successes and limitations of current computational methods and underscores how computational results can be used to inform experiments. These powerful simulation approaches in combination with experiments can provide unique insights into how large proteins fold and misfold, expanding our ability to predict and manipulate protein folding.

To function, structured proteins need to reproducibly fold to a unique three-dimensional structure in a biologically reasonable timescale. The observation that proteins reliably fold despite having astronomical numbers of possible conformations has been the impetus behind decades of experimental and theoretical folding studies (1–3). However, protein-folding pathways and folding intermediates are of interest not only in fundamental biophysics. Partially folded states expose surfaces that are normally buried. If these states are populated for extended periods of time, they may be recognized by elements of the cell’s protein quality control machinery that can assist in folding or target proteins for degradation (4–7). Further, in many protein-folding diseases, pathological mutant proteins populate partially folded nonnative conformations and these conformations may result in nonnative protein–protein associations or oligomerization. A detailed understanding of protein folding and misfolding pathways thus has the potential to aid in the development of therapeutic interventions that prevent misfolding or reduce the population of intermediates. Alternatively, in diseased cells, drugs could be designed that promote misfolding and drive cells into apoptosis.

A challenge to developing such a detailed understanding is posed by the transient nature of the intermediates. Even the most long-lived folding intermediates rarely persist beyond timescales of a few minutes at most. Experimentally, transient intermediate states have been detected and characterized using spectroscopic methods such as fluorescence and CD as well as by small-angle X-ray scattering and other scattering methods, and, although technical advances have allowed the detection of rare and excited states by NMR (8, 9) and X-ray crystallography (10), these molecularly detailed methods may not be widely available, or such experiments may not be sufficient to detect folding intermediates.
By contrast, computer simulations provide a detailed picture of protein folding not easily accessible to experiment. Specifically, protein-folding simulations can provide valuable, detailed, testable data on how proteins fold and misfold and may be used to formulate hypotheses on how protein folding might be manipulated. In the last decade, special purpose supercomputers such as ANTON (11) and massively distributed computing schemes such as Folding@home (12) have made it possible to simulate the folding of small proteins in all-atom detail using realistic empirical force fields, without the aid of any biasing forces. Whereas these simulations have provided invaluable insights, even special purpose, high-performance computing platforms are limited to simulating the folding of smaller chains (currently ~100 amino acids with folding times up to milliseconds). However, the median length of a protein is 532, 365, and 329 amino acids in eukaryotes, bacteria, and archaea, respectively (13), and folding times range from microseconds to tens of minutes. Even with anticipated continuing increases in computing power, simulating the folding of these larger slower folding proteins using standard molecular dynamics (MD) simulations will remain out of reach for the foreseeable future. In addition to the increased computational demands due to size, large slow folding proteins often fold through long-lived intermediates corresponding to deep local energy minima. For such proteins, even very long simulations are likely to simply observe the protein exploring limited conformational space within a single local minimum because transitions between minima are rare.

To simulate such rare transitions, numerous methods for accelerating or enhancing sampling during MD simulations have been developed (14–23). One particularly effective approach to efficiently simulate the folding of large proteins is to take advantage of the evidence that protein folding is to a large extent encoded by the contacts in the native, folded protein. Computational schemes in which the force field explicitly includes a biasing term favoring native contacts allow the simulation of protein folding with many orders of magnitude less computational effort than is required with conventional MD methods. These studies, including folding simulations of large proteins such as adenylate kinase (24, 25), GFP (26), TIM barrels (27), dihydrofolate reductase (28), a DNA polymerase (29), and serpins (30, 31), have been successful in generating folding pathways and intermediates that agree with experimental results and provide testable hypotheses on what intermediate states are likely to be populated during folding.

This review focuses on native-centric simulation methods that are applicable to the folding of large and slow-folding proteins. We consider two categories of techniques, both of which rely on the native protein contact map but have different levels of spatial resolution and chemical detail: (i) Gō and related structure-based models (SBMs) that provide knowledge on the effects of structure on folding and (ii) all-atom-based methods that take into account the effects of side-chain chemistry on folding and can therefore predict how mutations may affect folding.

As a prototypical case study, we discuss how both classes of methods have been used to simulate the folding of the canonical inhibitory serpin α1-antitrypsin (AAT) a 394-amino acid protein with folding times as long as tens of minutes (32–34). The interest in this specific protein resides in the fact that it has a topologically complex native structure, and the functional conformation is a kinetically trapped state, not the lowest free energy conformation (35–39). Furthermore, specific point mutations are known to enhance its misfolding propensity, giving rise to misfolding diseases (40, 41). We show how the two theoretical methods discussed above provide complementary results and how these results may be used to inform experiments, interpret experimental results, and generate hypotheses on how folding proteins may interact with the protein quality control machinery.

The importance of native interactions and the native-centric approach to protein folding

A physical basis and justification for native-centric approaches to modeling protein folding was provided by the energy landscape picture of protein folding and the principle of minimal frustration developed in the 1980s and 1990s (1). The effective potential energy (averaged over solvent degrees of freedom) as a function of chain conformation defines a protein’s energy landscape (Fig. 1). This multidimensional energy landscape results from multiple driving forces and constraints, including the drive to bury hydrophobic residues, to satisfy hydrogen bond donors and acceptors, and to solvate or pair charged residues and the constraints of chain connectivity and steric clashes. According to arguments adapted from the physics of disordered systems, random amino acid sequences will be characterized by irreconcilable conflicts between these multiple driving forces and constraints, termed energetic frustration, resulting in many unrelated structures of similar energy (42, 43).

The principle of minimal frustration states that, unlike random sequences, the sequences of naturally occurring proteins have been selected by evolution to minimize energetic conflicts between interactions in the native conformation (1). As a result, compared with alternative structures, the native structure is a better-optimized solution to the problem of satisfying the large number of competing interactions. The energy landscape of such a minimally frustrated protein resembles a high-dimensional funnel with the native state at the bottom. Although local minima and barriers still exist (the funnel is “rugged”), the global energetic bias toward the native structure promotes efficient folding (Fig. 1).

A funneled energy landscape implies that the potential energy of the system should decrease with increasing numbers of native contacts, suggesting that the number of native contacts formed should serve as a good reaction coordinate for folding. Early support for a native-centric picture of protein folding was provided by simulations of simplified lattice pro-

---

The abbreviations used are: MD, molecular dynamics; AAT, α1-antitrypsin; SBM, structure-based model; BF, bias functional; SCP5, self-consistent path sampling; RCL, reactive center loop; ER, endoplasmic reticulum; AWSEM, associative memory, water-mediated, structure and energy model; DHFR, dihydrofolate reductase; DLD, discontinuous loop domain; ABDD, adenosine-binding domain; WSME, Wako-Saito-Mun˜ oz-Eaton; AKE, adenylate kinase; HDX, hydrogen-deuterium exchange; PDB, Protein Data Bank; FEP, folding free energy profile(s).

5 The abbreviations used are: MD, molecular dynamics; AAT, α1-antitrypsin; SBM, structure-based model; BF, bias functional; SCP5, self-consistent path sampling; RCL, reactive center loop; ER, endoplasmic reticulum; AWSEM, associative memory, water-mediated, structure and energy model; DHFR, dihydrofolate reductase; DLD, discontinuous loop domain; ABDD, adenosine-binding domain; WSME, Wako-Saito-Mun˜ oz-Eaton; AKE, adenylate kinase; HDX, hydrogen-deuterium exchange; PDB, Protein Data Bank; FEP, folding free energy profile(s).
proteins (14, 44–47). More recently, Best et al. (48) analyzed ultralong unbiased all-atom simulations of several small (<100-aa) fast-folding proteins and found that during transitions between unfolded and folded states, native contacts persisted significantly longer than nonnative contacts.

**Gō models encode native structure and can be used to understand protein-folding dynamics**

The funneled shape of a protein’s energy landscape implies that there is a minimal chance that the polypeptide chain remains trapped in local energy minima (2, 49, 50). In the ideal funnel limit, proteins can smoothly flow from unfolded states at the top of the funnel to the folded ensemble. On the other hand, this process is associated with a large configurational entropy reduction, which can give rise to a significant free energy barrier. As a result, even in the minimally rugged funnels, protein folding remains a rare, thermally activated reaction.

Gō models of proteins further simplify the nature of the folding landscape by encoding the native structure of the protein in the potential energy and, for the most part, ignoring all nonnative interactions (14, 44, 51–53). This reduces the complexity of the calculations and makes the rare folding events computationally accessible. The term structure-based models (SBMs) has been used to refer to Gō-type models, which, in addition to native structure-derived terms, may contain knowledge-based terms that are derived from sequence information (15, 54), nonnative interactions (55), information from additional native structures (56), etc. However, because this review deals mostly with models that include few non-structure-derived interactions, we use the terms Gō models and SBMs interchangeably.

Gō models have commonly been simulated using either Monte Carlo or MD methods and are computationally relatively inexpensive. A further reduction in complexity can be achieved through coarse-graining to include either a single Cα bead or a few beads per residue (Fig. 2). Thus, Gō models are simple protein models that enable extensive sampling of potential energy landscapes. Additionally, because so few beads are involved, the data are easier to interpret. Further, standard enhanced sampling methods, such as replica exchange, used in MD simulations can be used when folding proteins with large barriers (57, 58). Gō-model simulations have been successfully used to

---

6 Please note that the JBC is not responsible for the long-term archiving and maintenance of this site or any other third party hosted site.
understand both the structural (folding mechanisms, intermediate populations, etc.) and the kinetic features (barrier heights, rates of different events, etc.) of protein folding (14, 51–53).

