CONTROLLING PROTEIN LANGUAGE MODELS WITH SYNTHETIC STRUCTURE DATA AUGMENTATION

Alex J. Lee[1,2,†], **Sarah Alamdari**[1], **Chentong Wang**[1,3†], **Reza Abbasi-Asl**[2], **Kevin K. Yang**[1], & **Ava P. Amini**[1,*]
[1]Microsoft Research, Cambridge, MA, USA
[2]University of California, San Francisco, San Francisco, CA, USA
[3]Westlake University, Hangzhou, Zhejiang, China
[†]Work done principally during an internship at Microsoft Research
[*]Corresponding author; email ava.amini@microsoft.com

ABSTRACT

The goal of *de novo* protein design is to leverage natural proteins to design new ones. Deep generative models of protein structure and sequence are the two dominant *de novo* design paradigms. Structure-based models can produce highly novel proteins but are constrained by data to produce proteins with a narrow range of topologies. Sequence-based design models produce more natural samples over a wider range of topologies, but with reduced novelty. Here, we propose a structure-based synthetic data augmentation approach to combine the benefits of structure and sequence in generative protein language models. We generated and characterized 240,830 *de novo* backbone structures and used these backbones to generate 45 million sequences for data augmentation. Models trained with structure-based synthetic data augmentation generate a shifted distribution of proteins that are more likely to express successfully in *E. coli*. We release the trained models as well as our complete synthetic dataset, BackboneRef.

1 INTRODUCTION

Deep generative models of proteins seek to generate new and novel proteins with desired functions through efficient exploration of a large design space. To design proteins with novel functionality, models must balance exploring new areas of design space, satisfying constraints learned from natural proteins, and maintaining the breadth of sequence, structure, and function seen in nature.

Deep generative models of proteins can be divided into structure-based and sequence-based approaches. Structure-based models excel at generating functional proteins with no homology to any natural sequence, and their generations are often highly stable (Watson et al., 2023; Ingraham et al., 2023). However, structure-based models are constrained by sparse and biased training data (Berman et al., 2000), consisting of static snapshots of protein structures. As a result, they have been shown to omit entire classes of proteins and structural elements from their generations, particularly disordered regions (Alamdari et al., 2023). In contrast, sequence-based models can utilize a much larger amount and diversity of training data and correspondingly sample distributions that better maintain the breadth of functions found in nature (Madani et al., 2023; Alamdari et al., 2023). However, they struggle to generate proteins that are both functional and truly novel. Recent efforts to combine the benefits of both approaches have focused on co-generation modeling strategies that utilize both sequence and structural information (Wang et al., 2024a; Hayes et al., 2025; Qu et al., 2024; Campbell et al., 2024; Wang et al., 2024b; Chu et al., 2024; Lu et al., 2024).

In this work, we propose a synthetic data augmentation approach to bridge the complementary benefits of protein sequence and structure information in deep generative models of proteins (Fig. 1A). We first used a structure-based generative model to sample 240,830 backbones unconditionally and characterized their novelty, diversity, and quality in order to evaluate their suitability for data augmentation (Fig. 1B-C). We performed fixed-backbone sequence design on filtered subsets of these backbones to generate a series of structure-based synthetic datasets, and then trained protein language models on mixtures of natural and synthetic sequence data by combining UniRef50 (Suzek

