nature machine intelligence

Article

https://doi.org/10.1038/s42256-026-01198-9

Conditional diffusion with locality-aware modal alignment for generating diverse protein conformational ensembles

Received: 27 February 2025

Accepted: 29 January 2026

Published online: 25 February 2026

Check for updates

Baoli Wang[1,5], Chenglin Wang[2,5], Jingyang Chen[3,5], Danlin Liu[2,4], Changzhi Sun[4], Jie Zhang[3] ✉, Kai Zhang[2,4] ✉ & Honglin Li[1,4] ✉

Recent advances in artificial intelligence have enabled accurate prediction of a protein's stable structure solely based on its amino acid sequence. However, capturing the complete conformational landscape of a protein and its dynamic flexibility remains challenging. Here we developed modal-aligned conditional diffusion (Mac-Diff), a score-based diffusion model for generating the conformational ensembles for unseen proteins. Central to Mac-Diff is an attention module that enforces a delicate, locality-aware alignment between the conditional view (protein sequence) and the target view (residue pair geometry) to compute highly contextualized features for effective structural denoising and generation. Furthermore, Mac-Diff leverages semantically rich sequence embedding from protein language models such as ESM-2 in enforcing the protein sequence condition that captures evolutionary, structural and functional information. Mac-Diff showed promising results in generating realistic and diverse protein structures. It successfully recovered conformational distributions of fast-folding proteins, captured multiple meta-stable conformations that were observed only in long MD simulation trajectories and efficiently predicted alternative conformations for allosteric proteins. We believe that Mac-Diff offers a useful tool to improve understanding of protein dynamics and structural variability, with broad implications for structural biology, structure-based drug design and protein engineering.

Proteins are fundamental building blocks of life and play an integral role in cellular and biological processes. Many proteins possess inherent flexibility that enables them to function through the interconversion of different conformational states with varying energy levels. This dynamic nature has profoundly determined protein functional repertoire in various contexts, including ligand interactions, enzymatic reactions and molecular evolutions. As a result, accurately generating the conformational ensembles for given protein sequences is vital for elucidating the mechanisms underlying their functionalities with wide applications in biological and medical sciences[1–3].

In the past, experimental structure determination primarily focused on single—or at best a few—discrete, static protein structures because of the substantial cost involved [4]. To achieve a comprehensive delineation, molecular dynamics (MD) simulations are widely used to generate coherent trajectories of protein conformations.

[1]School of Pharmacy, East China University of Science and Technology, Shanghai, China. [2]School of Computer Science and Technology, East China Normal University, Shanghai, China. [3]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai, China. [4]Innovation Center for Artificial Intelligence and Drug Discovery, School of Pharmacy, East China Normal University, Shanghai, China. [5]These authors contributed equally: Baoli Wang, Chenglin Wang, Jingyang Chen. ✉e-mail: jzhang080@gmail.com; kzhang@cs.ecnu.edu.cn; hlli@hsc.ecnu.edu.cn

Due to the high computational complexity and the requirement to simulate long timescales, MD simulations could be very time-consuming and resource demanding[5,6]. The emergence of AlphaFold2 has dramatically advanced the state-of-the-art of protein structure prediction. By integrating structural and co-evolutionary information through Evoformer attention in an end-to-end learning architecture, AlphaFold2 enables the faithful prediction of an individual (arguably the most probable) protein structure[7]. A number of variants[8–13] based on AlphaFold2 have been further proposed to produce multiple conformations for a protein by expanding the output of AlphaFold2 or by exploiting the evolutionary information. Examples include modifying multiple sequence alignment (MSA) depth through subsampling[8,9] or residue replacement[10], performing MSA cluster-level structure prediction[11], using state-specific structural templates[12] and employing multiple structural outputs as initialization for enhanced sampling[13]. Despite its great potential, how to fully recover the conformational heterogeneity observed in many proteins under standard AlphaFold2 inference protocols remains to be further explored[14,15]. More recently, Gao et al. proposed AF2Complex[16], which allows predicting alternative conformations that emerge in the presence of protein–protein interactions. Incorporating such partners enables the discovery of conformational states that are inaccessible when proteins are modelled in isolation, thereby substantially improving the prediction of biologically relevant structural ensembles[17,18]. This interaction-driven conformational heterogeneity offers an interesting perspective for extending AlphaFold2's capabilities towards modelling diverse, functionally relevant conformational ensembles.

In recent years, non-deterministic generative models drew considerable attention towards systematically generating the conformational ensembles of proteins. Early works focused on generative adversarial networks or variational autoencoders[19–21]. Then, a considerable amount of interest was drawn on diffusion models[22] due to their promising results in generating realistic samples from the distribution they are trained on. Examples include non-conditional diffusion models such as FoldingDiff[23] and Str2Str[24], and a number of conditional generative models using sequence representation from structure prediction models (as conditions) and various types of equivariant networks (for denoising), such as Eigenfold[25], ConfDiff[26], DiG[27] and BioEmu[28] and those using advanced flow models and diffusion transformers such as AlphaFlow[29] and IdpSAM[30]. Furthermore, diffusion models were also applied successfully in protein design tasks[31,32].

Most existing conditional diffusion methods[24–28] have relied closely on structure prediction models such as AlphaFold2[7], ESMFold[33] or OmegaFold[34] in terms of the protein geometric representation (for example, residue frames[7] and C_α atom coordinates[25]) and the denoising network architectures used in these methods (for example, invariant point attention[7,31]). Meanwhile, the initial sequence embedding was also obtained from these structure prediction models as the condition for generative models. For example, EigenFold used the residue embeddings from OmegaFold, both DiG and BioEmu used those from AlphaFold2, and ConfDiff used the sequence embedding from ESMFold. Although encouraging results have been observed, further exploration is needed to determine whether sufficient structural heterogeneity can be extracted from the sequence representations of structure prediction models. This is because many structure prediction models, under their default settings, were designed to predict a single structure for a given sequence, and so the resultant representations might be possibly biased towards a dominant structure that structure models tend to predict[35].

Here we present modal-aligned conditional diffusion (Mac-Diff), a score-based conditional diffusion model to generate realistic and diverse protein conformational ensembles. Mac-Diff performs iterative denoising on protein backbone geometries (target view) by continually receiving guidance from the protein sequence (conditional view). Here, instead of using structural prediction models like AlphaFold2, Mac-Diff adopts protein language models (PLMs) such as ESM-2[33] to obtain the initial representation of protein sequences as conditions. ESM-2 was trained by unsupervised masked language modelling based on massive protein sequence datasets, allowing it to capture a wide spectrum of information ranging from evolutionary patterns, structural motifs and functional properties to broader biological knowledge at different scales. This semantically rich sequence representation has shown great potential as a scalable and alignment-free alternative to capturing the conformational diversity of proteins in our study. Notably, the PLM-derived residue embedding has also been used in Chai-1[36] in its single-sequence mode, achieving strong structure prediction performance particularly for protein–ligand complexes and multimers.

Central to Mac-Diff is an attention module to bridge the gap between the conditional modality (protein sequence) and the target modality (residue geometry) called locality-aware modal alignment attention (LAMA-attention). Compared with text-to-image tasks[37] requiring only loose, unstructured alignment between text tokens and image pixels, LAMA-attention enforces a physically more delicate alignment between sequence and structure. In particular, while a direct correspondence between a specific residue and its own three-dimensional (3D) coordinates is trivial, the critical alignment lies between a residue in 3D space and its spatially interacting neighbours traced back to the input sequence. Capturing these interactions is the key to injecting useful sequence information into the target space for structural denoising. By restricting the attention field of each residue to its most likely local interacting environment, the locality-aware alignment between the two modalities was able to compute highly contextualized features in the target space to recover realistic and diverse protein structures. See Supplementary Note 1 for a comparison of the different levels of modal alignment for text-to-image generation and protein conformational ensemble generation, as well as discussions on the limitations of conventional cross-attention in the latter task.

Mac-Diff showed promising results in generating realistic conformational ensembles for given protein sequences. Empirically, Mac-Diff effectively recovered the conformational distribution of 12 fast-folding proteins from the benchmark test set[24,38] it has never seen before, in terms of a number of important evaluation metrics such as the Jensen–Shannon (JS) divergence on C_α–C_α distance distribution, radius of gyration distribution, and C_α–C_α distance distribution on top-two time-lagged independent components (TICs) analysis (TICA) components. Notably, the conformations generated by Mac-Diff exhibited greater diversity compared with competing methods while preserving a high level of accuracy in terms of ensemble distributions.

Furthermore, Mac-Diff demonstrated promising ability to predict alternative conformations that are potentially biologically relevant, even for proteins not encountered during training. For example, it recovered important conformational substates of bovine pancreatic trypsin inhibitor (BPTI) that were observed in long MD simulations of 1 ms, and also predicted the closed state and the open state of adenylate kinase (AdK), an allosteric protein involved in energy metabolism. Finally, Mac-Diff achieved a sampling speed approximately 3,000 times faster than conventional MD simulations (that is, over three orders of magnitude). Overall, we believe that Mac-Diff has the potential to improve our understanding of protein folding dynamics and provide insights into the intricate relationship between protein sequence, structure and function. The capability of Mac-Diff in predicting conformational heterogeneity will also be useful in applications of structure-based drug design and protein engineering.

Results

Figure 1 illustrates the overall design of Mac-Diff. Figure 1a is the backbone geometric representation used in Mac-Diff, including pairwise C_β distance, dihedral angle, planar angle and a padding channel, which

https://doi.org/10.1038/s42256-026-01198-9

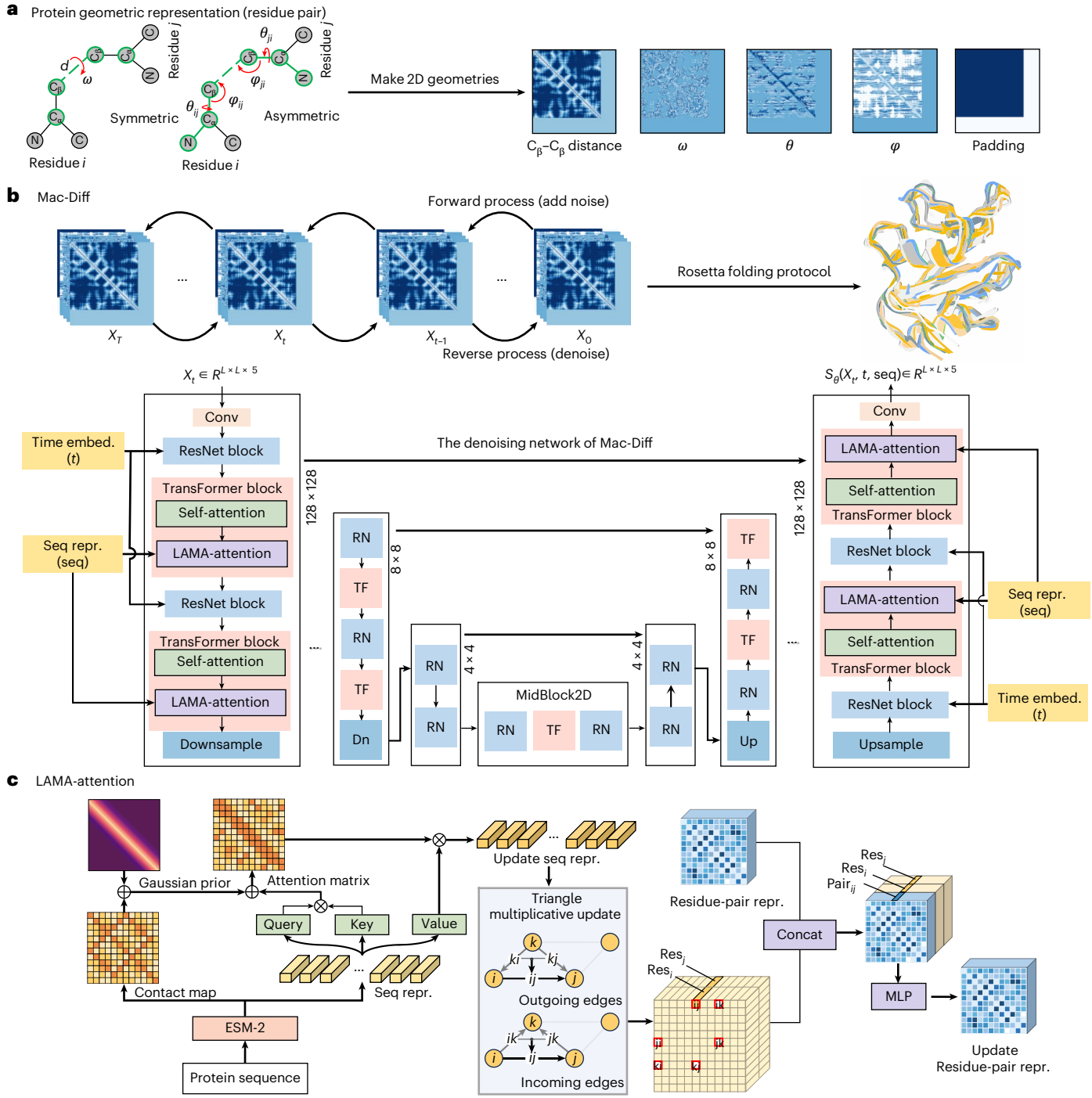


Fig. 1 | Overview of Mac-Diff architecture. a, Protein backbone geometric representation with L residues as an $L \times L \times 5$ tensor with pairwise C_β distance, dihedral angle ω along two C_β atoms, dihedral angles θ, and bond angles ϕ (direction of C_β atom of one residue in a reference frame centred on the other residue), and a padding channel indicating sequence length. **b**, Mac-Diff workflow. The forward diffusion process iteratively injects noise to geometric tensor, and the backward process performs iterative denoising. The denoising network is a U-Net structure with five downsampling/upsampling stages, each stage with a ResNet block and a TransFormer block (self-attention and LAMA-attention). **c**, LAMA-attention, allowing each residue to attend only to neighbouring residues with high contact probability, updating residue-pair representations with highly relevant, contextualized sequence features for denoising. repr., denotes representation; seq, protein sequence; Conv, convolutional layer; TF, TransFormer block; RN, ResNet block; Dn, downsampling stage.

are invariant to 3D rotation and translation. Figure 1b is the model overview. It is a score-based conditional generative model capable of recovering the conformational distribution of a protein by generating backbone geometric structures and converting them to atom-level coordinates through the Rosetta folding protocol. The forward diffusion process iteratively injects noise to the geometric tensor, and the backward process achieves iterative denoising. The denoising network is a U-Net structure with five downsampling/upsampling stages. Each stage contains a ResNet block to integrate time-step embeddings with residue-pair representations, and a TransFormer block to update residue-pair representations via self-attention and LAMA-attention. Figure 1c is the LAMA-attention. It enforces a well-controlled spatial

https://doi.org/10.1038/s42256-026-01198-9

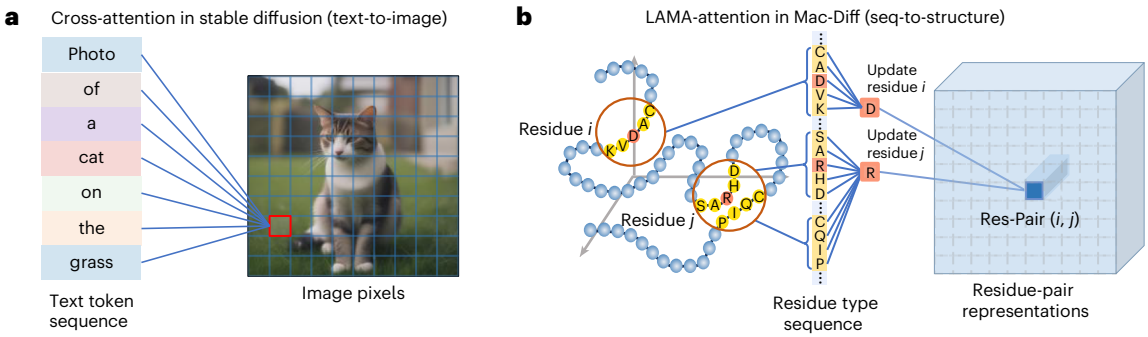


Fig. 2 | Schematic comparison of cross-attention (in text-to-image generation) and LAMA-attention (in protein conformational ensemble generation). a, In traditional cross-attention, each pixel in the generated image is potentially related to all tokens in the input text without prior algorithmic control on the attention field. **b**, In the LAMA-attention, each residue-pair (Res-Pair) representation *ij* is related to only those residues that are likely to interact with residue *i* and *j* biologically—a stronger, locality-aware spatial alignment between sequence and structure.

alignment between the sequence view and the structure view by forcing each residue to attend to only those neighbouring residues with high contact probability, to update residue-pair representations with highly relevant and contextualized sequence features for denoising. More detailed descriptions can be found in the Methods.