**Encoding structure in Go models**

Go models encode protein structure through two features: (i) the chain connectivity of the protein and (ii) contact interactions present in the native state (Fig. 2). Chain connectivity is encoded by having strong bonds (that cannot break during the folding simulation) between those beads that represent atoms or groups of atoms that are connected by chemical bonds in the protein. Additional local (along the backbone) interactions can include strong angular constraints between three consecutive beads (connected by bonds) and dihedral interactions between four consecutive beads. Strong dihedral interactions may be used to preserve chain chirality or other structural constraints, such as the planarity of rings. Weaker structure-derived or statistical dihedral interactions serve as a proxy for secondary structural propensity. These interactions are weak enough to break and form on the timescale of the simulation. Contact interactions are defined between two beads that do not interact through any local interactions and that are close in space in the folded structure of the protein. Two beads are considered to be close in space if either they or, alternately, the atoms that they represent are within a cutoff distance of each other. Other more complex ways of defining contact interactions exist and sometimes work better in folding simulations than the cutoff-based contact map (59). Contacts may also all have the same (homogeneous) strengths or be assigned contact strengths based on the atoms or the residues that are in contact. Finally, an excluded volume interaction is present between those beads that are not in contact in the native state. This ensures that beads do not pass through each other and the chain does not cross. Different flavors of Go models have been tested on a few proteins, and for the most part, as long as little to no frustration in the form of nonnative interactions is encoded, diverse features of protein folding calculated from the different models are both similar to each other and to experimental results (15, 60, 61). However, functional regions of proteins can be either energetically or topologically frustrated, and in such cases, different unfrustrated Go models can give different folding features (62, 63), and energetic frustration in the form of nonnative interactions (Fig. 2) needs to be explicitly included in the model (54, 64–66).

A key advantage of using a Go model is that it can be easily modified to include specific native or nonnative terms (increasing the complexity of the model) when data from the simplest model do not agree with a specific experimental observable, such as the population of an intermediate. When the addition of an extra term to the potential energy function leads to an agreement with experiments, then it can be concluded that that term leads to that specific feature. In other words, in some cases, extra terms can lead to a better understanding of the physical basis of experimental findings.

**Encoding frustration in Go models**

As stated earlier, the funneled energy landscape of structured proteins is a result of selection for protein sequences in which interactions present in the folded state are much more stabilizing than nonnative interactions, thus reducing frustration (2, 49, 50). However, residues that perform function in the folded state need to be conserved and cannot always be chosen to reduce energetic frustration. The interactions of functional residues (e.g. amino acids in the active site or at protein-protein interfaces) with residues that have been selected to reduce frustration and promote folding can lead to trapping during folding (67, 68). Such functional frustration can show up as an increase in complexity of the folded state due to the presence of additional functional secondary structural elements in the native state. This can lead to larger barriers to folding, backtracking during folding, etc. An alternate signature of functional frustration is a loss of contacts because functional amino acids are chosen to stabilize interprotein interactions rather than intra-protein interactions. This can lead to lower barriers to folding, stalling of folding and the population of intermediates, etc. Such effects of function, which induce localized changes in protein structure, are detectable even in Go models that do not encode frustration (67, 68).

However, for some proteins, nonnative interactions need to be explicitly encoded in the Go model for key folding features such as the population of a folding intermediate to agree with experiments (54, 64–66). Nonspecific nonnative interactions can be added to purely structure-based Go models by introducing an attractive interaction at a chosen interaction distance between all of those pairs of beads that are not in native contact (Fig. 2). Alternately, nonnative interactions may be added between selected groups of amino acids, such as the hydrophobic amino acids or the charged amino acids. Such nonspecific nonnative interactions, their forms, and their utility have been reviewed in detail elsewhere (55, 69).

Several proteins undergo conformational transitions, converting from one structural ensemble to a distinct structural ensemble upon ligand binding or chemical modification. In such cases, interactions that are “nonnative” in the unbound (or unmodified) structural ensemble are formed in the bound (or modified) ensemble. Such specific nonnative interactions can be appended to the Go model of the unbound structure (51, 56, 70). This class of models, termed dual-structure–based models, has often been used to understand conformational transitions of proteins but has rarely been used to understand folding. However, when a single large conformational transition dominates the function of a protein, which is the case for serpins, it is likely that simulations using such dual-structure–based models will be a computationally inexpensive way to capture the functional frustration present in the folding energy landscape.

**All-atom-enhanced sampling based on a native-centric biasing force**

An alternative strategy for simulating rare macromolecular transitions consists of retaining full all-atom resolution with chemically motivated realistic forces, but resorting to more sophisticated algorithms, possibly combined with additional approximations, to lower the cost of characterizing rare transitions. Although computationally more expensive, such methods have the advantage of accounting for the chemistry of the side chains and can be used to investigate the effects of muta-
Many such enhanced path-sampling techniques have been developed during the last 2 decades (for a recent review, see, for example, Ref. 20). Some of these methods are based on reconstructing the reaction kinetics from a statistical analysis of many short MD trajectories (an incomplete list includes transition interface sampling (17, 21), Markov state models (22), and milestoning (19)). This way, it is in principle possible to obtain predictions statistically consistent with plain MD simulations while massively distributing the computational load. In practice, however, these schemes still require huge computational resources and cannot be applied to slow, complex reactions, such as the folding or conformational changes of large proteins, which can occur on time scales of minutes to hours.

A computationally efficient way to further lower the computational cost of atomistic simulations consists of introducing biasing forces to promote escape from metastable states (e.g. see Refs. 18 and 71–74). One particularly useful biasing force is implemented in ratchet-and-pawl MD, in which a history-dependent force is only applied to prevent the system from backtracking (71, 72). In native-centric simulations, the collective coordinate of interest for backtracking is the contact map distance from the native state. In this implementation of ratchet-and-pawl MD, no biasing force is present as long as the protein is not moving away from the native state (e.g. as long as the total number of native contacts formed remains constant or increases or no nonnative contacts are formed). When the biasing force is on, the simulation is identical to a conventional MD simulation, and the motion of the system is determined entirely by the physical forces between atoms. When the system evolves in such a way that the total number of native contacts decreases or nonnative contacts are formed, then a biasing force is applied to discourage (but not absolutely prevent) the move from occurring. Unlike Gō type models, the energy landscape is not a smooth ideal funnel, and local minima and barriers still exist. However, the ratchet force facilitates escape from local minima and therefore allows for computationally efficient simulations of folding (Fig. 3).

**Figure 3. The BF method for simulating protein folding using all-atom force fields and ratchet-and-pawl MD.** A, schematic view of multiple folding trajectories from the BF approach to transition path sampling. In the BF approach, the force field, \( V(R) \), is a conventional, all-atom force field (e.g. Amber or CHARMM) plus the history-dependent ratchet-and-pawl bias allowing for the efficient production of multiple, trial folding trajectories (lines in the funnel). The ratchet-and-pawl (right) bias limits backtracking, and the least biased trajectories (red) are selected for analysis. (The ratchet figure is from Antoni Espinosa (commons.wikimedia.org/wiki/File:Trinquete.png). The funnel is from the Oas laboratory at Duke University (https://oaslab.com/Drawing_funnels.html). B, schematic explanation of the steps involved in implementing the BF method for folding simulations adapted from Wang et al. (31). In the initial step, the protein of interest is unfolded using high-temperature MD simulations. BF folding simulations are then performed using the force field defined above, and multiple trial folding trajectories are generated. Note that folding is not always successful, and some protein molecules fail to fold completely or misfold, and these results may be particularly pronounced for mutant proteins. Folding and misfolding trajectories with minimum biasing (yellow lines) are identified and analyzed.
One of the first challenges in introducing a bias is that if the reaction coordinate chosen for biasing is suboptimal, uncontrolled systematic errors may occur. A number of approaches have been proposed to keep these errors to a minimum. In particular, the bias functional (BF) method (74) relies on generating a large number of trial transition pathways using ratchet and pawl MD and subsequently scoring these trial pathways according to a specific penalty function. It can be shown that the paths corresponding to the lowest value of this penalty function are the most realistic, in the sense that they have the largest probability to occur in the absence of the biasing force.

Such post-processing of ensembles of possible transition pathways has the advantage of keeping the systematic error introduced by the biasing force to a minimum while enabling extremely slow and complex reactions to be simulated on typical computer clusters available to most computational biophysics/biochemistry laboratories.

The BF approach has been benchmarked against the results of plain MD simulations for the folding of small proteins (74) and directly against experimental data for folding kinetics (75, 76). It has since been applied to simulate very slow folding reactions of large proteins, including serpins (31) and even proteins with knotted native structures (65). These processes are far too slow to be simulated by plain MD, even by massively distributed computing or by resorting to the largest special-purpose supercomputing facility. The BF approach successfully predicted differences in the folding kinetics of two structurally homologous proteins by showing that one of these proteins had to overcome an additional free-energy barrier to reach the native state (75). In this case, the chemical information present in the atomistic description of the amino acids was required to distinguish between the folding pathways of the structurally homologous proteins.

Despite these promising results, it should be emphasized that if the collective coordinate (e.g. the contact map) adopted in the definition of the biasing force is not a good reaction coordinate, then the minimum-bias paths identified by the BF scheme will still be plagued by large systematic errors. In principle, these errors could affect the reliability of the BF calculations, in particular when they are applied to study the folding of large chains with complex folding mechanisms, where a biasing coordinate based on the distance from the native contact map may not be a good reaction coordinate.

To tackle this problem, an important improvement of the BF method, called self-consistent path sampling (SCPS) has been recently introduced (23). In this scheme, the biasing collective coordinate is not chosen a priori. Instead, it is calculated self-consistently, through an iterative procedure, starting from an initial guess. This way, at convergence, the dynamics is accelerated by a bias that acts along a direction set by a realistic reaction coordinate (technically, a parametrization of the so-called committor function). A number of numerical tests on toy systems designed to emphasize the problems of the BF method have shown that SCPS can lead to significant improvement. The computational cost of SCPS simulations is, however, about 1 order of magnitude larger than that of BF simulations. Although SCPS simulations are significantly more computationally challenging, they are still feasible using existing super-computing facilities or clusters based on a hybrid GPU-CPU computing platform.

Finally, it should be emphasized that unlike structure-based methods discussed above, BF and SCPS are intrinsically non-equilibrium methods. As a result, the statistical methods that are commonly adopted to estimate equilibrium properties (such as free energy barriers) from molecular simulations cannot be applied. Thus, most of the applications of BF and SCPS made to date are based on semiquantitative analyses. For example, studies have been performed to estimate the relative change in free-energy generated by specific point mutations. Recently, more elaborate statistical methods have been proposed for recovering equilibrium distributions from SCPS and BF nonequilibrium simulations. For example, a recently proposed scheme makes it possible to sample the Boltzmann distribution in the transition region by means of specific ratchet-and-pawl simulations (77). To date, however, these techniques have only been validated against MD simulations for simple systems, and further validation is required to assess their accuracy for realistic and biologically relevant systems.

Both Gō and BF computational methods have been used to simulate the folding of the human serpin AAT (30, 31). We therefore next describe the structural properties of AAT and highlight how these two approaches have led to complementary insights into the complex folding mechanisms of this inhibitory serpin.

