RESEARCH ARTICLE SUMMARY

PROTEIN SIMULATIONS

# Scalable emulation of protein equilibrium ensembles with generative deep learning

Sarah Lewis†, Tim Hempel†, José Jiménez-Luna†, Michael Gastegger†, Yu Xie†, Andrew Y. K. Foong†, Victor García Satorras†, Osama Abdin†, Bastiaan S. Veeling†, *et al.*

Full article and list of author affiliations: https://doi.org/10.1126/science.adv9817

**INTRODUCTION:** Proteins constitute the functional building blocks of life and are central to drug discovery and biotechnology. We now have technologies to determine protein sequence and predict protein structure at the genomic scale, but this is not the case for protein function. Protein function relies on dynamical mechanisms, particularly the transitions between long-lived protein structures (conformational states) and the association with and dissociation from other proteins and ligands (compositional states). The coupling between conformational and compositional state changes, and the probability of these states under a given set of conditions (temperature, solvation, concentration), determine "how proteins work" on a molecular scale. Although biophysical experiments and molecular dynamics (MD) simulations can reveal such structure-dynamics relationships with high accuracy, these methods suffer from low throughput.

**RATIONALE:** As a step toward solving this throughput challenge, we developed a biomolecular emulator (BioEmu) that samples the approximate equilibrium distribution of structures of single protein chains. BioEmu is a generative deep-learning system that can generate thousands of statistically independent structure samples per hour on a single graphics processing unit (GPU). BioEmu leverages AlphaFold to encode the protein sequence into a rich sequence-structure representation, which inputs into a diffusion model that efficiently samples three-dimensional structures. BioEmu was trained in three stages: It was pretrained on a processed version of the AlphaFold database (AFDB) in such a way as to incentivize the model to associate each protein sequence with a diverse set of structures. Training was continued on a vast dataset of MD simulations of thousands of proteins and more than 200 ms of aggregate simulation time. And finally, BioEmu was fine-tuned on more than 500,000 experimental protein stabilities using a technology developed here, property-prediction fine-tuning (PPFT).

**RESULTS:** We tested BioEmu on a variety of protein systems that are dissimilar from training proteins and benchmarked its performance on three tasks: (i) Predicting known conformational changes including large domain motions, local unfolding transitions, and the formation of cryptic binding pockets while achieving success rates of sampling the known references of between 55 and 90%. (ii) Emulating equilibrium distributions of both protein folding and native-state conformational transitions that can be generated by high-throughput MD simulation, demonstrating errors in free-energy differences below 1 kcal/mol and speedups of four to five orders of magnitude. (iii) Predicting experimentally measured stabilities of folded states of small proteins by directly generating equilibrium ensembles and explaining structure-stability relationships of mutants, achieving errors below 1 kcal/mol and correlation coefficients greater than 0.6 for both absolute folding free energies and folding free-energy changes of mutants.

**CONCLUSION:** BioEmu has various practical use cases, including complementing present MD simulation workflows, interpreting protein experiments in terms of structural mechanisms, identifying binding pockets and allosteric mechanisms in drug discovery, and generating ensembles for dynamical protein design. Our demonstration that the large upfront costs of MD simulation and experimental data generation can be amortized and that the prediction error decreases with an increasing amount of diverse training data indicates a path forward for predicting biomolecular function at the genomic scale. □

**Corresponding author. Frank Noé (franknoe@microsoft.com)** †These authors contributed equally to this work. **Cite this article as S. Lewis *et al.*, *Science* 389, eadv9817 (2025). DOI: 10.1126/science.adv9817**

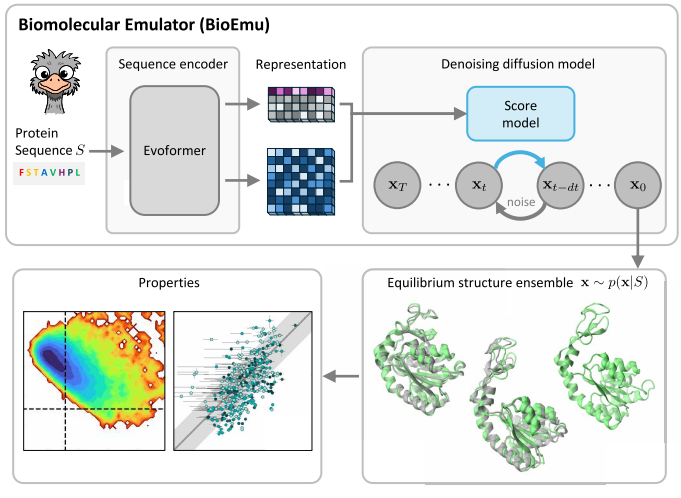


**Illustration of the BioEmu model and workflow.** BioEmu generates equilibrium protein structure ensembles by combining AlphaFold's sequence representation with a diffusion model trained on vast simulation and experimental data. These ensembles enable rapid computation of properties such as protein stability, achieving speeds that are orders of magnitude faster than MD simulation. [Emu illustration by F.N.]

PROTEIN SIMULATIONS

# Scalable emulation of protein equilibrium ensembles with generative deep learning

Sarah Lewis[1]†, Tim Hempel[1]†, José Jiménez-Luna[1]†, Michael Gastegger[1]†, Yu Xie[1]†, Andrew Y. K. Foong[1]†, Victor García Satorras[1]†, Osama Abdin[1]†, Bastiaan S. Veeling[1]†, Iryna Zaporozhets[1,2], Yaoyi Chen[1,2], Soojung Yang[1], Adam E. Foster[1], Arne Schneuing[1], Jigyasa Nigam[1], Federico Barbero[1], Vincent Stimper[1], Andrew Campbell[1], Jason Yim[1], Marten Lienen[1], Yu Shi[1], Shuxin Zheng[1], Hannes Schulz[1], Usman Munir[1], Roberto Sordillo[1], Ryota Tomioka[1], Cecilia Clementi[1,2,3], Frank Noé[1,2,3]*

Following the sequence and structure revolutions, predicting functionally relevant protein structure changes at scale remains an outstanding challenge. We introduce BioEmu, a deep learning system that emulates protein equilibrium ensembles by generating thousands of statistically independent structures per hour on a single graphics processing unit (GPU). BioEmu integrates more than 200 milliseconds of molecular dynamics (MD) simulations, static structures, and experimental protein stabilities using new training algorithms. It captures diverse functional motions—including cryptic pocket formation, local unfolding, and domain rearrangements—and predicts relative free energies with 1 kilocalorie per mole accuracy compared with millisecond-scale MD and experimental data. BioEmu provides mechanistic insights by jointly modeling structural ensembles and thermodynamic properties. This approach amortizes the cost of MD and experimental data generation, demonstrating a scalable path toward understanding and designing protein function.

Proteins and protein complexes are the functional building blocks of life and play a central role in drug development and biotechnology. Although next-generation sequencing and deep learning–based structure prediction tools (*1*–*4*) have revolutionized access to sequence and structure, scalable methods for exploring protein function remain elusive. A key driver of protein function is the ability to transition between distinct conformational states (i.e., sets of different structures), often coupled to the binding of ligands or other proteins. For example, actin's ability to form muscle fibers arises from its conformational dynamics, which is regulated by adenosine triphosphate (ATP) and adenosine diphosphate (ADP) (Fig. 1A).

Technologies in use now that quantitatively probe such conformational transitions and their coupling with binding states are not scalable. Single-molecule experiments can provide the full equilibrium distributions of observables such as intramolecular distances (*5*) but require bespoke molecular constructs and time-consuming data collection. Cryo–electron microscopy (cryo-EM) can resolve multiple conformational states of biomolecular complexes and their probabilities (*6*), but these experiments are time-consuming and costly. Molecular dynamics (MD) simulation is, in principle, a universal tool that allows both the structure and dynamics of biomolecules to be explored at all-atom resolution. However, biomolecular force fields are far from perfect, and the sampling problem makes studying protein folding or association with MD a feat of epic computational costs—even with special-purpose supercomputers or enhanced sampling methods (*7*, *8*). Machine-learned (ML) coarse-grained MD models have an opportunity to achieve similar accuracy as all-atom MD at two to three orders of magnitude lower computational cost (*9*, *10*) but are still under development.

The grand challenge to complete our understanding of protein function thus motivates the development of a technology that can elucidate protein conformational states and binding states, as well as their associated probabilities. This technology should ideally achieve an accuracy comparable to that of a converged MD simulation, or a cryo-EM experiment with multiconformation analysis, but it should only require a few hours of wall-clock time and cost no more than a few dollars per experiment. Boltzmann generators (*11*) have demonstrated that physics-based generative ML models can sample equilibrium distributions of arbitrary molecular energy functions; however, scaling such approaches to large macromolecules while maintaining high sample efficiency is challenging. Concurrently, data-based generative ML models, such as diffusion models, are now widely used in protein-structure prediction and design (*2*, *4*). Such models (*12*–*14*), as well as perturbation-based derivatives of AlphaFold (*15*, *16*), have also been shown to be capable of generating distinct protein structures and can be combined with MD simulation to alleviate the sampling problem (*17*). As of yet, generative ML systems have mainly demonstrated an ability to qualitatively sample distinct protein conformational states. A demonstration that generative ML can quantitatively match equilibrium ensembles and predict experimental observables is critical going forward (*18*).