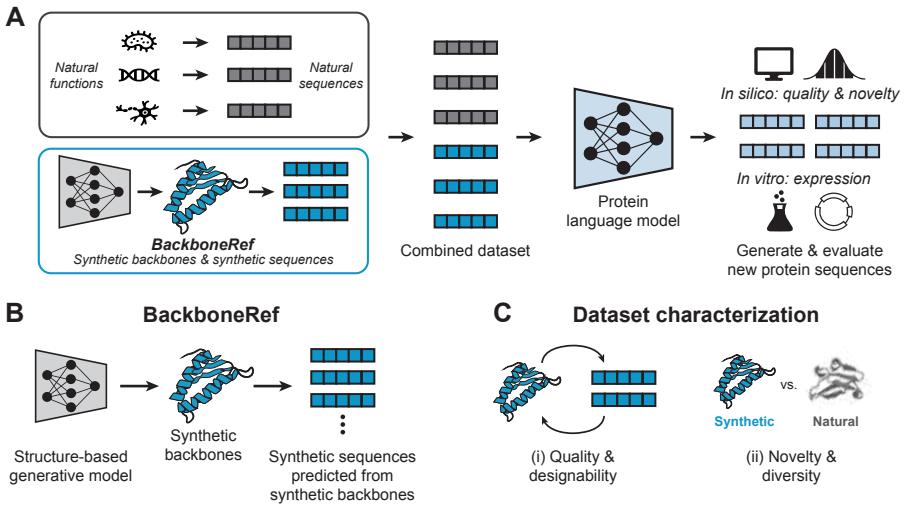


Figure 1: **BackboneRef enables structure-based synthetic data augmentation for protein language models.** **(A)** Pipeline for synthetic data augmentation and model training. Sequences of natural proteins (grey) and BackboneRef structure-based synthetic sequences (blue) are integrated into a combined dataset to train protein language models. Generated sequences are evaluated both *in silico* and *in vitro*. **(B)** BackboneRef includes 240,830 *de novo* designed backbones and 45,553,550 fixed-backbone designed protein sequences. **(C)** Characterization of the quality, novelty, and diversity of BackboneRef proteins at the structure and sequence levels.

et al., 2015) with each synthetic dataset. Models trained with structure-based synthetic data augmentation were slightly worse at modeling natural sequences. However, they generate a shifted distribution of proteins that are more likely to express successfully in *E. coli*. We release our synthetic dataset, called BackboneRef (BBR), including the original backbones, designed sequences, and predicted structures. Our findings demonstrate that structure-based synthetic data augmentation can be used to shift the output distribution of protein language models in desirable ways, providing a highly effective addition to protein engineering workflows.

2 Results

2.1 BackboneRef: a structure-based synthetic protein dataset

We hypothesized that structure-based synthetic data augmentation could combine the complementary benefits of protein structure and sequence information into sequence-only protein language models (Fig. 1A). To this end, we created a large-scale dataset of 240,830 synthetic backbone structures – to our knowledge, the largest set of generated structures to date – by sampling backbones unconditionally from RFdiffusion (Watson et al., 2023) (Fig. 1B). Structures were sampled according to the length distribution of UniRef50 (UniProt Consortium, 2025) to recapitulate the lengths of natural proteins but with a minimum length of 40 and maximum length of 512 for computational efficiency. Secondary-structure characterization revealed an enrichment of helical elements, versus disordered and β-stranded regions, in the synthetic backbones relative to natural structures (Fig. S1), consistent with prior reports (Lu et al., 2025; Alamdari et al., 2023; Lane, 2023).

We performed fixed-backbone sequence design for all 240,830 synthetic backbones, sampling 10 sequences per backbone at temperature 0.1 using ProteinMPNN (Dauparas et al., 2022) to produce c.a. 2.4 million sequences, which we use to characterize the dataset (Fig. 1B-C). Structures were predicted for all resulting sequences with OmegaFold to evaluate individual backbone quality via self-consistency RMSD (scRMSD) (Fig. 2A). 53.0% of the backbones exhibited average scRMSD below 2.0 (Fig. 2B). We then chose the sequence and predicted structure with the lowest scRMSD for each backbone to characterize the novelty and diversity of BackboneRef proteins.

