JCTC Journal of Chemical Theory and Computation

Article

Cite This: J. Chem. Theory Comput. 2018, 14, 3849−3858

pubs.acs.org/JCTC

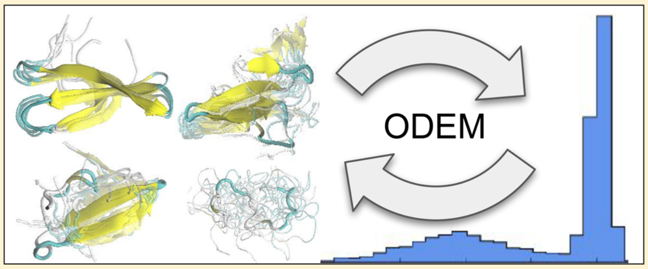
Learning Effective Molecular Models from Experimental Observables

Justin Chen,[†,§] Jiming Chen,[‡] Giovanni Pinamonti,[¶] and Cecilia Clementi*[,‡,¶,§,⊥]

[†]Department of Physics and Astronomy, Rice University, Houston, Texas 77005, United States
[‡]Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas 77005, United States
[¶]Department of Mathematics and Computer Science, Freie Universität, Berlin, Germany
[§]Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States
[⊥]Department of Chemistry, Rice University, Houston, Texas 77005, United States

Supporting Information

ABSTRACT: Coarse-grained models are an attractive tool for studying the long time scale dynamics of large macromolecules at a level that cannot be studied directly by experiment and is still out of reach for atomistic simulation. However, coarse models involve approximations that may affect their predictive power. We propose a modeling framework that allows us to design simplified models to accurately reproduce experimental observables. We demonstrate the approach on the folding mechanism of a WW domain. We show that when the correct coarsening resolution is used not only do the optimized models match the Reference model simulated experimental data accurately but additional observables not directly targeted during the optimization procedure are also reproduced. Additionally, the analysis of the results shows that localized frustration plays an important role in the folding mechanism of this protein and suggests that nontrivial aspects of the protein dynamics are evolutionary conserved.



INTRODUCTION

Recent advances in experimental technologies[1,2] as well as in high-performance techniques to simulate molecular systems at a microscopic level[3] have produced significant progress in the characterization of macromolecular systems and our understanding of complex biological processes such as protein folding, binding, and association.[4,5] Experiments and simulations offer complementary views on these processes, and their synergy is becoming essential to molecular studies. While experiments can produce accurate measurements of equilibrium and dynamical properties of biomolecular processes, they usually provide low resolution information; i.e., they only measure time evolution of a few order parameters or ensemble averages rather than the dynamical evolution of the full microscopic structure. On the other hand, microscopic simulations can, in principle, resolve details that are inaccessible to experiments. However, they involve approximations or empirical terms that usually lead to a systematic, but *a priori* unknown, bias.

Additionally, the time and length scales required to study the dynamic interplay of macromolecular complexes in very large biological processes, such as the ones bridging molecular and cellular mechanisms, still go far beyond what is possible to simulate with the most accurate atomistic molecular models,[3,6,7] even on special-purpose computers.[5,8] For this reason, coarse-grained models that "renormalize" groups of atoms into "effective" degrees of freedom have become a popular choice for the study of collective/organizing motions at time scales and system sizes inaccessible to atomistic simulations.[9−11] By sacrificing the atomistic details, coarse-grained models can explore significantly larger time and length scales. The justification for the use of reduced models lies on mathematical results[12] showing that, at least in principle, the behavior of macromolecular systems over long time scales is regulated by a small set of slow collective variables and not every single atomic degree of freedom is *per se* essential. However, in practice, the price of coarse-graining is usually paid with the introduction of additional approximations and uncontrolled biases, and aggressive coarse-graining can significantly exacerbate the problem of a direct and quantitative comparison and integration of simulation and experimental results.

Here, we instead take the following view: we want to find a "minimalist" molecular model that contains as much physical detail as necessary to model the process of interest but can be efficiently sampled and optimized to reproduce a given set of experimental measurements (and/or results from more accurate models). We propose a general and computationally efficient strategy to compute the effective potential associated with such constrained minimalist models. Our approach is based on the quantification of the agreement between simulation and experimental result by a "quality score" $Q(\epsilon)$ that depends on all the model parameters $\{\epsilon\}$ and takes the uncertainties in simulation and experiment into account. This

Received: February 19, 2018
Published: May 28, 2018

DOI: 10.1021/acs.jctc.8b00187

function $Q(\epsilon)$ is always bound between 0 and 1, where 1 indicates perfect agreement with the experiment. An optimal model can then be obtained by maximizing $Q(\epsilon)$. Different and heterogeneous sources of experimental data can be used to incorporate different measured properties. We name our approach as "Observable-driven Design of Effective Molecular models" (ODEM).

ODEM presents significant improvements with respect to previous iterative model optimization schemes (IMOSs). First, we take a conceptual step forward by proposing an IMOS based on Markov state models (MSMs). An implied condition for all IMOSs is that an equilibrium distribution can be sampled, which is an increasingly difficult requirement for larger biomolecular systems. Unlike previously proposed schemes,[13−16] using MSMs allow us to sample the global equilibrium distribution with short non-equilibrium trajectories that can be run in parallel, offering a significant speedup for iterative model optimization.[4] Furthermore, the concept of MSMs is also combined with the idea of ensemble reweighting.[17,18] Reweighting ensembles allows us to obtain an explicit expression for the quantitative agreement of the simulation results with experimental measurement and allows the model parameters to iteratively change in order to maximize such an agreement. Finally, the quantification of the agreement of a simulation model with experimental measurements as a function of the different model components can be used to test the importance of different physicochemical factors in determining the behavior under observation. This can in turn lead to the formulation of general principles regulating biomolecular processes, e.g., used in combination with bottom-up coarse-graining methodologies to evaluate the performance of different modeling choices, such as the coarse-grain resolution used.

As a proof of principle, and in order to control the sources of errors, here, we apply this approach on a system where "experimental measurements" of physical observables are simulated from long atomistic simulations. We show that by using a set of simulated single-molecule FRET measurements a coarse-grained model can be defined that reproduces the folding mechanism of protein FiP35 in significant detail, when compared to the Reference model atomistic simulation. Finally, by analyzing the frustrated interactions identified by our approach, we speculate about their functional role and show a possible link with coevolutionary correlations.

