ProteinSGM: Score-based generative modeling for de novo protein design

Jin Sub Lee
University of Toronto
Philip Kim (✉ pi@kimlab.org)
University of Toronto https://orcid.org/0000-0003-3683-152X

Article

Keywords:

Posted Date: October 10th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1855828/v1

ProteinSGM: Score-based generative modeling for *de novo* protein design

Jin Sub Lee[1,2] **and Philip M. Kim**[1,2,3,*]

[1]Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 3E1, Canada

[2]Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, ON M5S 3E1, Canada

[3]Department of Computer Science, University of Toronto, Toronto, ON M5S 3E1, Canada

[*]Correspondence: Philip M. Kim pi@kimlab.org

ABSTRACT

Score-based generative models are a novel class of generative models that have shown state-of-the-art sample quality in image synthesis, surpassing the performance of GANs in multiple tasks. Here we present ProteinSGM, a score-based generative model that produces realistic *de novo* proteins and can inpaint plausible backbones and functional sites into structures of predefined length. With *unconditional generation*, we show that score-based generative models can generate native-like protein structures, surpassing the performance of previously reported generative models. We apply *conditional generation* to *de novo* protein design by formulating it as an image inpainting problem, allowing precise and modular design of protein structure.

Introduction

Deep learning has significantly advanced protein engineering with robust methods for structure prediction and sequence design. AlphaFold 2 [1] allows researchers to access astronomically more protein structures while circumventing the strenuous effort of 3D structure determination, and recent sequence design methods [2, 3] show robust sequence recovery given protein backbone information. One fundamental challenge in protein design that is largely unaddressed is the design of novel backbones - that is, first, can we generate synthetic backbones that can be realized by a protein sequence, and second, can we discover novel folds not found in the native fold space (i.e. folds that are not found in SCOP [4] or CATH [5]? An extension of this problem is the task of *conditional* backbone generation: for a given functional site, can we generate novel yet compatible scaffolds for the functional site that will retain its activity? In this work, we seek to address these questions by leveraging recent advances in deep generative modeling.

Diffusion models [6, 7] have shown unprecedented success in various domains such as image synthesis [8], audio [9], text-to-image generation [10], graphs [11, 12], and 3D molecule generation [13]. Diffusion models define a forward diffusion process that perturbs data into noise and learns the reverse process that denoises random Gaussian noise into samples from the data distribution. Score-based generative modeling [7, 14] is a formulation of diffusion models that estimates the scores, or gradient of the log probability density with respect to data, of perturbed data in varying noise scales using a noise-conditional neural network. Sampling is performed by using the estimated scores with Langevin dynamics, which maps the Gaussian prior to the data distribution, thereby generating data from noise. Despite its success in many domains, its application to protein design is still limited [15, 16].

Here we present ProteinSGM, a continuous-time score-based generative model that generates *de novo* proteins with high sample quality and mode coverage. ProteinSGM learns to generate four matrices that fully describes a protein's backbone, which are used as smoothed harmonic constraints in the Rosetta minimization protocol. Sequence and rotamer design is performed using fixed-backbone `FastDesign`, and lastly the structure is relaxed with constraints to generate a low-energy full-atom structure. We show that ProteinSGM generates variable-length structures with a mean $<$ -2.9 REU per residue, indicative of native-like structures. Moreover, we show that an unconditionally trained model is able to effectively inpaint masked coordinates, allowing direct control over protein functional site, backbone, and/or length while still generating realistic structures that follow the user-defined constraints.

Results

In this work, we train a score-based generative model using image-like representations of protein structures, where each protein backbone is represented by inter-residue 6D coordinates [Figure 1A] as defined in trRosetta [17]. In brief, from each protein we calculate four matrices corresponding to Cβ-Cβ distances (hereafter referred to as d), ω and θ torsional angles, and φ planar angles that fully describe a protein backbone. These matrices constitute the 6D coordinates since φ and θ angles are asymmetric (ex. $\varphi_{ij} \neq \varphi_{ji}$ for residues i, j where $i \neq j$). For variable-length generation, the matrices are centered and an additional padding channel is added to differentiate the centered matrices from the padding residues. We observe that the model generates centered squares in most cases, and can be used in conditional generation for defining protein length. Taken altogether, a single protein structure is represented as a 128x128x5 tensor, with 4 channels corresponding to the 6D coordinates and one padding channel [Figure 1B].

We use the continuous-time framework of score-based generative modeling with stochastic differential equations [14], the first application to protein design to date [Figure 1C]. The model is trained to denoise realistic 6D coordinates from the Gaussian prior by estimation of the score function $\nabla_x \log p_t(x)$, which is used to solve the reverse-time SDE for mapping Gaussian noise into data (see *Methods*). The 6D coordinates are then subject to Rosetta minimization using `MinMover` for backbone minimization with constraints, `FastDesign` for fixed backbone sequence and rotamer design, and a final `FastRelax` constrained relaxation step to generate low-energy full-atom structures. After model training, we assess the model performance on unconditional generation to assess sample diversity and plausibility, and apply conditional generation by imputing masked input features for various protein design cases [Figure 1D].

Unconditional generation

6D coordinate analysis. Adjacent residues of proteins constrain the inter-residue internal coordinates and therefore exhibit specific inter-residue distributions. To verify that the model is learning natural biophysical constraints of proteins and effectively capturing these distributions, we generated 1000 samples with the fully-trained model and compared 6D coordinate distributions of adjacent residues to the distributions of the training data [Figure 2]. Across all d, ω, θ, ϕ distributions, we observe that the distributions match closely to those of the training set, suggesting that the model has learned to generate realistic and native-like 6D coordinates of varying lengths. We provide a few examples of the features generated by the model. For more generations, refer to Figure S2.

Structure analysis. We proceeded to generate 966 full-atom structures with the Rosetta protocol and compared their properties to minimized structures from the training set. We observe that the protein length distribution closely matches the distribution of the training set, though there is a reduced frequency of longer structures [Figure 3A]. More importantly, the generated structures have Rosetta energies comparable to those of native structures, with means of -2.9 REU per residue and -3.4 REU per residue, respectively [Figure 3B]. To the best of our knowledge, this is the first generative model reported to show negative Rosetta energies across all generated samples. To assess whether the model generates high-fidelity structures across different lengths, we split the generated samples across four categories based on length and compared the Rosetta energy distributions [Figure S3]. We observe that though there is a slight increase of the median Rosetta energy in longer structures, a two-sample Kolmogorov-Smirnov test between the shortest and longest length categories was not statistically significant, suggesting that the structural quality of samples is independent of protein length.

Next, we analyzed the secondary structure distributions using DSSP [18] between native and generated samples [Figure 3C, D]. We noticed that the generated samples, when compared to native structures, have a noticably lower mean proportion of beta sheets (0.09 vs 0.20) and higher mean proportion of alpha helices (0.44 vs 0.30), respectively. This is expected since beta sheets require specific local and global structural constraints for proper beta-sheet formation, while alpha helices are more dependent on local interactions between neighboring residues. To detect generalization and diversity of the samples, we measured pairwise TM-scores with TMalign [19] between each generated sample and structures from the training set [Figure 3E]. We observed that a subset of the structures exhibited TM-scores of < 0.5, indicative of the model learning to generate novel folds not present in the training set. Moreover, this indicates that ProteinSGM does not simply memorize structures found in the training set, which would strongly skew the max TM-score distribution to 1.0. When assessing the relationship between TM-score and Rosetta energy, we observed a strong negative correlation ($R^2 = -0.71$), which indicates that generated structures with similar folds to native structures generally resulted in lower Rosetta energy, and therefore increased structural plausibility [Figure S4]. To further assess structural stability, we ran 200-ns MD simulations for three selected structures [Figure S5] and observed that across all cases and replicates, the TM-scores of structures before and after MD simulation runs were greater than 0.5 and therefore roughly maintained the same fold, though some secondary structure variation was observed.

