	2.9	2.3	3.	2.6	2.6	2.7	2.1	2.3	2.2	3.4	1.	0.77
ln_kmp	2.7	2.4	2.7	2.3	2.3	2.	2.5	1.9	2.4	2.3	2.5	2.4	2.3	2.3	3.3	2.	2.2	1.6	2.7	4.1	1.

FIG. S3. **Correlation matrix for all geometrical parameters as well as folding rates.** The upper triangular elements are Kendall correlation coefficients. The lower triangular elements are the corresponding statistical significance values, which are represented as $-\log_{10}$ so that e.g. 4.5 corresponds to $p = 10^{-4.5} = 3.2\text{e-}5$. Red represents strong positive correlation; blue represents strong negative correlation. "_raw" indicates numbers taken from the raw trajectory, while "_smooth" indicates numbers taken from the smoothed trajectory. Trajectories are further divided into "_laminar" and "_turbulent" parts. Initial ensembles are either equilibrated "_equil", or pre-equilibration (energy minimized only or "_min"). Other parameters shown include ACO, protein length, GDT-TS, TM-score, natural log of the folding and unfolding rates in 0M denaturant, and natural log of relaxation rate at the transition midpoint.

TABLE S1. **Residual dipolar couplings for simulated unfolded ensembles**

Residue # / Protein	1L2Y	1VII	1IYT	2PDD	1ENH	1SHG	H-integrase	2CRO	1CSP	1PSF	1IMQ	1TIT	1APS	1BNI	ProT	1CBI	1XQ8	1A6N	2RN2	2A5E	1RA9
1	–	–	–	–	–	–	–	-1.5691	–	–	–	–	–	–	–	–	–	–	–	–	–
2	-1.20097	–	0.0695753	-0.994398	–	–	1.437	-0.540946	-0.824546	0.128958	1.21302	-1.879	0.524305	–	-4.25113	0.451055	-0.539563	-0.831028	3.55949	4.2314	7.85124
3	0.39806	–	1.14529	2.59899	–	–	1.134	-2.20707	1.07725	-0.556135	1.49642	3.91555	1.87431	–	2.81987	4.03064	0.426875	1.80827	-4.38336	–	4.60669
4	1.4542	–	1.36295	-0.362355	–	–	1.327	-0.211473	2.1211	-2.03885	-0.262484	2.76337	4.58189	-1.50272	1.09806	1.32216	1.43814	0.344984	0.222699	5.46418	-1.56911
5	2.65776	–	1.31978	–	0.982527	–	1.207	0.707974	0.597397	-1.68866	-2.08904	1.63847	–	-0.819469	-0.00887087	-3.02784	-0.851428	-0.0723738	2.73832	2.37276	-8.44104
6	4.57027	–	0.787021	-2.05346	1.30724	–	1.084	-2.40333	-1.30802	4.83925	3.04432	-3.52421	3.85805	-0.92213	-2.48489	2.76671	-0.416456	-1.95685	5.69503	1.17812	0.810924
7	2.30181	–	0.244609	-2.41515	2.18495	1.1237	0.915	0.0201088	-0.00946972	-0.122168	2.33934	–	2.3104	0.425677	-1.06234	-2.36406	-0.539593	-5.03617	5.17337	1.04303	-1.25106
8	1.68903	–	0.785388	-0.277751	0.946801	1.70021	1.022	-0.455145	0.193127	-1.06649	2.46577	0.399551	-2.80006	3.07539	0.10946	2.08001	0.524459	1.58953	-2.8033	-3.75138	-1.67887
9	1.05568	–	0.560431	1.06679	0.250395	-1.78392	-0.387	0.0979159	-0.693773	-1.09392	-0.597314	-1.31998	-4.6491	-4.5869	0.315381	-1.64945	0.310447	1.99627	-4.05379	6.09122	0.263696
10	2.83272	–	0.985069	-1.40685	0.482511	2.07912	0.654	1.48821	0.62837	2.07295	-2.3136	0.400664	-1.51489	-0.79828	-2.73552	6.56753	-0.939435	0.0236107	0.996013	2.64939	-0.237741
11	2.7626	–	0.266871	-0.489845	-0.493947	3.29125	0.177	0.937655	0.232886	0.68828	0.502791	-0.046752	-1.1434	-1.32657	2.26936	4.10214	0.342204	1.19834	-5.19705	–	4.30341
12	–	–	0.179666	0.728062	2.3332	0.0830201	0.525	0.19841	0.074307	0.692728	1.85041	-1.68675	-1.75345	2.25799	0.530482	-4.71094	0.594494	-3.65672	-2.08988	-1.91581	-4.05357
13	4.63063	–	1.59363	1.8002	-2.08479	2.17388	0.389	3.81979	2.47528	2.23115	0.833239	0.0671421	-2.27786	-1.28421	4.00106	-5.32779	2.08453	-4.87093	-8.58184	-0.357397	2.7709
14	3.11858	–	1.69752	0.0886982	-0.215748	3.41674	-0.742	-0.505325	2.77608	1.54966	-3.80354	-2.78362	1.04291	-0.149496	1.83578	-1.7766	0.962196	-5.65785	-1.82984	-3.63725	0.604975
15	2.82874	–	0.864542	1.39303	0.0177406	0.375117	-0.034	-1.73676	2.40655	1.37321	1.13482	-3.02328	0.894406	-3.42555	-0.416656	-0.789122	2.82479	4.71861	-0.938606	-2.97434	7.05702
16	3.58277	–	0.635864	1.76859	-0.162881	0.19981	0.060	-0.961419	1.45863	0.482047	1.13329	2.30351	-1.16927	0.621825	1.73969	0.625576	1.05823	-0.754129	-3.55883	2.61825	0.717463
17	–	–	-0.204336	2.02308	0.213467	1.26376	0.385	1.6594	1.54568	-0.828776	3.02684	-3.4253	-5.1509	-2.56145	-2.5699	-0.186402	0.6462	-1.61196	–	1.38589	2.84854
18	–	–	0.794609	2.50734	0.375353	-1.26872	0.193	0.263489	1.54677	-1.07266	0.483591	-2.31903	1.58753	-6.59986	-2.00012	5.50939	-1.04745	2.29565	-2.4396	-2.47432	7.15182
19	–	–	1.8997	1.3496	-1.76141	-0.848988	0.1	-0.0483082	-1.41394	-1.07056	-4.62187	-0.449866	0.285918	-2.97771	–	-1.18538	-0.13477	-1.65358	–	-0.500185	11.2895
20	1.08074	–	0.623642	1.5738	-0.336626	–	1.145	4.25784	-1.3063	-0.177914	-0.0484236	3.61254	3.16646	0.293165	0.864373	-8.5883	0.0646546	-3.88725	0.567284	-1.48469	-0.0466412
21	–	–	0.795773	0.148364	0.712803	-0.248159	1.226	3.14383	-2.02204	-2.17781	1.26035	3.22408	6.8616	–	0.107978	-4.52411	1.33923	-4.4918	-0.706427	2.60002	–