Incorporation of Experimental Data in a Computational Model. Coarse-grained models are usually designed either by means of some (qualitative or quantitative) model reduction from fine-grained or first-principle models[19−21] and/or by fitting the model parameters to reproduce desired properties.[10,22] While bottom-up approaches are rigorously based on the theory of statistical mechanics, they still involve several empirical choices, and the coarsening can still introduce significant approximations in the estimates of physical observables. A different class of reduced models, so-called structure-based models, eliminate frustration on a protein energy landscape. These models are frequently used as a sort of "'ideal gas'" for protein folding[23] and large macromolecular motion[24] and have provided considerable insight into the physical chemistry of these processes.[25] Even if they offer a solid zeroth order approximation, structure-based models by themselves are usually not generally applicable to study large conformational changes in macromolecules quantitatively. In addition, several methods have been proposed to use experimental data to correct the prediction from simulation for possible modeling biases.[6,15,18,26,27] Instead of using measured data to correct the results, here, we make the additional step of correcting the model itself. Approaches with a similar general philosophy have been previously proposed in the field of force-matching coarse-graining,[16] ultracoarse-graining,[19] chromosome modeling,[28−30] implicit solvent modeling,[31] and disordered proteins.[14]

Here, we propose ODEM as a general procedure for integrating disconnected pieces of evidence from different sources into a trained model that can then be used to investigate properties of the system not directly measurable experimentally, while also suggesting testable experimental hypotheses, generating a positive feedback loop. The basic idea is to first define a model that contains only the physicochemical components that are considered essential to reproduce some features of interest, e.g., the slow time scale processes in the macromolecular system. If the model is missing important ingredients, the ODEM procedure will quantify the inconsistency between what the model can reproduce and the set of measurements, thus providing a test for the hypotheses at the base of the modeling choices. Typically, the model Hamiltonian, $H(\epsilon)$, contains a sum of interaction potentials among the "'effective particles"' used to represent the system, e.g., C_α−C_β for coarse-grained protein representations. These energy terms contain parameters, such as force constants, bond lengths, effective charges, Lennard-Jones parameters, or others. We call the set of parameters ϵ, and they will be "'learned'" by the comparison with the available measurements. Specifically, the application presented here uses only two-body interactions in the nonlocal part of the Hamiltonian (see Supporting Information for details), and the parameters represent the weights of the linear combination of the different interaction terms in the Hamiltonian. However, the Hamiltonian could be designed to include different terms, such as multibody interactions.[32] In principle, the need for more complex interactions in the Hamiltonian representing a system of interest could be evaluated with ODEM by determining their contribution in the improvement of the accuracy of the model. All the parameters in the Hamiltonian are initially set to some (generally poor) guess, $\epsilon^{(0)}$.

Given this initial model, experimental observables computed from molecular dynamics (MD) simulations can be compared with the true experimental observation. The agreement between simulation results and experimental data is measured by a scoring function, $Q(\epsilon)$, expressed as a function of the model parameters, varying between 0 and 1, where the simulation is in perfect agreement with experiment for $Q = 1$, whereas $Q \approx 0$ indicates that the simulation has a poor agreement with experiment.

While the method can, in principle, account for arbitrary combinations of thermodynamic and kinetic measurements, we consider here the case of a given set of K thermodynamic measurements (f_1, ..., f_K)—e.g., specific residue−residue distance distributions estimated from FRET measurements—with associated uncertainties (Δf_1, ..., Δf_K).

An optimal set of parameters can then be found by the following iterative approach:

1. The initial parameter values $\epsilon^{(0)}$, defines a model Hamiltonian $H^{(0)}(\mathbf{x})$, where $\mathbf{x}$ indicates a configuration of the system (e.g., the Cartesian coordinates of the

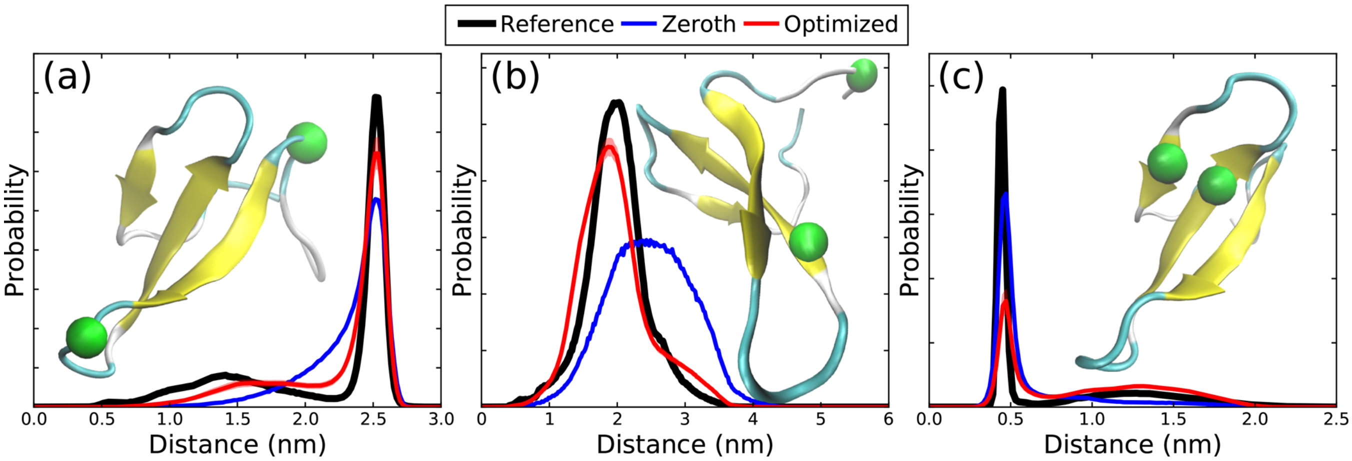


Figure 1. Pair distance distributions used as observables: (a) Pro6-Arg14, (b) Ser2-Arg17, and (c) Phe21-Ser28 distance for zeroth order, Reference, and Optimized models, sampled from the equilibrium distribution of the MSM computed on folding and unfolding trajectories of the protein. The deep red curve indicates the average over the 20 optimized models, while the (very narrow) pale red shaded area around it indicates the standard deviation over the different models. Green beads on the FiP35 native structure mark the Cα atoms between which the distance is calculated.

effective atoms of the protein system). We run MD simulations with $H^{(0)}(\mathbf{x})$.

2. We estimate the equilibrium distribution $\pi^{(0)}(\mathbf{x})$ for the model for the choice of parameters, $\epsilon^{(0)}$, from the MD data. Computationally, the equilibrium distribution is usually evaluated on discrete states, such that $\pi^{(0)}(\mathbf{x})$ is not a function but a vector $\boldsymbol{\pi}^{(0)}$. One practical approach is to discretize the state space into N clusters, S_i, $i = 1,\cdots,N$, to estimate a maximum-likelihood Markov State model (MSM)[33] on it and to compute its equilibrium distribution $\boldsymbol{\pi}^{(0)}$.[34,35] The use of an MSM allows us to obtain an equilibrium distribution from ensembles of short, nonequilibrium simulations.[36]
3. For each experimental measurement available, f_k, we use the distribution $\boldsymbol{\pi}^{(0)}$ to estimate the value, $e_k[\boldsymbol{\pi}^{(0)}]$, of the corresponding observable $g_k(\mathbf{x})$, e.g., the probability of a FRET distance. If the measurement f_k is an ensemble average of that observable, it can be obtained as

$$e_k[\boldsymbol{\pi}^{(0)}] = \sum_i \frac{\pi_i^{(0)}}{n_i} \sum_{\mathbf{x}\in S_i} g_k(\mathbf{x}) \tag{1}$$