Figure 2 illustrates the difference between Mac-Diff and stable diffusion—a popular diffusion framework for text-to-image generation[37]—in their attention modules. In stable diffusion, the attention between pixels (query) and words (keys) was dense and global, indicating that the alignment between pixels and words were unstructured, purely data-driven and without prior algorithmic controls (Fig. 2a). In comparison, LAMA-attention allowed focusing only on the interacting neighbours (their amino acid types) of a residue when updating its representation (Fig. 2b). This effectively narrows the attention field from the whole sequence to only a small fraction of useful residues in the conditional view, marking a key distinction from the global cross-attention module in stable diffusion.

We performed comprehensive evaluations of Mac-Diff in comparison with state-of-the-art approaches for generating realistic and diverse protein conformational ensembles, making use of both carefully curated training datasets and widely used public benchmark datasets. Two complementary categories of tasks were designed to assess model performance. First, we evaluated the ability of Mac-Diff to recover the underlying distribution of conformational ensembles and to identify key conformational substates, using MD trajectories of fast-folding proteins and the BPTI benchmark as reference standards[24,27,28,38,39]. Second, we assessed its capability to predict alternative conformations of proteins, leveraging experimentally validated 3D structures from a subset of the Cfold dataset[40] and a well-characterized model protein AdK.

The Mac-Diff model was trained using a pretraining–fine-tuning strategy. It was first pretrained on 619,045 experimentally determined protein sequence–structure pairs from the Protein Data Bank (PDB), and subsequently fine-tuned on 1,674 MD trajectories collected from two public sources: 371 trajectories from GPCRmd[41] and 1,303 trajectories from the Atlas of Protein Molecular Dynamics (ATLAS)[42] (see dataset details in 'Experimental settings' section in the Methods). Importantly, in both pretraining and fine-tuning stages, we excluded all proteins whose sequence identity with any test protein exceeded 40%, ensuring a rigorous evaluation of the model's generalization ability. Detailed settings and statistics of the training data are provided in the 'Training data' section in the Methods. For all other competing models, results were obtained without redundancy control, by directly using their released models trained on the original datasets (as retraining on de-redundant data would be prohibitively costly). These stringent comparative conditions ensure a rigorous and fair evaluation of Mac-Diff's performance.

Performance evaluation on fast-folding proteins benchmark

We first evaluated Mac-Diff in generating the conformational ensembles of fast-folding proteins. To evaluate how well the generative model reproduces the conformational distribution of the original MD data, we used the complete D. E. Shaw Research (DESRES) dataset, which includes 12 structurally diverse fast-folding proteins, along with the BPTI protein. These proteins have been widely used as benchmark for evaluating the quality of equilibrium distributions generated by computational models[24,26,38]. The reference conformation distribution for each protein is represented by 1,000 conformations sampled with fixed stride from their MD simulation trajectories, ranging from 100 μs to 1 ms to cover multiple folding or unfolding events[43]. See the detailed protein trajectory information in the 'Test data' section in the Methods and Supplementary Table 1.

We have generated 1,000 conformations using Mac-Diff and each of the competing methods to recover the protein conformational distribution, and examined the quality by assessing their deviation from the reference MD distribution. To provide a quantitative analysis, we computed the JS divergence on the following distributions, including: (1) pairwise C_α atom distance distribution (JS-PwD), which reflects the global spatial structure of the protein; (2) radius of gyration distribution (JS-Rg), which is based on the distance between each C_α atom to the centre of mass of the protein, reflecting the compactness of the protein; and (3) pairwise C_α atom distance matrices projected onto the top-two TICs (JS-TIC), which is commonly used to analyse slow dynamics of proteins MD trajectories[44–46]. The metrics used to quantify the generated conformational distributions are evaluated independently for each protein and subsequently averaged across the 12 fast-folding proteins. Detailed definitions of the three metrics can be found in the 'Evaluation metrics' section in the Methods.

As shown in Fig. 3a, Mac-Diff achieved competitive performance compared with existing diffusion- and flow-based models across all equilibrium metrics on fast-folding proteins. Averaged over 12 fast-folding proteins, Mac-Diff reduced errors in JS-PwD, JS-Rg and JS-TIC by approximately 18%, 22% and 5%, respectively, relative to the best competing method. In statistical testing (two-sided Wilcoxon test), Mac-Diff outperformed five of the competing methods under JS-PwD and JS-TIC ($P < 0.05$ with medium-to-large effect sizes) and outperformed all other competing methods under JS-Rg ($P < 0.05$ with medium-to-large effect sizes), as detailed in Supplementary Table 2. In Supplementary Tables 3–5, we provide more detailed comparative results on each of the 12 fast-folding proteins under these three metrics. These results showed the potential of Mac-Diff in recovering protein conformational distributions.

Next, we evaluated the trade-off between diversity and fidelity for different generative models. Diversity was quantified as the average pairwise dissimilarity (1 − lDDT − C_α) among the generated

https://doi.org/10.1038/s42256-026-01198-9

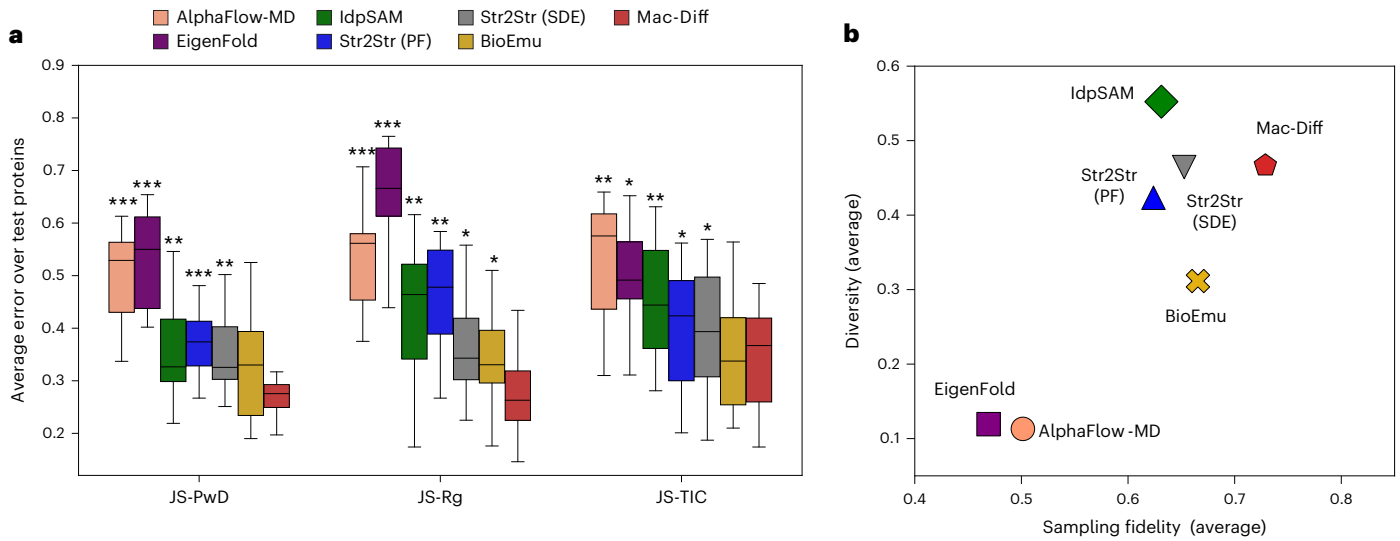


Fig. 3 | Performance of all competing methods in recovering the conformational distributions of the fast-folding proteins. a, The JS divergence between the generated conformational distribution and the reference MD distribution for 7 competing methods averaged over 12 fast-folding proteins[43], when considering the pairwise C_α atom distances (JS-PwD), the radius of gyration (JS-Rg) and the pairwise C_α atom distances matrices projected onto the top-two TICs (JS-TIC). Data were derived from a sample size of $n = 12$ independent test proteins, where each data point represents the mean of 3 independent runs. The box plots show the medians as the centre line, the 25th and 75th percentiles as lower and upper quartiles, and 1.5 times the interquartile range as whiskers. Asterisks above boxes indicate statistically significant differences compared with Mac-Diff, as determined by a two-sided Wilcoxon signed-rank test (*$P < 0.05$, **$P < 0.01$, ***$P < 0.0001$). **b**, Average diversity–fidelity trade-off for each model. Each point denotes the mean diversity and fidelity scores of a competing model across the 12 fast-folding proteins. Higher values indicate superior performance on both metrics.

conformations, where lDDT-C_α is the local distance difference test (lDDT) score, which quantifies the local structural similarity between two structures by evaluating the preservation of distances between C_α-atom distances within local neighbourhoods[47]. Fidelity was measured as $1 - JS_{PwD}$, where JS_{PwD} denotes the JS divergence between the generated conformational distribution and the reference MD distribution as defined in (equation (1)). For both metrics, higher values indicate better performance. However, diversity should not be interpreted in isolation and is meaningful only when the generated conformations also achieve sufficient structural accuracy. In Fig. 3b, we presented the diversity–fidelity scores of 7 models, each averaged across 12 fast-folding proteins in the test set. Points closer to the upper-right corner represent a balance between fidelity and diversity. Mac-Diff showed the highest average score across the two metrics, suggesting that it maintained both fidelity and diversity effectively. Detailed results for each of the 12 proteins are provided in Supplementary Fig. 1a.

Beyond the JS divergence, we also evaluated the deviation between the generated conformational ensembles and the MD data using the Earth Mover's Distance (EMD), considering both global and local structural differences in the C_α-atom structural space. Specifically, we constructed a cross-distance matrix between all pairs of structures across the generated samples and the MD ensemble, using 1 − TM-score[48] as a global structural difference metric and 1 − lDDT[47,49] as a local structural difference metric. The EMD was then computed from this cross-distance matrix via optimal transport (see detailed definitions in 'Evaluation metrics' section in the Methods). Unlike divergence-based measures that primarily quantify distributional overlap, EMD also accounts for the structural dis-similarity between individual conformations, making it particularly well suited for comparing ensembles of 3D protein structures. As shown in Supplementary Fig. 1b, Mac-Diff consistently achieved lower EMD values than other competing methods, indicating that its generated conformations more closely resemble the reference MD ensembles at both global and local structural levels. These results are consistent with the JS-divergence analyses, confirming the robustness of Mac-Diff under multiple complementary evaluation protocols.

To examine whether the generative models can capture diverse and biologically relevant conformations, we identified three representative conformational states—folded, loosely packed and unfolded—for each of the 12 fast-folding proteins and report the success rate of recovering each state in Supplementary Fig. 2 and Supplementary Table 6. The three metastable states for each protein were obtained following the clustering analysis procedure described by Zheng et al.[27]. Specifically, each MD trajectory was first projected onto a two-dimensional (2D) TICA space and then clustered into three dominant metastable states, representing distinct conformational basins along the folding landscape. For each cluster, up to 1,000 frames closest to the 2D cluster centroid were analysed using MDTraj[50] to derive the centroid structural pattern. The structure nearest to this centroid was selected as the representative conformation of that state. Supplementary Fig. 2 plotted the distribution of the generated conformations by projecting them onto the 2D TICA landscape, in which the folded, loosely packed and unfolded states were labelled as 1, 2 and 3, respectively. Supplementary Table 6 summarizes the recovery success of seven competing methods. Overall, Mac-Diff ranked highest and successfully captured 11 of 12 folded states, 10 of 12 loosely packed states and 2 of 12 unfolded states. It is worth noting that the unfolded states remain difficult to recover owing to their high structural heterogeneity.

We also computed the RMSF profiles to further assess the residue-level flexibility from different generative models. As shown in Supplementary Table 7, Mac-Diff achieved a median RMSF correlation of $r = 0.89$ with the reference MD data, ranking second among all seven competing methods and only slightly behind BioEmu ($r = 0.90$). These correlation values indicated that Mac-Diff could faithfully reproduce residue-level flexibility, capturing both major peaks and overall variations observed in the MD simulations for modelling fine-grained conformational dynamics.

We further used TICA to better visualize the conformational distributions generated from different models in a low-dimensional subspace. TICA is a dimension reduction technique for complex dynamic systems commonly applied in the study of protein folding

and unfolding processes[44]. Following the practice of Lu et al.[24], we have computed the top two TICA components from the pairwise C_α atom distance matrix that reflects the slow and dominant transitions among different protein conformational ensembles. Figure 4a visualizes the conformations produced by MD simulation and all competing models for α3D, Protein B and Homeodomain, in which each conformation generated was projected (as a dot) to the top two TICA components [44,45]. The JS-TIC error metric for each generative model is also shown in the subfigures. We observed that the conformations generated by Mac-Diff could better approach the reference MD data, both visually and in terms of the JS-TIC metric. The JS-TIC values from Mac-Diff were relatively 20%, 33% and 23% lower than those of the best competing method for the proteins α3D, Protein B and Homeodomain, respectively. The JS-TIC error metrics for the remaining nine fast-folding proteins, including Villin, BBA, Protein G, NTL9, WW domain, Lambda, Trp-cage, BBL and Chignolin, are presented in Supplementary Table 5.

