nature machine intelligence

Article

https://doi.org/10.1038/s42256-025-01059-x

Improving diffusion-based protein backbone generation with global-geometry-aware latent encoding

Received: 3 October 2024

Accepted: 12 May 2025

Published online: 18 June 2025

Check for updates

Yuyang Zhang[1,2,8], Yuhang Liu[3,4,8], Zinnia Ma[5,8], Min Li[6,7], Chunfu Xu[3,4] & Haipeng Gong[1,2]

The global structural properties of a protein, such as shape, fold and topology, strongly affect its function. Although recent breakthroughs in diffusion-based generative models have greatly advanced de novo protein design, particularly in generating diverse and realistic structures, it remains challenging to design proteins of specific geometries without residue-level control over the topological details. A more practical, top-down approach is needed for prescribing the overall geometric arrangements of secondary structure elements in the generated protein structures. In response, we propose TopoDiff, an unsupervised framework that learns and exploits a global-geometry-aware latent representation, enabling both unconditional and controllable diffusion-based protein generation. Trained on the Protein Data Bank and CATH datasets, the structure encoder embeds protein global geometries into a 32-dimensional latent space, from which latent codes sampled by the latent sampler serve as informative conditions for the diffusion-based backbone decoder. In benchmarks against existing baselines, TopoDiff demonstrates comparable performance on established metrics including designability, diversity and novelty, as well as markedly improves coverage over the fold types of natural proteins in the CATH dataset. Moreover, latent conditioning enables versatile manipulations at the global-geometry level to control the generated protein structures, through which we derived a number of novel folds of mainly beta proteins with comprehensive experimental validation.

De novo protein design is an intriguing and expanding field of research with the potential to venture into uncharted fold space, offering limitless opportunities for tailoring proteins to novel applications, including biomedical therapeutics[1–3], catalytic enhancement[4] and the development of innovative biological circuits[5,6]. Despite its vast potential, de novo protein design has long been recognized as a challenging task, due to the highly structured nature of protein data and the stringent requirements on geometric restraints[7].

Recent advances in diffusion models have substantially reshaped the field with their superior ability to generate novel, diverse and physically plausible structures. Although early efforts still relied on one-dimensional or two-dimensional (2D) protein representations[8–10], subsequent works tended to leverage the success in protein structure prediction tasks[11,12], building equivariant networks to directly learn physical priors in the Cartesian space[13–17].

Despite encouraging progress, several issues remain to be addressed. Although lacking systematic evaluation, there is evidence that some models, although trained on unbiased datasets such as the Protein Data Bank (PDB)[18] or CATH[19,20], struggle to generate protein backbones of certain fold classes[16]. This issue is evident when reviewing

A full list of affiliations appears at the end of the paper. e-mail: xuchunfu@nibs.ac.cn; hgong@tsinghua.edu.cn

https://doi.org/10.1038/s42256-025-01059-x

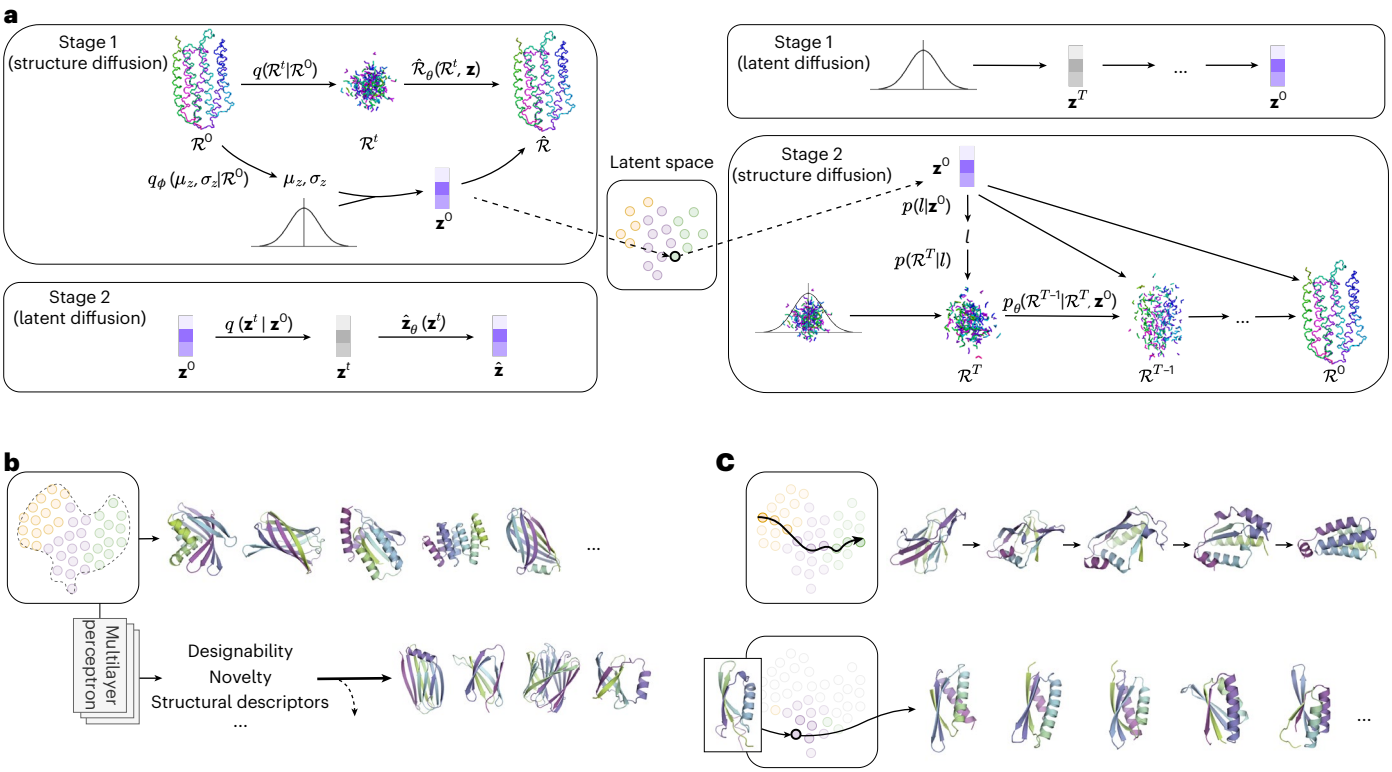


Fig. 1 | Overview of the work. a, Overall framework of TopoDiff. The training (left) and sampling (right) processes consist of two stages. In stage 1 of the training process, each protein structure is converted by the encoder module into a fixed-size, low-dimensional latent code **z** that captures the global geometry, whereas the perturbed structure $\mathcal{R}^t$ is sampled following the designed forward marginal distribution. The diffusion decoder is trained to predict the ground-truth structure $\mathcal{R}^0$ from $\mathcal{R}^t$ and **z**. In stage 2, the training focuses on learning the diffusion process in the latent representation space. During inference, a latent code is first sampled from the latent diffusion module, and then the structure decoder is used to generate structures conditioned on the sampled latent. **b**, Through unbiased sampling in the compact, continuous latent space, TopoDiff learns to generate samples with enhanced coverage of the natural fold space (top row). Otherwise, TopoDiff could be targeted to produce structures with the desired properties by reweighted latent sampling under the assistance of latent-based property predictors (bottom row). The example here shows a particular application for generating novel, designable mainly beta proteins. **c**, TopoDiff is capable of generating structures from a local distribution in the latent space. The two example applications are the simulation of a gradual transition between two structures through latent-based linear interpolation (top row) and the creation of structures with similar global geometry to a given reference through latent perturbation (bottom row).

all experimentally validated proteins generated by structure-based diffusion models so far, since they predominantly fall in the mainly alpha or alpha-and-beta classes. Furthermore, we note that the current widely used metrics, namely, designability, novelty and diversity, provide no indications of the extent to which the natural protein space has been covered. This gap further hinders the understanding and resolution of these issues. To improve the coverage of the generated samples over specific protein folds, previous works have used residue-level one-dimensional or 2D fold conditioning along with additional fine tuning to generate immunoglobulin domains with varied loop regions[21], or applied classifier guidance by training classifiers on specific protein classes[22]. Although these approaches are indeed capable of enhancing coverage for a particular group of proteins, their feasibility strictly depends on the clear and distinctive definition of this group as well as the presence of sufficient group-labelled training samples, which are the prerequisites for applying the finer-grained topology as conditioning, training a robust classifier to guide the gradient or model fine-tuning on specific data subsets for additional refinement. However, due to the limited and unbalanced amount of available annotations as well as the discreteness and subjectivity in annotation assignment[23,24], it is often impractical to apply the same strategy to achieve unbiased visiting of fold modes in the training set and effective expansion of the existing protein fold space simultaneously.

In this work, we focus on an important and general unsupervised problem setting: how to train the diffusion model to capture the underlying data distribution in an arbitrary dataset without the explicit requirement of annotations or prior understanding? We propose the framework of TopoDiff as a solution. First, we concurrently trained the diffusion-based structure generative model with a structure encoder of the encoder–decoder architecture. The encoder learns a fixed-size, continuous latent space that captures the high-level global geometry of proteins, whereas the generative module operates at the residue level for controllable sampling conditioned on the predefined latent encoding. Next, we trained a simple latent diffusion model to unbiasedly sample protein global geometries from this learned latent distribution and used the sampled latent to guide subsequent atomic-level protein structure sampling. This scheme not only effectively enhances the coverage of protein fold types in the dataset but also opens up a dimension for controllable generation. We then defined a coverage metric and conducted systematic evaluations on TopoDiff and other state-of-the-art models using this and other established metrics. Finally, we performed biological experiments to demonstrate the effectiveness of our method in improving the sampling of proteins of mainly beta topologies, a large class that is prevalent in nature but remains under-represented in prior de novo protein design efforts.

Results

Overview of TopoDiff

The overall framework of TopoDiff is illustrated in Fig. 1. The global structural properties of a protein, such as shape, fold and topology,

https://doi.org/10.1038/s42256-025-01059-x

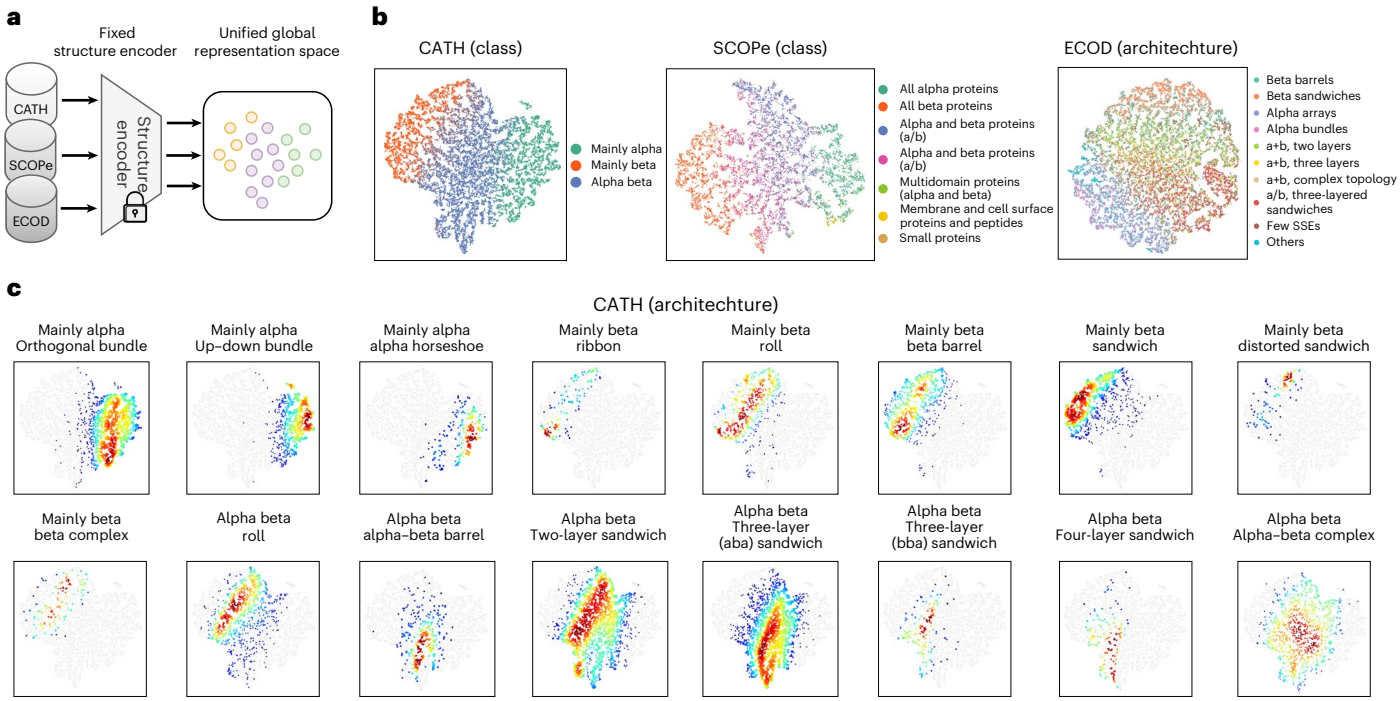


Fig. 2 | Analysis of TopoDiff's learned latent representations. a, Once trained, the fixed structure encoder maps the structures from different sources into a unified global representation space. **b**, Visualization of the latent space coloured by top-level classification hierarchies: CATH (class), SCOPe (class) and ECOD (architecture). Each point represents a structure, coloured according to its annotation. **c**, Kernel density estimation of specific CATH architectures within the latent space. Each subplot shows the density of structures within a particular architecture, illustrating its distribution across the latent space.

