Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Efficient exploration of conformational ensembles via protein interaction-informed deep learning

Zilin Ren

zilin.ren@outlook.com

Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences
https://orcid.org/0000-0002-3621-1024

Duo Liu
Northeast Normal University

Qianyi Jia
Northeast Normal University

Qiantong Jin
Northeast Normal University

Yubin Liu
Northeast Normal University

Chan Liu
National University of Defense Technology

Zhixiang Sui
Northeast Normal University

Luyao Han
Northeast Normal University

Yixiang Zhang
Northeast Normal University

Xin Yang
Northeast Normal University

Pingping Sun

Zhiguo Fu
Northeast Normal University

Chengkun Wu
National University of Defense Technology

Ming Ni
Institute of Health Service and Transfusion Medicine

Xiaochen Bo

Department of Advanced & Inter-disciplinary Biotechnology, Academy of Military Medical Sciences
https://orcid.org/0000-0003-1911-7922

Article

Keywords:

Posted Date: June 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9549778/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Efficient exploration of conformational ensembles via protein interaction-informed deep learning

Zilin Ren[1,2,5,*], Duo Liu[1,2,5], Qianyi Jia[1,2,5], Qiantong Jin[1,2,5], Yubin Liu[1,2,5], Chan Liu[3], Zhixiang Sui[1,2], Luyao Han[1,2], Yixiang Zhang[1,2], Xin Yang[1,2], Pingping Sun[2], Zhiguo Fu[2], Chengkun Wu[3,*], Ming Ni[4,*], Xiaochen Bo[4,*]

1 Changchun Veterinary Research Institute, Chinese Academy of Agricultural Sciences, State Key Laboratory of Pathogen and Biosecurity, Key Laboratory of Jilin Province for Zoonosis Prevention and Control, Changchun 130012, China;

2 Research Center of AI for Science, School of Information Science and Technology, Northeast Normal University, Changchun, Jilin 130117, China

3 College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan 410073, China

4. Institute of Health Service and Transfusion Medicine, Beijing 100850, China

5 These authors contributed equally: Zilin Ren, Duo Liu, Qianyi Jia, Yubin Liu, Qiantong Jin.

* To whom correspondence should be addressed. Email: zilin.ren@outlook.com; chengkun_wu@nudt.edu.cn; ni.ming@bmi.ac.cn; boxc@bmi.ac.cn

Abstract

Prediction of protein conformational ensembles is essential for understanding protein functions and advancing drug design. Molecular dynamics (MD) simulation is computationally intensive for prediction; generative artificial intelligence (AI) frameworks offer a promising alternative. To ensure model performance, existing methods continuously expand training data, including experimental structures, AlphaFold predictions, and MD trajectories; however, model performance still needs improvement. Here, we demonstrated that incorporating protein complex interactions enhances model performance more effectively than relying solely on single-chain structures. Supplementing single-chain data with a mere 10% of complex data improved the performance of monomer-trained models by 6.64% in PwD-JSD and 22.87% in Rg-JSD. Building upon this insight, we introduce DynoM, a generative framework trained on 5,502 in-house all-atom MD trajectories of protein complexes. DynoM demonstrates high robustness in generating monomeric ensembles; notably, for multi-domain proteins, the model effectively suppresses structural hallucinations in inter-domain positioning by capturing the long-range physical constraints inherent in complex interactions. Overall, our findings underscore the pivotal role of complex interactions in predicting accurate conformational ensembles and provide a robust tool for protein modeling research.

Introduction

Proteins are inherently dynamic macromolecules, and their biological functions are governed by the transitions between multiple conformational states[1,2]. Accurately characterizing these conformational ensembles is essential for deciphering the mechanisms and drug development[3,4]. To analyze these dynamic behaviors, researchers primarily rely on molecular dynamics (MD) simulation as the gold standard for exploring the protein conformational

landscape[5]. However, despite its ability to closely simulate reality, it often comes at the cost of huge computational resources[6,7]. The time scales required to observe protein conformational changes or the binding processes of complexes often extend into the microsecond or millisecond timescale, which could bring with a huge computational burden for high-throughput screening[8] or large macromolecular systems[9]. The gap between the need for dynamic insights and the enormous burden of physics simulations has spurred the development of generative artificial intelligence (AI) models for rapid conformational ensemble prediction.

Deep learning methods, represented by AlphaFold2 (AF2)[10] and RoseTTAFold[11], have already revolutionized the protein structure prediction field. Building on these advances, AlphaFold3 (AF3)[12] integrates a diffusion-based generative module. While the denoising process is still implemented as iterative regression of coordinates, the shift of the module reflects a transition in protein structure prediction from a regression task to a probabilistic generative task. Rather than seeking a single set of ground-truth coordinates, the frameworks (e.g., diffusion[13] and flow-based models[14]) aim to model the underlying probability distribution of atom-level protein structures. This provides a theoretical basis for sampling conformational ensembles.

Recently, a series of methods have been proposed, including unconditional diffusion models (e.g., FoldingDiff[15] and Str2Str[16]), conditional diffusion models (e.g., ConfDiff[17], BioEmu[18], IDPFold[19], Mac-Diff[20], Eigenfold[21], and DiG[22]), and flow-based models (e.g., AlphaFlow[23], F^3Flow[24], and P2DFlow[25]). The success of these generative frameworks implies that co-evolutionary signals captured by AF2 and ESM-2[26] can efficiently guide the generation of conformational ensembles. Indeed, methods using co-evolutionary information may face predictive challenges, in which models tend to mischaracterize flexible and rigid regions in information-sparse regions[27,28], and rely excessively on stochastic noise, leading to implausible structural hallucinations. Efforts to address these challenges primarily involve expanding training datasets through a two-stage training process: first, pre-training on large-scale static structures,

and then fine-tuning on MD trajectories (Fig. 1a). To maximize the diversity and volume of training data, it has become standard to use isolated single-chain structures extracted from protein complexes.

In this study, we demonstrated that protein complexes contain essential physical constraints derived from inter-chain interactions—priors that are fundamental to understanding protein dynamics. Artificially decoupling individual chains from their native interaction context limits models to learning static structural distributions rather than the underlying physics of conformational change. To validate this hypothesis, we present DynoM, a generative framework trained on static complex data and a newly generated dataset of all-atom trajectories encompassing 5,502 complex systems. During training, our observations revealed that while the incorporation of static complex data enhanced the fidelity of the conformational topology, MD-based fine-tuning functioned as a structural anchor to rectify non-physical motions, thereby systematically improving overall model performance. As a result, DynoM exhibited superior performance compared to existing state-of-the-art baselines. Interestingly, by addressing multi-domain monomers—a challenge rarely assessed in existing studies—DynoM captured long-range interactions and suppressed structural hallucinations, thus facilitating a more physically rigorous sampling of the conformational landscape. Furthermore, analyses of conformational transitions and the distributional shifts between bound and unbound states demonstrated that complex-derived constraints do not excessively restrict the model's ability to capture monomeric diversity. Overall, our work revealed that interfacial priors are an indispensable 'missing link' in current protein conformational ensemble prediction. DynoM can serve as a robust alternative in the protein modeling toolbox, and our newly generated dataset of 5,502 MD trajectories is provided to the research community facilitating further studies in protein conformational dynamics.

Results

Statistical analysis suggested that protein complexes could improve the efficiency of model learning

A comprehensive dataset was curated from the RCSB Protein Data Bank (PDB)[29] as of July 24, 2024, consisting of 64,591 complexes (52,001 heterodimers and 12,590 antigen-antibody complexes) and a ten-fold larger set of 648,749 single-chain proteins (see Methods). To ensure a non-redundant training set, sequences were clustered at a 70% sequence identity. This revealed a contrast in sequence redundancy between complexes and monomers: while the monomer set underwent a 12-fold reduction, the complex dataset exhibited a much higher degree of diversity with only a 3.1-fold decrease (Fig. 1b). Next, we assessed the flexibility of the representative structures within both the complex and monomer datasets. Following the methodology of Tang et al.[30], conformational flexibility was evaluated using the standardized radius of gyration, denoted as $R_g/N^{1/3}$, to eliminate the influence of chain length. The $R_g/N^{1/3}$ distribution for representative complexes is consistently higher than that of representative monomers (Fig. 1c). This indicates that complexes potentially encapsulate greater average flexibility, thereby offering more topological information for model learning.

Consequently, we built a diffusion-based generative framework capable of accommodating multi-chain inputs to integrate protein complexes into model training (Fig. 1d). Specifically, we employed a conditional diffusion architecture, where the modified template-free representation module derived from the AF3 architecture was used to encode the protein sequences and a positional encoding module was designed to differentiate input chain identities. For the diffusion module, the IPA-based transformer was utilized to iteratively reconstruct the protein backbone from random noise. The loss function comprises denoising score-matching loss, direct mean square error (MSE) loss on backbone coordinates, and local neighborhood loss on pairwise

atomic distances (≤ 8.5 Å). Additionally, the backbone steric clash loss was incorporated to prevent unphysical atomic overlaps within and between protein chains. To minimize computational overhead during training and inference, we optimized the number of diffusion blocks and denoising steps, finding that a configuration of four blocks and 200 steps provides sufficient sampling depth (Supplementary Figs. 1-2).

