Generating and evaluating diverse sequences for protein backbones

Yo Akiyama
Massachusetts Institute of Technology
yo_aki@mit.edu

Sergey Ovchinnikov
Massachusetts Institute of Technology
so3@mit.edu

Abstract

Generating diverse sequences for protein backbones remains an active challenge with important implications. *De novo* protein design typically requires screening large sets of diverse sequences to identify viable candidates under certain experimental conditions. Sequence design has also recently been employed to generate synthetic data for training models. Diverse sets of sequences can be trivially generated by increasing the sampling temperature of sequence design models; however, we find that the covariation between residues in these sequences do not recapitulate natural covariation or the structures for which they were designed. An alternative approach designs sequences for structural ensembles, motivated by previous studies demonstrating that natural sequence variation is strongly tied to structural variation rather than the constraints of a static backbone. RFdiffusion, with a reduced number of noising and denoising steps, has demonstrated the ability to diversify structures via learned potentials. Here, we compare sequences generated using single fixed backbones and partial RFdiffusion ensembles. Our analyses reveal that structural variation from RFdiffusion results in increased sequence diversity at a given sequence temperature without compromising AlphaFold2 designability metrics. Moreover, the covariance from partial diffusion MSAs better recapitulate natural covariation and contacts. Lastly, we propose a new approach to evaluate the quality of sequences, which tests AlphaFold2 self-consistency using shallow synthetic MSAs. This method enables evaluation of sequences for which the efficacy of the AlphaFold2 single-sequence self-consistency remains limited.

1 Introduction

Advancements in generative modeling have led to the development of automated pipelines for *de novo* protein design. These pipelines generally share a common procedure: sequence-free backbone generation followed by structure-conditioned sequence design, often generating and screening a large set of candidate sequences. At inference, structure-conditioned sequence design models, such as ProteinMPNN, predict the conditional probability distribution of the amino acid identity for a residue given the backbone structure and all other amino acid identities, wherever decoded; that is, $P(\text{seq}_i|\text{backbone}, \text{seq}_{\_i})$, where _i indicates the sequence for all resides excluding residue i [1]. Specifically, the categorical probability distribution of amino acid identities is computed via $\text{softmax}(a_k, \tau) = \frac{e^{a_k/\tau}}{\sum_{j=1}^{A} e^{a_j/\tau}}$, where τ is a sampling temperature hyperparameter. In order to generate a sequence for a given backbone, one can sample from this predicted probability distribution. As $\tau \to 1$, amino acid identities are sampled from the distribution inferred by the model, and as $\tau \to 0$, samples approach the argmax of the distribution.

Currently, diverse sets of sequences can be generated by increasing the sampling temperature, τ. Theoretically, when applied to natural backbones, sampling sequences directly from the inferred

Machine Learning for Structural Biology Workshop, NeurIPS 2024.

distribution should yield the natural sequence distribution. However, additional analysis is required to investigate whether this phenomenon arises in practice.

Another approach to generating diverse sequences is based on diversifying the target backbone. Indeed, by exploring structural and sequence diversity of natural proteins, studies have suggested that highly diverse sequences from a protein family often have non-trivial discrepancies ($\tilde{1}$-3Å RMSD) [2–4]. A previous study demonstrated the value of structural diversity for sequence design by applying Rosetta backrub to natural domains [5–7]. Briefly, backrub generates realistic ensembles by modeling the local conformational flexibility of crystallographic structures. Their resulting synthetic multiple sequence alignments (MSAs) better recapitulated natural pairwise covariation compared to those generated using the static crystal structure coordinates alone. Their results therefore highlighted the dominant role of structural constraints defined by backbone flexibility rather than a single static structure in driving sequence variability.