where n_i is the number of samples in S_i. Kinetic measurements can be similarly estimated from the MSM.
4. If we treat the measurements as statistically independent and normally distributed, we can compute the likelihood of the simulation prediction in light of the experimental measurements, $Q^{(0)} = Q(\epsilon^{(0)})$. [More sophisticated error models can be used that might better suit a particular observable.[37]] This is given as

$$\begin{aligned} Q &= \mathbb{P}[(e_1[\boldsymbol{\pi}^{(0)}], \ldots, e_K[\boldsymbol{\pi}^{(0)}])|\mathrm{Exp}] \\ &= \prod_k \mathcal{N}(e_k[\boldsymbol{\pi}^{(0)}]; f_k, \Delta f_k) \end{aligned} \tag{2}$$

where $\mathcal{N}(x;\, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. That is, each f_k observable is considered as the mean, μ, of a Gaussian distribution $\mathcal{N}(x;\, \mu, \sigma)$ and its associated uncertainty Δf_k is used for the variance σ of the distribution. The Gaussian is then evaluated on the value of the corresponding simulated observable, e_k, that is implicitly a function of the model parameters. Details on the estimate of e_k from simulations are presented in the Results section.
5. The equilibrium distribution $\boldsymbol{\pi}^{(1)}$ associated with a new parameter set $\epsilon^{(1)} = \epsilon^{(0)} + \delta\epsilon$, for a small change $\delta\epsilon$, can be estimated in each discrete set, S_i, by reweighing as[17,38]

$$\pi_i^{(1)} = \frac{\pi_i^{(0)}}{n_i} \sum_{x\in S_i} \exp[-\beta(H^{(1)}(\mathbf{x},\, \epsilon^{(1)}) - H^{(0)}(\mathbf{x},\, \epsilon^{(0)}))] \tag{3}$$

where $H^{(1)}$ indicates the Hamiltonian defined by the parameters $\epsilon^{(1)}$, and β is the inverse temperature. The reweighing allows us to obtain an expression for Q as a function of the model parameters $\epsilon^{(1)}$, by combining eqs 1, 2, and 3

$$Q^{(1)} = \mathbb{P}[(e_1[\boldsymbol{\pi}^{(1)}], \ldots, e_K[\boldsymbol{\pi}^{(1)}])|\mathrm{Exp}] \tag{4}$$

6. Equation 4 provides an explicit expression of the likelihood as a function of the model parameters. We can then make a step in the parameter space in the direction that maximizes $Q^{(1)}$, using a bounded gradient descent method (see Supporting Information). New simulations are run with the new parameters $\epsilon^{(1)}$, and a new maximum likelihood equilibrium distribution $\pi^{(1)}(\mathbf{x})$ is obtained. The steps above are repeated until convergence to a final Q^{opt}.

This formulation presents several improvements with respect to previously developed methods to incorporate experimental data into computational models.[13,14,31,39,40] As shown in the Supporting Information, the Q value is mathematically related to the reduced χ^2 score (χ^2_ν) which has been previously used.[14,40] Here, we introduce the incorporation of MSMs into the model (step 2 above), as well as the application of the method to distance probability data. The use of MSMs eliminates the need for long single trajectory equilibrium simulations, while the use of the advanced reweighting scheme TRAM (Transition-based Reweighting Analysis Method[17]) extends this capability when using enhanced-sampling techniques. TRAM only requires that a set of short simulations are locally equilibrated[17] to reproduce the global equilibrium distribution. This allows us to use a "divide and conquer" approach with many concurrent short simulations, drastically reducing the time required for sampling the conformational space of the system. We show in the following that the application of ODEM allows us to define a simple C_α−C_β CG model for the folding of FiP35 that reproduces the dynamics of the protein remarkably well when simulated with extensive all-atom simulations.

Details on the implementations of the steps above are provided in the Supporting Information.

RESULTS

Choice of Models and Observables. In order to demonstrate the efficacy of ODEM, we show here the application to a system where simulated experimental observables are generated from long atomistic simulations in explicit solvent, that is, the folding of WW domain FiP35 (data courtesy of D. E. Shaw Research group). This atomistic model is henceforth referred to as the Reference model.[8] The goal of ODEM is to design an optimal coarse-grained model able to reproduce the long time scale dynamics associated with the Reference model, by using only a small number of global observables measured on the latter. By using the Reference model, the performance of the optimized coarse model can be cross-validated by comparing more detailed features that have not been used for its optimization. In order to compare the dynamics, an MSM was estimated for the Reference model (details on the analysis and validation are provided in the Supporting Information). The resulting model is compatible with previous MSM analyses of this system.[41−43]

The distributions of three C_α−C_α pair distances (Pro6-Arg14, Ser2-Arg17, and Phe21-Ser28, Figure 1a, b, and c, respectively) are chosen as observables and estimated on the Reference model. Distribution of FRET probe distances can be inferred in single molecule FRET experiments by processing and denoising distributions of FRET efficiency over time.[44,45] By using simulated observables (measured in the Reference model), we avoid introducing spurious effects, such as the influence of the FRET probes, that can affect the performance of ODEM, and we are able to "'cleanly'" evaluate its efficiency. The first residue pair, Pro6-Arg14, is chosen as its distance correlates well with the slowest collective coordinate associated with the folding/unfolding dynamics, while the other two pairs of residues were chosen based on some structural intuition. Specifically, the Ser2-Arg17 distance was selected to monitor the flexibility of the first β-sheet, while the Phe21-Ser28 distance was chosen to provide information about the second β-sheet (see Supporting Information). In the following section, we discuss the results obtained when these distributions are used in the model optimization (see SI for full details).

It is important to compare the similarities and differences associated with the use of simulated observables instead of real experimental measurements. Here, we use the distributions of multiple backbone pair distances as our simulated observables. While FRET measurements with multiple probes may not be standard for small proteins, they are becoming increasingly common in the study of larger systems.[46−49] In the particular example presented here, three C_α−C_α pair distances are discretized with a bin spacing of 0.02 nm, where the probability of each bin, N_{count}/N_{norm}, is treated as the observable f_k, and the associated error Δf_k is taken to be $1000 \times \sqrt{N_{count}}/N_{norm}$, where each discrete bin in a distance distribution has N_{count} observed distances, and N_{norm} is the normalization of the probability distribution. The factor 1000 increases the statistical error significantly in order to take into account additional sources of errors present in real experimental data, such as systematic errors (i.e., cross-talk or background noise in FRET measurements). Some of these errors can be reduced by accurate postprocessing of the experimental measurements (i.e., denoising the FRET data), and several techniques have been proposed toward this goal.[50,51] The main source of error that is not explicitly accounted for in our model stems from the fact that real experimental FRET measurements report on the distance distributions between FRET probes, and that can be different from the distance distributions between backbone atoms in the protein that are used here as our simulated observables. In principle, the conversion between these sets of distances could be more accurately modeled when real FRET measurements are used in ODEM. For instance, FRET probes could be explicitly included, as has been done in coarse-grained models in recent work.[49] However, the introduction of FRET probes in the model would render the simulations more computationally demanding and would negate most of the benefits of using a CG model, especially in ODEM where multiple iterations are required. We note that experiments with multiple FRET probes have the additional advantage that they allow us to extrapolate the distribution of backbone pair distances from the measured FRET efficiencies with a relatively small error (about 2 Å),[48] and we expect this extrapolation to provide reliable data to reproduce extended conformational changes in large proteins with ODEM. Altogether, the additional sources of uncertainties associated with real experimental data could increase the error for the different measurements nonuniformly and that in turn could alter the position of the optimal solution in parameter space. While a more accurate incorporation of these additional errors requires a more complex model for Δf_k, the use of a large statistical error in the simulated measurements used here can partially account for them. Even if the results presented provide a simple test model for ODEM, we believe the method is still applicable to real experimental data in the future.