Static structures and dynamic trajectories provide complementary priors for ensemble modeling

To validate our hypothesis that complexes could provide richer structural priors for predicting conformational ensembles, we first conducted a comparative study on three models trained on different static structure data (Fig. 2a): Model_M (trained on 675k monomers), Model_C (trained on 62k complexes), and Model_A (trained on all structures). To evaluate the predictive performance of these models, we selected 30 monomer targets (length ranges from 132 to 479 AA) from the independent monomer test data (see Methods and Supplementary Table 1). Subsequently, all-atom MD simulations were performed for each protein, serving as the ground-truth for the test set. To assess the consistency of global characteristics between the predicted and the reference MD ensembles, we employed three commonly used metrics: PwD-JSD, R_g-JSD, and rRMSF (see Methods). The results showed that the model performance improved as training data increased, with Model_A exhibiting the best performance, followed by Model_M and Model_C, respectively (Pre-training stage in Fig. 2d-e, Supplementary Fig. 3). It is worth noting that the integration of complex data demonstrated superior data efficiency: even if the amount of complex data is only about 10% of that of the monomers, it can bring remarkable performance improvements.

As noted by Mu et al.[31] and Altis et al.[32], dihedral angles (Fig. 2b) provide a more robust description of internal and global protein motion compared to Cartesian coordinates. Thus, we

138 performed a time-lagged independent component analysis (TICA) on the backbone dihedrals of
139 the predicted ensembles and the MD data. In this analysis, the dihedral data were projected
140 onto the first two slowest time-lagged independent components (TICs). We then calculated the
141 TIC-JSD and TIC-RMWD (see Methods), to quantify the overlap between the MD-derived and
142 the predicted distributions in this subspace. Intriguingly, although the increased amount of
143 training data enabled Model_A to achieve better performance in global topological metrics, it did
144 not yield corresponding improvements in TIC-JSD (Pre-training stage in Fig. 2f, Supplementary
145 Fig. 3). This observation suggests that the heterogeneous training data of monomers and
146 complexes may introduce conflicting structural biases. Namely, while Model_A effectively
147 captures global characteristics, the conflicting structural biases between monomers and
148 complexes may hinder the model's ability to achieve precise local conformational motions.

149 Subsequently, we fine-tuned Model_C using the MD trajectories of protein complexes. As MD
150 data for complex systems are limited in public resources[33], we generated an in-house dataset of
151 5,502 protein complex systems. Each system was subjected to 100 ns all-atom MD simulations,
152 including 1,979 heterodimers and 3,523 antigen-antibody complexes (Supplementary Table 2).
153 To investigate the impact of the amount of trajectory data used for fine-tuning on model
154 performance, we fine-tuned Model_C using 1000, 2000, 3000, and all 5,502 trajectories
155 respectively (Fig. 2c). The results demonstrated that MD fine-tuning yielded systematic
156 improvements across all evaluation metrics (Fig. 2d-f). While fine-tuning with 1,000 to 3,000
157 trajectories yielded only marginal improvements, a more pronounced enhancement in model
158 performance was achieved only upon incorporating the full set of 5,502 trajectories. Analyses of
159 the first two slowest TICs demonstrate that MD fine-tuning enables the model to explore a
160 broader conformational space (Fig. 2g-h and Supplementary Fig. 4). As the trajectory volume
161 increased, the incidence of implausible dihedral angles decreased; this indicates that MD fine-

tuning allows the model to better approximate the underlying equilibrium distribution while functioning as a structural anchor to rectify non-physical motions.

Based on these insights into the complementary roles of static structures and MD trajectories, we proposed a generative model with optimal weights, named DynoM, for subsequent evaluations.

Interacting structures enhance fidelity in sampling equilibrium ensembles

To evaluate the performance of the complex-trained model, we benchmarked DynoM against three representative baselines for conformational ensemble prediction, including BioEmu (score-based diffusion), ConfDiff (conditional diffusion), and AlphaFlow (AF2-based flow-matching). As shown in Fig. 3a, DynoM involved remarkably less data to train the model than other baselines. Specifically, DynoM was trained using static complex structures consisting of 62,044 dimers and antigen-antibody complexes, which is only 4.85% to 34.58% of the data used by the other baseline models. Although all models employed an MD fine-tuning strategy, their trajectory data sources vary. ConfDiff and AlphaFlow leveraged the public ATLAS dataset comprising 1,390 protein chains[34], while BioEmu utilized a curated large-scale, in-house collection of peptide and monomer trajectories. In contrast, DynoM was fine-tuned using our newly generated dataset comprising 5,502 trajectories of dimer complexes, totaling over 550 μs.

Whereas current studies predominantly utilize relatively short proteins (mostly less than 300 amino acids)[35] from fast-folding proteins, BPTI, and ATLAS to evaluate models, we expanded upon our previous test set of 30 proteins to generate a test set containing 63 proteins of varying lengths and their 100 ns all-atom trajectories (Supplementary Table 1). Specifically, we supplemented our initial 30-protein set (132–479 AA) with 33 additional targets, comprising 7 small proteins (51–94 AA), 5 mid-sized proteins (268–355 AA), and 21 large-scale proteins

(531–775 AA). For these 63 targets, 100 ns all-atom MD trajectories were generated to serve as the ground truth for our comparative performance evaluations.

As shown in Fig. 3b-e, DynoM demonstrated superior performance compared with existing baselines across all equilibrium metrics in our independent test set. For the global topological metrics R_g-JSD, and PwD-JSD, DynoM outperformed all three baselines ($P < 0.05$, pairwise t-test). Regarding the dynamics-based metrics, TIC-RMWD and TIC-JSD, DynoM yielded the most robust results; this indicates its capacity to maximize coverage of the accessible conformational space (minimized TIC-JSD) while effectively suppressing the generation of non-physical or out-of-distribution states (minimized TIC-RMWD). Collectively, these results demonstrate that DynoM, which is trained exclusively on complex-derived data, exhibits exceptional data efficiency, thereby validating our hypothesis.

More importantly, to investigate how interfacial priors enhance the physical plausibility of conformational sampling, we evaluated the inter-domain consistency between the predicted and the reference MD ensembles. DynoM demonstrated superior performance relative to all baseline models, achieving highly accurate inter-domain positioning and effectively mitigating the structural hallucinations common in multi-domain monomeric predictions (Fig. 3f-h). For quantitative assessment, we calculated the deviation in average pairwise C_α-C_α distances between the predicted and the reference MD ensembles. As illustrated in the 2D heatmaps, the high consistency within individual protein domains is evident in the block-like patterns along the diagonal, reflecting the model's capacity to accurately preserve local fold integrity. Notably, the reduced deviations in off-diagonal regions highlight DynoM's superior proficiency in capturing long-range spatial interactions. To provide more intuitive visualizations of DynoM’s performance at the inter-domain level, we selected residue pairs spanning structural domains and analyzed their pairwise distance distributions. The histograms showed that DynoM exhibited extremely

high consistency with the dynamic behavior of molecular dynamics simulations. Moreover, we evaluated the performance consistency across the entire 63-protein test set for DynoM and the three baseline models (Supplementary Fig. 5). The results show that DynoM outperformed all three baselines in 23 out of 63 proteins and surpassed at least two baselines in an additional 24 cases (Supplementary Table 14), with comparable performance in the remaining cases. This performance enhancement is particularly evident for relatively large proteins, aligning with DynoM's effective integration of interfacial priors derived from protein complex data. Taken together, these results demonstrated DynoM's capacity to accurately capture equilibrium distributions across diverse structural scales ranging from single-domain to multi-domain monomers.

Sampling conformational changes of protein monomers

We subsequently addressed a pivotal concern: whether the integration of complex-derived constraints compromises the model's capacity to characterize the inherent conformational diversity of proteins in their monomeric states.

To address this, we assessed the capacity of DynoM to sample conformational transitions using a benchmark set of eleven proteins previously reported in BioEmu[18] and PathPre[36]. To ensure rigorous validation and mitigate potential data leakage, these targets were strictly filtered based on sequence similarity to our training set and their PDB release dates (Methods). The selected proteins encompass four distinct categories of conformational changes (Supplementary Table 15). The ensembles of the two proteins with local structural changes are displayed in Fig. 4a. Analysis of the two-dimensional free energy surfaces constructed within the backbone dihedral PCA space revealed that DynoM successfully identified local states that are highly matched to the MD-derived ground truth (Fig. 4b-c). Furthermore, DynoM demonstrated superior fidelity in predicting loop regions and secondary structures (Fig. 4c-d), outperforming BioEmu

(Supplementary Fig. 6). Similar analyses were conducted on the proteins with apo-holo transitions, rigid-body domain movements, and fold-unfold switches (Fig. 4e-g, Supplementary Figs. 7-10). For these two-state systems, DynoM was able to sample the multi-conformational landscape and capture both representative states. However, the predicted conformations tended to be clustered in the intermediate regions between the two states. This sampling bias likely reflects the lack of protein monomers with conformational changes in the training set.

Additionally, we assessed whether incorporating complex-derived constraints introduces systematic bias into the predicted conformational ensembles of proteins in their unbound states. Since antibody-antigen complexes were prevalent in our MD training set, we selected five systems for this analysis. We performed 100 ns all-atom MD simulations for both the bound complexes and the corresponding antigens in their unbound states. The TICA used on the backbone dihedrals of the MD data and the predicted ensembles exhibited the difference in antigen ensembles between the bound and unbound states. We next used DynoM to generate ensembles for the antigens in their unbound states and found that the predicted ensembles covered the conformational space for both the bound and unbound states (Supplementary Fig. 11). This suggested that training on complex-derived data does not over-constrain the model, enabling it to maintain the intrinsic conformational diversity of the proteins.