are closely related to its function[25,26] and dynamics[27,28], as well as lay the foundation for achieving a heuristic understanding of its molecular mechanisms. Conversely, although powerful, classic diffusion models constrain dimensionality to be proportional to the input, frustrating efforts to learn meaningful, compressed latent representations of global protein geometry[29]. With this in mind, the core focus of our framework is on the establishment, investigation and utilization of a fixed-size, low-dimensional latent representation that encodes the global structural characteristics of all proteins, building on recent advances in diffusion-based models. To achieve this aim, both training and sampling processes are divided into two stages. In stage 1 of the training process (Fig. 1a, left), we adopt a diffusion–variational autoencoder (VAE) formulation to harness the superior generative capability of diffusion models alongside the representational power of VAEs with a compressed and continuous latent space, thereby combining the inherent strengths of both architectures in a unified training process. As essential components of our framework, the resulting structure encoder and conditional diffusion decoder share a common fixed-dimensional latent space, which provides the protein-level encoding **z** for the global geometry. In stage 2, a latent diffusion module is trained to model the latent distribution, enabling unbiased sampling from this otherwise intractable space. During the sampling phase (Fig. 1a, right), the latent diffusion model first samples from the latent space, and a structure diffusion process is then performed for protein structure generation in the Cartesian space by conditioning on the sampled latent. The architecture design of TopoDiff enables multiple brand-new controllable sampling schemes (Fig. 1b,c).

Learned latent representation of the fold space

To gain an understanding of the latent space learned by TopoDiff, we first encoded all the structures in our CATH-60 training dataset into the 32-dimensional latent space and then applied *t*-distributed stochastic neighbour embedding (*t*-SNE)[30] for dimensionality reduction. As shown in Fig. 2b, these codes collectively form a compact and continuous manifold. In particular, even though no structure annotations were used during training, the resulting clusters perfectly coincide with the human curation of CATH class annotations, with each class clearly separable from the others even in 2D embedding. Moreover, we find that each CATH architecture cluster indeed exhibits a distinct spatial distribution (Fig. 2c and Supplementary Fig. 4). Many other intrinsic attributes of proteins, such as the secondary structure composition, chain length and radius of gyration, also display structured global or local distribution patterns within the manifold (Supplementary Fig. 3). These observations demonstrate that the model learns to partition over the training dataset in an unsupervised and highly interpretable manner.

Next, we tested the generalizability of this encoding method on unseen data. Specifically, we applied the same encoder trained on CATH to two other hierarchical structural classification datasets, namely, SCOPe[31,32] and ECOD[33] (Fig. 2a), which present evident distinctions in structural coverage and domain boundary definitions[34–36] due to their discrepant classification schemes[35,37]. Interestingly, the latent manifolds produced by the encoder align closely with the annotations of top-level hierarchies in both SCOPe and ECOD datasets (Fig. 2b and Supplementary Figs. 5 and 6). Consistent with previous analyses[34–36], the distinctive hierarchical organizations across different classification systems may essentially reflect different domain partitioning and class discretization within a common structure space. By learning a continuous and unsupervised representation, our method effectively bypasses these annotation inconsistencies across datasets.

The continuous, global-level latent space also offers an alternative view of the protein fold space compared with the established hierarchical, discrete organizations. In Supplementary Results 2, we provide a more thorough discussion of its potential advantages, focusing on revealing the continuous relationships between fold classes and identifying potentially inconsistent or ambiguous annotations.

https://doi.org/10.1038/s42256-025-01059-x

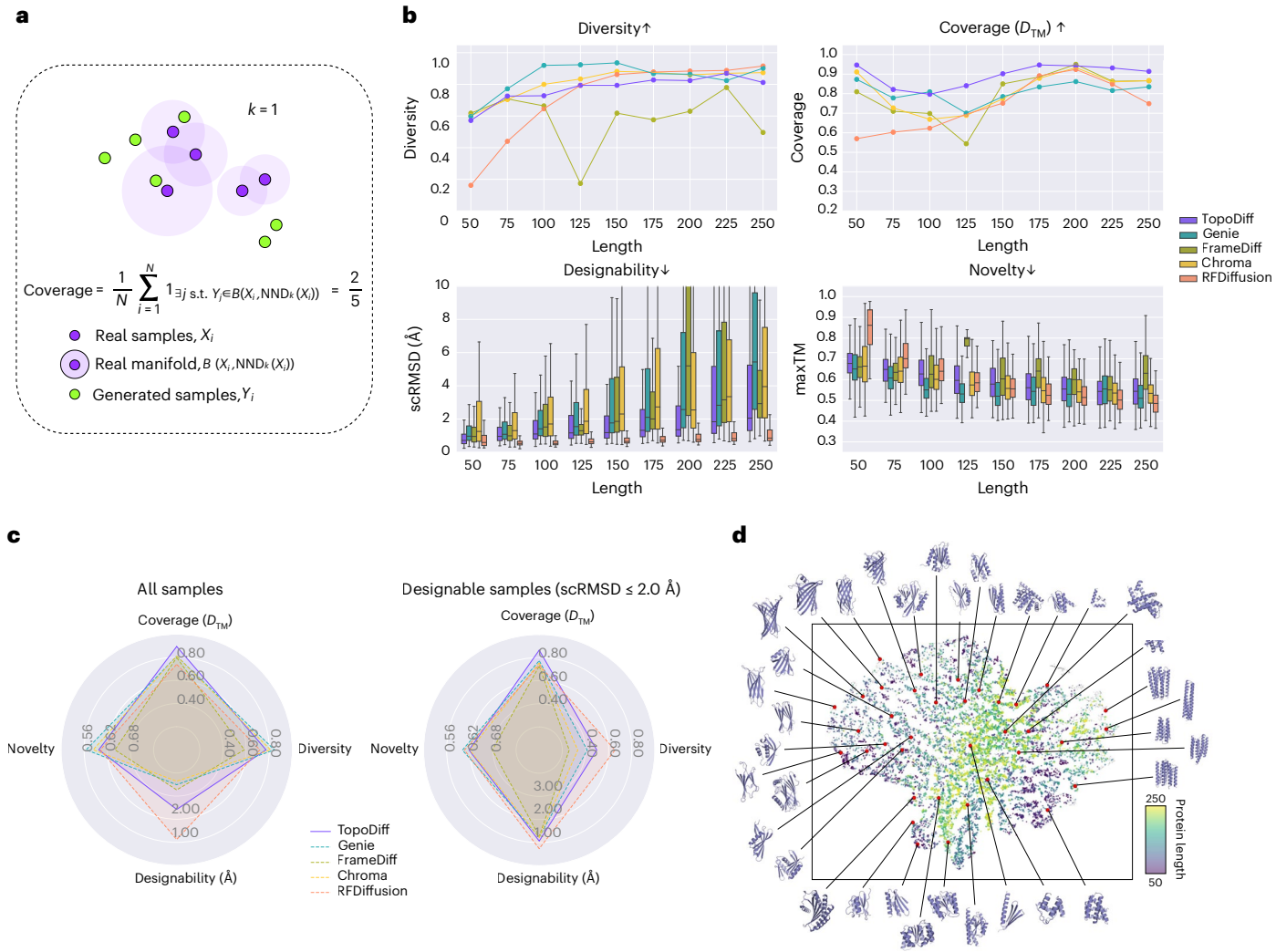


Fig. 3 | Evaluation of TopoDiff's generative performance for unconditional sampling. a, Illustration of the coverage metric used to quantify the extent to which the generated samples cover the natural protein fold space. Real samples X_i are compared with generated samples Y_j using a KNN approach to determine if each real sample is covered by the generated samples within a defined distance threshold. The coverage score is calculated as the proportion of real samples that have at least one generated neighbour within this threshold. **b**, Quantitative analysis of the generative performance across different models, measured by diversity, coverage, designability and novelty at varying protein lengths. The arrows besides the metrics indicate the desired direction of improvement. For novelty and designability metrics, box plots are used to depict the data distribution (n = 500 samples per length), showing median and quartiles, with whiskers extending to 1.5 times the interquartile range. **c**, Radar plots summarizing the average performance of each model on different metrics when considering all the generated samples (left) and only high-quality samples (right; scRMSD ≤ 2 Å). Each metric is averaged across all the sampled lengths. **d**, Projection of the sampled latent codes on the t-SNE dimension-reduced space of the CATH dataset. A total of 12,613 latent codes, corresponding to sampled structures with scRMSD ≤ 2 Å and maxTM < 0.7, are projected onto the original t-SNE space constructed from the CATH dataset (Fig. 2). Each coloured point represents a sampled latent code, with the colour indicating the sampled length (~50–250, from purple to green). The representative structures generated from the latent codes at different regions of the latent space are also visualized alongside the scatter plot, providing a visual inspection of the latent structure relationship.

Benchmark testing on unconditional sampling

We evaluate the performance of TopoDiff in unconditional sampling against several state-of-the-art diffusion-based generative models, including Genie[16], FrameDiff[15], Chroma[22] and RFDiffusion[17]. To ensure a comprehensive and robust evaluation across a wide range of protein sizes, for each model under evaluation, we randomly generated 500 samples at each fixed length in {50, 75, 100, 125, 150, 175, 200, 225, 250}, a series uniformly spanning the length range of our training data. On the basis of our experience, such a scheme can reveal the length-dependent behaviour in a model, which might be otherwise overlooked.

We notice that established metrics are dedicated to quantify either per-sample quality (for example, designability and novelty) or intrasample diversity (for example, diversity), but provide no information of how much the known fold space is covered by the generated samples, an indicator that quantitatively describes a model's capacity for unbiased sampling across existing data. Evidence from other fields shows that ignoring this metric will introduce selection bias towards models that sacrifice variation in favour of high-quality samples from a truncated subset of the sampling space[38]. Indeed, over the past decade, de novo protein design has been largely confined to alpha helix bundles and alpha–beta sandwiches[39,40], and diffusion-based models have not provided an immediate remedy to this biased trend, as experimental validations and applications still predominantly focus on these architectures[17,22,41]. To address this limitation, we adopted a coverage metric[42] to quantify the proportion of natural protein folds covered by samples produced from a generative model (Fig. 3a; Methods provides the definition and detailed implementation).