A C_α−C_β CG model is used here, where the C_α atoms of each residue represent the backbone and the C_β atom represents the center of mass of the side chains.[52] For the C_α−C_β Hamiltonian, we use a simple representation with local steric interactions (bonds, angles, and dihedrals potentials) and tertiary interactions in the form of Gaussian contact potentials (see Supporting Information for details). Very short simulations were performed at various temperatures (11 different simulations, one for each temperature 115, 116 K··· to 125 K) at each ODEM iteration and then reweighted to the temperature maximizing the agreement with observables (maximizing Q-value). The temperatures were chosen to sample structures near the folding temperature which is around 120 K due to the renormalization of the energy scale in the CG model.

ODEM Optimization. Figure 1 shows the agreement between the observables computed on the Reference model, the model used as the starting point, and the models resulting from the ODEM optimization. We use 20 different starting points for the ODEM optimization, generated by adding randomly distributed noise to each of the interaction energy parameters of a plain structure-based model for a C_α−C_β representation of the FiP35 protein (called the zeroth order model in the following). Each of the 20 starting points was iteratively optimized with ODEM for 10 iterations. We compare the Optimized coarse-grained models to the zeroth order model and to the Reference (all-atom) model. The parameters that are iteratively optimized by ODEM, as discussed above, are the strength of the interaction between each pair of effective atom types (C_α and C_β of all the different residue types). All models converge with an increase of 10 orders of magnitude in the likelihood Q^{opt} with respect to the zeroth order model (see Figure 2).

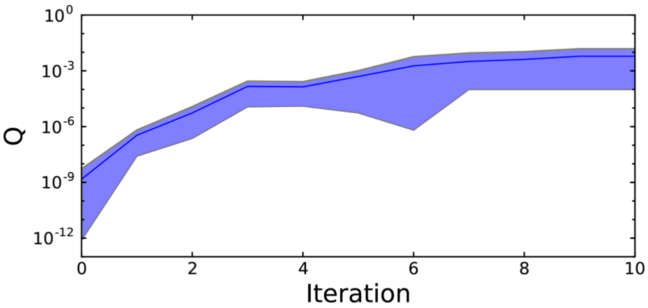


Figure 2. Convergence of the model likelihood, Q (see eq 4), computed over the three distance observables, for the $C_\alpha-C_\beta$ coarse-grained models (blue) as a function of the ODEM iterations. The average over the different optimized models is reported, while the shaded area indicates the standard deviation over the models.

Figure 1 shows that while the simulated FRET distributions computed on the zeroth order model are all quite far from the Reference model, in the Optimized models these observables are in agreement with the Reference values. Looking at the three distance distributions, we see (Figure 1) that while the zeroth order model has a single sharp peak in the Pro6-Arg14 and Phe21-Ser28 $C_\alpha-C_\alpha$ distances the distributions exhibit two-peaks for the Optimized models, mainly a new broad peak which closely resembles the corresponding behavior in the Reference model. The fact that the Optimized model cannot fit all three distances perfectly might be due to the functional form of the CG Hamiltonian, which is significantly less complex than an all-atom Hamiltonian. However, the overall agreement of the Optimized model to the Reference model shows remarkable improvement (10 orders of magnitude) in Figure 2. Q^{opt} converges to a value near 0.0061 in the final iteration. While this value may at first appear still low as a likelihood, it is important to note that it is obtained as a product of multiple Gaussians (see eq 2), and its geometric average per observable corresponds to a value of 0.988. An alternative way to compare the results is by means of the reduced χ^2_ν score. By the definition of Q^{opt}, the χ^2_ν score is proportional to $-\ln(Q^{opt})$ and decreases by a factor of four from the zeroth order to the optimized model.

In order to better understand the changes in the dynamics, an MSM analysis of the Reference, zeroth order, and Optimized models is performed to identify an Unfolded (U) and Folded (F) state based on PCCA+ membership,[53] and then a committor probability between U and F is used to define the Transition State (TS) for each model (see Supporting Information for details). While the parameter sets of the ODEM-optimized $C_\alpha-C_\beta$ models are not identical, they are all essentially equivalent in reproducing the folding process of the Reference model accurately. Figure 3 shows that while the zeroth order model is not able to reproduce the folding mechanism of the Reference model, all Optimized models capture its transition state with remarkable accuracy. The primary difference between the zeroth order model and the Reference model is that the first β-sheet is formed in the transition state of the zeroth order model but not in the Reference model. From previous studies on FiP35,[41,42] it was found that while folding can, in principle, initiate with the formation of either β-sheets the first β-sheet has a strong preference to form first, and this feature is absent in the zeroth order model's transition state but present in the optimized model. The only difference observed in the contacts maps of the different (U, F, TS) states between the Optimized and the Reference model consists in the formation of a few weak and transient non-native contacts in the unfolded state; in the Optimized models, these contacts tend to form between the first and second β-strands, while in the Reference model they are more delocalized in the contact map. However, this difference does not affect other properties of the unfolded state, such as the L_{trp} observable (simulated Trp spectroscopy) illustrated by Figure 4.