Discussion

In this study, we hypothesized that the protein complexes contribute to the prediction of protein monomer ensembles with greater efficiency than monomers. To validate this, we first demonstrated that the inclusion of a relatively small set of static protein complex data (62k, about 1/10 of monomer data) yields a more significant improvement in global structural metrics. Given that heterogeneous training data of monomers and complexes introduce structural biases in prediction, we thus fine-tuned the complex-trained model. And even with only 1000

trajectories, MD fine-tuning can significantly improve the performance. By acting as a structural anchor, MD fine-tuning re-weights the generative conformational space to align with the equilibrium distribution. Next, benchmarking on a curated set of 63 proteins showed that DynoM exhibited extremely high learning efficiency and can serve as a very robust alternative tool, compared to existing representative baselines. More importantly, owing to the learning of complex-derived interface constraints, DynoM effectively suppressed the structural hallucinations typically encountered in multi-domain monomer predictions. Furthermore, evaluations of multi-conformational generation and the modeling of antigens in bound/unbound states confirmed that these physical constraints derived from complex structures provide essential guidance without excessive constraints on conformational diversity of monomeric states.

Overall, a central premise of this work is that the physical priors derived from interactions can be leveraged to understand inter-chain dynamics. Naturally, this led us to explore whether DynoM could be directly applied to the prediction of protein-protein complexes. Preliminary evaluations were conducted using an in-house data set of 20 antigen-antibody targets, with 100 conformations generated for each target by DynoM and AF3, respectively. DynoM successfully distinguished rigid framework regions from flexible complementary determining regions (CDRs), exhibiting greater conformational diversity than AF3 (Fig. 5a). Regarding interface quality, AF3 demonstrated the predominant precision performance (Fig. 5b), achieving an average DockQ ≥ 0.23 in 8 targets and the best DockQ ≥ 0.23 in 15 targets. Interestingly, despite this gap in overall precision, DynoM identified plausible conformations in two specific cases where AF3 failed to resolve the correct binding mode even with 100 predicted conformations (Fig. 5c). We then quantified the predicted ensembles relative to the MD references using TM-score and interface RMSD (iRMSD) (Fig. 5d-e; Supplementary Fig. 12). While DynoM achieved a reasonable TM-score range of 0.6-0.8, the elevated iRMSD values highlighted significant

deviations in the interface residues. This suggested that although DynoM can roughly predict the overall complex structures, its interfacial precision is primarily constrained by the current training data volume. Further improvement requires the optimization of the loss function and the incorporation of interfacial constraints.

However, several limitations remain to be addressed in this work. First, a primary constraint lies in the scope of our training data, which only focuses on dimers and antibody-antigen complexes. Due to technical constraints on the model's maximum input length, larger multimeric assemblies have not yet been incorporated into the training process. Although we have optimized the memory usage for the existing framework, which supports sequence length up to 1250 AA (Supplementary Fig. 13), scaling the model to accommodate ultra-large complexes remains a significant challenge. Second, our model utilizes the IPA-based transformer, which is designed for backbone generation rather than end-to-end all-atom prediction. It lacks the advanced atomic encoders utilized by the state-of-the-art models such as Alphafold3[12], Boltz-1[37], and SimpleFold[38]. Our sidechain reconstruction is still performed by the all-atom coordinate reconstruction algorithm of AF2, which may limit the model's ability to capture fine-grained atomic interactions. Third, while our in-house test set provides valuable insights, its sample size is relatively limited for the validation study. Future efforts will involve expanding this benchmark to include a broader array of targets with diverse structural folds and varied assembly types. Finally, we did not include a comparative analysis with MSA sub-sampling methods, such as AF-Cluster[39], AFsample[40], or AFsample2[41]. Since our primary objective is to investigate how to refine generative frameworks, these methods are not considered in this study.

Building on the above findings, we propose three potential directions for future model optimization. First, the inclusion of protein complexes is essential for conformational ensemble prediction. With the relatively small volume of complex data (static structure and MD trajectories), the model still achieved a competitive performance. This suggests that the data

quality is more important than data quantity. This view aligns with BioEmu[18], whose training data consisted of a small core dataset (179,422 structures) curated from beyond 200 million structures from AFDB[42]. Meanwhile, complexes introduce interfacial priors that are inherently absent in monomeric structures. While current conditional generative models utilize co-evolutionary signals from MSAs or PLMs to sample conformations, these sequence-based priors—though reliable for conserved regions—often lead to structural hallucinations in regions with sparse evolutionary information. By leveraging protein complex data, our model incorporates interaction constraints that reduce these hallucinations, ensuring plausible conformational ensemble prediction.

Second, with the increase of amount and types in structural data and MD trajectories, how to build an efficient training strategy is quite essential. Our experiment demonstrated that the mixture of heterogeneous data allows the model to capture the correct 'average structure', it leads to biases in the backbone dihedral angle distributions. Therefore, training strategies for heterogeneous data urgently need optimization. Regarding the prevailing fine-tuning strategy, only one single conformation is randomly sampled from the MD ensembles and fed into the network in each training epoch. Since the number of fine-tuning epochs usually does not exceed a few dozen, this fine-tuning method prevents the model from exploring the structural diversity within MD ensembles, leading to underfitting of the model. We performed RMSD-based clustering and selected conformations with significant differences from different clusters to fine-tune the model. This fine-tuning regime yielded improved performance in terms of TIC-JSD (Supplementary Fig. 14), highlighting the importance of conformational diversity in the fine-tuning process. Therefore, developing new fine-tuning strategies will be essential for improving learning efficiency.

Third, accurate conformational ensemble prediction of biomolecular interactions is a crucial direction, which is of great significance for drug design and screening. Our proposed

architecture established a proof-of-concept for protein-protein complex dynamics prediction, but significant room for improvement remains, such as data accumulation and curation, atomic-level prediction, and robust interface modeling. Future work will focus on introducing geometric features and physicochemical priors[43,44] to enhance the modeling of complex dynamic systems. This will not only provide profound mechanistic insights into the nature of biomolecular interactions but will also catalyze a paradigm in high-throughput drug screening and rational drug design.

Methods

Data for model training

In this study, we utilized static structure data for pre-training and MD trajectories for fine-tuning. Thus, we described the data processing details below.

Protein monomer data. All available crystal structures with a resolution ≤ 5.0 Å were downloaded from the RCSB Protein Data Bank (as of 1st February 2025), totaling 213,260 entries. Following the data processing protocols of AlphaFold2[10] and RFdiffusion[45], both monomers and isolated chains extracted from complexes underwent a filtering and processing pipeline (Supplementary Note 1). As a result, a total of 647,859 single-chain protein structures were obtained. Based on their release dates, the dataset was divided into a training set of 613,872 structures released on or before July 23, 2024. Finally, the training sequences were clustered at a 70% sequence identity threshold using MMseqs2 (v17-b804f)[46], resulting in 51,349 representative clusters.

Protein complex data. Our complex data includes crystal structures of protein dimers and antigen-antibody complexes. According to the number of chains, protein dimers were directly retrieved from the previously downloaded 213,260 entries. And then, homodimers and

heterodimers were processed using the pipeline in Supplementary Note 2 and a total of 52,069 structures were obtained. For training, we included 50,852 structures released on or before July 23, 2024. Moreover, 9,532 antigen-antibody PDB IDs were retrieved from the SAbDab database[47] (as of May 16, 2025), and the corresponding structure files were downloaded from the PDB. Each complex containing multiple antibodies was disassembled into separate antigen-antibody structure files. Subsequently, these structures underwent screening and processing (see Supplementary Note 3 for details). Based on PDB release dates, antigen-antibody structures released on or before July 23, 2024, constituted the training set, totaling 11,192 structures.

In-house MD data. For model fine-tuning, an MD trajectory dataset comprising a total of 5,502 protein complexes was generated using GROMACS v2024.4[48] with the CHARMM36m force field[49]. Each system was built in a cubic box, solvated with TIP3P water, and neutralized with Na^+/Cl^-. Energy minimization consisted of 50,000 steepest-descent steps with backbone restraints (1,000 $kJ \cdot mol^{-1} \cdot nm^{-2}$), followed by 20,000 unconstrained conjugate-gradient steps to remove steric clashes. The systems were then equilibrated in two 100-ps stages (NVT followed by NPT) using a 1-fs time step and a leap-frog integrator. Temperature was fixed at 300 K using a V-rescale thermostat (τ_T = 1 ps), and pressure was maintained at 1 bar using a Parrinello-Rahman barostat (τ_P = 5 ps). Finally, 100-ns production runs were performed for each complex with a 2-fs time step, saving coordinates every 10 ps. Hydrogen bonds were constrained via LINCS, and long-range electrostatics were handled using PME.

Data for model evaluation

Independent monomer test data. A total of 63 protein monomers was constructed to evaluate our model, comprising 30 proteins (132-479 AA) and 33 additional proteins (51-775 AA). The first 30 proteins were derived from 7,287 structures released between February and July 2025.

And the rest 33 proteins were obtained from 5,964 entries deposited between July and November 2025. After preprocessing (Supplementary Note 1), we excluded targets sharing > 40% sequence identity with the training data to prevent data leakage. The retained targets were clustered at a 70% identity threshold and sampled by sequence length (Supplementary Table 1,8).

Multi-conformation data. To assess the multi-conformation sampling capabilities of DynoM, we retrieved PDB IDs from the BioEmu and PathPre datasets (Supplementary Table 15). Data curation followed the pipeline described in Supplementary Note 1. To prevent data leakage, we only retained proteins which had less than 40% identity with the training set sequences. As a result, only two targets from the BioEmu were obtained. Since some pairs of structures in the PathPre are not completely identical, we compared all possible pairs and only kept those pairs that were 100% identical. As a result, nine pairs were randomly selected with significant differences (RMSD $\geqslant$ 2.0 Å, TM-score $\leqslant$ 0.8).

Antigen-antibody complex data. Based on the filtering and processing pipeline described for the construction of the antigen-antibody complex training set, we obtained 1,398 antigen-antibody complexes to serve as a validation set (released between January 25, 2024, and May 16, 2025). From this collection, we randomly selected 20 structures to evaluate the performance of DynoM. The specific PDB IDs are listed in Supplementary Table 16.