As an orthogonal method to assess designability of structures, we design sequences from sampled backbones (prior to `FastDesign`) with the ESM-IF1 [3] sequence design model, AlphaFold2 structure prediction network [1], and the self-consistency TM (scTM) metric from [16] [Figure 3F]. We observe that 50.5% of the generations are designable given by scTM > 0.5, where a sequence (and predicted structure) with the same fold as the starting backbone can be generated. This is significantly higher than 11.8% reported in [16], which also faces helix chirality issues that further affects backbone designability. Chirality is a non-issue in our case since we use L-alanines in backbone minimization, and therefore all generations maintain proper handedness. We present a few generated structures in Figure 3G, and structures with TM-score < 0.5 and Rosetta energy per residue < -2.5 in Figure 3H, which represent high-fidelity structures with novel folds not found in the training set.

Conditional generation

To assess the model performance on conditional generation, we mask short spans (<8 residues) of different secondary structures (alpha helix, beta sheet, loop) in different structures and generate 50 designs for each case [Figure 4]. We superimpose *all* 50 designs onto the reference structure and measure RMSD and proportion of correct secondary structure to assess whether the model can reasonably inpaint the original secondary structure into the masked region. We observe that alpha helices and loops are easily inpainted by the model given that almost all generations carried the desired secondary structure. We also notice that beta-sheet formation is less consistent - though 49/50 designs still carried at least one beta-sheet annotation - than helix or loop generation, reiterating the difficulty of beta-sheet formation. Nonetheless, we observe that ProteinSGM can inpaint native secondary structural motifs back into highly constrained regions, suggesting that the model has learned biophysical constraints of structures.

Next, we present two practical protein design test cases for domain and scaffold inpainting. We used a recently published *de novo*-designed structure (PDB 2KL8) and masked out one helix domain of length 20 for input to the unconditionally trained model [Figure 5A]. We observed that across most generations, the model inpaints an alpha helix to the masked region. This suggests that the model has learned, given the global structure constraints, that a helix can reasonably fit inside this pocket albeit with slight structural differences. This facilitates sampling structures with near-native topologies to optimize a functional property of interest, a central task in protein design. PDB 7MRX represents the bacterial barstar-barnase complex, an extensively-studied protein complex for its tight binding kinetics. Barstar inhibits the barnase ribonuclease with a helix-loop domain of length 22, for which we sought to design novel scaffolds of varying length [Figure 5B]. We generated 100 designs of lengths 60, 85 (native length of barstar), and 128, and observed that the generations were diverse and maintained the desired helix-loop functional site of barstar, though the helix-loop domain in some generations were buried within its core or incorporated into other secondary structure domains. We also noticed that generations of length 60 generally exhibited higher motif RMSDs, most likely since the motif occupies a greater proportion of the structure, and therefore Rosetta minimization perturbs the motif to a greater degree when searching for low-energy structures. Nevertheless, this opens doors for the use of generative models to design variable-length proteins that maintain the native functional activity, such as downsizing bulky proteins for efficient cloning into gene expression systems. One potential application of scaffold inpainting for future exploration is the scaffolding of two disparate functional sites to generate synthetic bispecific proteins, which can be accomplished with ProteinSGM by imputation of scaffolds given two functional site descriptions.

Discussion

This work presents one of the first applications of diffusion models to protein design - and the first to use the continuous-time SDE framework - that generates viable structures and can inpaint realistic backbones. The model learns to generate realistic 6D coordinates for which a structure is minimized with Rosetta. We obtain structures of native-like Rosetta energies that indicate structural plausibility, and observe that the generated structures are novel and diverse by TM-score analysis. We apply the model for conditional generation, showing that the model can inpaint realistic domains and present novel scaffolds of varying lengths given functional site information.

Despite its high sample quality, one shortcoming of this approach is its computational cost, since sampling with the continuous-time diffusion model require many forward passes through the score network for solving the reverse SDE, and Rosetta relies on expensive MCMC procedures to traverse the energy landscape and find a local minima corresponding to a low-energy structure. For high-throughput assessments, we suggest using a lower number of trajectories and/or `PackRotamersMover` for rotamer design, which significantly decreases compute cost with a marginal increase in Rosetta energy of the final structures. Though here we design the full-length protein from polyalanines, we note that for inpainting tasks where partial sequence and structure information are known, the Rosetta pipeline can easily be adapted to specifically design the inpainted region and retain the native domain(s). We also provide results on an alternative approach to sequence and full-atom structure generation from backbones by employing sequence design followed by AlphaFold2 prediction as in [16]. This allows one to circumvent the expensive `FastDesign` step and only require the relatively inexpensive

`MinMover` for backbone generation from 6D coordinates, significantly increasing throughput.

Ultimately, end-to-end diffusion models for backbone, sequence, and rotamer design would be ideal since such models can directly learn the full protein structure with both sequence and side-chain information without the use of auxiliary methods such as Rosetta and AlphaFold. A promising early approach with discrete diffusion models was recently described in [15], which uses independently trained backbone, sequence, and rotamer diffusion models to sequentially generate the full-atom structure and sequence given coarse structural constraints. However, alternative approaches such as direct 3D coordinate generation with (E)3 equivariant networks [20] as in [16] extended to sequence and all-atom generation may be explored. One drawback of current generative models is the inability to model proteins larger than 256 residues, which limits practical applicability since the average eukaryotic protein length is > 400 residues. Moreover, current models are focused on scaffolding and inpainting single-chain structures and fail to capture inter-chain interactions, which may be especially useful for novel drug discovery (i.e. epitope-antibody binding). Hence, scalability of diffusion models to larger structures and multi-chain generation remain a significant challenge for future work.

Methods

Dataset curation

We use the CATH [5] 4.3 95% sequence similarity dataset to reduce redundancy and potential bias in specific folds. All 48,949 unique chains are filtered by 1. sequence length s of $40 \leq s \leq 128$, and 2. presence of all N, Cα, and C backbone heavy atom coordinates to yield 10,361 structures, each with one or more CATH-classified folds.

Score-based generative modeling

For this work, we use the continuous-time framework of score-based generative models with stochastic differential equations (SDE). Here we provide a brief overview of the method - for more detailed information, please refer to Song *et al.* [14].