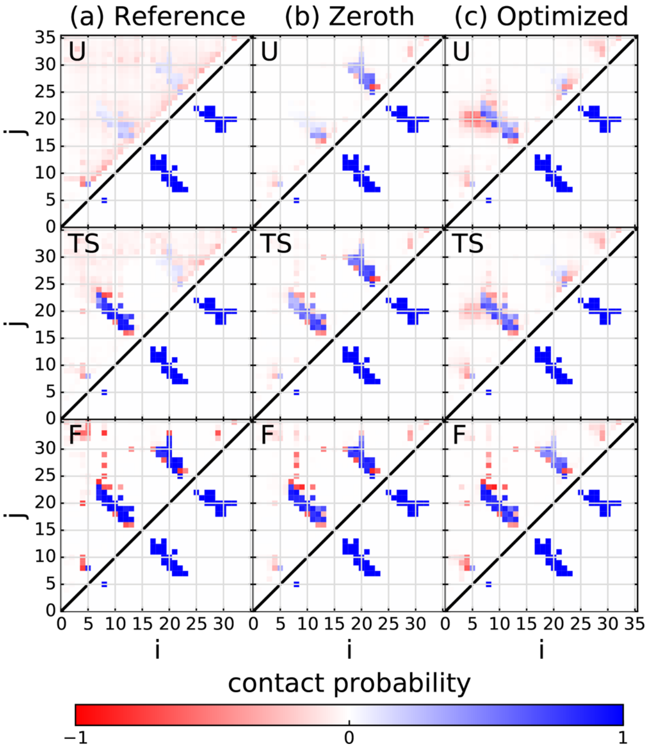


Figure 3. Average contact maps for the 20 models for states U (unfolded), TS (transition state), and F (folded): (a) Reference, (b) zeroth order, (c) Optimized. Different shades of blue (red) quantify the probability of formation of native (non-native) contacts. (a) The Reference model's U state shows a weak non-native contact formation and a weak residual β-sheet structure, while TS presents a formed first β-sheet with an unformed second β-sheet. (b) Unfolded configurations of the zeroth order models are all partially collapsed around the second β-sheet, producing a strong tendency to form the second β-sheet in the U state and near complete folding of both β-sheets in the TS. (c) Optimized models present a TS consistent with the one of the Reference model, where the first β-sheet is formed. Reference and Optimized models differ only in the formation of a few weakly formed non-native contacts in the U state.

In order to further gauge the ability of the Optimized coarse-grained models to reproduce the all-atom Reference model, an additional simulated experimental observable (not used in the

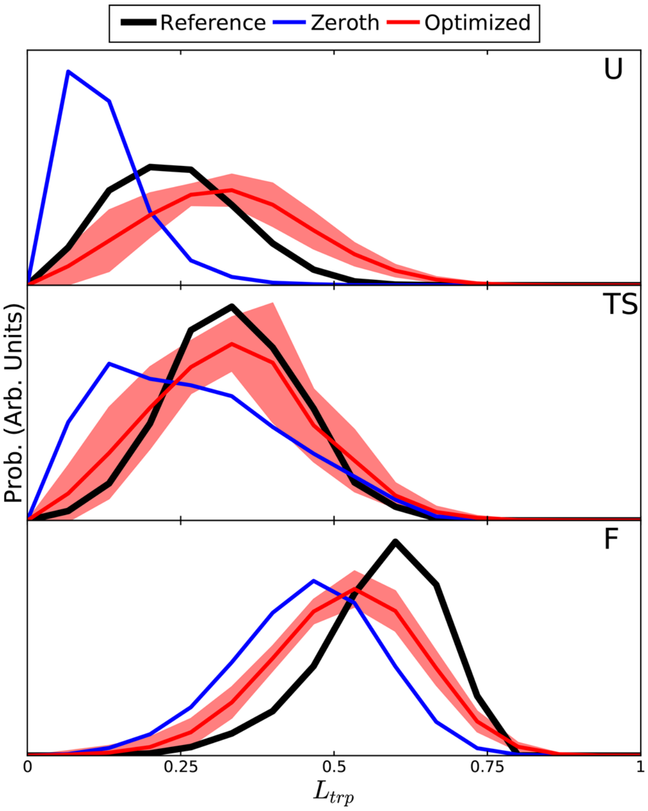


Figure 4. Histograms for the simulated Trp fluorescence measurements, $L_{trp}(t)$, for each model and for each state (U, TS, F). The red curve indicates the average over the 20 Optimized models, while the pale red shaded area indicates the standard deviation over the different models.

ODEM optimization) is used to compare the different states; to mimic the results of Trp spectroscopy, we monitor the formation of tertiary interactions involving residue W8, which is the only Tryptophan residue in the sequence of FiP35 used (see Supporting Information for details). We indicate this quantity as L_{trp}. As the CG beads do not correctly represent the true size and shape of residues, the number of contacts (both native and non-native) formed by the Tryptophan is used here as a proxy for the amount of solvent accessible area for the Tryptophan residues. It was previously shown[54] that the number of contacts formed by a Trp residue computed over a protein folding trajectory correlates well with Trp spectroscopy. In each (U, TS, F) state, the values of L_{trp} are histogrammed and compared for the Reference, zeroth order, and Optimized models in Figure 4. It is clear that while the results for the zeroth order model present significant deviation from the Reference model, particularly in the U and TS states, the Optimized models appear to more closely match the Reference value in all states. To quantify the difference, we report the χ^2_ν for each state and model independently, compared to the Reference model. While for the unfolded state the χ^2_ν score associated with the Optimized and zeroth order model are comparable (both around a value of 60,000), the shape of the distribution of the Optimized model is qualitatively much more similar to the corresponding Reference model for this state. For the transition state, the χ^2_ν score of the Optimized model is 1 order of magnitude smaller than the zeroth order model (114 and 1069, respectively), and it almost a factor of 2 smaller in the folded state (38,000 and 69,000, respectively). Note that the large values of the χ^2_ν scores are due to the small standard deviation of the Reference observables, estimated as $\sqrt{N_{count}}/N_{norm}$ over the MD long trajectories available for the Reference model. It is important to stress that, as W8 is not involved in any of the distance pairs used in the ODEM optimization, L_{trp} provides an independent validation test for the results of the procedure. Additionally, as W8 is located in the region in the contact map where a weak cluster of non-native contacts may form in the U state of the Optimized models, the matching of L_{trp} in the U state show that such a small probability of residual non-native structures does not significantly change the overall behavior of these models with respect to the Reference model.

The time scale associated with the folding process is slower for the Optimized models with respect to the zeroth order model; while the folding time of the zeroth order model is a factor of ≈4000 faster than the Reference model (when the time is measured in internal units in each model for a direct comparison), the average folding time over the different Optimized models is a factor of 2500 ± 200 faster than the Reference model. The folding slow-down during the ODEM optimization is due to the introduction of non-native interactions and energy heterogeneity.[55,56] The remaining ≈2500 factor speed up of the folding time scale of the Optimized models with respect to the Reference model arises from the reduction in degrees of freedom upon coarse-graining, which eliminates fluctuating forces and produces a much smoother overall energy landscape.[57,58] Because of this significant speedup, in addition to the reduced complexity of the coarse-grained model, multiple folding/unfolding events can be simulated for the Optimized models in a matter of a few CPU hours on standard computational facilities. For this reason, coarse-grained models optimized with experimental data provide an attractive alternative to characterize molecular processes still inaccessible to atomistic simulations.