Model architecture

Inspired by the models of protein conformation, we represent a template-free AF3 trunk to make the model compatible with the input and representation of multi-chain proteins, as shown in Fig. 1d.

The encoder module. This module consists of a template-free AF3 trunk module and a positional encoding module. The template-free AF3 trunk module encodes protein sequences and the MSA results to generate single and pair AF3 embeddings. In particular, the input sequences are first encoded by the Input Embedder and then are fed into the MSA module with MSA result, obtaining the single and pair embeddings. Next, single and pair embeddings are updated using the PairFormer module with 10 recycling iterations. The positional encoding module is used to generate single and pair positional embeddings. For single-chain protein input, the model generates a zero-indexed positional vector and a relative position matrix. For multi-chain input, the model generates a positional vector for each chain and concatenates them according to the chain order. Regarding the relative position matrix, values in off-diagonal regions of inter-chain are filled with zeroes. In the next step, the positional vector and the relative position matrix are processed using sinusoidal functions to obtain the single and pair positional embedding. Finally, the single and pair AF3 embeddings are concatenated to the single and pair positional embeddings, respectively.

The diffusion module. Before the single and pair embeddings are fed into the diffusion module, the diffusion timesteps are encoded using sinusoidal functions to obtain the sinusoidal time embedding. Both the single and pair embeddings are initialized with the corresponding embeddings and the sinusoidal time embedding. The pair embedding is additionally concatenated with the relative sequence distances, which are obtained by encoding the backbone translation vectors from random rigids through a Gaussian radial basis function. The concatenated embeddings (single and pair embeddings) are fused through a three-layer MLP, respectively. The diffusion module comprises four denoising blocks. Each block integrates Invariant Point Attention (IPA), a Transformer layer, and MLPs. The module employs IPA and Transformer to iteratively update the single embedding. Based on message-passing mechanism, an MLP is employed to further update the pair embedding via the single

embedding. Inspired by the backbone update and all-atom coordinate refinement modules from AF2, we employ an MLP to predict the coordinates of the backbone frames and the protein torsion angles directly from the updated single embedding. Overall, guided by single and pair embeddings, the diffusion module samples protein structures from random rigids. If the input sequence is a multi-chain protein, the diffusion module uses position vectors to distinguish the protein chain.

The loss functions. The loss functions consist of a primary denoising score matching loss and three auxiliary terms, including a direct MSE loss $\mathcal{L}_{bb}$ on backbone positions, a local neighborhood loss $\mathcal{L}_{2D}$ supervising pairwise distances, and a steric clash penalty $\mathcal{L}_{clash}$.

The diffusion model on $SE(3)^N$ has demonstrated powerful generative capability and has been widely applied to protein design and the prediction of protein conformational ensembles. To model protein structure dynamics, we parameterized the diffusion framework on $R^{3N} \times SO(3)^N$and optimized the network using a denoising score matching loss. For a given initial conformation $x_0 \sim q_0(x_0)$, the forward diffusion process of the model is defined by a stochastic differential equation (SDE):

$$dx_t = \mathbf{f}(x_t, t)dt + \mathbf{G}(t)d\mathbf{W}_t$$

where

$$\mathbf{f}(x_t, t) = \begin{bmatrix} f_{\text{trans}}(T_t, t) \\ f_{\text{rot}}(R_t, t) \end{bmatrix}, \mathbf{G}(t) = \begin{bmatrix} \sqrt{\beta_t} & \mathbf{0} \\ \mathbf{0} & \sqrt{\frac{d}{dt}\sigma_t^2} \end{bmatrix}$$

For the translation component $T_t \in \mathbb{R}^{3N}$ and rotation component $R_t \in SO(3)^N$, the forward diffusion trajectories are defined as follows:

$$dT_t = -\frac{1}{2}\beta_t P_t R_t dt + \sqrt{\beta_t} P_t dW_t \qquad (T_t \in \mathbb{R}^{3N})$$

446

$$dR_t = \sqrt{\frac{d}{dt}\sigma_t^2}\, dW_t^{SO(3)} \qquad (R_t \in SO(3)^N)$$

447 where dW_t is the standard Wiener process in Euclidean space, and $dW_t^{SO(3)}$represents the

448 Brownian motion on the $SO(3)$ manifold, satisfying $W_t \in \mathbb{R}^{3N}, W_t^{SO(3)} \in SO(3)^N$.

449 To control diffusion trajectories, we defined a linear noise schedule $\beta_t = \beta_{min} + (\beta_{max} - \beta_{min})t$,

450 and a logarithmic diffusion scale $\sigma(t) = log(te^{\sigma_{max}} + (1-t)e^{\sigma_{min}})$. Following empirical

451 protocols in geometric diffusion models [Ref framediff], the hyperparameters are set to $\beta_{min} =$

452 $0.1, \beta_{max} = 20.0$, $\sigma_{min} = 0.1$ and $\sigma_{max} = 1.5$.

453 Based on the SDEs defined above, we perturb the protein backbone's translations and rotations

454 by adding noise. The transition distribution for the translation component follows a Gaussian

455 distribution:

456 $$q_t(T_t|T_0) = N(T_t; \sqrt{\alpha_t}T_0, (1-\alpha_t)I)$$

457 where $\alpha_t := exp(-\int_0^t \beta(s)ds)$。For the rotation component, the noise distribution is defined on

458 the $SO(3)$ manifold via the Isotropic Gaussian on $SO(3)$ ($IGSO_3$):

459 $$p_t(R_t|R_0) = IGSO_3(R_t; R_0, \sigma_t^2) = \Sigma_\ell (2\ell+1) e^{(-\ell(\ell+1)/2\sigma_t^2)} sin((\ell+0.5)\omega_t/\ell)/sin(\omega_t/2)$$

460 where $\omega_t = Axis.angle(R_0 R_t)$. Correspondingly, the reverse time diffusion process follows

461 $$dx_t = [\mathbf{f}(x_t, t) - \mathbf{G}(t)\mathbf{G}(t)^T \nabla_{x_t} \log\, q_t(x_t)]dt + \mathbf{G}(t)d\overline{\mathbf{W}}_t$$

Considering the intractable marginal score function $\nabla \log\, q_t$, we trained a neural network s_θ to estimate the conditional score. For the translational score $s_\theta(T_t, t)$ and rotational scores $s_\theta(R_t, t)$, the denoising score matching loss functions are defined as

$$\mathcal{L}_{DSM}^{T} = \mathbb{E}_{t,T_0,T_t}[\lambda(t) \parallel s_\theta(T_t, t) - \nabla_{T_t} \log\, p_t(T_t|T_0) \parallel^2]$$

$$\mathcal{L}_{DSM}^{R} = \mathbb{E}_{t,R_0,R_t}[\lambda^r(t) \parallel s_\theta^r(R_t, t) - \nabla_{R_t} \log\, p_t(R_t|R_0) \parallel_{SO(3)}^2]$$

where $\lambda(t)$, $\lambda^r(t)$ is a time weighting function and the conditional score functions are defined by:

$$\nabla_T log p_t(T_t|T_0) = -(T_t - \sqrt{\alpha_t} T_0)/(1 - \alpha_t)$$

$$\nabla_R log p_t(R_t|R_0) = R_t(R_0^T R_t)\omega_t \nabla_\omega log IGSO_3(R_t; R_0, \sigma_t^2)$$

To prevent bond breakage and enhance the local structural quality of the generated protein, we added the mean squared error (MSE) loss on backbone positions $\mathcal{L}_{bb}$ and the local neighborhood loss on pairwise atomic distances $\mathcal{L}_{2D}$:

$$\mathcal{L}_{bb} = \frac{1}{4N} \sum_{n=1}^{N} \sum_{a \in \Omega} \parallel a_n^{(0)} - \hat{a}_n^{(0)} \parallel^2$$

$$\mathcal{L}_{2D} = \frac{1}{Z} \sum_{n,m=1}^{N} \sum_{a,b \in \Omega} \mathbb{I}\{d_{ab}^{nm} < 0.85\} \parallel d_{ab}^{nm} - \hat{d}_{ab}^{nm} \parallel^2,$$

$$Z = \left(\sum_{n,m=1}^{N} \sum_{a,b \in \Omega} \mathbb{I}\{d_{ab}^{nm} < 0.85\} \right) - N.$$

where $\mathbb{I}\{d_{ab}^{nm} < 0.85\}$ is an indicator variable to only penalize atoms that within 0.85 nm (i.e. 8.5 Å). Since the quality of generated protein complexes often suffer from inter-chain steric clashes, we incorporated an additional clash penalty $\mathcal{L}_{clash}$. This loss term is used to reduce both intra-

and inter-chain atomic clashes, thereby ensuring the physical plausibility of the generated protein structures:

$$\mathcal{L}_{clash} = \frac{1}{N_{clash}} \sum_{i=1}^{N_{nbpairs}} \max\left(d_{lit}^{i} - \tau - d_{pred}^{i}, \mathbf{0}\right)$$

where N_{clash} is the number of backbone atom pairs i that satisfy the condition $d_{lit}^{i} - \tau > d_{pred}^{i}$. Here, d_{lit}^{i} represents the van der Waals radii of corresponding backbone atoms, and τ is a tolerance factor set to 1.5 Å[50]. d_{pred}^{i} is the Euclidean distance of the atom pair i in the predicted structure.