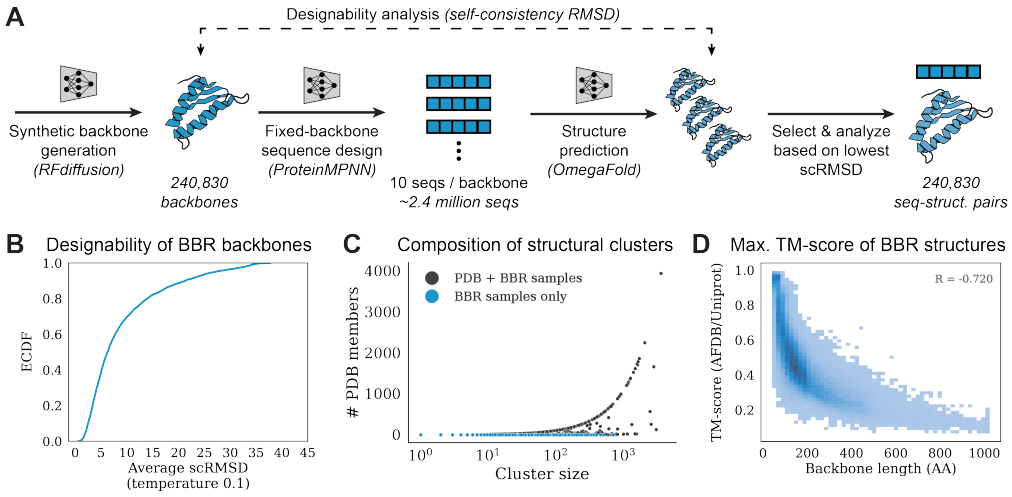


Figure 2: **BackboneRef structures are designable and novel. (A)** Workflow for generation, analysis, and filtering of BackboneRef synthetic structures and sequences. **(B)** Empirical CDF (ECDF) of average scRMSD per BackboneRef (BBR) backbone over 10 designed sequences (temperature 0.1) and OmegaFold-predicted structures, with the scRMSD computed between the original backbone and each predicted structure. **(C)** Scatterplot of cluster size (x-axis) vs. number of PDB members (y-axis) for each inferred structural cluster in the "BBRef + PDB" dataset (n=105,506 clusters). Each cluster is colored by whether it contains both natural PDB and synthetic BackboneRef samples (grey) or all synthetic BackboneRef samples (blue). (**D**) Distribution of maximum TM-score (y-axis), computed between each backbone's lowest scRMSD predicted structure searched against the AFDB/UniProt database using Foldseek, versus backbone length (x-axis).

To characterize the novelty of BackboneRef as a whole, we assessed the extent to which BackboneRef structures include folds not present in natural structures from the PDB (Berman et al., 2000). We clustered the lowest scRMSD predicted structures with the PDB using Foldseek (van Kempen et al., 2024) (Fig. 2C; Table S1). While 99.0% of clusters contained exclusively synthetic or exclusively natural structures, the 1.0% of mixed-membership clusters included 27.8% of synthetic structures. Nevertheless, the 240,830 BackboneRef synthetic backbones yielded 84,156 novel clusters with 277,699 structures (Table S1), indicating that generative models trained on the PDB can still produce large numbers of distinct, novel structures. To quantify whether BackboneRef samples were saturating in diversity, we conducted a rarefaction analysis. We randomly sub-sampled the combined BackboneRef and PDB dataset at different frequencies, performing one analysis by sub-sampling PDB and BackboneRef samples and one analysis only resampling BackboneRef structures. We then computed the number of distinct structural clusters present in each subsample derived from the original Foldseek clustering(Fig. S2). Cluster diversity did not saturate in either curve, suggesting that scaling BackboneRef could lead to additional novel structures.

We also evaluated the novelty of individual BackboneRef structures by using Foldseek to find the closest structural match in the 'AFDB/UniProt' database (Varadi et al., 2024), which comprises AlphaFold2 (Jumper et al., 2021) predictions of every protein in UniProt (UniProt Consortium, 2025) (Fig. 2D). 57.3% of synthetic BackboneRef structures had a maximum TM-score less than 0.5 to any natural structure (Fig. 2D), consistent with the observation that 62.7% of synthetic structures were in BackboneRef-only clusters (Fig. 2C). Together, these results suggest that BackboneRef synthetic structures are designable, diverse, and novel.

2.2 Augmenting protein language models with BackboneRef

Having characterized the quality and novelty of BackboneRef 's synthetic backbones, we next sought to utilize BackboneRef to augment sequence-only protein language models with structure-based synthetic data (Fig. 1A). While we evaluated backbones using sequences designed at $T = 0.1$ following the convention in structure-based protein design, for augmentation, we aimed to select a temperature for fixed-backbone sequence design that balanced sequence diversity and quality. To

choose this temperature, we designed sequences for 5,000 randomly-selected backbones at $T = \{0.1, 0.2, 0.3, 0.4, 0.5, 0.8, 1\}$ and evaluated the designability of the structures predicted from those sequences. We observed a steep increase in scRMSD between $T = 0.2$ and $T = 0.3$ (Fig. S3). Therefore, we selected $T = 0.2$ as the temperature for augmentation and designed 185 sequences per backbone at $T = 0.2$, yielding an additional 45,553,550 synthetic sequences. Duplicate sequences were filtered out and resampled so that the number of sequences per backbone was roughly equal.

To evaluate the effect of synthetic data from BackboneRef on protein language model training, we filtered BackboneRef for quality and novelty to assess how these attributes affected downstream model performance. For quality filtering, we selected backbones with average scRMSD below 2Å, retaining 127,633 backbones (BBR-sc). For novelty filtering, we included backbones with a maximum TM-score to natural proteins of at most 0.5, retaining 138,044 backbones (BBR-novel).