22	–	–	0.32331	-0.453531	-1.59693	0.141536	0.61	-2.97459	-0.123488	-1.62394	0.575919	-0.20238	4.65864	-0.20015	0.4242	-0.467028	1.05249	2.8925	1.38974	-5.26039	1.45802
23	–	–	-0.944872	1.4295	4.87986	-1.88826	0.934	-0.489017	2.74614	-1.2455	2.4165	0.0439072	8.24584	-1.73075	-2.42777	0.0121877	1.42112	-0.939347	5.52734	3.70798	7.1247
24	–	–	0.140028	1.30643	0.248159	-2.12731	0.24	2.5734	0.539378	2.46356	-0.399564	-0.0450783	-2.43738	4.17042	-0.998663	5.74571	-1.10195	-2.99081	-1.69806	3.3241	7.43897
25	–	–	1.19906	2.57571	-0.0914239	1.42815	0.239	1.31716	1.17015	1.34719	0.581645	1.00811	-2.8742	-1.63612	-0.0592197	3.15284	-0.250749	3.84614	-2.68	1.56236	–
26	–	–	1.3134	-0.43921	0.431398	0.552822	0.476	-0.246863	2.97663	2.34684	-0.812759	0.629649	-3.04623	3.08057	2.02542	3.22104	-0.932446	0.94447	2.26596	2.98922	6.29544
27	–	–	0.954295	0.0809602	1.41538	1.40301	0.549	-1.89882	2.66639	-0.218245	1.77479	-0.343591	-3.26203	1.50854	0.157703	-1.63017	1.84997	3.06934	3.29991	10.0568	10.124
28	–	–	0.891221	1.33503	-0.0448683	-0.190029	0.354	3.91325	1.06125	0.46602	1.54525	–	1.11396	0.991807	-0.271808	-3.00209	1.01121	3.58905	1.304	1.67051	2.03858
29	–	–	0.0043216	3.41444	0.0873672	-2.04725	1.871	4.42733	-1.08026	2.42592	4.95101	0.736707	0.789465	-2.27682	-0.985964	-0.258894	0.618966	10.6859	-0.544225	-1.19016	2.36422
30	–	–	0.453929	3.4855	0.904731	1.25962	0.914	2.13359	2.23385	2.4472	2.50622	2.30806	3.13687	0.736478	-1.40581	-1.71975	0.40239	0.402466	-1.59529	-6.44766	-4.46908
31	–	–	1.1492	1.52996	-0.174858	-5.463	0.186	0.753765	0.813651	4.63261	-1.45209	2.37027	4.43744	2.80701	1.41503	-3.85991	-0.609551	2.11534	-0.28637	-3.27209	2.40361
32	–	–	0.92556	1.031	1.39472	-0.0512971	0.697	0.563854	0.880923	-0.255535	2.71346	2.49155	7.28746	4.18982	-0.296597	-7.54424	-0.0452876	-1.32397	2.44849	-0.976273	0.656402
33	–	–	1.08833	0.589377	4.09435	0.807099	–	-0.890134	-0.463991	2.12941	-0.140817	1.26647	1.93514	3.9223	1.69925	1.30155	0.265073	-2.92774	-3.72994	3.16327	-7.06344
34	–	–	1.05034	0.882641	0.637987	0.278438	–	1.66608	0.890182	3.36702	-2.79394	0.809245	1.90294	2.50102	-1.91232	-4.60668	0.302759	-3.63134	-1.42384	-0.967988	4.05797
35	–	–	0.329927	1.91919	-0.174375	0.543768	1.797	5.28493	0.0938099	–	-0.915539	-1.8258	-0.978027	3.45195	0.059167	2.3087	2.68378	1.97369	3.26621	-5.01868	3.96706
36	–	–	0.601581	-1.05247	0.554351	0.446876	1.043	-0.370419	-3.90537	2.06382	-1.64414	3.52551	-1.52937	2.08076	2.24152	3.39575	1.16586	-1.80253	-0.307414	5.39485	6.31698
37	–	–	1.02857	1.34491	0.826614	2.65287	0.807	-0.493547	0.186489	3.09452	-3.49384	2.1097	-3.73577	6.28808	1.21567	3.18988	0.937943	–	-2.00674	2.97427	-6.19108
38	–	–	0.338806	1.84969	2.674	1.51639	1.181	0.677838	3.06278	-0.0808882	-5.72293	1.7982	2.3307	3.94689	-1.88137	-1.15548	-0.255013	-1.07172	-0.565351	–	8.98468
39	–	–	0.821209	0.720452	-0.793705	-1.36112	0.602	1.24296	-0.138597	2.3445	-0.294876	1.22354	0.407381	-1.73081	0.781585	–	-0.756527	1.43258	-0.97362	-2.15669	–
40	–	–	2.42732	1.51246	-0.698889	0.865549	0.561	2.57706	-2.36943	0.327144	2.95361	–	-2.93909	-2.73379	-1.60805	-0.0417923	0.0809978	1.40953	1.23057	-3.94396	-1.38113
41	–	–	0.0816983	0.444741	0.916891	3.6201	2.118	0.597907	0.752691	3.73207	1.65613	0.854433	2.60847	5.08444	0.971883	-1.72818	0.22414	1.48167	-3.09933	–	1.21439
42	–	1.14918	0.331248	-0.646099	0.744237	-0.924413	2.163	–	2.28422	1.2137	3.61892	0.912625	-1.31994	5.79841	-0.483561	2.85884	0.542907	1.52242	-1.71412	1.95756	3.18452
43	–	4.24258	–	0.85857	1.50244	2.36145	0.030	2.23122	-1.50812	3.85731	1.74574	3.61847	-1.69386	-3.25236	2.99082	1.42657	-1.58645	0.888637	3.2545	2.34816	0.398271
44	–	1.07446	–	–	0.439529	1.63807	0.215	2.14824	-1.09705	1.7476	1.98357	-0.238993	-0.207952	-0.554893	-0.993631	7.19845	-0.927773	-0.710809	0.347179	-3.88849	10.5227
45	–	-0.814377	–	–	-0.018985	1.3284	0.029	-0.120939	1.36887	0.899852	2.46483	–	2.58156	2.86715	4.5101	4.88271	-1.90189	4.86834	1.41638	-8.44703	3.66109
46	–	1.2838	–	–	-2.56636	-0.979336	0.089	-0.871379	-0.447792	0.469019	2.51136	-0.573601	1.63511	3.785	0.79781	-2.97579	0.511996	3.91488	2.83472	-0.143142	-2.71291
47	–	1.10832	–	–	-2.33549	-0.195017	1.704	1.27552	-0.132149	1.10811	–	-0.720912	0.727386	–	-0.679636	-0.671792	1.68546	-1.24438	1.01798	2.88025	1.92382
48	–	1.45109	–	–	1.29318	-3.23228	1.36	2.60691	-0.649458	1.71338	0.574139	-1.27693	-0.874635	0.427927	1.28238	5.29399	-1.05406	9.90688	-6.53271	–	2.3869
49	–	1.55391	–	–	-0.201167	1.27708	0.642	1.58204	-2.03677	2.7911	0.519411	-0.447081	-2.19228	0.250715	2.67211	1.23933	0.937784	9.56806	-1.17234	-1.56994	-9.07703
50	–	3.24052	–	–	-0.619423	-0.0762328	1.286	-2.05964	-3.00491	-0.711653	2.4972	0.265622	-0.855438	2.28173	-2.39817	2.62111	1.61208	9.27884	6.33661	-3.75247	1.5461