The first row in Fig. 3b presents an evaluation of diversity and coverage. TopoDiff is comparable with the other methods with regard to diversity, but prevails over all of them with regard to coverage. RFDiffusion shows notable deficiencies in both metrics for short proteins in the length of [50, 150], a range covering over 60% of natural domains in the CATH database. Similarly, FrameDiff also exhibits some degree of length-dependent fluctuations. Genie excels at generating highly diverse samples, but with a slightly lower coverage. To further investigate the specific advantage of TopoDiff over the others in covering natural fold types, we analysed the sample-wise binary coverage indicators and found that our model could cover a substantially larger number of mainly beta-strand fold classes, a group of topologies that other methods typically under-represent (Supplementary Fig. 13).

The second row in Fig. 3b shifts to the evaluation of designability and novelty. Regarding designability (in terms of self-consistent root mean square distance (scRMSD) as defined in the Methods), TopoDiff demonstrates advantages over most models excluding RFDiffusion (which has a significantly larger parameter size (Supplementary Table 9)), across the entire sampled length range. For novelty (in terms of maxTM as defined in the Methods), TopoDiff shows a steady intermediate value along the chain lengths, indicating a fair balance between the coverage of known folds and the generalizability to novel ones.

We further compare the model performance by averaging metrics across all the sampled lengths (Fig. 3c), where the metrics are calculated using all the samples (left) and using samples with high designability (that is, scRMSD ≤ 2 Å) only (right). TopoDiff undoubtedly improves the overall coverage in both cases, indicating that this improvement truly arises from designable samples. As mentioned in the Methods, the computation of coverage relies on the definition of distance between a pair of structures (equation (5)). To supplement the evaluation in Fig. 3b,c in which the average TM-score[43] between the query and target structures is uniformly used for the distance definition (equation (6)), we also present results with distances computed by a third-party model[44] (equation (7)) in Supplementary Fig. 12. Clearly, the overall relative trend among the tested methods is preserved and the advantage of TopoDiff is consistent, indicating the robustness of our coverage computation on the choice of distance definition. Overall, despite a slight inferiority to RFDiffusion, particularly at longer lengths, the designability of TopoDiff surpasses the other methods. Furthermore, TopoDiff generates samples at least three times faster than RFDiffusion (Supplementary Table 9), which enables the production of more diverse and designable backbones within the same time. In addition, we also evaluated TopoDiff against several recently proposed methods, including FoldFlow[45], Genie2 (ref. 46) and FoldFlow2 (ref. 47). Detailed results and thorough discussions are provided in Supplementary Results 4.

Finally, we seek to further explore the sampling space of TopoDiff. From the 22,500 latent structure pairs generated in this benchmark, we selected 12,613 with high designability and novelty (scRMSD ≤ 2 Å and maxTM < 0.7), and projected these latents onto the CATH-encoded *t*-SNE subspace (Fig. 3d). To investigate the relationship between the latent codes and generated structures, we subsampled the latent codes spanning the manifold and displayed their corresponding structures around the scatter plot. As expected, these codes spread over the *t*-SNE-reduced manifold, providing a strong foundation for the enhanced coverage of natural fold space by the sampled structures. A closer examination of the sampled structures reveals that the spatial arrangement of secondary structure elements (SSEs) is highly correlated with the position of the conditioning latent code, closely reflecting the original distribution of CATH training samples (Fig. 2b,c). This observation demonstrates that concurrent training not only prompts the encoder to capture global geometric information in the latent codes but also allows the diffusion decoder to leverage this information to generate structures with matching spatial features.

Controllable generation with the learned latent space

In Fig. 4a, we illustrate how different modules developed in this work can be integrated to enhance the controllability of structure sampling. The latent code could be sampled from the latent sampler shown in the previous section or alternatively harvested directly from the encoder's posterior based on a specific input structure. Moreover, additional latent classifiers can be trained to predict properties of interest and used to tune the sampled latent distribution through classifier guidance or rejection sampling. Once a latent is selected, structure sampling can be conditioned on this code and optionally on additional residue-level information, achieving simultaneous constraining over the global geometry and localized atomic details.

Indeed, many measurable properties of the generated samples, like the proportion of SSEs, novelty and designability, exhibit distinct spatial patterns in the latent space, reflecting their inherent correlation with the global protein geometry. Hence, we explored the possibility of tuning the model performance by reweighting the latent sampling regions using pretrained classifiers that predict the desired properties. In particular, this strategy effectively enables the fine tuning of the model performance by simple manipulation at the low-dimensional latent level, which, unlike exhaustive sampling at the atomic level in Cartesian space, consumes negligible additional time. We focused on the trade-off between designability and novelty via rejection sampling (Methods provides the implementation details) and created three model variants: a designability-prior variant (using a designability classifier), a novelty-prior variant (using a novelty classifier) and an all-round variant (using both classifiers). As expected, the designability-prior variant improves the designability at the expense of novelty, whereas the novelty-prior variant shows the opposite trend (Fig. 4b and Supplementary Fig. 17). Interestingly, the all-round variant achieves balanced improvement over the base model, with novelty and diversity significantly increased whereas designability and coverage sustained.

In addition to general unconditional sampling in the latent space, latent codes could also be sampled from any local region of interest. We first demonstrate structure generation around a query structure, based on latent sampling from a local distribution around its latent encoding. We selected representative query proteins with a wide variety of architectures from the training dataset and randomly sampled five structures without further cherry-picking for each of them. On the basis of a side-by-side comparison (Fig. 4c and Supplementary Fig. 18), the generated structures, in general, share a very similar SSE spatial arrangement to the query one with occasionally improved proportions of regular secondary structures, but always present a considerable level of diversity in the exact connectivity and topology.

Subsequently, we present controlled structure generation based on interpolation between pairs of query latent codes. In this experiment, we collected a variety of latent pairs distantly located on the latent manifold and linearly interpolated ten sequential intermediate latent codes between each pair. Figure 4d shows the variations of generated structures along different latent trajectories, with each row representing a distinctive latent pair. The second row presents the gradual transition from a mainly beta roll to an all-alpha helix bundle, where adjacent structures share a certain degree of similarity (TM-score > 0.45), indicating an overall smooth transition process, although the two terminal structures have completely different SSE spatial arrangements (TM-score = 0.27). Similarly, the first row illustrates the interpolation from a structure of orthogonal helices to a beta sandwich. More examples are provided in Supplementary Fig. 19.

Moreover, since the latent encoding provides a coarse control over the global geometry, it is possible to incorporate additional residue-level conditions to impose control at a finer level of granularity. We demonstrate this capacity with a motif scaffolding experiment. Specifically, we selected three representative motifs from the RFDiffusion study[17], including a single helix, a beta hairpin, and a mixed pair

https://doi.org/10.1038/s42256-025-01059-x

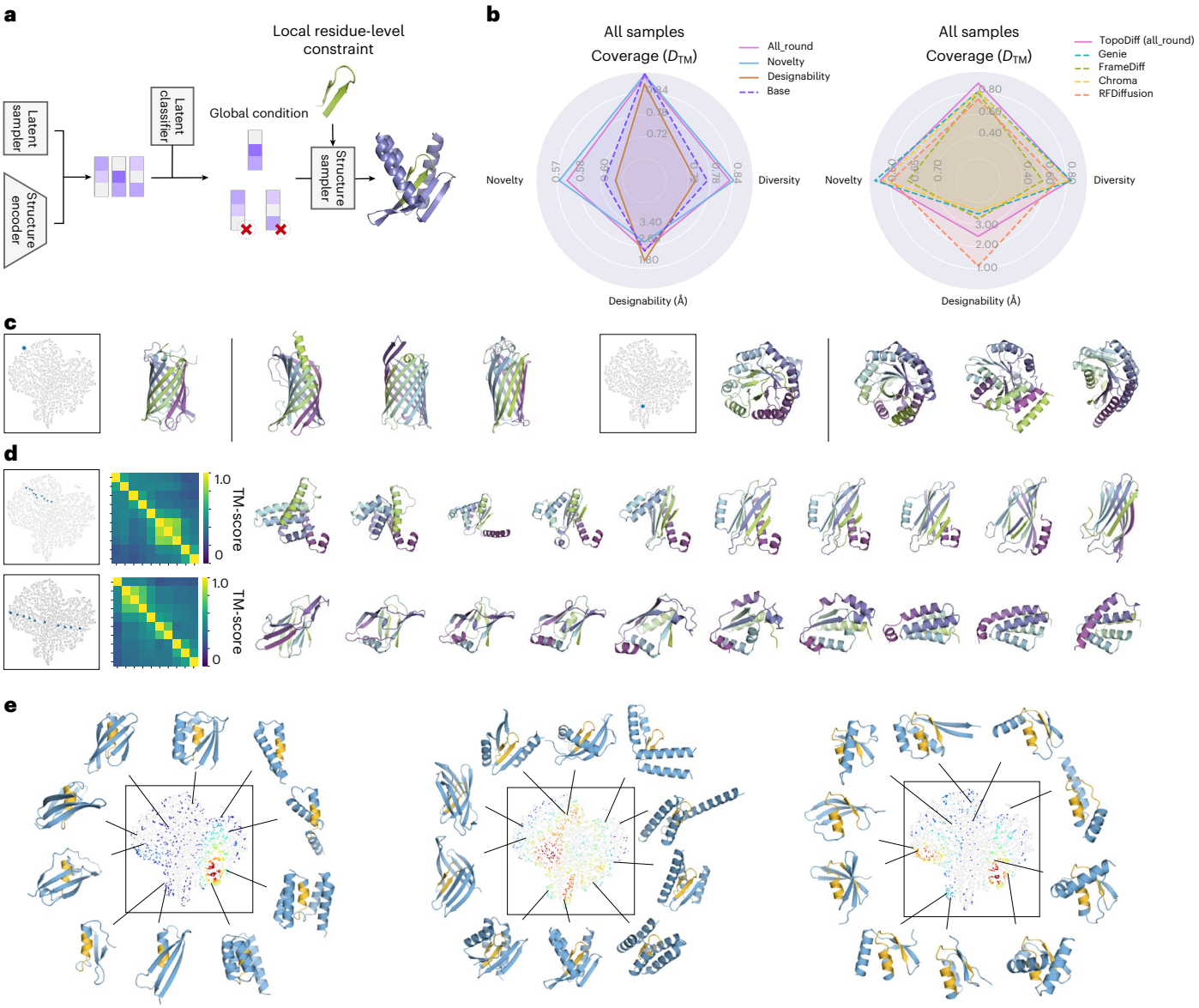


Fig. 4 | Exploring controllable protein structure generation with TopoDiff. **a**, Overview of the controllable generation process. The latent code could be generated using the latent diffusion module or sampled from a local distribution. Downstream structure sampling is controlled by the acceptance probability of the sampled latent codes (equation (9)). The accepted latent code, acting as a global condition, can be combined with local residue-level constraints to generate structures that meet specific design criteria. **b**, Radar plots comparing the performance of different model variants derived from the latent-space rejection sampling: base, novelty-prioritized, designability-prioritized and all-round. Left: three model variants are compared with the base model. Right: all-round model variant is compared with other baseline methods. All the metrics are computed in the same way as that in Fig. 3c. **c**, Examples of structures generated by resampling based on the latent code of a reference structure. The leftmost structure represents the reference structure selected from the CATH dataset, with the right one showing randomly generated structures without cherry-picking. **d**, Visualization of structures generated by linear interpolation between two latent codes in the latent space. Each row depicts a different interpolation process, with the projected latent trajectory on the latent space (left), a TM-score distance matrix showing pairwise similarity between the sampled structures (middle) and a visualization of these sampled structures (right). **e**, Motif scaffolding experiment. For each motif case, latent codes corresponding to successful designs are plotted on the central scatter plot, with each point coloured based on the kernel density estimation of successful designs in the local vicinity. Representative structures sampled from different regions of the latent space are shown around the scatter plot, with the query residue-level motifs highlighted in yellow and the rest are coloured blue.

