From sequence to Boltzmann weighted ensemble of structures with AlphaFold2-RAVE

Bodhi P. Vani,[1, a)] Akashnathan Aranganathan,[2, a)] Dedi Wang,[2] and Pratyush Tiwary[3, b)]
[1)] *Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20740, USA*
[2)] *Biophysics Program and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA*
[3)] *Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA*

(Dated: May 2022)

In the short time since it has appeared, AlphaFold2 (AF2) has been widely adopted as a new standard in accurate and fast protein structure prediction starting from any arbitrary sequence of amino acids. However, AF2 maps a single sequence to a single structure, and even with recently proposed modifications that add conformational diversity, it is arguably devoid of thermodynamics. In this working paper we demonstrate an efficient protocol that uses the structural diversity from AF2 as a starting point to perform Artificial Intelligence augmented enhanced molecular dynamics simulations. Specifically we use the "Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE)" method as post-processing on AF2, and thus the protocol shown here is called AlphaFold2-RAVE. These simulations expand upon the results from AF2 ranking them as per their correct Boltzmann weights. This schema for going from sequence to Boltzmann weighted ensemble of structures is demonstrated here for a small cold-shock protein, and will be expanded to include many more sequences together with an easy-to-use open-source code.

While traditionally protein secondary and tertiary structure prediction has relied on experimental techniques such as cryo-EM, x-ray crystallography and NMR studies, 2021 saw a change in the status quo with AlphaFold2 (AF2)[1]. It has surpassed the accuracy of other prediction models and offers a seemingly robust tool for structure prediction directly from the sequence of constituent amino acids. AF2 relies largely on the idea that the conservation of residues across evolutionary protein sequences is likely to be correlated with three-dimensional euclidean distances. Building on this idea, AF2 generates a multiple sequence alignment (MSA) of evolutionarily related sequences, which helps identify residues that have co-evolved thereby facilitating structure prediction. While AF2 indeed represents a significant change in the field of structural biology, there are a few key limitations of AF2 that have been recently described in the literature[2–4], with no clear solution to these problems yet.

In this working paper, we first summarize these central limitations with AF2, and then show how these can be surmounted with the use of Artificial Intelligence (AI) augmented Molecular Dynamics (MD) methods[5]. We emphasize that this is a work in progress which will be supplemented as we add more proteins demonstrating the strength and generalizability of our protocol.

There are at least two central limitations to AF2 which we address here. The first limitation is that in its original most-cited incarnation, AF2 is a single structure prediction method for any given sequence. This problem was partly solved with the simple realization that reducing the size of the input MSA increases the conformational diversity explored in AF2[3,6]. This procedure however does not provide any notion of relative probabilities of these alternate conformations, and many of them could in fact be physically improbable. Being able to assign Boltzmann weights would instantly lead to ruling out unphysical models generated by reducing the MSA length as well as rank them as per their thermodynamic propensities. The second limitation is that AF2 fails in predicting changes in protein structure due to missense mutations[2]. Here as well one would expect that reducing the MSA length could give glimpses into the possible alternate conformations for a mutated protein, but once again without Boltzmann weights.

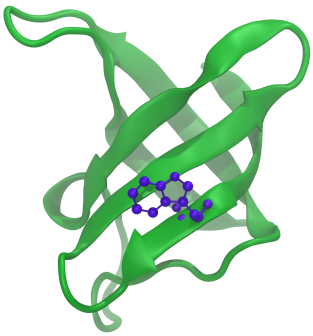

FIG. 1. Cold shock protein 1HZB with residue Trp8 highlighted in its most populated configuration.

In principle, long-enough MD simulations could directly help characterize the thermodynamics of the diverse conformations generated for any given protein through reducing MSA length with AF2, thereby sur-

[a)] These two authors contributed equally.
[b)] Corresponding author. Email: ptiwary@umd.edu