## Model

We developed a biomolecular emulator called BioEmu—a generative deep-learning system designed to approximately sample from the equilibrium distribution of protein conformations. BioEmu uses a similar model architecture as Distributional Graphormer (DiG) (*12*) but employs a distinct training approach. Starting from the input protein sequence, single and pair representations of the sequence are computed using the AlphaFold2 evoformer (*1*). This sequence representation serves as input to a denoising diffusion model that generates protein structures (Fig. 1, B and C, and materials and methods). Sequence encoding is performed only once per protein, and an efficient integration scheme enables structure generation in as few as 30 to 50 denoising steps (fig. S11 and materials and methods). As a result, 10,000 independent protein structures from the learned equilibrium distribution can be sampled within minutes to a few hours on a single graphics processing unit (GPU), depending on their size.

A major challenge in training BioEmu is the absence of a single high-quality data source for protein equilibrium distributions, owing to the aforementioned challenges with experimental methods and MD (*19*). We therefore integrated training data from different, complementary sources. BioEmu was pretrained on a clustered version of the AlphaFold database (AFDB) using a data augmentation strategy that encourages it to sample diverse conformations (Fig. 1, D and E, and materials and methods). In a second stage, we continued training the model on more than 200 ms of all-atom MD data of thousands of small to medium-sized proteins (Fig. 1D, table S1, and materials and methods). To mitigate the sampling problem, MD data were reweighed toward equilibrium using either Markov state models (*20*) or weights from experimental data, when possible (materials and methods). Finally, we fine-tuned the model on 500,000 sequences of the MEGAscale dataset (*21*), a large-scale, homogeneous collection of in vitro protein stability measurements (Fig. 1D). As the MEGAscale dataset does not contain structures, we developed a new algorithm called property-prediction fine-tuning (PPFT) that can efficiently incorporate experimental measurements into diffusion model training (see section "Predicting protein stabilities" and materials and methods). To ensure

[1]AI for Science, Microsoft Research. [2]Freie Universität Berlin, Berlin, Germany. [3]Department of Chemistry, Rice University, Houston, TX, USA. ***Corresponding author. Email: franknoe@microsoft.com** †These authors contributed equally to this work.

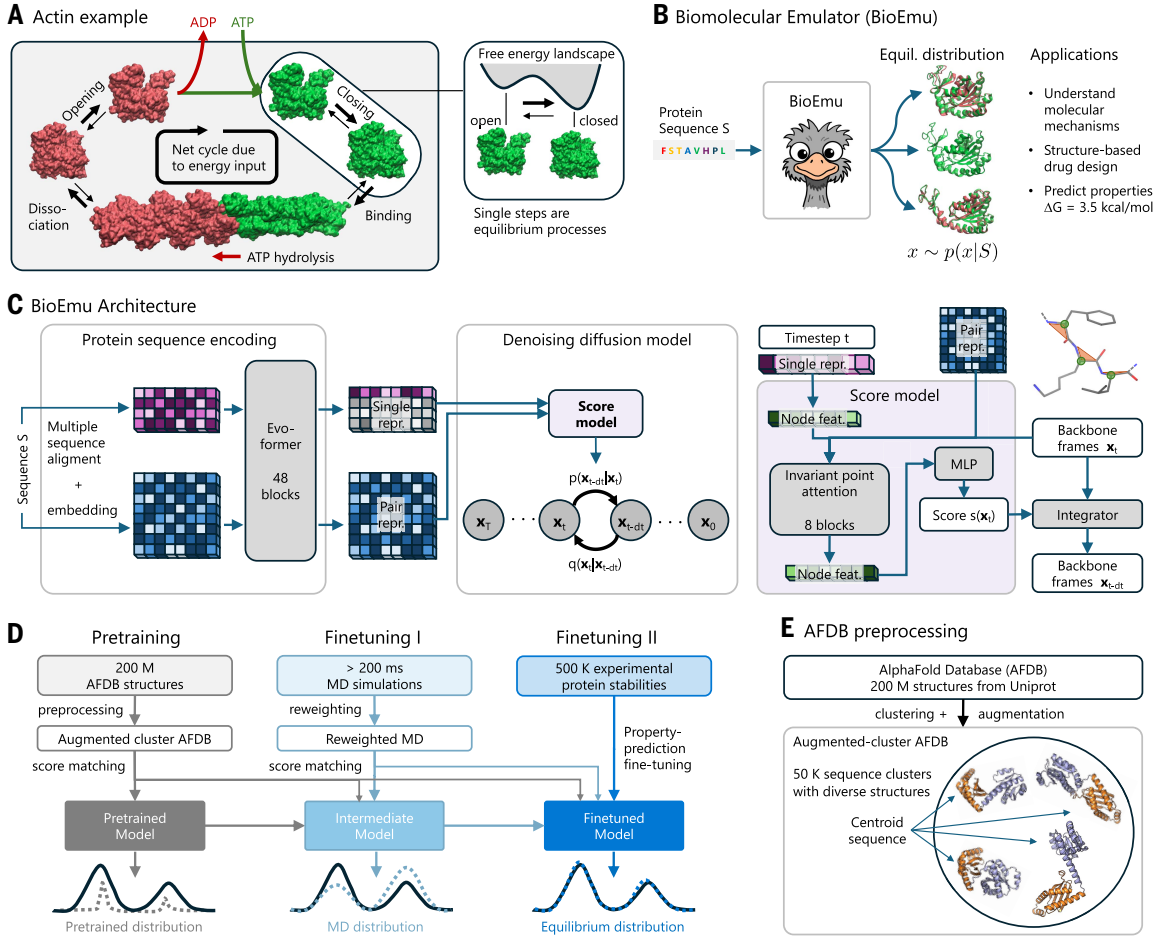


**Fig. 1. Overview of model and architecture.** (**A**) Actin as a representative example of protein function driven by conformational changes. Actin filament formation depends on an open-close transition of monomers, which is controlled by ADP and ATP binding. (**B**) Given a protein sequence, BioEmu samples protein structures from an approximate equilibrium distribution, from which properties such as free-energy differences can be computed. [Emu illustration by F.N.] (**C**) ML model architecture consisting of protein sequence encoder, denoising diffusion model, and score model. The protein structure is represented using coarse-grained backbone frames. (**D**) Data integration and model training pipeline. (**E**) Data processing pipeline for pretraining. MLP, multilayer perceptron; repr., representation; feat., features.

generalization, we filtered our training set such that no protein had more than 40% sequence similarity to a predefined holdout set, including all reported test proteins of at least 20 residues or longer. The term "BioEmu" refers to the fine-tuned model, trained on AFDB, MD simulations, and experimental protein-stability measurements. Subsequent results used this model unless otherwise indicated.

## Sampling conformational changes related to protein function

We consider the ability to qualitatively sample distinct, biologically relevant conformations as a foundation for building a quantitative equilibrium sampler. Therefore, we first tested whether BioEmu could predict known conformational changes and compared this capability with AFCluster (*15*), AlphaFlow (*13*), DiG (*12*), and uniform multiple-sequence alignment (MSA) subsampling (*22*) as representative baseline methods (materials and methods). To benchmark BioEmu's ability to capture functionally relevant structural variability, we curated four benchmark sets comprising around 100 proteins with experimentally validated transitions (figs. S1 to S4). The first set, OOD60, assesses sequence generalization. The remaining three sets target specific types of conformational change: domain motions, local unfolding transitions, and cryptic pocket formations.

OOD60 is designed to test strong generalization: Its proteins were deposited in the Protein Data Bank (PDB) after the AlphaFold2 cutoff date and share no more than 60 and 40% sequence similarity with the AlphaFold2 monomer model and BioEmu training sets, respectively. OOD60 includes various challenging cases such as large-scale conformational changes induced by binding with other biomolecules (fig. S1). BioEmu significantly outperformed all baselines on this benchmark (fig. S7). For the other benchmarks, BioEmu matched or exceeded baseline performance, except in predicting apo states of cryptic pockets, where AFCluster performed best. The performance gap was especially pronounced for proteins outside the AlphaFold2 training set (fig. S7).