In Supplementary Fig. 2, we adopted a complementary evaluation to quantify the distributional overlap in the 2D TICA space, using the recall metric defined by Zheng et al.[27]. Specifically, the 2D TICA space was discretized into a 50 × 50 grid. Each grid cell was labelled as ground-truth positive if it contained at least one structure from the reference simulation, and as sampled (or not sampled) if it contained at least one (or no) structure generated by the model. The recall was then computed as $Recall = \frac{TP}{TP+FN}$, where TP and FN denote true positives and false negatives. Supplementary Fig. 2 reported the recall values of all the seven competing methods by comparing their sampled distribution with the reference MD distribution on the 2D TICA space. Overall, the average recall of Mac-Diff and Str2Str (SDE) across the 12 test proteins was 0.403 and 0.401, respectively—the two highest among all 7 competing methods.

Figure 4b shows the probability map of pairwise residue contact relationships for two proteins—WW domain (featuring a three-stranded antiparallel β-sheet in the folded structure) and Homeodomain (featuring three α-helices in the folded structure)—as recovered by all competing models. Here, the *ij*th entry in the probability map represents the probability of contact between the *i*th and *j*th residues across all generated conformations, using a threshold of 10 Å (ref. 38). To better visualize the quality of the probability map recovered by each method, we plotted the difference between true/recovered probability map. We observed that Mac-Diff learned both local contact patterns and global structural patterns along the sequence[51]. AlphaFlow-MD and EigenFold tended to predict stable native structures as opposed to those contact patterns arising from unfolded but physically feasible structures. IdpSAM generated contact patterns that more closely resemble those of intrinsically disordered proteins, such as the β-sheet-related contact patterns observed in the protein WW domain (Fig. 4b, magenta box). This may stem from the fact that IdpSAM was primarily trained on intrinsically disordered proteins. Str2Str tended to emphasize certain weak contact patterns between the second β-sheet and the C-terminus residues (Fig. 4b, magenta box). See more visualization results in Supplementary Fig. 3. In Supplementary Table 8, we report the root mean square error (RMSE) of the residue contact probability map for all the competing methods, in which Mac-Diff had the lowest average approximation error across the 12 fast-folding proteins. This illustrates the effectiveness of Mac-Diff in faithfully capturing key contact patterns during protein folding.

The Homeodomain protein belongs to a class of evolutionarily conserved proteins with three α-helices. Commonly observed 3D structures are compact, with low JS-Rg values (the average distance between Cα atoms and the protein centroid, effectively representing the protein's radius), in which the first and last helices form intrachain contacts through hydrogen bonding between side-chain atoms. We carefully examined the contact map recovered by Mac-Diff, in particular its subblock corresponding to the residues from the first and the last helices. As can be seen in Fig. 4b and Supplementary Fig. 3, Mac-Diff accurately captures contact probabilities, producing a spatially compact 3D structure that is consistent with experimental observations.

Conformational substates prediction for the BPTI

To further examine the capacity of Mac-Diff in capturing protein conformational changes, we used BPTI, a globular protein of 58 residues, for a case study. In the literature[52], Shaw et al. found that BPTI had five structurally distinct representation clusters in a long MD simulation of 1 ms. Supplementary Table 9 reports the JS-divergence metrics for seven competing methods on the BPTI, in which Mac-Diff ranked second to BioEmu on JS-PwD and JS-Rg, and was comparable to Str2Str on JS-TIC. In Fig. 5, we further visualized the representative conformation for each of these five clusters, as well as the conformations generated from all the competing methods, by projecting their pairwise C_α atom distances into the top two TICs (TICA components) of the reference MD trajectories[24], corresponding to the slowly varying MD of greatest interest[44–46]. We also plotted contours of the 2D density of MD-simulated conformations using a kernel density estimator, with bandwidth determined by the default Scott's rule.

In Fig. 5, we observed that, for EigenFold and AlphaFlow-MD, most of the generated conformations were closer to cluster 1. In fact, capturing protein conformations across long timescales poses a major challenge. For Str2Str (SDE) and Str2Str (PF), the generated conformations well covered four conformational substates, except for substate 3, which is separated from the others by relatively steep energy barriers on the energy landscape. The conformations generated by IdpSAM were relatively dispersed across the five clusters, indicating that it tended to explore the energy landscape more uniformly. This may be attributed to the fact that IdpSAM was primarily trained on intrinsically disordered proteins that exhibit greater structural flexibility and occupy a wider range of conformational states[30]. Mac-Diff generated samples that were relative closer to all five substates. Using the RMSD between each cluster's representative conformation and the best-matching generated structure (after standard translation- and rotation-based calibration[24]), Mac-Diff achieved the lowest RMSDs for clusters 3, 4 and 5 (2.18 Å, 1.26 Å and 0.84 Å, respectively) and the second-lowest RMSDs for clusters 1 and 2 (0.81 Å and 1.55 Å). Across all eight competing methods, Mac-Diff attained the lowest average RMSD of 1.33 Å. See Supplementary Table 10 for the RMSDs of all the competing methods relative to the five conformation clusters.

We also applied the Bonferroni correction to control the family-wise error rate and ensure the robustness of the comparison. For each method, the lowest RMSD values across the five reference clusters were considered, and the measurement was repeated three times to account for variability. We then performed Mann–Whitney *U* tests comparing each method against Mac-Diff, followed by Bonferroni correction. As shown in Supplementary Fig. 4, Mac-Diff ranks among the top tier, comparable to AlphaFlow-MD, EigenFold and BioEmu, with no statistically significant differences ($P < 0.05$).

To verify the validity of the generated conformational ensembles of the BPTI by Mac-Diff, we further plotted the bond length distribution and backbone dihedral angle distribution in Supplementary Fig. 5a,b. We observed that Mac-Diff achieved good agreement with the MD reference data: the mean bond length of 1,000 conformational structures (3.81 Å) is very close to both the mean bond length from the reference MD (3.85 Å) and the ideal bond length (previous study on protein structure quality assessment in PDB demonstrates that the distance between consecutive C_α atoms is distributed normally with a mean of 3.8 Å (ref. 53)). Moreover, the distribution of the distances between non-bonded C_α atoms that are at least four residues apart in the protein sequence are also shown in Supplementary Fig. 5a (right). The lower bound of the non-bonded atom distances by Mac-Diff is 3.49 Å, which is very close to that derived from the reference MD (3.48 Å). This suggested that the pairwise geometric representations generated by Mac-Diff were sufficiently realistic to guide Rosetta-based reconstruction, resulting in atom-level backbone coordinates that successfully avoid self-crossings.

https://doi.org/10.1038/s42256-026-01198-9

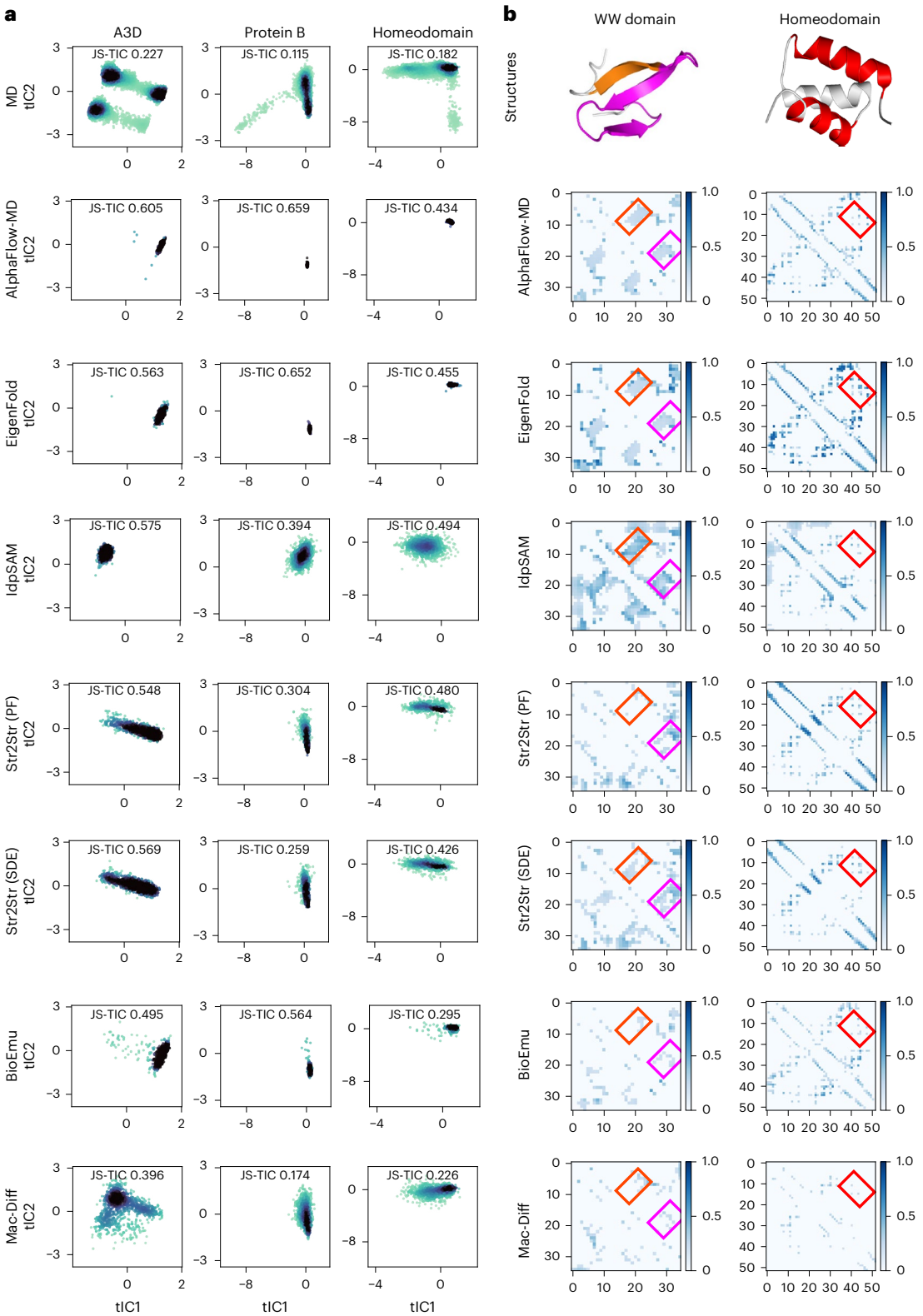


Fig. 4 | Performance of all competing methods in recovering protein conformational distributions and residue contact probabilities. **a**, Conformations generated by MD simulations and seven generative models, projected onto the top two TICA components for α3D, Protein B and Homeodomain. Each dot represents a conformation and is colour coded by its density (the darker the point, the higher the density). The JS-TIC lower bound from intratrajectory MD is shown in the MD ensemble as a reference. **b**, Absolute differences in residue contact probability maps for the WW domain and Homeodomain, comparing seven competing methods with the MD-derived reference. Sparser or lighter pixels indicate smaller differences. Contact patterns of the β-sheets in the WW domain (boxed in orange and magenta) and the α-helix–α-helix interaction in the Homeodomain (boxed in red) are shown in the upper triangles. The folding structures at the top are marked with the corresponding colours.

https://doi.org/10.1038/s42256-026-01198-9

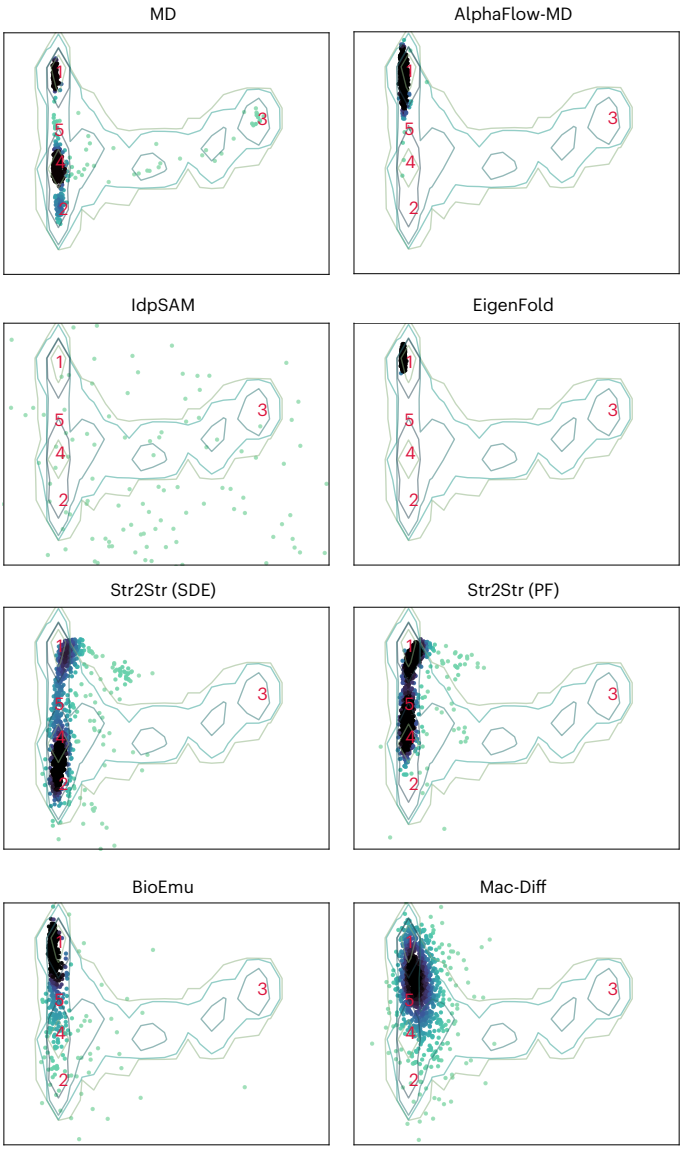


Fig. 5 | Generated conformations (dots) from seven competing methods, along with the reference MD distribution for BPTI, visualized in the top two TICA components most relevant to protein folding dynamics. The density contours of the MD reference distribution are plotted in each result. Five key conformation clusters are marked by numbers 1–5 in red colour.

In Supplementary Fig. 5b, we evaluated the torsion angles ϕ and ψ generated by Mac-Diff in terms of the Ramachandran map[54], which describes the key rotational freedom around backbone peptide bonds. Results showed that most of the generated angle pairs fell into the region of right-handed α-helix conformations, β-sheet conformations and left-handed α-helix conformations, closely matching MD trajectories. Taken together, these results suggested that Mac-Diff generated feasible and informative pairwise geometric constraints, enabling downstream folding modules to recover valid conformations that capture key substates while preserving a favourable balance between structural diversity and fidelity.