Inhibitory serpins: A prototype for folding large, topologically complex proteins

Inhibitory serpins are the most common inhibitors of serine and cysteine proteases and are found in all kingdoms of life (78). Large conformational changes of these two-domain, ~400-amino-acid-long, topologically complex proteins are required for both regulation and function (Fig. 4). The ubiquity of this multidomain protein family, the functionally required metastability of the active structure, and human diseases associated with serpin misfolding motivated both Gō- and BF-based folding simulations of AAT, the canonical human inhibitory serpin (30, 31).

Inhibitory serpins regulate target proteases by mechanically deforming the protease active site (79–81). The energy for this mechanical process is stored in the metastable, stressed serpin conformation characterized by the solvent-exposed reactive center loop (RCL), which acts as bait for proteases (Fig. 4). The initial stages in interactions between the RCL and the Ser or Cys target protease are the same as those for a normal protease substrate; the target protease docks to the RCL, the acyl intermediate with a covalent bond between the catalytic Ser in the protease and the RCL is formed, and the peptide bond is cleaved. In inhibitory serpins, RCL cleavage leads to insertion of the cleaved RCL into β sheet A, the central β sheet, as a sixth, central strand, translocating the covalently attached protease ~70 Å relative to the serpin, increasing the distance between the catalytic Ser and His residues in the protease catalytic triad from ~3 to ~6 Å and trapping the acyl intermediate with its covalent bond between the catalytic Ser and the cleaved RCL (80–83). This reaction results in a protease-serpin covalent complex containing an inhibited, mechanically deformed pro-
Inhibitory serpin structure and function. Shown is an active, metastable AAT structure (PDB entry 1QLP) (88) with a solvent-accessible RCL (purple). The structure is colored from blue to pink from the N to the C terminus. The α/β domain (CATH domain 2) is shown in blue (residues 23–190) and yellow (residues 290–340), whereas the mainly β domain (CATH domain 1), which includes the solvent-exposed RCL (residues 341–361), is shown in green (residues 191–290), purple (RCL), and pink (residues 362–394). Spontaneous insertion of the RCL into sheet A remodels the domains, adding the RCL to the α/β domain, resulting in the lower-free-energy inactive latent state (PDB entry 1I2Z (39)), and the latency transition is important for regulating the activity of some serpins (36). Cleavage of the RCL by target serine and cysteine proteases results in the formation of an acyl enzyme bond between the protease active site and the RCL,

cleavage of the RCL by target serine and cysteine proteases results in the formation of an acyl enzyme bond between the protease active site and the RCL,
rather than in specific sequences, implying that Co-based SBMs should be sufficient to explain this phenomenon.

To address this question, Giri Rao and Gosavi (30) used Gō models to simulate folding of human AAT to the metastable, active, and latent conformations. Proteins that undergo large conformational transitions between stable structures have often been studied using dual-structure–based models in which favorable energies are included for contacts present in either of the two conformations, leading to an energy landscape with two alternative minima. However, to address the basis of serpin folding to a metastable conformation, a different approach was taken. Two independent Gō model-folding simulations were performed. In one, the native (target) state was taken to be the metastable active structure, and in the other, the native state was the more stable latent structure (see Figs. 4 and 5).

The Gō-model simulations of AAT folding to the metastable, active structure found that the mainly β domain (CATH domain 1), including most of sheets B and C but lacking the two C-terminal β strands, 4 and 5B, folds early, whereas the α/β domain remains largely unfolded (Fig. 5). This partially folded structure with a largely folded β domain comprises a major intermediate along the folding pathway consistent with the experimental observations that folding to the active, metastable conformation involves at least one intermediate (32–34, 39, 87, 90, 91). Subsequently, the α/β domain folds, and, in one of the last folding steps, strands 4 and 5B are incorporated into the mainly β domain. This order of events is in good agreement with available data on AAT-folding kinetics from hydrogen/deuterium exchange coupled to MS (33), fast photochemical oxidation coupled to MS (34), and tryptophan fluorescence spectroscopy (32).

Simulations of folding to the latent state showed significant differences from folding to the metastable structure (Fig. 5). In the latent state, the inserted RCL is strand 4A in the α/β domain, hydrogen-bonding with strands 3 and 5A (Figs. 4 and 5). During folding to the latent state, stable contacts between the RCL (4A) and β strand 5A formed early, and consolidation of these two strands resulted in concerted folding of the mainly β and α/β domains, with no significantly populated intermediate states along the folding pathway. This lack of an intermediate is consistent with experimental observations that unfolding of serpin species with cleaved, inserted RCLs appears to be two-state (90).

This simultaneous folding of both domains observed for folding to the latent state resulted in a large concerted loss of conformational entropy. This entropy loss leads to a folding free energy barrier that is higher than that seen for the stepwise folding pathway that leads to the active, metastable structure (Fig. 5). Based on these results, it is suggested that if RCL-strand 5A contacts form early during folding, they will subsequently break and allow folding to proceed along the lower free energy barrier pathway to the active metastable structure.

This finding, that folding to the active, metastable state results from entropic contributions to folding barriers is experimentally testable. Giri Rao and Gosavi (30) suggested incorporating a disulfide bond between the RCL and strand 5A. Under oxidizing conditions with an intact disulfide bond, this AAT variant should fold directly to the latent state with slower kinetics than would be observed for folding of the reduced AAT variant to the active, metastable state.

These results demonstrate how simple, structure-based models can be used to address folding conundrums even for large proteins with complicated topologies, such as serpins.
Investigating the effects of mutations on serpin folding using enhanced sampling based on biased all-atom simulations

Because all-atom biased MD simulations retain the chemistry of the side chains they allow for the investigation of how sequence specific factors (e.g. mutations) affect folding. These investigations are particularly important for proteins such as serpins, where mutations can lead to disease-associated misfolding and aggregation (41, 84). Therefore, Wang et al. (31) used the BF method to examine how known pathological mutations associated with human disease perturb AAT folding.

The most common AAT mutation linked to severe disease is the Z mutation, E342K, which converts a Glu-Lys salt bridge at the base the RCL to a repulsive Lys-Lys interaction. In vitro, folding Z populates an aggregation-prone intermediate that can persist for hours (92), and Z unfolds from the native to the intermediate state significantly faster than does WT (93). In cells, this folding defect results in severe misfolding, degradation of nascent Z chains, and the formation and accumulation of insoluble polymers in the endoplasmic reticulum (ER) (94). The resulting low level of circulating AAT leads to lung disease, whereas accumulation of polymers in hepatocytes can lead to cell death and liver disease (41). The structure(s) of the partially folded Z AAT species that mediate polymer formation is therefore of considerable medical interest, but its high aggregation propensity has hindered structural studies.

The serpin fold consists of two nonsequential domains with extensive interdomain and nonlocal contacts (Fig. 4). Perhaps unsurprisingly, in all-atom BF folding simulations for WT, Z, and other AAT variants, folding began with the independent formation of local, sequential structural units (Fig. 6). For WT AAT, subsequent successful folding involved at least two pathways in which these structural units dock to each other in a defined order. In the major pathway, strands 4 and 5B at the C terminus and the α helices at the N terminus docked last. The N-terminal helical region is highly frustrated (68). Thus, the finding that the last step in folding is docking of the N-terminal α helices highlights the difficulties in folding frustrated regions. This result, that the C-terminal β strands 4 and 5B are incorporated into the AAT structure prior to incorporation of the N-terminal α helices in the major folding pathway, is a novel, experimentally testable prediction of the BF simulations.

In BF simulations, folding of the pathological Z mutant diverged from WT folding relatively early despite the use of the same all-atom force field (Amber99) and ratchet-and-pawl native-centric biasing force. In other words, the E342K mutation was sufficient to drive the variant to a nonnative structure.

In WT AAT folding simulations, interactions were formed between β strands 5 and 6A in the α/β domain and sheets B and C in the mainly β domain. This interdomain association occurred early in the folding as the local structural elements began to dock with each other, and once formed, these interdomain interactions were preserved for the remainder of the WT folding pathway. In Z folding simulations, this association failed to occur, and as a result, the majority of simulations led to final structures in which β sheets in both domains failed to form correctly.

In cellular studies, a number of AAT mutations have been made to rescue or better understand Z misfolding (95, 96). Folding simulations for these and other variants suggest that
unfavorable electrostatic interactions and steric clashes both play a role in Z misfolding and that there are a number of different ways to rescue Z misfolding.

Experimental studies of the kinetics of AAT folding have mainly focused on WT AAT (32–34) with limited data on the kinetics of Z unfolding (92, 93). The results of the BF simulations are in good agreement with data from kinetic and equilibrium AAT folding studies. Importantly, these simulations make new, testable predictions on how AAT variants fold and misfold (e.g., structural predictions for likely long-lived intermediates in the Z misfolding/folding pathways) and how Z misfolding might be rescued.

Comparing all-atom biased simulations with Gō-model predictions

An important issue to address is how the results of BF protein-folding simulations, performed using all-atom force fields, compare with the results of MD simulations based on simplified native centric (Gō) force fields. For small globular proteins, typically characterized by native structure with a simple native topology, both the BF and Gō-model approximation schemes lead to remarkably similar folding mechanisms (73). On the other hand, some discrepancies seem to emerge for medium-sized proteins (e.g., consisting of more than 100 amino acids) and for proteins with nontrivial native topology (65, 75).

A prototypical case of disagreement is the folding of the two evolutionarily related bacterial colicin immunity proteins, Im7 and Im9. These chains have nearly identical α-helical native structures; thus, Gō-type models cannot distinguish between them and predict identical folding kinetics (66). However, a number of kinetic folding experiments have shown that at neutral pH, Im7 populates a folding intermediate and shows three-state kinetics, whereas Im9 shows two-state folding kinetics (97–99). This difference is due to transient nonnative interactions that stabilize the folding intermediate in Im7 and can be resolved in Gō-type models augmented with sequence-dependent nonnative hydrophobic interactions (66). However, these Gō-type simulations required several model iterations before the correct model was arrived at. In contrast, the Im7 folding intermediate was correctly observed in BF simulations, based on the CHARMM36 force field in explicit water without any modifications to the BF method (75).