The forward noising process, given some data $\mathbf{x}$ and time t, is defined by the following general-form SDE:

$$\delta \mathbf{x} = \mathbf{f}(\mathbf{x},t)\delta t + g(t)\delta \mathbf{w},$$

for $\mathbf{f}(\cdot,t)$ is the drift coefficient, $\mathbf{g}(t)$ is the diffusion coefficient, and $\delta \mathbf{w}$ is white Gaussian noise. Intuitively, input data $\mathbf{x}$, given a small time step δt, is noised by normally distributed random values with mean $\mathbf{f}(\mathbf{x},t)\delta t$ and variance $\mathbf{g}^2(t)\delta t$. Given a forward SDE, a corresponding reverse-time SDE can be formulated that requires knowledge of the score $\nabla_x \log p_t(\mathbf{x})$, or the gradient of the log probability density with respect to data, defined as follows:

$$\delta \mathbf{x} = \left[\mathbf{f}(\mathbf{x},t) - g^2(t)\nabla_x \log p_t(\mathbf{x})\right] \delta t + g(t)\delta \mathbf{w}$$

Therefore, once we define the forward noising process and learn an approximate score function, we can solve the reverse-time SDE and denoise random samples from a prior distribution into realistic samples. We use the VESDE (Variance Exploding SDE) discretization corresponding to the score matching objective with Langevin MCMC sampling and defined by the following SDE:

$$\delta \mathbf{x} = \sqrt{\frac{\delta[\sigma^2(t)]}{\delta t}} \delta \mathbf{w},$$

where $\sigma(t) = \sigma_{\min}(\sigma_{\max}/\sigma_{\min})^t$ given a user-defined lower and upper bound noise variance $\sigma_{\min}$ and $\sigma_{\max}$, respectively. To solve the reverse-time SDE, we estimate the score $\nabla_x \log p_t(x)$ with a score network $s_\theta(x,t)$. The score network is trained using a weighted denoising score matching objective [21], defined as follows:

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

where $\lambda(t)$ is a positive weighting function, and $p_t(x(t)|x(0))$ corresponds to a perturbation kernel that perturbs clean sample $x(0)$ to noisy sample $x(t)$. For a Gaussian kernel, $\nabla_x \log p_t(x(t)|x(0)) = [x(0) - x(t)]/\sigma^2(t)$.

We retained the RefineNet [22] architecture for score estimation as used in [14] with multi-head self-attention in the 16x16 resolution block. We reasoned that using the RefineNet architecture for the score network is justified since the task is presented as an image synthesis problem, and the input data resembles image-like data with 5 channels. The model is trained with a single NVIDIA A100 GPU using a batch size of 16 and learning rate 1×10^{-4} until loss convergence.

Sampling is performed by using Predictor-Corrector sampling (Reverse diffusion with Langevin dynamics), which uses the numerical SDE solver (Euler-Maruyama) to provide an initial prediction for a given denoising step, and is further refined using Langevin MCMC and predicted scores from the score network. We apply conditional generation for imputation of masked features akin to image inpainting in [14] for the task of protein design.

Rosetta minimization

To obtain structures from the inter-residue 6D coordinates, we use an adaptation of the trRosetta minimization protocol. While trRosetta and a previously reported VAE-based approach [23] use distograms, or probabilities of each distance/angular bin, to fit a spline function to generate smoothed constraints for Rosetta minimization, ProteinSGM directly generates the distance and angular values. We then use the HARMONIC function for d and φ and CIRCULARHARMONIC for ω and θ, with the mean corresponding to the generated value and the standard deviation set to 2.0 for d and 20° for φ, ω, and θ. This allows reproducible generation of the structure given a set of 6D coordinates while still relaxing the constraints enough to generate realistic structures [Figure S1].

Since we do not have sequence information for the generated matrices, we use the padding channel to obtain the matrix boundaries L by L of the generated matrices, and a polyalanine chain of length L for the minimization protocol. We also use the upper triangle for the d and ω matrices since they are symmetric, and do not include any constraints that are $d > 12\text{Å}$ apart.

To generate backbone from constraints, we use `MinMover` with 5 rounds of minimization and select for the pose with lowest energy using a coarse-grained centroid energy function. Each round of minimization progressively uses short, medium, and long-range constraints by sequence separation, and randomly changes the φ and ψ backbone torsional angles by up to 10 degrees. Once a tentative backbone has been generated, we use `FastDesign` for fixed-backbone design to sample different sequences and rotamers that can fit the generated backbone. Then we idealize problematic regions and perform full-atom relaxation of the structure using the matrix constraints. The entire process is repeated for 10 independent trajectories, and the lowest energy structure is selected as the final structure.

All Rosetta protocols were written with `pyRosetta` 4.0 [24].

Sequence design and AlphaFold prediction

As an alternative method to survey designability of generated backbones, we adapt the self-consistency TM (scTM) metric from [16]. After backbone selection from `MinMover`, instead of passing the backbone to `FastDesign`, we use the ESM-IF1 [3] sequence design model with temperature = 1 to sample 10 sequences for each backbone, and predict structures for each sequence with AlphaFold2 using the `model_1_ptm` parameters, 10 recycling iterations, and without any MSA input information. We calculate the TM-score between each AlphaFold-predicted structure and the starting backbone, and the highest TM-score across all 10 structures is designated as the scTM score for the given backbone. Backbones with scTM > 0.5 are considered designable structures, since at least one sequence can be designed for which a predicted structure has the same fold (TM > 0.5) as the starting backbone.

MD analysis

We use molecular dynamics (MD) simulations with `OpenMM` [25] to assess structural stability for selected structures. First, we remove all heterogens (non-amino acid residues) and replace non-standard residues with their standard counterpart. Then, we create a solvent box with 1nm padding around the structure and add Na+ and Cl- ions to neutralize the net charge of the system. We run a 200-ns MD simulation using the `AMBER14` force field and `TIP3P-FB` explicit water model, and particle mesh Ewald for long-range electrostatic interactions. To solve the equations of motion, we use the `LangenvinMiddleIntegrator` with temperature at 300K, friction coefficient at 1ps^{-1}, and 4 fs time steps. Analysis is performed with MDAnalysis [26] and VMD [27].

References

[1] J. Jumper *et al.*, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 26, 2021. DOI: 10.1038/s41586-021-03819-2.

[2] A. Strokach, D. Becerra, C. Corbi-Verge, A. Perez-Riba, and P. M. Kim, "Fast and flexible protein design using deep graph neural networks," *Cell Systems*, vol. 11, no. 4, 402–411.e4, Oct. 21, 2020. DOI: 10.1016/j.cels.2020.08.016.

[3] C. Hsu *et al.*, "Learning inverse folding from millions of predicted structures," *bioRxiv*, Apr. 10, 2022. DOI: 10.1101/2022.04.10.487779.

[4] A. Andreeva, E. Kulesha, J. Gough, and A. G. Murzin, "The SCOP database in 2020: Expanded classification of representative family and superfamily domains of known protein structures," *Nucleic Acids Research*, vol. 48, pp. D376–D382, D1 Jan. 8, 2020. DOI: 10.1093/nar/gkz1064.

[5] I. Sillitoe *et al.*, "CATH: Increased structural coverage of functional space," *Nucleic Acids Research*, vol. 49, pp. D266–D273, D1 Jan. 8, 2021. DOI: 10.1093/nar/gkaa1079.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," no. arXiv:2006.11239, Dec. 16, 2020. arXiv: 2006.11239[cs,stat].

[7] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," no. arXiv:1907.05600, Oct. 10, 2020. arXiv: 1907.05600[cs,stat].

[8] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," no. arXiv:2105.05233, Jun. 1, 2021. arXiv: 2105.05233[cs,stat].

[9] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro, "DiffWave: A versatile diffusion model for audio synthesis," no. arXiv:2009.09761, Mar. 30, 2021. arXiv: 2009.09761[cs,eess,stat].

[10] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," no. arXiv:2204.06125, Apr. 12, 2022. DOI: 10.48550/arXiv.2204.06125. arXiv: 2204.06125[cs].

[11] C. Niu, Y. Song, J. Song, S. Zhao, A. Grover, and S. Ermon, "Permutation invariant graph generation via score-based generative modeling," no. arXiv:2003.00638, Mar. 1, 2020. arXiv: 2003.00638[cs, stat].

[12] J. Jo, S. Lee, and S. J. Hwang, "Score-based generative modeling of graphs via the system of stochastic differential equations," no. arXiv:2202.02514, Feb. 16, 2022. arXiv: 2202.02514[cs].