Recently, RFdiffusion with partial noising and denoising schedules has been used to sample diverse and idealized backbones for binder design [8–10]. Rather than applying the full noising and denoising process across T steps to generate a random backbone, $P(\text{structure})$, the partial diffusion protocol takes a structure as input and uses a smaller number of steps ($t < T$) to generate a similar structure from a noisy version of the initial structure, $P(\text{structure}^* | \text{structure}_{initial} + \epsilon)$, where ϵ increases with t (Fig. 1). This protocol leverages RFdiffusion's learned potentials and may generate realistic ensembles similar to the Rosetta backrub method.

Here, we compare sets of sequences generated using the partial RFdiffusion workflow against those designed using a fixed backbone using the same 40 structurally diverse domains used in the backrub design study. We measure sequence diversity and evaluate the degree to which covariation reflects natural coevolution and structural constraints. We then evaluate individual sequences using single-sequence AlphaFold2 (AF2) self-consistency metrics and propose a new AF2-based approach to evaluate sets of designs, which provides a framework for comparing sequence design methods and circumvents issues regarding AF2's inability to predict structure from a single-sequence. [11].

2 Results

2.1 Structural diversity from partial RFdiffusion increases sequence diversity for a given ProteinMPNN temperature

Starting from the crystallographic structures of the 40 domains from the backrub design study, we first perform a grid search of ProteinMPNN sampling temperatures ($\tau \in \{0.01, 0.1, 0.2, ..., 1\}$) and RFdiffusion steps ($T \in \{0, 1, 5, 7, 10, 12, 15, 20\}$, where $T = 0$ corresponds to a single fixed backbone) to characterize sequence diversity along these two axes (Fig. 1). We limit the maximum number of steps to $T = 20$ since structures generated beyond 20 steps reflect different folds (TM-score < 0.5) (Fig. S1). We find that while sequence design using fixed backbones and ProteinMPNN sampling temperature $\tau = 1$ results in a maximum average sequence identity of 31%, the diversity of sequences designed using partial RFdiffusion ensembles exceeds this level at substantially lower sampling temperatures (Fig. 2A). For example, sequences generated by setting $T = 15$ and $\tau = 0.4$ surpasses this diversity with an average of 28%. The remaining analyses focus on evaluating these diverse sets and their individual sequences.

2.2 Synthetic MSAs generated using partial RFdiffusion ensembles recapitulate natural covariation and contacts

The backrub design study revealed that structural constraints play a dominant role in shaping residue covariation in natural sequences [5]. We reason that reliable sequence design of natural backbones should also recapitulate natural covariation since similar structural constraints are imposed. Therefore, in order to characterize synthetic MSAs generated from the fixed backbone and partial diffusion design protocols, we first compute the overlap between highly covarying residues in the designed sequences and natural sequences, consistent with the analysis performed in the Rosetta backrub design study. Similar to the backrub designed sequences, synthetic MSAs generated using partial RFdiffusion yields greater overlap in covarying residues than the MSAs generated from a fixed backbone (Fig. S2). The median percent overlap maximizes at 33% setting $T = 15$ and $\tau = 0.3$ ($p = 10^{-9}$). Notably, at a comparable sequence diversity, a sampling temperature of 1 on a fixed