We then quantified the distribution-level divergence from natural proteins for each set of associated synthetic sequences. We used the Fréchet ProtBert distance (FPD) (Alamdari et al., 2023), the earth-mover's distance between ProtBert (Elnaggar et al., 2022) embeddings for the different sequence sets, to quantify this divergence. A lower FPD indicates closer distributional similarity. Compared to the unfiltered dataset, filtering for novelty reduced FPD (Table 1) while filtering for self-consistency increased FPD.

Table 1: **Distributional divergence of BackboneRef sequences relative to natural protein sequences,** measured by Fréchet ProtBert Distance (FPD) to UniRef50 (lower is more similar).

Dataset	FPD
BBR-unfiltered	4.78
BBR-novel	4.34
BBR-sc	5.01

We trained 170M-parameter autoregressive protein language models on UniRef50 (UR50) augmented by 10 million sequences from BackboneRef-unfiltered (BBR), BackboneRef-novel (BBR-novel), or BackboneRef-sc (BBR-sc), (Table 2). We used a recent hybrid transformer-state space model architecture, Jamba (Lieber et al., 2024). We trained baseline models on UniRef50 alone and on UniRef90 with sampling by UniRef50 cluster, roughly equivalent to augmenting UniRef50 with natural sequences. Note that UniRef90 has roughly 2.5x more sequences than the BackboneRef-augmented datasets.

Table 2: **Cross-entropy (CE) and FPD for protein language models trained on various datasets.** The CE is computed on held-out UniRef50 sequences. The FPD is between held-out UniRef50 sequences and unconditional generations from the model trained on the listed dataset.

Training dataset	CE	FPD
UR90	2.44	0.70
UR50	2.45	0.74
UR50 + BBR-unfiltered	2.46	1.04
UR50 + BBR-novel	2.46	1.00
UR50 + BBR-sc	2.46	0.97

Structure-based synthetic data augmentation resulted in only slightly worse cross-entropy on the UniRef50 validation set, indicating that these models still learn the natural sequence distribution (Table 2). However, the FPD was much higher for sequences generated from models trained using synthetic data augmentation (Table 2). The model trained on data filtered for novelty (BBR-novel) resulted in the highest generation FPD. This model and the unfiltered synthetic data model (BBR-unfiltered) produced higher FPD than the self-consistency filtered dataset. We note that the FPD of generations was much lower than the FPD of the datasets themselves (Table 1). The observed differences in generation FPD suggest that structure-based synthetic data augmentation can result in a distribution shift in the generated sequences.

2.3 DATA AUGMENTATION WITH BACKBONEREF IMPROVES IN VITRO EXPRESSION RATES

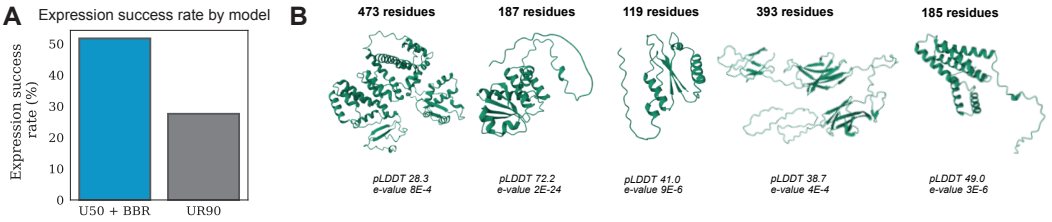


Figure 3: **Training on synthetic data from BackboneRef improves *in vitro* expression rates of designed proteins. (A)** *In vitro* expression rates for generations from models trained on a synthetically augmented dataset (U50+BBR, UR50+BBR-novel, blue) versus on only natural sequences (UR90, grey). **(B)** Predicted structures and metrics for select successfully-expressed proteins from the U50+BBR model. e-values were determined via Foldseek query against AFDB-UniProt.

Finally, we experimentally validated the effects of structure-based synthetic data augmentation on the *in vitro* expressibility of generated sequences. We selected 29 samples from each of the UniRef90 and UR50+BBR-novel models for *in vitro* validation. These sequences were 60-1500 residues long and without predicted signal peptides, mitochondrial transit peptides, or transmembrane helices; no filtering based on predicted structural properties was done. Generations were expressed in two *E. coli* strains each (BL21-AI, a derivative of the BL21 line, and BLR(DE3)); a protein was considered to be expressed successfully if expression in at least one strain was detected at the correct molecular weight via SDS-PAGE. Generations from the model trained on the UR50 + BBR-novel dataset were more likely to express (15/29; UR50+BBR) relative to generations from the UniRef90 model (8/29; UR90), a 1.875 fold increase (Fig. 3A). Select proteins generated by the UR50+BBR model that expressed successfully in *E. coli* are visualized in Figure 3B.