[13] E. Hoogeboom, V. G. Satorras, C. Vignac, and M. Welling, "Equivariant diffusion for molecule generation in 3d," no. arXiv:2203.17003, Mar. 31, 2022. arXiv: 2203.17003[cs,q-bio,stat].

[14] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, “Score-based generative modeling through stochastic differential equations,” no. arXiv:2011.13456, Feb. 10, 2021. arXiv: 2011.13456[cs,stat].

[15] N. Anand and T. Achim, “Protein structure and sequence generation with equivariant denoising diffusion probabilistic models,” no. arXiv:2205.15019, May 26, 2022. arXiv: 2205.15019[cs,q-bio].

[16] B. L. Trippe *et al.*, *Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem*, Jun. 8, 2022. arXiv: 2206.04119[cs,q-bio,stat].

[17] J. Yang, I. Anishchenko, H. Park, Z. Peng, S. Ovchinnikov, and D. Baker, “Improved protein structure prediction using predicted interresidue orientations,” *Proceedings of the National Academy of Sciences*, vol. 117, no. 3, pp. 1496–1503, Jan. 21, 2020. DOI: 10.1073/pnas.1914677117.

[18] W. Kabsch and C. Sander, “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features,” *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983. DOI: 10.1002/bip.360221211.

[19] Y. Zhang, “TM-align: A protein structure alignment algorithm based on the TM-score,” *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, Apr. 11, 2005. DOI: 10.1093/nar/gki524.

[20] V. G. Satorras, E. Hoogeboom, and M. Welling, “E(n) equivariant graph neural networks,” no. arXiv:2102.09844, Feb. 16, 2022. arXiv: 2102.09844[cs,stat].

[21] P. Vincent, “A connection between score matching and denoising autoencoders,” *Neural Computation*, vol. 23, no. 7, pp. 1661–1674, Jul. 2011. DOI: 10.1162/NECO_a_00142.

[22] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” no. arXiv:1611.06612, Nov. 24, 2016. arXiv: 1611.06612[cs].

[23] Z. Lin, T. Sercu, Y. LeCun, and A. Rives, “Deep generative models create new and diverse protein structures,” *Machine Learning in Structural Biology, NeurIPS*, 2021.

[24] S. Chaudhury, S. Lyskov, and J. J. Gray, “PyRosetta: A script-based interface for implementing molecular modeling algorithms using rosetta,” *Bioinformatics*, vol. 26, no. 5, pp. 689–691, Mar. 1, 2010. DOI: 10.1093/bioinformatics/btq007.

[25] P. Eastman *et al.*, “OpenMM 7: Rapid development of high performance algorithms for molecular dynamics,” *PLOS Computational Biology*, vol. 13, no. 7, e1005659, Jul. 26, 2017. DOI: 10.1371/journal.pcbi.1005659.

[26] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, “MDAnalysis: A toolkit for the analysis of molecular dynamics simulations,” *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011. DOI: 10.1002/jcc.21787.

[27] W. Humphrey, A. Dalke, and K. Schulten, “VMD – Visual Molecular Dynamics,” *Journal of Molecular Graphics*, vol. 14, pp. 33–38, 1996.

Author Contributions

J.S.L. and P.M.K. conceptualized the work, J.S.L. developed the results and performed the analyses, and J.S.L. and P.M.K. wrote the manuscript. P.M.K. supervised the work and acquired funding.

Competing Interests

P.M.K. is a co-founder and consultant to multiple companies, including Resolute Bio, Oracle Therapeutics and Navega Therapeutics and serves on the scientific advisory board of ProteinQure. J.S.L declares no competing interests.

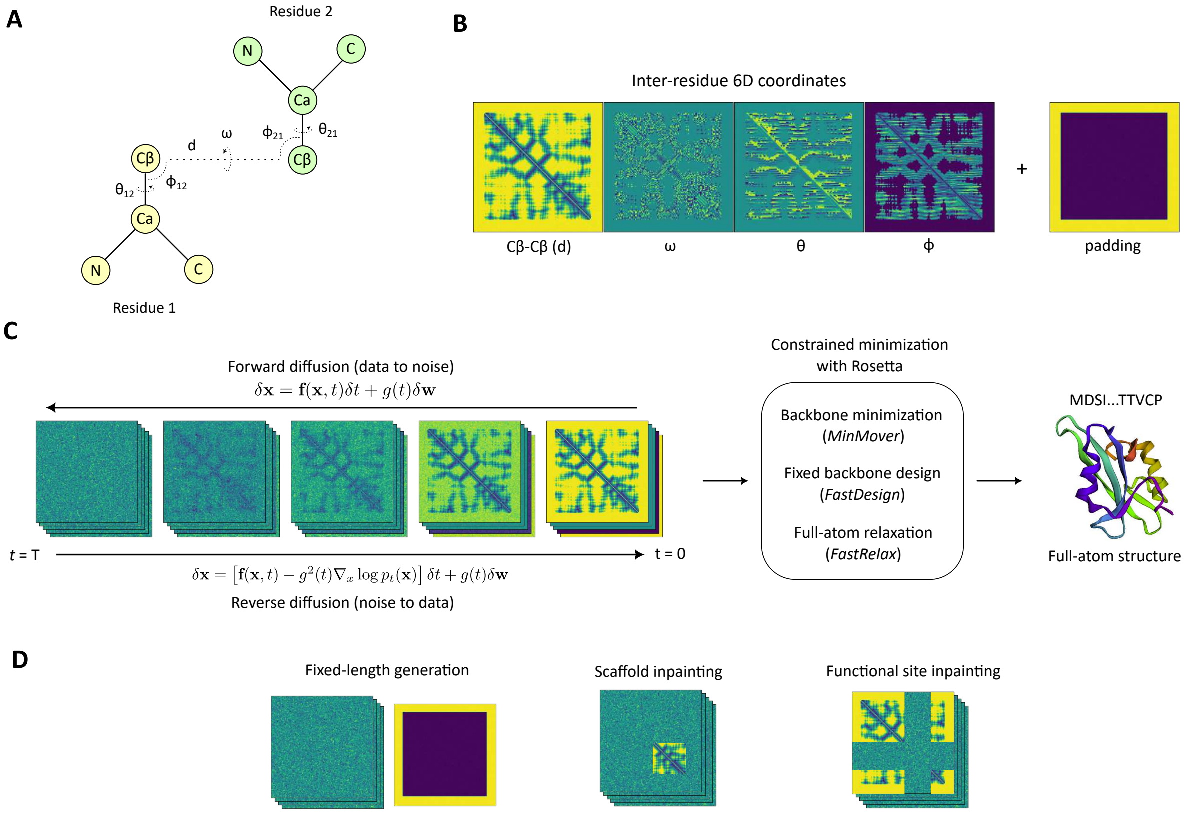
A
Residue 2
N
C
Ca
ϕ_{21}
θ_{21}
ω
d
Cβ
Cβ
θ_{12}
ϕ_{12}
Ca
N
C
Residue 1
B
Inter-residue 6D coordinates
Cβ-Cβ (d)
ω
θ
ϕ
+
padding
C
Forward diffusion (data to noise)
$\delta \mathbf{x} = \mathbf{f}(\mathbf{x}, t)\delta t + g(t)\delta \mathbf{w}$
t = T
t = 0
$\delta \mathbf{x} = \left[\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_x \log p_t(\mathbf{x})\right]\delta t + g(t)\delta \mathbf{w}$
Reverse diffusion (noise to data)
Constrained minimization with Rosetta
Backbone minimization (*MinMover*)
Fixed backbone design (*FastDesign*)
Full-atom relaxation (*FastRelax*)
MDSI...TTVCP
Full-atom structure
D
Fixed-length generation
Scaffold inpainting
Functional site inpainting

Figure 1. Model overview. (A) Inter-residue 6D coordinates d, ω, θ, ϕ between two residues. (B) Input features (6D coordinates and padding channel) used to describe a given protein structure. (C) A diffusion model is trained to generate realistic samples from noise by learning a reverse "denoising" process given the forward diffusion process that perturbs data to noise. The generated 6D coordinates are used as input for constrained minimization with Rosetta, which performed fixed-backbone design and full-atom relaxation to yield a protein structure corresponding to the 6D coordinate constraints. (D) A few applications for protein design using conditional generation with masked input features.