51	–	1.27557	–	–	-1.53258	-1.46888	1.084	-0.356227	-0.33415	-2.40459	-1.19406	-2.34154	-1.71209	3.9786	0.436231	7.08867	1.43169	3.07836	0.0684012	-4.71033	-3.34489
52	–	-1.19346	–	–	0.391958	0.784009	0.129	1.50148	-1.38661	0.98044	-0.603276	-2.74494	0.0112315	1.267	-1.92729	1.56882	-1.33658	1.42976	-3.12497	-4.44597	-4.16204
53	–	2.03219	–	–	1.3107	0.269559	0.123	0.805559	-1.47299	3.1874	-0.139077	-0.820277	-3.59365	1.11654	–	-2.07656	-1.39793	-2.22217	2.93528	0.325653	–
54	–	1.06272	–	–	2.537	–	0.863	0.187745	-0.755869	0.0574092	-1.34076	-1.03958	–	0.604684	0.371795	-1.19391	-0.0791192	-8.42281	3.28614	-1.44399	-3.95955
55	–	-0.166487	–	–	-0.0200994	0.299744	1.264	-1.19803	3.80156	1.9388	2.39451	1.69779	3.0034	-2.78302	-0.895858	-5.09434	1.4317	-6.75404	-0.966903	-1.97495	–
56	–	0.504661	–	–	4.39472	1.89473	0.53	–	1.26332	1.15507	–	0.0989164	2.40646	0.0706082	0.853434	0.718873	2.52787	-7.77141	4.94497	-0.288529	-3.44868
57	–	0.0645924	–	–	–	1.67714	0.545	0.955154	1.381	0.155918	2.5584	-0.704114	2.45745	-2.90428	2.70449	-1.99227	1.41318	4.07929	0.704025	-1.1351	-7.16558
58	–	2.67439	–	–	–	0.133489	0.791	2.07363	–	1.36877	3.98936	1.56214	-1.42086	-1.83826	1.59545	-0.449618	2.22333	10.3327	3.15267	-4.73997	-6.42761
59	–	1.9821	–	–	–	0.88076	1.751	1.25744	3.31147	1.76717	0.858924	1.17566	-0.925203	2.16345	2.72828	-4.70737	0.846851	-3.70324	0.342122	0.430928	-12.0782
60	–	0.768943	–	–	–	-0.777047	0.653	3.18415	1.8532	-0.0155209	-0.345867	0.855336	-2.55237	-0.304275	-0.664922	-4.01091	0.14505	2.73247	1.48922	1.42053	7.09109
61	–	0.629142	–	–	–	1.70014	–	1.59921	-2.146	-1.90752	-2.83745	-1.42422	0.00685923	2.11122	0.61166	0.456532	0.826333	2.88326	1.09892	4.18574	3.64198
62	–	–	–	–	–	1.90635	–	-5.60326	-0.403681	3.61182	-2.95727	0.5423	-2.7985	0.882075	-2.05896	-5.76272	1.35308	-1.4861	-0.816008	7.69178	1.00297
63	–	-0.477159	–	–	–	–	–	-0.352668	1.01065	3.19884	0.759902	0.144409	-2.07732	5.30568	-1.50784	-0.367627	-0.157847	-1.00533	-0.493597	2.62044	3.33303
64	–	-0.187756	–	–	–	–	–	–	3.53801	0.0793634	–	0.755879	1.85224	–	2.56989	3.30526	1.89729	1.14684	-3.83113	2.39114	2.74837
65	–	0.945166	–	–	–	–	–	–	2.30922	-4.12782	2.79189	-0.0687829	1.3658	-0.536665	1.19113	4.17505	1.12879	3.15792	0.79782	-1.83324	1.09639
66	–	-2.03803	–	–	–	–	–	–	2.97507	1.4164	-0.587218	-3.60925	0.905438	-2.87963	2.63999	-4.25762	1.35948	1.7322	2.67357	9.15581	–
67	–	0.362234	–	–	–	–	–	–	2.33392	0.913452	-2.01114	-1.74749	1.72291	-1.44589	4.86812	0.582691	-0.926723	7.96995	-1.15699	3.55663	-5.32409
68	–	1.22166	–	–	–	–	–	–	–	2.51625	-0.569326	0.258659	3.23585	-2.98719	3.79079	3.99842	1.27467	-0.764305	-6.48749	0.30992	-9.18699
69	–	3.34928	–	–	–	–	–	–	–	2.22948	0.966005	3.31209	4.00801	-0.734785	4.32929	0.601762	0.937033	-2.32249	-5.11209	-2.19053	1.38211
70	–	2.63969	–	–	–	–	–	–	–	–	0.524414	0.285134	0.107812	2.84082	-0.0712289	-1.0661	-0.763899	-1.54489	0.133452	–	8.28051
71	–	1.70165	–	–	–	–	–	–	–	–	0.0587791	1.97206	–	-0.179834	-0.0805811	-6.30321	0.529013	-3.55489	0.868047	-2.66795	-1.23359
72	–	1.92142	–	–	–	–	–	–	–	–	2.96579	-1.55829	2.0024	0.671071	-1.45092	2.9336	2.26985	-3.90768	-3.34241	-3.67746	-1.81505
73	–	2.92531	–	–	–	–	–	–	–	–	3.88512	-1.53116	-3.67306	-0.296636	1.27436	1.43733	0.952505	0.573464	-2.41071	2.78	-6.34869
74	–	1.3559	–	–	–	–	–	–	–	–	-0.130153	-0.241304	1.75976	-3.10539	3.77928	4.00759	0.279824	6.21825	-2.50639	3.42651	-5.34912
75	–	-1.32746	–	–	–	–	–	–	–	–	-0.781441	-0.60174	0.985948	3.11461	2.06242	3.57432	2.01127	2.73888	4.81896	–	4.07746
76	–	-0.755575	–	–	–	–	–	–	–	–	-0.220291	4.42091	-1.88924	-0.148835	0.455383	-0.359256	0.820443	3.0696	-4.73408	-4.02223	-5.76218
77	–	–	–	–	–	–	–	–	–	–	-4.1271	3.20853	4.14636	-1.80826	1.71604	4.75915	1.3457	2.33372	6.49896	-2.55737	0.514412
78	–	–	–	–	–	–	–	–	–	–	3.34733	2.49807	2.76122	-2.82065	2.84017	2.84025	0.685558	-3.89615	0.502939	-1.20004	6.92538
79	–	–	–	–	–	–	–	–	–	–	1.53943	2.01452	-1.30251	-4.17822	1.61538	6.65046	-0.37269	2.48305	3.662	-2.06142	-0.371324
80	–	–	–	–	–	–	–	–	–	–	0.215405	-0.109195	-2.78821	0.830852	2.55097	6.41251	1.5075	1.00362	-0.83437	2.00675	2.93916
81	–	–	–	–	–	–	–	–	–	–	1.86675	-2.6947	-1.94069	-2.2926	-0.545079	-2.29605	0.605682	-0.00763689	-2.35683	–	4.4819
82	–	–	–	–	–	–	–	–	–	–	-1.13112	-2.59765	2.88514	-0.777566	-0.427326	2.88874	0.725763	0.713311	-0.380237	4.06926	1.20625
83	–	–	–	–	–	–	–	–	–	–	1.51561	-1.72667	-1.04298	-2.25025	0.226548	5.16944	-0.370497	9.7321	0.0347259	1.57734	-2.28029
84	–	–	–	–	–	–	–	–	–	–	-4.89521	3.14725	0.477736	0.744667	1.91933	-1.95548	-0.510516	2.37357	3.86647	0.821329	-12.2256
85	–	–	–	–	–	–	–	–	–	–	0.193524	0.816124	0.775937	2.93104	-0.169384	–	0.977541	-2.65019	1.20545	0.761619	-4.53314