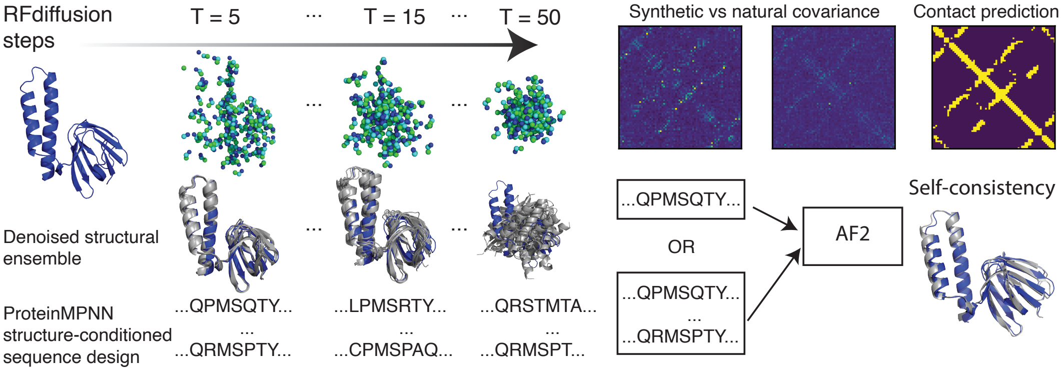

Figure 1: Partial diffusion design protocol and evaluation overview

backbone results in a median overlap of 19%, highlighting a substantial decrease in covariation similarity ($p = 10^{-15}$). Interestingly, the overlap using partial RFdiffusion is generally lower than the median overlap of 38% from the backrub design study ($p = 0.01$). We note that Rosetta backrub uses all-atom contexts to generate realistic conformations, which may result in more realistic ensembles. Further studies are needed to explore this trend.

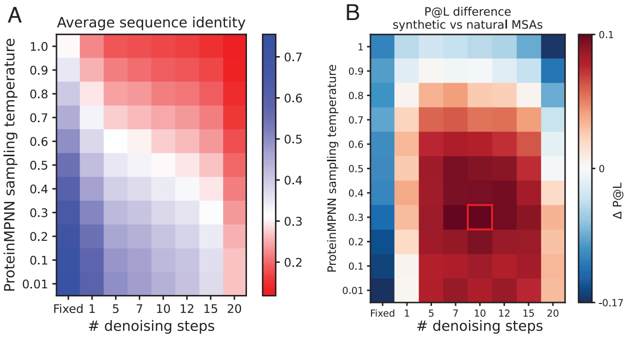

Figure 2: (A) Average sequence identity across design parameters; (B) average difference in P@L between synthetic and natural MSAs

Whereas the synthetic MSAs were generated strictly using structural constraints imposed by the partial RFdiffusion ensembles, natural sequences arise from a variety of constraints, including structure, function, and stability among many others. We therefore reason that the source of remaining mismatch between highly covarying pairs may be due to differences in these constraints. In order to predict contacts from these MSAs, we compute the inverse covariance and apply average product correction for each MSA. Indeed, by computing the precision of the top L covarying pairs of residues for contact prediction (P@L), we find that the covariance from partial diffusion MSAs more strongly reflect contacts compared to covariance in natural MSAs (Fig. 2B). The increased P@L further supports that differences in covariation between natural and partial diffusion generated synthetic MSAs reflect differences in constraints and the structurally idealized characteristic of the designed sequences. Notably, contact prediction using fixed backbone MSAs resulted in significantly lower P@L compared to natural MSAs across all temperatures (FDR < 0.05), suggesting that fixed backbone design weakens expected second order relationships.

2.3 Sequence design using partial RFdiffusion ensembles maintains AlphaFold2 self-consistency at high levels of sequence diversity

Next, we aim to measure compatibility between the target structure and individual sequences, referred to as designability. In order to evaluate designs, we employ the widely used AF2 self-consistency test on a random subset of 50 sequences per domain [12, 11]. We first generate 50 sequences for each of the 40 domains applying ProteinMPNN across varying temperatures on fixed backbones. RMSD between the predicted structure using AF2 single-sequence mode and the crystal structure increases monotonically with sampling temperature, consistent with the hypothesis that sequence-structure compatibility deteriorates as sampling temperature increases (Fig. 3A).