Capturing alternative conformations of allosteric proteins

In this section, we evaluate the ability of Mac-Diff to capture multiple functionally relevant conformational states using AdK from *Escherichia coli* and proteins from the Cfold40 test set. This analysis emphasizes functional features that may be obscured in a single static structure. The Cfold40 set comprises 40 proteins from the Cfold dataset[40], each with fewer than 256 residues and minimal missing regions. Every protein is annotated with two alternative conformations, with a TM-score between them below 0.8. Additional data details are provided in the 'Test data' section in the Methods.

We included eight state-of-the-art generative models as competing methods in our comparisons, including (1) Naive AlphaFold2[7]; (2) AlphaFold2 dropout variant (AlphaFold2_dropout)[10]; (3) MSA subsampling[8], an AlphaFold2-based method generating protein alternative conformations by performing MSA subsampling; (4) AF-Cluster[11], an AlphaFold2-based method generating multiple conformations by MSA clustering; (5) AlphaFlow-MD[29], which fine-tuned the weights of AlphaFold2 on OpenProteinSet and MD datasets under flow matching framework; (6) DiG[27], a diffusion model trained on both PDB and more than one thousand MD trajectories with initial residue embeddings from AlphaFold2's Evoformer; (7) Boltz-2[55], a structural biology foundation model demonstrating state-of-the-art performance in structure and affinity prediction; and (8) BioEmu[28], a state-of-the-art diffusion model based on the invariant point attention TransFormer as the structure prediction module, trained on tens of millions of protein conformations. In Supplementary Table 11, we summarize how these competing models selected their training data and how they derived the residue embedding vectors.

We first compared the performance of Mac-Diff and the eight competing methods on the allosteric protein AdK and benchmark Cfold40. All methods were allowed to generate 100 conformations per protein, and all MSAs used in this evaluation were fetched from the ColabFold server[56]. Detailed parameter settings of the nine competing methods are given in Supplementary Note 7. In Fig. 6a, we report the sampling results on the allosteric protein AdK. The native structure of AdK comprises three key domains: a relatively rigid CORE domain and two highly flexible domains, NMP and LID, which undergo conformational transitions between open and closed states during ATP-to-ADP catalysis[57]. We observed that Mac-Diff exhibited high conformational diversity, effectively covering both the open and closed states, with best TM-scores of 0.85 and 0.97, respectively. Considering the average of these two TM-scores, Mac-Diff jointly ranked highest with MSA subsampling, whose two TM-scores were 0.98 and 0.84. In comparison, AlphaFold2, AlphaFold2-dropout and DiG produced samples largely centred on the closed conformation. AlphaFlow-MD and Boltz-2 produced conformations that were closer to the open state. BioEmu captured the closed state well, but its recovery of the open state was less satisfactory. Taken together, these results indicate that Mac-Diff can effectively explore the conformational heterogeneity of AdK. Notably, increasing the sampling size generally improves the coverage of representative conformational states across generative models. For example, when BioEmu and DiG sample tens of thousands of conformations, as in their original settings, they more accurately recover both conformational states of AdK. In our benchmark, all competing methods were evaluated with a uniform sampling size of 100 conformations per protein to ensure fairness and computational comparability across models with different sampling efficiencies.

In Fig. 6b,f, we report sampling results on five representative proteins from Cfold40. These proteins include the tandem bromodomains of human TATA-binding protein-associated factor-1 (TAF1)[58,59], the b′-a′ domains of protein disulfide isomerase (PDI)[60], the substrate-binding domains (SBDs) of the transmembrane protein OpuA[61], and the tandem VHS and FYVE domains of hepatocyte growth factor-regulated tyrosine kinase substrate (HGS-Hrs), as well as a surface layer protein. As shown, Mac-Diff successfully generated diverse conformational states that recover both of the two experimentally determined reference structures for each protein. Quantitatively, Mac-Diff achieved TM-scores >0.8 for all five representative proteins, with RMSD <3 Å for four proteins, whereas other competing methods recovered at most four proteins under the TM-score criterion and three proteins under the RMSD criterion.

Figure 6g shows the distributions of TM-scores and RMSDs across all 40 proteins in Cfold40 for all competing methods. On average,

Mac-Diff achieved a TM-score of 0.85, close to the best competing method (0.87 for BioEmu). In terms of RMSD, Mac-Diff attained an average of 3 Å, comparable to BioEmu (2.9 Å). The detailed averages for each method are reported in Supplementary Table 12. Considering the number of successfully recovered alternative conformations across all 40 proteins, Mac-Diff recovered 56 out of 80 conformations with TM-scores >0.8, ranking third—slightly behind BioEmu (60/80) and MSA subsampling (57/80). For RMSD, Mac-Diff recovered 55 out of 80 conformations with RMSD <3 Å, jointly ranking first with BioEmu. Detailed results for each method are provided in Supplementary Fig. 6 and Supplementary Table 13.

After evaluating the distributions of TM-score and RMSD for the best predicted conformations, we further applied the Wilcoxon signed-rank test to assess the statistical significance of the differences between Mac-Diff and eight competing methods. Using both metrics, the comparison was conducted on the top three generated conformations for each reference structure (40 proteins × 2 conformations × 3 predictions). Results are summarized in Supplementary Table 14. For TM-score, Mac-Diff showed substantial improvements over MSA subsampling and DiG, while differences with AlphaFold2, AlphaFold2_dropout, AF-Cluster, AlphaFlow-MD, Boltz-2 and BioEmu are not statistically significant. For RMSD, Mac-Diff achieved notably lower deviations than AlphaFold2, AF-Cluster, AlphaFlow-MD, DiG and Boltz-2.

In Supplementary Table 15, we compute the minimum of the two TM-scores between each protein's two representative conformations and their best-matching structures generated by the models, averaged across all 40 proteins in Cfold40. This provides a worst-case assessment of ensemble coverage. Mac-Diff achieved an average minimum TM-score of 0.79 (jointly ranking second with MSA subsampling) and an average maximum RMSD of 4.2 Å (also ranking second), demonstrating strong coverage of the reference ensembles. While BioEmu demonstrated slightly superior performance compared with Mac-Diff, it was trained on tens of millions of protein structures—approximately ten times the size of our dataset—without any redundancy removal against the test proteins. Conversely, note that Mac-Diff achieved robust ensemble coverage without relying on representations from structure prediction models (for example, AlphaFold2), highlighting its potential as an alternative approach for sampling diverse protein conformations.

In Supplementary Table 16, we present a protein-wise structural recovery analysis, categorizing each of the 40 proteins in Cfold40 into 3 groups based on whether both, only one or neither of the experimentally determined conformations were successfully recovered by the generative models. Two complementary criteria were used: TM-score ≥0.8 or RMSD ≤3 Å. In terms of the success rate (recovery of both experimental conformations per protein), Mac-Diff achieved 50% under the TM-score criterion and 42.5% under the RMSD criterion—ranking second highest among nine competing methods and only slightly behind BioEmu. Regarding the failure rate (recovery of neither conformation), Mac-Diff achieved 10% under TM-score ≥0.8—ranking joint third with BioEmu[28]—and 5% under RMSD ≤3 Å, representing the lowest failure rate among all methods. These results demonstrate the robustness of Mac-Diff in capturing protein conformational diversity across a structurally diverse benchmark set.

To further assess whether the generated conformations adequately captured structural diversity and protein-level flexibility, we computed two complementary metrics. First, the mean absolute error (MAE) between the true conformational diversity ($TM_{conf1/conf2}$) and that of the generated conformations (TM_{var}) was used to quantify how accurately the model reproduces conformational variations within a protein. Second, protein-level flexibility was assessed by computing the MAE between residue fluctuations of the true and generated structures, whose definitions can be found in the 'Evaluation metrics' section in the Methods. As reported in Supplementary Table 17, Mac-Diff ranked fourth in reproducing conformation diversity, and ranked second in modelling protein-level residue flexibility among all the nine competing methods. Third, the protein-level correlation of residue flexibility was also calculated to evaluate whether the patterns of residue fluctuations are preserved, with Mac-Diff achieving a correlation of 0.6—ranking second among nine competing methods (the highest correlation was 0.62). Together, the MAE- and correlation-based results provided comprehensive evidence of Mac-Diff's competitiveness in capturing both conformational diversity and residue flexibility.

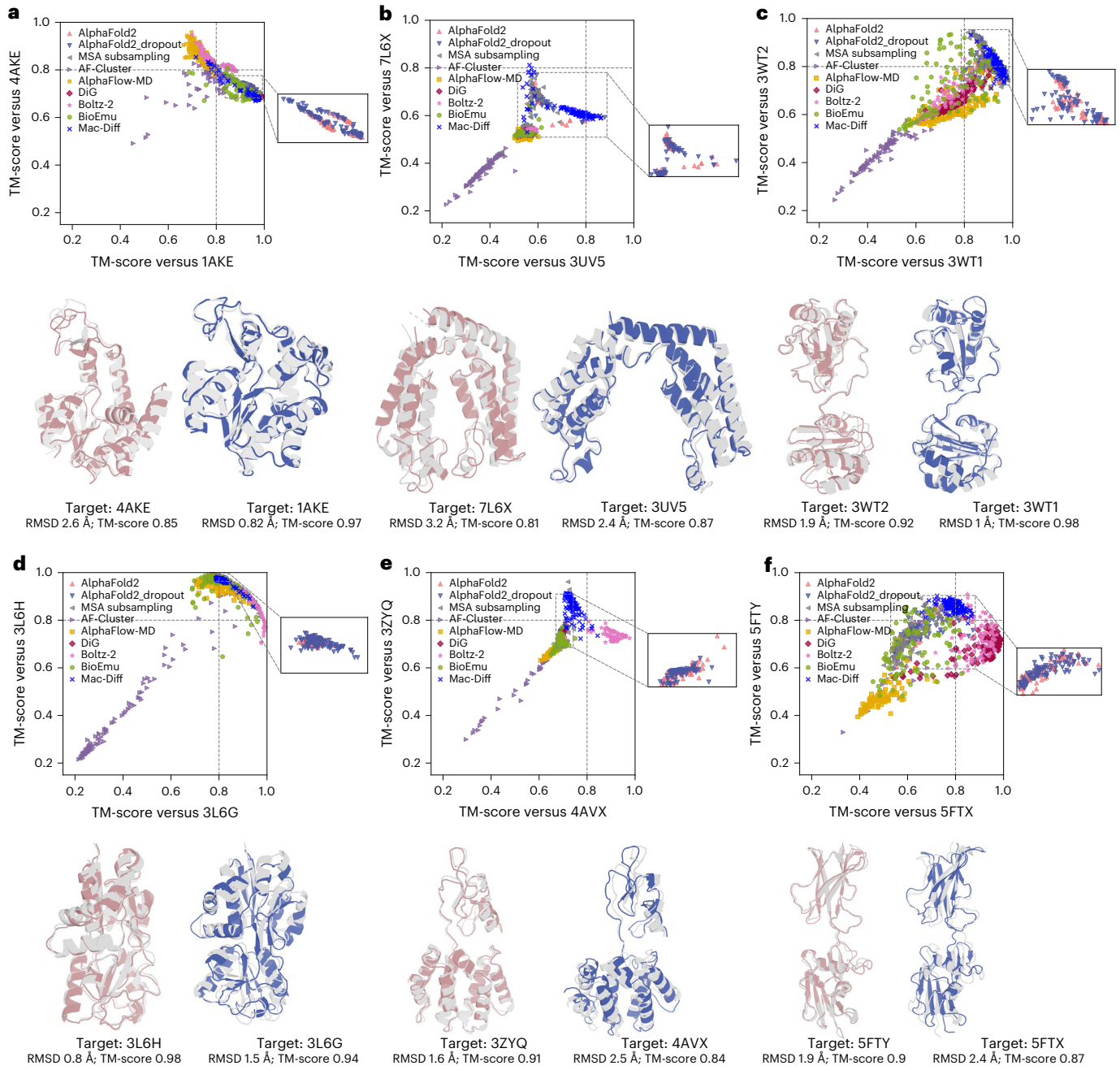
Despite its overall strong performance, Mac-Diff exhibited slightly lower accuracy than BioEmu on the Cfold40 test set. To explore the underlying causes, we investigated two potential factors. First, we examined the influence of the ESM-2 sequence embedding by computing the pseudo-log-likelihood derived from this embedding[62] and correlating it with the TM-scores across proteins in Cfold40 (for each protein, the average of the best TM-scores for the two reference conformations was used). Supplementary Fig. 7 showed the scatterplot of pseudo-log-likelihood versus TM-score, revealing a Pearson correlation score of $r = 0.45$ ($P = 0.004$). This statistically significant correlation indicates that sequences with higher ESM-2 likelihoods tend to produce more accurate conformational ensembles, suggesting that ESM-2 embeddings partially capture the sequence–structure relationship. However, the moderate level of correlation also implies that sequence embeddings alone may not fully account for the complex structural and functional constraints, and so incorporating structure-aware or MSA-informed priors, such as the Evoformer embedding used in BioEmu, may further improve generative performance. Second, it should be noted that BioEmu was trained on a substantially larger and more diverse dataset—approximately ten times the size of ours. Moreover, while our model was trained and fine-tuned strictly on datasets de-redundant with respect to the test proteins with sequence identity threshold below 0.4, BioEmu was evaluated directly using its publicly released model without de-redundancy processing against the test data, leading to higher training data familiarity. Notably, 8 out of the 40 test proteins in Cfold40 have a sequence identity exceeding 90% with BioEmu's training data. These two factors jointly contributed to BioEmu's marginally higher accuracy on the Cfold40 benchmark.

Mac-Diff sampling speed

In this section, we compare the sampling speed of Mac-Diff, implemented under the variance-preserving SDE (VPSDE) framework with the denoising diffusion implicit model (DDIM) sampler [63], against conventional MD simulations on a single NVIDIA A100 graphics processing unit (GPU). As shown in Supplementary Table 18, when evaluated against the full 1,000-μs MD trajectory, Mac-Diff achieves a roughly three orders of magnitude speed-up while outperforming a 100-μs MD simulation on the JS-PwD and JS-Rg metrics, and attaining comparable performance on the JS-TIC metric (0.441 versus

Fig. 6 | Sampling performance of Mac-Diff and eight competing methods on the allosteric protein AdK and five representative proteins from the Cfold40 test set. a–**f**, The 2D plots (top) show 100 sampled structures from each method, with the two axes representing the TM-scores relative to the two experimentally determined conformational states: AdK (**a**), TAF1 (**b**), PDI (**c**), SBDs of OpuA (**d**), tandem VHS and FYVE domains of HGS-Hrs (**e**), and surface layer protein (**f**). The overlays (bottom) highlight the Mac-Diff-generated structures (blue/claret) that achieved the highest TM-scores relative to the two reference structures, with corresponding TM-score and RMSD values also indicated. **g**, TM-score and RMSD distributions. Data were derived from a sample size of $n = 80$ independent protein structures, comprising 2 distinct conformations for each of the 40 test proteins. The box plots show the medians as centre lines, the 25th and 75th percentiles as lower and upper quartiles, and 1.5 times the interquartile range as whiskers. The average values of TM-score and RMSD are depicted as black diamonds.

https://doi.org/10.1038/s42256-026-01198-9



0.438). Unlike diffusion methods that directly output backbone structures, Mac-Diff reconstructs all-atom models via Rosetta protocols. Although this step adds computational overhead, PyRosetta-based parallelization greatly improves throughput when generating multiple samples. As shown in Supplementary Table 19, the Rosetta reconstructions time for generating the atom-level conformations for a batch of 1,000 structures of an input sequence accounts for about 12.5% of the runtime of the complete Mac-Diff pipeline. Overall, the combination of VPSDE and fast sampling enables the generation high-quality samples with substantially fewer denoising steps, making it computationally efficient for Mac-Diff to generate large-scale conformational ensemble.

