Rescuing Off-Equilibrium Simulation Data through Dynamic Experimental Data with dynAMMo

Citation for the original published paper (version of record):
Kolloff, C., Olsson, S. (2023). Rescuing Off-Equilibrium Simulation Data through Dynamic Experimental Data with dynAMMo. Machine Learning: Science and Technology, 4(4).
http://dx.doi.org/10.1088/2632-2153/ad10ce

(article starts on next page)

Machine Learning: Science and Technology

PAPER • **OPEN ACCESS**

Rescuing off-equilibrium simulation data through dynamic experimental data with dynAMMo

To cite this article: Christopher Kolloff and Simon Olsson 2023 *Mach. Learn.: Sci. Technol.* **4** 045050

IOP Publishing Mach. Learn.: Sci. Technol. 4 (2023) 045050 https://doi.org/10.1088/2632-2153/ad10ce

MACHINE LEARNING Science and Technology

OPEN ACCESS

RECEIVED 21 July 2023

REVISED 7 November 2023

ACCEPTED FOR PUBLICATION 29 November 2023

PUBLISHED 12 December 2023

PAPER

Rescuing off-equilibrium simulation data through dynamic experimental data with dynAMMo

Christopher Kolloff and Simon Olsson*

Chalmers University of Technology, Department of Computer Science and Engineering,Rännvägen 6, 412 58 Gothenburg, Sweden

* Author to whom any correspondence should be addressed.

E-mail: simonols@chalmers.se

Keywords: Markov state models, dynamic Augmented Markov models, protein dynamics, molecular dynamics simulations, relaxation dispersion spectroscopy, chemical exchange

Supplementary material for this article is available online

Abstract

Long-timescale behavior of proteins is fundamental to many biological processes. Molecular dynamics (MD) simulations and biophysical experiments are often used to study protein dynamics. However, high computational demands of MD limit what timescales are feasible to study, often missing rare events, which are critical to explain experiments. On the other hand, experiments are limited by low resolution. We present dynamic augmented Markov models (dynAMMo) to bridge the gap between these data and overcome their respective limitations. For the first time, dynAMMo enables the construction of mechanistic models of slow exchange processes that have been not observed in MD data by integrating dynamic experimental observables. As a consequence, dynAMMo allows us to bypass costly and extensive simulations, yet providing mechanistic insights of the system. Validated with controlled model systems and a well-studied protein, dynAMMo offers a new approach to quantitatively model protein dynamics on long timescales in an unprecedented manner.

1. Introduction

Understanding the triad of protein structure–function–dynamics is of paramount importance in many fields, including biochemistry, biophysics, and medicine [1–7]. Thanks to extensive studies of the bovine pancreatic trypsin inhibitor (BPTI), for example, we now understand the atomic details of its essential role in inhibiting serine proteases [8]. This knowledge has been possible by combining findings from the fields of x-ray crystallography [9] and nuclear magnetic resonance (NMR) [10–13] with molecular dynamics (MD) simulations [14–16]. However, reconciling experimental and simulation data in a systematic manner often poses problems due to technical and resource limitations. Enabling such a merger would yield a significant opportunity for quantitative structural biology and biophysics.

Typically, we model dynamic experiments [17], such as NMR relaxation dispersion, single molecular Förster resonance energy transfer, dynamic neutron scattering, or x-ray photon correlation spectroscopy using simple n-site jump models [18–22]. This approach yields forward and reverse exchange rates for the different states as well as site populations. However, modeling dynamics this way beyond a two-state exchange is challenging, due to experimental limitations and poor timescale separation. In effect, we are limited to highly simplified models of the complex underlying dynamics of our data where the structure of states often remain elusive or ambiguous [23–28].

Over the past few decades, MD simulations have become increasingly popular in the field of biophysics, providing atomistic insights into the behavior of biological systems at high temporal and spatial resolutions [29–31]. Force field models are steadily improving in quality and their scope is boarding to include disordered proteins and nucleic acids [32–34]. Although not broadly available, the development of special-purpose computers, like Anton [35–37], makes it possible to study millisecond timescale molecular processes. Graphics processing unit (GPU)-accelerated simulations as well as distributed computing

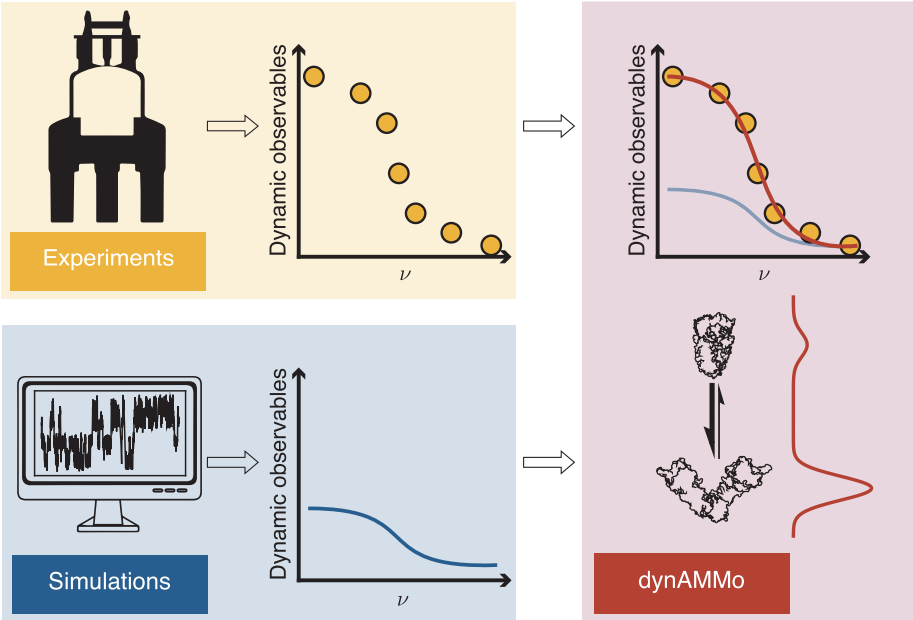

Figure 1. Schematic of dynamic augmented Markov models (dynAMMo). Simulations often fail to reproduce experimentally observed dynamic observables, such as NMR relaxation dispersion data due to force field inaccuracies and finite simulation length. Experiments, on the other hand, are often noisy and often do not reveal the mechanistic details of kinetic exchange. With dynAMMo, we take into account, both, simulation and experimental data to obtain a single mechanistic kinetic model.

initiatives, such as Folding@home [38] or GPUGrid are more widely available allow us to access processes on the micro- to millisecond timescale, in particular when large ensembles of simulations are analyzed using Markov state models (MSMs) [39–43]. MSMs represent the *molecular kinetics* fully: the relevant structural states, their thermodynamic weights and their mutual exchange rates [44–50]. MSMs have enabled studies of many biological processes, such as protein folding, enzymatic activity, or protein–protein interactions [42, 51–57].

Despite these advances in force field accuracy and simulation technologies, we often observe a systematic discrepancy between the experimental values predicted from MD trajectories and experimental data [58–61]. The origin of these discrepancies is two fold. First, imperfections in the force field models remain which lead to skewed populations and altered dynamics. Second, simulations still do not cover the range of biological timescales of interest. Methods like transition [62] and discrete path sampling [63], transition interface sampling [64], milestoning [49, 65], metadynamics [66], flooding [67], or replica exchange [68, 69] offer potential avenues to bridge this timescale gap. The advent of deep generative neural network-based surrogates further provides new venues for overcoming the sampling problems [70, 71]. However, each of these approaches have inherent limitations in their scope or rely on extensive manual intervention. Together, these limitations prevent us from directly comparing to many experiments and thus gaining a mechanistic interpretation of our data.