of helix and strand. In this experiment, we first unconditionally sampled a series of latent codes that fit into the designed length, and then applied these codes and the specified motifs as the joint conditioning to generate the structures. Figure 4e presents the density plots of all the successful designs (that is, scRMSD ≤ 2 Å and motif scRMSD ≤ 1 Å) on the latent manifold, alongside representative sampled structures from different regions. In each case, the successful designs exhibit a non-uniform distribution across the latent manifold, achieving high populations in regions in which the latent information and the motif are coherent. When structures inferred by the semantic latent conflict with the motif itself, the model seeks to find a compromised solution, although with increased difficulty. Taking the two-strand 4ZYP motif (Fig. 4e, middle) as an example, even in the mainly alpha region of the manifold, we occasionally sampled successful designs in which the beta hairpin is encompassed by helices achieving various global shapes. Hence, when combined with local constraints, the latent code acts as a global prompt to guide the model in exploring various architectures and topologies rather than always sampling from the preferred regions.

https://doi.org/10.1038/s42256-025-01059-x

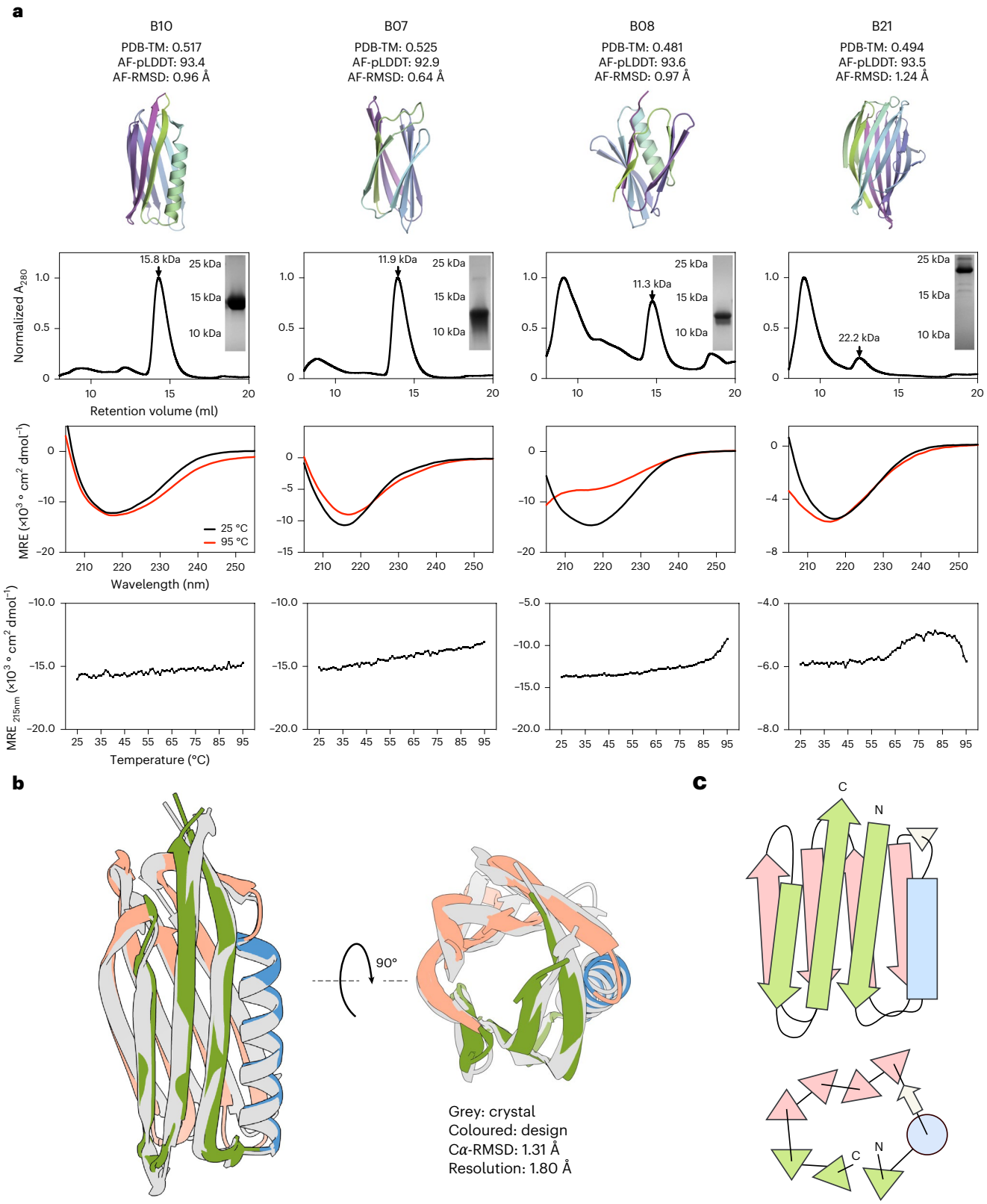


Fig. 5 | Experimental validation of novel mainly beta protein designs. **a**, Experimental characterizations of the selected candidates (B10, B07, B08 and B21). The first row displays the design models with their corresponding metrics: PDB-TM (maximal TM-score to PDB), AF-RMSD (scRMSD compared with the best model of AlphaFold2) and AF-pLDDT (pLDDT of the best model of AlphaFold2). The second row shows the SEC profiles (with expected monomeric peaks indicated by arrows) and the monomer band observed on the sodium dodecyl sulfate–polyacrylamide gel electrophoresis gel. The third row presents the CD spectra of the purified proteins at 25 °C (black) and 95 °C (red), demonstrating secondary structure content and thermostability. The fourth row displays the temperature dependence of the CD signal at 215 nm. No unfolding transition is observed at temperatures up to 95 °C for B10 and B07, suggesting their high thermostability. MRE, mean residue ellipticity. **b**, X-ray crystal structure (grey) of design B10 matches closely with the backbone structure (coloured) generated by TopoDiff (RMSD, 1.31 Å). **c**, Topological diagrams of design B10 in the side and top views.

Experimental validation of generated mainly beta proteins

We seek to test our TopoDiff model in the real-world design scenarios, focusing on the discovery of novel mainly beta proteins, a large class of proteins that are commonly found in nature but remain under-represented in de novo designed proteins (Extended Data Fig. 1 shows the statistical analysis on the Protein Design Archive database[48] and Supplementary Results 6 provides a thorough discussion on the scientific significance of this task). Following step-wise filtering described in the Methods, we finally obtained 403 backbones with 950 sequence designs, from which we further selected three designs per sampled length, resulting in 21 candidates for subsequent experimental validation. As shown in Supplementary Fig. 20, these designs are highly diverse in beta-strand arrangements and exhibit notable novelty compared with known structures. All the designs feature >50% of residues in beta strands and <20% in alpha helices, with more than half of the designs consisting exclusively of beta strands and coils. Moreover, the global packing of these designs predominantly relies on the formation of beta sheets with numerous non-local interactions, making manual blueprint design exceptionally challenging. Regarding designability, all the designs are predicted to be sufficiently foldable, as assessed by ESMFold and AlphaFold2. In particular, 16 out of the 21 designs have five AlphaFold2 models achieving predicted local distance difference test (pLDDT) > 85% and scRMSD < 1.75 Å, which are considerably strong indicators of successful design[49].

Following in silico selection, we obtained synthetic genes encoding the 21 selected designs for subsequent wet-laboratory experimental validation. Nine of these designs present a soluble expression in *Escherichia coli* (Supplementary Fig. 21), allowing for efficient follow-up purification by nickel-affinity chromatography and size exclusion chromatography (SEC; Supplementary Fig. 22a). The SEC profiles of designs B07 and B10 exhibit distinct monomer peaks, whereas the others form a mixture of soluble aggregates and monomers (Fig. 5a and Supplementary Fig. 22b). Six of the nine expressed designs display the anticipated circular dichroism (CD) spectra for beta-sheet-rich proteins (Fig. 5a and Supplementary Fig. 22c). A summary of these experimental results is provided in Supplementary Table 11.

Here we highlight four designs that attain separable monomeric states and the correct CD spectra (Fig. 5a). Specifically, B07 and B10 demonstrate high thermostability up to 95 °C, whereas B08 and B21 also exhibit sufficient resistance to heat denaturation, with melting temperatures of approximately 80 °C and 65 °C, respectively. Moreover, all of them have the PDB-TM values (that is, the maximum TM-score to PDB structures) lower than or close to 0.5, supporting their novelty in topology. We further determined the structure of B10 with X-ray crystallography (Supplementary Fig. 23 and Supplementary Table 12). The solved structure closely matches the original backbone generated by TopoDiff (Fig. 5b), with a C_α-root mean square distance (RMSD) of 1.31 Å. This 125-residue mainly beta protein is composed of eight beta strands and one alpha helix. One novel structural feature is the packing of the alpha helix into the crossover region between two beta sheets (Fig. 5b), where the helix extends outwards and pushes the adjacent beta strands of the two sheets apart, creating a unique triangular geometric arrangement in the top view (Fig. 5c). Interestingly, this compact triangular arrangement is unseen in natural proteins, with the closest structures in PDB being either beta barrels or two-layer sandwiches (Supplementary Fig. 20). Structural analysis of the other monomeric designs is provided in Supplementary Results 7.6.

Discussion

In this work, we propose an unsupervised framework that builds on the current state-of-the-art diffusion generative model, enabling the concurrent learning of an encoder to capture a low-dimensional global structural representation and a conditional diffusion module to leverage this information for controllable generation. Noticeably, different from methods[50,51] that engage length-dependent latent diffusion to facilitate protein structure generation, the latent representation in TopoDiff is set as fixed dimensional (analogous to a branch of computer vision models[52–54]), which enables the sampling and manipulation of various protein global geometries in a universal latent space. Hence, the introduction of this fixed-size latent encoding into the formulation of diffusion-based protein structure generation not only facilitates human interpretation/understanding of the data distribution and generation process but also improves the coverage of the protein fold space and keeps the other performance metrics competitive. Moreover, through the VAE architecture, the latent space is confined to a low dimensionality with strong continuity, enforcing a coarse constraint over the global geometry without hindering the discovery of novel folds during the generation process. The effectiveness of our unique model design has been validated through ablation studies (Supplementary Results 8). On the basis of this design, we also propose a number of brand-new, versatile control schemes for protein structure generation through simple latent-level manipulation. The latent-level control provides a useful supplement to established residue-level conditions like the residue-wise SSE and pairwise adjacency[13,17], which, although valuable, require substantial domain expertise and supposedly limit the sampling space. Eventually, when applied to a widely recognized challenging design task, the design of mainly beta proteins with novel backbone topologies, our approach allows the diffusion-based generation of mainly beta or even full-beta novel proteins that have been validated by firm experimental evidence, without relying on any human predesign.

Considering the broad application of small, single-domain proteins in nowadays practical protein design and engineering[4,55–60], the current version of TopoDiff focuses on the structure generation of proteins with a length of ≤256 residues. However, due to its unsupervised nature, the whole framework is inherently generalizable to longer proteins, potentially by enlarging the parameter capacity in networks. In Supplementary Results 9, we show some preliminary explorations on the incorporation of the flow-matching technique into the diffusion formula, as well as the scalability of this model on parameter size. Alternatively, our framework can be customized for the user-defined categories of proteins, allowing for the learning of class-specific representations alongside a specialized generative model. On the other hand, unlike mainstream design methods and our method, which generate backbone coordinates and protein sequence in a step-wise manner, all-atom protein generation has been proposed to further improve the sequence-structure consistency for the protein under design[61,62]. Although immature, these prior works indicate the possible direction of future improvement for our method.

Methods

Overall model design

In alignment with previous studies[13–17], we represent the protein as a list of rigid transformations (residue clouds) in the $SE(3)^N$ space. Briefly, for a sequence of length l, each residue is parameterized as the collection of the translation of its C_α atom, denoted as $\mathbf{x}_i \in \mathbb{R}^3$, and its orientation, uniquely defined by the coordinates of three backbone atoms (C, C_α, N) and denoted as $\mathbf{r}_i \in SO(3)$. Collectively, we denote the whole sequence as $\mathcal{R} = \{(\mathbf{x}_i, \mathbf{r}_i)\} \in SE(3)^l$.