3 DISCUSSION

This work demonstrates that structure-based synthetic data augmentation can shift protein language model generations towards desirable traits. By training on novel, high-quality synthetic data, we induce protein language models to generate distributions that are further from the distribution of natural sequences (higher FPD scores) while increasing the expression rate from 27.6% to 51.7%.

The output distributions of generative models are determined by their inductive biases and training data. For proteins, training data consists of amino-acid sequences and structural coordinates. While structures are richer than sequence, sequencing data available for training is more abundant. Previous approaches for combining information from sequences and structures include repurposing a structure prediction module (Lu et al., 2025; Lisanza et al., 2024), designing a model architecture that can use both (Hayes et al., 2025), or predicting structures from sequence (Huguet et al., 2024).

In contrast to prior approaches, we use sequences derived from a structure-based generation pipeline to augment sequence-only generative models. Our structure-based synthetic protein dataset, BackboneRef, is the largest and most systematic sampling of sequences from a structure-based generative model. Data augmentation using BackboneRef effectively transfers information from the PDB during training while maintaining the simplicity of sequence-based generative modeling.

In summary, we present an open-source and experimentally-validated pipeline for combining information from sequence and structure databases into a protein language model. We achieve these results with relatively small models trained on modest compute. BackboneRef-novel contains just 138,044 backbones with an average of 74 sequences per backbone. Our rarefaction analysis indicates that we have not exhausted all possible novel and unique synthetic backbones; scaling the augmentation dataset and the model could lead to increased diversity and quality of generations. We hope that this work illuminates a new avenue to improve protein design.

REFERENCES

Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Alex X Lu, Nicolo Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. September 2023.

H M Berman, J Westbrook, Z Feng, G Gilliland, T N Bhat, H Weissig, I N Shindyalov, and P E Bourne. The protein data bank. *Nucleic Acids Res.*, 28(1):235–242, January 2000.

A. Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design. *arXiv preprint arXiv:2402.04997*, 2024.

A. E. Chu, Jinho Kim, Lucy Cheng, Gina El Nesr, Minkai Xu, Richard W. Shuai, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 121(27): e2311500121, 2024.

Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J Ragotte, Lukas F Milles, Basile IM Wicky, Alexis Courbet, Rob J de Haas, Neville Bethel, et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):7112–7127, October 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv [cs.LG]*, December 2023.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S Molina, Neil Thomas, Yousuf A Khan, Chetan Mishra, Carolyn Kim, Liam J Bartie, Matthew Nemeth, Patrick D Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, pp. eads0018, January 2025.

Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al. Sequence-augmented se (3)-flow matching for conditional protein backbone generation. *arXiv preprint arXiv:2405.20313*, 2024.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, Shan Tie, Vincent Xue, Sarah C Cowles, Alan Leung, João V Rodrigues, Claudio L Morales-Perez, Alex M Ayoub, Robin Green, Katherine Puentes, Frank Oplinger, Nishant V Panwar, Fritz Obermeyer, Adam R Root, Andrew L Beam, Frank J Poelwijk, and Gevorg Grigoryan. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, November 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

Thomas J Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, pp. 1–4, 2023.

Opher Lieber, Barak Lenz, Hofit Bata, Gal Cohen, Jhonathan Osin, Itay Dalmedigos, Erez Safahi, Shaked Meirom, Yonatan Belinkov, Shai Shalev-Shwartz, Omri Abend, Raz Alon, Tomer Asida, Amir Bergman, Roman Glozman, Michael Gokhman, Avashalom Manevich, Nir Ratner, Noam Rozen, Erez Shwartz, Mor Zusman, and Yoav Shoham. Jamba: A hybrid transformer-mamba language model. *arXiv [cs.CL]*, March 2024.

S. L. Lisanza, Jacob Merle Gershon, Samuel W. K. Tipps, Jeremiah Nelson Sims, Lucas Arnoldt, Samuel J. Hendel, Miriam K. Simma, Ge Liu, Muna Yase, Hongwei Wu, Claire D. Tharp, Xinting Li, Alex Kang, Evans Brackenbrough, Asim K. Bera, Stacey Gerben, Bruce J. Wittmann, Andrew C. McShan, and David Baker. Multistate and functional protein design using RoseTTAFold sequence space diffusion. *Nature Biotechnology*, 2024.