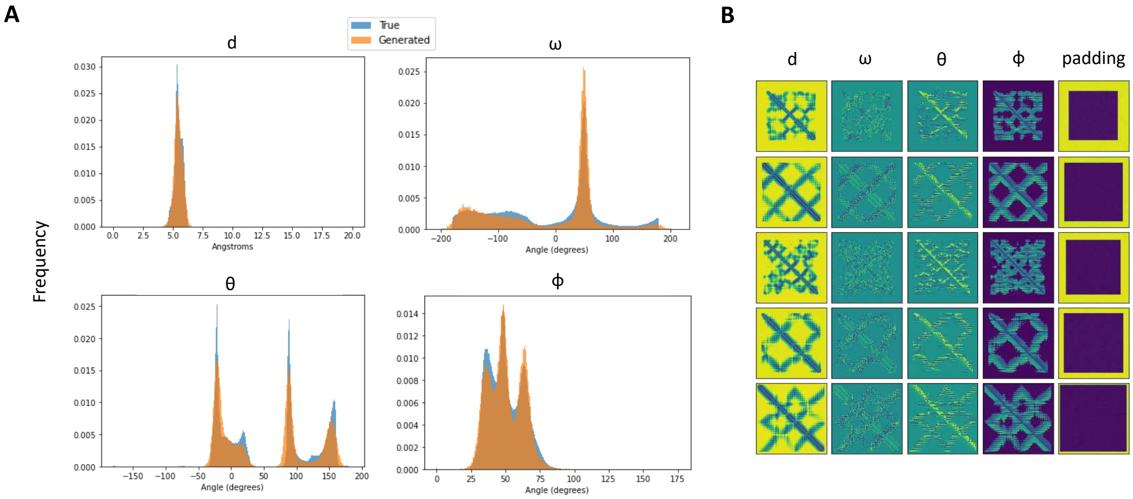

Figure 2. 6D coordinate analysis. (A) 1000 samples were generated with the model and compared to features in the training set. d, ω, θ, and ϕ distributions of true (blue) vs generated (orange) samples show significant overlap, suggesting that the model has learned native-like constraints of interresidue 6D coordinates. (B) A few examples of the generated matrices.

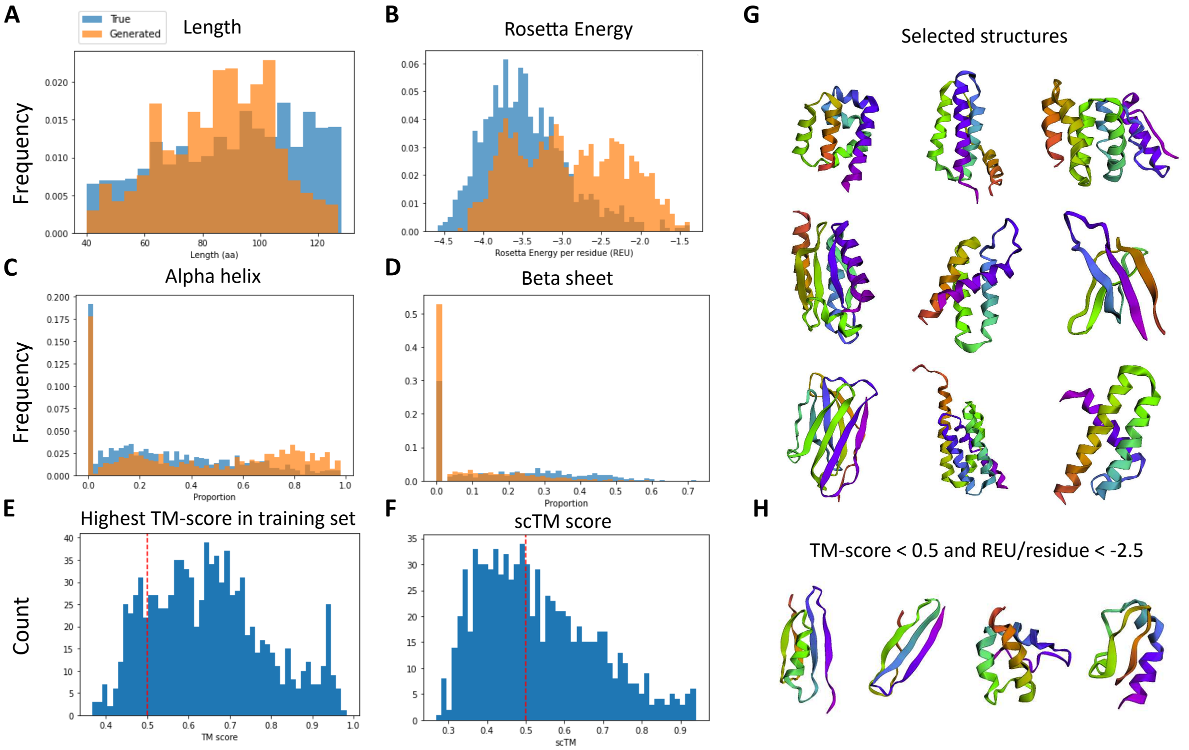

Figure 3. Structural analysis. 966 generated samples were minimized, designed, and relaxed with Rosetta to produce full-atom structures. (A) We see a clear concordance of the length distribution between generated samples and the training set, though the model generates longer proteins with reduced frequency. (B) Rosetta energies of the structures when compared to relaxed native structures show significant overlap with all generations exhibiting negative Rosetta energies, indicative of structural viability. Proportions of (C) alpha helices and (D) beta sheets were measured between the generations and samples from the training set. The model tends to generate more helices and less beta sheets than native structures, which is expected since beta sheets are more difficult to generate due to long-range structural constraints. (E) Max TM-scores were measured for each generation compared to all structures in the training set as a quantitative measure of generalization to novel folds not found in the training set. (F) An orthogonal assay with the ESM-IF1 sequence design model, AlphaFold2, and the scTM metric shows that 50.5% of the structures (scTM > 0.5) can be realized by a protein sequence. (G) A few examples of generated structures, including (H) ones that have novel folds (TM-score < 0.5) and viable (Rosetta energy per residue < -2.5). More generations can be found in Figure S2.