The domain motion benchmark consists of proteins that undergo large-scale motions as part of their functional cycle (Fig. 2A). In the open-close transition of adenylate kinase, the closed state brings the substrates together to catalyze the ATP + AMP ⇌ 2ADP reaction. Single-molecule experiments have confirmed that opening and closing occurs reversibly on timescales of tens of microseconds when the substrates are bound (*23*). BioEmu predicts a range of open and closed states, including close matches with crystallographic structures. A second example is the open-close transition of LAO-binding protein, which is required to bind and release lysine, arginine, and ornithine for transport across membranes as part of the ATP-binding cassette protein family. Another interesting example of domain motions is that of the receptor module that regulates the concentration of cyclic

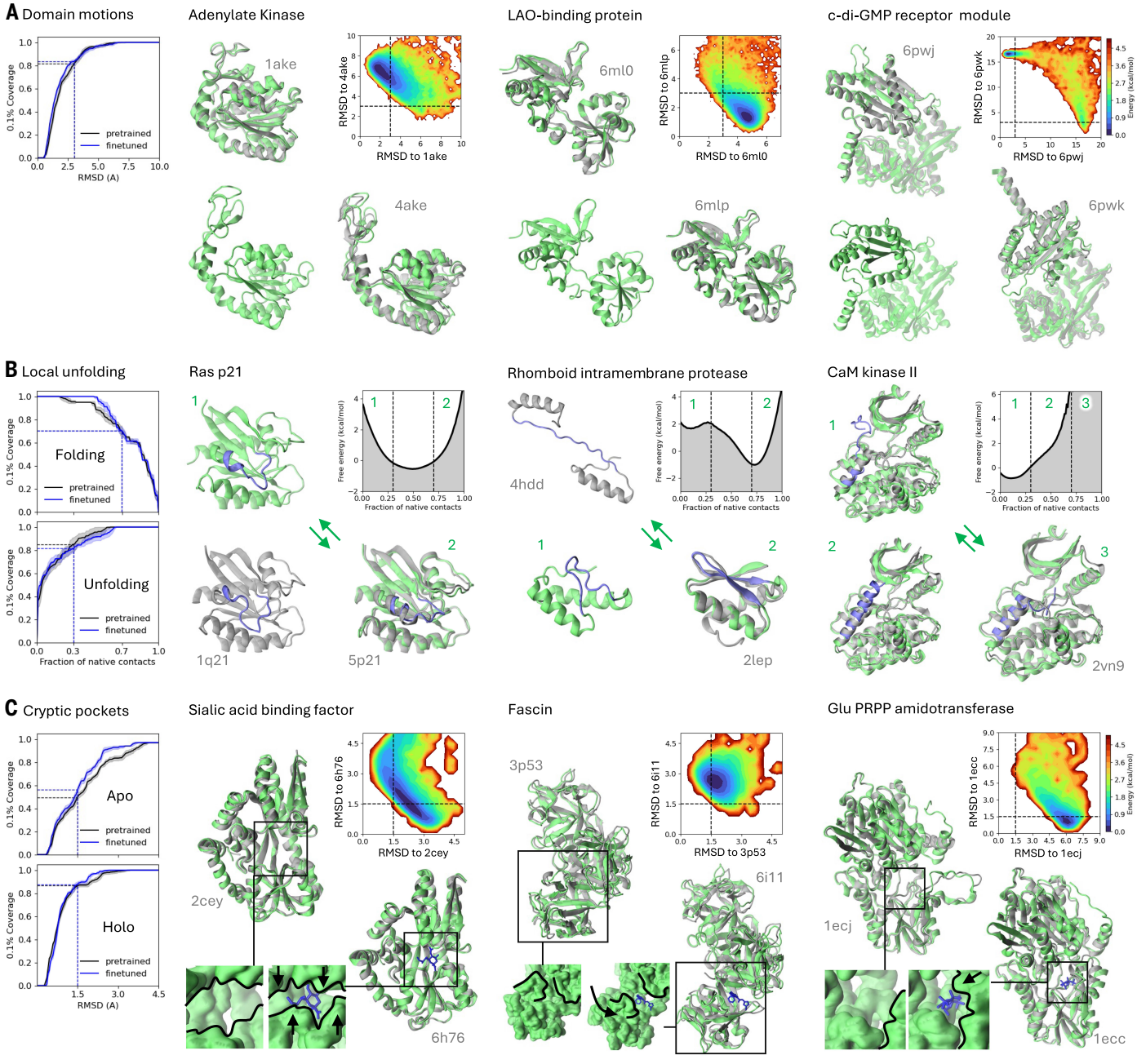


**Fig. 2. BioEmu samples functionally distinct protein conformations.** (**A**) Large-scale domain motions. (**B**) Local unfolding or unbinding of parts of the protein. (**C**) Formation of cryptic binding pockets that are not present in the apo ground state. The left column shows the coverage of pretrained and fine-tuned BioEmu models, defined as the percentage of reference structures that are sampled by at least 0.1% of samples within a given distance. Global and local RMSD are used for domain motions and cryptic pocket formation benchmarks, respectively, and the fraction of native contacts for local unfolding. Successful coverage of reference states is defined by probability density left of and below the dashed lines in the energy landscape plots of (A) and (C) and outside the dashed lines in (B). BioEmu sampled structures are shown in green, PDB structures in gray, and key secondary structure elements in blue. CaM, calcium-calmodulin dependent; LAO, lysine arginine ornithine. See table S4 for 12-letter PDB codes and original citations.

di–guanosine monophosphate (cyclic-di-GMP) in bacteria. In this case, one domain undergoes a large-scale rotation and repacks to the other domain with a completely different contact pattern. See fig. S2 for 19 further examples. Overall, BioEmu predicted 83% of the reference experimental structures with ≤3-Å root mean square deviation (RMSD) (Fig. 2A), indicating the model's ability to predict which protein regions are rigid and which are flexible, as well as which resulting motions can occur.

Next, we considered local unfolding transitions, in which part of a protein chain unfolds or detaches from its main structure as part of a signaling pathway (Fig. 2B). Predicting local unfolding challenges the model to correctly rank the relative stabilities of a protein's fold. A famous example of local unfolding is Ras p21, a protein whose mutants are often linked to cancer development; the local unfolding of Ras p21 is a conformational switch that signals cell growth (*24*). In its active state, stabilized by binding guanosine triphosphate (GTP), the switch II region forms a short α helix, which partially unfolds in the inactive guanosine diphosphate (GDP)–bound state. Rhomboid intramembrane protease is a more complex case that involves domain swapping. Its monomeric form features a globular conformation, whereas in its

dimeric form, the central β sheet unfolds and the helices of the two monomers bind to each other. Finally, CaM kinase II presents an autoinhibition mechanism, wherein the N terminus binds into the active site. BioEmu correctly predicted the local unfolding of these structure elements and sampled 70% of the folded and 81% of the locally unfolded states across 20 protein examples (Fig. 2B and fig. S3).

As a final class of conformational changes, we considered the formation of pockets that are absent in the apo PDB structure but can form to bind a small molecule (Fig. 2C). Such "cryptic" binding pockets can be discovered with high-performance MD simulation (*25*, *26*), but the millisecond timescales often involved in the spontaneous opening of such pockets make MD on commercial hardware rarely viable for in silico drug-discovery pipelines. We curated 34 cases of experimentally validated formation of cryptic binding pockets from the literature (fig. S4). The sialic acid binding factor presents a case where a large opening in the apo state can partially close and form a binding site for the ligand. Fascin is a four-domain protein where two domains can rotate with respect to each other to reveal a binding site. In Glu PRPP amidotransferase, part of the chain is unfolded in the apo state and can fold into a structure that completes the binding site for the ligand. To ensure capturing subtle changes, we defined success by a very strict 1.5-Å RMSD threshold to the apo and holo reference structures. Surprisingly, the model had a strong preference for holo states, successfully predicting the cryptic pocket in 86% of cases, whereas it only succeeded in predicting 56% of the apo structures, indicating further room for improvement (Fig. 2C). We hypothesize that the model may be picking up a bias implicit in the embeddings—proteins may have only one or a few apo structures deposited in the PDB, but it is common to find multiple structures of the same protein with different small molecules bound.

To conclude, we conducted several tests to confirm that BioEmu's multiconformation prediction is a hallmark of generalization rather than memorization of sequence-structure pairs. BioEmu's ability to predict multiple conformations depends only weakly on the sequence similarity between the query molecule and the training set, with the clearest trend seen for domain motions where the final performance is reached at 30% sequence similarity (fig. S5 and materials and methods). As our test proteins contain examples that overlap with the AlphaFold2 training set, we ablated whether multiconformation prediction performance stems from trivial extraction of the information already present in the AlphaFold2 evoformer embeddings. To this end, we trained an end-to-end version of the model that uses MSA information directly, instead of pretrained embeddings, on a training set that is at most 40% similar in sequence to any protein belonging to the multiconformation benchmarks. The resulting model showed similar performance in the previously mentioned multiconformation benchmarks to that of the fine-tuned BioEmu (fig. S6 and materials and methods).