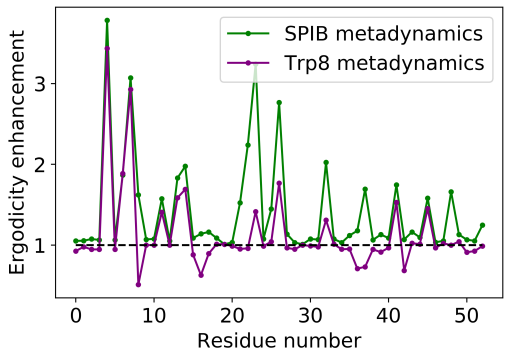

FIG. 2. Ergodicity enhancement in CSP residues. To examine the improvement in sampling of sidechain configurations, we compute the volume sampled in χ spaces for each residue and plot this for metadynamics performed by (i) directly biasing the sidechain dihedrals χ_1 and χ_2 of the Trp8 sidechains (purple) and (ii) biasing along the information bottleneck our methodology obtains (green). Both volumes are normalized by volume sampled by unbiased trajectories of the same length.

mounting both of the limitations above. However, MD simulations, even with the best available supercomputers, are limited in the accessible timescales, barely reaching microseconds with days or weeks of simulation, while events of interest, such as conformational changes, most often occur at timescales ranging from microseconds to seconds. Without observing multiple back-and-forth transitions between possible competing conformations, MD can not assign Boltzmann weights to them. Observing these conformational changes in MD is computationally intractable, and gathering statistically significant data to predict energetics even more so. This has given rise to a rich class of methods known as enhanced sampling algorithms[7], with the goal of driving sampling to rare regions of configurational space while retaining the ability to reweight and extract accurate statistics in a computationally efficient manner. However, often such enhanced sampling methods require prior mechanistic knowledge of the slow conformation changes of interest, also known as the reaction coordinate[8].

In this work, we propose the AlphaFold2-RAVE scheme as a first step towards solving this problem. In this scheme we supplement AlphaFold2 with a post-processing protocol using the machine learning method "State Predictive Information Bottleneck (SPIB)"[9]. SPIB belongs to the "Reweighted Autoencoded Variational Bayes for Enhanced Sampling (RAVE)" family of methods[10,11]. SPIB, and more generally RAVE, uses an autoencoder-themed AI framework to learn the reaction coordinate from limited sampling in an iterative manner (see Methods) wherein every iteration of MD is biased to enhance fluctuations along the learned reaction coordinate. The reaction coordinate is expressed as a past-future information bottleneck, i.e. the most parsimonious lower dimensional manifold for embedding the dynamics such that the future state of the system can be predicted as accurately as possible. The states are learnt by the decoder, but are representative of physically relevant state such as, for example a protein's conformation. How far into the future one is aiming to predict is known as the time lag, and is an important parameter in SPIB (see Methods).

The use of such a time lag allows one to account for the inherently dynamic personalities of proteins[12]. Reweighting procedures[13] are then used to obtain Boltz-

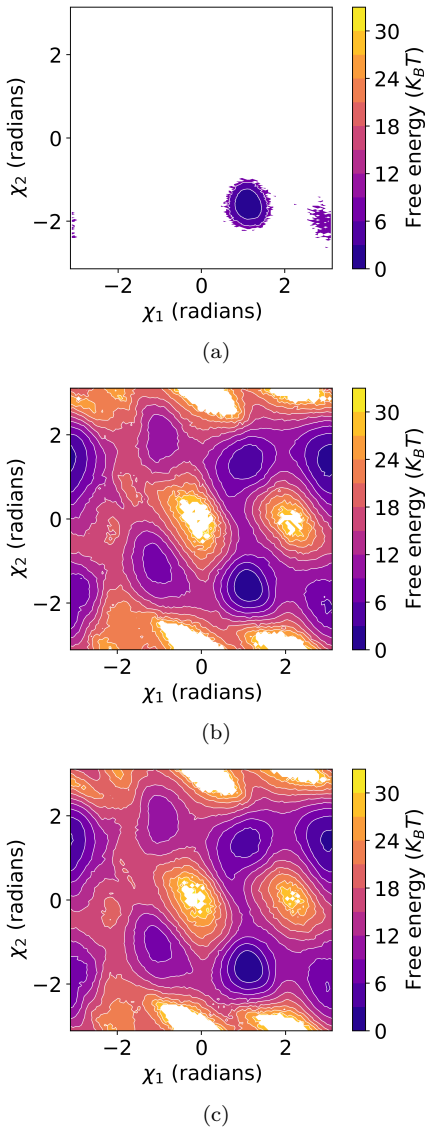

FIG. 3. PMFs along sidechain angles of the Trp8 residue which exhibits six metastable states. In (a) we show calculations from 20 ns long unbiased trajectory initialized from XRD structure. In (b) we show results of metadynamics performed on sidechain angles of the Trp8 residue. In (c) we show results from metadynamics performed using information bottleneck obtained without *a priori* knowledge of Trp8 metastability. Colorbars for all 3 figures are shown with energies in units of k_BT.

mann weights for different conformations sampled during the enhanced MD performed by enhancing fluctuations along the information bottleneck.