86	–	–	–	–	–	–	–	–	–	–	-0.392022	-0.449818	4.12041	5.21675	0.856596	2.35895	-0.264062	1.04932	0.75696	0.356702	1.5176
87	–	–	–	–	–	–	–	–	–	–	–	3.40926	-2.32089	0.254911	3.41655	-0.265115	1.34427	-3.71988	4.82749	1.94619	3.04599
88	–	–	–	–	–	–	–	–	–	–	–	1.47592	-4.42841	2.51111	0.54027	8.7353	0.813966	–	-4.16258	5.54342	-0.643031
89	–	–	–	–	–	–	–	–	–	–	–	-1.04931	-2.14413	2.13471	3.46612	-4.03127	1.31565	2.39463	-2.09331	2.28448	–
90	–	–	–	–	–	–	–	–	–	–	–	–	-5.69494	-0.211526	2.5273	0.0784508	-1.794	-1.76856	0.267472	1.23489	2.67956
91	–	–	–	–	–	–	–	–	–	–	–	–	-3.21129	-1.17735	2.8319	0.357263	0.250572	-4.04135	1.02154	3.84384	-2.1856
92	–	–	–	–	–	–	–	–	–	–	–	–	-4.83281	-1.05863	2.57824	0.0944185	0.779376	6.67095	1.86505	-3.6124	3.69259
93	–	–	–	–	–	–	–	–	–	–	–	–	2.92069	0.445034	1.04984	-0.853558	0.358519	1.98578	0.0510663	-4.21005	-1.23097
94	–	–	–	–	–	–	–	–	–	–	–	–	-0.647902	-2.0981	3.04816	10.4581	1.33949	0.830457	4.01724	0.54649	1.90939
95	–	–	–	–	–	–	–	–	–	–	–	–	-0.353085	1.06129	3.25511	2.29411	1.67673	-5.94956	-2.55078	3.2354	-7.78808
96	–	–	–	–	–	–	–	–	–	–	–	–	-0.599602	1.12022	-2.1328	2.43611	-1.06015	3.58602	1.06122	1.57363	0.881971
97	–	–	–	–	–	–	–	–	–	–	–	–	-2.75619	0.409947	-1.13697	1.31547	0.862399	7.77587	–	-3.05437	-6.99624
98	–	–	–	–	–	–	–	–	–	–	–	–	4.79751	-2.72549	-4.32188	0.511155	-1.51626	-4.85535	4.0243	-5.923	0.193941
99	–	–	–	–	–	–	–	–	–	–	–	–	–	-7.07189	-3.86863	2.05929	1.40762	0.355307	4.31917	-7.39905	-12.8391
100	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0555754	-3.24822	1.67706	0.141456	–	5.17711	-4.30855	-12.0786
101	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.90601	-1.25751	2.29743	-1.79643	-2.78348	0.0146584	-1.33614	-1.01625
102	–	–	–	–	–	–	–	–	–	–	–	–	–	0.371234	-0.774461	0.427564	0.506414	7.2653	5.62636	-0.276065	2.29992
103	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.555054	-0.12619	-3.27401	0.382208	-3.35731	4.08454	2.61325	3.14432
104	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.517573	-0.730305	4.26685	-1.07921	-9.05965	4.99493	0.235256	-1.18445
105	–	–	–	–	–	–	–	–	–	–	–	–	–	0.627948	-0.413686	–	-0.791881	-0.863927	1.3751	4.03739	–
106	–	–	–	–	–	–	–	–	–	–	–	–	–	1.05408	-2.65068	2.21129	2.46101	-7.97374	-3.52649	-5.35344	2.60717
107	–	–	–	–	–	–	–	–	–	–	–	–	–	1.28093	3.06743	2.2073	-0.505192	-2.30192	-7.75142	-6.77037	0.220618
108	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.91921	0.780559	0.377037	–	-2.38051	4.28983	-3.26066	-10.6079
109	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.234402	0.757403	3.29858	-0.365209	1.9616	6.30182	-1.80853	4.74927
110	–	–	–	–	–	–	–	–	–	–	–	–	–	0.908116	-0.87604	6.33646	-0.418395	1.48693	0.868581	-0.693395	7.41259
111	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.852653	-0.619191	-0.274826	-3.43234	-1.24238	3.32511	4.63608
112	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.46961	4.35422	2.28942	-7.08529	-3.19613	-2.27799	5.66256
113	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.00005	7.41925	1.08564	4.06982	-2.16803	-11.2903	-2.40529
114	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.68876	2.20299	-2.30957	4.01861	-3.27972	–	-2.98984
115	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.947068	-0.52992	1.03515	0.830778	0.301052	-0.925065	3.28179
116	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.05875	3.33463	2.2362	-0.504526	-3.73391	-3.43011	6.05667
117	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-3.6294	1.91669	–	1.27114	-3.14509	1.74916	4.94672
118	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.953793	2.4815	2.05304	5.9383	3.46202	2.38137	3.84337
119	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.53996	8.58428	1.2422	-1.45973	3.20521	-1.12574	-0.523586
120	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.475928	4.50173	–	–	1.47233	2.9424	0.995182
121	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.01096	-5.72566	1.14904	-3.47345	3.65595	9.53581	11.0252
122	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.2455	0.283151	-0.652687	1.41668	3.3399	0.417027	-4.15693
123	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.187213	-1.73158	1.02233	1.52675	1.69555	0.594501	-0.404043
124	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.579784	0.326689	2.0075	-4.08127	3.13487	-0.928096	-0.220318
125	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.89394	-2.07068	3.87203	2.2494	3.80732	3.66205	6.46925
126	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0371603	-0.252629	0.340996	1.09918	0.801374	1.90674	–
127	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.22801	-1.65714	1.10401	-4.09637	-2.70734	-6.31592	-5.74389