Differences between Gō-type and biased dynamics simulations were also observed in the folding of the smallest known polypeptide chain that folds into a topologically knotted native structure (65). However, remarkable agreement between these two approaches was recovered after the effective potential in the Gō model was modified to include nonnative interactions that implicitly account for the hydrophobic/hydrophilic property of the amino acids (100). These results for both the knotted protein and Im7 emphasize that, as stated previously (see “Encoding frustration in Gō models”), inclusion of nonnative interactions is sometimes required to correctly simulate folding using Gō-type models (Fig. 2).

In the case of serpin folding, the Gō model (30) and the BF approach (31) are complementary and show considerable agreement. For WT AAT, both sets of simulations found that the mainly β domain folds early and that there is a highly populated intermediate state in which the β domain is mostly folded, whereas the α/β domain is still largely unformed (Figs. 5 and 6). However, even at this stage in the all-atom BF simulations, significant amounts of local secondary structure were formed in the α/β domain.

WT AAT is at least a three-state folder (32–34, 91), and this is reflected in both sets of simulations. In the Gō-model simulations, a highly populated intermediate is formed when the fraction of native contacts, Q, is ~0.4 (Fig. 5), and in the BF simulations, similar intermediates are populated between Q values of 0.5 and 0.7 (local minima 1 and 2 in Fig. 6). The somewhat larger fraction of native contacts in the BF simulations may be because folding was carried out at room temperature, whereas the Gō-model simulations were carried out at a higher temperature, T\text{fold}, the temperature at which the native and unfolded states are equally populated. In addition, because the native-centric BF approach discourages backtracking, structural elements may be overstabilized. The approach to this intermediate state is different for the two simulations; in the BF simulations, nonnative contacts between sheets B and C in the β domain and a strand from sheet A in the α/β domain must be resolved before these intermediates can form. These stabilizing nonnative interactions are not present in the Gō model.

As noted above, positioning of the highly mobile RCL is critical for correct serpin folding to the active, metastable state (Fig. 4). Although the timing is slightly different, in both sets of simulations, premature insertion of the RCL into sheet A is prevented by the formation of contacts between the RCL and sheets B and C locking the C-terminal end of the RCL into place before the β domain has completed folding in agreement with the results of kinetic folding experiments (34). Interestingly, based on Frustratometer (68, 101) calculations of frustration, the C-terminal end of the RCL shows a lower degree of frustration than does the N-terminal end, again emphasizing that frustrated regions can be more difficult to fold. Both Gō-model and BF simulations also agree that packing of the C-terminal strands 4 and 5B, which complete the formation of the β domain, is a late event in folding, a finding that is supported by both pulsed oxidative footprinting (34) and fragment complementation studies (102).

Thus, as demonstrated by investigations of AAT folding, Gō-type and all-atom approaches can provide complementary insights into protein folding, and method selection will depend on the question being asked.

Simulating the folding of other large proteins

In addition to AAT, folding simulations have been performed for a number of large multidomain proteins using both SBMs and all-atom methods (24–29, 103). The domain architecture in multidomain proteins may be classified as sequential, proteins in which the domains are translated sequentially from the N to the C terminus, and nonsequential or discontinuous, proteins such as serpins in which one or more domain is discontinuously translated. Often, the folding of proteins with sequential domains is hypothesized to also be sequential, but more complicated folding has also been observed (104–106). A special class of sequential multidomain proteins are the repeat
proteins, and the folding and misfolding of this class of proteins has recently been reviewed elsewhere (107, 108).

Another promising native-centric approach to study the folding and misfolding of multidomain proteins is AWSEM-MD (associative memory, water-mediated, structure and energy model) developed by Wolynes and colleagues (16, 109). AWSEM is a coarse-grained protein force field in which a native-centric bias can be introduced to varying degrees, ranging from a full Gō-type potential based on the native structure to a potential with local conformational preferences derived from fragment libraries (16). Zheng et al. (109) used AWSEM-MD to study the folding of fused dimers of either SH3 domains or Ig domains from human titin. Prior to these simulations, single-molecule studies on Ig domain folding had shown that the monomers could domain-swap (a process in which adjacent monomers form intramolecular contacts that mimic the corresponding intramolecular native contacts in the monomer) and that monomers with identical sequences were more likely to domain-swap than monomers with divergent sequences (110). Domain swapping of monomers with identical sequences was also observed in force unfolding experiments and simulations of polyubiquitin chains (111).

The AWSEM-MD simulations performed by Zheng and et al. (109) suggested that misfolding of repeat proteins could occur in multiple ways, including the experimentally observed domain swapping and an alternate mechanism involving the formation of short stretches of intermolecular β-strands that formed amyloid-like contacts. Subsequently, the amyloid-like misfolding mechanism was supported by both ensemble and single-molecule kinetic folding experiments (112). Unlike domain-swapped contacts, these amyloid-like contacts had no counterpart in the native structure. More generally, Zheng et al. (109) found that the frequency of misfolding through both mechanisms was reduced by lowering the sequence identity between monomers in the fused dimer, a finding that supports the hypothesis that evolution has disfavored high sequence identity between neighboring domains in repeat proteins as a means of minimizing misfolding (64). These findings are important not just for repeat proteins but also for proteins such as serpins where misfolding of some disease-associated mutants is proposed to result in domain-swapped oligomers (113, 114) as well as for amyloidogenic proteins.

Unlike repeat proteins, many sequential multidomain proteins are made of diverse structural units, decreasing the probability of misfolding by domain swapping. Jin Wang and colleagues (29) used SBMs to simulate the unfolding and folding of DNA polymerase IV from Sulfolobus solfataricus (DPO4), a 342-amino-acid-long protein. With the exception of the N-terminal eight amino acids, which form a strand in the palm domain, DPO4 is a four-domain sequential protein with the finger, palm, thumb, and little finger domains in order from the N to the C terminus (Fig. 7A). In the SBM simulations, DPO4 folded via six parallel pathways, a “divide and conquer” strategy where each domain could fold independently and some, but not all, domain interfaces helped template the folding of other domains. Wang and co-workers suggest that this divide and conquer strategy can lead to multiple folding intermediates but that the formation of each intermediate decreases the degrees
of freedom, thereby speeding the folding of multidomain proteins.

Simulations of the folding of the protein dihydrofolate reductase (DHFR) and a circular permutant by Inanami et al. (28) provide further insight into the divide and conquer folding strategy. The 159-amino acid DHFR has two discontinuous domains, a discontinuous loop domain (DLD) (residues 1–37 and 107–159) and a continuous adenosine-binding domain (ABD) (residues 38–106) (Fig. 7B), whereas the circular permutant resolves the discontinuities resulting in two sequential domains. To simulate the folding of DHFR and other proteins with discontinuous domains, Inanami and co-workers had to modify the Wako-Saito-Muñoz-Eaton (WSME) method (119–121), a kind of SBM. They then simulated the folding of both WT DHFR and the circular permutant using their extended WSME approach.

In the simulations, folding occurred through two major intermediates, one in which the continuous ABD domain was mostly formed and the second in which the DLD domain was mostly formed. In the WT protein, formation of the DLD was entropically unfavorable due to the loop closure penalty that arises when the N and C termini are brought close together. Thus, the DLD first pathway was disfavored, and the folding pathway in which the ABD forms before the DLD dominated. In the circular permutant, the entropic penalty for forming the DLD was much reduced, and the two pathways (ABD first or DLD first) had similar probabilities analogous to the divide and conquer folding observed for DPO4. The DLD loop closure penalty is similar to the barrier for folding to the serpin divide and conquer folding observed for DPO4. The DLD loop closure penalty is similar to the barrier for folding to the serpin.

The 159-amino acid DHFR has two discontinuous domains, a discontinuous loop domain (DLD) (residues 1–37 and 107–159) and a continuous adenosine-binding domain (ABD) (residues 38–106) (Fig. 7B), whereas the circular permutant resolves the discontinuities resulting in two sequential domains. To simulate the folding of DHFR and other proteins with discontinuous domains, Inanami and co-workers had to modify the Wako-Saito-Muñoz-Eaton (WSME) method (119–121), a kind of SBM. They then simulated the folding of both WT DHFR and the circular permutant using their extended WSME approach.

In the simulations, folding occurred through two major intermediates, one in which the continuous ABD domain was mostly formed and the second in which the DLD domain was mostly formed. In the WT protein, formation of the DLD was entropically unfavorable due to the loop closure penalty that arises when the N and C termini are brought close together. Thus, the DLD first pathway was disfavored, and the folding pathway in which the ABD forms before the DLD dominated. In the circular permutant, the entropic penalty for forming the DLD was much reduced, and the two pathways (ABD first or DLD first) had similar probabilities analogous to the divide and conquer folding observed for DPO4. The DLD loop closure penalty is similar to the barrier for folding to the serpin latent state (30), and both of these studies reiterate the importance of entropic folding barriers.

Escherichia coli adenylate kinase (AKE) is a complex nonsequential protein composed of two continuous domains (LID and NME) inserted into a discontinuous domain (CORE) split into three parts (Fig. 7C). SBM simulations of AKE found, in agreement with bulk thermodynamic experiments, that AKE folds cooperatively (25). This cooperativity was enabled by the insertion of the smaller less stable domains into the larger more stable domain. The smaller domains transiently folded and unfolded, stabilizing only upon folding of the larger discontinuous domain, whose folding constrained the termini of the smaller domains into near-native conformations and reduced the entropic cost of folding these domains. Similar entropic considerations go into the physics-based sequential collapse model in which the closure of loops with lengths on the order of 65 amino acids, too large to crowd the side chains and too small to have very large entropic penalties, are postulated to be one of the earliest events in protein folding (122). This nonlocal loop formation would encourage the formation of local structure, such as foldons (123, 124). Application of the sequential collapse-based method to AKE agreed with the results of time-resolved FRET experiments (122) and the previous SBM folding simulations.

Results of the SBM simulations suggested that for AKE, the number of intermediates and in turn folding pathways could be modulated by increasing the stabilities of the inserted domains and making their folding independent of the discontinuous domain (25). In separate SBM simulations, the stabilities of the inserted domains were tuned by introducing energetic heterogeneity in the strengths of the contacts stabilizing the smaller domains (24). The results of these simulations, specifically both the number of folding intermediates and the number of folding pathways, broadly matched with experiments. Further tuning of domain stabilization and folding pathways was achieved by mutating AKE residues in or near the NME domain (125). Upon experimentally testing the effects of these mutations, it was found that stabilizing the NME domain mainly preserved WT-like folding while shunting a minor fraction (9%) of the folding flux through a pathway that bypassed folding intermediates and folded directly to the native state. Although the effects of the specific mutations used in the experiments on folding have not been tested using SBMs, the experimental results only partially agree with predictions from the SBM simulations with homogeneous contact strengths. This partial agreement between experimental and computational folding results provides clues on how to improve simulation approaches. For instance, heterogeneous contact strengths may be necessary to predict the specifics of intermediate populations in AKE folding. As shown for other proteins, such as Im7 (66) and the knotted protein (100) mentioned earlier, SBMs may need to be supplemented with nonnative interactions (see Fig. 2) to improve agreement with experiment.