## Emulating MD equilibrium distributions

A major motivation for developing BioEmu was to overcome the well-known sampling problem in MD: Simulating the full range of protein conformations and estimating their equilibrium probabilities often requires extensive MD simulations, on the order of 100 μs to 10 ms (*7*, *8*, *27*). These timescales are necessary to capture rare but functionally important transitions, yet achieving them is computationally demanding, if not prohibitive, even with specialized supercomputers (*28*) or large-scale distributed simulations integrated using statistical models (*8*, *29*). Here, we assessed BioEmu's ability to emulate the equilibrium distribution that would be sampled with extensive MD simulations. To this end, we amassed all-atom simulations of proteins with a total aggregated simulation time of more than 200 ms (table S1), which were used for fine-tuning BioEmu (Fig. 1D).

Before analyzing the model trained on the full dataset, we first tested whether our machine-learning architecture and training protocol permit learning to emulate long-timescale MD equilibrium distributions. To this end, we used D. E. Shaw Research (DESRES) simulations of 12 fast-folding proteins generated on the Anton supercomputer (*7*). For each protein, we fine-tuned a separate model on the other 11 proteins and evaluated it on the held-out 12th, an approach known as leave-one-out cross-validation (see materials and methods). This setup ensures that each test case is evaluated independently of its training data and avoids bias from arbitrary train-test splits in this small dataset. As expected, the AFDB-pretrained model predicted the native state but performed poorly in sampling the full free-energy surface (fig. S9). Surprisingly, however, the "DESRES–fine-tuned models," each trained on only 11 fast folders, predicted free-energy surfaces on the held-out proteins that closely matched the MD ground truth (Fig. 3A and fig. S9). For all proteins, the model predicted both the native as well as the unfolded states with similar shapes on the free-energy landscape. In many cases, several or all folding intermediates visible on the two-dimensional (2D) free-energy surface were predicted (Fig. 3A and fig. S9): For beta-beta-alpha protein (BBA), both MD and the DESRES–fine-tuned models predicted the existence of an intermediate with the α helix formed and the β sheet broken. For protein G, both MD and the DESRES–fine-tuned model sampled intermediates with half of the β sheet still formed, whereas the other half and most of the helix were broken. For homeodomain, MD and the model agreed in the prediction of an intermediate with only one helix turn unwound, whereas the unfolded states still featured some degree of helical propensity. There was excellent agreement of the predicted secondary structure propensities with the MD data (Fig. 3A). Quantitatively, the mean absolute error between the MD and model 2D free-energy landscapes was only 0.74 kcal/mol, ranging from 0.30 kcal/mol for BBA to 1.63 kcal/mol for λ-repressor, which is on the order of differences expected from two different classical MD force fields (*30*, *31*).

We compared the computational costs of MD data generation and BioEmu in GPU-hours (on a NVIDIA Titan V; Fig. 3A, top right). For all BioEmu results shown here, we drew 10,000 samples, which incurred computational costs of less than 1 GPU-minute for Chignolin to around 1 GPU-hour for λ-repressor. For MD, we considered the cost for generating the DESRES simulations, whose lengths were chosen to include roughly 10 folding-unfolding transitions. The MD costs then ranged from 2000 GPU-hours for Chignolin to more than 100,000 GPU-hours for NTL9, resulting in a speedup of BioEmu over MD of four to five orders of magnitude. We also note that for most proteins shown here, performing sufficiently long MD simulations to directly observe folding and unfolding in single trajectories is still not possible on consumer-grade hardware but instead requires a much more complex methodological framework (*29*, *32*).

The main BioEmu model was fine-tuned on more than 200 ms of MD simulations with Amber force fields at or near a temperature of 300 K (table S1). We chose to combine data from slightly different simulation conditions because each of these MD models is inherently imperfect, and we regarded experimental data as being more reliable for weighing between conformations (Fig. 1D). Differences in the simulation conditions of our own generated data were intentional; for example, AMBER ff99sb-disp (*33*) was chosen to avoid spuriously misfolded states produced by other force fields in the context of protein folding (materials and methods). A large fraction of training data, 46 ms, is dedicated to 1100 CATH domains, common building blocks of protein structure (*34*) (materials and methods). We designated 17 CATH systems with a simulation time of more than 100 μs as a test set and report statistics comparing MD and model distributions (Fig. 3B, fig. S10, and materials and methods). Similar to DESRES simulations, BioEmu predicts the native state with local fluctuations and typically several other substates and structures sampled by MD. Most secondary structure propensities matched well (Fig. 3B). We observed a free-energy mean absolute error over the converged test set of 0.9 kcal/mol, which was again comparable to the differences expected between different MD force fields.

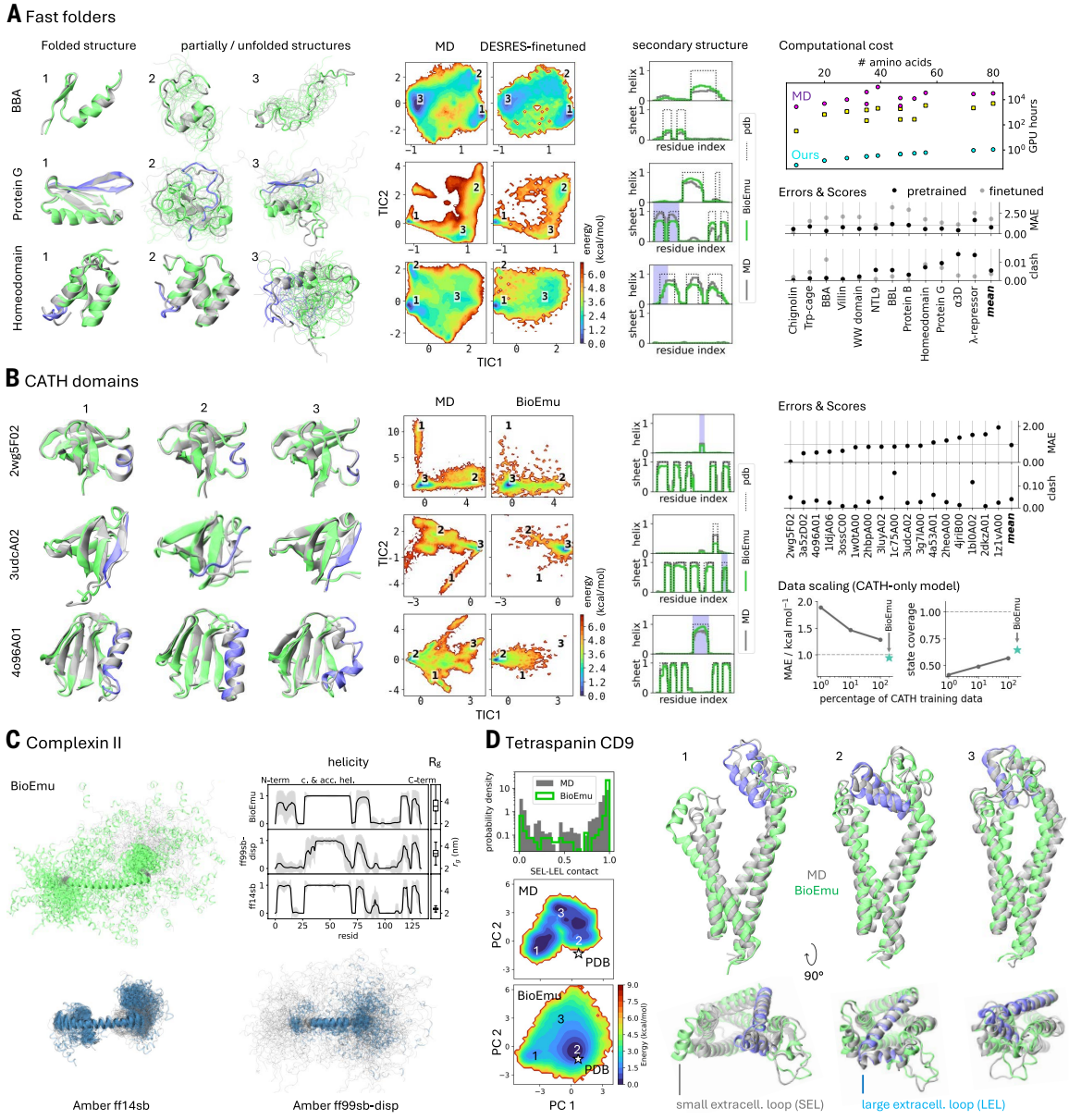


**Fig. 3. BioEmu achieves fast emulation of all-atom MD equilibrium distributions.** (**A**) DESRES fast-folding proteins. From left to right are representative structures (green, model; gray, MD; blue, regions of interest), free-energy surfaces over slowest time-lagged independent components (TIC) (*54*), secondary structure propensities, computational cost for MD (magenta, full DESRES data; yellow, single folding-unfolding roundtrip) versus model (cyan, 10,000 samples), and mean absolute error (MAE) of free-energy differences and fraction of unphysical model samples. (**B**) CATH domains. Structures, free-energy surfaces, and errors are as in (A). Shown at the bottom right is data scaling for the specialized CATH-only model with free-energy MAE and state coverage as a function of training data. The cyan star indicates BioEmu. (**C**) Complexin II: Structures, helix content, and radius of gyration ($R_g$) as predicted by BioEmu versus two all-atom force fields. (**D**) Tetraspanin CD9 results from BioEmu and MD (*37*). Shown is the open-closed transition, represented as a histogram in logarithmic scale of small and large extracellular loop (SEL-LEL) contacts, as defined in (*37*). Two-dimensional principal components (PC) analysis of $\exp(-d_{ij})$ of Ca-Ca distances $d_{ij}$ between SEL and LEL. The open star marks the experimental structure (6k4j).