A distinctive step we took as compared with previous works[13–17] is the introduction of a protein-level latent variable $\mathbf{z} \in \mathbb{R}^{C_z}$ with the fixed-size dimensionality of $C_\mathbf{z} = 32$, which encodes the essential information about the global geometry of the underlying structure. Unlike the length-dependent representation of protein structures adopted in previous works, the fixed-dimensional representation is independent of protein size and, thus, allows the mapping of all the protein structures into a uniform latent space. The continuous nature of this latent space further supports the diffusion-based latent sampling among different protein topologies. In particular, in this work, we intentionally restrict this latent space to a low dimensionality,

aiming to identify the principal degrees of freedom for the description of protein global geometry. By this means, the latent-level manipulations only impose coarse controls over the protein global geometry without enforcing strong restrictions on the atomic-level local structures.

We further model the joint distribution $p(\mathcal{R}, \mathbf{z})$ in a hierarchical fashion. With this formulation, we essentially decompose the generation process into two steps, namely, the acquisition of $\mathbf{z}$ and the subsequent structure-space generation process conditioned on it:

$$p(\mathcal{R}, \mathbf{z}) = p(\mathbf{z})\, p(\mathcal{R}|\mathbf{z}). \tag{1}$$

The conditional generation in the structure space is further modelled as a diffusion-based process, through which our model is trained to learn the complex data distribution (Supplementary Fig. 1 shows the model architecture). In brief, we define the forward process of the diffusion model as a non-learnable Markovian process that gradually introduces noises to a protein structure $\mathcal{R}^0$ towards a predefined prior distribution p_T within T steps:

$$q(\mathcal{R}^{1:T}|\mathcal{R}^0) = \prod_{t=1}^{T} q(\mathcal{R}^t|\mathcal{R}^{t-1}), \tag{2}$$

whereas the reverse process is also a Markovian process that learns to remove the noise signal conditioned on the latent variable $\mathbf{z}$ as

$$p_{\boldsymbol{\theta}}(\mathcal{R}^{0:T}|\mathbf{z}) = p_T(\mathcal{R}^T) \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\mathcal{R}^{t-1}|\mathcal{R}^t, \mathbf{z}), \tag{3}$$

where q and p are the probability distributions for the forward and reverse processes, respectively, with θ denoting the model parameters. The training objective is to make an accurate prediction $\hat{\mathcal{R}}_\theta$ for the ground truth $\mathcal{R}^0$, which is, for simplicity, denoted here as the reconstruction loss $\mathcal{L}_{\text{recon}_{\boldsymbol{\theta}}}(\mathcal{R}^0, \hat{\mathcal{R}}_\theta(\mathcal{R}^t, t, \mathbf{z}))$.

Although the current protein structure classification systems[19,20,32,63] partially align with our definition and expectation of $\mathbf{z}$, several notable drawbacks preclude them as optimal representations. First, all classification systems implicitly assume protein folds as discrete islands in the structure space, but overlook the connections between these discrete collections[64]. To enforce this discreteness in annotation assignments, the systems must rely on subjective and arbitrary criteria, which often lead to inconsistencies between systems and frequently cause confusion and controversy[24,34,64]. Second, the explicit requirement of manual annotations negates the possibility of scaling up to the whole known structure space or repurposing to some smaller and specialized dataset, where the annotation is lacking and the learning of representations is still an intriguing question.

To tackle this challenge, we seek to learn a continuous representation of $\mathbf{z}$ directly from the training data. Indeed, alternative to a discrete view, some past studies also suggested that the complete fold space—and consequently, the space of de novo designed proteins—should be considered geometrically continuous[23,65,66]. In light of past attempts[67–71], we incorporate an E(3)-equivariant structure encoder (with parameters $\boldsymbol{\phi}$) to infer $\mathbf{z}$ from the input ground-truth structure following the posterior distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathcal{R}^0)$ (Supplementary Methods 1.1 shows the discussion on the equivalence of E(3) and SE(3) equivariance in the reprensentation of natural proteins). We model $p(\mathbf{z})$ with an isotropic Gaussian prior, and add a Kullback–Leibler divergence loss term to encourage a continuous latent structure, effectively shaping our model into a VAE-like framework. Combined with the diffusion model, our final training objective is

$$\begin{aligned}&\arg\min_{\boldsymbol{\phi},\boldsymbol{\theta}} \mathbb{E}_{\mathcal{R}^0 \sim p_{\text{data}}(\mathcal{R})} \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathcal{R}^0)} \mathbb{E}_{t \sim \mathcal{U}\{1,T\}} \mathbb{E}_{\mathcal{R}^t \sim q(\mathcal{R}^t|\mathcal{R}^0)} \\ &\left[\mathcal{L}_{\text{recon}_{\boldsymbol{\theta}}} + \beta \text{KL}\left(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathcal{R}^0) \parallel p(\mathbf{z})\right)\right].\end{aligned} \tag{4}$$

With this training scheme, we concurrently trained an encoder to capture the essential information about the global geometry of a structure and a decoder to sample in the structure space by conditioning on it. The considerably low dimensionality of the latent and distribution regularization effect (by the Kullback–Leibler divergence loss) jointly force the encoder to capture several crucial degrees of freedom that account for the variation within the dataset, simplifying the latent distribution. By encoding all the training samples into the latent space, we further trained a latent diffusion model to capture this distribution and eventually achieve unconditional sampling of the structure space $p(\mathcal{R})$ through the unconditional sampling of $p(\mathbf{z})$ and the conditional generation of $p(\mathcal{R}|\mathbf{z})$ (equation (1)).

Despite the analogy to a class of latent diffusion models[72–74] in the use of latent diffusion, our framework does not strictly follow their architecture. In Supplementary Methods 1.3, we provide a detailed discussion of the connections between our method and latent diffusion models, along with additional domain-specific motivations and benefits.

Dataset and training summary

We prepared two datasets for the training of TopoDiff: the PDB monomer set and the CATH-60 set. The PDB monomer set was directly collected from the PDB[18], and the CATH-60 dataset was constructed based on the S60 non-redundant domain list from the CATH 4.3 release[20]. For the training of the structure diffusion module, we used a strategy that considers both model performance and training efficiency. Specifically, we first trained the structure diffusion module alone on the PDB monomer set to learn a good generative prior to the protein structure space, and in the later phase, concurrently trained this base model with a randomly initialized structure encoder in the aforementioned architecture. A detailed introduction of the dataset preparation process and the training timeline is provided in Supplementary Methods 1.4 and 1.5, with the essential parameters listed in Supplementary Tables 1–3.

Visualizing structure representation of different databases

We collected domain-level classification datasets from three different sources: CATH[19,20], SCOPe[32] and ECOD[33]. The structure encoder was trained exclusively on the CATH dataset, with the preparation and processing details outlined in Supplementary Methods 1.4. The other two datasets, SCOPe and ECOD, were used solely during the inference stage, and we restricted the included structures to single-chain proteins with a maximum of 256 residues. Detailed information for each dataset, including the version, structure count and classification hierarchy, is provided in Supplementary Table 6.

To get the representation of each structure, we extracted the coordinates of C_α atoms and inferred through the trained encoder. For dimension reduction, we applied the t-SNE[30] algorithm to compute the transformed 2D embeddings of the latent representations. Specifically, we used the implementation from openTSNE[75], with the L2 distance metric and a perplexity of 50.

To visually assess the agreement of the learned latent space with human annotations, we coloured the t-SNE scatter plot according to different annotation hierarchies. For the top hierarchy of each database, we used a discrete colour map to differentiate categories with distinct colours. For the second hierarchy, such as the architecture level in the CATH database, we created subplots for each architecture category, colouring the samples according to the kernel density estimation[76,77] of that category on the 2D latent space, providing a clear visualization of each category's distribution.

Definition of evaluation metrics for structure generation

To comprehensively benchmark the quality of structures generated by various diffusion models, we evaluated the samples with a series of metrics, each emphasizing a distinct aspect.

Coverage. We adapt the coverage metric initially proposed in ref. 42 to measure the extent to which a model can cover the natural protein space. Briefly, we first constructed a K-nearest neighbour (KNN) manifold of all real samples (that is, natural proteins) and then measured the fraction of real samples whose neighbourhoods encompass at least one fake sample (that is, structures produced by the generative models). Formally, for N given samples, the coverage over the target sample distribution is defined as

$$\begin{aligned} \mathrm{NND}_k(X_i) &:= D(X_i, X_{\mathrm{NN}(X_i,k)}) \\ B(x,r) &:= \{\, y | D(x,y) < r \} \\ \text{Coverage} &:= \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\exists j \text{ s.t } Y_j \in B(X_i, \mathrm{NND}_k(X_i))} \end{aligned}, \tag{5}$$

where $\{X_i\}$ denotes the set of real samples, $\{Y_i\}$ denotes the set of all fake samples, $\mathrm{NND}_k(X_i)$ stands for the distance from X_i to its kth nearest neighbour in $\{X_i\}$ excluding itself and $B(x, r)$ refers to the hypersphere around x with a radius of r.

On the basis of this definition, we need to define a function $D(\cdot,\cdot)$ to measure the distance between two arbitrary structures. Specifically, we need to compute at least the distance between X_i and the kth nearest neighbours in $\{X_j | j \neq i\}$ to construct the KNN manifold at the given point, and then use the distance between X_i and its first nearest neighbour in $\{Y_i\}$ to decide whether X_i is covered by an arbitrary fake sample. The final coverage metric is obtained by averaging the binary indicators over $\{X_i\}$. In this work, we computed the metric with two different distance definitions to demonstrate flexibility and robustness.

The first definition is the complement of the TM-score[43,78] of the two compared structures. To make the function symmetric with respect to the order of inputs, we define it based on the average of the query TM-score and the target TM-score:

$$D_{\mathrm{TM}} = 1 - \frac{\mathrm{TM}_{\mathrm{query}}(\mathrm{query}, \mathrm{target}) + \mathrm{TM}_{\mathrm{target}}(\mathrm{query}, \mathrm{target})}{2}. \tag{6}$$

This distance definition is a natural choice as the TM-score is designed to measure the global structural similarity and has been widely used for structure evaluation[79]. However, due to its use of dynamic programming and heuristic iterative algorithms to refine for optimal solutions, the computation of this metric is generally non-parallelizable and will be exceptionally slow when comparing a sampled structure to all natural structures. Therefore, we also provide a faster-speed alternative, taking the distance defined by a third-party structure searching model[80], which uses supervised contrastive learning to learn an embedding vector (**emb**) of each protein for structure comparison:

$$D_{\mathrm{Progres}} = 1 - \mathbf{emb}_{\mathrm{query}}^{\top} \cdot \mathbf{emb}_{\mathrm{target}}. \tag{7}$$

In practice, the pairwise distance between the natural structures is precomputed with both approaches for reuse.

The choice of the hyperparameter k of KNN should be determined before the computation of the metric. Following the recommendation in the original study[42], we chose k based on the principle that a sample size equivalent to the artificial samples from the very same distribution of real samples could achieve a coverage close enough to 1. To do this, we initially randomly sampled 500 natural chains as the pseudo-query set, used the remaining chains as samples from the target distribution and then computed the coverage of the pseudo-query set against the target distributions for different choices of k. On the basis of such a scheme, we ultimately used $k = 100$ for all experiments, although we found that different choices of k generally do not alter the relative rankings of the evaluated models (Supplementary Fig. 11).

When comparing the samples of a fixed length to the natural protein distribution, at each sampling length l, we considered all natural protein structures in the CATH-40 dataset[20] lying within the interval of $[l - 25, l + 25]$.

Diversity. To compute the diversity of N samples, we first used TM-align[78] to compute the pairwise TM-scores. Then, we clustered the samples with a cut-off of 0.6. The proportion of the total clusters to the total number of samples N was reported as diversity: a higher score generally indicates that the generated samples are more diverse.