We next ask whether highly diverse sets of sequences from the partial RFdiffusion protocol maintain AF2 self-consistency. To this end, we compare sequences from the maximally diverse set from fixed backbone design ($\tau = 1$) to a comparably diverse set generated from 15 RFdiffusion steps and sampling temperature of 0.4 (average sequence identity = 31% and 29%, respectively). Notably, predictions for sequences generated by the partial diffusion generally remain closer to the target structure ($p = 2 * 10^{-16}$) (Fig. 3B, 3C). Of the 40 domains, 10 have a median RMSD < 2Å using the partial diffusion protocol, while none have a median RMSD that passes this standard threshold when using a fixed backbone with $\tau = 1$. Moreover, the partial diffusion protocol generated at least 1 sequence with RMSD < 2Å for 29 domains compared to 12 for fixed backbone design. Among the domains with the most self-consistent sequences is chloroplastic m-type thioredoxin, for which the median RMSD is 1.7Å, maintaining the structure around the active site (Fig. 3D). The structure is entirely ablated in predictions from fixed backbone design, with no predictions less than 2Å RMSD. Overall, single-sequence AF2 self-consistency experiments suggest the dramatic increase in compatibility of diverse sequences to their target backbones using partial RFdiffusion ensembles.

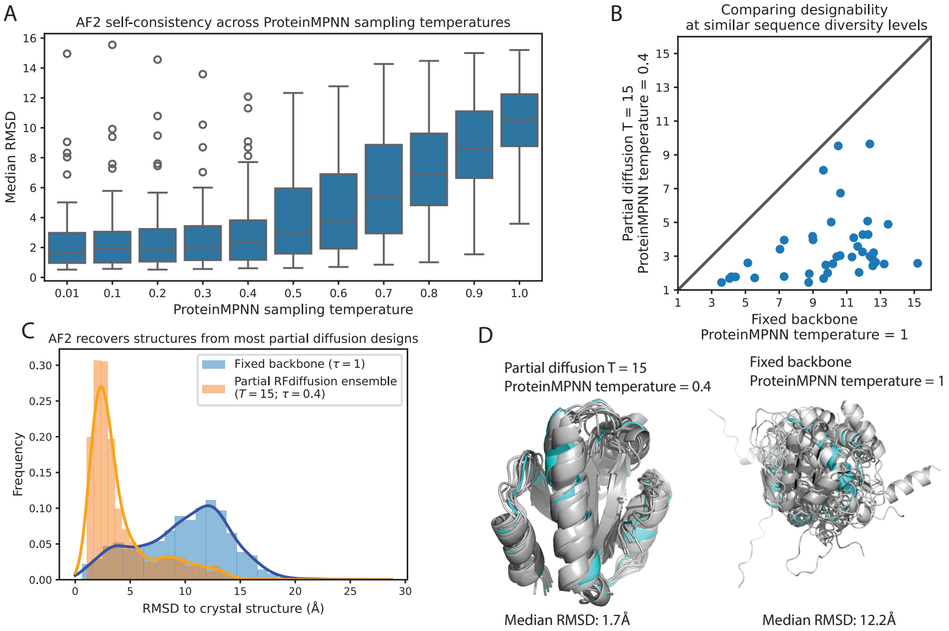

Figure 3: (A) Median RMSD for each of the 40 domains across ProteinMPNN sampling temperatures; (B) median RMSD between AF2 predictions and crystal structure for each domain using partial RFdiffusion ($T = 15, \tau = 0.4$) and fixed backbone design ($\tau = 1$); (C) distribution of RMSD comparing fixed backbone design and partial diffusion; (D) structural alignment of five AF2 predictions for each of two experiments superimposed to the crystal structure (blue; PDB id: 1FB0)

2.4 Evaluating large designs using AlphaFold2 prediction with shallow MSAs

AF2 single-sequence self-consistency scores are widely used for in-silico evaluation, yet they often fail for naturally occurring proteins [13]. Given that self-consistency for designs generally decreases sharply for larger designs (more than 300 residues), we randomly select seven natural, high confidence (median $\mathrm{pLDDT} > 0.85$) AF2 structures of proteins with at least 750 residues [9, 13–15]. We find that none can be predicted with AF2 in the absence of an MSA (TM-score < 0.5). While the number tested remains small and more thorough analyses are needed to establish a relationship between sequence length and AF2 single-sequence self-consistency, these results motivate new evaluation methods that do not fail for natural sequences. Building on the notion that the diversity reflected in designed sequences should be consistent with structural constraints, we ask whether AF2 self-consistency using MSAs can address this problem.