In this first benchmark test, we apply our approach on the small 66-residue cold shock protein (CSP; PDB ID: 1HZB), shown in Figure 1. We show that without using any *a priori* information regarding residues that are known to exhibit conformational metastability, we are able to obtain different competing conformations together with their accurate Boltzmann weights. Specifically, CSP has known rotameric metastability in its eighth residue (Trp8), exhibiting 6 metastable states in its χ_1 and χ_2 torsions as per fluorescence spectroscopy[14]. To demonstrate the power of our method, we assume no *a priori* knowledge of Trp8, or any other residues with metastable rotamers, and at every stage of the algorithm, we use all available 125 protein χ torsional angles. Without using any preference towards the χ_1 and χ_2 dihedrals of Trp8, our protocol not only correctly ranks the 6 metastable states along these two dihedrals, but also further samples conformations along the other 52 residues with sidechains with increased ergodicity, as measured here by the fraction of dihedral configuration space explored and shown in Figure 2. Here, we see that our protocol not only samples the Trp8 residue conformations comparably to metadynamics directly on that residue, but also samples more extensively for all residues with sidechains, which directed metadynamics fails to do.

Specifically, our protocol begins with the MSA modification to AF2 through which we first obtain a conformationally diverse ensemble of structures, albeit without corresponding Boltzmann weights. This gives rise to an ensemble comprising 2560 structures. We choose a large set of order parameters, such as all sidechain dihedral angles (see Methods) to characterize this ensemble, and perform a preliminary spatial clustering using SPIB with a lag time of 0. Choosing representative structures initialized from different clusters, we run multiple short unbiased molecular dynamics trajectories. We then use these trajectories as inputs for SPIB once again, this time using the past-future information bottleneck with a finite time lag[9,15]. Finally, we run biased metadynamics simulations using the information bottleneck learnt from this iteration of SPIB. This final run can be used to directly identify different side chain conformations and obtain their equilibrium free energies by appropriate reweighting.

While Figure 2 shows enhanced exploration along all dihedrals, this is no guarantee that such sampling when reweighted to account for the bias will give us correct Boltzmann distributions. Next we demonstrate that indeed this is the case by considering Trp8. In Figure 3 we show the potentials of mean force (PMFs) along Trp8 χ angles obtained from unbiased and biased metadynamics simulations respectively. Figure 3(a) shows unbiased sampling for 250ns initialized from the NMR structure, where only two metastable basins are sampled. The trajectory even fails to sample the second of the two free energy minima. Figure 3(b) shows PMF obtained from metadynamics performed by biasing the sidechain angles of the Trp8 residue. Figure 3(c) shows results