Finally, although Mac-Diff achieves substantial speed-up over MD simulations and generates structurally diverse ensembles, it does not capture the temporal continuity inherent in MD trajectories. By contrast, MD simulations provide richer kinetic insights into protein behaviour. To bridge this gap, kinetic information can be approximated through simulation-based resampling of the generated ensemble[64]. As a complementary direction, future work will explore autoregressive diffusion frameworks that integrate data-driven learning with physical priors to directly generate temporally continuous and physically plausible trajectories.

Discussion

We presented Mac-Diff, a conditional diffusion model to recover protein conformational ensembles. Mac-Diff is characterized by a locality-aware attention module that aligns the residue neighbourhood across the conditional and the target views in a more delicate manner. In particular, the interacting neighbours of each residue, which is key to acquiring structural representations for effective structure denoising, were determined carefully via the combination of three sources, namely, (1) an isotropic Gaussian kernel that emphasizes proximal residue relations along the one-dimensional (1D) primary chain, (2) the contact map derived from ESM-2, providing long-range residue dependencies as indirect structural constraints, and (3) fine-grained residue embeddings that encode both biochemical identities of residues and contextualized sequence features. This comprehensive integration enables the model to extract rich conditional signals from the input sequence and to precisely control the alignment between the conditional and target views in a locality-aware manner. Empirically, the LAMA-attention-based diffusion showed promising results in recovering protein conformational distribution and identifying key conformational substates on fast-folding proteins and BPTI benchmark. Furthermore, Mac-Diff successfully predicted alternative structures for allosteric proteins AdK even without relying on MSA. Therefore, our method holds profound significance for drug discovery, particularly in identifying potential metastable states and advancing structure-based drug design by targeting allosteric effects and transient conformations, thereby paving the way for the development of highly specific therapeutic agents.

Our use of PLM-derived residue embeddings as input features aligns with Chai-1[36], a foundation model recognized for its strong performance in predicting protein–ligand complexes and protein multimers. While Chai-1 was primarily trained (and performs optimally) using MSA to capture co-evolutionary information, it also achieves competitive performance in single-sequence mode by leveraging PLM-derived embeddings. Taken together, this observation and our findings underscore the value of sequence-only pretraining in generating informative sequence representations for modelling both protein structural configurations and conformational heterogeneity.

Recently, diffusion models have gained considerable interest in biomolecular structure prediction, with methods such as AlphaFold3[65] and Boltz-1[66] achieving strong performance in modelling dominant conformations of protein–protein, protein–RNA and protein–ligand complexes. These methods rely on extensive structural context (for example, MSAs and experimentally determined templates) and generate atomic coordinates directly via diffusion in Cartesian space. By contrast, our method leverages PLM-derived residue embeddings as the primary input features to capture essential structural diversity and heterogeneity without external context. Furthermore, instead of predicting coordinates directly, we designed the LAMA-attention module to learn informative pairwise geometric constraints at the residue level.

There are several directions we plan to explore in the future. First, Mac-Diff was fine-tuned on only two publicly available MD databases with scanty timescales (from 100 ns to 500 ns) and limited protein diversity (1,674 proteins). The short MD simulations on these proteins may capture only small structural fluctuations, making it challenging to make predictions on larger proteins with higher degrees of structural freedom. Therefore, we plan to collect additional MD simulation data spanning larger timescales (μs to ms) from the scientific literature and curate them into large MD trajectory databases with unified, standardized simulation protocols. In light of these limitations, it is important to note that, while MD simulations are widely used in the literature as a surrogate for experimental ensembles in evaluating generative model quality, the MD data themselves inherently exhibit variability and uncertainty. Consequently, statistical divergence measures computed against such data should be interpreted with caution to prevent overinterpretation of the results.

Second, following the success of trRosetta and ProteinSGM in protein structure prediction and protein design, we will investigate the combination of the diffusion model and the energy minimization protocols in a complete end-to-end optimization framework[67,68] to improve the fidelity of the conformations generated.

Third, while Mac-Diff effectively sampled diverse conformations, identifying biologically meaningful states from the conformational ensemble remains a central challenge in protein conformation generation. Existing strategies—such as clustering or energy-based ranking are often suboptimal without ground-truth labels. Future directions would include incorporating experimental constraints or functional annotations into diffusion models to improve the interpretability and relevance of sampled states. Furthermore, while our sampling speed is substantially faster than MD simulations, it may be slower than some diffusion models that directly generate full-atom coordinates in Cartesian space[65]. To further improve both the quality and efficiency of Mac-Diff, we plan to explore more compact protein structure representations such as SE(3) backbone[31,69] and use advanced diffusion samplers such as EDM[70].

Finally, we plan to investigate whether generative models can be extended to capture conformational variation arising not only from intrinsic MD but also from sequence-level divergence across orthologs in different species, thereby providing a unified framework for understanding both the dynamic flexibility and the evolutionary diversity of protein structures.

Methods

Datasets and evaluation metrics

High-quality protein simulation data are very important for training and evaluating diffusion models for generating diverse and realistic structural ensembles of proteins. Below, we provide a detailed description of the datasets used, their sources and preprocessing steps, as well as the evaluation metrics.

Training data. The Mac-Diff model was developed using a pretraining–fine-tuning strategy: it was first pretrained on PDB structure data and then fine-tuned on MD trajectories to capture the conformational diversity reflected in both datasets. In the following, we introduce the details of the datasets.

Protein structure data. The structure data used in this Article were drawn from the PDB data source[71]. We have collected the proteins deposited in

the PDB on or before 1 May 2022 as the training set. The protein structure deposited between 2 May 2022 and 2 January 2023 were used as the validation set. To fully capture the conformational heterogeneity of each protein, we extracted all the available protein chains rather than preserving only the first chain available in the PDB database.

MD simulation data. The first MD dataset used for training was GPCRmd[41]. We collected MD trajectories that were released in GPCRmd before May 2023, including 595 G-protein-coupled receptors (GPCRs), the most common targets of approved drugs. The following MD simulation protocols were used in GPCRmd. First, the initial protein structures for simulation were obtained from the PDB, in which missing residues were recovered by MODELLER[72], steric clashes were removed, and residue protonation and tautomeric states were assigned using PROPKA[73]. Next, every GPCR structure was embedded within a lipid bilayer, and TIP3P water molecules and 0.15 mol l^{-1} NaCl ions were added. Finally, the parameters required in the simulation were obtained from the CHARMM36m force field, and each system underwent 3× 500 ns (total 1.5 µs) independent production runs under 310 K temperature and constant pressure with a 4-fs time step. Detailed simulation protocols are available via GitHub at ref. 74, and the GPCRmd datasets are publicly available at ref. 41. Overall, the GPCRs' sequence lengths range from 255 to 501, and the average sequence length is 317 for the 371 GPCR proteins after removing redundant sequences.

The second MD dataset is the ATLAS database[42], which includes MD simulation trajectories for 1,390 non-membrane protein chains with sequence lengths of at least 38 residues and spatial resolution below 2 Å. The ATLAS dataset was built from the PDB by removing proteins with identical structures/domains, or with more than ten consecutive residues missing. Only those chains with sufficient length (38 residues) and spatial resolution (2 Å and below), and only those satisfying the MolProbity's quality will be preserved. The initial structure used in MD simulation did not take into account water or ligand molecules; protein structures with fewer than five consecutive residues missing were modelled with MODELLER[72], and those with six to ten consecutive residues missing were modelled with AlphaFold2[7]. The initial protein structures were then placed in a triclinic box, with TIP3P water used for solvated and 0.15 mol l^{-1} NaCl ions used to establish the simulation system. All-atom MD simulations were performed with GROMACS software and parameters was generated with the CHARMM36m force field. Each protein include 3× 100 ns (total 300 ns) with different initial velocity, and the time step was fixed to 2 fs. The LINCS algorithm was used to constrain hydrogen atoms, and the particle-mesh Ewald method was applied for long-range electrostatic interactions.

The preprocessing of the two MD datasets included the following procedures. First, we excluded proteins that were difficult to handle due to extreme sequence lengths and whose MD trajectory information could not be extracted. Second, to avoid extracting redundant conformational structures from the short MD trajectories, we adopted the following protocol. To extract structure features from proteins simulation trajectories, we clustered every MD simulation trajectory into 100 clusters by TTClust[75] and randomly picked 10 conformations from each cluster. Finally, we obtained training and validation splits of 1,674 and 80 MD trajectories, respectively. More details are given in Supplementary Note 4, and the MD trajectories used in this Article are available via GitHub at can be found at ref. 76. Together, We constructed a combined MD dataset of 1,674 proteins by merging GPCRmd and ATLAS after preprocessing.

To ensure a rigorous evaluation of the generative model's generalization, we excluded from both the PDB and MD training datasets any proteins sharing more than 40% sequence identity with any test protein, as determined using MMseqs2, when training the Mac-Diff model[77], similar to the filtering strategy used in ESMFold[33]. Supplementary Table 11 reported the training data used by Mac-Diff and other competing methods for a complete reference. These state-of-the-art competing methods mainly used the PDB dataset as training data for generating the protein conformations.

Test data. We chose two different test datasets, corresponding to two types of evaluation task, as detailed below.

Task 1. To evaluate how well the generative model reproduces the conformational distribution of the original MD data, we used the complete DESRES dataset, which includes 12 structurally diverse fast-folding proteins, along with the BPTI protein. The MD simulation trajectories of these proteins were conducted by the DESRES group using specialized supercomputer, which are among the most well-recognized studies in the field of MD[43,52]. These trajectories provide valuable insights into the mechanisms of protein folding, illustrating rapid conformational changes and highlighting the complex dynamics involved in this fundamental biological process. The MD simulation trajectories of 12 fast-folding proteins used for evaluation were obtained by requesting them from the authors[43], which range from 100 µs to 1 ms in length, each including at least 10 folding–unfolding events. A 1-ms MD simulation trajectory of the BPTI was also obtained by request from the authors[52]. These test sequences have much longer MD trajectories (100 µs to 1 ms) than those in the ATLAS dataset used for training (100 ns), thereby enabling a rigorous assessment of the temporal generalization performance of the generative models. To evaluate the distribution of the generated 1,000 conformations with reference MD trajectories, we used a fixed stride to extract 1,000 frames for every protein. The C_α atoms were selected for the next analysis.

Task 2. In this task, we assessed the ability of generative models to capture essential conformational diversity by predicting biologically relevant alternative protein states. We evaluated two datasets. The first comprises the allosteric protein AdK, which has experimentally determined open and closed conformations[78]. The second is the Cfold dataset, containing 245 proteins, each protein annotated with two biologically relevant and structurally diverse conformations (TM-score <0.8), which were systematically selected through family-level sequence and structure comparisons combined with clustering analysis. Many Cfold proteins exhibit conformational changes—including hinge motions, rearrangements and fold switches—reflecting intrinsic structural flexibility and potential functional adaptability, such as ligand interactions[40]. For our study, we focused on proteins with fewer than 256 residues and minimal missing regions (<8 consecutive residues), resulting in a subset of 40 proteins with fully resolved alternative conformations, denoted Cfold40. The PDB IDs of the 40 test proteins with sequence information are available via GitHub at ref. 76, and the ground-truth structures are available via Zenodo at ref. 79 (ref. 40).

Evaluation metrics. Appropriation metrics are crucial in evaluating the capacity of generative models in recovering realistic protein structures and the underlying conformational distributions. Following the literature[19,24,38], we have selected a number of widely used metrics, including:

(1) JS divergence for pairwise Euclidean distances of C_α atoms (JS-PwD). Only C_α atoms separated by at least four residues are considered. To compare the distribution of pairwise distances of C_α atoms from a generative model with the reference MD samples, we quantify the pairwise distance distribution as a histogram with 50 bins, and a pseudo-count value ε was set to 10^{-6} for smoothing the distribution[38]. Then, the JS-PwD was computed by

$$\mathrm{JS}(P||Q) = \frac{1}{2}(\mathrm{KL}(P||Q) + \mathrm{KL}(Q||P)), \quad (1)$$

where P and Q are histogram or probability distributions of the reference MD data and the generated data, respectively, and an approximate

Kullback–Leibler divergence (KL distance) between two histograms is computed by

$$\mathrm{KL}(P||Q) = \sum_{k}^{N} P_k \log\left(\frac{P_k}{Q_k}\right),$$

where k represents the bin index and N is the total number of bins. To avoid overrepresentation of local structure, the distance matrix of C_α atoms was calculated with a diagonal offset larger than three.

(2) JS divergence on the radius of gyration (JS-Rg). The radius of gyration is calculated using the root mean square distance of C_α atoms to the centre of mass r_{CM} as

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i (r_i - r_{CM})^2}{\sum_{i=1}^{N} m_i}}, \quad (2)$$

where m_i is the mass of the C_α atom for residue i, r_i is the position of the C_α atom for residue i, and r_{CM} is calculated as

$$r_{CM} = \sqrt{\frac{\sum_{i=1}^{N} m_i r_i}{\sum_{i=1}^{N} m_i}}. \quad (3)$$

Having computed the R_g for each conformation in the distribution, we can then obtain the histogram of R_g and compute the JS divergence between the reference/MD distribution and the generated distribution as in equation (1).

(3) JS divergence on the top-two TICs (JS-TIC). TIC-Analysis is a useful analysis tool to project trajectory data onto their maximum-autocorrelation directions[44,45], corresponding to the reaction coordinates that explain the slowest conformational changes and are highly relevant for protein (un)folding. To compute the JS-TIC score, we projected the pairwise C_α distance matrices of both MD and generated conformations onto the first two dominant TICs of the MD trajectories using the Deeptime library[80]. Following Str2Str[24], each TIC dimension was discretized into 50 equally spaced bins with a pseudo-count of $\varepsilon = 10^{-6}$ for smoothing. We then calculated the JS divergence between the resulting histogram distributions of MD and model generated conformations for each TIC dimension and averaged the results to obtain the final JS-TIC score. We also visualized the density (contours) of the conformational distribution using Gaussian kernel density estimation (bandwidth determined by the default Scott's rule). Overall, this metric assesses not only the fidelity of a generative model in capturing long-timescale dynamics and relative state populations, but also facilitates visualization of its deviations from the MD trajectory distribution within the 2D tIC subspace.