The integration of simulation and experimental data is a big challenge with a long history [72]. Previous methods include *post hoc* reweighing or sub-selection of simulations data [73, 74], modeling kinetics with generated ensembles [75], biasing simulations with stationary experimental data [76–85] and dynamic experimental data [86, 87], and building MSM using experimental and simulation data [88], augmented Markov models (AMMs). AMMs combine stationary experimental observables with simulation data to correct for bias in MD data, which improved agreement with complementary stationary and dynamic data [88]. However, AMMs cannot take into account dynamic data, and consequently also cannot deal with situations where our simulations have not sampled processes which are important to explain the data.

Here, we present dynAMMo, a new approach that accounts for stationary and dynamic experimental data, such as $R_{1\rho}$ or Carr–Purcell–Meiboom–Gill (CPMG) relaxation dispersion experiments [20, 89, 90], to estimate a Markov model (figure 1). By combining constrained optimization with the principle of maximum entropy, we are able to correctly recover experimental timescales from biased simulations and are able to model exchange between states not seen in the simulation data as long as the states themselves are known. To our knowledge, this is the first method that enables the construction of mechanistic models of protein dynamics, even when rare events remain unobserved in the MD data. This achievement is made possible through the dynamic experimental data that complement the simulations by reporting on the exchange

processes that have not been sampled by the simulations. With dynAMMo, we therefore address the aforementioned issue in MD simulations by circumventing the, often very costly, need for (reversibly) sampling rare events in order to establish a kinetic mechanistic model. Also, we can show that our method fails and thus does not overfit when one or more relevant states are missing. The method therefore broadens the scope for future research in understanding the complex dynamics of biomolecular systems in general and brings the field closer to the development of data-driven models that accurately capture the underlying mechanisms-of-action.

2. Theory

2.1. dynAMMo

MSMs are based on the discretization of the state space Ω of a molecule into n states. By following the traversal of an MD trajectory through these states we can estimate the transition probabilities p_{ij} from states i to states j through the analysis of transition counts $c_{ij}(\tau)$ with a lag time τ [39, 40, 43]. The resulting transition matrix $\mathbf{T}(\tau) \in \mathbb{R}^{n\times n}$ encodes the *molecular kinetics* of the system, including the populations of the states and the rates of exchange between them. This information is accessible through the spectral components of $\mathbf{T}(\tau)$, the eigenvectors $\mathbf{R}$ and eigenvalues $\boldsymbol{\lambda}$, as well as the stationary distribution $\boldsymbol{\pi}$ (see supporting information, 'Theory') [43].

AMMs [88] aim to estimate an MSM which matches stationary experimental observables, such as NMR ^3J-coupling or residual dipolar coupling data that probe the 'true' Boltzmann distribution, by reweighing the relative populations of the states through the maximum entropy principle. Even though this approach does not directly take into account information about the kinetic rates between the states, Olsson and Noé observe that the integration of stationary observables has an effect on the prediction of dynamic observables, such as $R_{1\rho}^{\text{ex}}$ relaxation dispersion. However, in general we cannot expect AMMs to match dynamic experimental data, nor can they consider cases in which not all states of the MSM are connected.

Here, we address these limitations with dynAMMo that combine simulation data, in the form of one or more count matrices $\mathcal{C}$ (supplementary information, algorithm 1, line 1), and dynamic and stationary experimental observables $\mathbf{o}^{\text{exp}}$ to a single kinetic model. By combining these sources of information, we aim to obtain a more accurate and comprehensive representation of biomolecular dynamics. Using experimental data that report on the conformational exchange kinetics, we can directly estimate the forward and reverse rate constants of switching from one state to another. Unlike Brotzakis *et al*, no prior knowledge of the kinetic rates are required nor are we limited to a two-state exchange.

2.2. Connection between experiments and simulations

For each experimental observable $\mathbf{o}^{\text{exp}}$, we assume that there is a corresponding observable function $f(\cdot)$ (*forward model*) available that maps all configurations, $\mathbf{x} \in \Omega$, to a complex or real vector space, $\mathcal{V}$, however, often just a scalar, e.g. a distance or a chemical shift. For MSMs, we can average these values over the n Markov states yielding $\mathbf{a} \in \mathcal{V}^n$ [88]. Here, we focus on dynamic experiments, where we measure time correlations of these observables either directly or through a transformation. From an MSM, $\mathbf{T}(\tau)$, we can compute the time correlation of f:

$$\langle f(\mathbf{x}(0))^\top f(\mathbf{x}(k\tau))\rangle \approx o^{\text{dynamic}}(k) = \mathbf{a}^\top \boldsymbol{\Pi}\mathbf{T}(\tau)^k \mathbf{a}, \tag{1}$$

where $\boldsymbol{\Pi}$ is the diagonal matrix of the stationary distribution $\boldsymbol{\pi}$, and $^\top$ is the transpose or the complex conjugate. We can compare this quantity directly to experimentally measured counterparts and thereby use it to drive the estimation of MSMs. Many dynamic observables, however, are transformations of the time-correlation, rather than the time-correlation itself. This includes, among others, CPMG and $R_{1\rho}^{\text{ex}}$ relaxation dispersion, which measure the convolution of the time-correlation function with a spin-lock field. Here we assume fast chemical exchange, and use previously described closed-form expression for Markov models [91, 92] to predict data and drive MSM estimation (see supporting information, 'Theory' for a more detailed explanation).

2.3. Estimation of dynAMMo

We estimate dynAMMo models by optimizing a loss function which includes the transition counts c_{ij} from states i to j and the sum of the mean-square difference between the predictions and experimental data, $\mathcal{D}$, of the lth observable at the kth lag-time,

$$\arg\min_{\hat{\boldsymbol{\lambda}},\hat{\mathbf{R}},\hat{\boldsymbol{\pi}}} \mathcal{L}\left(\hat{\mathbf{T}}(\tau) \mid \mathcal{D}, \mathbf{C}(\tau)\right) = -\sum_{ij} c_{ij}\log p_{ij} + \sum_{l,k}\left(o_{l,k}^{\text{pred, dyn}} - o_{l,k}^{\text{exp, dyn}}\right)^2. \tag{2}$$

Here, p_{ij} is the probability of transitioning between states i and j. The loss is computed with respect to the spectrum of $\hat{\mathbf{T}}(\tau)$, i.e. the eigenvalues $\hat{\boldsymbol{\lambda}}$, eigenvectors $\hat{\mathbf{R}}$, and the stationary distribution $\hat{\boldsymbol{\pi}}$ and is subject to several constraints. To estimate $\hat{\boldsymbol{\lambda}}$ and $\hat{\mathbf{R}}$, we use gradient descent with additional orthogonality constraints for optimization of the eigenvectors. Rather than enforcing orthogonality directly with a penalty term in the loss function, dynAMMo optimizes the eigenvectors on the Stiefel manifold through Riemannian optimization [93]. Following AMMs, we further have the option to include stationary experimental observables as described previously [88]. The estimation procedure as well as the theoretical details are explained in more detail in supporting information, 'Theory'.