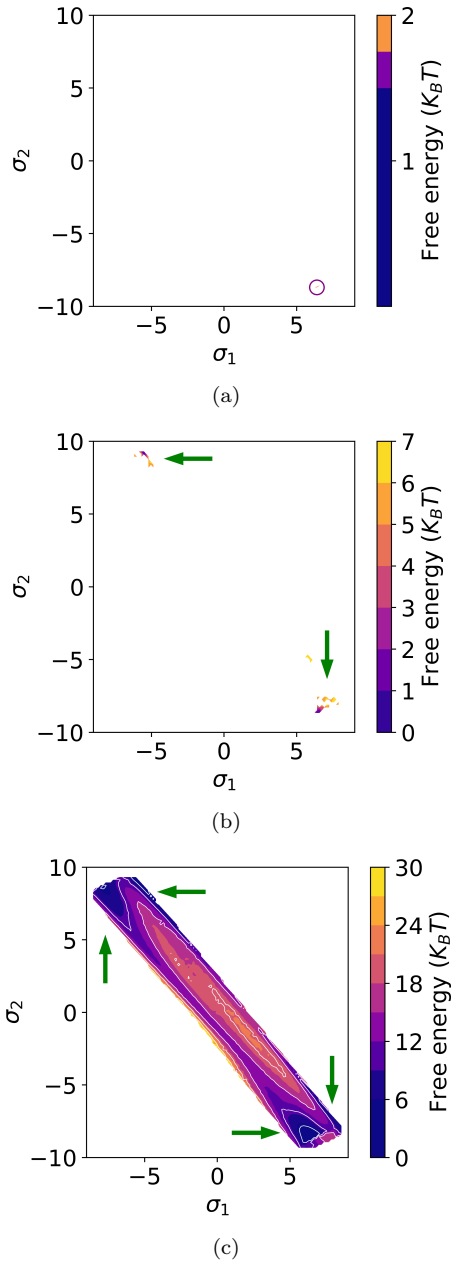

FIG. 4. PMFs projected along the two-dimensional information bottleneck denoted σ_1 and σ_2 learnt from our scheme. Projections using different methods are shown. In (a) we show projection in this space obtained from direct use of AF2, which does not show much diversity. In (b) we show the distribution with reduced MSA length which now shows some diversity, indicated with thick arrows, however without any notion of relative thermodynamic stability. In (c) we show the results from AlphaFold2-RAVE, where both conformational diversity and correct relative metastability can be seen. Respective colorbars indicate free energies measured as negative *log* of probabilities. The arrows indicate free energy basins identified on the PMFs.

from metadynamics performed by biasing the information bottleneck obtained without *a priori* knowledge of Trp8 metastability.

In Figure 4 we show different PMFs projected along the two-dimensional information bottleneck (labeled σ_1 and σ_2). Figure 4(a) shows that when simply using AF2 as is, no significant conformational diversity is obtained. Figure 4(b), obtained with AF2 after reducing MSA length, shows some conformational diversity but the weights are not correct. Figure 4(c) obtained after AlphaFold2-RAVE shows both richer conformational diversity as well as correct Boltzmann weights, which we validate in Figure 3 for Trp8.

To conclude, in this working paper we have proposed the AlphaFold2-RAVE protocol. This combines the strengths of AlphaFold2 with the all-atom resolution enhanced sampling powers of RAVE in a way that allows us to go from sequence to an ensemble of conformations ranked with their correct Boltzmann weights. We want to emphasize that perhaps instead of RAVE we could have in principle used some other enhanced sampling protocol. We prefer RAVE and specifically its SPIB variant due to the minimal amount of hand-tuning it requires in terms of pre-knowledge of the reaction coordinate for driving the enhanced sampling, or the number of metastable states in the system. In a similar vein, instead of AlphaFold2 we could have used some other experimental[16] or even computational[17,18] approach to generate an initial dictionary of possible competing conformations. We choose AlphaFold2 because it is easy and relatively quick to use, with higher accuracy than other computational predictions. In this first working paper, we demonstrated the protocol on the small 66-residue cold shock protein and showed how we could obtain correct thermodynamic sampling, despite restricting ourselves to no prior information, and in fact demonstrating higher generality. We will be complementing this working paper gradually with many more systems and also release a complete open-source code to be hosted at `github.com/tiwarylab`. Individual components of the code can already be found at `github.com/sokrypton/ColabFold` and `github.com/tiwarylab/State-Predictive-Information-Bottleneck`.

Acknowledgments

Research reported in this publication was supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number R35GM142719. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Deepthought2, MARCC, and XSEDE[19] (projects CHE180007P and CHE180027P) for providing computational resources used in this work.

Notes

The authors declare the following competing financial interest: P.T. is a consultant to Schrodinger, Inc.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021.

[2] Gwen R. Buel and Kylie J. Walters. Can alphafold2 predict the impact of missense mutations on structure? *Nature Structural & Molecular Biology*, 29(1):1–2, Jan 2022.

[3] Diego del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with alphafold2. *eLife*, 11:e75751, mar 2022.

[4] Anowarul Kabir, Toki Inan, and Amarda Shehu. Analysis of alphafold2 for modeling structures of wildtype and variant protein sequences. In *Proceedings of 14th International Conference*, volume 83, pages 53–65, 2022.