Designability. To assess the designability of a given sample, we first used ProteinMPNN[81] to sample eight amino acid sequences with a temperature of 0.1. Subsequently, the sequences were fed to ESMFold[82] to infer the structures. The minimum RMSD of the inferred structures to the given sample was reported as scRMSD: a smaller value generally implies that the sample is more designable.

Novelty. To assess the novelty of a given sample, we began by using Foldseek[83] to query the sample against the CATH-40 dataset[20] with the parameter '-a 1 –exhaustive-search 1 -e inf -c 0.5 –alignment-type 1'. As Foldseek uses a slightly different implementation of TM-align[78], we subsequently selected the top 25 matches from the query results with the highest TM-scores and recomputed the alignment with TM-align[78]. The highest TM-score to the chains in the dataset was reported as maxTM (length normalized by the sampled structures), representing the novelty of a sample: a higher score generally implies that the sample is less novel.

In the experimental validation section, to assess novelty against the full PDB scope, we used Foldseek[83] to query the entire PDB (data up to February 2023) with the parameters '-a 1 –exhaustive-search 0 -e inf -c 0.5 –alignment-type 1' (with exhaustive pairwise search disabled). The top 25 matches were then realigned using TM-align, and the highest query TM-score was reported.

Benchmark on unconditional sampling

All benchmark experiments were conducted with TopoDiff model v. 1.1.2 (Supplementary Table 1). For each model evaluated in the benchmark testing, we randomly sampled 500 structures at each length of {50, 75, 100, 125, 150, 175, 200, 225, 250}. We first computed metrics based on all 500 samples at each length to get a series of values reflecting the length-dependent performance for each model. Since some sampled structures tended to exhibit considerably low designability as well as high novelty and diversity due to structural defects, leading to an overestimation of these metrics, we also implemented an additional step to filter the samples by only preserving those with scRMSD ≤ 2 Å and recomputed the metrics at each length. Finally, for each metric, we averaged out along all the sampled lengths to present an indication of its overall performance. We used the length-averaged metrics to draw the radar plot (Fig. 3c).

Co-visualization of latent space and generated structures

To project the latent codes of sampled structures onto the t-SNE dimension-reduced manifold, we used the transform method implemented in openTSNE[75] to embed the newly sampled latents into the same space that we used for the CATH dataset visualization. Briefly, following the same basic principles as conventional t-SNE, the positions of the background embeddings were kept fixed, and each sampled latent was optimized independently with respect to them. We only selected the latent codes associated with the fairly designable and novel samples (that is, scRMSD ≤ 2 Å and CATH-maxTM < 0.7). Finally, we also co-visualized some sampled structures spanning the entire latent manifold, providing a visual context on the sampled structures and their spatial relationships with respect to the latent codes.

Tuning model preference by latent-space rejection sampling

To tune the sampled latent distribution, we began by training latent classifiers that predict the structural properties of the given latent

codes (Supplementary Methods 1.6). We further used these latent classifiers to conduct rejection sampling over the sampled latents, obtaining variants of the model with distinct sampling preferences. Recall that $p_\theta(\mathbf{z})$ is the base distribution; we unconditionally sample from the latent diffusion module. Our goal is to reweight the sampled latent distribution using a set of trained classifier functions $f_1(\mathbf{z}),\ldots,f_n(\mathbf{z})$ and corresponding thresholds $c_1,\ldots,c_n$ based on an acceptance probability function $\alpha(\mathbf{z})$, such that the resulting latent distribution $p_{\text{accept}}(\mathbf{z})$ (and eventually the sampled structure distribution) would be shifted towards our intended preference.

Specifically, for each variant, we first defined a series of length-dependent thresholds in terms of the designability (scRMSD) and novelty (maxTM) of the sampled structures. For simplicity, we only considered three possible variants: designability-prior (sampling highly designable structures with the designability classifier), novelty-prior (sampling highly novel structures with the novelty classifier) and all-round (sampling structures with fairly high designability and novelty when collectively using both classifiers). The exact thresholds we used at various sample lengths are summarized in Supplementary Table 4.

Intuitively, we want to sample the latent codes with a higher probability when they match the expectation, and lower if not. To achieve this aim, we define an acceptance probability function $\alpha(\mathbf{z})$, requiring that $\mathbf{z}$ is accepted with the probability of 1 only if approved by all classifiers and with the probability of 0.1 if denied by any classifier:

$$\alpha(\mathbf{z}) = \begin{cases} 1, & \text{if } f_i(\mathbf{z}) \le c_i \text{ for all } i = 1 \ldots n, \\ 0.1, & \text{if } f_i(\mathbf{z}) > c_i \text{ for any } i = 1 \ldots n. \end{cases} \tag{8}$$

The resulting probability density function after applying the rejection sampling procedure is

$$p_{\text{accept},\theta}(\mathbf{z}) = \frac{p_\theta(\mathbf{z}) \cdot \alpha(\mathbf{z})}{Z_{\text{accept}}}, \tag{9}$$

where $p_\theta(\mathbf{z})$ is the original latent distribution sampled from the latent diffusion module,

$\alpha(\mathbf{z})$ is the combined acceptance probability as defined above and Z_{accept} is the normalization constant, ensuring that $p_{\text{accept},\theta}(\mathbf{z})$ integrates to 1.

This formulation ensures that the resulting distribution $p_{\text{accept},\theta}(\mathbf{z})$ reflects the collective influence of all the classifiers. By training all modules once and setting different combinations of these thresholds, we shifted the sampled distribution to match our expectations as closely as possible.

For each variant, we sampled 500 structures per sample length using the given latent sampling strategy, and computed all metrics following the definition introduced in the benchmark section.

Sampling of structures with similar global geometry

To sample structures sharing similar global geometry with respect to a reference structure, we encoded the query structures in our CATH training set using the trained encoder, and randomly resampled five structures with each inferred latent code. We selected reference structures with a wide variety of their SSE compositions and spatial arrangements, and listed their original structures and the five non-cherry-picked samples side by side.

Sampling of structures with latent code interpolation

For the interpolation experiment, we first randomly chose a number of candidate pairs of samples from the CATH training set. For each candidate pair, we conducted a linear interpolation to get ten latent codes evenly located between the two termini. Finally, we randomly selected a chain length between 75 and 150 residues, and sampled a structure with each latent code.

For structure sampling, we implemented with a setting that minimizes the stochasticity introduced into the reverse sampling process (Supplementary Table 5). Specifically, in the $\mathbb{R}^3$ space, we used the denoising diffusion implicit models[84] formulation instead of the default denoising diffusion probabilistic models[13], where the reverse sampling process is reformulated in a noise-free way that $\mathcal{R}^{t-1}$ is deterministically derived given $\mathcal{R}^t$, $\hat{\mathcal{R}}_\theta$ and the noise schedule. Similarly, in the SO(3) space, we used a score scale of 1 or 2 and a noise scale of 0, such that the reverse sampling process would also introduce no additional stochasticity. By using this reverse sampling setting, we ensured that the generated structures are solely determined by the initial state of noise sampled from the $\text{SE}(3)^l$ space and the conditioned latent code. For the structure generation of each series of interpolated latent codes, we further fixed the random seed used for the initial-state sampling, so that the only difference between the generation of each structure is the latent code itself. By this means, we could ensure that the generated structures are more consistent with the latent code they are conditioned on and reduce unwanted variability caused by sampling stochasticity.

Sampling of structures with motif scaffolding

For the motif scaffolding experiment, we selected three representative motifs from the RFDiffusion work[17], whose original structures adopt distinct secondary structures (mainly alpha, mixed of alpha and beta, and mainly beta). We largely followed the experimental setting introduced in the original study[17]. Specifically, each design case was first associated with a detailed setting on the motif information and constraints, such as the total length of designs, motif definition and sequential position of motif on designs. Subsequently, we randomly sampled 5,000 valid combinations of constraints and latent codes, and used these combinations for structure sampling. For each generated structure, we designed eight sequences with ProteinMPNN[81] and folded them with ESMFold[82]. The designs were marked successful if global scRMSD ≤ 2 Å and motif scRMSD ≤ 1 Å. We kept these settings as close as possible with respect to the original study[17], except for two differences. First, for the two-strand 4ZYP motif, since the originally designed length (~30–50) was too short to adopt most of the stable mainly beta topologies, we kept the motif fixed and increased the sample length to ~90–110. Second, during sequence design, we did not freeze the motif sequence as our structure diffusion module was not specifically trained to be aware of the motif sequence information. The final motif design constraints are summarized in Supplementary Table 7.

For each motif, we gathered all the successful designs and projected the corresponding latent codes onto the *t*-SNE dimension-reduced plot. The scatter plot is visualized as a heat map, with each point coloured according to the kernel density estimation[76,77] of the successful designs on the *t*-SNE space. We also co-visualized the generated structures from each region of the manifold to demonstrate the diversity and controllability of the global geometry. In the structural presentation, the motif part is coloured in yellow and the rest, in sky blue.

In silico design of novel mainly beta proteins

Due to the asynchronous development of the model, all the experiment validations in this section were conducted using TopoDiff model v. 1.1.1 (Supplementary Table 1). All the supplementary modules (for example, latent diffusion module, latent classifiers and so on) used in this experiment were also trained based on this version. Briefly, we first sampled a number of protein backbones using the aforementioned rejection sampling strategy to focus the sampling on mainly beta proteins with good novelty and designability. We then adopted a three-stage filtering pipeline to generate the sequence designs and screen for designs with exceptionally good in silico quality.

Specifically, we began by generating novel protein backbones through four rounds of sampling. For each round of sampling, we first sampled 17,500 backbones (2,500 samples from each of the seven protein lengths between 50 and 200 residues), and applied a stringent

filtering criterion to identify some of the most promising mainly beta designs in terms of novelty and designability. In the first round of sampling, we conducted a completely unconditional sampling, resulting in 33 backbones with 70 sequence designs. In the three subsequent rounds, we gradually incorporated latent classifiers to reweight the sampled latent distribution based on the predicted beta ratio, novelty and designability. This approach effectively enforced in silico enrichment towards latent codes that were likely to decode into novel, designable mainly beta proteins, yielding a fourfold increase in successful designs when using all three classifiers.

Finally, out of the 403 resulting backbone with 950 sequence designs, we selected 21 designs ranging from 50 residues to 200 residues for experimental validation. The structure sampling and the in silico selection processes are schematically summarized in Supplementary Fig. 2 and are elaborated in detail in Supplementary Methods 2.5.

Protein expression and purification

Genes encoding the selected designs were synthesized from GenScript, cloned onto the pET-29b(+) vector using GoldenGate assembly (NEB, R3733L and M0202T), verified with Sanger sequencing, and transformed into the BL21 (DE3) *E. coli* strain (Tsingke, TSC-E06). Protein expression was induced in the Terrific Broth medium with 50 μg ml^{-1} of kanamycin overnight at 37 °C. Protein expression was confirmed by western blotting against the N-terminal His tag (Proteintech, 66005, SA00001-1). Bacterial cells were harvested by centrifugation, resuspended and homogenized in lysis buffer (50 mM of phosphate, 300 mM of NaCl, 30 mM of imidazole, pH 8.0), and lysed by sonication. Lysates were cleared by centrifugation (4 °C, 12,000*g*, 60 min). The supernatants were loaded onto Ni-NTA columns (Solarbio, P2010) pre-equilibrated with a lysis buffer. The columns were washed with ten column volumes of lysis buffer, followed by ten column volumes of high-salt wash buffer (50 mM of phosphate, 1 M of NaCl, 30 mM of imidazole, pH 8.0), five column volumes of high-imidazole wash buffer (50 mM of phosphate, 300 mM of NaCl, 50 mM of imidazole, pH 8.0) and eluted in an elution buffer (50 mM of phosphate, 300 mM of NaCl, 500 mM of imidazole, pH 8.0).