(4) Structural diversity and fidelity. These are two complementary metrics for evaluating the quality of the generated conformational ensemble relative to MD data. The diversity score quantifies structural variability based on the pairwise lDDT-scores within the ensemble, defined as

$$\text{diversity} = 1 - \frac{1}{N_{\text{pairs}}} \sum_{i<j} \mathrm{lDDT}(s_i, s_j), \quad (4)$$

where N is the number of conformations generated by the model, $N_{\text{pairs}} = N(N-1)/2$ is the total number of unique conformation pairs, and s_i and s_j are the ith and jth generated structures, respectively. The lDDT score quantifies local structural similarity by evaluating the preservation of inter-residue C_α distances within local neighbourhoods[47]. Because lDDT is highly sensitive to local structural deviations, it can capture subtle conformational differences that may be missed by global metrics. We therefore adopted lDDT to quantify the structural diversity within each generated conformational ensemble.

The Fidelity score, by contrast, evaluates how well the generated conformations reproduce the reference distribution obtained from MD simulations. It is computed as the JS divergence between the distributions of pairwise C_α atom distances in the generated and reference ensembles, as in equation (1).

(5) The EMD based on structural dissimilarities. We used the EMD to quantify deviations between the conformational ensemble generated by the model and that obtained from MD simulations. Specifically, we first constructed a cross-distance matrix whose entries represent the structural dissimilarities between each MD conformation and each model-generated conformation. Two complementary measures were used to quantify structural differences: global dissimilarity, defined as 1 − TM-score[48], and local dissimilarity, defined as 1 − lDDT-score[47,49]. The EMD between the two ensembles (each sampled with 1,000 conformations) was then computed via optimal transport using the Python Optimal Transport library[81], providing a distribution-level metric of how well the generated ensemble matches the reference MD ensemble. The EMD values were averaged over all 12 proteins in the fast-folding protein dataset for comparison across methods.

(6) Root of mean squared error on the residue contact probability map ($\mathrm{RMSE}_{\text{cmap}}$). The residue contact probability map p measures the probability that each pair of residues contact with each other across all the conformations for a given protein, based on the distance between C_α atoms with a threshold d_r of 10 Å following Arts et al.[38]. Specifically, p_{ij} is computed as proportion of conformations in which residue i and residue j is below the distance threshold, as

$$p_{ij}^{(l)} = \frac{1}{N} \sum_{l=1}^{N} \mathbb{1}\left(d_{ij}^{(l)} \le d_r\right), \quad (5)$$

where l is the conformation index, and N is the number of all the conformations from MD data or from generative models. Then, we can measure the discrepancy between the reference probability map and the generated probability map by the root of mean squared error,

$$\mathrm{RMSE}_{\text{cmap}} = \sqrt{\frac{1}{N_{\text{pairs}}} \sum_{i<j} (\log(p_{ij}) - \log(\hat{p}_{ij}))^2}, \quad (6)$$

where $N_{\text{pairs}} = N(N-1)/2$ is the number of residue pairs for a protein. Here, we take the log of the probabilities following the same setting in the literature[19], which, by mapping the probabilities in [0, 1] to the log-space in the range $(-\infty, 0]$, can better quantify the difference between two probabilities.

(7) Average pairwise TM-score (TM_{var}) is another measure of the diversity of a conformational ensemble in terms of the average pairwise TM-scores, defined as

$$\mathrm{TM}_{\text{var}} = \frac{1}{N_{\text{pairs}}} \sum_{i<j} \mathrm{TM}(s_i, s_j), \quad (7)$$

where $N_{\text{pairs}} = N(N-1)/2$ is the number of the unique conformation pairs, and s_i and s_j is the ith and jth generated structure, respectively. The TM-score is a widely used metric for evaluating structural similarity between protein conformations. It is robust to variations in protein length and sensitive to overall fold topology, making it highly effective at determining whether two structures share the same fold[48,82]. A higher value of TM_{var} indicates lower diversity.

(8) MAE diversity, the absolute deviation between the diversity of model-generated samples (TM_{var} in equation (7)) and the true diversity (the TM-score between the two key conformational states, $\mathrm{TM}_{\text{conf1/conf2}}$), averaged over N test proteins, as

$$\text{MAE diversity} = \frac{1}{N} \sum_{n=1}^{N} \left| \mathrm{MT}_{\text{var}}^{(n)} - \mathrm{MT}_{\text{conf1/conf2}}^{(n)} \right|. \quad (8)$$

Following Jing et al.[25], the TM-score between the two ground-truth conformations, $TM_{conf1/conf2}$, serves as a reference measure of a protein's conformational diversity (for cases annotated with two key states), where higher values correspond to lower diversity.

(9) Residue flexibility. The flexibility of a residue quantifies how different the relative position of this residue can be across different, but globally aligned conformations of this protein. For the mth residue in a protein, it was defined as

$$\mathrm{RMSF}_m = \sqrt{\frac{1}{N_{\mathrm{pairs}}}\sum_{i<j}(d_{ij})^2}, \tag{9}$$

where $i,j = 1, 2, \ldots, N$ are the index of N observed conformations for each protein, $N_{\mathrm{pairs}} = N(N-1)/2$ is the number of conformation pairs, and d_{ij} is the Euclidean distance between residue m in the ith conformation and residue m in the jth conformation after global conformation alignment. Following Jing et al.[25], we assessed the consistency of flexibility amplitudes by computing residue-level flexibility correlations at the protein level and then reporting the mean correlation across all test proteins.

(10) MAE flexibility, the absolute deviation of the residue-level flexibility (RMSF in equation (9)), is defined as

$$\text{MAE flexibility} = \frac{1}{L}\sum_{i=1}^{L}\left|\mathrm{RMSF}_i^{\mathrm{true}} - \mathrm{RMSF}_i^{\mathrm{pred}}\right|. \tag{10}$$

where L is the number of residues in the protein of interest. In practice, the protein-level MAE flexibility is computed for each protein and then averaged over all test proteins.

Score-based conditional generative model

Score-based generative models[83] have shown potential in generating high-quality proteins[84]. Here, we use a continuous-time score-based generative model to encapsulate protein conformational ensembles conditioned on protein sequence.

In the forward process, the model incrementally introduces Gaussian noise to protein geometric representations as illustrated in Fig. 1a, ultimately mapping it to a prior distribution. Given the protein geometric representation $\mathbf{x}$ and a time step t, the forward process can be modelled as

$$\mathrm{d}\mathbf{x} = f(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{11}$$

where $f(\mathbf{x}, t)$ is the drift coefficient, $g(t)$ is known as the diffusion coefficient and w is Brownian motion. We use the variance-preserving SDE discretization, and the forward process can be defined as

$$\mathrm{d}\mathbf{x} = -\frac{1}{2}\beta(t)\mathbf{x}\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}\mathbf{w}, \tag{12}$$

where $\beta(t)$ is the noise scale schedule in time step t.

In the reverse process, the model iteratively denoises the backbone geometric structure representation. The reverse process can be modelled as

$$\mathrm{d}\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\nabla_x \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\bar{\mathbf{w}}, \tag{13}$$

where $\nabla_x \log p_t(\mathbf{x})$ is the score function in time step $t \in [0, T]$, and $\bar{\mathbf{w}}$ is Brownian motion in the reverse time direction. Given the protein sequence seq as the condition, Mac-Diff estimates the score function $\nabla_x \log p_t(\mathbf{x})$ using the following score matching training objective[83,85]:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_t\left\{\lambda(t)\mathbb{E}_{x(0)}\mathbb{E}_{x(t)|x(0)}[\| s_\theta(x(t), t, \mathrm{seq}) - \nabla_x \log p_t(x(t)|x(0))\|_2^2]\right\}. \tag{14}$$

Here, $\mathbb{E}$ denotes the 'expectation' over a data distribution. Specifically, $\mathbb{E}_t$ is the expectation over the time variable t drawn from the uniform distribution Uniform(0, T); $\lambda(t) \in \mathbb{R}^+$ is a positive weighting function corresponding to time t; $\mathbb{E}_{x(0)}$ represents the expectation over the true data distribution $p_0(x)$ at time 0; $\mathbb{E}_{x(t)|x(0)}$ denotes the expectation over the conditional distribution $p_t(x(t)|x(0))$, and this distribution defines how the clean data are perturbed to a noisy version $x(t)$ at timestep t, so that the model can be trained to denoise across all noise levels. The $s_\theta()$ term is the score-based denoising network, and $\nabla_x \log p_t(x(t)|x(0))$ is the score of the conditional distribution $p_t(x(t)|x(0))$ at time t.

Modal-aligned conditional diffusion. Figure 1a is the backbone geometric representation of the protein with L residues, which is a $L \times L \times 5$ tensor including: inter-residue distances between C_β atoms, symmetry dihedral angle ω ($\omega_{ij} = \omega_{ji}$ for the ijth residue pair), asymmetry dihedral angle θ and planar angle ϕ ($\theta_{ij} \neq \theta_{ji}$, $\phi_{ij} \neq \phi_{ji}$) for the ijth residue pair). Moreover, to accelerate the training process and generate protein structures for sequences of varying lengths, a padding channel is used to indicate the actual sequence length. The pairwise geometric tensor has the advantage of being invariant to global 3D rotation and translation.

Figure 1b illustrates the Mac-Diff workflow. Given a protein sequence, Mac-Diff implements a conditional diffusion in which the forward process gradually transforms pairwise geometric representations (tensors) into Gaussian noise and the reverse process uses a U-Net as the building block of the score-based denoising network, following Song et al.[83]. Here, the $L \times L \times 5$ residue-pair tensor is first fed into a convolutional layer with a 3×3 kernel, which transforms the 5-dimensional vector into a higher-dimensional feature with 128 channels, thereby enriching its latent geometric representation. The enriched representation then goes through five downsampling stages and five upsampling layers with skip connections across each downsampling/upsampling pair. Inside each layer, a ResNet block and a TransFormer block interleave with each other twice and end with either a 2×2 mean pooling with a stride of 2 for downsampling, or a 2×2 nearest-neighbour interpolation for upsampling. Each ResNet block has two 3×3 convolutional layers. The first convolutional layer extracts features from the residue-pair tensor, which is then mixed with the denoising time-step embedding. The mixed features are subsequently passed to the second convolutional layer. Each TransFormer block includes a self-attention block and a LAMA-attention block. The self-attention block captures the global dependencies among the L^2 residue-pair representations in the geometric tensor. The LAMA-attention module then injects protein sequence information into the target view ($L \times L \times 128$ geometric tensor) in such a manner that locally interacting residues are preserved across these two views.

Figure 1c is the LAMA-attention module. It injects protein sequence information into the structural domain by carefully aligning the neighbourhood structure in both views. First, each residue in the input sequence is encoded by the embedding from ESM-2[33]. Then, a relational matrix is used to comprehensively quantify residue relations in both the sequence space and the 3D structure space to update residue and residue-pair representations. This relational matrix is composed of three sources: (1) a homogeneous Gaussian function encouraging adjacent residues in the 1D sequence to interact with each other in 3D space; (2) the contact map matrix computed by the ESM-2, which can be deemed as a prediction of the contact probabilities between residues (by exploiting huge collection of sequences in the pre-training stage and a small number of proteins structures in the regression stage)[33]; and (3) the residue-level attention matrix based on the residue embedding features via ESM-2 (residue features are turned to keys and queries to compute the attention matrix). These three matrices are combined through a weighted summation to form the final residue-level relational matrix, which is then used to update each residue's features via the standard Q–K–V attention mechanism. The attention-updated residue features are then passed through the Triangular Multiplicative Update module from AlphaFold2[7] and subsequently concatenated with the noisy $L \times L \times 128$ geometric tensor to form a complete pairwise representation. This representation serves

as an intermediate feature in the denoising network, which is ultimately used to recover the backbone geometric parameters.

Overall, the contact map serves as a coarse, global prior distilled from the attention structure of PLMs to capture averaged residue–residue dependencies and relational patterns. By contrast, the residue embedding provides a fine-grained, context-aware representation that incorporates biochemical identity, evolutionary covariation and implicit structural tendencies, thereby complementing the contact map with enriched local information. The diffusion architecture integrates these two sources of information as a reliable conditional guide to enhance the pairwise geometric representations while simultaneously performing iterative denoising. Together, the global prior and local representations are integrated in Mac-Diff to generate diverse and physically realistic conformational ensembles.

Locality-aware modal alignment attention. Generating 3D structures for a given sequence requires more accurate spatial (neighbourhood) alignment between these two modalities than compared with generating images from descriptive texts. In the latter case, the alignment between the two modalities (text tokens and image pixels) is unstructured and is typically not under prior algorithmic controls. In comparison, when mapping a protein sequence to a 3D-structural space, we would require an accurate estimation of the spatial neighbourhood to identify which residues/amino acid types in the sequence domain are spatially close to an interested residue in the structural domain due to physical and chemical interactions. This locality-aware modal alignment enables the computation of informative structural features for the denoising network in the diffusion model and is key to injecting sequence information into the structural domain. In the following, we describe in detail the LAMA-attention module.

Suppose we are given an amino acid sequence with L residues, as $S = \{s_1, s_2, \ldots, s_L\}$, where $s_i \in \mathbb{R}^{1\times f_d}$ is the embedding vector for the ith residue. We use an $L \times L \times 5$ tensor $\mathbb{A}$ to represent the (noisy version) of the latent geometric representations of the protein structure. In particular, it can be organized as $\mathbb{A} = \{a_{ij}\}_{L\times L}$, where each 'fibre' $a_{ij} \in 1 \times 5$ is the pair representation of the ith and jth residue and can be vividly deemed as a pixel of the latent geometric representations $\mathbb{A}$. The goal of the reverse process in the diffusion model is to iteratively improve the noisy version of $\mathbb{A}$ so that it can finally be used to recover the protein structure ensembles. During the reverse process, the denoising network is used to estimate the score function and progressively reduce the noise at each step, ultimately reconstructing a clean version of the protein geometric representation.

Considering that $\mathbb{A}$ includes key geometric parameters of the protein structure, that is, pairwise C_β distance, dihedral angle and planar angle, it is critical to enrich each a_{ij} by contextualized information from residue i and residue j, as well as those residues that are in close vicinity to the interested residue pair. To achieve this, we will first encode the contextual information of residues i and j using the self-attention scheme, and then through the triangular multiplicative updates, as described below.