A. X. Lu, Wilson Yan, Sarah A Robinson, Kevin K Yang, Vladimir Gligorijevic, Kyunghyun Cho, Richard Bonneau, Pieter Abbeel, and Nathan Frey. Generating all-atom protein structure from sequence-only training data. *bioRxiv*, pp. 2024–12, 2024.

Tianyu Lu, Melissa Liu, Yilin Chen, Jinho Kim, and Po-Ssu Huang. Assessing generative model coverage of protein structures with SHAPES. *bioRxiv*, 2025. doi: 10.1101/2025.01.09.632260. URL https://www.biorxiv.org/content/early/2025/01/14/2025.01.09.632260.

Ali Madani, Ben Krause, Eric R Greene, Subu Subramanian, Benjamin P Mohr, James M Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, et al. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023.

Wei Qu, Jiawei Guan, Rui Ma, Ke Zhai, Weikun Wu, and Haobi Wang. P(all-atom) is unlocking new path for protein design. *bioRxiv*, pp. 2024–08, 2024.

Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

UniProt Consortium. UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.*, 53(D1):D609–D617, January 2025.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L M Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.*, 42(2):243–246, February 2024.

Mihaly Varadi, Damian Bertoni, Paulyna Magana, Urmila Paramval, Ivanna Pidruchna, Malarvizhi Radhakrishnan, Maxim Tsenkov, Sreenath Nair, Milot Mirdita, Jingi Yeo, Oleg Kovalevskiy, Kathryn Tunyasuvunakool, Agata Laydon, Augustin Žídek, Hamish Tomlinson, Dhavanthi Hariharan, Josh Abrahamson, Tim Green, John Jumper, Ewan Birney, Martin Steinegger, Demis Hassabis, and Sameer Velankar. AlphaFold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.*, 52(D1):D368–D375, January 2024.

John M Walker. *The proteomics protocols handbook*. Humana Press, Totowa, NJ, 2005.

Chentong Wang, Sarah Alamdari, Carles Domingo-Enrich, Ava Amini, and Kevin K. Yang. Towards deep learning sequence-structure co-generation for protein design, 2024a.

X. Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Dplm-2: A multimodal diffusion protein language model. *arXiv preprint arXiv:2410.13782*, 2024b.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, Basile I M Wicky, Nikita Hanikel, Samuel J Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620 (7976):1089–1100, August 2023.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999.

A APPENDIX

SUPPLEMENTARY METHODS

BACKBONEREF GENERATION AND CHARACTERIZATION

Synthetic backbone generation We used RFdiffusion (checkpoint "Base_ckpt") Watson et al. (2023) for all backbone generations. In order to increase applicability of analyses comparing backbones with natural sequences, we sampled the length distribution of sequences in UniRef50 (January 2024). We generated proteins with minimum length 40 and maximum length 512 for computational efficiency.

Secondary structure analysis For all secondary structure analyses, we predicted 3-class (loops, helices, and strands) type assignment using DSSP (Kabsch & Sander, 1983). We performed this analysis directly on the backbones, rather than on the sequence-designed and folded structures. For visualization of percentage strandedness and helicity for each backbone we averaged the count of the given annotation class over sequence.

Self-consistency analysis For each backbone, we used ProteinMPNN (Dauparas et al., 2022) at temperature 0.1 to design 10 sequences per backbone. We then used OmegaFold (Wu et al., 2022) to predict a structure for each sequence, and then computed the scRMSD between the C_α positions of the original backbone and the predicted structures.

We also separately conducted fixed-backbone sequence design with temperatures 0.2, 0.3, 0.4, 0.6, 0.8, and 1.0, designing ten sequences per backbone per temperature prior to structure prediction using OmegaFold. scRMSD was then computed between each backbone's predicted structures at the the various temperatures and the original backbone.

Foldseek novelty search Rather than using Foldseek's (van Kempen et al., 2024) 'easy-search' directly on backbones in C_α mode, we used the lowest scRMSD structure from our self-consistency analysis and used Foldseek to query these structures against the AlphaFold/UniProt database. We report the maximum TM-score returned per query; this represents the closest structural match to the query structure.