The total training loss $\mathcal{L}_{total}$ is defined as the weighted sum of the denoising score matching loss and the three auxiliary losses:

$$\mathcal{L}_{total} = \mathcal{L}_{DSM}^{T} + 0.5 * \mathcal{L}_{DSM}^{R} + 0.25 * \{t < (T_F = 0.25)\}(\mathcal{L}_{bb} + \mathcal{L}_{2D}) + 0.3 * L_{clash}$$

Model training

The DynoM model adopted a pretraining-fine-tuning strategy, in which DynoM was first pretrained on static protein structure data and subsequently fine-tuned on the in-house MD trajectories.

In the pretraining stage, the model learned the prior distribution of static protein structures, thereby capturing the inherent geometric rule of natural protein folds. To de-duplicate highly similar protein structures and improve the model's generalization ability, we implemented a cluster-based sampling strategy during training. Specifically, we initially clustered the sequences in the training set at a 70% similarity threshold. In each training epoch, we randomly selected n clusters (n = batch size) and sampled one sequence from each cluster as the model input. Meanwhile, we use mmseqs (v17-b804f) with default parameters for multi-sequence alignment

search, which is implemented through the colabfold_search from ColabFold[51]. The databases used for the search include UniRef30 (v2103), PDB70 (v220313), and ColabFoldDB (v202108). In the fine-tuning stage, we optimized the pre-trained model using an in-house MD dataset. We used GROMACS to sample the conformations at 50-ps intervals. Consequently, each 100-ns trajectory yielded 2,000 conformations for the subsequent fine-tuning. In each training epoch, we randomly select one conformation from 2,000 conformations for each trajectory and use it for model training.

To address missing coordinates in experimentally determined structures, we implemented a masking mechanism that zeros out the loss contributions for these unobserved residues. This strategy effectively prevents unresolved regions from interfering with gradient updates. We used the Adam optimizer to optimize the model parameters, with an initial learning rate of 3e-4 and zero weight decay. To improve training stability and accelerate convergence, we employed a learning rate schedule based on a linear warmup followed by adaptive decay. Specifically, a linear warmup was applied for the first 5,000 steps to stabilize early training. Subsequently, the learning rate was halved (decay factor = 0.5) once the validation loss plateaued for 10 consecutive epochs. We trained the model using eight NVIDIA A100 GPUs (80 GB). The hyperparameter settings are provided in Supplementary Table 17.

Evaluations metrics

Jensen–Shannon divergence of the radius of gyration (R_g-JSD). The radius of gyration (R_g) is calculated using the mass-weighted root-mean-square distance r_{cm} from the C_α atom to its center of mass r_{cm} as

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i |\mathbf{r}_i - \mathbf{r}_{\mathrm{cm}}|^2}{\sum_{i=1}^{N} m_i}}$$

where N is the total number of C_α atoms, m_i is the mass of the i-th C_α atom, r_i is the coordinate of the i-th C_α atom, and r_{cm} is calculated as

$$r_{cm} = \frac{\sum_{i=1}^{N} m_i r_i}{\sum_{i=1}^{N} m_i}$$

To compare the difference between the distribution of reference MD and the predicted, we quantify the R_g data as a histogram with 50 bins after computing the R_g for each conformation in the distribution. Then we calculate the Jensen-Shannon Divergence (JSD), defined as:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M)$$

where P and Q represent the probability distributions of the reference data and the generated data, respectively, $M = \frac{1}{2}(P + Q)$, and D_{KL} denotes the Kullback-Leibler divergence.

Jensen–Shannon divergence on pairwise C_α atomic distance distributions (PwD-JSD). Initially, the predicted structure is aligned with the reference MD structure. For each unique pair of C_α atoms, its distance distribution is quantified as a histogram with 50 bins and normalized to a probability distribution. The JSD between each predicted distribution and the reference MD distribution is then calculated independently. Finally, the overall PwD-JSD score is calculated by

$$PWD - JSD = \frac{1}{N_{pairs}} \sum_{k=1}^{N_{pairs}} JSD(P_k \parallel Q_k)$$

where k represents the index of the specific C_α atom pair, and N_{pairs} is the total number of unique C_α atom pairs.

Pearson correlation of Root Mean Square Fluctuation (rRMSF). The predicted structure is aligned with the reference structure based on common backbone atoms to eliminate global translational and rotational motions. The RMSF of the i-th residue C_α atom is defined as

$$RMSF_i = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\|r_i(t) - \bar{r}_i\|^2}$$

where T is the total number of frames in the ensemble, $r_i(t)$ represents the coordinates of the i-th C_α atom in conformation t after spatial alignment, and $\bar{r}_i$ is its mean position averaged over the entire trajectory.

To evaluate the consistency of residue-level flexibility between the predicted and the reference MD, the Pearson correlation coefficient between their respective RMSF curves is calculated

$$r_{pearson} = \frac{\sum_{i=1}^{L}(RMSF_i^{pred} - \overline{RMSF}^{pred})(RMSF_i^{ref} - \overline{RMSF}^{ref})}{\sqrt{\sum_{i=1}^{L}(RMSF_i^{pred} - \overline{RMSF}^{pred})^2 \sum_{i=1}^{L}(RMSF_i^{ref} - \overline{RMSF}^{ref})^2}}$$

where L represents the total number of matched C_α atoms, $RMSF_i^{pred}$ and $RMSF_i^{ref}$ are the RMSF values of residue i from the model and the MD reference, respectively. $\overline{RMSF}$ represents the mean RMSF value across all residues within the respective ensemble.

Jensen-Shannon divergence on the top-two TICs (TIC-JSD). TICA is a linear transformation method used for dimensionality reduction, which projects trajectory data onto its slowest components by maximizing the autocorrelation of the transformed coordinates over a specified lag time. To construct the feature space, we use the MDTraj library (v1.10.3) to extract the

backbone dihedral angles (φ, the dihedral of C–N–Cα–C, and ϕ, the dihedral of N–Cα–C–N) as input features from the conformations aligned with the reference structure. Periodic boundary condition corrections are performed by converting these angles into sin and cos. Next, the TICA model was fitted solely on the reference MD trajectory features using the MSMBuilder (v3.8.0) after feature extraction. Both the MD conformations and the predicted conformations are projected onto the first two dominant independent principal components (TIC 1 and TIC 2). To compare the difference between the two distributions, the two-dimensional projection space was discretized into a histogram containing 50 bins. A pseudo-count (ϵ=1e-12) is added to smooth the probability distributions and prevent zero-division. The difference between the two-dimensional probability distributions of the reference data (P) and the predicted data (Q) is measured using JSD, computed as

$$TIC - JSD = (JSD(P \parallel Q))^2$$

Root-Mean-Square Wasserstein Distance (TIC-RMWD). To focus on the local differences of TICs, we introduce the root-mean-square Wasserstein distance (RMWD) as a complement to TIC-JSD. The one-dimensional Wasserstein distance (W_1) was independently calculated for the projections along the TIC 1 and TIC 2 axes. RMWD is defined as the squared mean of these two distances

$$RMWD = \sqrt{\frac{W(X_{ref}, X_{gen})^2 + W(Y_{ref}, Y_{gen})^2}{2}}$$

where X_{ref} and Y_{ref} represent the marginal distributions of the reference MD along the respective TICs, and X_{gen} and Y_{gen} denote those of the generated ensembles.

Data availability

The datasets generated and analyzed for training and evaluating DynoM in this study are available on GitHub at https://github.com/RPDGroup/DynoM. For the multi-conformation data, the original structures were derived from the datasets proposed by BioEmu (https://github.com/microsoft/bioemu-benchmarks) and PATHpre (https://zenodo.org/doi/10.5281/zenodo.13337019). The additional 20 antigen-antibody complexes were retrieved from the PDB database. Detailed descriptions of the data cleaning, sequence alignment, and standardization processes are provided in Supplementary Note 1-3. The final complete list of 9 pairs of multi-conformation sequences and 20 antigen-antibody complex structures is provided in Supplementary Table 15-16.

Code availability

Our model was developed using Python, PyTorch, Numpy, and PyTorch-Lightning. We used Biopython, MDTraj[52], MSMBuilder and Protenix to extract the initial protein sequence and structural representation. The data analysis and plotting for the experiment were performed using Python, NumPy, Matplotlib, MDTraj, seaborn, SciPy, pandas, and Biopython. Protein structure visualization and rendering were performed using Pymol v.3.1.6.1 (https://github.com/schrodinger/pymol-open-source)[53] and ChimeraX v.1.9[54]. All Rosetta protocols were implemented under PyRosetta4[55]. And the source code, pre-trained models, and evaluation pipelines for DynoM developed in this study are available on GitHub at https://github.com/RPDGroup/DynoM.

Reference

1. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
2. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat Chem Biol* **5**, 789–796 (2009).
3. Nussinov, R. & Tsai, C.-J. Allostery in disease and in drug discovery. *Cell* **153**, 293–305 (2013).
4. Ma, M. *et al.* N-terminal PEGylation enhances organophosphorus hydrolase catalysis for a promising fast and long-acting prophylactic candidate. *Journal of Hazardous Materials* **488**, 137336 (2025).
5. Karplus, M. & McCammon, J. A. Molecular dynamics simulations of biomolecules. *Nat Struct Mol Biol* **9**, 646–652 (2002).
6. Shaw, D. E. *et al.* Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science* **330**, 341–346 (2010).
7. Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics. *J. Chem. Phys.* **151**, 070902 (2019).
8. De Vivo, M., Masetti, M., Bottegoni, G. & Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **59**, 4035–4061 (2016).
9. Perilla, J. R. *et al.* Molecular dynamics simulations of large macromolecular complexes. *Current Opinion in Structural Biology* **31**, 64–74 (2015).
10. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
11. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*.

12. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* 1–3 (2024) doi:10.1038/s41586-024-07487-w.
13. Yim, J. *et al.* SE(3) diffusion model with application to protein backbone generation. Preprint at https://doi.org/10.48550/arXiv.2302.02277 (2023).
14. Yim, J. *et al.* Fast protein backbone generation with SE(3) flow matching. Preprint at https://doi.org/10.48550/arXiv.2310.05297 (2023).
15. Wu, K. E. *et al.* Protein structure generation via folding diffusion. *Nat Commun* **15**, 1059 (2024).
16. Lu, J., Zhong, B., Zhang, Z. & Tang, J. Str2Str: A Score-based Framework for Zero-shot Protein Conformation Sampling. in (2023).
17. Wang, Y. *et al.* Protein Conformation Generation via Force-Guided SE(3) Diffusion Models. Preprint at https://doi.org/10.48550/arXiv.2403.14088 (2024).
18. Lewis, S. *et al.* Scalable emulation of protein equilibrium ensembles with generative deep learning. 2024.12.05.626885 Preprint at https://doi.org/10.1101/2024.12.05.626885 (2024).
19. Zhu, J. *et al.* Accurate Generation of Conformational Ensembles for Intrinsically Disordered Proteins with IDPFold. *Advanced Science* **n/a**, e11636.
20. Wang, B. *et al.* Conditional diffusion with locality-aware modal alignment for generating diverse protein conformational ensembles. *Nat Mach Intell* 1–20 (2026) doi:10.1038/s42256-026-01198-9.
21. Jing, B. *et al.* EIGENFOLD: GENERATIVE PROTEIN STRUCTURE PREDICTION WITH DIFFUSION MODELS. (2023).
22. Zheng, S. *et al.* Predicting equilibrium distributions for molecular systems with deep learning. *Nat Mach Intell* **6**, 558–567 (2024).
23. Jing, B., Berger, B. & Jaakkola, T. AlphaFold Meets Flow Matching for Generating Protein Ensembles.

24. Li, S. *et al.* F^{3}low: Frame-to-Frame Coarse-grained Molecular Dynamics with SE(3) Guided Flow Matching. Preprint at https://doi.org/10.48550/arXiv.2405.00751 (2024).
25. Jin, Y. *et al.* P2DFlow: A Protein Ensemble Generative Model with SE(3) Flow Matching. *J. Chem. Theory Comput.* **21**, 3288–3296 (2025).
26. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
27. van Breugel, M., Rosa e Silva, I. & Andreeva, A. Structural validation and assessment of AlphaFold2 predictions for centrosomal and centriolar proteins and their complexes. *Commun Biol* **5**, 312 (2022).
28. Chowdhury, R. *et al.* Single-sequence protein structure prediction using a language model and deep learning. *Nat Biotechnol* **40**, 1617–1623 (2022).
29. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242 (2000).
30. Tang, Q.-Y., Ren, W., Wang, J. & Kaneko, K. The Statistical Trends of Protein Evolution: A Lesson from AlphaFold Database. *Mol Biol Evol* **39**, msac197 (2022).
31. Mu, Y., Nguyen, P. H. & Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* **58**, 45–52 (2005).
32. Altis, A., Nguyen, P. H., Hegger, R. & Stock, G. Dihedral angle principal component analysis of molecular dynamics simulations. *J Chem Phys* **126**, 244111 (2007).
33. Mokhtari, O., Bignon, E., Khakzad, H. & Karami, Y. DynaRepo: the repository of macromolecular conformational dynamics. *Nucleic Acids Res* **54**, D393–D401 (2026).
34. Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C. & Galochkina, T. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research* **52**, D384–D392 (2024).
35. Shen, Y. *et al.* ConfRover: Simultaneous Modeling of Protein Conformation and Dynamics via Autoregression. Preprint at https://doi.org/10.48550/arXiv.2505.17478 (2025).

36. Hu, Y. *et al.* Exploring Protein Conformational Changes Using a Large-Scale Biophysical Sampling Augmented Deep Learning Strategy. *Advanced Science* **11**, 2400884 (2024).
37. Wohlwend, J. *et al.* Boltz-1 Democratizing Biomolecular Interaction Modeling. *bioRxiv* 2024.11.19.624167 (2025) doi:10.1101/2024.11.19.624167.
38. Wang, Y., Lu, J., Jaitly, N., Susskind, J. & Bautista, M. A. SimpleFold: Folding Proteins is Simpler than You Think. Preprint at https://doi.org/10.48550/arXiv.2509.18480 (2025).
39. Wayment-Steele, H. K. *et al.* Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).
40. Wallner, B. AFsample: improving multimer prediction with AlphaFold using massive sampling. *Bioinformatics* **39**, btad573 (2023).
41. Kalakoti, Y. & Wallner, B. AFsample2 predicts multiple conformations and ensembles with AlphaFold2. *Commun Biol* **8**, 373 (2025).
42. Fleming, J. *et al.* AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. *Journal of Molecular Biology* **437**, 168967 (2025).
43. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat Methods* **17**, 184–192 (2020).
44. Stebliankin, V. *et al.* Evaluating protein binding interfaces with transformer networks. *Nat Mach Intell* **5**, 1042–1053 (2023).
45. Bennett, N. R. *et al.* Atomically accurate de novo design of antibodies with RFdiffusion. *Nature* **649**, 183–193 (2026).
46. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**, 1026–1028 (2017).
47. Olsen, T. H., Boyles, F. & Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Sci* **31**, 141–146 (2022).

48. Van Der Spoel, D. *et al.* GROMACS: fast, flexible, and free. *J Comput Chem* **26**, 1701–1718 (2005).
49. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **14**, 71–73 (2017).
50. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. 2021.10.04.463034 Preprint at https://doi.org/10.1101/2021.10.04.463034 (2022).
51. Kim, G. *et al.* Easy and accurate protein structure prediction using ColabFold. *Nat Protoc* **20**, 620–642 (2025).
52. McGibbon, R. T. *et al.* MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (2015).
53. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. (2015).
54. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).
55. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

Funding sources

This work was supported by the National Natural Science Foundation of China (32501102).

Author contributions

Z.R., D.L., Q.Y.J., and Y.L. conceived the concept of this study and designed the workflow. Z.R. and D.L. trained the DynoM network architecture. Y.L. and Y.Z. prepared the static protein structure data, while Q.T.J., C.L., and L.H. curated the MD simulation trajectory data. Q.Y.J.,

D.L., Q.T.J., and Z.S. contributed to the analysis and interpretation of the results. Z.R., D.L., Q.Y.J., Q.T.J., X.Y., P.S., Z.F., C.W., M.N., and X.B. assisted in the experimental design. Z.R., M.N., D.L., Q.Y.J., Q.T.J., and Y.L. drafted the manuscript and prepared the figures. All authors provided critical feedback, helped shape the research, and revised the paper.

Competing interests

The authors declare no competing interests.