3. Results and discussion

3.1. Enforcing dynamic experimental constraints on a kinetic model rescues biased simulation data

To demonstrate the power of dynAMMo, we will first examine our model by applying it to two one-dimensional energy potentials: the Prinz potential [43] (figure 2, brown background) and the three-well potential (figure 2, teal background). In both model systems, there is one slow transition as well as one or more faster transition(s) that we aim to model. The Prinz potential has four states with comparable populations with one slow transition between the first and last two states, whereas the three-well potential has two fast-interchanging low-energy states and one state with a high energy barrier. In both model systems, we used the slowest eigenvectors as an observable function as they encode near perfect information about the slowest process.

We will first examine the scenario where we have biased simulations, which we compare with the experimental data derived from a 'ground truth' model. In this case, all states were reversibly sampled in the simulations. However, due to, for example, force field inaccuracies and finite sampling, the timescales of exchange between the processes and the thermodynamics of the system do not correspond to the 'true' ensemble. By investigating the free energy profiles of the two systems (figure 2(A)/(J)), we see that the populations of the MSM (blue) match the ground truth well (yellow). Consequently, we would not expect reweighing considering only the stationary observables, as is done in AMMs, will not have a big effect on the prediction of the dynamic observables since the timescales show significant discrepancies (figure 2(B)/(K)) between the MSM and the ground truth. Using dynAMMo, we can perfectly match the slowest timescales (figure 2(B)/(K)). We find a similar mismatch between the MSMs estimated on the biased data and the 'ground truth data' (figure 2(C)/(L)). Since our observable function inherently informs about the slowest process we find that, dynAMMo does not substantially modify the timescale of the faster processes (figure 2(B)). This implies that the model does not introduce unnecessary bias into the estimation if it is not reflected in the observable. As opposed to the Prinz potential, we find that the predicted kinetics in the model trained on biased data from the three-well potential (figure 2, panels (J)–(L)) are accelerated compared to the ground truth (figure 2(K)). This mismatch in timescales manifests itself as poor agreement with the observable time correlation functions (figure 2(L)) between the ground truth (yellow) and the naive MSM (blue). Integrating the correlation function data and the biased simulation data with dynAMMo, we are in agreement with the ground-truth data and match the underlying rates.

3.2. Disconnected simulations can be combined to a single Markov model using dynamic constraints

Many systems are characterized by timescales which are impractical to sample with statistical confidence. However, we may have access to multiple experimental structures of each of the states in isolation, some of which we can sample transitions between, others which are infeasible to sample. An example is the bovine BPTI, for which numerous studies have reported slow millisecond timescale dynamics [94–97], and a millisecond long MD simulation only sampled the suspected slow transition once [16]. In many other cases, sampling such a transition in an unbiased fashion remains impractical.

To test such a scenario, we designed two experiments where we discard the transition counts of the slowest transition (figures 2(D) and (E)/(M) and (N)), gray dotted line) and split the trajectory in two. We build two MSM corresponding to the, now, disconnected subregions of the state space (supporting information, 'Materials and Methods'). After reweighing the populations using the stationary AMM procedure [88], we perfectly match the model (red) and the ground truth (yellow) populations (figure 2(D)/(M)). Despite not having prior information on the slowest process, we can correctly identify the missing timescale using dynamic experimental observables (figure 2(E)/(N)). Our model bridges the two sides, even in the absence of observed transitions between them. This discovery is guided by the correlation function, i.e. the observable (figure 2(F)/(O)). The correlation function indirectly holds this information (equation (S4)), as a slower exchange process corresponds to a slower decay in the correlation function. Since we can match, both, the kinetics and the thermodynamics of the systems, we can also fit the observable prediction (figure 2(F)/(O) red, solid) to the ground truth (figure 2(F)/(O) yellow, dashed). Using

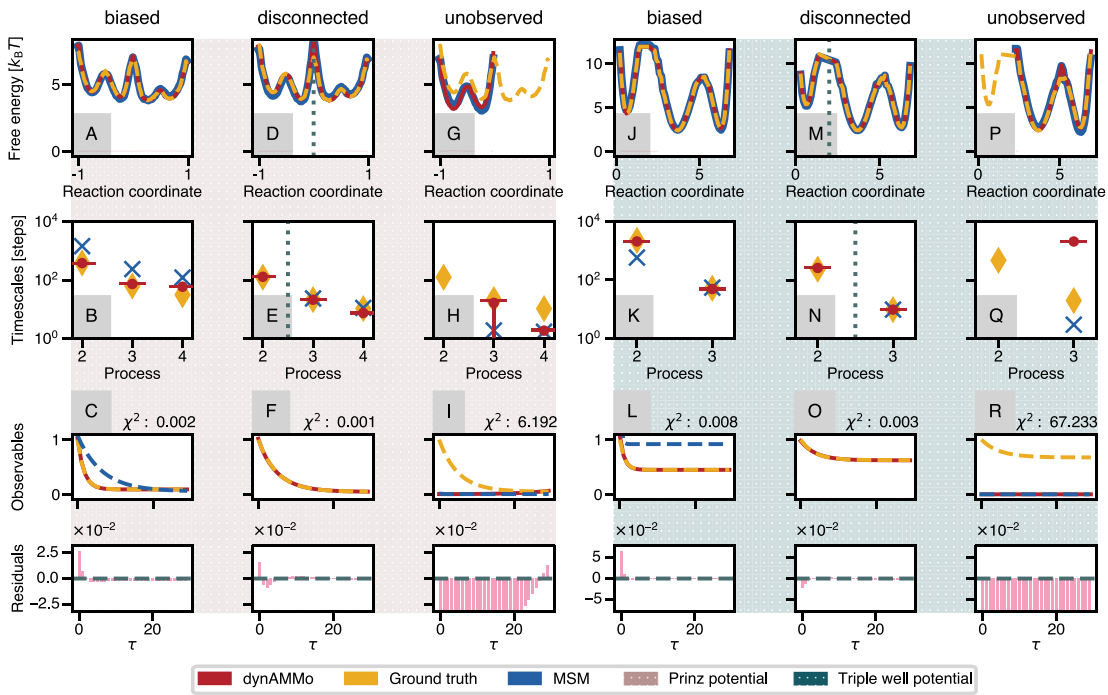

Figure 2. Overview of model system benchmark results. Three different scenarios using two model systems, the Prinz potential (brown background) and the three-well potential (teal background) are shown. The three different scenarios include biased simulations (A)–(C), (J)–(L), disconnected trajectories (D)–(F), (M)–(O), and unobserved or missing states (G)–(I), (P)–(R). Each of the scenarios show the free energy potential, timescales of exchange, and observable plots. χ^2 values and residuals between the model predictions and the ground truth data are also shown for the observable plots. The model results are shown in red, the 'ground truth' experimental data in yellow, and the naive MSM predictions in blue. ΔG values of the slowest transition for the different systems are given in the supporting information (table S4).

experimental dynamic observables, we can thereby merge disconnected simulation statistics and estimate a single model which faithfully reproduces all the available data.