Foldseek clustering To perform our clustering analysis, we again selected the best scRMSD structure of 10 candidates. Using Foldseek, these structures were then clustered along with all structures from the PDB (209,850 structures). For PDB files with multiple chains, we selected the first one.

MODEL TRAINING WITH STRUCTURE-BASED SYNTHETIC DATA AUGMENTATION

Pretraining data preparation We first generated an initial set of 185 sequences per backbone at temperature 0.2, producing a dataset with 44,553,550 sequences. To create the BBR-unfiltered dataset, we randomly selected 42 sequences per backbone at temperature 0.2, yielding 10,114,860 sequences, removed exact duplicates, and then subsampled randomly to produce a dataset of 10M sequences. To create the BBR-sc dataset, we removed any backbones with average scRMSD score greater than 2Å, leaving 127,633 backbones. We randomly selected 80 sequences per backbone at temperature 0.2, removed exact duplicates, and then subsampled randomly to produce a dataset of 10M sequences. To create the BBR-novel dataset, we removed any backbones with maximum TM-score to any structure in the PDB larger than 0.5, leaving 138,044 backbones. We randomly sampled 74 sequences per backbone at temperature 0.2, removed exact duplicates, and then subsampled randomly to produce a dataset of 10M sequences.

Model training For all experiments, we used a 170M parameter Jamba (Lieber et al., 2024) architecture and trained with an autoregressive objective. Specifically, the model has 24 layers; every eighth layer is a transformer module, with the remainder being Mamba modules (Gu & Dao, 2023). Every other Mamba module uses mixture of experts with 16 experts instead of a dense layer. The model dimension is 256, with intermediate layer widths of 1024 inside each transformer block. All training runs used 8 NVIDIA A100 or 8 NVIDIA H100 GPUs. We used an inverse square root scheduler with linear warmup of 10,000 steps, max learning rate of 4e-4, and no weight decay. We

used the Adam optimizer with betas (0.9, 0.999). All models were trained for 76k steps on 8 H100 or A100 GPUs with adaptive batch sizes where every batch had at most 360k tokens per GPU.

We constructed combined natural-synthetic datasets by taking the union of a fixed training split of UniRef50 from January 2024 (63,662,039 sequences) to each of the synthetic datasets described in the "Pretraining data preparation" section. This produced training datasets of $\sim$73M sequences each. We used the same training and testing splits for each experiment. We also trained baseline models using UniRef50 and UniRef90 (184,520,055 sequences), with sampling by UniRef50 clusters. For the UniRef90 model, one randomly-chosen member per UniRef50 cluster was seen per epoch.

MODEL EVALUATION

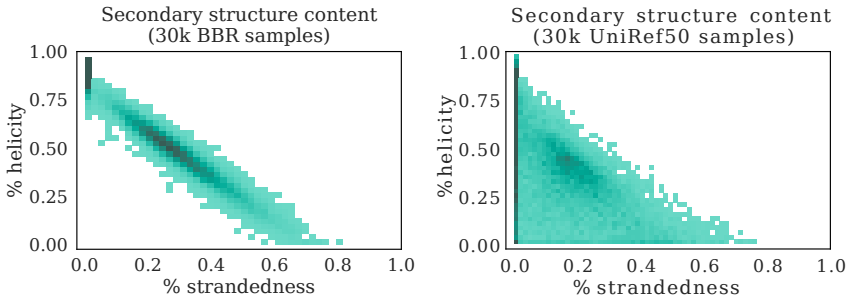
Fréchet protein distance (FPD) To estimate the divergence between distributions of protein sequences, we sampled 1024 sequences from each distribution, embedded them with the ProtBert (Elnaggar et al., 2022) model, and computed the earth mover's distance

$$\|\mu - \mu'\|^2 + \text{Tr}\left(\Sigma + \Sigma' - 2(\Sigma\Sigma')^{\frac{1}{2}}\right) \tag{1}$$

where μ and μ' are average vectors in the ProtBert embedding space and Σ and Σ' are covariance matrices for the two sets of embeddings. Tr refers to the trace of this matrix.