We choose $s_i \in \mathbb{R}^{1\times 1280}$ as the vector representation for the ith residue from the ESM-2 model, which is supposed to capture the contextual information of each residue at the sequence level, along with the broader evolutionary, structural and biological constraints derived from the protein family. They are transformed to keys, values and queries through the transform matrices W_k, W_v and W_q as follows:

$$q_i = W_q s_i, k_i = W_k s_i, v_i = W_v s_i. \tag{15}$$

In computing the pairwise attention matrix for the L residues, we considered the hybrid of three sources of relations as follows:

$$W_{ij} = \lambda \times \frac{\exp(q_i^\top k_j/\sqrt{f_d})}{\sum_{j=1}^{L}\exp(q_i^\top k_j/\sqrt{f_d})} + (1-\lambda)(\alpha \times C_{ij} + \beta \times \exp(-\| i-j\|^2/2h^2)). \tag{16}$$

The W matrix is normalized so that each row sums up to 1. The first term is the inner-product-based attention score between residues i and j based on their respective embeddings. The second term C_{ij} is the contact map matrix in ESM-2 that is obtained by mixing the attention matrices across different attention heads through a logistic regression using a dozen protein structures as a weak supervision, which can be deemed as a comprehensive estimate of the spatial contact probability and functional relationship of residue i and j by comprehensively taking into account evolutionary, structural and biological information[86]. The third term is a Gaussian matrix in which the interaction between two residues i and j decays smoothly with respect to their distance along the 1D sequence. Here, λ, α and β are weights assigned to these three terms. The bandwidth h of the Gaussian is a small value so as to enforce the spatial closeness between residues that are very close to each other in the sequence. This simple physical prior serves as a regularization term so that the resultant structures generated are more physically feasible.

Compared with traditional positional encoding, the contact map from ESM-2 introduces a more informative, probabilistic estimate of the spatial and functional dependencies between residues into the attention matrix W. This allows updating the representation of each residue based on its most likely local environment, providing highly informative contextual features for the denoising network. In this sense, our attention module can be more locality-aware than the standard cross-attention in current diffusion models. Empirically, this is more effective in generating realistic conformational ensembles than conventional cross-attention (see Supplementary Note 1 for more detailed comparisons).

After residue features are updated by the attention mechanism with the W matrix in equation (16), the information of any two residues in the input sequence is concatenated to form a 2D residue-pair representation. Specially, we use z_{ij} to denote the ij-residue pair representation, as

$$z_{ij} = \text{Concat}\,[Wv_i,\ Wv_j]. \tag{17}$$

Then, we apply the triangular multiplicative update module, originally introduced in the triangular attention mechanism of AlphaFold2, on top of the pair representation z_{ij} in equation (17). This update enhances the modelling of geometric relationships among residue triplets, which enables richer structural reasoning and facilitates modelling non-local interactions and forming correct global folds. The update is denoted as

$$\tilde{z}_{ij} = \text{Triangular\_Multiplicative\_Update}(z_{ij}). \tag{18}$$

In this step, each ij-pair representation z_{ij} is updated by sequentially integrating information from a third residue, k, through outgoing and incoming edge updates. These updates explicitly exploit the pairwise interactions within the i–j–k triplet: the edges ik and jk are used for the outgoing update, while the edges ki and kj are used for the incoming update (see Supplementary Note 6 for details).

After updating the residue-pair representation via triangular multiplicative updates, we integrate the resulting $\tilde{z}_{ij}$ (equation (18)) with the residue-pair representation a_{ij} in the geometric tensor $\mathbb{A}$ (after it has passed through the self-attention module of the TransFormer block) for subsequent processing. Specifically, the two representations are concatenated and passed through a fully connected layer with GELU activation to promote effective information exchange between them, as follows:

$$\tilde{a}_{ij} = \text{Linear}(\text{GELU}(\text{Linear}(\text{Concat}[a_{ij}, \tilde{z}_{ij}]))). \tag{19}$$

The finalized residue-pair representation $\tilde{a}_{ij}$ integrates both sequence and geometric information of residues i and j, serving as a comprehensive feature for recovering inter-residue distances and orientations (angles). Subsequently, it is passed either to a

downsampling layer, implemented as a 2 × 2 mean pooling operation with a stride of 2 in the UNet's downsampling layer, or to an upsampling layer with the same configuration, thereby completing the full information flow.

The LAMA-attention module can generate highly effective and contextualized residue-pair features, which are plugged in the TransFormer block to fully integrate with its self-attention module for denoising.

Folding by energy minimization with predicted restraints. After obtaining the fully denoised pairwise geometric representation ($L \times L \times 4$ tensor) that depicts residue-level protein backbone structure through Mac-Diff, we then transformed it to a all-atom level protein structure using Rosetta through minimization of the energy potential. Rosetta is a versatile software suite for computational biology, particularly designed for protein modelling and the structures prediction of macromolecules[87].

We first introduce the detailed energy minimization procedures used in task I (recovering protein conformational distributions). Following Yang et al.[88], we used the trRosetta folding protocol. We first transformed the predicted continuous values in the $L \times L \times 4$ tensor into one-hot-encoded vector. For example, the pairwise distances between C_β atoms were quantized into 37 bins (36 equal-sized bins in the range 2–20 Å, along with an additional indicator on whether the pair of residues are in contact or not), thus representing each pairwise distance as a 37-dimensional vector. Similarly, the dihedral angles ω and θ, and the bond angles ϕ were converted into 24-, 24- and 12-dimensional one-hot vectors, respectively. Subsequently, we applied the Gaussian function $\exp(-x^2/2\sigma^2)$ to transform the one-hot (binned) representations into smoother probabilities, with the bin centre as the Gaussian mean and a bandwidth parameter σ for controlling the level of smoothing. By doing this, Rosetta can search from a potentially larger space of geometric parameter combinations to locate a folding configuration with lower energy potential.

To determine a suitable smoothing parameter (σ), we have examined a number of candidate choices and selected the best one by evaluating their resultant JS-PwD errors on the 17 proteins from the validation set. For the pairwise residue distance channel, σ was selected from 1 to 5 with intervals of 1, and the value giving rise to the conformation set with the lowest divergence with the real conformation distribution (JS-PwD) was selected. The variance parameters for the other three channels (bond angles ϕ channel, dihedral angle ω channel and dihedral angles θ channel) were all chosen to be the same as those of the distance channel to reduce the computational cost of generating samples from the diffusion models. Empirically, σ was set to 1 in the task of generating protein conformational ensembles.

Next, the obtained probability distribution of every residue pair was converted to energy potential as used in trRosetta. Specifically, the probability value of the last distance bin was set to a reference state, so that each binned distance was transferred to a distance score:

$$\text{score}^d(i) = -\ln(p_i) + \ln((d_i/d_N)^\alpha p_N), i = 1, 2, \ldots, N, \quad (20)$$

where $N = 37$ is the total number of distance bins, p_i and p_N is the probability of ith and the Nth distance bin, respectively, d_i and d_N is the distance for the ith and the Nth distance bin, and α is fixed at 1.57 in the normalized distance term[88].

For the two dihedral angles ω and θ and the bond angle ϕ, the equations used to convert their angle probability distributions (of dimension $L \times L \times 24$ or $L \times L \times 12$) to orientation scores were similar to those used in obtaining the distance scores, but without normalized terms.

$$\text{score}^o(i) = -\ln(p_i) + \ln(p_N), i = 1, 2, \ldots, N. \quad (21)$$

The pairwise residue distances and the three pairwise orientations of each residue pair were converted to energy potential by spline function and used as restraint to folding protein structures. Specifically, for the potential of distance d, angle ϕ and torsion angles ω and θ, the AtomPair restraint, Angle restraint and Dihedral restraint were applied, respectively. Next, a coarse-grained model was built with MinMover in Rosetta, using L-BFGS for optimization with 1,000 iterations and convergence cut-off of 0.0001. Other restraints such as ramachandran (rama), the omega and the steric repulsion van der Waals forces (vdw) and the centroid backbone hydrogen bonding (cen_bb) were also included. Overall, the model was built by short-, medium- and long-range restraints together, and the centroid models with the lowest energy was chosen for the next full-atom relaxation using the FastRelax protocol. The ref2015 scoring function and a probability threshold of 0.15 were used in the relax protocol. The final model was selected by both Rosetta energy and restraint scores.

In the second evaluation task of predicting alternative conformations of allosteric proteins in Cfold40, we no longer used the pairwise distances/angles to fit splines to generate energy potential for Rosetta energy minimization, considering the larger conformational fluctuations in protein conformations. Instead, following the strategy of ProteinSGM[84], we set the distance d and dihedral ϕ value as the mean of the HARMONIC functions and set the ω and θ values as the mean of the CIRCULAR-HARMONIC functions. The standard deviation was set to 2.0 Å for d and 20° for θ, ω and ϕ. This allowed reasonable structures to be generated with loose constraints given distance and angle values. Here, we ran minimization five times with MinMover and picked the centroid model with the lowest energy. No constraints were enforced regarding distances greater than 12 Å. Again, the models were built by short-, medium- and long-range restraints in every minimization. We then performed ten rounds of full atomic relaxation on each centroid model using FastRelax, and the final model was picked by the ref2015 scoring function with the lowest energy.

Experimental settings

In the training stage, we used the U-Net architecture and used an AdamW optimizer with a weight decay of 0.02 for all models. The sequence embedding was obtained from a pretrained PLM ESM-2, together with its contact-map matrix (that was computed by mixing the attention matrices in each attention head of ESM-2 using a dozen protein structures as a weak supervision[33]). The bandwidth in the diagonal Gaussian was fixed as 2, and the weight α and β in equation (16) was set at 0.3 and 0.7, respectively. The weight λ in equation (16) was optimized in the learning process, with its value initialized to 0.0003. The U-Net architecture included five downsampling/upsampling stages with skip connections across each downsampling/upsampling pair. At each resolution level, a ResNet block and a Transformer block are interleaved twice, followed by either a downsampling operation (2 × 2 mean pooling with a stride of 2) or an upsampling operation (2 × 2 nearest-neighbour interpolation). The noise schedule hyperparameters β_{min} and β_{max} used in equation (12) were set 0.1 and 20, respectively. The number of time steps, N, was fixed to 2,000 in the training stage. In the sampling stage, we used the DDIM[63] sampling method with 50 steps for denoising, which offers an efficient sampling while preserving the generation quality.

We adopted a pretraining–fine-tuning strategy, in which Mac-Diff was first pretrained on PDB structural data to learn generalizable priors from the diverse repertoire of experimentally resolved folded states and subsequently fine-tuned on MD trajectories to capture the dynamic conformational variability and rare transitions intrinsic to protein motions. This two-stage paradigm not only enhances computational efficiency during training and inference, but also improves performance by jointly leveraging complementary structural information from static PDB data and dynamic MD simulations.

In the pretraining stage, we constructed two versions of the Mac-Diff model using PDB structural data, tailored to proteins of different sequence lengths: (1) Mac-Diff-128, trained on 619,045 cropped

protein segments containing 40–128 residues, and (2) Mac-Diff-256, trained on 579,476 cropped segments containing 80–256 residues. This split strategy maximized the use of limited computational resources and improved training efficiency. During testing, proteins with fewer than 128 residues were evaluated with Mac-Diff-128, while those with 129–256 residues were evaluated with Mac-Diff-256. For simplicity, we collectively refer to these pretrained variants as Mac-Diff-PDB. All Mac-Diff-PDB models were trained using residue embeddings from the ESM-2-650M model (embedding dimension 1,280).

The Mac-Diff-128 model was trained with downsampling resolution {128, 64, 32, 16, 8} in the diffusion architecture, with a mini-batch size of 48, and altogether 12 epochs with learning rate schedule of $\{1e-3\times3, 4e-4\times3, 2e-4\times3, 1e-4\times3\}$. The Mac-Diff-256 model was trained with downsampling resolution {256, 128, 64, 32, 16, 8}, mini-batch size of 16, and 16 epochs with a fixed learning rate $\{5e-4\times4, 1e-4\times12\}$.

In the fine-tuning stage, the pretrained Mac-Diff models were fine-tuned on MD trajectories to further capture protein conformational diversity. Specifically, we fine-tuned both Mac-Diff-128 and Mac-Diff-256 on MD trajectories. For Mac-Diff-128, we used a mini-batch size of 48 sequence–conformation pairs and trained for a total of 6 epochs (approximately 420,000 iterations). The protein sequences from MD data were standardized to a fixed size of 128 residues (zero-padding shorter sequences and random-cropping longer ones), with downsampling resolutions 128, 64, 32, 16, 8 and a fixed learning rate of 1×10^{-5}. For Mac-Diff-256, the same fine-tuning protocol was adopted, except that the input sequences were standardized to a fixed size of 256 residues, and downsampling resolutions 256, 128, 64, 32, 16, 8 were applied. A mini-batch size of 16 was used, and training was carried out for a total of 840,000 iterations (8 epochs in total).

In the inference stage, when comparing the capability of Mac-Diff with other models for generating alternative conformations, we used source codes of the following methods: (1) AlphaFold2 and AlphaFold2_dropout (using ColabFold, available at https://github.com/sokrypton/ColabFold), (2) AF-Cluster (available at https://github.com/HWayment-Steele/AF_Cluster), (3) MSA-subsampling (available at https://github.com/delalamo/af2_conformations), (4) AlphaFlow (available at https://github.com/bjing2016/alphaflow), (5) DiG (available at https://github.com/microsoft/Graphormer/tree/dig-v1.0/distributional_graphormer) (6) Boltz-2 (available at https://github.com/microsoft/bioemu) (7) BioEmu (available at https://github.com/microsoft/bioemu).

MD simulation and Mac-Diff setting for sampling speed assessment

The initial structure of protein WW domain was obtained from the authors[43]. For MD simulations, the protein WW domain (562 atoms) was solvated with TIP3P water molecules (16,297 atoms) and neutralized with 3 chloride ions to establish the complete simulation system. The Amber14SB force field was used for the parameters of protein residues. OpenMM v.8.2.0 was used to carry out the following MD simulation process. First, the system was minimized for 1,000 steps, followed by heating to 300 K over 10,000 steps under NVT conditions. Next, the system underwent 100 ns of production runs under NPT conditions with a 2-fs integration timestep on an NVIDIA A100 GPU. The runtime of this NPT stage was used to estimate the speed of the traditional MD simulation. The temperature was controlled using Langevin dynamics, and pressure was maintained with a Monte Carlo barostat. Long-range electrostatic interactions were treated using the particle mesh Ewald algorithm, and a 10 Å cut-off was used for van der Waals and short-range interactions. During the inference stage of the Mac-Diff model, we set the batch size to 100 on an NVIDIA A100 80G GPU, and estimated the model's sampling time for generating 1,000 conformations of the WW domain.

Data availability

MD trajectories of GPCRMD and ATLAS datasets are available at https://www.gpcrmd.org/dynadb/datasets/ (ref. 41) and https://www.dsimb.inserm.fr/ATLAS (ref. 42), respectively. We collected 1,390 trajectories from the ATLAS datasets and 371 trajectories from the GPCRmd datasets. The experimental structures deposited before 1 May 2023, were used for the training phase and can be downloaded from the PDB dataset (https://www.rcsb.org)[71]. The reference MD trajectories for fast-folding proteins and BPTI are available from the authors of refs. 43,52. Sequence files for the 12 fast-folding proteins and BPTI, as well as PDB IDs for Adk and the Cfold40 test set, are availiable via GitHub at https://github.com/Paulie-ai/Mac-Diff (ref. 76) and Zenodo at https://doi.org/10.5281/zenodo.17936479 (ref. 89). The corresponding structural files can be retrieved from the PDB using their respective accession code: PDBID (https://www.rcsb.org/structure/PDBID). Source data are provided with this paper.