Analysis of Solutions' Space. As reported above, all the ODEM Optimized models present similar thermodynamics and kinetic behavior, although the corresponding parameters are not identical. Figure 5 reports the average value (upper diagonal) and standard deviation (lower diagonal) for the parameters regulating the strength of the residue−residue interactions over all Optimized models. How the different interactions between the C_α and C_β pseudoatoms contribute to the strength of the residue−residue interactions is shown in Figure S1 in the Supporting Information. The locations of some of the residue−residue interactions that are consistently strongly attractive in all models coincide with, or are near to, native interactions (contacts highlighted by a black box in the contact maps in Figure 5). However, two interesting features emerge. First, while about half of the native interactions are always attractive, the other half are not positive in every model and can be found to be repulsive in different Optimized models. These fluctuating interactions are not necessary for the correct folding of this protein. Additionally, a cluster of strongly attractive interactions are found in a non-native region of the contact map, between residues 4−9 and residues 16−22. The propensity of the Optimized models to form non-native contacts in this region in the unfolded state (Figure 3) is clearly connected to this cluster of interactions. Many of the interactions that are found consistently strong in the Optimized models involve residue W8; the participation of this residue in both native and non-native strong interactions makes it

DOI: 10.1021/acs.jctc.8b00187

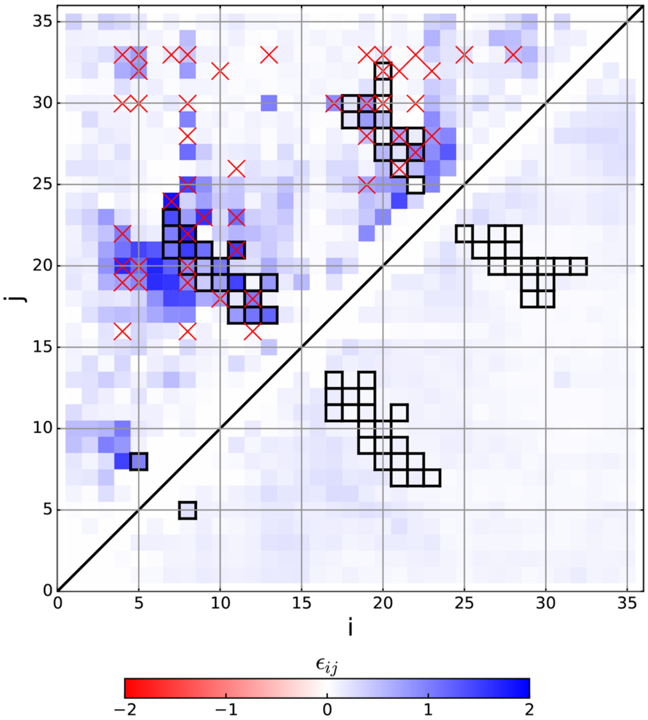


Figure 5. Statistics of the parameters for the ODEM optimized C_α−C_β models obtained from 20 different initial values. The average value (upper diagonal) and standard deviation (lower diagonal) of the interaction strength between two residues (i,j) is reported for each pair, where the strength is averaged in the case both C_α−C_α and C_β−C_β interactions exist. Darker shades of blue indicate stronger attractive interactions, as quantified by the color map. The highlighted black boxes mark the position of native interactions, while red crosses mark the positions of strongly coupled pairs found from mfDCA. A breakdown of each individual interaction in terms of C_α−C_α and C_β−C_β contributions (including repulsive interactions) is available in the Supporting Information.

significantly frustrated. In particular, the interactions of W8 with the loop at the opposite end of its β-strand can not be satisfied by small perturbations of the native structure and compete with native interactions. Because of its persistent presence in all the Optimized models, this frustration is most probably inherited from the Reference model, and it may be an important ingredient to reproduce the folding mechanism correctly, that is, to form the first β-sheet rather than the second one in the transition state (see Figure 3). Indeed, both β-sheets are equally likely to be formed in the transition state of the completely unfrustrated zeroth order model, while this symmetry is broken in the Reference and Optimized models.

In order to test the role of these strong and frustrated interactions on FiP35, we use mean-field Direct Coupling Analysis (mfDCA) to identify residue pairs which are coevolutionary linked.[59] FiP35 belongs to the WW domain family; a multiple sequence alignment of this protein family (Pfam ID: PF00397) is obtained from the Pfam domain database[60] and used as input for mfDCA to obtain the 48 residue pairs with highest coevolutionary coupling. The number 48 is selected as follows: 50 pairs were obtained from the DCA, corresponding to around 1.5 times the length of the MSA. This is a typical cutoff employed in applications of DCA. As two of the pairs among these 50 included an amino acid present in the MSA but not in FiP35's specific sequence, 48 pairs were used. However, results are robust if a reduced set of pairs is used (see Supporting Information, Figure S3). These strongly coupled pairs are marked with red crosses in Figure 5. Very interestingly, both clusters of persistently strong non-native interactions in the ODEM optimized parameter sets are among the interaction pairs with highest coupling. An additional cluster of strongly coupled pairs emerge from mfDCA and involves the interactions of the C terminus (residues 30−35) with the rest of the protein. The promiscuity of the interactions of the C terminus in this protein family is not surprising as it is a flexible linker that can connect the domain to different structures in larger complexes. On the contrary, the existence of a cluster of strongly coupled pairs in the region defined by the interactions between residues 4−9 and residues 16−22 is highly nontrivial. In particular, W8 is one of the residues most frequently found among the pairs with high coevolutionary coupling, and most of the non-native interactions obtained by mfDCA are close to the interactions emerging from ODEM as persistently strong across the different Optimized models.

While the ODEM optimization is designed to reproduce the protein dynamics, mfDCA identifies pairs of residues most likely to interact by means of coevolutionary considerations. As frustrated interactions promote the formation of non-native contacts, they are shaping the folding mechanism but not driving the folding directly and are therefore absent in any structure-based model built on this protein domain. This result suggests that at least some aspects of the folding dynamics of FiP35 are not directly encoded in its native structure and are preserved in the WW domain family.

CONCLUSIONS

We propose ODEM, a theoretical/modeling framework to integrate multiple experimental measurements (that could come from heterogeneous sources) in the definition of optimal coarse-grained models, for the study of molecular processes over long time and length scales. We have illustrated the approach by using simulated FRET measurements to "learn" different coarse-grained models, using a structure-based model as starting point, to reproduce the folding mechanism of protein FiP35, as obtained by extensive atomistic simulations. We note that a similar philosophy was previously used to generate a model for the study of intrinsically disordered proteins.[14]

The results show that a set of C_α−C_β models can be optimized to faithfully characterize the slow dynamics of this protein. The analysis of the different optimized C_α−C_β models reveals that some localized frustration is important to shape the folding mechanism of this protein. Additionally, frustrated interactions appear evolutionarily conserved in the protein family and suggest a link between folding dynamics and evolutionary preserved features that may have important functional implications.

The ODEM approach is particularly relevant in light of the recent advances in experimental key technologies that are now offering an unprecedented view into the spatial and temporal organization of proteins and larger macromolecular complexes in the cellular environment. As experiments are still lacking the ability to simultaneously resolve structure and dynamics in a way that would permit a full characterization of the dynamical mechanisms of large proteins, computer simulations present a powerful complement to experiment. Because of their empirical nature, coarse-grained models are generally less predictive, and it is often unclear a priori whether a chosen coarse-grained model possesses sufficient physicochemical detail to model the protein interactions under investigation. The ODEM approach allows us to design molecular models constrained to be consistent with available experimental data, in order to overcome the intrinsic shortcomings of the coarse-graining modeling approach.

