Seq2MSA: A Language Model for Protein Sequence Diversification

Pascal Sturmfels
University of Washington
psturm@uw.edu

Roshan Rao, Robert Verkuil, Zeming Lin, Ori Kabeli,
Tom Sercu, Adam Lerer, Alexander Rives
Meta AI Research
{rmrao, rverkuil, zlin, orik, tsercu, alerer, arives}@meta.com

Abstract

Diversification libraries of protein sequences that contain a similar set of structures over a variety of sequences can help protein design pipelines by introducing flexibility into the starting structures and providing a range of starting points for directed evolution. However, exploring the sequence space is computationally challenging: the vast majority of sequence space is non-viable, and even of those sequences that do fold to well-formed protein structures, it is challenging to find the fraction that maintain a similar fold class to a given protein. In this work, we propose to use an encoder-decoder language model, trained on a novel Seq2MSA task, that can create diversification libraries of any input protein. In particular, using our model, we are able to generate sequences that maintain structural similarity to a target sequence while pushing below 40% sequence identity to any protein in UniRef. Our diversification pipeline has the potential to aid in computational protein design by providing a diverse set of starting points in sequence space for a given functional or structural target.

1 Introduction

Modeling natural diversity around a target sequence has broad relevance for experimental and computational design of proteins. Generative models of protein sequences have been shown to provide rational single-site mutations that improve the efficiency of directed evolution Hie et al. [2022]. More generally, modeling the space of natural sequences in a family has the potential to not only improve the efficiency of protein design, but also may help design trajectories escape local optimization minima by recapitulating useful modifications learned from evolutionary history. The ability to model the evolutionary landscape around a given sequence allows protein designers to search the manifold of natural sequences to optimize an objective function, while avoiding 'adversarial' parts of sequence space that may be arrived at via unconstrained search.

We present a novel protein language model, Seq2MSA, which is capable of consistently generating highly diverse sequences (sequence identity $< 40\%$) from an initial seed sequence while maintaining structural similarity. The Seq2MSA model is trained with a novel task where the model is provided with a seed sequence and is trained to generate sequences from a Multiple Sequence Alignment (MSA) built from that seed. We validate the model's generations on held out proteins both using a single-sequence folding model as an oracle and aggregation scores derived from rosetta [Alford et al., 2017]. In addition, we confirm that a significant portion of generated sequences are novel both with respect to the entire pool of generated proteins and the language model's entire training set.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Related Work

There exist many wet-lab methods for sequence diversification, often based around single or multi-codon mutagenesis at the DNA level [Liu and Cropp, 2013, Shivange et al., 2009, Arkin and Youvan, 1992, Minshull et al., 2004, Wong et al., 2006, Ruff et al., 2013]. Such methods often consist of multiple rounds of mutation and functional screening and can be both time consuming and expensive.

More closely related to this work, there have been several recent computational approaches for protein sequence diversification, including using a large autoregressive protein language model fine-tuned on specific protein families [Madani et al., 2021], a sequence-to-sequence denoising model [Gligorijevic et al., 2021], and a fitness prediction model combined with multiple iterations of single-site mutations [Bryant et al., 2021]. Unlike existing methods, our model: (1) is general purpose for any protein sequence, and does not need to be fine-tuned on specific families, (2) allows for arbitrary insertions and deletions rather than just single-site mutations and (3) creates a structurally similar but not identical ensemble of sequences. This latter point is important in comparison to inverse-folding methods, which may generate an ensemble of sequences that fold to a specific backbone, but do not allow any flexibility or diversity in that backbone [Hsu et al., 2022, Dauparas et al., 2022].

More broadly, there is a wide range of recent literature on training large language models on protein sequences, including Bepler and Berger [2019], Rives et al. [2021], Rao et al. [2019], Elnaggar et al. [2021], Madani et al. [2020]. In particular, both Madani et al. [2021] and Ferruz et al. [2022] train large autoregressive models to generate protein sequences unconditionally and conditioned on functional tags. In our work, we focus specifically on generating proteins within a family of a given seed protein. Related to our task, Rao et al. [2021] develop a transformer trained on MSAs. Unlike this work, our model does not directly take an MSA as input, and instead seeks to generate sequences from the MSA as output. As a result, our model does not require MSA generation at inference time.

3 Methods

Here we describe the core task and pipeline we use to generate diverse sequence ensembles. Training and architectural details, as well as datasets used, are described in Sections A.1 and A.2.

3.1 Multiple Sequence Alignment

A multiple sequence alignment (MSA) is a core object in biological sequence analysis that groups related proteins in a matrix whose columns represent homologous or structurally related amino acids. In brief, a multiple sequence alignment is typically created by searching a "seed" sequence against a large database of reference sequences using a search algorithm such as HHBlits [Steinegger et al., 2019]. The search tool returns target sequences that are *aligned* to the seed. For each index in the seed sequence, the target sequence can either have a corresponding match state residue at that index, or a gap token indicating that the target sequence is missing a residue at this index. Additionally, residues in the target sequence that do not match to any index in the seed sequence are called inserted residues. See Chatzou et al. [2016] for a more complete overview.

3.2 The Seq2MSA Task

Our proposed task is to generate sequences in an MSA given the seed sequence that was used to generate that MSA. Formally, given a seed sequence and members of its corresponding MSA, at training time we sample a single target sequence from the MSA uniformly at random. The loss at each iteration is the cross entropy between the target sequence and the model output given the seed sequence, where cross entropy is computed over all tokens equally [Vaswani et al., 2017].