128	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.52297	3.08281	–	2.13281	–	1.81142	4.44377
129	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.09806	0.644942	0.675831	6.61387	-0.0783861	-2.28529	2.51395
130	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	5.85185	0.291658	-1.78234	4.91196	6.71768	–
131	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.35082	-2.77127	2.38396	1.11351	-3.36347	7.2155
132	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.69186	-1.26575	-0.629425	4.43336	-0.134846	-3.34341
133	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.14611	1.06636	-6.04096	2.30102	3.49593	3.05602
134	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.89186	1.78484	4.05952	1.8114	-0.454838	-1.46456
135	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.9305	0.823686	1.8519	0.909133	-5.42033	0.764839
136	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6.90614	0.0547263	-4.45281	4.86073	-0.914674	-0.932102
137	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.480198	-5.73364	-0.0909367	0.614712	0.212586
138	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.02656	1.89722	5.35102	7.35929
139	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.42554	3.64769	1.21594	3.21411	11.1421
140	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.23233	-2.3141	0.950164	-3.39997	-6.50521
141	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-6.86395	-2.51275	-0.732706	10.5325
142	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.78381	4.87705	2.70378	5.66074
143	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.05758	1.30195	5.55075	-1.68058
144	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.949193	–	2.74023	-1.90818
145	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.85879	1.06601	0.192716	4.30454
146	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-0.117921	2.01595	-3.35656	-1.54631
147	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.173873	1.04558	-1.29215	-2.057
148	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.26186	-2.06908	0.654297	-0.720169
149	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1.73327	-0.842005	3.01922	5.1999
150	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.5191	-1.66737	3.17488	-1.80882
151	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.28351	-1.89452	–	0.286043
152	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.0599771	2.58275	2.4903
153	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-1.9864	3.06349	1.09988
154	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.72287	-0.0754692	-0.968498
155	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	4.184	–	6.1223
156	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	-2.24693	10.2986
157	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	0.212724
158	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2.65695
159	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3.23437

TABLE S2. **Chemical shift values for simulated unfolded ensembles**

Residue # / Protein	1L2Y	1TTI	1IYT	2A5E	1ENH	1RA9	H-integrase	1XQ8	1CSP	1PSF	1IMQ	1SHG	1APS	1BNI	ProT	1CBI	1VII	1A6N	2PDD	2CRO	2RN2
1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	68.	–
2	55.2	60.6	52	54.8	–	62.	61.8	51.6	52.	59.8	55.3	–	61.4	–	49.6	58.2	–	54	58.3	56.	57.7
3	57.5	58.	56.5	63.8	–	61.8	55.5	63.	55.3	55.2	63.4	–	49.2	–	60.5	60.9	–	64.7	49.7	68.7	55.4
4	62.1	64.7	58.9	50.7	66.5	54.1	55.4	56.2	44.4	55.4	54.8	–	57.9	61.6	60.2	50.4	–	59	59.6	59.6	56.6
5	60.2	56.9	56.3	50.6	61.8	62.1	60.3	55.9	59.2	48.2	60.2	–	63.6	51.9	58.5	48.	–	48.7	66.6	52.1	66.1
6	66.9	58.3	53.3	42.3	61.8	50.4	55.1	57.	62.	64.2	57.4	–	55.2	65.5	55.7	62.8	–	55.4	59.8	59.5	56.4
7	56.2	68.3	58.5	66.	53.5	49.7	56.2	44.8	54.4	54.8	61.	58.8	59.4	58.2	54.8	62.7	–	59.7	66.1	51.6	65.7
8	56.2	57.8	65.3	59.5	59.8	53.5	44.9	58.7	61.9	67.3	67.6	59.5	68.5	53.1	59.3	52.	–	57.2	57.3	53.1	57.4
9	56.8	60.2	47	54.	70.2	50.5	65.3	67.6	55.1	56.7	51.3	67.6	64.5	44.5	61.8	61.7	–	56	55.6	63.6	62.3
10	43.7	47.1	56.5	54.3	69.	64.1	53.2	57.7	50.2	65.8	55.5	52.1	53.7	59.9	52.7	54.6	–	58.5	64.4	55.2	55.2
11	40.2	63.9	61.3	62.8	56.9	52.8	55.7	53.9	64.8	57.4	65.	55.0	56.2	51.5	51.4	60.4	–	52.7	53.1	65.8	46.8
12	62.5	51.9	64.3	63.4	53.7	59.	48.9	56.3	57.1	58.6	60.5	54.8	59.1	56.4	56.6	57.5	–	56.7	62.4	56.4	55.2
13	61.1	61.2	57.9	52.5	53.8	62.1	60.2	57.	56.	51.9	50.4	58.1	65.	56.3	58.7	57.6	–	61	55.4	50.8	65.
14	54.6	64.9	57.5	56.1	52.2	60.8	61.1	45.3	44.	60.2	55.1	55.4	57.3	54.1	68.4	52.8	–	61.2	53.9	53.4	57.1
15	44.8	61.6	56.7	68.6	59.3	49.1	54.2	61.5	55.3	62.	58.4	58.5	44.3	58.8	54.5	55.2	–	54.4	43.	64.3	44.2
16	57.	43.7	55.4	53.6	58.	57.7	59.3	66.	45.6	65.2	64.7	56.2	55.	64.	57.7	54.7	–	60.1	65.2	67.5	49.6