Code availability

Source code for the Mac-Diff model, inference scripts and model weights are available via GitHub at https://github.com/Paulie-ai/Mac-Diff (ref. 76) and Zenodo at https://doi.org/10.5281/zenodo.17936479 (ref. 89). Our model was developed with Python, PyTorch, Numpy and flash-attention. We used biopython, mdtraj and fair-esm to extract initial protein sequence and structure representations. More details can be found in the code repository. Data analysis and plot for experimental were conducted using Python, Numpy, Matplotlib, MDTraj, seaborn, SciPy, pandas and Biopython. Protein structure visualization and rendering were done with Pymol v.2.5.2 (https://github.com/schrodinger/pymol-open-source) and ChimeraX[90]. All Rosetta protocols were implemented under PyRosetta4[91].

References

1. Wei, G., Xi, W., Nussinov, R. & Ma, B. Protein ensembles: how does nature harness thermodynamic fluctuations for life? The diverse functional roles of conformational ensembles in the cell. *Chem. Rev.* **116**, 6516–6551 (2016).
2. Xie, T., Saleh, T., Rossi, P. & Kalodimos, C. G. Conformational states dynamically populated by a kinase determine its function. *Science* **370**, eabc2754 (2020).
3. Bahar, I., Lezon, T. R., Yang, L.-W. & Eyal, E. Global dynamics of proteins: bridging between structure and function. *Annu. Rev. Biophys.* **39**, 23–42 (2010).
4. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338**, 1042–1046 (2012).
5. Shaw, D. E. et al. Anton 2: raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. In *Proc. SC'14 International Conference for High Performance Computing, Networking, Storage and Analysis* (eds Damkroger, T. & Dongarra, J.) 41–53 (IEEE, 2014).
6. Yang, Y. I., Shao, Q., Zhang, J., Yang, L. & Gao, Y. Q. Enhanced sampling in molecular dynamics. *J. Chem. Phys.* **151**, 070902 (2019).
7. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
8. Del Alamo, D., Sala, D., Mchaourab, H. S. & Meiler, J. Sampling alternative conformational states of transporters and receptors with alphafold2. *eLife* **11**, e75751 (2022).
9. Monteiro da Silva, G., Cui, J. Y., Dalgarno, D. C., Lisi, G. P. & Rubenstein, B. M. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. *Nat. Commun.* **15**, 2464 (2024).
10. Stein, R. A. & Mchaourab, H. S. SPEACH_AF: sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLoS Comput. Biol.* **18**, e1010483 (2022).
11. Wayment-Steele, H. K. et al. Predicting multiple conformations via sequence clustering and AlphaFold2. *Nature* **625**, 832–839 (2024).

12. Heo, L. & Feig, M. Multi-state modeling of G-protein coupled receptors at experimental accuracy. *Proteins* **90**, 1873–1885 (2022).
13. Vani, B. P., Aranganathan, A., Wang, D. & Tiwary, P. AlphaFold2-RAVE: from sequence to Boltzmann ranking. *J. Chem. Theory Comput.* **19**, 4351–4354 (2023).
14. Ourmazd, A., Moffat, K. & Lattman, E. E. Structural biology is solved—now what? *Nat. Methods* **19**, 24–26 (2022).
15. Saldaño, T. et al. Impact of protein conformational diversity on AlphaFold predictions. *Bioinformatics* **38**, 2742–2748 (2022).
16. Gao, M., Nakajima An, D., Parks, J. M. & Skolnick, J. Af2complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* **13**, 1744 (2022).
17. Gao, M., An, D. N. & Skolnick, J. Deep learning-driven insights into super protein complexes for outer membrane protein biogenesis in bacteria. *eLife* **11**, e82885 (2022).
18. Gao, M. & Skolnick, J. Predicting protein interactions of the kinase Lck critical to T cell modulation. *Structure* **32**, 2168–2179 (2024).
19. Janson, G., Valdes-Garcia, G., Heo, L. & Feig, M. Direct generation of protein conformational ensembles via machine learning. *Nat. Commun.* **14**, 774 (2023).
20. Mansoor, S., Baek, M., Park, H., Lee, G. R. & Baker, D. Protein ensemble generation through variational autoencoder latent space sampling. *J. Chem. Theory Comput.* **20**, 2689–2695 (2024).
21. Tian, H. et al. Explore protein conformational space with variational autoencoder. *Front. Mol. Biosci.* **8**, 781635 (2021).
22. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proc. International Conference on Machine Learning* (eds Bach, F. & Blei, D.) 2256–2265 (PMLR, 2015).
23. Wu, K. E. et al. Protein structure generation via folding diffusion. *Nat. Commun.* **15**, 1059 (2024).
24. Lu, J., Zhong, B., Zhang, Z. & Tang, J. Str2Str: a score-based framework for zero-shot protein conformation sampling. In *Proc. 12th International Conference on Learning Representations* 24678–24709 (ICLR, 2024).
25. Jing, B. et al. Eigenfold: generative protein structure prediction with diffusion models. In *Proc. ICLR 2023-Machine Learning for Drug Discovery Workshop* (OpenReview.net, 2023).
26. Wang, Y. et al. Protein conformation generation via force-guided SE(3) diffusion models. In *Proc. 41st International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) 56835–56859 (PMLR, 2024).
27. Zheng, S. et al. Predicting equilibrium distributions for molecular systems with deep learning. *Nat. Mach. Intell.* **6**, 558–567 (2024).
28. Lewis, S. et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **389**, eadv9817 (2025).
29. Jing, B., Berger, B. & Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. In *Proc. 41st International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) 22277–22303 (PMLR, 2024).
30. Janson, G. & Feig, M. Transferable deep generative modeling of intrinsically disordered protein conformations. *PLoS Comput. Biol.* **20**, e1012144 (2024).
31. Yim, J. et al. Se (3) diffusion model with application to protein backbone generation. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 40001–40039 (PMLR, 2023).
32. Guo, Z. et al. Diffusion models in bioinformatics and computational biology. *Nat. Rev. Bioeng.* **2**, 136–154 (2024).
33. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
34. Wu, R. et al. High-resolution de novo structure prediction from primary sequence. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.21.500999 (2022).
35. Chakravarty, D. et al. Alphafold predictions of fold-switched conformations are driven by structure memorization. *Nat. Commun.* **15**, 7296 (2024).
36. Boitreaud, J. et al. Chai-1: Decoding the molecular interactions of life. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.10.615955 (2024).
37. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10684–10695 (IEEE, 2022).
38. Arts, M. et al. Two for one: diffusion models and force fields for coarse-grained molecular dynamics. *J. Chem.Theory Comput.* **19**, 6151–6159 (2023).
39. Noé, F., Olsson, S., Köhler, J. & Wu, H. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
40. Bryant, P. & Noé, F. Structure prediction of alternative protein conformations. *Nat. Commun.* **15**, 7328 (2024).
41. Rodríguez-Espigares, I. et al. GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods* **17**, 777–787 (2020).
42. Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J.-C. & Galochkina, T. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res.* **52**, D384–D392 (2024).
43. Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
44. Naritomi, Y. & Fuchigami, S. Slow dynamics in protein fluctuations revealed by time-structure based independent component analysis: the case of domain motions. *J. Chem. Phys.* **134**, 065101 (2011).
45. Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).
46. Schwantes, C. R. & Pande, V. S. Improvements in markov state model construction reveal many non-native interactions in the folding of ntl9. *J. Chem. Theory Comput.* **9**, 2000–2009 (2013).
47. Mariani, V., Biasini, M., Barbato, A. & Schwede, T. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**, 2722–2728 (2013).
48. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
49. Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
50. McGibbon, R. T. et al. Mdtraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.* **109**, 1528–1532 (2015).
51. Wang, S., Sun, S., Li, Z., Zhang, R. & Xu, J. Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* **13**, e1005324 (2017).
52. Shaw, D. E. et al. Atomic-level characterization of the structural dynamics of proteins. *Science* **330**, 341–346 (2010).
53. Chakraborty, S., Venkatramani, R., Rao, B. J., Asgeirsson, B. & Dandekar, A. M. Protein structure quality assessment based on the distance profiles of consecutive backbone C_α atoms. *F1000Research* **2**, 211 (2013).
54. Ramachandran, G., Ramakrishnan, C. & Sasisekharan, V. Stereochemistry of polypeptide chain configurations. *J. Mol. Biol.* **7**, 95–99 (1963).
55. Passaro, S. et al. Boltz-2: towards accurate and efficient binding affinity prediction. Preprint at *bioRxiv* https://doi.org/10.1101/2025.06.14.659707 (2025).
56. Kim, G. et al. Easy and accurate protein structure prediction using ColabFold. *Nat. Protoc.* **20**, 620–642 (2025).

57. Whitford, P. C., Gosavi, S. & Onuchic, J. N. Conformational transitions in adenylate kinase: allosteric communication reduces misligation. *J. Biol. Chem.* **283**, 2042–2048 (2008).
58. Moriniere, J. et al. Cooperative binding of two acetylation marks on a histone tail by a single bromodomain. *Nature* **461**, 664–668 (2009).
59. Filippakopoulos, P. et al. Histone recognition and large-scale structural analysis of the human bromodomain family. *Cell* **149**, 214–231 (2012).
60. Inagaki, K., Satoh, T., Itoh, S. G., Okumura, H. & Kato, K. Redox-dependent conformational transition of catalytic domain of protein disulfide isomerase indicated by crystal structure-based molecular dynamics simulation. *Chem. Phys. Lett.* **618**, 203–207 (2015).
61. Wolters, J. C. et al. Ligand binding and crystal structures of the substrate-binding domain of the ABC transporter OpuA. *PLoS ONE* **5**, e10361 (2010).
62. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
63. Song, J., Meng, C. & Ermon, S. Denoising diffusion implicit models. In *Proc. International Conference on Learning Representations* (eds Hofmann, K. & Oh, A.) 14205–14224 (ICLR, 2021).
64. Noé, F. & Rosta, E. Markov models of molecular kinetics. *J. Chem. Phys.* **151**, 190401 (2019).
65. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
66. Wohlwend, J. et al. Boltz-1: democratizing biomolecular interaction modeling. Preprint at *bioRxiv* https://doi.org/10.1101/2024.11.19.624167 (2024).
67. Ingraham, J., Riesselman, A., Sander, C. & Marks, D. Learning protein structure with a differentiable simulator. In *Proc. International Conference on Learning Representations* (OpenReview.net, 2018).
68. Ingraham, J. B. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
69. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
70. Karras, T., Aittala, M., Aila, T. & Laine, S. Elucidating the design space of diffusion-based generative models. *Adv. Neural Inf. Process. Syst.* **35**, 26565–26577 (2022).
71. Berman, H. M. et al. The protein data bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
72. Webb, B. & Sali, A. Comparative protein structure modeling using Modeller. *Curr. Protoc. Bioinformatics* **54**, 5–6 (2016).
73. Olsson, M. H., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. Propka3: consistent treatment of internal and surface residues in empirical pka predictions. *J. Chem. Theory Comput.* **7**, 525–537 (2011).
74. Rodriguez-Espigares, Ismael. MD-protocol. *GitHub* https://github.com/GPCRmd/MD-protocol (2025).
75. Tubiana, T., Carvaillo, J.-C., Boulard, Y. & Bressanelli, S. Ttclust: a versatile molecular simulation trajectory clustering program with graphical summaries. *J. Chem. Inf. Model.* **58**, 2178–2182 (2018).
76. Wang, B. et al. Mac-Diff. *GitHub* https://github.com/Paulie-ai/Mac-Diff (2025).
77. Steinegger, M. & Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
78. Müller, C., Schlauderer, G., Reinstein, J. & Schulz, G. E. Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4**, 147–156 (1996).
79. Bryant, P. Structure prediction of alternative protein conformations. *Zenodo* https://doi.org/10.5281/zenodo.10837082 (2024).
80. Hoffmann, M. et al. Deeptime: a Python library for machine learning dynamical models from time series data. *Mach. Learn. Sci. Technol.* **3**, 015009 (2021).
81. Flamary, R. et al. Pot: Python optimal transport. *J. Mach. Learn. Res.* **22**, 1–8 (2021).
82. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**, 889–895 (2010).
83. Song, Y. et al. Score-based generative modeling through stochastic differential equations. In *Proc. International Conference on Learning Representations* 240–275 (OpenReview.net, 2021).
84. Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nat. Comput. Sci.* **3**, 382–392 (2023).
85. Song, Y., Garg, S., Shi, J. & Ermon, S. in *Uncertainty in Artificial Intelligence* 574–584 (PMLR, 2020).
86. Wang, Z. & Xu, J. Predicting protein contact map using evolutionary and physical constraints by integer programming. *Bioinformatics* **29**, i266–i273 (2013).
87. Rohl, C. A., Strauss, C. E., Misura, K. M. & Baker, D. in *Methods in Enzymology* Vol. 383, 66–93 (Elsevier, 2004).
88. Yang, J. et al. Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl Acad. Sci. USA* **117**, 1496–1503 (2020).
89. Wang, B. et al. Mac-diff v1.0.0: conditional diffusion with locality-aware modal alignment for generating diverse protein conformational ensembles. *Zenodo* https://doi.org/10.5281/zenodo.17936479 (2025).
90. Pettersen, E. F. et al. Ucsf chimerax: structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
91. Chaudhury, S., Lyskov, S. & Gray, J. J. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. *Bioinformatics* **26**, 689–691 (2010).

Acknowledgements

We thank C. Tian for the data collection and helpful discussions. This work was supported in part by the National Key Research and Development Program of China (grant number 2022YFC3400501); the National Natural Science Foundation of China (grant numbers 82425104 to H.L.; 62276099 to K.Z.).

Author contributions

K.Z. and H.L. conceived the concept and designed the workflow for this study. K.Z. and J.Z. designed the locality-aware modal alignment attention module. B.W., C.S., C.W. and J.C. designed the diffusion model. B.W., C.W. and J.C. implemented the Mac-Diff network architecture. B.W., C.W., J.C. and D.L. performed computational experiments. B.W. and D.L. prepared all data and contributed to the analysis and interpretation of results. B.W., K.Z., D.L., C.S. and J.Z. assisted with designing the experiments. K.Z., B.W., C.W., D.L., J.Z. and H.L. wrote the paper text and prepared figures and tables. All authors provided critical feedback, helped shape the research and analysis and revised the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-026-01198-9.

Correspondence and requests for materials should be addressed to Jie Zhang, Kai Zhang or Honglin Li.

Peer review information *Nature Machine Intelligence* thanks Haipeng Gong, Timothy Jenkins and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026