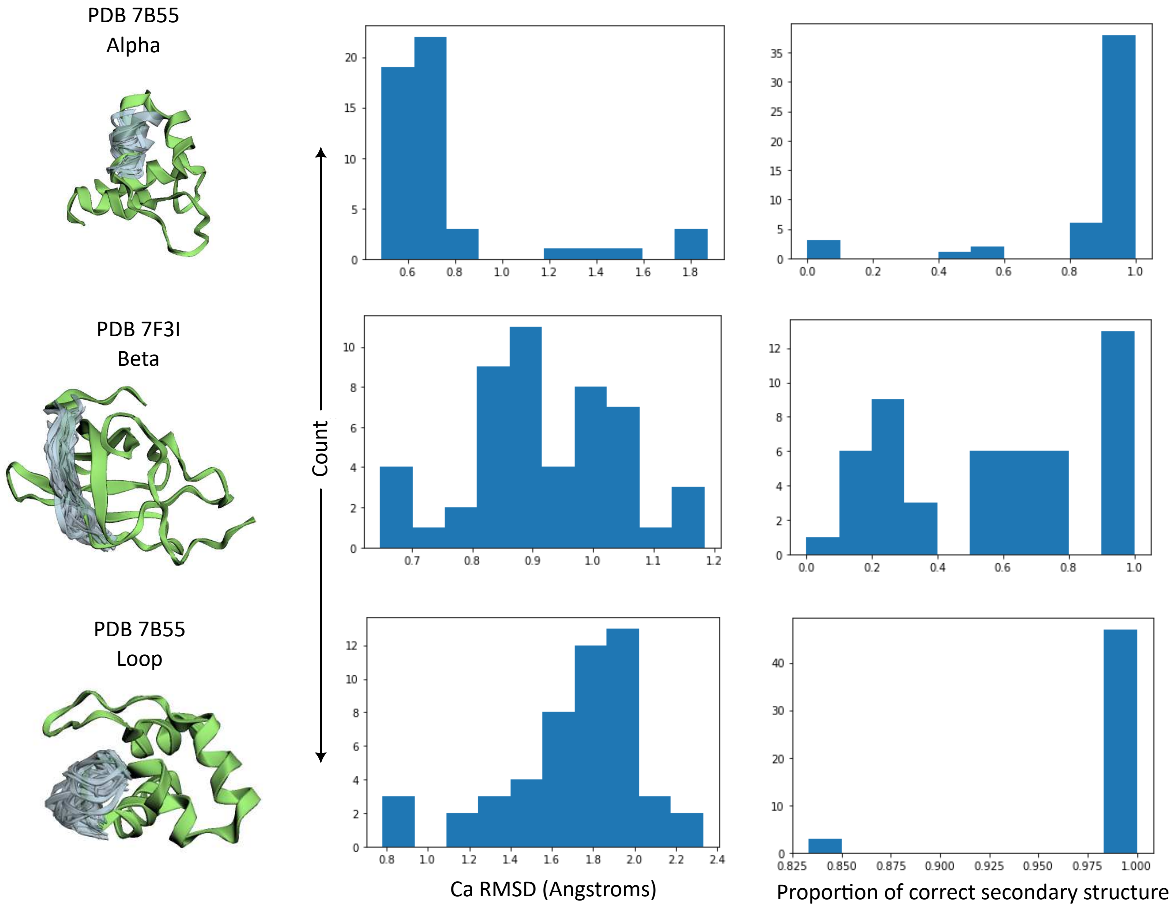

Figure 4. 50 designs were generated for three highly constrained inpainting cases to assess the model's ability to inpaint the native secondary structure. We observe that alpha helices and loops more consistently inpainted than beta sheets by the model given the higher proportions of recovered secondary structures. We also see that sampling different loops causes a greater increase in overall Ca RMSD than other secondary structures due to its increased flexibility, therefore perturbing the structure to a greater extent.

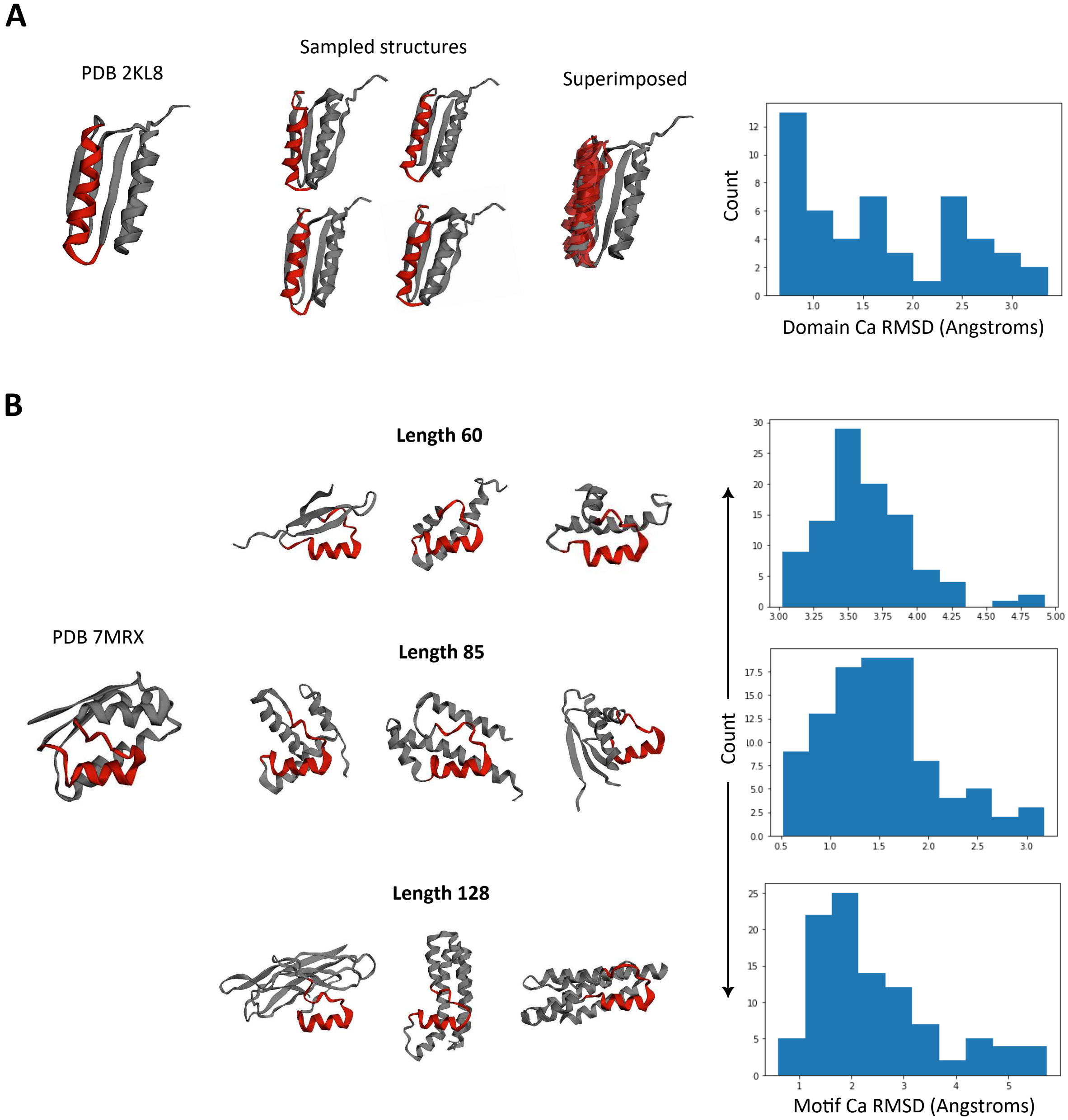

Figure 5. Protein design test cases. (A) As an example of domain inpainting, we ask the model to inpaint one helical domain of PDB 2KL8. Across 50 designs, the model can inpaint slightly varied helices of similar shape to the original structure, where some designs show close domain RMSDs to the native structure while others are more varied. (B) When tasked with inpainting novel scaffolds for a given functional site of PDB 7MRX, the model can generate viable scaffolds of varying length that retains the canonical barstar helix-loop domain.