3.3. dynAMMo does not overfit when relevant states are missing

Next, we consider the case where one or more states that contribute to a measurable experimental signal is missing from the MD simulation data. This situation is common in MD simulation studies as the simulation time is often insufficient to sample all the relevant states, and is an edge case related to the 'disconnected' situation discussed above. However, contrary to the previous case, we do not have all structural information to support a reliable prediction of the observable here. We therefore anticipate that dynAMMo is unable to yield a model perfectly fitting these data. To simulate this scenario, we discard transition counts from one state completely. Concretely, this procedure discards simulation data about states above 0 (figures 2(G)–(I)) in the Prinz potential and states below 2.5 in the asymmetric triple well potential (figures 2(P)–(R)). In this case, models built with dynAMMo cannot match (figure 2(I)/(R)) the data, which translates into missing timescales (figure 2(H)/(Q)). Mismatching predictions indicate that the model is failing, which suggests that one or more states that give rise to a measurable signal are missing.

3.4. A mechanistic model of BPTI disulfide isomerization dynamics with dynAMMo

To test how our model performs in a realistic scenario, we turned to BPTI as a protein system. BPTI is a 58 residue protein whose dynamics has been extensively studied, both experimentally [95, 96, 98, 99] and computationally [16, 100]. BPTI is known to have micro- to millisecond conformational exchange [95, 96, 101], centered mainly on different isomerizations of the disulfide bond between Cys14 and Cys38. In addition, there is a 1 ms long MD trajectory available at a temperature of 300 K [16]. In this simulation, all known major conformations of BPTI are sampled, and the transitions between them show a distinct separation of timescales. Analysis of the trajectory shows conformational exchange in the fast microsecond regime [16, 88, 102], which is much faster than what the experimentally determined rates are suspected to reflect these processes. The discrepancy between the experiments and the simulation data makes BPTI an ideal test case for demonstrating dynAMMo as an avenue to reconcile the data.

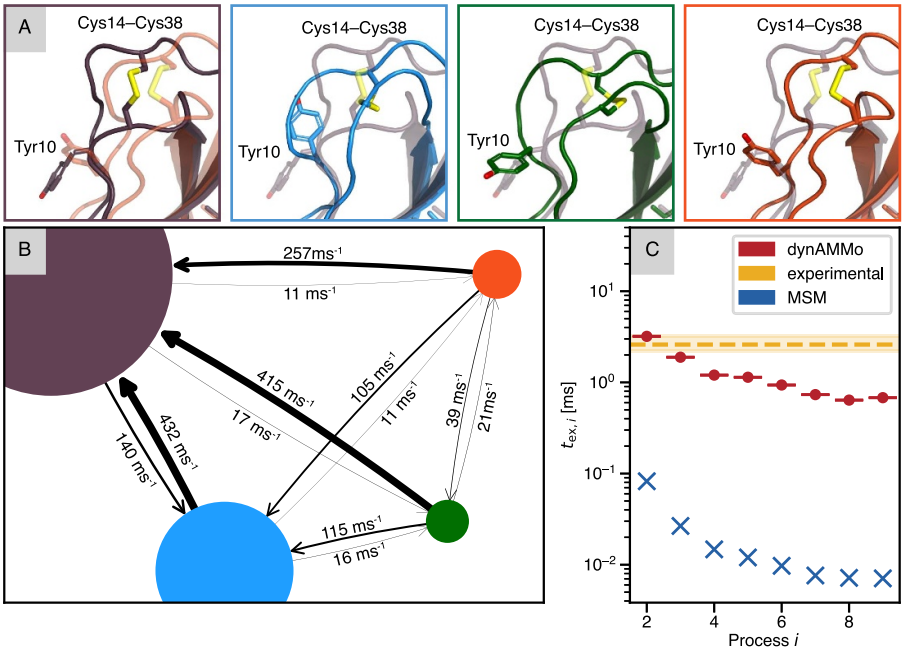

Figure 3. Integrating BPTI simulations and CPMG NMR data to build a quantitative kinetic model with dynAMMo. (A) Representative structures of the major BPTI states with the structural characteristics highlighted. Aromatics are shown for orientation. (B) Kinetic network between macrostates. The colors of the nodes correspond to the structural representatives shown above. The size of the nodes and the arrow widths are proportional to the size of the populations as well as the reaction rate, respectively. Rates are shown above/below the arrows in ms^{-1}. (C) Timescales of exchange as a function of the slowest processes. dynAMMo is shown in red and the MSM from the simulation data is colored blue. The time constant of the experimentally determined exchange is shown in yellow with the standard deviation shown as shaded area.

By integrating simulation [16] and experimental CPMG data [94] from NMR spectroscopy with dynAMMo, we build a kinetic model of BPTI (figure 3). In line with previous analyses [88, 103], we used time-lagged independent component analysis [104] to define a low dimensional space which we discretized into 384 states and aggregated into four metastable structural states. Consistent with previous analyses, the major structural substates display isomerization of the disulfide bridges between residues 14 and 38 (figure 3(A)). We show the most populated states colored purple and light blue (a total population of approximately 90%), while the remaining population is shared by the green and orange states (figure 3(A)). The state connectivity is dense and rates vary across an order of magnitude (figure 3(B)), which we show with arrows of varying thickness between the states colored by identity and scaled by their relative populations. The slowest rates correspond to the transitions to the two minor states (supporting information, figure S6(a)) and we expect to occur in the low millisecond regime. The implied timescales computed from our dynAMMo model are systematically shifted compared to those of the naive MSM that only takes simulation statistics into account (figure 3(C)). For the MSM, the slowest processes are barely on the order of hundreds of microseconds (dark blue crosses) [91]. On the other hand, the slowest implied timescale estimated by dynAMMo is on the order of magnitude of approximately 3.2 ms. This timescale matches well with those estimated from experimental data at the same temperature using a two-state fit, where Millet *et al* determined a chemical exchange of the order of 2–3 ms [94].

In figure 4 we show representative examples of some key observables. The plots show CPMG relaxation dispersion curves, which measure the effect of chemical exchange on $^{15}N^{H}$ spins. The chemical shift predictions that were used as observables for the backbone amides were obtained by the PPM algorithm [105]. The experimental CPMG data are shown in yellow, whereas the predictions of the model are shown in red and the predictions were scaled according to the values reported in the supporting information (figure S7). All observables show an excellent overall agreement with the data, suggesting that the underlying model is capable of explaining the data in a meaningful way (see supporting information, figures S8 and S9). We note that all relevant residues involved in the exchange display a relatively strong dispersion, which we are able to perfectly match using dynAMMo. This observation strengthens the argument that the conformations sampled in the MD simulation constitute the relevant configurations needed to explain the experimental

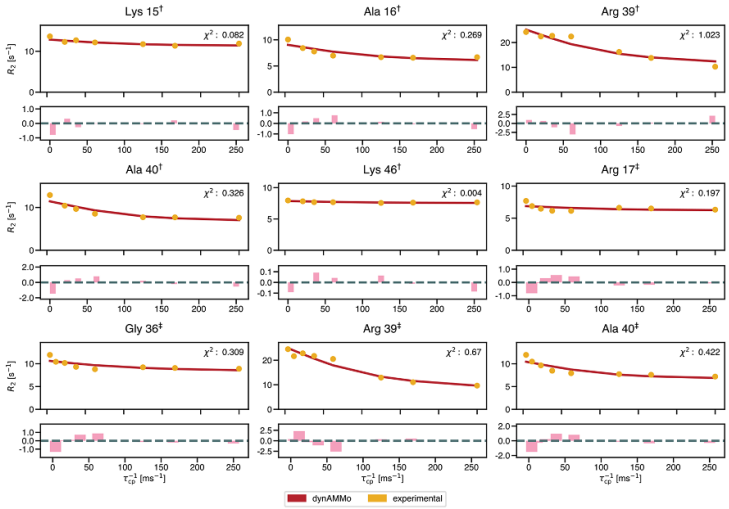

Figure 4. CPMG relaxation dispersion data of BPTI $^{15}N^{H}$ spins. Nine representative examples of CPMG plots are shown. The fitted model (red, solid) is in high agreement with the experimental data (yellow circles). χ^2 values between the prediction and the experiments are shown for each subplot. The dagger † refers to the dataset that has been recorded at a Larmor frequency of 600 MHz. Conversely, the double dagger ‡ refers to the dataset recorded at 500 MHz [94]. Residuals between the experiments and predictions are shown in pink.