SEC

Eluted protein samples from Ni-NTA purification were concentrated using Amicon Ultra 3 kDa MWCO concentrators (Merck Millipore, UFC900308) to a final volume of 1.2 ml, and were further purified by SEC using a Superdex 75 increase 10/300 GL column (Cytiva, 29148721). Fractions were collected for further analysis, according to the light absorption at 280 nm and the results of sodium dodecyl sulfate–polyacrylamide gel electrophoresis.

CD

CD spectra from 190 nm to 280 nm were recorded at 25 °C and 95 °C using a 1-mm-path-length cuvette using a Chirascan Plus CD spectrometer (Applied Photophysics). The background spectra were acquired across the same spectral range and manually subtracted. The processed CD data (in millidegrees) were further normalized by the sample concentration and the sequence length to derive the mean residue ellipticity. Temperature melts were conducted in 1 °C steps (heating rate of 1 °C min^{-1}) by measuring the signal of samples prepared at 0.2 mg ml^{-1} in phosphate-buffered saline buffer (25 mM of phosphate, 150 mM of NaCl, pH 7.4) at a wavelength of 215 nm.

Crystallization and structure determination

Crystallization was screened at 16 °C by sitting-drop vapour diffusion using commercial kit sets. For the design B10, before screening, the protein was concentrated to 10.6 mg ml^{-1}. Crystals of diffraction quality were obtained under two conditions: (1) 30% (w/v) of PEG-3350, 0.1 M of Tris-HCl, 0.2 M of NaCl, pH 8.5; (2) 30% (w/v) of PEG-400, 0.2 M of sodium citrate tribasic dihydrate, 0.1 M of Tris-HCl, pH 8.5. The collected crystals were then flash cooled in liquid nitrogen without cryoprotection. All data were collected on 02U1 at the Shanghai Synchrotron Radiation Facility and processed with HKL2000[85]. Phenix[86] was used for molecular-replacement structure determination and subsequent refinement, using AlphaFold2[11] prediction as the initial model. Manual rebuilding was performed in Coot[87], and all the molecular graphics were generated with PyMOL[88] and ChimeraX[89].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The dataset used for model training, along with the trained model weights, benchmark data and protein designs selected for experimental validation, is available via Zenodo at https://zenodo.org/records/13879811 (ref. 90). The crystal structure models have been deposited in the Protein Data Bank (accession codes 9KGZ and 9KGY). Source data are provided with this paper.

Code availability

The TopoDiff model is implemented in PyTorch. Full scripts (including the training code) and guidance for utilizing the model are available via GitHub at https://github.com/meneshail/TopoDiff/tree/main (ref. 91). A reproducible code capsule of TopoDiff is available via CodeOcean at https://doi.org/10.24433/CO.8705528.v1 (ref. 92).

References

1. Chevalier, A. et al. Massively parallel de novo protein design for targeted therapeutics. *Nature* **550**, 74–79 (2017).
2. Silva, D.-A. et al. De novo design of potent and selective mimics of IL-2 and IL-15. *Nature* **565**, 186–191 (2019).
3. Roy, A. et al. De novo design of highly selective miniprotein inhibitors of integrins avβ6 and avβ8. *Nat. Commun.* **14**, 5660 (2023).
4. Yeh, A. H.-W. et al. De novo design of luciferases using deep learning. *Nature* **614**, 774–780 (2023).
5. Langan, R. A. et al. De novo design of bioactive protein switches. *Nature* **572**, 205–210 (2019).
6. Chen, Z. et al. De novo design of protein logic gates. *Science* **368**, 78–84 (2020).
7. Pan, X. & Kortemme, T. Recent advances in de novo protein design: principles, methods, and applications. *J. Biol. Chem.* **296**, 100558 (2021).
8. Wu, K. E. et al. Protein structure generation via folding diffusion. *Nat. Commun.* **15**, 1059 (2024).
9. Ni, B., Kaplan, D. L. & Buehler, M. J. Generative design of de novo proteins based on secondary-structure constraints using an attention-based diffusion model. *Chem* **9**, 1828–1849 (2023).
10. Lee, J. S., Kim, J. & Kim, P. M. Score-based generative modeling for de novo protein design. *Nat. Comput. Sci.* **3**, 382–392 (2023).
11. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
12. Baek, M. et al. Efficient and accurate prediction of protein structure using RoseTTAFold2. Preprint at *bioRxiv* https://doi.org/10.1101/2023.05.24.542179 (2023).
13. Anand, N. & Achim, T. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. Preprint at https://arxiv.org/abs/2205.15019 (2022).
14. Luo, S. et al. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. In *Proc. 36th International Conference on Neural Information Processing Systems* (eds Koyejo, S. et al.) 9754–9767 (Curran Associates Inc., 2022).

15. 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 (JMLR.org, 2023).
16. Lin, Y. & AlQuraishi, M. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. In *Proc. 40th International Conference on Machine Learning* (eds Krause, A. et al.) 20978–21002 (PMLR, 2023).
17. Watson, J. L. et al. De novo design of protein structure and function with RFDiffusion. *Nature* **620**, 1089–1100 (2023).
18. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
19. Orengo, C. A. et al. CATH—a hierarchic classification of protein domain structures. *Structure* **5**, 1093–1108 (1997).
20. Sillitoe, I. et al. CATH: increased structural coverage of functional space. *Nucleic Acids Res.* **49**, D266–D273 (2021).
21. Bennett, N. R. et al. Atomically accurate de novo design of single-domain antibodies. Preprint at https://doi.org/10.1101/2024.03.14.585103 (2024).
22. Ingraham, J. B. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
23. Sadreyev, R. I., Kim, B.-H. & Grishin, N. V. Discrete-continuous duality of protein structure space. *Curr. Opin. Struct. Biol.* **19**, 321–328 (2009).
24. Pascual-García, A., Abia, D., Ortiz, A. R. & Bastolla, U. Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLoS Comput. Biol.* **5**, e1000331 (2009).
25. Martin, A. C. et al. Protein folds and functions. *Structure* **6**, 875–884 (1998).
26. Hegyi, H. & Gerstein, M. The relationship between protein structure and function: a comprehensive survey with application to the yeast genome 1. *J. Mol. Biol.* **288**, 147–164 (1999).
27. Micheletti, C. Prediction of folding rates and transition-state placement from native-state geometry. *Proteins* **51**, 74–84 (2003).
28. Wang, J. & Panagiotou, E. The protein folding rate and the geometry and topology of the native state. *Sci. Rep.* **12**, 6384 (2022).
29. Luo, C. Understanding diffusion models: a unified perspective. Preprint at https://arxiv.org/abs/2208.11970 (2022).
30. Maaten, L. v. d. & Hinton, G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540 (1995).
32. Hubbard, T. J., Murzin, A. G., Brenner, S. E. & Chothia, C. SCOP: a structural classification of proteins database. *Nucleic Acids Res.* **25**, 236–239 (1997).
33. Cheng, H. et al. ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.* **10**, e1003926 (2014).
34. Day, R., Beck, D. A., Armen, R. S. & Daggett, V. A consensus view of fold space: combining SCOP, CATH, and the Dali Domain Dictionary. *Protein Sci.* **12**, 2150–2160 (2003).
35. Csaba, G., Birzele, F. & Zimmer, R. Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Struct. Biol.* **9**, 23 (2009).
36. Schaeffer, R. D., Kinch, L. N., Pei, J., Medvedev, K. E. & Grishin, N. V. Completeness and consistency in structural domain classifications. *ACS Omega* **6**, 15698–15707 (2021).
37. Mura, C., Veretnik, S. & Bourne, P. E. The *Urfold*: structural similarity just above the superfold level? *Protein Sci.* **28**, 2119–2126 (2019).
38. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J. & Aila, T. Improved precision and recall metric for assessing generative models. In *Proc. 33rd International Conference on Neural Information Processing Systems* (eds Wallach, H. M. et al.) 3927–3936 (Curran Associates Inc., 2019).
39. Listov, D., Goverde, C. A., Correia, B. E. & Fleishman, S. J. Opportunities and challenges in design and optimization of protein function. *Nat. Rev. Mol. Cell Biol.* **25**, 639–653 (2024).
40. Chu, A. E., Lu, T. & Huang, P.-S. Sparks of function by de novo protein design. *Nat. Biotechnol.* **42**, 203–215 (2024).
41. Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **384**, eadl2528 (2024).
42. Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y. & Yoo, J. Reliable fidelity and diversity metrics for generative models. In *Proc. 37th International Conference on Machine Learning* (eds Daumé, H. & Singh, A.) 7176–7185 (JMLR.org, 2020).
43. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57**, 702–710 (2004).
44. Greener, J. G. & Jamali, K. Fast protein structure searching using structure graph embeddings. *Bioinform. Adv.* **5**, vbaf042 (2025).
45. Bose, A. J. et al. *Proc. 12th International Conference on Learning Representations* (OpenReview.net, 2024).
46. Lin, Y., Lee, M., Zhang, Z. & AlQuraishi, M. Out of many, one: designing and scaffolding proteins at the scale of the structural universe with Genie 2. Preprint at https://arxiv.org/abs/2405.15489 (2024).
47. Huguet, G. et al. Sequence-augmented SE(3)-flow matching for conditional protein generation. In *Advances in Neural Information Processing Systems 37* (eds Globerson, A. et al.) 33007–33036 (Curran Associates, Inc., 2024).
48. Chronowska, M., Stam, M. J., Woolfson, D. N., Di Costanzo, L. F. & Wood, C. W. The Protein Design Archive (PDA): insights from 40 years of protein design. *Nat. Biotechnol.* **43**, 669–671 (2024).
49. Hermosilla, A. M., Berner, C., Ovchinnikov, S. & Vorobieva, A. A. Validation of de novo designed water-soluble and transmembrane β-barrels by in silico folding and melting. *Protein Sci.* **33**, e5033 (2024).
50. Liu, Y., Chen, L. & Liu, H. Diffusion in a quantized vector space generates non-idealized protein structures and predicts conformational distributions. Preprint at *bioRxiv* https://doi.org/10.1101/2023.11.18.567666 (2023).
51. Fu, C. et al. A latent diffusion model for protein structure generation. In *Proc. Second Learning on Graphs Conference* (eds Villar, S. & Chamberlain, B.) 29:1–29:17 (PMLR, 2024).
52. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at https://arxiv.org/abs/2204.06125 (2022).
53. Preechakul, K., Chatthee, N., Wizadwongsa, S. & Suwajanakorn, S. *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022).
54. Kim, S. W. et al. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2023).
55. Praetorius, F. et al. Design of stimulus-responsive two-state hinge proteins. *Science* **381**, 754–760 (2023).
56. Berger, S. et al. Preclinical proof of principle for orally delivered Th17 antagonist miniproteins. *Cell* **187**, 4305–4317.e18 (2024).
57. Glögl, M. et al. Target-conditioned diffusion generates potent TNFR superfamily antagonists and agonists. *Science* **386**, 1154–1161 (2024).
58. Huang, B. et al. Designed endocytosis-inducing proteins degrade targets and amplify signals. *Nature* **638**, 796–804 (2024).
59. Baker, D. et al. De novo designed proteins neutralize lethal snake venom toxins. *Nature* **639**, 225–231 (2024).
60. An, L. et al. Binding and sensing diverse small molecules using shape-complementary pseudocycles. *Science* **385**, 276–282 (2024).
61. Chu, A. E. et al. An all-atom protein generative model. *Proc. Natl Acad. Sci. USA* **121**, e2311500121 (2024).