***In vitro* characterization of unconditional generations** Starting from 200 generations each from the UR50 + BBR-novel and UR90 models (400 total), we selected 29 sequences from each model for expression and further experimental characterization. Sequences were filtered out from the original set if they were < 60 amino acids or > 1500 amino acids. We also filtered out sequences predicted to have a signal peptide, mitochondrial transit peptide, or a transmembrane helix. Lastly, sequences were also filtered out if they exhibited a GRAVY score > 0, an instability index > 70, or if the pI was more than ± 2 units of the buffer pH. These metrics were calculated using ProtParam (Walker, 2005). After filtering, there were 47 UR50+BBR sequences and 66 UR90 sequences remaining. We randomly selected 29 from each set. These sequences were then expressed in two *E. coli* strains (BL21-AI, a derivative of BL21, and BLR(DE3)). Protein abundance was quantified using SDS-PAGE, and a protein was considered successfully expressed if any protein was detected at the correct molecular weight by SDS-PAGE in the yield from any of the three strains.

SUPPLEMENTARY FIGURES

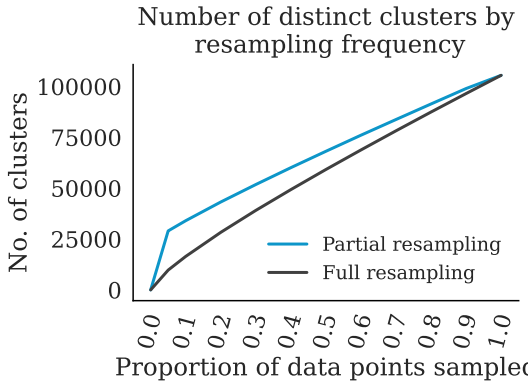


(a) (b)

Figure S1: **Structural diffusion models preferentially generate helical structures.** (**A**) Histograms of percentage of residues called as stranded (x-axis) or helical (y-axis) for (**A**) $n = 30,000$ generated synthetic backbones from BackboneRef and (**B**) $n = 30,000$ randomly-selected structures for sequences in UniRef50, with structures downloaded from AlphaFold-DB.

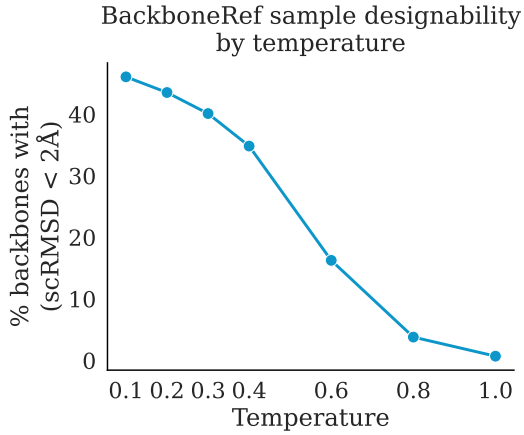
Table S1: **Numbers of clusters and structures for inferred clusters of different compositions.**

Cluster composition	Number clusters (% clusters)	Number structures (% structures)
PDB only	21,350 (20.2%)	172,981 (38.4%)
PDB + BBRef	1,035 (1.0%)	103,656 (23.0%)
BBRef only	83,121 (78.8%)	174,043 (38.6%)



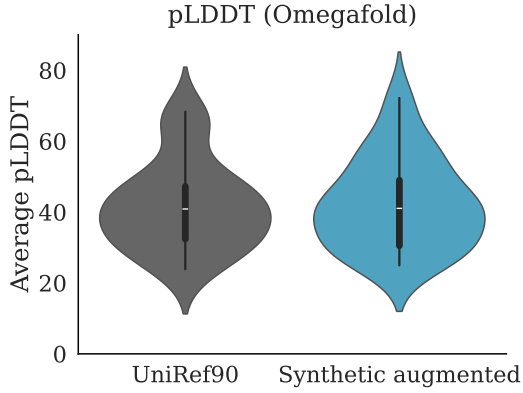
(a)

Figure S2: **Rarefaction analysis of inferred structural clusters from Foldseek.** Foldseek was used to cluster a dataset comprised of structures downloaded from the PDB concatenated to lowest scRMSD predicted structures from BBR. Structures were resampled at varying frequency (either across the whole dataset, grey, or only resampling BackboneRef samples, blue), and the number of distinct Foldseek clusters was computed. Note that we do not recompute clusters at different frequencies and instead resample the initial set of inferred Foldseek clusters.



(a)

Figure S3: **Average percentage of designable samples from BackboneRef by temperature,** with the percentage of designable BackboneRef samples computed as those with average scRMSD < 2Å over 10 predicted structures at a given temperature.



(a)

Figure S4: **pLDDT distributions of OmegaFold-predicted structures** for unconditional generations from models trained with U50+BBR-novel (n=29, blue) and without (n=29, grey) synthetic data augmentation.