We first note that structural signal is not entirely ablated from these synthetic MSAs, even for fixed backbone design (Fig. S3). The Evoformer module of AF2 computes the mean across the outer product of each sequence. Therefore, given an MSA with sufficient depth, AF2 should extract the proper structural signal from noisy MSAs (i.e. law of large numbers). We therefore test self-consistency across varying depths of synthetic MSAs from 2 to 128, binned by powers of 2. Furthermore, in order to avoid structural refinement from learned potentials and to emphasize extraction of structural properties directly from the MSA, we disable AF2 recycling and templates.

Using the seven large natural sequences for which the structures AF2 cannot predict in single-sequence mode, we first generated synthetic MSAs for each using the fixed backbone method with sampling temperature $\tau = 1$ and partial RFdiffusion with $T = 10$ and $\tau = 0.3$. The resulting MSAs have comparable diversity with average sequence identity 25% across the seven proteins. We note that the natural MSAs have slightly higher average sequence diversity (38%) when subsampled for 128 sequences each using hhfilter [16]. Interestingly, we find that while AF2 is able to recover the correct fold (TM-score > 0.5) using only 8 sequences on average from partial RFdiffusion and natural MSAs, none of the folds are recovered for the fixed backbone MSAs at this depth (Fig. 4). This result is consistent with our previous analyses, which suggest decreased compatibility between sequences designed with ProteinMPNN temperature $\tau = 1$ and their target structure. Moreover, at this MSA depth, the TM-score for predictions using partial diffusion MSAs are significantly higher than those using natural MSAs ($p = 0.02$). Again, we reason that these sequences generated using a low ProteinMPNN sampling temperature are structurally idealized, whereas natural sequences are optimized for many other factors. Overall, our analyses highlight the potential value of using synthetic MSAs for AF2 self-consistency tests for evaluating the compatibility between designed sequences and target structure.

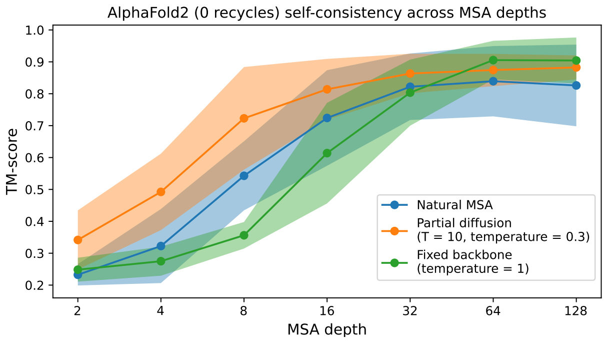

Figure 4: Comparing the median TM-score of AF2 predictions across MSA depths

3 Discussion

Sequence design using structural ensembles generated via partial diffusion offers a promising approach to generate large, diverse sets of sequences for a given backbone. We find that the covariances stored in the diversity of these sequences better recapitulates natural covariation arising from structural constraints. Single-sequence AF2 self-consistency highlighted i) the monotonic deterioration of designability along increasing sampling temperatures; and ii) that the partial diffusion protocol maintains strong self-consistency metrics at high sequence diversity levels. While single-sequence AF2 self-consistency evaluations have been widely adopted, its efficacy as a general test for designability is complicated by its failure to validate natural sequences. We therefore introduced a variant of the self-consistency metric using shallow MSAs, which distinguished natural, partial diffusion, and fixed backbone sequences. Future work can explore whether these differences are linked to experimental success rates and can be used to evaluate sequence design methods.

Crucially, these analyses further support the hypothesis that sequence diversity is best achieved via structural diversity. While current structure-conditioned sequence design models approximate the probability distribution based on a static structure, $P(\text{sequence}|\text{structure})$, structural constraints applied to protein sequences may be more accurately represented by a set of structures, $P(\text{sequence}|\text{structure(s)})$. Experimental validation is needed to further substantiate whether differences in computational evaluations are reflected in the success rates of designed sequences.