The ultimate goal of ODEM is its application to model the dynamics of protein systems with larger length and time scales, which are difficult to study with all-atom molecular dynamic simulations. We expect several aspects of the application of ODEM discussed here to seamlessly translate to the study of larger systems. For instance, we expect the speed up of time scales observed in the model of FiP35 to be similarly significant in larger proteins. In addition to the speed up in simulation time due to the reduced complexity of the model, coarse-grained models also significantly decrease the time scales by eliminating the fast fluctuating degrees of freedom, time scales by eliminating the fast fluctuating degrees of freedom,[57,58] allowing us to sample large conformational changes faster. We also expect the use of MSMs and TRAM in ODEM to be very effective in the modeling of larger systems. One difference in the application of ODEM to larger systems consists in the fact that the Hamiltonian model may have more parameters that can be optimized. If a larger number of parameters in the Hamiltonian could introduce more flexibility to reproduce the experimental data, it could also increase the computational cost or reduce the interpretability of the model. This problem can be addressed technically by using parallel computing and more efficient optimization algorithms or conceptually by selecting a specific subset of parameters to optimize, guided by physical intuition or other modeling consideration. The possibility to integrate experimental data from multiple sources into mechanistic and quantitative molecular models can provide a solid bridge between simulation and experiment for the complete characterization of complex molecular processes.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jctc.8b00187.

Details about the ODEM implementation and validation and additional technical details. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: cecilia@rice.edu.

ORCID

Jiming Chen: 0000-0001-5232-5382

Cecilia Clementi: 0000-0001-9221-2358

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to D. E. Shaw Research for sharing the protein folding simulation data. We are indebted to Frank Noé (FU Berlin) for his insightful suggestions and his assistance with the MSM and HMM analyses. We thank Lorenzo Boninsegna, Alex Kluber, Feliks Nüske, and Simon Olsson for a critical reading of the manuscript. This work was supported by the National Science Foundation (CHE-1265929, CHE-1738990, and PHY-1427654), the Welch Foundation (C-1570), and the Einstein Foundation in Berlin. Simulations have been performed on the computer clusters of the Center for Research Computing at Rice University, supported in part by the Big-Data Private-Cloud Research Cyberinfrastructure MRI-award (NSF Grant CNS-1338099), and on the clusters of the Department of Mathematics and Computer Science at Freie Universität, Berlin.

■ REFERENCES

(1) Moerner, W. E. Single-molecule spectroscopy, imaging, and photocontrol: Foundations for super-resolution microscopy. *Rev. Mod. Phys.* **2015**, *87*, 1183−1212.

(2) Binshtein, E.; Ohi, M. D. Cryo-Electron Microscopy and the Amazing Race to Atomic Resolution. *Biochemistry* **2015**, *54*, 3133−3141.

(3) Pan, A. C.; Weinreich, T. M.; Piana, S.; Shaw, D. E. Demonstrating an Order-of-Magnitude Sampling Enhancement in Molecular Dynamics Simulations of Complex Protein Systems. *J. Chem. Theory Comput.* **2016**, *12*, 1360−1367.

(4) Plattner, N.; Doerr, S.; De Fabritiis, G.; Noé, F. Complete protein−protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **2017**, *9*, 1005−1011.

(5) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334*, 517−520.

(6) Cavalli, A.; Camilloni, C.; Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **2013**, *138*, 094112.

(7) Balsera, M. A.; Wriggers, W.; Oono, Y.; Schulten, K. Principal Component Analysis and Long Time Protein Dynamics. *J. Phys. Chem.* **1996**, *100*, 2567−2572.

(8) Shaw, D. E.; Maragakis, P.; Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Eastwood, M. P.; Bank, J. A.; Jumper, J. M.; Salmon, J. K.; Shan, Y.; Wriggers, W. Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **2010**, *330*, 341−346.

(9) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10−15.

(10) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S.-J. The MARTINI Coarse-Grained Force Field: Extension to Proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819−834.

(11) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73−93.

(12) Noé, F.; Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: Theory and methods. *Curr. Opin. Struct. Biol.* **2017**, *43*, 141−147.

(13) Matysiak, S.; Clementi, C. Optimal Combination of Theory and Experiment for the Characterization of the Protein Folding Landscape of S6: How Far Can a Minimalist Model Go? *J. Mol. Biol.* **2004**, *343*, 235−248.

(14) Norgaard, A. B.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **2008**, *94*, 182−192.

(15) Pitera, J. W.; Chodera, J. D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445−3451.

(16) Dannenhoffer-Lafage, T.; White, A. D.; Voth, G. A. A Direct Method for Incorporating Experimental Data into Multiscale Coarse-Grained Models. *J. Chem. Theory Comput.* **2016**, *12*, 2144−2153.

(17) Wu, H.; Paul, F.; Wehmeyer, C.; Noé, F. Multiensemble Markov models of molecular thermodynamics and kinetics. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, E3221−E3230.

(18) Olsson, S.; Wu, H.; Paul, F.; Clementi, C.; Noé, F. Combining experimental and simulation data of molecular processes via augmented Markov models. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, 8265−8270.

(19) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining

method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.

(20) Zhang, Z.; Lu, L.; Noid, W. G.; Krishna, V.; Pfaendtner, J.; Voth, G. A. A Systematic Methodology for Defining Coarse-Grained Sites in Large Biomolecules. *Biophys. J.* **2008**, *95*, 5073−5083.

(21) Chaimovich, A.; Shell, M. S. Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.* **2011**, *134*, 094112.

(22) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: Protein Structure Prediction Using Coarse-Grained Physical Potentials and Bioinformatically Based Local Structure Biasing. *J. Phys. Chem. B* **2012**, *116*, 8494−8503.

(23) Clementi, C.; Nymeyer, H.; Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and ″en-route″ intermediates for protein folding? an investigation for small globular proteins. *J. Mol. Biol.* **2000**, *298*, 937−953.

(24) Okazaki, K.-i.; Koga, N.; Takada, S.; Onuchic, J. N.; Wolynes, P. G. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2006**, *103*, 11844−11849.

(25) Ratje, A. H.; et al. Head swivel on the ribosome facilitates translocation via intra-subunit tRNA hybrid sites. *Nature* **2010**, *468*, 713−716.

(26) Morcos, F.; Chatterjee, S.; McClendon, C. L.; Brenner, P. R.; López-Rendón, R.; Zintsmaster, J.; Ercsey-Ravasz, M.; Sweet, C. R.; Jacobson, M. P.; Peng, J. W.; Izaguirre, J. A. Modeling Conformational Ensembles of Slow Functional Motions in Pin1-WW. *PLoS Comput. Biol.* **2010**, *6*, 1−13.

(27) Marinelli, F.; Faraldo-Gómez, J. Ensemble-Biased Metadynamics: A Molecular Simulation Method to Sample Experimental Distributions. *Biophys. J.* **2015**, *108*, 2779−2782.