data. We stress that dynAMMo uses all observables to fit one global kinetic model. Therefore, the predictions of the relaxation dispersion curves for the backbone nitrogens differ only by the observable used for each residue. Here, we demonstrate how dynAMMo can be used to combine experimental NMR relaxation data with simulation data from MD simulations. Therefore, we obtain a detailed mechanistic explanation of how the different metastable states observed in the MD trajectory contribute to the experimentally probed chemical exchange.

3.5. Quantitative molecular kinetics from disconnected simulation statistics with dynAMMo

Above we saw how dynAMMo could recover the correct kinetics on a controlled test system. To evaluate whether dynAMMo generalizes to more complex protein systems, we establish a similar benchmark, systematically removing simulation statistics that connect the major and minor populated states of BPTI. For this case, we similarly find that dynAMMo can quantitatively recover the exchange rates between the disconnected states (figure 5(A)), and recover the implied timescales accurately (figure 5(B)), despite the minor discrepancies between the timescales in the connected and disconnected case (figures 3 and 5). The discrepancies observed, although noticeable, are on the same order of magnitude. We are to expect these due to a combination of limited data and data uncertainty. The MSM in this case is missing the slowest process (figure 5(A), dashed cross), however, dynAMMo can recover this process and quantitatively predict the timescale. We show a detailed analysis of this scenario in the supporting information (figure S6(b)).

4. Conclusion

Here we have introduced dynAMMo, a new method to improve the accuracy of mechanistic biomolecular models by incorporating dynamic experimental measurements to correct for biases in the kinetics and thermodynamics of MD simulations. However, most intriguingly, dynAMMo also allows us to build quantitatively predicted model of molecular kinetics even in the absence of simulation statistics on (slow) conformational transitions. We show the performance of dynAMMo across two well-controlled benchmark systems and later deploy it to two realistic scenarios using data from molecular simulations and NMR spectroscopy on the protein BPTI. It is essential to highlight that while dynAMMo offers significant advancements, the robustness of the model still depends on the quality of the initial MD and experimental data. In cases where MD simulations inadvertently do not sample certain rare events or states, we can only build a (useful) dynamic AMM if all states are known but only some transitions are missing ('disconnected' model case). The model fails if one or more important state are missing, offering important insights into the conformational space of the molecule. This observation reinforces the need for comprehensively sampling the relevant conformational states, either through simulations or with experimental techniques. Further, since dynAMMo balances experimental and simulation data through the principle of maximum entropy, we

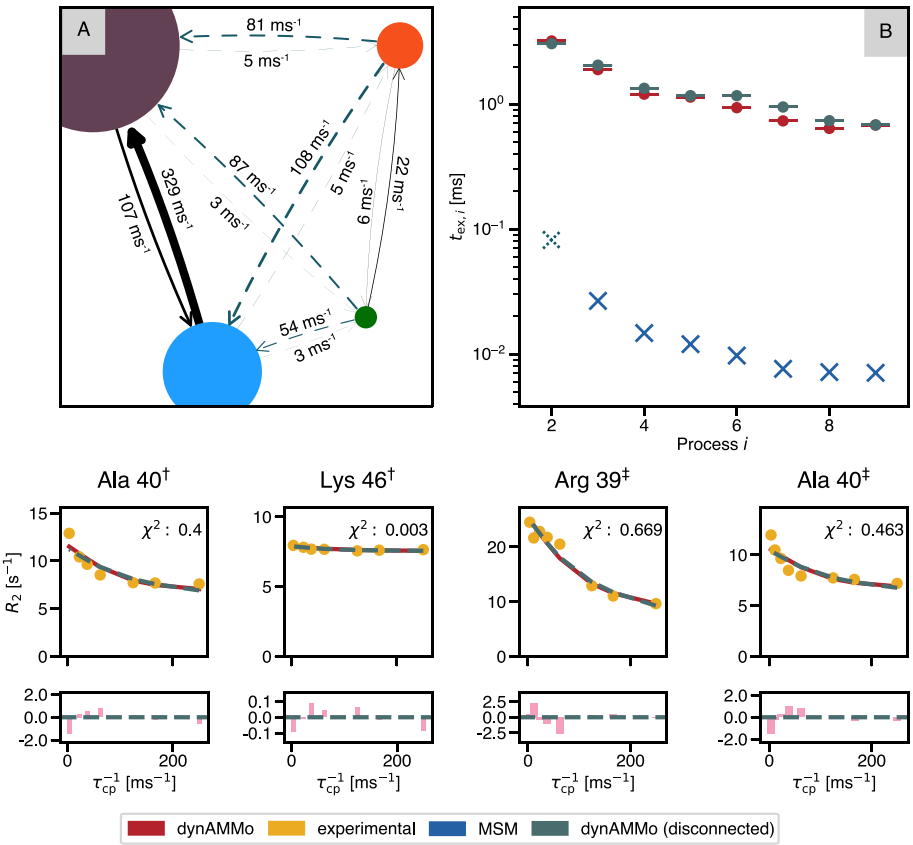

Figure 5. Overview of simulating disconnected case using BPTI simulations and CPMG data. (A) Kinetic network between the disconnected states. All transitions between the purple/light blue and orange/green clusters were removed (indicated with dashed gray arrows) and two MSM were built using only the within-states trajectory data. The colors of the states correspond to the clusters shown in figure 3(A). (B) Implied timescale plot of the disconnected model (gray), the connected model (red), and the MSM (blue) for comparison. The presumed timescale of exchange that was removed in this scenario is indicated as a dashed cross. Lower panel: four representative examples of CPMG relaxation dispersion predictions of selected backbone nitrogens. The predictions of the disconnected case (dashed gray) are plotted together with the predictions of the connected scenario (red) for comparison. χ^2 values are shown with respect to the predictions of the disconnected case and the experimental data (yellow).

approach a compromise from an information theoretical perspective. Therefore, although multiple models could potentially explain the data, we identify the one that requires minimal perturbation from the simulation data to align with the available experimental evidence. As such, dynAMMo opens up the possibility of salvaging sparsely sampled simulation data sampled using biased force fields and repurpose them to build quantitatively predictive models for structural biology.

5. Materials and methods

The estimator is implemented in Python and uses PyTorch [106] and DeepTime [107] as the main analysis and modeling tools. The estimation procedure and theory details are provided in supporting information, 'Theory'. The code will be made available on https://github.com/olsson-group/dynAMMo. All figures showing molecular structures were made using PyMol [108]. All plots were generated using Matplotlib [109]. Additional results, such as the slowest estimated eigenvectors, loss function, and stationary distribution are reported in the supporting information for all model systems (figures S4 and S5) and BPTI (figure S6), respectively.