17	63.2	57.6	56.7	51.1	53.	55.5	56.2	52.3	57.	58.8	58.9	55.9	61.3	61.5	57.3	56.2	–	63.2	55.3	65.2	59.1
18	66.8	59.7	62.4	64.7	54.3	55.6	58.2	53.3	65.4	52.3	55.8	59.7	57.6	58.4	47.7	57.3	–	59.7	60.1	67.8	51.2
19	64.7	53.8	60.8	52.1	55.	52.7	59.5	50.3	57.	47.3	59.6	65.8	41.7	59.1	63.7	57.3	–	51.9	52.9	61.2	66.6
20	–	57.8	56.5	56.2	56.5	60.3	59.4	57.	68.8	51.9	65.3	66.8	66.9	60.	59.5	55.5	–	56.1	58.1	53.5	45.2
21	–	57.5	51.3	49.9	49.8	61.2	57.1	54.7	56.4	68.1	61.7	55.	59.1	62.9	64.8	51.1	–	64.3	59.9	53.7	43.3
22	–	59.7	59.3	57.2	59.6	60.8	55.7	65.5	44.1	45.6	62.1	59.8	60.5	52.7	52.5	59.4	–	52.5	55.4	60.7	64.5
23	–	63.7	50	45.8	47.5	55.2	65.4	54.6	59.9	61.	57.6	60.4	56.3	54.6	51.6	48.1	–	43.3	44.	57.8	46.8
24	–	57.6	66.6	52.3	55.9	53.6	55.9	58.9	46.9	67.2	53.1	61.9	60.3	59.	50.5	61.3	–	54.3	64.5	59.4	58.4
25	–	55.2	42.7	67.7	53.3	60.6	50.7	46.5	51.1	56.4	49.7	60.6	56.8	62.9	58.6	54.2	–	48.5	46.4	42.7	58.5
26	–	61.2	64.5	58.8	57.7	54.8	58.3	62.7	66.9	64.9	59.3	59.3	50.8	63.7	56.2	52.8	–	59.3	60.3	56.	55.6
27	–	61.4	53.2	59.4	65.5	58.4	52.7	53.	58.	60.2	60.1	58.3	57.1	53.3	67.4	52.7	–	56.8	53.7	59.8	56.9
28	–	62.2	59.7	60.3	59.5	52.7	67.5	60.8	54.1	60.2	59.8	44.2	60.7	62.2	65.6	56.7	–	62.5	47.	53.6	59.9
29	–	54.8	46.8	57.	53.2	50.2	51.8	49.9	61.5	55.	65.	55.	59.5	57.4	64.2	56.5	–	52.1	59.4	61.5	57.5
30	–	66.9	53.6	56.1	58.6	61.9	55.5	54.3	57.8	62.7	56.4	62.4	49.	53.4	52.	58.6	–	62.7	63.2	57.	42.6
31	–	56.2	66.1	56.5	54.7	57.3	53.8	47.1	63.5	45.6	54.8	50.8	61.5	54.4	64.2	61.3	–	52.8	55.3	60.9	60.1
32	–	44.1	67.8	52.3	55.4	56.1	56.4	53.	51.3	66.6	54.5	67.9	59.2	49.6	60.8	51.2	–	55.1	55.	60.6	60.9
33	–	60.1	46.2	60.1	60.2	57.6	65.3	63.	62.5	59.3	53.3	55.6	58.9	61.2	64.2	62.9	–	62.4	60.7	51.4	53.7
34	–	57.9	59.8	53.1	54.2	52.7	63.0	58.3	56.6	63.7	67.5	54.3	44.1	46.	54.8	51.5	–	53.7	56.1	61.6	61.8
35	–	55.7	56.7	45.1	62.5	64.1	64.2	54.5	43.3	62.7	55.6	51.4	66.	54.7	55.8	50.6	–	54.5	62.5	61.2	57.9
36	–	56.7	63.	51.1	67.6	53.8	65.5	42.5	58.2	59.7	62.2	62.9	64.5	64.7	59.2	54.8	–	59.9	53.7	49.5	60.2
37	–	52.	44.4	55.1	53.4	52.8	53.2	62.5	46.6	61.	63.4	62.9	43.7	52.4	55.2	62.3	–	68.1	54.7	45.7	53.6
38	–	46.5	46.1	62.9	58.3	60.5	57.2	60.8	58.	61.6	60.8	48.9	71.6	61.	57.6	52.7	–	54.1	54.2	57.	44.2
39	–	56.2	66.1	50.8	47.8	63.6	64.7	58.8	55.3	53.3	55.7	55.2	69.9	58.	52.5	65.8	–	60.9	56.	65.7	59.
40	–	62.7	62.9	52.8	56.5	65.6	62.3	60.5	59.3	58.1	56.8	54.2	56.4	47.7	56.3	52.3	–	56	50.8	55.4	67.8
41	–	57.3	61.4	63.6	48.8	60.5	59.7	47.2	55.2	57.2	57.8	70.6	55.4	53.2	59.4	66.6	–	54.2	43.5	61.1	59.3
42	–	63.4	–	55.4	63.	59.	51.5	61.4	56.2	53.4	55.2	77.8	61.3	59.4	63.3	54.8	58.8	55.4	42.9	59.4	58.6
43	–	58.1	–	69.	46.2	43.5	66.7	53.3	59.8	67.3	58.1	51.8	59.5	50.7	61.6	63.8	65.9	55.5	–	54.7	63.2
44	–	58.1	–	60.7	60.9	61.2	60.0	63.6	45.4	57.4	66.4	60.1	58.4	54.5	57.8	56.9	56.1	54.8	–	61.6	56.2
45	–	62.2	–	44.8	65.3	55.9	53.5	58.8	62.1	68.4	57.	55.8	44.6	64.5	56.9	53.1	52.6	55.4	–	53.6	56.8
46	–	51.8	–	57.7	55.	59.8	60.	54.3	53.	56.4	59.6	68.2	67.4	50.8	50.5	53.7	52.7	57.1	–	61.	55.1
47	–	58.9	–	60.4	58.1	61.6	61.5	46.9	67.9	45.5	62.4	53.2	71.6	73.2	54.2	46.3	59.3	54.4	–	60.4	63.
48	–	58.9	–	64.6	61.7	61.4	59.2	61.7	63.6	56.4	64.5	53.2	65.1	42.7	53.3	51.6	55.2	56.9	–	57.2	54.4
49	–	63.2	–	64.1	57.1	61.1	53.9	61.9	60.3	57.5	46.3	59.1	44.7	58.	46.7	57.2	58.7	58.5	–	52.1	58.2
50	–	59.6	–	55.4	54.	60.9	53.9	60.9	55.	45.7	65.2	64.9	55.9	62.1	59.2	58.8	58.4	57.8	–	59.9	62.5
51	–	59.	–	62.8	60.4	42.5	45.5	47.2	61.2	67.4	55.7	42.3	65.9	61.9	55.1	54.2	58.8	62.3	–	50.3	59.5
52	–	53.6	–	56.7	57.	54.6	57.3	63.1	66.5	52.9	60.	59.3	59.9	47.	51.	59.4	44.8	56.5	–	63.4	52.9
53	–	41.1	–	62.	54.3	57.5	51.6	50.9	55.8	47.6	61.9	66.5	38.3	45.	63.6	57.4	53.2	53.9	–	50.8	62.9
54	–	58.9	–	56.8	53.7	55.5	60.7	65.6	45.7	45.5	57.5	66.8	68.4	51.6	52.1	61.7	64.6	58.5	–	59.3	60.2
55	–	54.4	–	45.9	54.2	60.8	58.6	64.3	54.9	57.2	56.6	58.7	55.7	61.3	53.4	55.1	56.	53.5	–	55.2	52.1
56	–	59.3	–	59.6	–	43.9	46.6	49.4	56.3	53.1	58.	50.	56.4	58.3	44.	62.2	70.1	57.2	–	75.	53.4
57	–	63.1	–	49.5	–	55.6	57.2	60.7	43.9	63.6	62.5	59.3	56.9	59.1	52.2	62.	52.8	55.2	–	65.4	56.3
58	–	61.8	–	56.7	–	54.	62.2	57.2	64.6	51.7	57.6	60.2	62.8	55.9	53.3	63.8	57.9	63.2	–	69.6	51.2
59	–	63.5	–	65.	–	53.1	52.	63.6	57.2	52.5	45.4	56.7	51.7	60.3	55.2	56.9	53.2	54.8	–	55.	57.6
60	–	53.4	–	49.6	–	62.4	–	59.4	51.7	55.6	55.9	58.3	64.6	58.1	57.	65.3	57.9	55.9	–	51.7	54.1
61	–	58.1	–	55.4	–	58.6	–	56.6	52.4	53.1	55.2	56.8	59.	41.4	54.9	62.8	53.7	52.4	–	78.6	60.6
62	–	53.	–	61.4	–	66.4	–	57.4	52.3	60.	57.2	–	55.2	55.5	52.5	57.1	66.	56.1	–	41.2	68.1
63	–	64.1	–	54.2	–	61.2	–	66.9	63.3	53.8	67.8	–	74.1	59.3	48.	60.	54.7	59.5	–	–	58.4
64	–	57.9	–	51.4	–	66.9	–	60.7	62.7	60.7	60.2	–	65.	64.3	61.1	49.7	79.7	57.4	–	–	56.