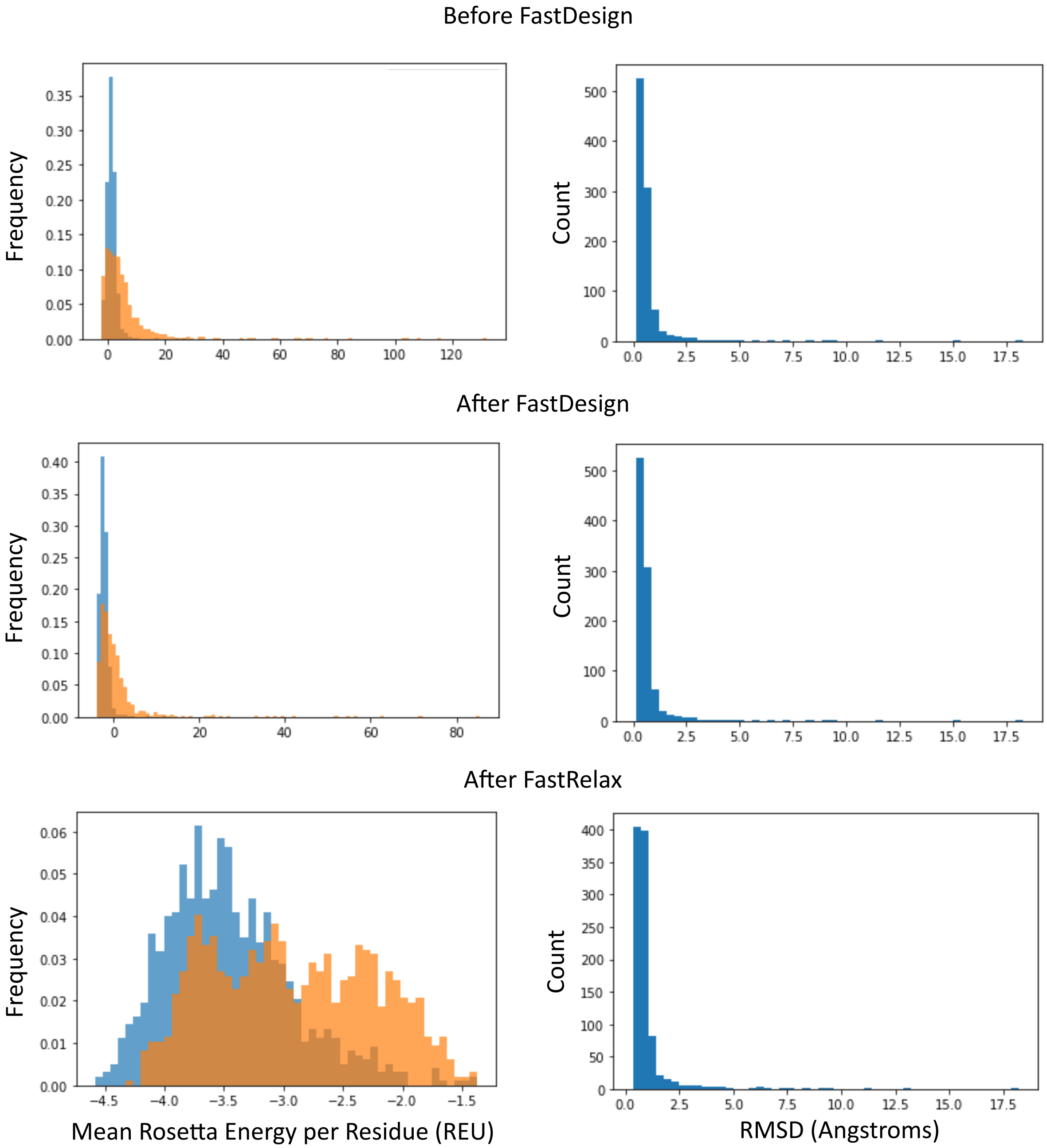

Figure S1. Rosetta energies and Cα RMSDs with Rosetta pipeline. (left) Rosetta energy before `FastRelax` remains relatively high for generated structures, but full-atom relaxation with constraints allows effective energy minimization to bring energies closer to the true distribution. (right) Minimized structures from 6D coordinates were compared with their native structures to assess fidelity of the Rosetta protocol. We observe that across all steps, mean Cα RMSDs are lower than 1Å, suggesting the protocol is suited for reproducible generation of structures from 6D coordinates.

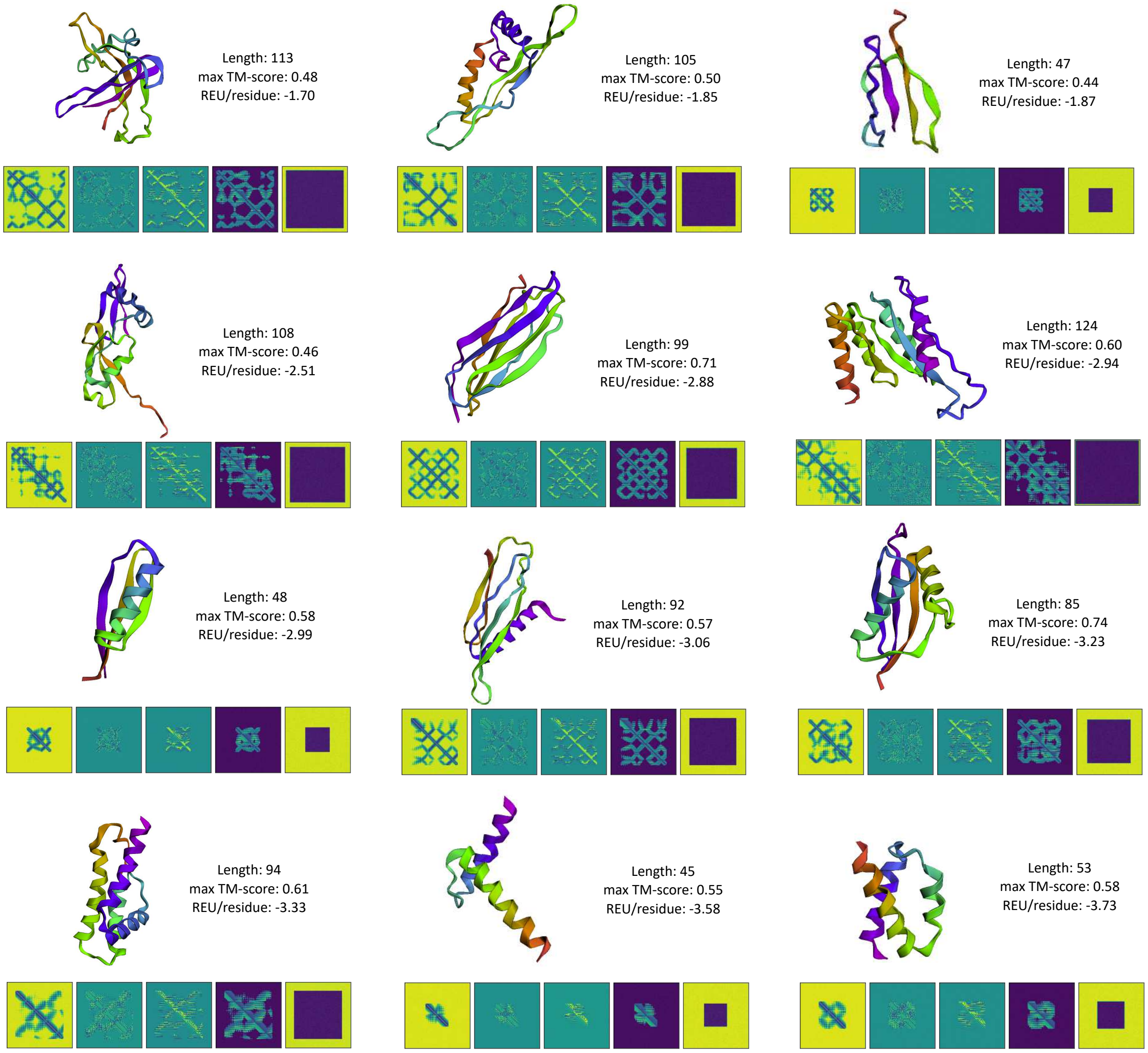

Figure S2. Examples of generated 6D coordinates and corresponding structures. 12 randomly selected structures are ranked by decreasing Rosetta energy (top-left to bottom-right).