Figures

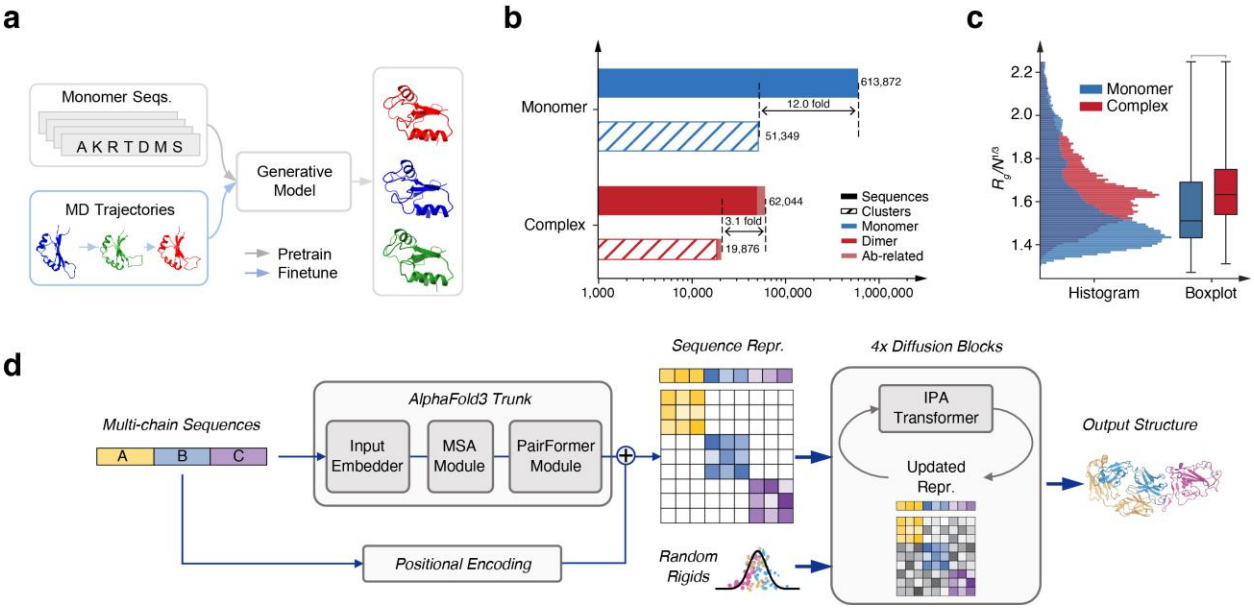
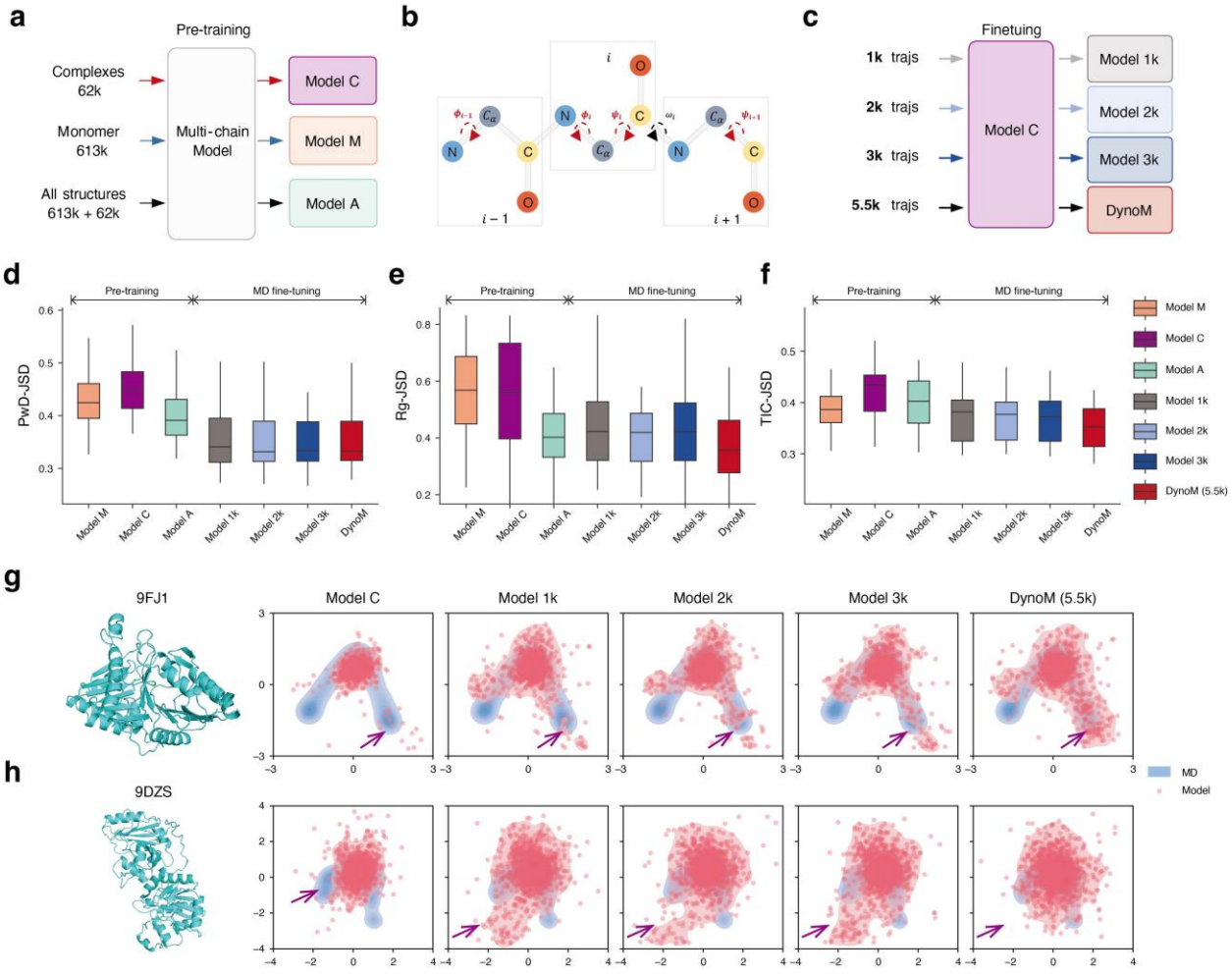


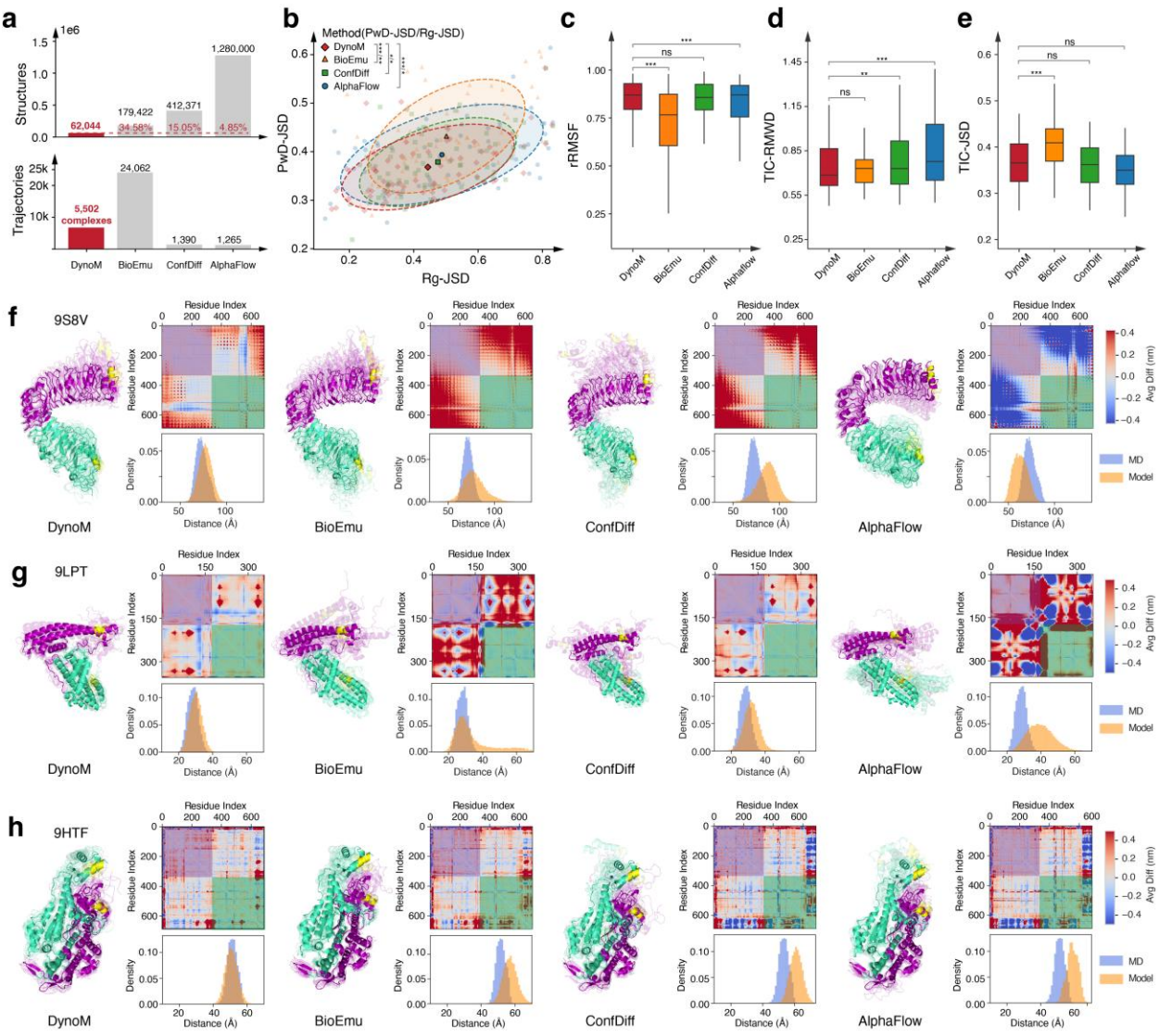
Fig. 1 | **Overview of the study design. a,** The prevailing training strategy for conformational ensemble prediction model: pre-training on large-scale static structures, followed by fine-tuning on MD trajectories. **b,** The total counts of sequences and non-redundant clusters for monomer and complex structures collected from PDB. Sequences clustering at 70% identity were conducted to remove redundancy, leading to a 12.0-fold (91.64%) reduction for monomers and a 3.1-fold (67.98%) reduction for complexes. **c,** Distributions of the standardized radius of gyration ($R_g/N^{1/3}$) for the monomer and complex datasets. **d,** Schematic of the architecture built in this study. Multi-chain sequences are processed through an AlphaFold3-based trunk and a positional encoding module to generate feature representations. These representations guide a four-block diffusion process to sample final structures from random rigid bodies.



738

739 **Fig 2.** | **Comparative performance study on model trained with different datasets. a**, Three
740 models were trained separately using different static structures, including Model_C (trained on
741 62k complexes), Model_M (trained on 613k monomers), and Model_A (trained on both complex
742 and monomer data). **b**, Schematic of protein backbone dihedral angles (ϕ and ψ). **c**, Four
743 models were fine-tuned separately on Model_C using various number of MD trajectories (1k, 2k,
744 3k, and 5,502). **d**, The JS Divergence of pairwise Cα-Cα atomic distance distributions (PwD-
745 JSD) between the predicted and the reference MD. **e**, The JS Divergence of radius of gyration
746 distributions (Rg-JSD) between the predicted and the reference MD. **f**, The JS Divergence on
747 the top-two TICs of backbone dihedral angle distributions between the predicted and the
748 reference MD. **g-h**, For two protein cases (PDB IDs: 9FJ1 and 9DZS), the left panels show the