65	–	57.2	–	64.6	–	60.5	–	52.8	56.5	54.6	61.5	–	60.2	44.2	55.	58.5	55.9	52.1	–	–	67.6
66	–	42.9	–	57.4	–	61.9	–	64.1	54.4	57.5	46.	–	57.2	61.2	56.8	57.4	57.6	59.3	–	–	51.5
67	–	53.8	–	47.3	–	42.7	–	42.7	–	65.1	62.7	–	55.1	58.1	53.2	63.7	61.	65.1	–	–	58.4
68	–	61.7	–	50.3	–	63.7	–	47.3	–	61.1	61.2	–	63.3	46.1	51.7	43.8	54.1	63.1	–	–	63.4
69	–	43.	–	56.1	–	57.	–	51.9	–	–	50.	–	46.7	54.3	53.7	57.7	67.3	50.9	–	–	67.3
70	–	56.	–	66.2	–	53.5	–	62.7	–	–	61.8	–	65.1	66.6	55.5	44.4	59.8	64.4	–	–	55.9
71	–	62.5	–	52.	–	56.4	–	68.4	–	–	61.1	–	64.4	68.2	67.2	57.8	53.6	52.1	–	–	66.1
72	–	64.9	–	59.8	–	64.6	–	63.2	–	–	61.5	–	66.5	59.2	54.8	57.2	60.2	61.3	–	–	61.1
73	–	63.	–	51.8	–	64.7	–	44.8	–	–	54.1	–	57.7	58.5	60.2	57.6	57.6	45.6	–	–	69.6
74	–	57.1	–	54.7	–	64.8	–	63.6	–	–	74.5	–	60.3	51.4	53.5	58.	40.5	52.1	–	–	62.5
75	–	46.2	–	61.3	–	60.7	–	60.2	–	–	56.8	–	62.1	57.7	64.5	66.	57.8	61.7	–	–	58.9
76	–	48.4	–	54.4	–	57.5	–	56.2	–	–	51.1	–	53.	63.8	57.1	64.9	–	59.7	–	–	62.3
77	–	58.2	–	61.8	–	60.	–	62.1	–	–	51.5	–	59.9	54.8	55.	53.9	–	62.9	–	–	49.7
78	–	52.2	–	57.1	–	63.8	–	51.2	–	–	57.6	–	65.9	60.1	47.6	44.5	–	56.4	–	–	62.6
79	–	55.1	–	64.1	–	52.9	–	58.3	–	–	45.2	–	57.	65.	45.8	57.7	–	56.5	–	–	63.1
80	–	67.4	–	55.1	–	60.2	–	57.8	–	–	57.6	–	55.3	67.6	53.1	54.4	–	43.5	–	–	56.1
81	–	54.2	–	61.7	–	51.2	–	60.4	–	–	64.9	–	67.8	46.5	59.3	60.6	–	53.9	–	–	69.1
82	–	52.	–	67.1	–	67.3	–	62.4	–	–	44.1	–	56.	57.1	59.	57.8	–	58	–	–	68.3
83	–	53.3	–	55.7	–	49.4	–	57.2	–	–	55.8	–	53.2	60.7	56.7	55.5	–	55.7	–	–	61.8
84	–	56.8	–	50.3	–	50.3	–	44.	–	–	53.8	–	59.	57.9	58.9	55.4	–	52.6	–	–	47.9
85	–	60.2	–	53.2	–	63.	–	51.5	–	–	58.1	–	60.6	61.2	59.5	63.	–	53.9	–	–	64.3
86	–	59.1	–	51.2	–	45.3	–	48.3	–	–	–	–	60.2	54.4	58.8	60.2	–	52.1	–	–	54.2
87	–	61.4	–	56.7	–	54.2	–	58.4	–	–	–	–	58.5	56.9	65.	79.2	–	57.3	–	–	56.
88	–	56.4	–	65.5	–	63.8	–	62.8	–	–	–	–	60.1	60.6	45.2	56.8	–	66.3	–	–	60.4
89	–	–	–	41.9	–	64.9	–	48.8	–	–	–	–	50.9	58.8	53.3	54.2	–	53.4	–	–	45.6
90	–	–	–	60.8	–	55.6	–	55.5	–	–	–	–	57.7	65.8	45.9	54.6	–	54.6	–	–	90.8
91	–	–	–	51.9	–	59.9	–	53.3	–	–	–	–	59.2	61.1	55.5	52.1	–	52.5	–	–	52.3
92	–	–	–	58.7	–	51.4	–	64.4	–	–	–	–	57.7	59.2	55.5	58.3	–	59.2	–	–	60.7
93	–	–	–	63.4	–	63.	–	42.8	–	–	–	–	56.4	50.7	59.1	59.5	–	56.2	–	–	50.
94	–	–	–	54.7	–	64.	–	61.2	–	–	–	–	60.5	60.5	56.1	61.8	–	53.3	–	–	51.9
95	–	–	–	64.7	–	47.1	–	64.9	–	–	–	–	58.8	59.	42.9	57.2	–	66.2	–	–	55.7
96	–	–	–	65.2	–	45.4	–	59.	–	–	–	–	70.5	60.7	55.5	64.1	–	54.5	–	–	65.8
97	–	–	–	57.1	–	42.2	–	54.1	–	–	–	–	58.1	58.6	63.4	62.5	–	59.9	–	–	57.8
98	–	–	–	54.	–	56.2	–	57.1	–	–	–	–	–	61.6	50.2	59.1	–	59.2	–	–	62.6
99	–	–	–	57.4	–	50.6	–	57.7	–	–	–	–	–	63.4	57.3	57.7	–	64.3	–	–	54.
100	–	–	–	53.5	–	63.4	–	57.7	–	–	–	–	–	58.	57.3	58.6	–	62.8	–	–	54.4
101	–	–	–	43.9	–	53.8	–	44.6	–	–	–	–	–	55.1	52.1	54.8	–	61.2	–	–	64.4
102	–	–	–	53.	–	56.4	–	59.	–	–	–	–	–	59.6	56.9	46.7	–	55.7	–	–	52.5
103	–	–	–	55.8	–	58.3	–	52.3	–	–	–	–	–	58.9	60.3	53.2	–	55	–	–	55.9
104	–	–	–	55.	–	58.9	–	53.7	–	–	–	–	–	56.2	56.4	44.9	–	54.2	–	–	55.2
105	–	–	–	52.4	–	62.	–	56.9	–	–	–	–	–	59.2	63.7	64.	–	54.1	–	–	58.1
106	–	–	–	62.2	–	56.3	–	47.1	–	–	–	–	–	57.4	49.	56.2	–	58.2	–	–	55.3
107	–	–	–	58.5	–	53.5	–	49.5	–	–	–	–	–	67.6	55.	62.4	–	61	–	–	54.9
108	–	–	–	56.8	–	60.8	–	66.6	–	–	–	–	–	58.2	56.4	54.1	–	66.7	–	–	53.6
109	–	–	–	52.8	–	59.9	–	53.6	–	–	–	–	–	60.3	53.4	66.5	–	55.1	–	–	46.9
110	–	–	–	100.	–	53.2	–	56.3	–	–	–	–	–	–	57.6	55.8	–	51.2	–	–	52.6
111	–	–	–	44.2	–	65.9	–	44.7	–	–	–	–	–	–	56.8	62.5	–	61.1	–	–	50.8
112	–	–	–	56.6	–	55.9	–	60.	–	–	–	–	–	–	59.	60.6	–	61.1	–	–	42.2
113	–	–	–	55.1	–	63.6	–	57.8	–	–	–	–	–	–	54.5	54.7	–	58.1	–	–	57.6
114	–	–	–	65.2	–	56.	–	57.4	–	–	–	–	–	–	57.6	51.2	–	64	–	–	66.9
115	–	–	–	63.	–	60.6	–	52.3	–	–	–	–	–	–	53.2	53.7	–	59.5	–	–	63.4
116	–	–	–	51.1	–	52.1	–	56.7	–	–	–	–	–	–	51.5	50.6	–	57.5	–	–	62.8
117	–	–	–	56.8	–	50.7	–	65.1	–	–	–	–	–	–	55.2	57.5	–	59.4	–	–	40.9
118	–	–	–	53.7	–	58.2	–	65.1	–	–	–	–	–	–	68.	56.7	–	61.9	–	–	57.1
119	–	–	–	56.4	–	64.5	–	54.8	–	–	–	–	–	–	55.5	60.9	–	57.9	–	–	60.5
120	–	–	–	60.7	–	57.1	–	63.5	–	–	–	–	–	–	60.6	56.5	–	64.1	–	–	75.8
121	–	–	–	59.2	–	44.7	–	57.3	–	–	–	–	–	–	57.1	59.1	–	44.9	–	–	68.