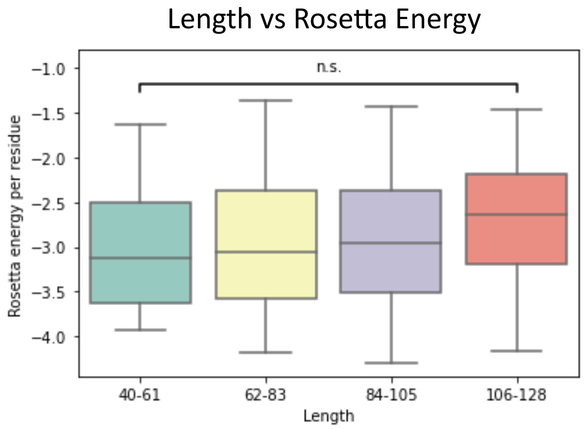

Figure S3. All generated structures are separated into four categories by length, and the Rosetta energy distributions are compared to assess the model performance on variable-length structures. We do observe a noticeable increase in Rosetta energy for longer structures, but a two-sample Kolmogorov-Smirnov test between Rosetta energies of length 40-61 and those of length 106-128 was statistically insignificant ($p > 0.05$), suggesting that the two energy distributions come from the same distribution and therefore Rosetta energies are independent of protein length.

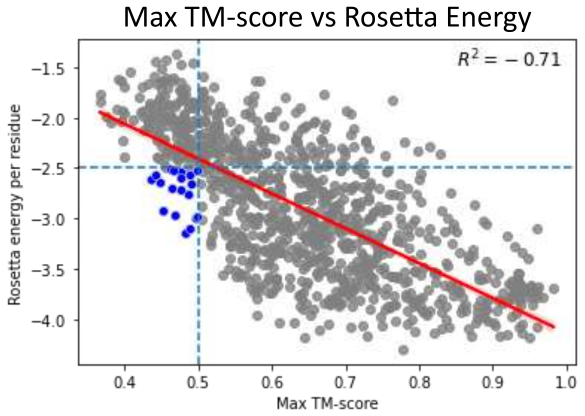

Figure S4. All generated structures are plotted by highest TM-score found in training set by Rosetta energy per residue. There is a strong negative correlation of $R^2 = 0.71$, indicating that stronger similiarity to a native structure generally results in higher structural viability. However, we notice a subset of generated structures that have TM-score < 0.5 and Rosetta energy per residue < -2.5, which indicates that the model can generalize to high-fidelity structures not found in the training set.

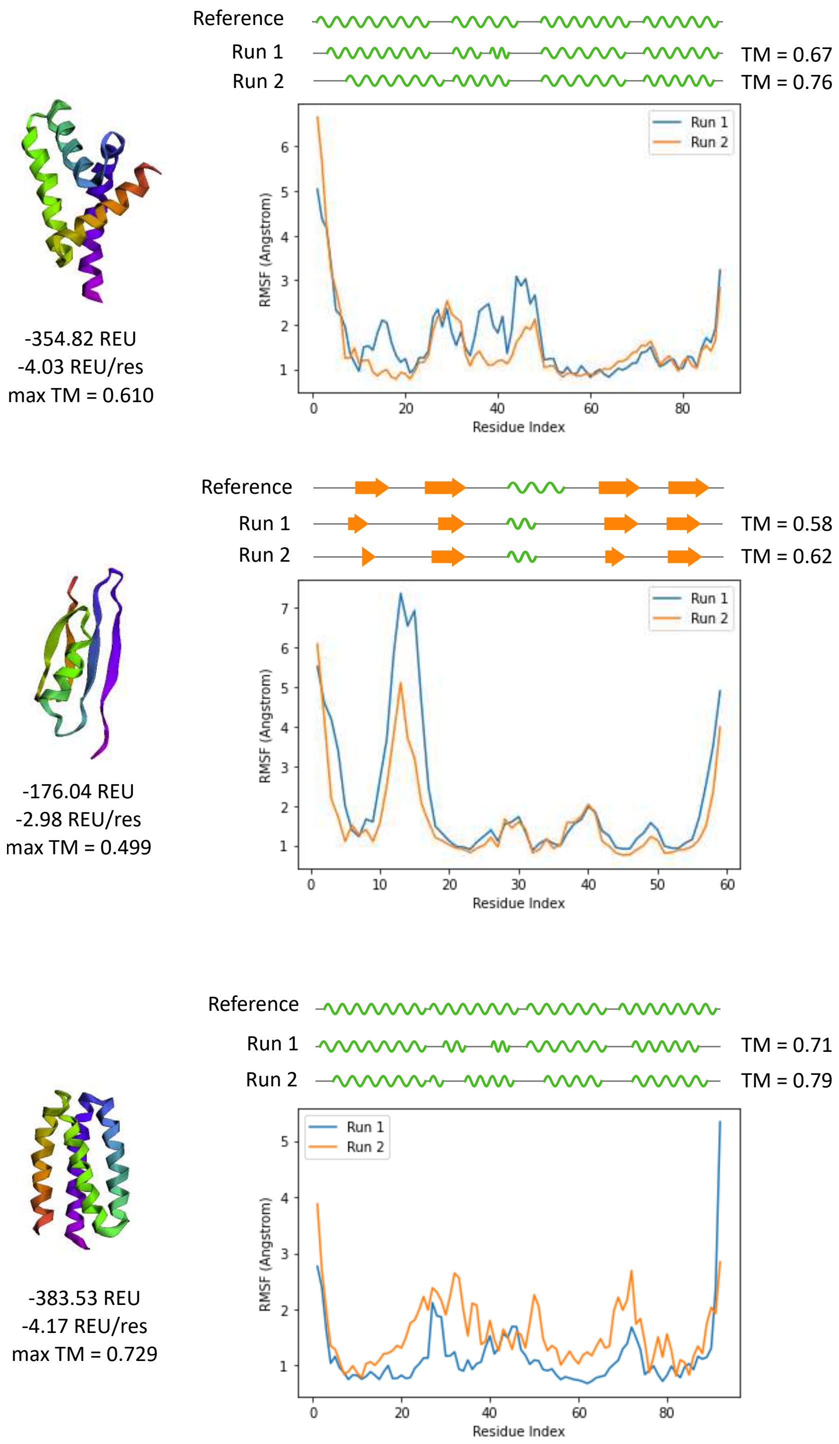

Figure S5. MD simulation on three structures to assess structural stability. We run two replicates of MD simulations on two helical structures (top and bottom) and one alpha-beta structure with max TM-score < 0.5 (middle) to assess fold stability. We observe that generally, the overall secondary structure is maintained across all samples with RMSF peaks corresponding to unstructured domains. Green lines represent alpha helices, and orange arrows represent beta sheets, and the black lines represent unstructured domains.