In contrast, for the 443-amino-acid-long E. coli cell division protein SufI (also called ftsP), the ability to access a number of parallel folding pathways leads to misfolding (103). SufI is composed of three sequential domains (Fig. 7D), and Ignatova and colleagues (126) showed experimentally that stretches of slow translation are required for proper SufI folding. Tanaka et al. (103) used their SBM to simulate the folding of full-length SufI and mimicked co- translational folding by vectorially synthesizing SufI at various rates during the simulations. For the full-length protein and for fast vectorial synthesis, SufI folded via multiple pathways, many of which resulted in long-lived misfolded species. In contrast, vectorial synthesis regimes that allowed for folding during synthesis restricted the number of possible pathways and increased the efficiency of folding. These results are consistent with recent work showing that rare codon regions, which may attenuate translation, are conserved across species (127, 128). In general, the balance between translation rates and folding is protein-dependent. For proteins such as SufI, translational attenuation at particular places in the sequence can aid in folding, whereas for other proteins, fast translation can reduce the probability of populating misfolded intermediates (129, 130).

In summary, when larger proteins consist of multiple small domains with minimal interdomain interfaces, the folding of individual domains may be only partially dependent on or completely independent of the folding of other domains, and this can lead to multiple pathways. Interdomain interfaces can template the folding of neighboring domains and reduce the number of pathways that are available to the protein. Evolution may further modulate the folding of such sequential multidomain proteins by having divergent amino acid sequences in adjacent domains or by tuning translation rates. However, on the whole, conclusions drawn from single-domain proteins can be applicable to the folding of the individual domains of such modular proteins.
multidomain proteins. In contrast, in multidomain proteins, where either the domains are connected through multiple linkers or there is an extensive interface between them, protein folding depends on the structure of the entire protein. In such proteins, topological frustration created by the need to fold a complex structure may stall folding and reduce free energy barriers but enable the population of folding intermediates. Alternatively, the entropy loss required to fold the entire protein in an all or nothing fashion can lead to large folding energy barriers.

**Harnessing the predictive power of folding simulations**

Existing experimental folding data are often used to validate the results of folding simulations; as shown for AKE, where simulations motivated the design of AKE mutants and single-molecule FRET (smFRET) studies of their folding (125), the conformational ensembles along the simulated folding trajectories can be used to design new kinetic folding experiments and to help interpret experimental results. As has been demonstrated for near-UV CD spectra calculated from BF folding simulations (76), these conformational ensembles may be used to predict how average protein properties change during folding. However, different conformational ensembles can lead to similar average properties, and the real power of the conformational ensembles available from folding simulations lies in comparisons with and prediction of results from methods that can elucidate how conformational distributions change as a protein folds.

Many methods, including time-resolved fluorescence, lifetime FRET, smFRET, \(^{13}F\)NMR, and EPR, use site-specific labels that report on the local environment around the label or changes in the relative distance between labels. For these local probes, simulations may be used to make informed decisions as to where to locate these local reporters to maximize signal changes as the protein folds and to predict the experimental results. Simulations may also be helpful in interpreting observed signals. For example, for smFRET, all-atom conventional MD simulations performed on the Anton supercomputer have been used to explain observed folding and unfolding rate constants (131). More recently, Schug and co-workers (132) used coarse-grained SBMs that explicitly included the FRET dyes to simulate smFRET folding data, resulting in good agreement with experimental results.

Although label-dependent methods can provide information on how the local environment of the probe changes as the protein folds, other methods, including NMR, pulsed oxidative labeling coupled to MS (133, 134), and hydrogen-deuterium exchange MS (HDX-MS) or NMR (135), provide conformationally sensitive data with peptide to single-residue resolution. Particularly for NMR experiments, experimental data may be incorporated into simulations to help provide a molecular interpretation of the experimental results (e.g., see Ref. 136). In addition, conformational distributions from folding simulations may be used to predict time-resolved experimental results. For example, oxidative labeling of amino acid side chains depends on side-chain solvent accessibility (133, 134), and the time evolution of the probability that a given residue is solvent-accessible may be simply computed from conformational ensembles along the folding trajectory.

For folded proteins, deuterium uptake curves from HDX-MS experiments have been predicted with reasonable accuracy from all-atom or coarse-grained simulations even on relatively short submicrosecond timescales (137, 138). Similarly, conformational ensembles along folding pathways, such as those shown in Figs. 5 and 6 for AAT, could be used to predict the results of kinetic folding experiments monitored by HDX-MS. However, conformational ensembles from native-centric simulations may not capture structural fluctuations, possibly leading to underestimation of exchange. An alternative approach would be to better account for structural fluctuations by subjecting one or more conformations from the major long-lived intermediates identified in the simulations to 0.1–1-μs MD simulations. Deuterium uptake at short pulse-labeling timescales (typically 5–10 s in the case of manual labeling) could then be predicted from ensembles harvested directly from the folding simulations or from the MD-generated ensembles, allowing direct comparisons between predicted and observed folding intermediates.

In addition to predicting experimental results, the simulations may be used to test the effects of mutations as demonstrated for both AAT (31) and AKE (25). Furthermore, as has been demonstrated for AKE (125), the simulations may be used as the basis for designing and experimentally characterizing new mutations.

**Extrapolating from folding simulations to folding in more complicated environments**

As discussed above, the results of folding simulations are generally compared with experimental results from the folding of purified, full-length proteins. But, physiologically, proteins are synthesized and fold in the crowded and complicated intracellular milieu (139–141). In cells, proteins may fold co-translationally as they are synthesized and both co- and post-translationally proteins interact with the protein homeostasis machinery, including modifying enzymes (e.g., kinases and oligosaccharyltransferases) and molecular chaperones. Although folding simulations are performed in a much simpler environment, the insights generated into how the protein sequence folds are still relevant to in-cell folding.

Folding pathways and intermediates generated using the computational methods described above may allow the formulation of specific hypotheses regarding which structural regions and motifs of a protein interact with molecular chaperones and other components of the protein homeostasis machinery. Because folding simulations are usually performed on full-length proteins, we focus on what is known about post-translational interactions between folding proteins and cellular quality control. (For more comprehensive reviews of the protein homeostasis machinery, see Ref. 7.) To our knowledge, there have not been rigorous attempts to extrapolate substrate-chaperone interactions from folding simulations; nonetheless, a general strategy is summarized below.

The first step is to use protein-folding simulations to predict and structurally characterize the conformational ensembles of the relevant long-lived intermediates. Next, the solvent accessibility of regions that are likely to bind molecular chaperones...
or other components of the protein homeostasis machinery can be determined.

For the bacterial Hsp70 molecular chaperones DnaK and BiP, the Hsp70 resident in the ER, seven amino acid sequence motifs that are likely to bind in the Hsp70 substrate-binding site have been determined using peptide arrays combined with Hsp70 structures and computational methods (142, 143). Using the available webservers (Limbo (switchlab.org/bioinformatics/limbo) for DnaK and BipPred (https://www.ncbi.nlm.nih.gov/pubmed/27287023) for BiP), likely Hsp70 binding sites can be determined from the protein sequence. Then the relative solvent accessibility of these regions could be determined for the conformational ensembles of folding intermediates from simulations. Similarly, recent work (144) suggests that binding motifs for the ER-resident Hsp40s Erdj4 and Erdj5 as well as the mammalian ER-resident Hsp110 ortholog Grp170 may be identified using the TANGO algorithm (tango.crg.es) (145) to develop high aggregation propensity for proteins that fold and mature in the ER, these sequence motifs may also be used to try to identify likely binding sites for molecular chaperones. Other classes of molecular chaperones, including the chaperonin GroEL (146), may preferentially recognize dynamic, frustrated protein regions (147). Frustrated regions in folding simulations may be identified using the Frustratometer (101, 148), again allowing the identification of possible accessible recognition sites in the conformational ensembles of likely long-lived folding intermediates. Although these proposed methods are relatively primitive and not applicable to all chaperones (149), they have the potential to provide structural information on transient folding intermediates that may be recognized by chaperone networks.

These predictions may also be experimentally tested either in vitro for folding experiments performed in the presence of molecular chaperones or in cells. Interactions between the chaperones and folding proteins may be captured using cross-linking, and the location of the interactions may be mapped using MS (150). This combination of simulations and experiments has the potential to aid in understanding how chaperones and other components of the protein homeostasis machinery recognize and interact with folding and misfolding client proteins.

Another area where native centric folding simulations can find application is in understanding the basis of diseases caused by protein misfolding. In misfolding diseases, such as AAT deficiency, ALS, and Alzheimer’s disease, toxic gain-of-function phenotypes are correlated with aggregate accumulation, and oligomerization and aggregation are believed to proceed from specific partially folded or misfolded states (151). Such oligomerization-prone states could potentially be targeted by small molecules. High-throughput screening is one strategy for identifying such compounds, but for a rational design strategy to be pursued, knowledge of the structure of these pathological protein states, ideally at the level of atomic detail, is needed. In the AAT folding simulations described above, it was encouraging to see that the propensity of disease-associated mutants to misfold correlated with observed disease severity in patients. Similar correlations between MD trajectories, mutations, and disease severity have been observed for superoxide dismutase 1 and ALS (152). These results suggest that the misfolded states generated computationally may adequately represent the pathological misfolded states formed in cells. Although this hypothesis requires further experimental validation, if correct, then these computationally generated misfolded states could serve as targets for in silico drug design. We suggest that such a combination of folding simulations, in silico small molecule screening/design, and experiments could serve as a general approach for identifying potential therapies for diseases linked to protein misfolding.

Conclusions and future directions

Native centric simulation methods have made it possible to extend the computation of detailed folding pathways and intermediates beyond the small fast-folding model proteins that have traditionally been favored by protein-folding researchers. Collectively, comparisons between studies suggest that the results of all-atom approaches, such as BF/SPCS and native-centric structure-based (Gō-type) model (SBM) folding simulations share many common features, likely reflecting the importance of native contacts. SBMs are particularly helpful when interrogating how the same or similar chains fold to different structures, as demonstrated by simulating serpin folding to the native and latent conformations (30). BF/SPCS and other all-atom methods can address similar questions while extending the reach of simulations to questions of chemistry.