5.1. Benchmark model systems

The DeepTime implementation of the four-state Prinz potential and the three-well potential datasets was used to simulate the two benchmark systems [107]. The parameters used to simulate the trajectories are reported in the supporting information (table S1). The estimation of the dynAMMo were carried out as outlined in section 3. Each scenario uses different parametrizations of the potential and a table with an overview is listed in the supporting information (table S2). Chapman–Kolmogorov tests have been performed on all MSMs used in this study (supporting information figures S1 and S2).

5.2. BPTI

The estimation and analysis of the BPTI dynAMMo were conducted as described in section 3. Chapman–Kolmogorov tests were conducted on the MSMs used here to ensure validity of the models supporting information (figure S3). The estimation parameters of the two scenarios are listed in the supporting information (table S3). Further details are provided in supporting information, 'Materials and Methods'.

Data availability statement

No new data were created or analyzed in this study.

Acknowledgments

The authors would like to thank D E Shaw Research for sharing the BPTI simulation and Arthur G Palmer III, for sharing the raw NMR relaxation dispersion data. C K thanks Shanawaz Ahmed for fruitful discussions and for sharing a preliminary implementation of the Cayley transform for the eigenvector estimation. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program funded by the Knut and Alice Wallenberg Foundation.

Author contributions

C K and S O conceptualized, designed, and performed research as well as wrote the manuscript; C K analyzed and visualized the data and performed statistical analysis; S O provided supervision, project administration and funding acquisition.

ORCID iDs

Christopher Kolloff https://orcid.org/0000-0003-0304-3818
Simon Olsson https://orcid.org/0000-0002-3927-7897

References

[1] Arber W and Linn S 1969 DNA modification and restriction *Annu. Rev. Biochem.* **38** 467–500
[2] Antonini E and Brunori M 1970 Hemoglobin *Annu. Rev. Biochem.* **39** 977–1042
[3] Poretsky L and Kalin M F 1987 The gonadotropic function of insulin *Endocr. Rev.* **8** 132–41
[4] Wullschleger S, Loewith R and Hall M N 2006 TOR signaling in growth and metabolism *Cell* **124** 471–84
[5] Monod J, Wyman J and Changeux J P 1965 On the nature of allosteric transitions: a plausible model *J. Mol. Biol.* **12** 88–118
[6] Koshland D E, Némethy G and Filmer D 1966 Comparison of experimental binding data and theoretical models in proteins containing subunits *Biochemistry* **5** 365–85
[7] Sanger F 1958 *Chemistry of insulin* (Nobel Lecture) (available at: www.nobelprize.org/uploads/2018/06/sanger-lecture.pdf)
[8] Ascenzi P, Bocedi A, Bolognesi M, Spallarossa A, Coletta M, De Cristofaro R and Menegatti E 2003 The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein *Curr. Protein Pept. Sci.* **4** 231–51
[9] Wlodawer A, Walter J, Huber R and Sjölin L 1984 Structure of bovine pancreatic trypsin inhibitor. Results of joint neutron and x-ray refinement of crystal form II *J. Mol. Biol.* **180** 301–29
[10] Wagner G, DeMarco A and Wüthrich K 1976 Dynamics of the aromatic amino acid residues in the globular conformation of the basic pancreatic trypsin inhibitor (BPTI) *Biophys. Struct. Mech.* **2** 139–58
[11] Berndt K D, Güntert P, Orbons L P and Wüthrich K 1992 Determination of a high-quality nuclear magnetic resonance solution structure of the bovine pancreatic trypsin inhibitor and comparison with three crystal structures *J. Mol. Biol.* **227** 757–75
[12] Peng J W and Wagner G 1994 [20] Investigation of protein motions via relaxation measurements *Nuclear Magnetic Resonance, Part C* (*Methods in Enzymology* vol 239) (Academic) pp 563–96
[13] Smith P E, van Schaik R C, Szyperski T, Wüthrich K and van Gunsteren W F 1995 Internal mobility of the basic pancreatic trypsin inhibitor in solution: a comparison of NMR spin relaxation measurements and molecular dynamics simulations *J. Mol. Biol.* **246** 356–65
[14] van der Spoel D, van Buuren A R, Tieleman D P and Berendsen H J C 1996 Molecular dynamics simulations of peptides from BPTI: a closer look at amide-aromatic interactions *J. Biomol. NMR* **8** 229–38
[15] Daggett V and Levitt M 1992 A model of the molten globule state from molecular dynamics simulations *Proc. Natl Acad. Sci. USA* **89** 5142–6
[16] Shaw D E, Maragakis P, Lindorff-larsen K, Piana S, Shan Y and Wriggers W 2010 Atomic-level characterization of the structural dynamics of proteins *Science* **330** 341–7
[17] Grimaldo M, Roosen-Runge F, Zhang F, Schreiber F and Seydel T 2019 Dynamics of proteins in solution *Q. Rev. Biophys.* **52** 1–63
[18] Trott O and Palmer A G 2004 Theoretical study of R1ρ rotating-frame and R2 free-precession relaxation in the presence of n-site chemical exchange *J. Magn. Reson.* **170** 104–12
[19] Koss H, Rance M and Palmer A G 2017 General expressions for R1ρ relaxation for N-site chemical exchange and the special case of linear chains *J. Magn. Reson.* **274** 36–45
[20] Palmer A G and Massi F 2006 Characterization of the dynamics of biomacromolecules using rotating-frame spin relaxation NMR spectroscopy *Chem. Rev.* **106** 1700–19