To understand whether our model's ability to sample accurate equilibrium distributions is limited by training data or model expressivity, we trained three models with the same architecture as BioEmu from scratch, using only 1, 10, and 100% of the CATH systems in the training dataset. We observed decreased free-energy errors and an increased coverage of the conformations sampled by MD as the amount of training proteins increased (Fig. 3B, bottom right), suggesting that the model can be further improved by adding more training data. Notably, the fine-tuned model's error on the same test set was further reduced to 0.9 kcal/mol, demonstrating the potential benefit of pretraining and integrating multiple datasets even if they do not use identical simulation conditions.

Finally, we evaluated BioEmu for two case studies that involve larger proteins: complexin II (134 amino acids) and tetraspanin CD9 (225 amino acids). Complexin II is an intrinsically disordered protein (IDP) from the neurotransmitter release apparatus (*35*). IDPs tend to be difficult to sample with MD; however, BioEmu can efficiently emulate a flexible ensemble of complexin II structures (Fig. 3C) while reproducing known secondary structure elements such as the central and accessory helices (*35*, *36*). Achieving convergence of IDPs of this size with all-atom MD is unpractical. At an order of magnitude higher computational cost than with BioEmu, we conducted ~5 μs of MD simulations with all-atom MD, which are most likely not converged but already

display qualitatively different behavior: The AMBER ff14sb force field produced a very rigid compact structure with a small radius of gyration and little to no variation in secondary structure content, whereas AMBER ff99sb-disp tended to destabilize known secondary structure elements (Fig. 3C).

Tetraspanin CD9, the second case study, plays a role in cell adhesion and fusion (Fig. 3D). The large extracellular loop of CD9 is part of our OOD60 test set (fig. S1) in which our pretrained model samples both crystallographic reference structures (PDB entries 6rlo and 6rlr), whereas the BioEmu model fine-tuned on MD data samples 6rlo but discards 6rlr. This is consistent with the observation that both structures exist in crystal environments; however, 6rlr cannot be realized in a folded monomeric protein and is therefore correctly discarded when fine-tuning BioEmu (fig. S12, A to C). We also sampled the full-length CD9 structure, which has less than 40% sequence similarity to both the BioEmu and AlphaFold training sets (Fig. 3D). In agreement with MD simulations of previous work (*37*), BioEmu predicted the widely open state 1 and closed state 2 and similar contact distributions between the small and large extracellular loops as reported in (*37*). A principal components analysis revealed that BioEmu and MD sample similar sets of conformations (Fig. 3D). MD predicts an experimentally unknown metastable closed state 3, which is unstable in BioEmu. BioEmu samples closed structures (state 2) that are very similar to the experimental structure 6k4j (1.9-Å RMSD), whereas the closest MD sample has an RMSD of 4.6 Å to the crystal structure (fig. S12D).

## Predicting protein stabilities

Understanding protein stability is crucial for various applications in molecular biology, drug design, and biotechnology. From a modeling point of view, predicting a protein's stability is a specific case of predicting the equilibrium probabilities of its different conformational states, and these all arise from the same underlying biophysics. We therefore wanted to train BioEmu so that the proportion of samples in folded and unfolded states matches the experimentally measured protein stability. We classify protein structures as folded or unfolded based on their fraction of native contacts and define the folding free-energy as $\Delta G = G_{\text{folded}} - G_{\text{unfolded}}$ (materials and methods). To facilitate protein stability prediction, BioEmu was trained on 502,442 mutant sequences generated from 361 wild types, a subset of the more than 674,000 experimental measurements in the MEGAscale dataset (materials and methods) (*21*). We evaluated BioEmu on a test set of randomly chosen mutants from 95 wild types. For a subset of 271 wild-type proteins and 21,458 mutants, we additionally conducted a total of 25 ms of all-atom MD simulations of the folded and unfolded states (materials and methods). To address MD sampling and force-field issues, we weighed the folded and unfolded samples based on the experimentally measured protein stabilities (materials and methods). To accelerate training convergence and leverage the large number of MEGAscale measurements, we developed the PPFT algorithm (Fig. 4A and materials and methods), which integrates experimental expectation values, such as protein stabilities, into diffusion model training without requiring protein structures. PPFT uses fast approximate sampling with only eight denoising steps, which we observed to be sufficient to confidently predict whether each sampled structure will be classified as folded or unfolded. By comparing the mean foldedness of sampled structures with experimental measurements and backpropagating the error, our model could be efficiently trained to match experimental protein stabilities.

BioEmu achieved a mean absolute error of less than 0.9 kcal/mol and a Spearman correlation coefficient of approximately 0.6 for the MEGAscale test proteins (Fig. 4B). The BioEmu ensembles also correlated well with stability changes of point mutants, as measured by the change of folding free energy $\Delta\Delta G$, achieving mean absolute errors of less than 0.8 kcal/mol and a Spearman correlation coefficient above 0.6 (Fig. 4C). Errors of approximately 1 kcal/mol were achieved for test proteins that had 40% sequence similarity with the training set, but the best performance was obtained for test sets that included sequences with 50% or greater similarity (Fig. 4, B and C). As there are only 361 distinct wild-type sequences in the MEGAscale training data, it is likely that generalization performance can be further improved with a training dataset that is more diverse in protein sequence space.

To check whether BioEmu makes physically reasonable predictions outside the MEGAscale set of proteins, we tested it on proteins that are known to be very stable or unstable. We first selected stable proteins from ProThermDB (*38*) with a $\Delta G < -8$ kcal/mol (materials and methods). Our model consistently sampled these proteins in their folded states with a fraction of native contacts that always exceeded 0.65 (Fig. 4D). To test whether our model systematically predicts IDPs as unfolded, we used the CALVADOS test set (*39*). Most proteins sampled displayed a radius of gyration ($R_g$) similar to that of random coil structures and mostly larger than that of typical folded proteins (Fig. 4E). Unlike other models (*40*, *41*), ours has not been directly trained on IDPs; nonetheless, it provides zero-shot predictions of $R_g$ that correlate well with experimental measurements (Fig. 4E).

In comparison to black-box methods that predict protein stability directly from sequences (*42*–*45*), BioEmu has competitive or superior prediction accuracy. However, in contrast to a black-box prediction of $\Delta G$, we can analyze the structure ensemble generated by our model to reveal insights on mutation-caused stability changes. To illustrate this point, in Fig. 4F, we show mutants of the design protein HHH_rd1_0335 and PDB entry 2JWS. In HHH_rd1_0335, the mutation I7P ($Ile^{7}$→Pro) leads to a destabilization of the first helix, as indicated by the model's prediction of a $\Delta\Delta G$ of 1.8 kcal/mol compared with the experimental 2.1 kcal/mol. The analysis shows a decrease in average helicity that particularly affects the helix where the mutation is located. For 2jws, the mutation I24D ($Ile^{24}$→Asp) in the middle helix results in partial unfolding, with the model predicting a $\Delta\Delta G$ of 2.1 kcal/mol, which closely matches the experimental value of 2.9 kcal/mol. This mutation replaces a hydrophobic residue with a negatively charged aspartate, disrupting core stability and leading to a localized structural change. These analyses highlight BioEmu's ability to correlate predictions of thermodynamics with structural causes, which is not possible with black-box prediction models.

## Discussion

In this work, we introduce BioEmu, a generative machine-learning system that can approximately sample the equilibrium distributions of proteins and through that explore two key aspects of molecular function: protein conformations and their equilibrium probabilities. We have shown that the system can sample experimentally known structures of proteins undergoing a variety of conformational changes, approximate the equilibrium distributions of extensive MD simulations, and predict experimentally measured protein stabilities within errors of 1 kcal/mol. The cost of running inference is on the order of 1 GPU-hour per computational experiment, which is many orders of magnitude less than running MD simulations, even if enhanced sampling methods are invoked, and orders of magnitude cheaper than experiments that can provide detailed structure-function relationships. Nonetheless, there are further opportunities to reduce BioEmu's inference cost. Conditional flow-matching (*46*) can be used to generate protein structures using even fewer integration steps (*47*). The computational cost of evaluating the transformer network in the score model (Fig. 1C) can potentially be reduced by leveraging sparse or low-rank attention mechanisms.