For example, these all-atom methods can be used to predict or explain how differences in the primary structure alter folding pathways (75). Additionally, important differences between all-atom and SBM approaches are likely to arise for reactivity in which transient nonnative interactions play an important role in restricting the exploration of the configuration space. Thus, the choice of method is likely to depend on computational resources and efficiencies as well as the question being asked.

All-atom biased simulations have been successfully applied to elucidate the mechanism through which a specific small molecule accelerates the PAI-1 serpin latency transition (153) and to predict how point mutations alter the misfolding propensity (31). Future applications are likely to include explicitly simulating interactions between proteins and small molecules as well as simulating the formation of protein oligomers during and after folding.

SBMs have already been used to simulate the complex conformational dynamics of ribosomes (154), the membrane insertion of proteins (155), and co-translational folding (103). Additionally, in the absence of a structure, protein or DNA sequence derived information and experimental constraints have been used in conjunction with coarse-grained and structure-based models to understand protein conformations and assembly (52, 156, 157), chromosome folding (158), etc. With increasing computational power, and the accumulation of biological data, we expect that both structure- and data-derived models of larger biological machines will be used to simulate and understand folding, assembly, and dynamics in cells and other physiologically relevant environments (159–161). We expect that such simulations will aid in the analysis of biological experiments at diverse length scales. We also hope that such detailed computational models will stimulate new and interesting experiments.

The increasing availability of affordable graphics-processing units capable of running MD codes brings both native-centric simulations and simulations based on a native biasing force.
within the reach of many research groups at reasonable cost. (Also, freely distributed software packages such as SMOG (and the associated web server) (162), CafeMol (163), Go-Kit (164), and eSBMTools (165) are available to facilitate the use of Gö-type simulations by nonexperts.) We hope this review will convince both experimentalists and MD simulation experts that native-biased simulations have come of age, can bridge the timescale gap between experiments and simulations, and are just as easily accessible to everyone as atomistic MD simulations. Importantly, these methods can aid in the interpretation of experiments and generate novel testable hypotheses.

There are insights to be gained from both types of native-biased simulations in the folding of large proteins, and such insights may be hard to come by in experiments. Altogether, these computational platforms can provide useful tools for translational research aiming to identify new therapeutic strategies. Much remains to be done to expand the application of native structure-based techniques at both the larger and the atomistic length scales, and we hope this review will inspire further developments in the models, methods, and applications of these promising simulation techniques.

Acknowledgments—We thank current and former members of the Faccioli, Gershenson, Gosavi, and Wintrobe research groups for their research on folding large proteins and helpful discussions.

References

1. Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997) Theory of protein folding: the energy landscape perspective. Annu. Rev. Phys. Chem. 48, 545–600 CrossRefMedline

2. Dill, K. A., and MacCallum, J. L. (2012) The protein-folding problem, 50 years on. Science 338, 1042–1046 CrossRefMedline

3. Gruebele, M., Dave, K., and Sukhenk, S. (2016) Globular protein folding in vitro and in vivo. Annu. Rev. Biophys. 45, 233–251 CrossRefMedline

4. Hebert, D. N., and Molinari, M. (2007) In and out of the ER: protein folding, quality control, degradation, and related human diseases. Physiol. Rev. 87, 1377–1408 CrossRefMedline

5. Hartl, F. U., and Hayer-Hartl, M. (2009) Converging concepts of protein folding in vitro and in vivo. Nat. Struct. Mol. Biol. 16, 574–581 CrossRefMedline

6. Labbadia, J., and Morimoto, R. I. (2015) The biology of proteostasis in aging and disease. Annu. Rev. Biochem. 84, 435–464 CrossRefMedline

7. Dubnikov, T., Ben-Gedalya, T., and Cohen, E. (2017) Protein quality control in health and disease. Cold Spring Harb. Perspect. Biol. 9, a023523 CrossRefMedline

8. Jang, Y., and Kalodimos, C. G. (2017) NMR studies of large proteins. J. Mol. Biol. 429, 2667–2676 CrossRefMedline

9. Sekhar, A., and Kay, L. E. (2019) An NMR view of protein dynamics in health and disease. Annu. Rev. Biophys. 48, 297–319 CrossRefMedline

10. Hekstra, D. R., White, K. I., Socolich, M. A., Henning, R. W., Šrajer, V., and Ranganathan, R. (2016) Electric-field-stimulated protein mechanics. Nature 540, 400–405 CrossRefMedline

11. Piana, S., Klepeis, J. L., and Shaw, D. E. (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr. Opin. Struct. Biol. 24, 98–105 CrossRefMedline

12. Pande, V. S. (2014) Understanding protein folding using Markov state models. Adv. Exp. Med. Biol. 797, 101–106 CrossRefMedline

13. Wang, M., Kurland, C. G., and Caetano-Anollés, G. (2011) Reductionive evolution of proteomes and protein structures. Proc. Natl. Acad. Sci. U.S.A. 108, 11954–11958 CrossRefMedline

14. Gó, N. (1976) Statistical mechanics of protein folding, unfolding and fluctuation. Adv. Biophys. 9, 65–113 CrossRefMedline

15. Karanicoias, J., and Brooks, C. L., 3rd (2003) Improved Gö-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. J. Mol. Biol. 334, 309–325 CrossRefMedline

16. Papan, G. A., and Wolynes, P. G. (2017) AWSEM-MD: from neural networks to protein structure prediction and functional dynamics of complex biomolecular assemblies. Coarse-grained Modeling of Biomolecules (Papoian, G. A., ed) pp. 121–190, CRC Press, Inc., Boca Raton, FL

17. Bolhuis, P. G., Chandler, D., Delgado, C., and Geissler, P. L. (2002) Transition path sampling: throwing ropes over rough mountain passes, in the dark. Annu. Rev. Phys. Chem. 53, 291–318 CrossRefMedline

18. Barducci, A., Bonomi, M., and Pannirivel, M. (2011) Metadynamics. Wiley Interdiscip. Rev. Comput. Mol. Sci. 1, 826–843 CrossRefMedline

19. Belo-Rivas, J. M., and Elber, R. (2016) Simulations of thermodynamics and kinetics on rough energy landscapes with milestoneing. J. Comput. Chem. 37, 602–613 CrossRefMedline

20. Elber, R. (2017) A new paradigm for atomically detailed simulations of kinetics in biophysical systems. Q. Rev. Biophys. 50, e8 CrossRefMedline

21. Cabriolu, R., Skjelbred Refnes, K. M., Bolhuis, P. G., and van Erp, T. S. (2017) Foundations and latest advances in replica exchange transition interface sampling. J. Chem. Phys. 147, 152722 CrossRefMedline

22. Husic, B. E., and Pande, V. S. (2018) Markov state models: From an art to a science. J. Am. Chem. Soc. 140, 2386–2396 CrossRefMedline

23. Orioli, S., a Beccara, S., and Faccioli, P. (2017) Self-consistent calculation of protein folding pathways. J. Chem. Phys. 147, 064108 CrossRefMedline

24. Li, W., Terakawa, T., Wang, W., and Takada, S. (2012) Energy landscape and multiroute folding of topologically complex proteins adenylate kinase and 2ouf-knot. Proc. Natl. Acad. Sci. U.S.A. 109, 17789–17794 CrossRefMedline

25. Giri Rao, V. V. H., and Gosavi, S. (2014) In the multi-domain protein adenylate kinase, domain insertion facilitates cooperative folding while accommodating function at domain interfaces. PLoS Comput. Biol. 10, e1003938 CrossRefMedline

26. Reddy, G., Liu, Z., and Thrurunali, D. (2012) Denaturation-dependent folding of GFP. Proc. Natl. Acad. Sci. U.S.A. 109, 17832–17838 CrossRefMedline

27. Halloran, K. T., Wang, Y., Arora, K., Chakravarthi, S., Irving, T. C., Bilsel, O., Brooks, C. L., 3rd, and Matthews, C. R. (2019) Frustration and folding of a TIM barrel protein. Proc. Natl. Acad. Sci. U.S.A. 116, 16378–16383 CrossRefMedline

28. Inanami, T., Terada, T. P., and Sasi, M. (2014) Folding pathway of a multidomain protein depends on its topology of domain connectivity. Proc. Natl. Acad. Sci. U.S.A. 111, 15969–15974 CrossRefMedline

29. Wang, Y., Chu, X., Suo, Z., Wang, E., and Wang, J. (2012) Multidomain protein solves the folding problem by multifunnel combined landscape: theoretical investigation of a Y-family DNA polymerase. J. Am. Chem. Soc. 134, 13755–13764 CrossRefMedline

30. Giri Rao, V. V. H., and Gosavi, S. (2018) On the folding of a structurally complex protein to its metastable active state. Proc. Natl. Acad. Sci. U.S.A. 115, 1998–2003 CrossRefMedline

31. Wang, F., Orioli, S., Ianeselli, A., Spagnoli, G., a Beccara, S., Gershenson, A., Faccioli, P., and Wintrobe, P. L. (2018) All-atom simulations reveal how single-point mutations promote serpin misfolding. BioJ. 114, 2083–2094 CrossRefMedline

32. Kim, D., and Yu, M. H. (1996) Folding pathway of human 1-antitrypsin. J. Biol. Chem. 271, 15969–15974 CrossRefMedline

33. Tsutsui, Y., Dela Cruz, R., and Wintrobe, P. L. (2012) Folding mechanism of the metastable serpin 1-antitrypsin. Proc. Natl. Acad. Sci. U.S.A. 109, 4467–4472 CrossRefMedline

34. Stocks, B. B., Sarkar, A., Wintrobe, P. L., and Konermann, L. (2012) Early hydrophobic collapse of α1-antitrypsin facilitates formation of a metastable state: insights from oxidative labeling and mass spectrometry. J. Mol. Biol. 423, 789–799 CrossRefMedline

35. Honeycutt, J. D., and Thirumalai, D. (1990) Metastability of the folded states of globular proteins. Proc. Natl. Acad. Sci. U.S.A. 87, 3526–3529 CrossRefMedline
36. Mottonen, J., Strand, A., Symersky, J., Sweet, R. M., Danley, D. E., Geoghegan, K. F., Gerard, R. D., and Goldsmith, E. J. (1992) Structural basis of latency in plasminogen activator inhibitor-1. *Nature* **355**, 270–273 CrossRef Medline