122	–	–	–	44.3	–	54.9	–	54.4	–	–	–	–	–	–	56.8	56.3	–	54.2	–	–	53.9
123	–	–	–	57.5	–	64.7	–	55.4	–	–	–	–	–	–	54.2	44.9	–	55.2	–	–	41.3
124	–	–	–	57.4	–	57.9	–	49.9	–	–	–	–	–	–	56.2	51.5	–	48.7	–	–	70.5
125	–	–	–	58.6	–	60.4	–	59.	–	–	–	–	–	–	62.9	50.5	–	49.4	–	–	43.2
126	–	–	–	62.6	–	56.4	–	57.3	–	–	–	–	–	–	60.3	54.7	–	58.1	–	–	47.8
127	–	–	–	47.6	–	56.4	–	57.5	–	–	–	–	–	–	56.1	64.9	–	52.2	–	–	58.1
128	–	–	–	59.3	–	55.9	–	58.5	–	–	–	–	–	–	55.4	63.7	–	56.7	–	–	69.
129	–	–	–	58.9	–	51.7	–	67.6	–	–	–	–	–	–	–	59.7	–	46.4	–	–	53.4
130	–	–	–	52.7	–	63.	–	58.1	–	–	–	–	–	–	–	58.3	–	50.6	–	–	45.8
131	–	–	–	60.5	–	52.3	–	53.9	–	–	–	–	–	–	–	58.8	–	54.8	–	–	56.4
132	–	–	–	51.1	–	50.6	–	44.7	–	–	–	–	–	–	–	61.9	–	49.8	–	–	59.1
133	–	–	–	50.9	–	65.8	–	60.5	–	–	–	–	–	–	–	61.4	–	57.1	–	–	64.7
134	–	–	–	51.8	–	57.8	–	56.	–	–	–	–	–	–	–	65.2	–	51.5	–	–	52.
135	–	–	–	45.2	–	67.1	–	51.4	–	–	–	–	–	–	–	55.1	–	56.2	–	–	55.5
136	–	–	–	43.8	–	66.6	–	61.4	–	–	–	–	–	–	–	–	–	55	–	–	60.1
137	–	–	–	62.5	–	56.9	–	59.8	–	–	–	–	–	–	–	–	–	59.1	–	–	56.9
138	–	–	–	61.2	–	63.4	–	61.7	–	–	–	–	–	–	–	–	–	58.6	–	–	57.8
139	–	–	–	41.	–	55.3	–	54.6	–	–	–	–	–	–	–	–	–	55.5	–	–	57.4
140	–	–	–	63.3	–	59.5	–	–	–	–	–	–	–	–	–	–	–	53.6	–	–	55.4
141	–	–	–	53.4	–	58.9	–	–	–	–	–	–	–	–	–	–	–	60.7	–	–	49.8
142	–	–	–	54.5	–	53.3	–	–	–	–	–	–	–	–	–	–	–	61.1	–	–	53.3
143	–	–	–	52.	–	53.9	–	–	–	–	–	–	–	–	–	–	–	51.7	–	–	53.4
144	–	–	–	53.5	–	56.1	–	–	–	–	–	–	–	–	–	–	–	50.3	–	–	58.5
145	–	–	–	62.7	–	52.	–	–	–	–	–	–	–	–	–	–	–	54.8	–	–	75.3
146	–	–	–	55.7	–	57.9	–	–	–	–	–	–	–	–	–	–	–	67.9	–	–	52.6
147	–	–	–	55.7	–	54.7	–	–	–	–	–	–	–	–	–	–	–	56.8	–	–	62.3
148	–	–	–	51.6	–	60.4	–	–	–	–	–	–	–	–	–	–	–	56.5	–	–	48.4
149	–	–	–	54.4	–	57.8	–	–	–	–	–	–	–	–	–	–	–	60.1	–	–	63.6
150	–	–	–	46.5	–	59.4	–	–	–	–	–	–	–	–	–	–	–	45.4	–	–	48.
151	–	–	–	65.9	–	60.6	–	–	–	–	–	–	–	–	–	–	–	–	–	–	61.6
152	–	–	–	56.4	–	58.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	63.4
153	–	–	–	54.7	–	55.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	65.
154	–	–	–	61.4	–	56.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	51.1
155	–	–	–	62.7	–	61.3	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
156	–	–	–	–	–	56.1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
157	–	–	–	–	–	59.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
158	–	–	–	–	–	55.	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
159	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

TABLE S3. **Discriminating 2-state from 3-state folders**

Parameter	p value[a]
⟨D_raw_min⟩	9.0e-3
⟨D_raw_equil⟩	7.2e-3
⟨D_smooth_min⟩	6.5e-3
⟨D_smooth_equil⟩	6.8e-3
⟨D_raw_min_laminar⟩	4.0e-2
⟨D_raw_equil_laminar⟩	5.2e-3
⟨D_smooth_min_laminar⟩	4.5e-3
⟨D_smooth_equil_laminar⟩	3.0e-4
⟨D_raw_min_turbulent⟩	1.9e-2
⟨D_raw_equil_turbulent⟩	4.1e-3
⟨D_smooth_min_turbulent⟩	2.2e-2
⟨D_smooth_equil_turbulent⟩	3.2e-2
⟨D_NC⟩	6.7e-3
⟨RMSD⟩	1.2e-2
ACO	1.3e-1
Length	1.3e-3
⟨GDT-TS⟩	4.5e-3
⟨TM-Score⟩	1.0e-1

[a] Statistical significance (p-values) for each metric in this table is determined by a t-test with null hypothesis that values of a given metric for the 2-state and 3-state folders come from independent random samples with normal distributions having equal means and equal but unknown variances, against the alternative that the means are not equal. The values in this table are plotted in figure 12 of the main text.