[21] Lindner B, Yi Z, Prinz J H, Smith J C and Noé F 2013 Dynamic neutron scattering from conformational dynamics. I. Theory and Markov models *J. Chem. Phys.* **139** 175101
[22] Möller J, Sprung M, Madsen A and Gutt C 2019 X-ray photon correlation spectroscopy of protein dynamics at nearly diffraction-limited storage rings *IUCrJ* **6** 794–803
[23] Hiller S 2019 Chaperone-bound clients: the importance of being dynamic *Trends Biochem. Sci.* **44** 517–27
[24] Schiffrin B, Calabrese A N, Devine P W, Harris S A, Ashcroft A E, Brockwell D J and Radford S E 2016 Skp is a multivalent chaperone of outer-membrane proteins *Nat. Struct. Mol. Biol.* **23** 786–93
[25] Burmann B M, Wang C and Hiller S 2013 Conformation and dynamics of the periplasmic membrane-protein-chaperone complexes OmpX-Skp and tOmpA-Skp *Nat. Struct. Mol. Biol.* **20** 1265–72
[26] Thoma J, Burmann B M, Hiller S and Müller D J 2015 Impact of holdase chaperones Skp and SurA on the folding of β-barrel outer-membrane proteins *Nat. Struct. Mol. Biol.* **22** 795–802
[27] Gauto D F, Macek P, Malinverni D, Fraga H, Paloni M, Sučec I, Hessel A, Bustamante J P, Barducci A and Schanda P 2022 Functional control of a 0.5 MDa TET aminopeptidase by a flexible loop revealed by MAS NMR *Nat. Commun.* **13** 1927
[28] Neudecker P, Robustelli P, Cavalli A, Walsh P, Lundström P, Zarrine-Afsar A, Sharpe S, Vendruscolo M and Kay L E 2012 Structure of an intermediate state *Science* **336** 362
[29] Kruse A C *et al* 2012 Structure and dynamics of the M3 muscarinic acetylcholine receptor *Nature* **482** 552–6
[30] Rosenbaum D M *et al* 2011 Structure and function of an irreversible agonist-β2 adrenoceptor complex *Nature* **469** 236–42
[31] Lindorff-Larsen K, Piana S, Dror R O and Shaw D E 2011 How fast-folding proteins fold *Science* **334** 517–20
[32] Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, De Groot B L, Grubmüller H and MacKerell A D 2016 CHARMM36m: an improved force field for folded and intrinsically disordered proteins *Nat. Methods* **14** 71–73
[33] Lindorff-Larsen K, Piana S, Palmo K, Maragakis P, Klepeis J L, Dror R O and Shaw D E 2010 Improved side-chain torsion potentials for the Amber ff99SB protein force field *Proteins Struct. Funct. Bioinform.* **78** 1950–8
[34] Smith J S, Isayev O and Roitberg A E 2017 ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost *Chem. Sci.* **8** 3192–203
[35] Shaw D E *et al* 2009 Millisecond-scale molecular dynamics simulations on Anton *Proc. Conf. on High Performance Computing Networking, Storage and Analysis* pp 1–11
[36] Shaw D E *et al* 2014 Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer *Int. Conf. for High Performance Computing, Networking, Storage and Analysis, SC* pp 41–53
[37] Shaw D E *et al* 2021 Anton 3: twenty microseconds of molecular dynamics simulation before lunch *Int. Conf. for High Performance Computing, Networking, Storage and Analysis, SC* vol 1
[38] Voelz V A, Pande V S and Bowman G R 2023 Folding@home: achievements from over 20 years of citizen science herald the exascale era *Biophys. J.* **122** 1–12
[39] Bowman G R, Beauchamp K A, Boxer G and Pande V S 2009 Progress and challenges in the automated construction of Markov state models for full protein systems *J. Chem. Phys.* **131** 124101
[40] Bowman G R, Voelz V A and Pande V S 2011 Atomistic folding simulations of the five-helix bundle protein λ6-85 *J. Am. Chem. Soc.* **133** 664–7
[41] Lane T J, Bowman G R, Beauchamp K, Voelz V A and Pande V S 2011 Markov state model reveals folding and functional dynamics in ultra-long MD trajectories *J. Am. Chem. Soc.* **133** 18413–9
[42] Voelz V A *et al* 2012 Slow unfolded-state structuring in Acyl-CoA binding protein folding revealed by simulation and experiment *J. Am. Chem. Soc.* **134** 12565–77
[43] Prinz J-H, Wu H, Sarich M, Keller B, Senne M, Held M, Chodera J D, Schütte C and Noé F 2011 Markov models of molecular kinetics: generation and validation *J. Chem. Phys.* **134** 174105
[44] Kenkre V M, Montroll E W and Shlesinger M F 1973 Generalized master equations for continuous-time random walks *J. Stat. Phys.* **9** 45–50
[45] Montroll E W and Liebowitz J L 1962 *Studies in Statistical Mechanics* (Wiley) (https://doi.org/10.1002/bbpc.19830870527)
[46] Zwanzig R 1983 From classical dynamics to continuous time random walks *J. Stat. Phys.* **30** 255–62
[47] Resibois P 1963 On the equivalence between two generalized master equations *Physica* **29** 721–41
[48] Nicolis G and Nicolis C 1988 Master-equation approach to deterministic chaos *Phys. Rev.* A **38** 427–33
[49] Schütte C, Fischer A, Huisinga W and Deuflhard P 1999 A direct approach to conformational dynamics based on hybrid Monte Carlo *J. Comput. Phys.* **151** 146–68
[50] Swope W C, Pitera J W and Suits F 2004 Describing protein folding kinetics by molecular dynamics simulations. 1. Theory$^{+}_{-}$ *J. Phys. Chem.* B **108** 6571–81
[51] Wassman C D *et al* 2013 Computational identification of a transiently open L1/S3 pocket for reactivation of mutant p53 *Nat. Commun.* **4** 1–9
[52] Noé F, Schütte C, Vanden-Eijnden E, Reich L and Weikl T R 2009 Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations *Proc. Natl Acad. Sci. USA* **106** 19011–6
[53] Qiao Q, Bowman G R and Huang X 2013 Dynamics of an intrinsically disordered protein reveal metastable conformations that potentially seed aggregation *J. Am. Chem. Soc.* **135** 16092–101
[54] Raich L, Meier K, Günther J, Christ C D, Noé F and Olsson S 2021 Discovery of a hidden transient state in all bromodomain families *Proc. Natl Acad. Sci.* **118** e2017427118
[55] Chakrabarti K S *et al* 2022 A litmus test for classifying recognition mechanisms of transiently binding proteins *Nat. Commun.* **13** 3792
[56] Liebl K and Zacharias M 2023 The development of nucleic acids force fields: from an unchallenged past to a competitive future *Biophys. J.* **122** 1–11
[57] Tan D, Piana S, Dirks R M and Shaw D E 2018 RNA force field with accuracy comparable to state-of-the-art protein force fields *Proc. Natl Acad. Sci. USA* **115** E1346–55
[58] Lindorff-Larsen K, Maragakis P, Piana S, Eastwood M P, Dror R O and Shaw D E 2012 Systematic validation of protein force fields against experimental data *PLoS One* **7** 1–6
[59] 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
[60] Henriques J, Cragnell C and Skepö M 2015 Molecular dynamics simulations of intrinsically disordered proteins: force field evaluation and comparison with experiment *J. Chem. Theory Comput.* **11** 3420–31