[5] Yihang Wang, Joao Marcelo Lamim Ribeiro, and Pratyush Tiwary. Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Current Opinion in Structural Biology*, 61:139–145, 2020.

[6] Avner Schlessinger and Massimiliano Bonomi. Artificial intelligence: Exploring the conformational diversity of proteins. *eLife*, 11:e78549, apr 2022.

[7] Pratyush Tiwary and Axel van de Walle. A review of enhanced sampling approaches for accelerated molecular dynamics. *Multiscale materials modeling for nanomechanics*, pages 195–221, 2016.

[8] Baron Peters. Chapter 20 - reaction coordinates and mechanisms. In Baron Peters, editor, *Reaction Rate Theory and Rare Events Simulations*, pages 539–571. Elsevier, Amsterdam, 2017.

[9] Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154(13):134111, 2021.

[10] João Marcelo Lamim Ribeiro, Pablo Bravo, Yihang Wang, and Pratyush Tiwary. Reweighted autoencoded variational bayes for enhanced sampling (rave). *The Journal of chemical physics*, 149(7):072301, 2018.

[11] Yihang Wang, João Marcelo Lamim Ribeiro, and Pratyush Tiwary. Past–future information bottleneck for sampling molecular reaction coordinate simultaneously with thermodynamics and kinetics. *Nature Communications*, 10(1):3573, Aug 2019.

[12] Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450(7172):964–972, 2007.

[13] Pratyush Tiwary and Michele Parrinello. A time-independent free energy estimator for metadynamics. *The Journal of Physical Chemistry B*, 119(3):736–742, Jan 2015.

[14] Samuel L. C. Moors, Mario Hellings, Marc De Maeyer, Yves Engelborghs, and Arnout Ceulemans. Tryptophan rotamers as evidenced by x-ray, fluorescence lifetimes, and molecular dynamics modeling. *Biophysical journal*, 91(3):816–823, Aug 2006.

[15] Shams Mehdi, Dedi Wang, Shashank Pant, and Pratyush Tiwary. Accelerating all-atom simulations and gaining mechanistic understanding of biophysical systems through state predictive information bottleneck. *Journal of Chemical Theory and Computation*, 18(5):3231–3238, May 2022.

[16] Ka Man Yip, Niels Fischer, Elham Paknia, Ashwin Chari, and Holger Stark. Atomic-resolution protein structure determination by cryo-em. *Nature*, 587(7832):157–161, 2020.

[17] Yinglong Miao and J Andrew McCammon. Gaussian accelerated molecular dynamics: Theory, implementation, and applications.

In *Annual reports in computational chemistry*, volume 13, pages 231–278. Elsevier, 2017.

[18] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[19] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, and Nancy Wilkins-Diehr. Xsede: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74, 2014.

[20] Milot Mirdita, Sergey Ovchinnikov, and Martin Steinegger. Colabfold - making protein folding accessible to all. *bioRxiv*, 2021.

[21] Omar Valsson, Pratyush Tiwary, and Michele Parrinello. Enhancing important fluctuations: Rare events and metadynamics from a conceptual viewpoint. *Annual review of physical chemistry*, 67:159–184, 2016.

[22] Giovanni Bussi and Alessandro Laio. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics*, 2(4):200–212, 2020.

[23] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of computational chemistry*, 24(16):1999–2012, 2003.

[24] Mark James Abraham, Teemu Murtola, Roland Schulz, Szilárd Páll, Jeremy C. Smith, Berk Hess, and Erik Lindahl. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25, 2015.

[25] Berk Hess, Henk Bekker, Herman JC Berendsen, and Johannes GEM Fraaije. Lincs: a linear constraint solver for molecular simulations. *Journal of computational chemistry*, 18(12):1463–1472, 1997.

[26] Massimiliano Bonomi, Davide Branduardi, Giovanni Bussi, Carlo Camilloni, Davide Provasi, Paolo Raiteri, Davide Donadio, Fabrizio Marinelli, Fabio Pietrucci, Ricardo A. Broglia, and Michele Parrinello. PLUMED: A portable plugin for free-energy calculations with molecular dynamics. *Computer Physics Communications*, 180(10):1961–1972, 2009.

[27] Gareth A. Tribello, Massimiliano Bonomi, Davide Branduardi, Carlo Camilloni, and Giovanni Bussi. PLUMED 2: New feathers for an old bird. *Computer Physics Communications*, 185(2):604–613, 2014.