These results carry implications beyond de-novo protein design. In order to scale deep learning models, recent studies have explored whether augmenting training datasets using samples from generative models could improve model performance [17–22]. Interestingly, ESM3 was trained using synthetic sequence-structure pairs generated using a structure-conditioned sequence design model on structures from the AlphaFold database and ESM Metagenomic atlas. Our results motivate further investigation into the consequences of training protein language models on synthetic data.

References

[1] J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning-based protein sequence design using proteinmpnn. *Science*, 2022. doi: 10.1126/science.add2187.

[2] Cyrus Chotia and Arthur Lesk M. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 1986.

[3] Burkhard Rost. Twilight zone of protein sequence alignments. *Protein Engineering, Design and Selection*, 1999. doi: https://doi.org/10.1093/protein/12.2.85.

[4] Mireia Olivella, Angel González Wong, Leonardo Pardo, and Xavier Deupi. Relation between sequence and structure in membrane proteins. *Bioinformatics*, 2013. doi: https://doi.org/10.1093/bioinformatics/btt249.

[5] Noah Ollikainen and Tanja Kortemme. Computational protein design quantifies structural constraints on amino acid covariation. *PLOS Computational Biology*, 2013. doi: 10.1371/journal.pcbi.1003313.

[6] Ian W. Davis, W. Bryan Arendall, David C. Richardson, and Jane S. Richardson. The backrub motion: How protein backbone shrugs when a sidechain dance. *Structure*, 2006. doi: https://doi.org/10.1016/j.str.2005.10.007.

[7] Colin A. Smith and Tanja Kortemme. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *Journal of Molecular Biology*, 2008. doi: https://doi.org/10.1016/j.jmb.2008.05.023.

[8] Susana Vázquez Torres, Philip J. Y. Leung, and Preetham Venkatesh et. al. Diffusion protein binders to intrinsically disordered proteins. *biorxiv*, 2024. doi: 10.1101/2024.07.16.603789.

[9] 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*, 2023. doi: https://doi.org/10.1038/s41586-023-06415-8.

[10] Caixuan Liu, Kejia Wu, and Hojun Choi et. al. Diffusion protein binders to intrinsically disordered proteins. *biorxiv*, 2024. doi: 10.1101/2024.07.16.603789.

[11] Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *International Conference on Learning Representations (ICLR)*, 2023. doi: https://doi.org/10.48550/arXiv.2206.04119.

[12] 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*, 2021. doi: https://doi.org/10.1038/s41586-021-03819-2.

[13] Yeqing Lin, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2. *arxiv*, 2024. doi: https://doi.org/10.48550/arXiv.2405.15489.

[14] Alexander Chu, Jinho Kim, Lucy Cheng, and Po-Ssu Huang. An all-atom protein generative model. *Proceedings of the National Academy of Sciences*, 2024. doi: https://doi.org/10.1073/pnas.2311500121.

[15] Mihaly Varadi, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, Oana Stroe, Gemma Wood, Agata Laydon, Augustin Žídek, Tim Green, Kathryn Tunyasuvunakool, Stig Petersen, John Jumper, Ellen Clancy, Richard Green, Ankur Vora, Mira Lutfi, Michael Figurnov, Andrew Cowie, Nicole Hobbs, Pushmeet Kohli, Gerard Kleywegt, Ewan Birney, Demis Hassabis, and Sameer Velankar. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 2021. doi: https://doi.org/10.1093/nar/gkab1061.

[16] Martin Steinegger, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. Hh-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 2019.

[17] 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 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. *biorxiv (Cold Spring Harbor Laboratory)*, 2024. doi: https://doi.org/10.1101/2024.07.01.600583.

[18] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2404.01413.