37. Lawrence, D. A., Olson, S. T., Palaniappan, S., and Ginsburg, D. (1994) Engineering plasminogen activator inhibitor 1 mutants with increased functional stability. *Biochemistry* **33**, 3643–3648 CrossRef Medline

38. Auer, S., Miller, A. M., Karvov, S. V., Dobson, C. M., Karplus, M., and Vendruscolo, M. (2007) Importance of metastable states in the free energy landscapes of polypeptide chains. *Phys. Rev. Lett.* **99**, 178104 CrossRef Medline

39. Im, H., Woo, M.-S., Hwang, K. Y., and Yu, M.-H. (2002) Interactions causing the kinetic trap in serpin protein folding. *J. Biol. Chem.* **277**, 46347–46354 CrossRef Medline

40. Stein, P. E., and Carrell, R. W. (1995) What do dysfunctional serpins tell us about molecular mobility and disease? *Nat. Struct. Biol.* **2**, 96–113 CrossRef Medline

41. Greene, C. M., Marciniak, S. J., Teckman, J., Ferrarotti, I., Brantly, M. L., Lomas, D. A., Stoller, J. K., and McElvenay, N. G. (2016) αl-Antitrypsin deficiency. *Nat. Rev. Dis. Primers* **2**, 16051 CrossRef Medline

42. Bryngelson, J. D., and Wolynes, P. G. (1987) Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. U.S.A.* **84**, 7524–7528 CrossRef Medline

43. Shakhnovich, E. I., and Gutin, A. M. (1989) Formation of unique structural motifs for protein folding and dynamics. *Int. J. Mol. Sci.* **167–195** CrossRef Medline

44. Taketomi, H., Ueda, Y., and Gō, N. (1975) Studies on protein folding, unfolding and fluctuations by computer simulation. I. The effect of specific amino acid sequence represented by specific inter-unit interactions. *Int. J. Pept. Protein Res.* **7**, 1–12 CrossRef Medline

45. Camacho, C. J., and Thirumalai, D. (1995) Modeling the role of disulfide bonds in protein folding: entropic barriers and pathways. *Proteins* **22**, 27–40 CrossRef Medline

46. Klimov, D. K., and Thirumalai, D. (2001) Multiple protein folding nuclei and the transition state ensemble in two-state proteins. *Proteins* **43**, 465–475 CrossRef Medline

47. Gin, B. C., Garahan, J. P., and Geissler, P. L. (2009) The limited role of nonnative contacts in the folding pathways of a lattice protein. *J. Mol. Biol.* **392**, 1303–1314 CrossRef Medline

48. Best, R. B., Hummer, G., and Eaton, W. A. (2013) Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17874–17879 CrossRef Medline

49. Bryngelson, J. D., Onuchic, J. N., Socci, N. D., and Wolynes, P. G. (1995) Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins* **21**, 167–195 CrossRef Medline

50. Wolynes, P. G., Onuchic, J. N., and Thirumalai, D. (1995) Navigating the folding routes. *Science* **267**, 1619–1620 CrossRef Medline

51. Hills, R. D., Jr., and Brooks, C. L., 3rd (2009) Insights from coarse-grained Gō models for protein folding and dynamics. *Int. J. Mol. Sci.* **10**, 889–905 CrossRef Medline

52. Noel, J. K., and Onuchic, J. N. (2012) The many faces of structure-based potentials: from protein folding landscapes to structural characterization of complex biomolecules. In *Computational Modeling of Biological Systems* (Dokholyan, N. V., ed) pp. 31–54, Springer US, Boston

53. Hyeon, C., and Thirumalai, D. (2011) Capturing the essence of folding from biased simulations. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18760–18765 CrossRef Medline

54. Azia, A., and Levy, Y. (2009) Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. *J. Mol. Biol.* **393**, 527–542 CrossRef Medline

55. Yadahalli, S., Hemanth Giri Rao, V. V., and Gosavi, S. (2014) Modeling non-native interactions in designed proteins. *Isr. J. Chem.* **54**, 1230–1240 CrossRef

56. Whitford, P. C., Miyashita, O., Levy, Y., and Onuchic, J. N. (2007) Conformational transitions of adenylyl kinase: switching by cracking. *J. Mol. Biol.* **366**, 1661–1671 CrossRef Medline

57. Okamoto, Y. (2004) Generalized-ensemble algorithms: enhanced sampling techniques for Monte Carlo and molecular dynamics simulations. *J. Mol. Graph. Model.* **22**, 425–439 CrossRef Medline
80. Huntington, J. A., Read, R. J., and Carrell, R. W. (2000) Structure of a serpin-protease complex shows inhibition by deformation. *Nature* 407, 923–926 CrossRef Medline

81. Dementiev, A., Dobó, J., and Gettins, P. G. W. (2006) Active site distortion is sufficient for proteinase inhibition by serpins: structure of the covalent complex of α1-proteinase inhibitor with porcine pancreatic elastase. *J. Biol. Chem.* 281, 3452–3457 CrossRef Medline

82. Dementiev, A., Simonovic, M., Volz, K., and Gettins, P. G. W. (2003) Canonical inhibitor-like interactions explain reactivity of α1-proteinase inhibitor Pittsburgh and antithrombin with proteinases. *J. Biol. Chem.* 278, 37881–37887 CrossRef Medline

83. Ye, S., Cech, A. L., Belmares, R., Bergstrom, R. C., Tong, Y., Corey, D. R., Kanost, M. R., and Goldsmith, E. J. (2001) The structure of a Michaelis serpin-protease complex. *Nat. Struct. Biol.* 8, 979–983 CrossRef Medline

84. Gooptu, B., and Lomas, D. A. (2009) Conformational pathology of the serpins: themes, variations, and therapeutic strategies. *Annu. Rev. Biochem.* 78, 147–176 CrossRef Medline

85. Huntington, J. A. (2006) Shape-shifting serpins—advantages of a mobile mechanism. *Trends Biochem. Sci.* 31, 427–435 CrossRef Medline

86. Hansen, M., Busse, M. N., and Andreasen, P. A. (2001) Importance of the amino-acid composition of the shudder region of plasminogen activator-inhibitor-1 for its transitions to latent and substrate forms. *Eur. J. Biochem.* 268, 6274–6283 CrossRef Medline

87. Zhang, Q., Buckle, A. M., Law, R. H. P., Pearce, M. C., Cabrita, L. D., Lloyd, G. J., Irving, J. A., Smith, A. I., Ruzyla, K., Rossjohn, J., Bottomley, S. P., and Whistock, J. C. (2007) The N terminus of the serpin, tennin, functions to trap the metastable native state. *EMBO Rep.* 8, 658–663 CrossRef Medline

88. Elliott, P. R., Pei, X. Y., Dafforn, T. R., and Lomas, D. A. (2000) Topography of a 2.0 Å structure of α1-antitrypsin reveals targets for rational drug design to prevent conformational disease. *Protein Sci.* 9, 1274–1281 CrossRef Medline

89. Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, A., Ashford, P., Orengo, C. A., and Sillitoe, I. (2017) CATH: an expanded resource to identify domains in force-quench polyubiquitin folding are domain swapped. *Proc. Natl. Acad. Sci. U.S.A.* 110, 1680–1685 CrossRef Medline

90. Borgia, M. B., Borgia, A., Steward, A., Nettels, D., Wunderlich, B., and Clarke, J. (2010) Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* 474, 662–665 CrossRef Medline

91. Xia, F., Thirumalai, D., and Gräter, F. (2011) Minimum energy compact structures in force-quench polyubiquitin folding are domain swapped. *Proc. Natl. Acad. Sci. U.S.A.* 108, 6963–6968 CrossRef Medline

92. Borgia, A., Kemplen, K. R., Borgia, M. B., Soranno, A., Shammas, S., Wunderlich, B., Nettels, D., and Clarke, J. (2015) Transient misfolding dominates multidomain protein folding. *Nat. Commun.* 6, 8861 CrossRef Medline

93. Eng, S., Kielbassa, J., Le, J. M., and Whisstock, J. C. (2011) Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* 474, 662–665 CrossRef Medline

94. Yu, M. H., Lee, K. N., and Kim, J. (1995) The Z type variation of human α1-antitrypsin causes a protein folding defect. *Nat. Struct. Biol.* 2, 363–367 CrossRef Medline

95. Knaupp, A. S., Levin, V., Robertson, A. L., Pearce, M. C., and Bottomley, S. P. (2010) Kinetic instability of the serpin Z α1-antitrypsin promotes aggregation. *J. Mol. Biol.* 396, 375–383 CrossRef Medline

96. Lomas, D. A., Evans, D. L., Finch, J. T. J., and Carrell, R. W. (1992) The mechanism of Z α1-antitrypsin accumulation in the liver. *Nature* 357, 605–607 CrossRef Medline

97. Brantly, M., Courtney, M., and Crystal, R. G. (1988) Repair of the secretion defect in the Z form of α1-antitrypsin by addition of a second mutation. *Science* 242, 1700–1702 CrossRef Medline

98. Siwers, R. N., Hardick, C. P., and Woo, S. L. (1989) Disruption of the 290–342 salt bridge is not responsible for the secretory defect of the PiZ α1-antitrypsin variant. *J. Biol. Chem.* 264, 2997–3001 CrossRef Medline

99. Ferguson, N., Capaldi, A. P., James, R., Kleanthous, C., and Radford, S. E. (1999) Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* 286, 1597–1608 CrossRef Medline

100. Capaldi, A. P., Shastry, M. C., Kleanthous, C., Roder, H., and Radford, S. E. (2001) Ultrarapid mixing experiments reveal that Im7 folds via an on-pathway intermediate. *Nat. Struct. Biol.* 8, 68–72 CrossRef Medline

101. Capaldi, A. P., Kleanthous, C., and Radford, S. E. (2002) Im7 folding mechanism: misfolding on a path to the native state. *Nat. Struct. Biol.* 9, 209–216 CrossRef Medline

102. Wallin, S., Zeldovich, K. B., and Shakhnovich, E. I. (2007) The folding mechanics of a knotted protein. *J. Mol. Biol.* 368, 884–893 CrossRef Medline