[28] Massimiliano Bonomi, Giovanni Bussi, Carlo Camilloni, Gareth A. Tribello, Pavel Banáš, Alessandro Barducci, Mattia Bernetti, Peter G. Bolhuis, Sandro Bottaro, Davide Branduardi, Riccardo Capelli, and Paolo Carloni. Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods*, 16(8):670–673, aug 2019.

METHODS

A. AlphaFold2

To generate a diverse structural ensemble, we made a key change to the publicly available Colabfold notebook[20]. The input featurization for the AF2 algorithm is in the form of two MSA clusters. Namely, the first is used as the input to the main AF2 network, and the second is used to initialize the pair representation matrix, and also used as an input. We decrease the size of both MSA clusters to (16, 32), and (8, 16) respectively. As also reported in Ref. 3, we find that restricting the information in this featurization leads to revealing structures that are significantly different from the crystal structure, while providing hints to metastability. The other computational choices and parameters are as follows. We use the model dropout feature, set number of recycles to 1, and use 128 random seeds– since AF2 generates 5 structures per random seed, we obtain 640 structures for each set of MSA hyperparameters.

B. Metadynamics

Algorithms developed to enhance sampling can be characterized in two ways: driving sampling by adding "biases" in the form of energetic potentials, or by splitting and stratification methods that focus on simulating multiple trajectory fragments and resampling them with biased probabilities[7]. In practice, the former is found to be preferable for enthalpic barriers, while the latter is more effective for entropic barriers. Our scheme is in theory agnostic to the sampling method used, as long as it is possible to recover unbiased statistics to obtain potentials of mean forces using the method. Here, we choose to use an easily implementable and relatively fast method, metadynamics[21]. Metadynamics functions by depositing Gaussian biases intermittently along the simulation so that regions of the collective variable space that have been traversed become less probable the more they are sampled, and the system is forced to sample rarer regions. This results in a time dependant bias that is recorded and can be used to generate a PMF along the collective variables used for bias. Additionally, it has been shown that an on-the-fly bias can be computed independently[13] to calculate a weighted histogram along any arbitrary collective variable. However, in practice the choice of collective variable is difficult and essential. Biasing irrelevant collective variables or missing crucial slow collective variables[22] could lead to no motion away from initial metastable state, or to sampling unphysical regions of configurational space.

C. State predictive information bottleneck

Here, we use SPIB iteratively with biased or enhanced dynamics to learn the reaction coordinate[9]. SPIB is essentially a past-future information bottleneck protocol which takes time resolved trajectory data in a high dimensional order parameter space. It then predicts a latent space of desired dimensionality and identifies states in this space. Here we used a 2-dimensional information bottleneck which we denote $\{\sigma_1, \sigma_2\}$. The protocol

requires an initial identification of state labels which is then refined iteratively. The model learnt through SPIB consists of an encoder, which transforms the high dimensional input into a latent space, and a decoder, which uses this latent space to predict state labels after a pre-set time-lag. Ideally, the latent space identified as the information bottleneck would correspond with our traditional understanding of reaction coordinates. This was demonstrated to be true for a model system in 9, where the latent space corresponds very well with the committor as defined in transition path theory. The information bottleneck is found by optimizing a loss function that maximizes the latent space's ability to predict the state of the system after a pre-set time-lag, while reducing the input dimensionality to the given bottleneck dimensionality. SPIB is capable of using both a linear or a non-linear encoder. In this work, we use a linear encoder with a non-linear decoder. For CSP, the encoder was a 2-d linear combination of 125 order parameters arising from the sidechain dihedral angles χs of all 66 residues. We have found that while there are regions of state space that have zero sampling, and we aim to push the simulation into those regions, a linear encoder is far superior as a non-linear encoder will often overfit to sampled regions while producing unphysical results in unsampled regions.

D. Other simulation details

The protein is represented by the AMBER03 force field[23]. The simulations are performed at 300 K with the leap-frog integrator in GROMACS 5.1.4[24]; LINCS was used to constrain the lengths of bonds to hydrogen atoms[25]; the step size was 2 fs. We used PLUMED 2.4[26–28] to extract collective variables.