[19] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. The curse of recursion: Training on generated data makes models forget. *arXiv*, 2023.

[20] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2401.16380.

[21] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J. Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2304.08466.

[22] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *arXiv*, 2023. URL `https://doi.org/10.48550/arXiv.2306.00984`.

[23] S.D. Dunn, L.M. Wahl, and G.B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 2007.

[24] Justas Dauparas, Haobo Wang, Avi Swartz, Peter Koo, Mor Nitzan, and Sergey Ovchinnikov. Unified framework for modeling multivariate distributions in biological sequences. *arXiv*, 2019.

[25] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 2022.

4 Supplementary material

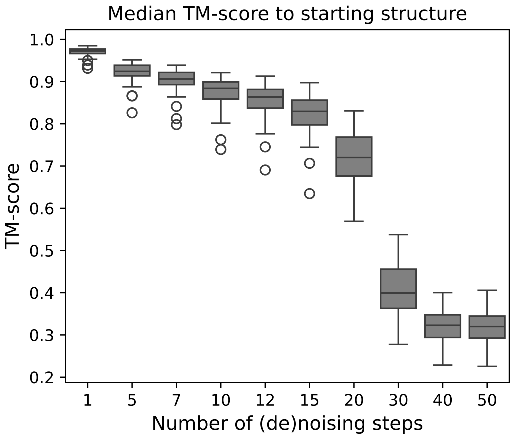

Figure S1: Median TM-score of structures generated using partial RFdiffusion and the crystal structure across diffusion steps.

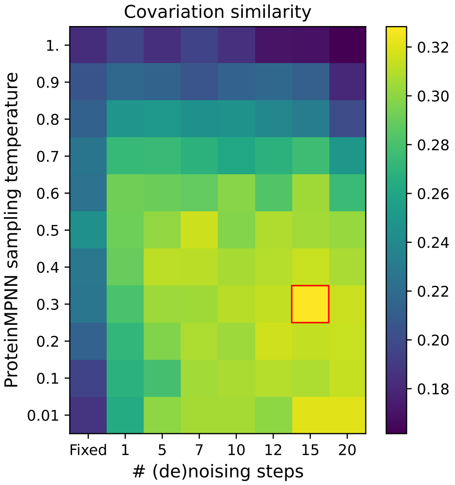

Figure S2: Median proportion of overlapping covarying residues between synthetic and natural MSAs. Red box indicates the maximum similarity experiment.

4.1 Covariance analyses

To compute the overlap of the top covarying pairs of residues in synthetic and natural MSAs, we follow the analysis described in the backrub design study [5]. Briefly, we first compute the mutual information between residues, then subtract the background mutual information due to noise and shared phylogeny (APC) [23]. These values are then converted to Z-scores per column and multiplied per pair of residues. The final score is the square root of the absolute value of this score. Pairs with scores greater than two standard deviations above the mean are considered to be the top covarying pairs. Covariation overlap is computed as $2C/(A+B)$, where C is the number of shared covarying pairs, and A and B are the total number of top covarying pairs in the natural and synthetic MSAs, respectively. For these analyses, we limit to residues without gaps in the natural MSA, consistent with the pre-processing performed in the backrub design study. Additional details can be found in the corresponding manuscript.

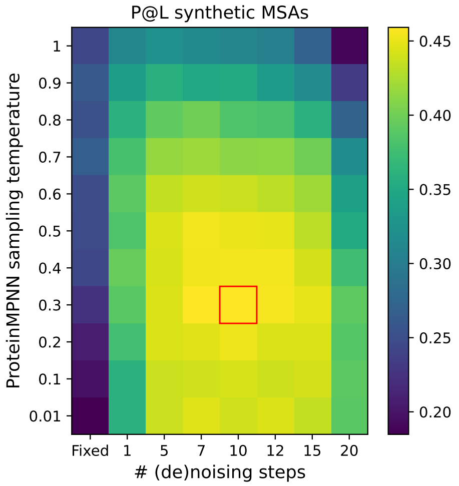

Figure S3: Average precision of the top L covarying pairs of residues for contact prediction (P@L) across design experiments. Red outline indicates highest P@L