BioEmu and MD simulation are complementary: Our system was trained on large amounts of MD simulation data for soluble proteins, and within this scope, it has shown that it can approximate MD distributions at a tiny fraction of the MD simulation costs. However, BioEmu is not designed to generalize beyond this scope; for example, membrane environments and small-molecule ligands are not represented in either the model or the training data, and reliable predictions cannot be

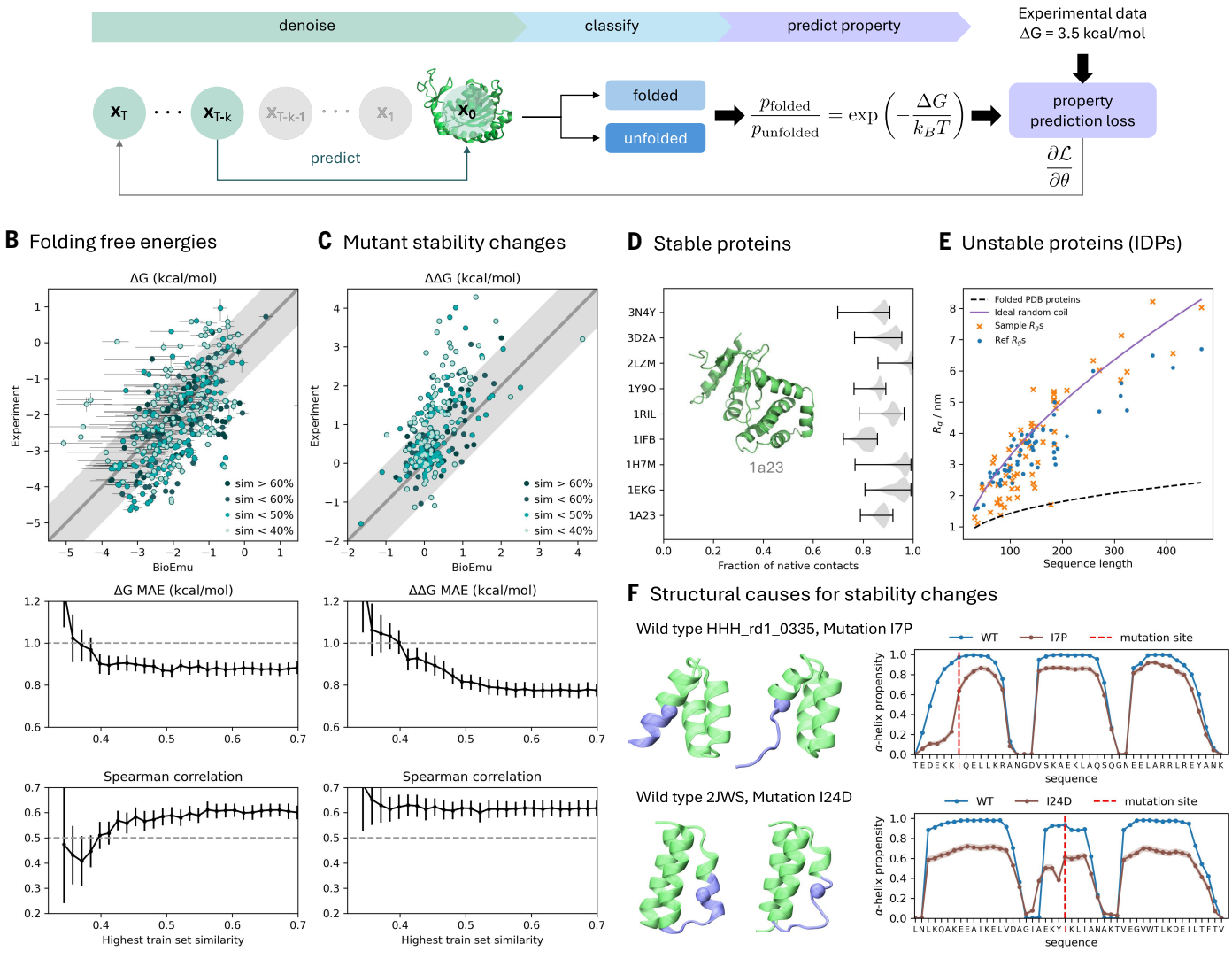


**Fig. 4. Prediction of experimentally measured protein stabilities.** (**A**) PPFT algorithm for fine-tuning a pretrained diffusion model to match experimentally measurable properties such as the protein's stability. (**B**) Comparison of experimental measurements of folding free energies (*21*) with model predictions, generated by direct sampling and counting of folded and unfolded states for test proteins, the mean absolute error (MAE), and the Spearman correlation coefficient as a function of the sequence similarity between test and training proteins. The gray shaded region represents an error range of ±1 kcal/mol. (**C**) Same as (B) for the change in folding stability upon point mutation. (**D**) Validation that very stable proteins that are not included in the MEGAscale experimental dataset are consistently predicted as folded. Bars and shading indicate the minimum, maximum, and distribution of sampled values, respectively. (**E**) Validation that IDPs reported in (*39*) and (*41*) are predicted as unfolded. Comparison of the radius of gyration ($R_g$) for model (orange crosses), experimental measurement (blue dots), and Flory scaling (*55*). (**F**) Analysis of the effect of two destabilizing mutants on the folded structures as predicted by the model: HHH_rd1_0335 with mutation I7P and 2JWS with mutation I24D. See table S4 for 12-letter PDB codes and original citations. WT, wild type.

assumed when such factors play a key role in the process. By contrast, MD can be readily generalized to such conditions, though it remains constrained by the sampling problem. Our system can be used to generate an approximation to the equilibrium distribution, and MD trajectories can be launched from a BioEmu ensemble to obtain chemically accurate all-atom structures, refine the distribution, and even compute dynamical properties. We therefore do not expect emulators such as BioEmu to make MD simulation obsolete; rather, we anticipate that MD will increasingly serve as a data generation and validation tool. A similar shift of roles is already in progress for other simulator-emulator pairs such as quantum chemistry methods and machine-learned force fields.

An important limitation of BioEmu is that it generates distributions entirely empirically, whereas MD simulation uses potential energy functions, which are connected to equilibrium distributions and expectation values by statistical mechanics. If direct access to a reduced potential energy function $u(x)$ was available that is consistent with the generated distribution by $p(x) \propto e^{-u(x)}$, it could be used for reweighting and making rigorous enhanced sampling simulations available through the emulator. BioEmu can potentially also be improved by going beyond score matching and using energy or forces information from MD force fields at training time, as considered in Boltzmann generators (*11*) and variational force matching (*9*, *48*).

Although BioEmu samples approximate equilibrium distribution, it does not model protein dynamics, which is done by MD and other methods (*49*), nor does its training incorporate dynamical information as it does for Markov state models (*20*). A pragmatic approach to generate a dynamic ensemble is to predict starting points with BioEmu and launch MD simulations from them. A starting point for a more

principled methodology that can both model dynamics and exploit dynamical information in the training data is the implicit transfer operator approach (*50*).

We have demonstrated that by using the PPFT method developed here, BioEmu can be efficiently fine-tuned on experimental data. In this work, we have chosen to do that for protein stabilities using the MEGAscale dataset because it presents a very favorable trade-off of large data scale and quantitative reliability. However, in principle, PPFT can be used to fine-tune BioEmu and other diffusion models to match any set of experimental observables, including nuclear magnetic resonance data, small-angle x-ray scattering, fluorescence measurements, and so on. Being able to fine-tune the generated ensemble to arbitrary experimental data is an important advantage compared with MD force fields: These can also be tuned to fit experimental data (*51*), but the processes that give rise to the experimental observables must be sampled during the training process, a task that is tedious or even unfeasible for observables that involve complex rare events, such as folding free energies.

A widely used feature of AlphaFold is its ability to predict confidence in an output structure, and a similar confidence module for BioEmu would be highly desirable. Training a confidence module is relatively straightforward in AlphaFold, which serves a single prediction task (structure) and relies on a single ground truth dataset (the PDB), whereas equilibrium structure ensembles serve multiple downstream tasks and observables and no universal ground truth dataset is available. Confidence prediction or uncertainty quantification of arbitrary observables thus remains an important future research direction and may leverage ongoing research in the deep-learning community (*52*). Even a rough notion of model confidence in observables, such as free-energy differences, could be exploited to improve training data efficiency: In the MD community, Markov state models and other kinetic models have been used to guide MD data production in an active learning loop (*8*, *53*), and a similar approach could be implemented to have BioEmu request new MD or experimental data that are most likely to increase model confidence.

Another limitation of the present system is that it emulates single protein chains under a fixed thermodynamic condition of 300 K. A proper emulator for proteins requires conditioning on experimentally and biologically relevant parameters such as temperature and pH and needs to be able to model multiple interacting molecules, as proteins rarely have a function on their own. We envision two ways of achieving this: (i) training on additional MD simulation data at relevant thermodynamic ranges and (ii) incorporating relevant additional experimental data (e.g., melting curves for temperature) into fine-tuning strategies.