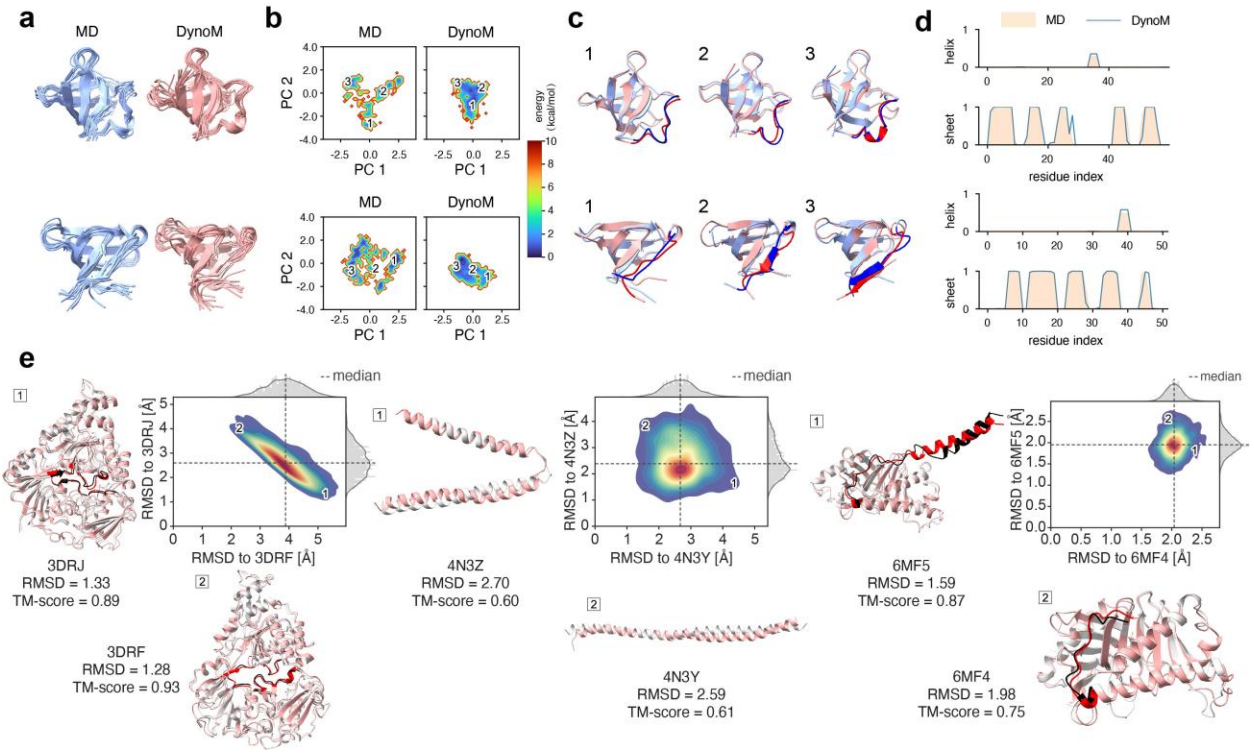
reference structures. The right panels show the conformations generated by MD and models with varying amounts of fine-tuning data, projected onto the first two time-lagged independent components. The blue contours represent the reference MD, while the red scatter points represent the predicted conformations.



754

755 **Fig. 3 | Performance of DynoM and three other representative generative models on an**
756 **independent test set comprising 63 proteins. a,** The number of static protein structures and
757 MD trajectories used in different methods. The upper bar plot shows the number of static
758 structures used for pre-training, with percentages indicating the relative proportion of the DynoM
759 dataset (62,044 structures) compared to BioEmu, ConfDiff, and AlphaFlow. The lower bar plot
760 shows the number of MD trajectories used for fine-tuning. **b**, The scatter plot shows the PwD-
761 JSD and Rg-JSD of four methods for 63 proteins. Centroids (symbols with black outlines)
762 represent the mean performance across the test set. Statistical significance is determined by a

one-sided paired t-test (ns, not significant; *P < 0.05, **P < 0.01, ***P < 0.001). **c-e**, The box plots of rRMSF (c), TIC-RMWD (d) and TIC-JSD(e) between the predicted and the reference MD. **f-h,** Conformations generated by four models for three long protein monomers: 9S8V (**f**), 9LPT (**g**) and 9HTF (**h**). The 2D heatmap illustrates the residue-to-residue deviation of average pairwise Cα-Cα distances between the predicted and ground-truth MD ensembles. Red indicates that the distance is overestimated, and blue indicates that the distance is underestimated. The density plots show the pairwise distance distribution of residue pairs in the specific regions (marked in yellow on ensemble plots) between the reference MD (blue), and the predicted (orange).



772

773 **Fig. 4 | DynoM samples proteins with multiple conformations. a,** Conformation ensembles
774 generated by MD (blue) and DynoM (light red) on the two protein cases (PDB IDs: 2WG5F and
775 3UDCA) with local structural changes. **b,** Free energy surfaces (in kcal/mol) of ground truth MD
776 and DynoM in the space of the first two principal components, three representative
777 conformations were selected and labeled as 1-3. **c,** Representative conformations 1 to 3. Blue
778 and light blue represent the conformations of MD; red and light red represent the conformations
779 generated by DynoM. **d,** The frequency distribution plot shows the secondary structure (α-helix
780 and β-sheet) of each residue calculated from 1000 conformations of the reference MD (orange)
781 and the predicted (blue curve), respectively. **e,** Performance of DynoM in sampling diverse
782 conformational transitions. The 2D density plots illustrate the distribution of the generated
783 ensemble relative to two distinct reference structures (RMSD on x- and y-axes). Left to right:
784 apo-holo state transitions (PDB IDs: 3DRJ and 3DRF), rigid-body domain motions (PDB IDs:
785 4N3Z and 4N3Y), and folding-unfolding transitions (PDB IDs: 6MF4 and 6MF5).

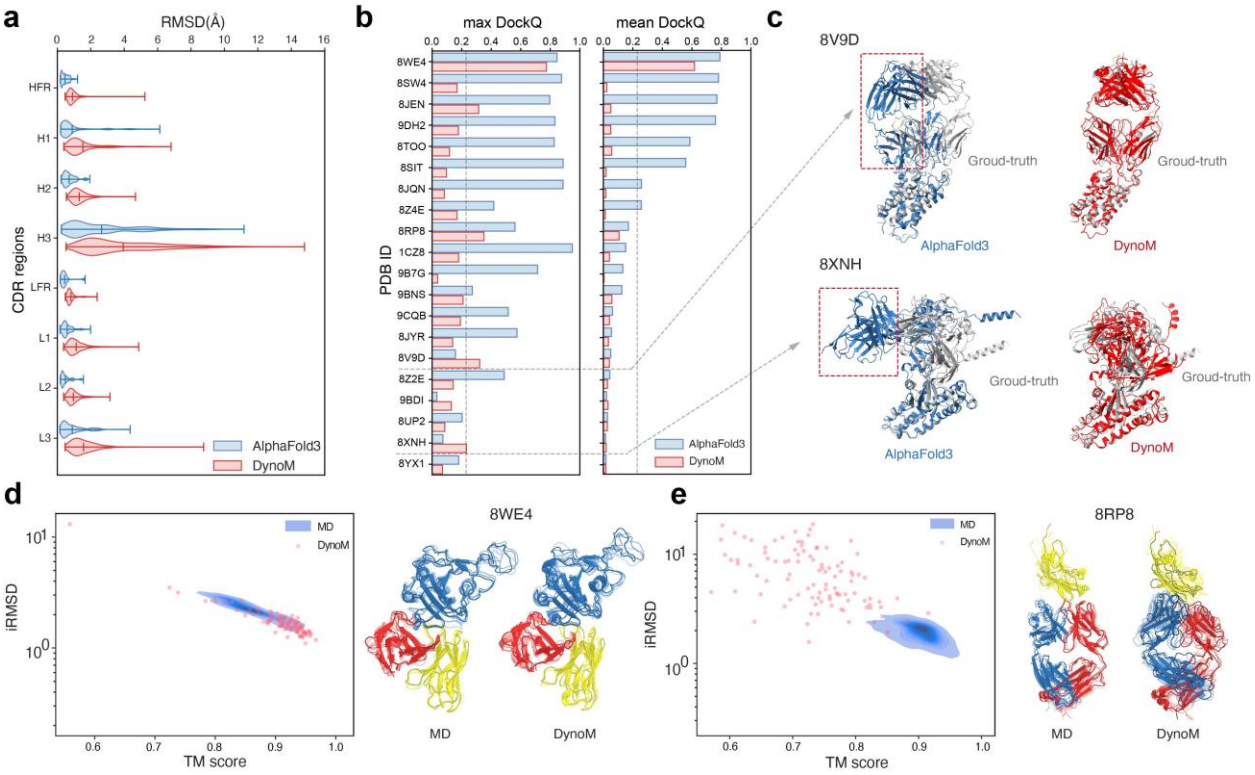


Fig. 5 | **Exploration of conformational ensemble prediction for protein complexes using DynoM. a,** RMSD distributions of predicted conformational ensembles for 20 antigen-antibody (Ag-Ab) complexes. For each complex, 100 conformations were generated by DynoM and AlphaFold3, yielding a total of 2,000 structures per method. The evaluation assesses the heavy and light chain framework regions alongside the complementarity-determining regions (H CDR 1-3 and L CDR 1-3). CDR RMSDs were calculated based on Chothia loop definitions following the structural superposition of corresponding framework regions. Center markers in the violin plots indicate the mean values. **b,** Mean and maximum DockQ scores for conformations predicted by the two methods across the 20 Ag-Ab complexes. The gray horizontal line represents DockQ=0.23. **c,** Representative protein structures predicted by AlphaFold3 and DynoM (PDB IDs: 8XNH, 8V9D). Blue: AlphaFold3 predictions; Red: DynoM predictions; Gray: experimentally resolved ground-truth. **d-e,** The TM score-iRMSD distributions of the predicted (red points) and the reference MD (blue contours). In the left scatter plot, each red dot

represents a DynoM predicted conformation, and the blue contour maps show the density of conformational sampling distribution obtained from MD simulation. The right-side shows conformations selected from the predicted and the reference MD, where the predicted structures were randomly selected from the red dots falling within the MD range. (PDB IDs: 8WE4, 8RP8).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarytable1.csv
- Supplementarytable2.csv
- Supplementarytable8.csv
- Supplementarytable14.csv
- Supplementarytable16.csv
- SupplementaryMaterials.docx