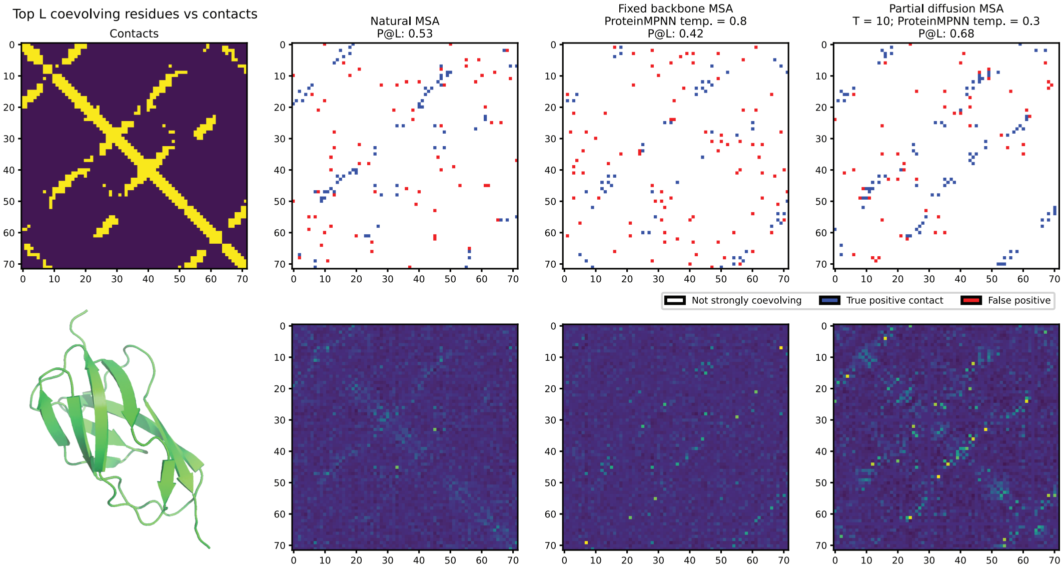

Figure S4: Contact prediction from natural and synthetic MSAs for human EBF1 IPT/TIG domain (PDB id: 3MQI). Fixed backbone and partial diffusion MSAs have comparable average sequence identities (40% and 38%, respectively. Top row shows contacts from the crystal structure, defined as residues within 8Å, followed by the top L (length) covarying residues from the inverse covariance calculation with average product correction (APC). Red points indicate covarying residues that are not contacts, while blue points indicate true positive contacts. Bottom row shows the cartoon representation of the crystallographic structure, followed by the inverse covariance matrix after computing the L2-norm of the sum across the second and fourth dimensions of the $L \times 20 \times L \times 20$ tensor and applying APC.

We predict residue contacts from MSAs using the top L couplings from the inverse covariance method, where L corresponds to the length of the protein [24]. Additional details can be found in the corresponding manuscript.

4.2 Structure conditioned sequence design using ProteinMPNN

All ProteinMPNN sequence design is performed using the ColabDesign implementation, using the v_48_020 model.

4.3 AlphaFold2 structure prediction and self-consistency evaluation

For all AlphaFold2 predictions, we use the ColabFold implementation with random seed set to 42 [25]. For single-sequence self-consistency tests, we used default settings for ColabFold setting "msa-mode" to "single_sequence". For the shallow-depth MSA self-consistency evaluations, we randomly sample a certain number of sequences to generate synthetic MSAs. We subsampled the natural MSA using hhfilter with a minimum coverage of 80%, maximum sequence identity of 90%, and specifying the "diff" parameter to the depth of the MSA [16]. If the depth exceeded the specified depth, we randomly sampled from the remaining sequences. We selected the top prediction from the 5 models based on maximum pTM.