An important future direction is to extend BioEmu's modalities by incorporating multiple protein chains and ligands, which are already included in recent biomolecular structure prediction systems (*2*, *4*). Presently, oligomer and ligand binding state are implicit, which may cause biases in the training data to show up in the sampled distribution. A hallmark of this may be BioEmu's preference to predict holo over apo structures in the cryptic pocket benchmark (Fig. 2C). Such biases in the sampling distribution can be avoided by explicitly conditioning the prediction of the structure ensemble to ligands, which is, however, hampered by the lack of training data. Although we have shown that the ability to accurately emulate the equilibrium distributions of small proteins increases with more training data, the sampling problem limits MD as a data-generation engine. The development of highly scalable experimental techniques is key for training machine-learning models that can predict conformational and binding states of large biomolecular complexes, as well as the subtle differences of binding affinities that underlie biomolecular function.

Our results demonstrate that the large upfront costs of MD simulation and experimental data generation can be amortized by a deep-learning emulator whose prediction error decreases with an increasing amount of high-quality training data. This indicates a path forward for predicting biomolecular function at the genomic scale.

Materials and methods are available in the supplementary materials.

## REFERENCES AND NOTES


1. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021). doi: 10.1038/s41586-021-03819-2; pmid: 34265844
2. J. Abramson *et al.*, Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024). doi: 10.1038/s41586-024-07487-w; pmid: 38718835
3. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021). doi: 10.1126/science.abj8754; pmid: 34282049
4. R. Krishna *et al.*, Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024). doi: 10.1126/science.adl2528; pmid: 38452047
5. F. Ritort, Single-molecule experiments in biological physics: Methods and applications. *J. Phys. Condens. Matter* **18**, R531–R583 (2006). doi: 10.1088/0953-8984/18/32/R01; pmid: 21690856
6. X.-C. Bai, G. McMullan, S. H. Scheres, How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, 49–57 (2015). doi: 10.1016/j.tibs.2014.10.005; pmid: 25544475
7. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011). doi: 10.1126/science.1208351; pmid: 22034434
8. N. Plattner, S. Doerr, G. De Fabritiis, F. Noé, Complete protein-protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9**, 1005–1011 (2017). doi: 10.1038/nchem.2785; pmid: 28937668
9. J. Wang *et al.*, Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* **5**, 755–767 (2019). doi: 10.1021/acscentsci.8b00913; pmid: 31139712
10. N. E. Charron *et al.*, Navigating protein landscapes with a machine-learned transferable coarse-grained model. *Nat. Chem.* **17**, 1284–1292 (2025). doi: 10.1038/s41557-025-01874-0; pmid: 40681718
11. F. Noé, S. Olsson, J. Köhler, H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019). doi: 10.1126/science.aaw1147; pmid: 31488660
12. S. Zheng *et al.*, Predicting equilibrium distributions for molecular systems with deep learning. *Nat. Mach. Intell.* **6**, 558–567 (2024). doi: 10.1038/s42256-024-00837-3
13. J. Jing, B. Berger, T. Jaakkola, "AlphaFold meets flow matching for generating protein ensembles" in *Proceedings of the 41st International Conference on Machine Learning* (JMLR, 2024), pp. 22277–22303.
14. Z. Qiao, W. Nie, A. Vahdat, T. F. Miller III, A. Anandkumar, State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.* **6**, 195–208 (2024). doi: 10.1038/s42256-024-00792-z
15. H. K. Wayment-Steele *et al.*, Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024). doi: 10.1038/s41586-023-06832-9; pmid: 37956700
16. P. Bryant, F. Noé, Structure prediction of alternative protein conformations. *Nat. Commun.* **15**, 7328 (2024). doi: 10.1038/s41467-024-51507-2; pmid: 39187507
17. B. P. Vani, A. Aranganathan, D. Wang, P. Tiwary, AlphaFold2-RAVE: From sequence to Boltzmann ranking. *J. Chem. Theory Comput.* **19**, 4351–4354 (2023). doi: 10.1021/acs.jctc.3c00290; pmid: 37171364
18. A. Aranganathan, X. Gu, D. Wang, B. Vani, P. Tiwary, Modeling Boltzmann-eighted structural ensembles of proteins using artificial intelligence–based methods. *Curr. Opin. Struct. Biol.* **91**, 103000 (2025). doi: 10.1016/j.sbi.2025.103000; pmid: 39923288
19. R. E. Amaro *et al.*, The need to implement FAIR principles in biomolecular simulations. *Nat. Methods* **22**, 641–645 (2025). doi: 10.1038/s41592-025-02635-0; pmid: 40175561
20. J.-H. Prinz *et al.*, Markov models of molecular kinetics: Generation and validation. *J. Chem. Phys.* **134**, 174105 (2011). doi: 10.1063/1.3565032; pmid: 21548671
21. K. Tsuboyama *et al.*, Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* **620**, 434–444 (2023). doi: 10.1038/s41586-023-06328-6; pmid: 37468638
22. D. del Alamo, D. Sala, H. S. Mchaourab, J. Meiler, Sampling alternative conformational states of transporters and receptors with AlphaFold2. *eLife* **11**, e75751 (2022). doi: 10.7554/eLife.75751; pmid: 35238773
23. H. Y. Aviram *et al.*, Direct observation of ultrafast large-scale dynamics of an enzyme under turnover conditions. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 3243–3248 (2018). doi: 10.1073/pnas.1720448115; pmid: 29531052
24. F. B. Fromowitz *et al.*, ras p21 expression in the progression of breast cancer. *Hum. Pathol.* **18**, 1268–1275 (1987). doi: 10.1016/S0046-8177(87)80412-4; pmid: 3315956
25. J. B. Greisman *et al.*, Discovery and validation of the binding poses of allosteric fragment hits to protein tyrosine phosphatase 1b: From molecular dynamics simulations to x-ray crystallography. *J. Chem. Inf. Model.* **63**, 2644–2650 (2023). doi: 10.1021/acs.jcim.3c00236; pmid: 37086179
26. M. I. Zimmerman *et al.*, SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. *Nat. Chem.* **13**, 651–659 (2021). doi: 10.1038/s41557-021-00707-0; pmid: 34031561