62. Campbell, A., Yim, J., Barzilay, R., Rainforth, T. & Jaakkola, T. Generative flows on discrete state-spaces: enabling multimodal flows with applications to protein co-design. In *Proc. 41st International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) 5453–5512 (JMLR.org, 2024).
63. Dietmann, S. et al. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res.* **29**, 55–57 (2001).
64. Xu, J. & Zhang, J. Impact of structure space continuity on protein fold classification. *Sci. Rep.* **6**, 23263 (2016).
65. Skolnick, J., Arakaki, A. K., Lee, S. Y. & Brylinski, M. The continuity of protein structure space is an intrinsic property of proteins. *Proc. Natl Acad. Sci. USA* **106**, 15690–15695 (2009).
66. Woolfson, D. N. et al. De novo protein design: how do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.* **33**, 16–26 (2015).
67. Greener, J. G., Moffat, L. & Jones, D. T. Design of metalloproteins and novel protein folds using variational autoencoders. *Sci. Rep.* **8**, 16189 (2018).
68. Hawkins-Hooker, A. et al. Generating functional protein variants with variational autoencoders. *PLoS Comput. Biol.* **17**, e1008736 (2021).
69. Guo, X., Du, Y., Tadepalli, S., Zhao, L. & Shehu, A. Generating tertiary protein structures via interpretable graph variational autoencoders. *Bioinform. Adv.* **1**, vbab036 (2021).
70. Eguchi, R. R., Choe, C. A. & Huang, P.-S. Ig-VAE: generative modeling of protein structure by direct 3D coordinate generation. *PLoS Comput. Biol.* **18**, e1010271 (2022).
71. Lai, B., McPartlon, M. & Xu, J. End-to-end deep structure generative model for protein design. Preprint at *bioRxiv* https://doi.org/10.1101/2022.07.09.499440 (2022).
72. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2022).
73. Podell, D. et al. *Proc. 12th International Conference on Learning Representations* (OpenReview.net, 2024).
74. Esser, P. et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. 41st International Conference on Machine Learning* (eds Salakhutdinov, R. et al.) 12606–12633 (JMLR.org, 2024).
75. Poličar, P. G., Stražar, M. & Zupan, B. openTSNE: a modular Python library for *t*-SNE dimensionality reduction and embedding. *J. Stat. Softw.* **109**, 1–30 (2024).
76. Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization* 1st edn (Wiley, 1992).
77. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
78. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
79. Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K. & Moult, J. Critical assessment of methods of protein structure prediction (CASP)-Round XV. *Proteins* **91**, 1539–1549 (2023).
80. Greener, J. G. & Jamali, K. Fast protein structure searching using structure graph embeddings. *Bioinform. Adv.* **5**, vbaf042 (2022).
81. Dauparas, J. et al. Robust deep learning-based protein sequence design using ProteinMPNN. *Science* **378**, 49–56 (2022).
82. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
83. Van Kempen, M. et al. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2023).
84. Song, J., Meng, C. & Ermon, S. *Proc. 9th International Conference on Learning Representations* (OpenReview.net, 2021).
85. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* (ed. Carter, C. W. Jr) 307–326 (Elsevier, 1997).
86. Adams, P. D. et al. *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
87. Emsley, P. & Cowtan, K. *Coot*: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
88. The PyMOL Molecular Graphics System (Schrödinger, LLC, 2015).
89. Meng, E. C. et al. UCSF ChimeraX: tools for structure building and analysis. *Protein Sci.* **32**, e4792 (2023).
90. Zhang, Y. et al. Improving diffusion-based protein backbone generation with global-geometry-aware latent encoding. Preprint at *bioRxiv* https://doi.org/10.1101/2024.10.05.616664 (2024).
91. Zhang, Y. meneshail/TopoDiff: v1.1.0. *GitHub* https://github.com/meneshail/TopoDiff/tree/main (2025).
92. Zhang, Y., Liu, Y., Ma, Z., Li, M. & Chunfu, X. CodeOcean release of 'TopoDiff: improving diffusion-based protein backbone generation with global-geometry-aware latent encoding', version 1. *CodeOcean* https://doi.org/10.24433/CO.8705528.v1 (2025).

Acknowledgements

This work has been supported by the Ministry of Science and Technology of China (no. 2023YFF1204400 to H.G.), the National Natural Science Foundation of China (no. 32171243 to H.G.) and the Beijing Frontier Research Center for Biological Structure. We thank the staff of beamlines BL02U1, BL10U2, BL18U1 and BL19U1 at the Shanghai Synchrotron Radiation Facility as well as the X-ray crystallography platform, National Protein Science Facility, Tsinghua University, for assistance in the X-ray diffraction data collection and analysis. We thank J. Hu, Z. Zhu, Y. Xue and C. Song for helpful discussions.

Author contributions

Y.Z. and H.G. conceived the study. Y.Z. and Z.M. designed and implemented the model. Y.Z. and Z.M. performed the in silico experiments and analysed the results. Y.Z. designed the candidate proteins for experimental validation. Y.L. designed, executed and analysed all the wet-laboratory experiments. H.G. supervised the development of the model and the result analysis. C.X. supervised the design of the candidate proteins and wet-laboratory experiments. M.L. contributed to the X-ray structure determination. Y.Z. drafted the initial paper. Y.Z., Y.L. and Z.M. created the final figures. All authors contributed to writing and improving the paper, and approved the submission.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s42256-025-01059-x.

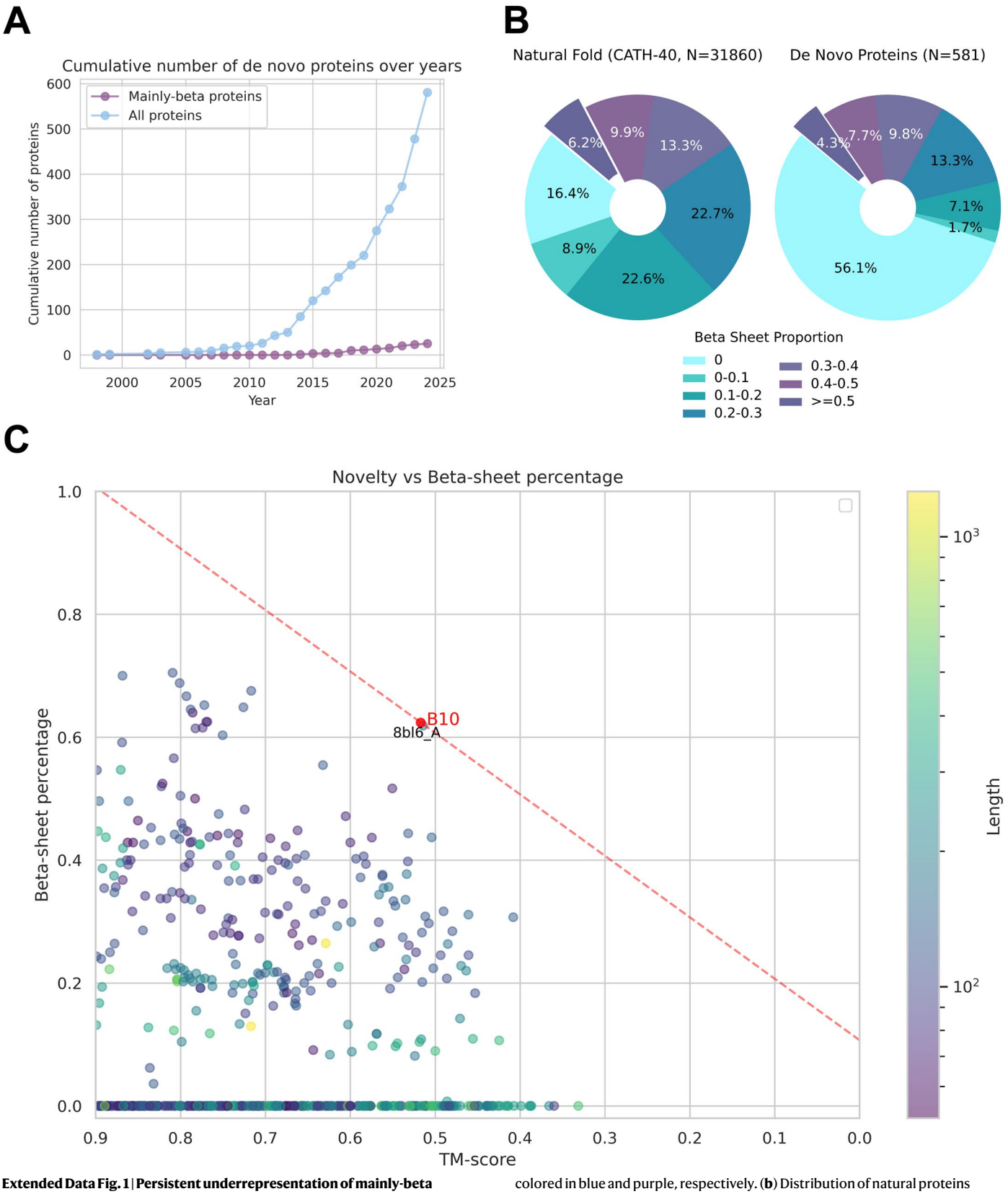
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-025-01059-x.

Correspondence and requests for materials should be addressed to Chunfu Xu or Haipeng Gong.

Peer review information *Nature Machine Intelligence* thanks Zhuoran Qiao, Limei Wang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

[1]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China. [2]Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing, China. [3]National Institute of Biological Sciences, Beijing, China. [4]Tsinghua Institute of Multidisciplinary Biomedical Research, Tsinghua University, Beijing, China. [5]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [6]National Center for Protein Sciences, Beijing, China. [7]X-ray Crystallography Facility, Technology Center for Protein Sciences, Tsinghua University, Beijing, China. [8]These authors contributed equally: Yuyang Zhang, Yuhang Liu, Zinnia Ma. ✉e-mail: xuchunfu@nibs.ac.cn; hgong@tsinghua.edu.cn

https://doi.org/10.1038/s42256-025-01059-x



Extended Data Fig. 1 | Persistent underrepresentation of mainly-beta proteins in *de novo* design. All *de novo* designed proteins deposited in PDB up to September 2024 were collected from the PDA database, and filtered to exclude small peptides (length ≤ 50) as well as designs originating from sequence mutations or redesigns of naturally occurring backbones (maximum TM-score to PDB ≥ 0.9). (**a**) Cumulative number of de novo protein design entries over the time. General proteins and mainly-beta proteins (with beta ratio ≥ 0.5) are colored in blue and purple, respectively. (**b**) Distribution of natural proteins of the CATH dataset (left) and de novo designed proteins (right) based on the proportion of beta sheets. (**c**) Scatter plot of all de novo designed proteins, where the horizontal axis represents novelty (maximum TM-score to PDB) and the vertical axis represents the proportion of beta sheets. Each protein is denoted as a point, colored based on protein length. Detailed discussion of these data could be found in Supplementary Results 6.3.

nature portfolio

Corresponding author(s): Haipeng Gong and Chunfu Xu

Last updated by author(s): Apr 27, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The TopoDiff model is implemented in PyTorch. Full scripts (including the training code) and guidance for utilizing the model are available at https://github.com/meneshail/TopoDiff/tree/main.
Data analysis	Python 3.8.7, PyTorch 2.0.1, openTSNE 1.0.1, POT (Python Optimal Transport) 0.9.3, seaborn 0.12.2, NumPy 1.23.5, SciPy 1.10.1, Foldseek (c7e4a37856b49438eaee03bbfcdf1588cbce0695), TM-tools 0.0.3

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The dataset used for model training, along with the trained model weights, benchmark data and protein designs selected for experimental validation, are available at https://zenodo.org/records/13879812.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size was determined based on the previously reported success rate of de novo protein design and the number of diverse designs that successfully met our computational filtering criteria.
Data exclusions	No data were excluded from this study.
Replication	Protein purification, SEC, SDS-PAGE and CD were performed at least twice for each experiment. For the proteins that yield crystal structure, the preparation of crystallization samples was repeated twice using different batches of pure proteins on separate days. All replication attempts were successful.
Randomization	Randomization is not relevant to the study.
Blinding	Blinding is not relevant to the study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	His-Tag Monoclonal Antibody (Mouse, proteintech, 66005, 1:2000 in TBST) HRP-conjugated Goat Anti-Mouse IgG(H+L)(Goat, proteintech, SA00001-1, 1:5000 in TBST)

Validation	The antibodies were used to comfirm the expression of selected designs by Western blot against N-terminal his-tag. All antibodis were purchased from commercial sources. The verification statements can be found at the manufacturers' website.