(28) Giorgetti, L.; Galupa, R.; Nora, E. P.; Piolot, T.; Lam, F.; Dekker, J.; Tiana, G.; Heard, E. Predictive polymer modeling reveals coupled fluctuations in chromosome conformation and transcription. *Cell* **2014**, *157*, 950−963.

(29) Di Pierro, M.; Zhang, B.; Aiden, E. L.; Wolynes, P. G.; Onuchic, J. N. Transferable model for chromosome architecture. *Proc. Natl. Acad. Sci. U. S. A.* **2016**, *113*, 12168−12173.

(30) Zhang, B.; Wolynes, P. G. Topology, structures, and energy landscapes of human chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6062−6067.

(31) Sanyal, T.; Shell, M. S. Coarse-grained models using local-density potentials optimized with the relative entropy: Application to implicit solvation. *J. Chem. Phys.* **2016**, *145*, 034109.

(32) Ejtehadi, M. R.; Avall, S. P.; Plotkin, S. S. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101*, 15088−15093.

(33) Prinz, J.-H.; Wu, H.; Sarich, M.; Keller, B.; Senne, M.; Held, M.; Chodera, J. D.; Schütte, C.; Noé, F. Markov models of molecular kinetics: generation and validation. *J. Chem. Phys.* **2011**, *134*, 174105.

(34) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **2014**, *25*, 135−144.

(35) Scherer, M. K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernàndez, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525−5542.

(36) Nüske, F.; Wu, H.; Prinz, J.-H.; Wehmeyer, C.; Clementi, C.; Noé, F. Markov state models from short non-equilibrium simulations—Analysis and correction of estimation bias. *J. Chem. Phys.* **2017**, *146*, 094104.

(37) Hughes, I.; Hase, T. *Measurements and Their Uncertainties: A Practical Guide to Modern Error Analysis*; Oxford University Press: Oxford, U.K., 2010.

(38) Mey, A. S. J. S.; Wu, H.; Noé, F. xTRAM: Estimating Equilibrium Expectations from Time-Correlated Simulation Data at Multiple Thermodynamic States. *Phys. Rev. X* **2014**, *4*, 041018.

(39) Matysiak, S.; Clementi, C. Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J. Mol. Biol.* **2006**, *363*, 297−308.

(40) Li, D.-W.; Brüschweiler, R. Iterative Optimization of Molecular Mechanics Force Fields from NMR Data of Full-Length Proteins. *J. Chem. Theory Comput.* **2011**, *7*, 1773−1782.

(41) Boninsegna, L.; Banisch, R.; Clementi, C. A data-driven perspective on the hierarchical assembly of molecular structures. *J. Chem. Theory Comput.* **2018**, *14*, 453−460.

(42) Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theory Comput.* **2015**, *11*, 5947−5960.

(43) McGibbon, R. T.; Pande, V. S. Learning Kinetic Distance Metrics for Markov State Models of Protein Conformational Dynamics. *J. Chem. Theory Comput.* **2013**, *9*, 2900−2906.

(44) Keller, B. G.; Kobitski, A.; Jäschke, A.; Nienhaus, G. U.; Noé, F. Complex RNA Folding Kinetics Revealed by Single-Molecule FRET and Hidden Markov Models. *J. Am. Chem. Soc.* **2014**, *136*, 4534−4543.

(45) Taylor, J. N.; Landes, C. F. Improved Resolution of Complex Single-Molecule FRET Systems via Wavelet Shrinkage. *J. Phys. Chem. B* **2011**, *115*, 1105−1114.

(46) Kilic, S.; Felekyan, S.; Doroshenko, O.; Boichenko, I.; Dimura, M.; Vardanyan, H.; Bryan, L. C.; Arya, G.; Seidel, C. A. M.; Fierz, B. Single-molecule FRET reveals multiscale chromatin dynamics modulated by HP1*α*. *Nat. Commun.* **2018**, *9*, 235.

(47) Dimura, M.; Peulen, T. O.; Hanke, C. A.; Prakash, A.; Gohlke, H.; Seidel, C. A. Quantitative FRET studies and integrative modeling unravel the structure and dynamics of biomolecular systems. *Curr. Opin. Struct. Biol.* **2016**, *40*, 163−185.

(48) Hellenkamp, B.; Wortmann, P.; Kandzia, F.; Zacharias, M.; Hugel, T. Multidomain structure and correlated dynamics determined by self-consistent FRET networks. *Nat. Methods* **2017**, *14*, 174−185.

(49) Reinartz, I.; Sinner, C.; Nettels, D.; Stucki-Buchli, B.; Stockmar, F.; Panek, P. T.; Jacob, C. R.; Nienhaus, G. U.; Schuler, B.; Schug, A. Simulation of FRET dyes allows quantitative comparison against experimental data. *J. Chem. Phys.* **2018**, *148*, 123321.

(50) Shuang, B.; Cooper, D.; Taylor, J. N.; Kisley, L.; Chen, J.; Wang, W.; Li, C. B.; Komatsuzaki, T.; Landes, C. F. Fast Step Transition and State Identification (STaSI) for Discrete Single-Molecule Data Analysis. *J. Phys. Chem. Lett.* **2014**, *5*, 3157−3161.

(51) Murphy, R. R.; Danezis, G.; Horrocks, M. H.; Jackson, S. E.; Klenerman, D. Bayesian Inference of Accurate Population Sizes and FRET Efficiencies from Single Diffusing Biomolecules. *Anal. Chem.* **2014**, *86*, 8603−8612.

(52) Cheung, M.; Finke, J.; Callahan, B.; Onuchic, J. Exploring the interplay between topology and secondary structural formation in the protein folding problem. *J. Phys. Chem. B* **2003**, *107*, 11193−11200.

(53) Noé, F.; Schütte, C.; Vanden-Eijnden, E.; Reich, L.; Weikl, T. R. Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19011−19016.

(54) Das, P.; Wilson, C. J.; Fossati, G.; Wittung-Stafshede, P.; Matthews, K. S.; Clementi, C. Characterization of the folding landscape of monomeric lactose repressor: Quantitative comparison of theory and experiment. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 14569−14574.

(55) Plotkin, S. S.; Onuchic, J. N. Structural and energetic heterogeneity in protein folding. I. Theory. *J. Chem. Phys.* **2002**, *116*, 5263−5283.

(56) Das, P.; Matysiak, S.; Clementi, C. Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10141−10146.

(57) Izvekov, S.; Voth, G. A. Modeling real dynamics in the coarse-grained representation of condensed phase systems. *J. Chem. Phys.* **2006**, *125*, 151101.

(58) Matysiak, S.; Clementi, C.; Praprotnik, M.; Kremer, K.; Delle Site, L. Modeling diffusive dynamics in adaptive resolution simulation of liquid water. *J. Chem. Phys.* **2008**, *128*, 024503.

DOI: 10.1021/acs.jctc.8b00187

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

(60) Punta, M.; et al. The Pfam protein families database. *Nucleic Acids Res.* **2012**, *40*, D290−D301.