27. T. J. Lane, D. Shukla, K. A. Beauchamp, V. S. Pande, To milliseconds and beyond: Challenges in the simulation of protein folding. *Curr. Opin. Struct. Biol.* **23**, 58–65 (2013). doi: 10.1016/j.sbi.2012.11.002; pmid: 23237705
28. D. E. Shaw *et al.*, Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010). doi: 10.1126/science.1187409; pmid: 20947758
29. J. D. Chodera, F. Noé, Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* **25**, 135–144 (2014). doi: 10.1016/j.sbi.2014.04.002; pmid: 24836551
30. R. B. Best, J. Mittal, Free-energy landscape of the GB1 hairpin in all-atom explicit solvent simulations with different force fields: Similarities and differences. *Proteins* **79**, 1318–1328 (2011). doi: 10.1002/prot.22972; pmid: 21322056
31. D. F. Hahn, V. Gapsys, B. L. de Groot, D. L. Mobley, G. Tresadern, Current state of open source force fields in protein-ligand binding affinity predictions. *J. Chem. Inf. Model.* **64**, 5063–5076 (2024). doi: 10.1021/acs.jcim.4c00417; pmid: 38895959
32. A. Laio, M. Parrinello, Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12562–12566 (2002). doi: 10.1073/pnas.202427399; pmid: 12271136
33. P. Robustelli, S. Piana, D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4758–E4766 (2018). doi: 10.1073/pnas.1800690115; pmid: 29735687
34. I. Sillitoe *et al.*, CATH: Increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021). doi: 10.1093/nar/gkaa1079; pmid: 33237325
35. J. Malsam *et al.*, Complexin suppresses spontaneous exocytosis by capturing the membrane-proximal regions of VAMP2 and SNAP25. *Cell Rep.* **32**, 107926 (2020). doi: 10.1016/j.celrep.2020.107926; pmid: 32698012
36. Q. Zhou *et al.*, The primed SNARE-complexin-synaptotagmin complex for neuronal exocytosis. *Nature* **548**, 420–425 (2017). doi: 10.1038/nature23484; pmid: 28813412
37. R. Umeda *et al.*, Structural insights into tetraspanin CD9 function. *Nat. Commun.* **11**, 1606 (2020). doi: 10.1038/s41467-020-15459-7; pmid: 32231207
38. R. Nikam, A. Kulandaisamy, K. Harini, D. Sharma, M. M. Gromiha, ProThermDB: Thermodynamic database for proteins and mutants revisited after 15 years. *Nucleic Acids Res.* **49**, D420–D424 (2021). doi: 10.1093/nar/gkaa1035; pmid: 33196841
39. G. Tesei *et al.*, Conformational ensembles of the human intrinsically disordered proteome. *Nature* **626**, 897–904 (2024). doi: 10.1038/s41586-023-07004-5; pmid: 38297118
40. G. Tesei, K. Lindorff-Larsen, Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *Open Res. Eur.* **2**, 94 (2023). doi: 10.12688/openreseurope.14967.2; pmid: 37645312
41. J. Zhu *et al.*, Precise generation of conformational ensembles for intrinsically disordered proteins via fine-tuned diffusion models. bioRxiv 2024.05.05.592611 [Preprint] (2024); https://doi.org/10.1101/2024.05.05.592611.
42. J. Ouyang-Zhang, D. Diaz, A. Klivans, P. Krähenbühl, Predicting a protein's stability under a million mutations. *Adv. Neural Inf. Process. Syst.* **36**, 76229–76247 (2023).
43. M. Cagiada, S. Ovchinnikov, K. Lindorff-Larsen, Predicting absolute protein folding stability using generative models. *Protein Sci.* **34**, e5233 (2025). doi: 10.1002/pro.5233; pmid: 39673466
44. P. Notin *et al.*, ProteinGym: Large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* **36**, 64331–64379 (2023).
45. T. Widatalla, R. Rafailov, B. Hie, Aligning protein generative models with experimental fitness via Direct Preference Optimization. bioRxiv 2024.05.20.595026 [Preprint] (2024); https://doi.org/10.1101/2024.05.20.595026.
46. Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, M. Le, Flow matching for generative modeling. arXiv:2210.02747 [cs.LG] (2022).
47. J. Yim *et al.*, Fast protein backbone generation with SE(3) flow matching. arXiv:2310.05297 [q-bio.QM] (2023).
48. W. G. Noid *et al.*, The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244114 (2008). doi: 10.1063/1.2938860; pmid: 18601324
49. L. Klein *et al.*, Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. *Adv. Neural Inf. Process. Syst.* **36**, 52863–52883 (2023).
50. M. Schreiner, O. Winther, S. Olsson, Implicit transfer operator learning: Multiple time-resolution models for molecular dynamics. *Adv. Neural Inf. Process. Syst.* **36**, 36449–36462 (2023).
51. T. Fröhlking, M. Bernetti, N. Calonaci, G. Bussi, Toward empirical force fields that match experimental observables. *J. Chem. Phys.* **152**, 230902 (2020). doi: 10.1063/5.0011346; pmid: 32571067
52. J. Gawlikowski *et al.*, A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* **56**, 1513–1589 (2023). doi: 10.1007/s10462-023-10562-9
53. H. Sidky, W. Chen, A. L. Ferguson, Machine learning for collective variable discovery and enhanced sampling in biomolecular simulation. *Mol. Phys.* **118**, e1737742 (2020). doi: 10.1080/00268976.2020.1737742
54. G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, F. Noé, Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013). doi: 10.1063/1.4811489; pmid: 23822324
55. H. Hofmann *et al.*, Polymer scaling laws of unfolded and intrinsically disordered proteins quantified with single-molecule spectroscopy. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 16155–16160 (2012). doi: 10.1073/pnas.1207719109; pmid: 22984159
56. S. Lewis *et al.*, Molecular dynamics dataset for "scalable emulation of protein equilibrium ensembles with generative deep learning": Octapeptides. Zenodo (2025); https://doi.org/10.5281/zenodo.15641199.
57. S. Lewis *et al.*, Molecular dynamics dataset for "scalable emulation of protein equilibrium ensembles with generative deep learning": CATH domains. Zenodo (2025); https://doi.org/10.5281/zenodo.15629740.
58. S. Lewis *et al.*, Molecular dynamics dataset for "scalable emulation of protein equilibrium ensembles with generative deep learning": MegaSim. Zenodo (2025); https://doi.org/10.5281/zenodo.15641184.
59. G. M. Visani, W. Galvin, M. Pun, A. Nourmohammad, "H-Packer: Holographic rotationally equivariant convolutional neural network for protein side-chain packing" in *Proceedings of the 18th Machine Learning in Computational Biology Meeting*, vol. 240 of *Proc. Mach. Learn. Res.*, D. A. Knowles, S. Mostafavi, Eds. (PMLR, 2024), pp. 230–249.
60. P. Eastman *et al.*, OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comput. Biol.* **13**, e1005659 (2017). doi: 10.1371/journal.pcbi.1005659; pmid: 28746339
61. BioEmu benchmarks, Github (2025); https://github.com/microsoft/bioemu-benchmarks.
62. S. Lewis *et al.*, Supplementary data for "scalable emulation of protein equilibrium ensembles with generative deep learning". Zenodo (2025); https://doi.org/10.5281/zenodo.15672282.


## ACKNOWLEDGMENTS

We thank the entire Microsoft Research AI for Science team and thank, in particular, the following individuals for valuable support and discussions: C. Liu, P. Jin, T.-Y. Liu, C. M. Bishop, B. Kruft, J. Köhler, T. Vogels, M. Segler, R. van den Berg, M. Federici, S. Markou, M. Riechert, and L. Sun. We thank M. Steinegger (Seoul National University) for his invaluable help in publishing the BioEmu inference code on ColabFold and T. Nishizawa (University of Tokyo) for sharing the CD9 simulation data. Furthermore, we thank colleagues from FU Berlin for valuable discussions and advice, particularly N. E. Charron, K. Elez, and A. S. Pasos Trejo. Additionally, we thank G. R. Bowman and A. M. Trent (University of Pennsylvania) as well as S. Singh (MSKCC) for their help with Folding@Home data. We also thank N. Plattner (University of Basel) for her support. **Funding:** All work was funded by Microsoft. **Author contributions:** Data curation, generation, and benchmarks: S.L., T.H., J.J.-L., I.Z., Y.C., C.C., F.N.; Software: S.L., T.H., J.J.-L., M.G., Y.X., A.Y.K.F., V.G.S., O.A., B.S.V., Y.C., S.Y., A.S., J.N., F.B., V.S., A.C., J.Y., M.L., H.S., R.T.; Methodology: S.L., T.H., J.J.-L., M.G., Y.X., A.Y.K.F., V.G.S., O.A., B.S.V., I.Z., Y.C., S.Y., A.E.F., A.S., J.N., F.B., V.S., A.C., J.Y., M.L., Y.S., S.Z., R.T., F.N.; Results generation, analysis, validation, and visualization: S.L., T.H., J.J.-L., M.G., Y.X., A.Y.K.F., V.G.S., O.A., B.S.V., I.Z., Y.C., A.E.F., R.T., C.C., F.N.; Writing: S.L., T.H., J.J.-L., M.G., Y.X., A.Y.K.F., V.G.S., O.A., I.Z., Y.C., R.T., C.C., F.N.; Project administration: R.S., U.M.; Software lead: S.L., B.S.V.; Simulation data lead: T.H.; Project lead and conceptualization: F.N. **Competing interests:** A US patent application for the algorithms and training methods described in this paper has been filed under the title "Prediction of protein structure ensembles." The following authors are listed as inventors in the patent application: S.L., T.H., J.J.-L., M.G., Y.X., A.Y.K.F., V.G.S., O.A., B.S.V., S.Y., A.S., and F.N. **Data and materials availability:** As presently no standard database for MD data exists [see (*19*) for a discussion], MD data were either downloaded from various sources or generated for this work. See table S1 and materials and methods for a description of all datasets with download links. The MD training data produced as part of this study are available at Zenodo under the permissive Community Data License Agreement CDLA 2.0; see (*56*–*58*). The DESRES data for fast-folding proteins cannot be redistributed according to the DESRES data license. Researchers are requested to contact DESRES directly for obtaining this dataset and a license. The BioEmu model and inference code are available under MIT license at https://github.com/microsoft/bioemu and can be installed using "pip install bioemu" for scalable execution on Linux desktops or clusters. Internal training code is not published because it is specialized to a Microsoft-internal Azure cloud environment. To facilitate the development of a community training code for BioEmu, the pretraining AFDB cluster definition and a sample implementation of PPFT are available in the same GitHub repository. A simple way to test the model is available through https://github.com/sokrypton/ColabFold. Side-chain reconstruction and relaxation of generated structures for downstream analyses are also provided through the hpacker (*59*) and OpenMM (*60*) packages, respectively. Code for benchmarking BioEmu or other models for their ability to sample multiple protein conformations, free-energy surfaces, and folding free energies (Figs. 2, 3B, and 4A and figs. S1 to S4) is provided under MIT license at https://github.com/microsoft/bioemu-benchmarks (*61*). The version of the BioEmu code and model used in this paper, as well as the BioEmu model samples used for generating the figures, has been archived at Zenodo (*62*). **License information:** Copyright © 2025 the authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original US government works. https://www.science.org/about/science-licenses-journal-article-reuse

## SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adv9817
Materials and Methods; Figs. S1 to S12; Tables S1 to S4; Algorithms 1 and 2; References (*63*–*126*); MDAR Reproducibility Checklist

Submitted 17 January 2025; accepted 27 June 2025; published online 10 July 2025
10.1126/science.adv9817