103. Parra, R. G., Schafer, N. P., Radusky, L. G., Tsai, M.-Y., Gutzovsky, A. B., Wolyne, P. G., and Ferreiro, D. U. (2016) Protein Frustration: a tool to localize energetic frustration in protein molecules, now with electrostatics. *Nucleic Acids Res.* 44, W356–W360 CrossRef Medline

104. Dolmer, K., and Gettins, P. G. W. (2012) How the serpin α1-proteinase inhibitor folds. *J. Biol. Chem.* 287, 12425–12432 CrossRef Medline

105. Braselmann, E., Chaney, J. L., and Clark, P. L. (2013) Folding the proteinase inhibitor folds. *J. Biol. Chem.* 288, 6274–6283 CrossRef Medline

106. Liu, K., Maciuba, K., and Kaiser, C. M. (2019) The ribosome cooperates to prevent conformational disease. *J. Mol. Biol.* 605–607 CrossRef Medline

107. Tanaka, T., Hori, N., and Takada, S. (2015) How co-translational folding of multi-domain protein Is affected by elongation schedule: molecular simulations. *PLoS Comput. Biol.* 11, e1004536 CrossRef Medline

108. Wako, H., and Saitô, N. (1978) Statistical mechanical theory of the protein conformation. I. General considerations and the application to homopolymers. *J. Phys. Soc. Jap.* 44, 1931–1938 CrossRef

109. Wako, H., and Saitô, N. (1978) Statistical mechanical theory of the protein conformation. II. Folding pathway for protein. *J. Phys. Soc. Jap.* 44, 1939–1945 CrossRef

JBC REVIEWS: Simulating the folding of large proteins
121. Muñoz, V., and Eaton, W. A. (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures. Proc. Natl. Acad. Sci. U.S.A. 96, 11311–11316 CrossRef Medline

122. Bergasa-Caceres, F., Haas, E., and Rabitz, H. A. (2019) Nature’s shortcut to protein folding. J. Phys. Chem. B 123, 4463–4476 CrossRef Medline

123. Jacobs, W. M., and Shakhnovich, E. I. (2017) Evidence of evolutionary control in the endoplasmic reticulum. Proc. Natl. Acad. Sci. U.S.A. 114, 11434–11439 CrossRef Medline

124. Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1997) Interpreting hydrogen-deuterium exchange events in proteins using at- cal oxidation of proteins (FPOP): a powerful mass spectrometry based structural proteomics tool. Nat. Struct. Mol. Biol. 14, 2368–2376 CrossRef Medline

125. Kantaev, R., Riven, I., Goldenzweig, A., Barak, Y., Dym, O., Peleg, Y., Albeck, S., Fleishman, S. J., and Haran, G. (2018) Manipulating the folding landscape of a multidomain protein. J. Phys. Chem. B 122, 11030–11038 CrossRef Medline

126. Zhang, G., Hubalewska, M., and Ignatova, Z. (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. Nat. Struct. Mol. Biol. 16, 274–280 CrossRef Medline

127. Chaney, J. L., Steele, A., Carmichael, R., Rodriguez, A., Specht, A. T., Ngo, K., Li, J., Emrich, S., and Clark, P. L. (2017) Widespread position-specific conservation of synonymous rare codons within coding sequences. PLoS Comput. Biol. 13, e1005531 CrossRef Medline

128. Sharma, A. K., and O’Brien, E. P. (2018) Non-equilibrium coupling of protein structure and function to translation-elongation kinetics. Curr. Opin. Struct. Biol. 49, 94–103 CrossRef Medline

129. O’Brien, E. P., Ciryam, P., Vendruscolo, M., and Dobson, C. M. (2014) Understanding the influence of codon translation rates on cotranslational protein folding. Acc. Chem. Res. 47, 1536–1544 CrossRef Medline

130. Panchenko, A. R., Luthey-Schulten, Z., and Wolynes, P. G. (1996) Predicting the membrane-triggered conformational transition of an 266 protein using simplified models. Methods 294, 11969–11979 CrossRef Medline

131. Englander, S. W., Mayne, L., Bai, Y., and Sosnick, T. R. (1997) Hydrogen exchange: the modern legacy of Lindermuth-Lang. Protein Science 6, 1101–1109 CrossRef Medline

132. Fossat, M. J., Dao, T. P., Jenkins, K., Dellarole, M., Yang, Y., McCallum, S. A., Garcia, A. E., Barrick, D., Roumestand, C., and Royer, C. A. (2016) High-resolution mapping of a repeat protein folding free energy landscape. Biophys. J. 111, 236–2376 CrossRef Medline

133. fanot, M. J., Dao, T. P., Jenkins, K., Dellarole, M., Yang, Y., McCallum, S. A., Garcia, A. E., Barrick, D., Roumestand, C., and Royer, C. A. (2016) High-resolution mapping of a repeat protein folding free energy landscape. Biophys. J. 111, 236–2376 CrossRef Medline

134. Destaillats, E. M., Samant, R. S., and Frydman, J. (2017) Mechanisms and functions of spatial protein quality control. Annu. Rev. Biochem. 86, 97–122 CrossRef Medline

135. Adams, B. M., Oster, M. E., and Hebert, D. N. (2019) Protein quality control in the endoplasmic reticulum. Protein Sci. 38, 317–329 CrossRef Medline

136. Van Durme, J., Maurer-Stroh, S., Gallardo, R., Wilkinson, A., Rousseau, F., and Schymkowitz, J. (2009) Accurate prediction of DNA-binding protein detachment via homology modelling and experimental data. PLoS Comput. Biol. 5, e1000475 CrossRef Medline

137. Schneider, M., Rosam, M., Glaser, M., Patronov, A., Shah, H., Back, K. C., Daake, M. A., Buchner, J., and Antes, I. (2016) BiP:Pred: Combined sequence- and structure-based prediction of peptide binding to the Hsp70 chaperone BiP. Proteins 84, 1390–1407 CrossRef Medline

138. Behnke, J., Mann, M. J., Scrucca, F.-L., Feige, M. J., and Hendershot, L. M. (2016) Members of the Hsp70 family recognize distinct types of sequences to execute ER quality control. Mol. Cell 63, 739–752 CrossRef Medline

139. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J., and Serrano, L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat. Biotechnol. 22, 1302–1306 CrossRef Medline

140. Bandypadhyay, B., Goldenzweig, A., Unger, T., Adato, O., Fleishman, S. J., Unger, R., and Horovitz, A. (2017) Local energetic frustration affects the dependence of green fluorescent protein folding on the chaperonin GroEL. J. Biol. Chem. 292, 20583–20591 CrossRef Medline

141. Gámez, A., Yuste-Checa, P., Desviant, L. R., Brasil-Montiano, Á., Horvitz, H. R., and Bardwell, J. C. (2017) Chaperone-client interactions: non-specificity engenders multifunctionality. J. Biol. Chem. 292, 12010–12017 CrossRef Medline

142. Sinz, A. (2018) Cross-linking/mass spectrometry for studying protein structures and protein-protein interactions: where are we now and where should we go from here? Angew. Chem. Int. Ed. Engl. 57, 6390–6396 CrossRef Medline

143. Gómez, A., Yuste-Checa, P., Beris, S., Briso-Montiano, Á., Desviant, L. R., Ugarte, M., Pérez-Cerdá, C., and Pérez, B. (2018) Protein misfolding diseases: prospects of pharmacological treatment. Clin. Genet. 93, 450–458 CrossRef Medline

144. Das, A., and Plotkin, S. S. (2013) Mechanical probes of SOD1 predict systematic trends in metal and dimer affinity of ALS-associated mutants. J. Mol. Biol. 425, 850–874 CrossRef Medline

145. Cazzolli, G., Wang, F., a Beccara, S., Gershenson, A., Faccioli, P., and Wintrode, P. L. (2014) Serpin latency transition at atomic resolution. Proc. Natl. Acad. Sci. U.S.A. 111, 15414–15419 CrossRef Medline

146. Levi, M., Noel, J. K., and Whitford, P. C. (2019) Studying ribosome dynamics with simplified models. Methods 162, 128–140 CrossRef Medline

147. Giri Rao, V. V. H., Desikan, R., Ayappa, K. G., and Gosavi, S. (2016) Capturing the membrane-triggered conformational transition of an α-helical pore-forming toxin. J. Phys. Chem. B 120, 12064–12078 CrossRef Medline

148. Morcos, F., Jana, B., Hwa, T., and Onuchic, J. N. (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc. Natl. Acad. Sci. U.S.A. 110, 20533–20538 CrossRef Medline

149. Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc. Natl. Acad. Sci. U.S.A. 108, E1293–E1301 CrossRef Medline

150. Zhang, B., and Wolynes, P. G. (2015) Topology, structures, and energy landscapes of human chromosomes. Proc. Natl. Acad. Sci. U.S.A. 112, 6062–6067 CrossRef Medline

151. McGuffee, S. R., and Ecolco, A. H. (2010) Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput. Biol. 6, e1000694 CrossRef Medline

152. Earnest, T. M., Cole, J. A., and Luthey-Schulten, Z. (2018) Simulating biological processes: stochastic physics from whole cells to colonies. Rep. Prog. Phys. 81, 052601 CrossRef Medline
161. Yu, I., Mori, T., Ando, T., Harada, R., Jung, J., Sugita, Y., and Feig, M. (2016) Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. *eLife* **5**, e19274 Medline

162. Noel, J. K., Levi, M., Raghunathan, M., Lammert, H., Hayes, R. L., Onuchic, J. N., and Whitford, P. C. (2016) SMOG 2: a versatile software package for generating structure-based models. *PLoS Comput. Biol.* **12**, e1004794 CrossRef Medline

163. Kenzaki, H., Koga, N., Hori, N., Kanada, R., Li, W., Okazaki, K., Yao, X.-Q., and Takada, S. (2011) CafeMol: a coarse-grained biomolecular simulator for simulating proteins at work. *J. Chem. Theory Comput.* **7**, 1979–1989 CrossRef Medline

164. Neelamraju, S., Wales, D. J., and Gosavi, S. (2019) Go-Kit: a tool to enable energy landscape exploration of proteins. *J. Chem. Inf. Model.* **59**, 1703–1708 CrossRef Medline

165. Lutz, B., Sinner, C., Heuermann, G., Verma, A., and Schug, A. (2013) eSBMTools 1.0: enhanced native structure-based modeling tools. *Bioinformatics* **29**, 2795–2796 CrossRef Medline

**JBC REVIEWS: Simulating the folding of large proteins**