[61] Robustelli P, Piana S and Shaw D E 2018 Developing a molecular dynamics force field for both folded and disordered protein states *Proc. Natl Acad. Sci. USA* **115** E4758–66
[62] Dellago C, Bolhuis P G, Csajka F S and Chandler D 1998 Transition path sampling and the calculation of rate constants *J. Chem. Phys.* **108** 1964–77
[63] Wales D J 2002 Discrete path sampling *Mol. Phys.* **100** 3285–305
[64] Moroni D, van Erp T S and Bolhuis P G 2004 Investigating rare events by transition interface sampling *Physica* A **340** 395–401
[65] Faradjian A K and Elber R 2004 Computing time scales from reaction coordinates by milestoning *J. Chem. Phys.* **120** 10880–9
[66] Laio A and Parrinello M 2002 Escaping free-energy minima *Proc. Natl Acad. Sci.* **99** 12562–6
[67] Grubmüller H 1995 Predicting slaw structural transitions in macromolecular systems: conformational flooding Helmut *Phys. Rev.* E **52** 2893
[68] Swendsen R H and Wang J S 1986 Replica Monte Carlo simulation of spin-glasses *Phys. Rev. Lett.* **57** 2607–9
[69] Pasarkar A P, Bencomo G M, Olsson S and Dieng A B 2023 Vendi sampling for molecular simulations: diversity as a force for faster convergence and better exploration *J. Chem. Phys.* **159** 10
[70] Noé F, Olsson S, Köhler J and Wu H 2019 Boltzmann generators: sampling equilibrium states of many-body systems with deep learning *Science* **365** 1–11
[71] Schreiner M, Winther O and Olsson S 2023 Implicit transfer operator learning: multiple time-resolution surrogates for molecular dynamics *Advances in Neural Information Processing Systems*
[72] Bottaro S and Lindorff-Larsen K 2018 Biophysical experiments and biomolecular simulations: a perfect match? *Science* **361** 355–60
[73] Leung H T A, Bignucolo O, Aregger R, Dames S A, Mazur A, Bernèche S and Grzesiek S 2016 A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content *J. Chem. Theory Comput.* **12** 383–94
[74] Capelli R, Tiana G and Camilloni C 2018 An implementation of the maximum-caliber principle by replica-averaged time-resolved restrained simulations *J. Chem. Phys.* **148** 184114
[75] Smith C A, Mazur A, Rout A K, Becker S, Lee D, de Groot B L and Griesinger C 2020 Enhancing NMR derived ensembles with kinetics on multiple timescales *J. Biomol. NMR* **74** 27–43
[76] Cavalli A, Camilloni C and Vendruscolo M 2013 Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle *J. Chem. Phys.* **138** 094112
[77] Boomsma W, Ferkinghoff-Borg J and Lindorff-Larsen K 2014 Combining experiments and simulations using the maximum entropy principle *PLoS Comput. Biol.* **10** 1–9
[78] Olsson S, Strotz D, Vögeli B, Riek R and Cavalli A 2016 The dynamic basis for signal propagation in human Pin1-WW *Structure* **24** 1464–75
[79] Beauchamp K A, Pande V S and Das R 2014 Bayesian energy landscape tilting: towards concordant models of molecular ensembles *Biophys. J.* **106** 1381–90
[80] Pitera J W and Chodera J D 2012 On the use of experimental observations to bias simulated ensembles *J. Chem. Theory Comput.* **8** 3445–51
[81] Bonomi M, Camilloni C, Cavalli A and Vendruscolo M 2016 Metainference: a Bayesian inference method for heterogeneous systems *Sci. Adv.* **2** 1–9
[82] Hummer G and Köfinger J 2015 Bayesian ensemble refinement by replica simulations and reweighting *J. Chem. Phys.* **143** 243150
[83] Olsson S, Frellsen J, Boomsma W, Mardia K V and Hamelryck T 2013 Inference of structure ensembles of flexible biomolecules from sparse, averaged data *PLoS One* **8** 1–7
[84] Lindorff-Larsen K, Best R B, DePristo M A, Dobson C M and Vendruscolo M 2005 Simultaneous determination of protein structure and dynamics using cryo-electron microscopy *Nature* **433** 128–32
[85] White A D and Voth G A 2014 Efficient and minimal method to bias molecular simulations with experimental data *J. Chem. Theory Comput.* **10** 3023–30
[86] Faidon Brotzakis Z, Vendruscolo M and Bolhuis P G 2021 A method of incorporating rate constants as kinetic constraints in molecular dynamics simulations *Proc. Natl Acad. Sci. USA* **118** e2012423118
[87] Rudzinski J F, Kremer K and Bereau T 2016 Communication: consistent interpretation of molecular simulation kinetics using Markov state models biased with external information *J. Chem. Phys.* **144** 051102
[88] Olsson S, Wu H, Paul F, Clementi C and Noé F 2017 Combining experimental and simulation data of molecular processes via augmented Markov models *Proc. Natl Acad. Sci. USA* **114** 8265–70
[89] Meiboom S and Gill D 1958 Modified spin-echo method for measuring nuclear relaxation times *Rev. Sci. Instrum.* **28** 688–90
[90] Luz Z and Meiboom S 1963 Nuclear magnetic resonance study of the protolysis of trimethylammonium ion in aqueous solution-order of the reaction with respect to solvent *J. Chem. Phys.* **39** 366–70
[91] Olsson S and Noé F 2017 Mechanistic models of chemical exchange induced relaxation in protein NMR *J. Am. Chem. Soc.* **139** 200–10
[92] Xue Y, Ward J M, Yuwen T, Podkorytov I S and Skrynnikov N R 2012 Microsecond time-scale conformational exchange in proteins: using long molecular dynamics trajectory to simulate NMR relaxation dispersion data *J. Am. Chem. Soc.* **134** 2555–62
[93] Wen Z and Yin W 2013 A feasible method for optimization with orthogonality constraints *Math. Program.* **142** 397–434
[94] Millet O, Loria J P, Kroenke C D, Pons M and Palmer A G 2000 The static magnetic field dependence of chemical exchange linebroadening defines the NMR chemical shift time scale *J. Am. Chem. Soc.* **122** 2867–77
[95] Grey M J, Wang C and Palmer A G 2003 Disulfide bond isomerization in basic pancreatic trypsin inhibitor: multisite chemical exchange quantified by CPMG relaxation dispersion and chemical shift modeling *J. Am. Chem. Soc.* **125** 14324–35
[96] Massi F, Johnson E, Wang C, Rance M and Palmer A G 2004 NMR R1ρ rotating-frame relaxation with weak radio frequency fields *J. Am. Chem. Soc.* **126** 2247–56
[97] Weininger U, Brath U, Modig K, Teilum K and Akke M 2014 Off-resonance rotating-frame relaxation dispersion experiment for ^{13}C in aromatic side chains using L-optimized TROSY-selection *J. Biomol. NMR* **59** 23–29
[98] Denisov V P and Halle B 1995 Protein hydration dynamics in aqueous solution: a comparison of bovine pancreatic trypsin inhibitor and ubiquitin by oxygen-17 spin relaxation dispersion *J. Mol. Biol.* **245** 682–97
[99] Brooks B and Karplus M 1983 Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor *Proc. Natl Acad. Sci. USA* **80** 6571–5
[100] Wagner G 1983 Characterization of the distribution of internal motions in the basic pancreatic trypsin inhibitor using a large number of internal NMR probes *Q. Rev. Biophys.* **16** 1–57

[101] Wagner G, Müller N, Wüthrich K, Bodenhausen G, Sorensen O W, Ernst R R and Rance M 1985 Exchange of two-spin order in nuclear magnetic resonance: separation of exchange and cross-relaxation processes *J. Am. Chem. Soc.* **107** 6440–6
[102] Noé F and Clementi C 2015 Kinetic distance and kinetic maps from molecular dynamics simulation *J. Chem. Theory Comput.* **11** 5002–11
[103] Scherer M K, Trendelkamp-Schroer B, Paul F, Pérez-Hernández G, Hoffmann M, Plattner N, Wehmeyer C, Prinz J H and Noé F 2015 PyEMMA 2: a software package for estimation, validation and analysis of Markov models *J. Chem. Theory Comput.* **11** 5525–42
[104] Pérez-Hernández G, Paul F, Giorgino T, De Fabritiis G and Noé F 2013 Identification of slow molecular order parameters for Markov model construction *J. Chem. Phys.* **139** 015102
[105] Li D W and Brüschweiler R 2012 PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles *J. Biomol. NMR* **54** 257–65
[106] Paszke A *et al* 2019 PyTorch: an imperative style, high-performance deep learning library *Advances in Neural Information Processing Systems* vol 32
[107] Hoffmann M *et al* 2022 Deeptime: a Python library for machine learning dynamical models from time series data *Mach. Learn.: Sci. Technol.* **3** 015009
[108] Schrödinger L 2015 The PyMOL molecular graphics system, version 2.0 (available at: https://pymol.org/) (Retrieved 4 January 2021)
[109] Hunter J D 2007 Matplotlib: a 2D graphics environment *Comput. Sci. Eng.* **9** 90–95