We tokenize the seed sequence following the amino acid alphabet from Rives et al. [2021]. There are two variants of the Seq2MSA task depending on how the target sequence is tokenized. In the first variant, called *unaligned*, we remove all alignment information from the target sequence: we discard gap tokens and treat insertions the same as match states. In the second variant, called *aligned*, we maintain the alignment information given in the multiple sequence alignment. We double the vocabulary size by distinguishing between match state and insertion state residues in the target sequence, and add an additional token for gaps. We compare the two variants in Section A.3;

we generally find that using alignment information produces sequences that have higher predicted structural stability and use this variant of the task for all experiments below.

3.3 Diversification Pipeline

We predict the structure of each protein in our evaluation set using a single-sequence folding model as an oracle [Lin et al., 2022]. We discard proteins that have a pTM < 0.8. For the remainder of proteins, we generate 2k proteins using the Seq2MSA model. During generation, we sample directly from the model distribution without using strategies like nucleus or low-temperature sampling to improve quality at the cost of diversity [Vijayakumar et al., 2016]. We find that prompting the model with the first five residues of the seed sequence reduces the prevalence of generations consisting of only small aligned fragments, and do so for all experiments. We filter generations for pTM > 0.8 and TM-score between the predicted structures of the seed and generated sequences (TM-to-seed) > 0.8. We then sort the remaining sequences by increasing sequence identity percentage to the seed, where we define sequence identity percentage using definition 3 from May [2004]. We keep the top 100 proteins farthest away from the seed and discard the remaining generations.

4 Results

Using the diversification pipeline outlined in the previous section, we find that: (1) we are able to diversify a broad range of structurally held-out sequences down to as low as < 30% sequence identity, (2) these sequences are not only diverse relative to the original sequence but also to the model's entire training set, and (3) our model's best generations not only align well to the structure of the original protein, but also show similar characteristics on a range of Rosetta-based aggregation metrics.

4.1 Structural Similarity

Under our folding oracle, our Seq2MSA models generates at least 10 sequences with pTM > 0.8 and TM-to-seed > 0.8 for 103 out of the 117 seed proteins, and over 100 such sequences for 78 of the 117 seed proteins. This indicates that our diversification pipeline is able to generate high quality structures for a majority of proteins it was run on. Moreover, it indicates that our language model can generate sequences that preserve the structure of the input seed sequence.

To gain more confidence in the predicted structures of our generated sequences, we compute two metrics of stability for our generated proteins and compare them to original seed sequence: the SAP score and the hydrophobic solvent accessible surface area (SASA). Both metrics are based on the number of hydrophobic residues exposed at the surface of the protein and indicate aggregation propensity of the protein. See Coventry [2021] for a complete definition. Figure 1 shows that both metrics correlate well between each seed protein and its respective generations, and also indicate that the generated proteins tend to expose only a small proportion of hydrophobic residues, which we would expect of proteins with clearly defined secondary structure elements. We run additional sanity checks of our single-sequence folding oracle against AlphaFold 2 (AF2) in Section A.5 - in general we find high agreement between our oracle and AF2 in single sequence mode.

4.2 Sequence Diversity

Out of 117 proteins, our pipeline is able to successfully generate diverse sequence ensembles for more than half of them: counts are in Table 1 and Figure 5. For comparison, several recent methods for diversification using deep models fail to generate any functional sequences at this sequence identity range [Madani et al., 2021, Gligorijevic et al., 2021]. Figure 5 depicts a more detailed breakdown of the number of sequences that pass our computational criteria. Figure 2 depicts a representative example: an anti-cD40 DARPin protein domain (pdb id: 7P3I_2, chain B). In particular, this example demonstrates that the generated proteins preserve major secondary structure elements, but contain diversity in the variable loop regions, including variations in both length and placement.

To more stringently verify that our sequences are novel, we perform three sanity checks. First we verify that the sequence identities computed by our aligned model are close to the optimal alignment. Second, we verify that each generated sequence is diverse with respect to the entire pool of generated

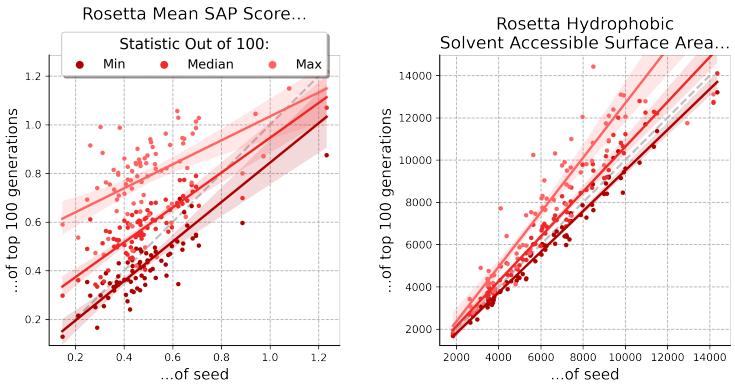

Figure 1: We run the diversification pipeline described in Section 3.3, which leaves 100 generations per seed sequence that satisfy pTM > 0.8 and TM-to-seed > 0.8. We sort the generations under each seed by the respective metric, and plot points for the minimum, median and maximum values on the y axis. The generated sequences share similar characteristics to their seed sequence, indicating the generated sequences will have stable structures similar to their respective seed sequences.

Sequence Identity Cutoff	$\leq 20\%$	$\leq 40\%$	$\leq 60\%$	$\leq 80\%$	$\leq 100\%$
Count	12	63	81	100	103

Table 1: The number of sequences in the evaluation set for which our model generates at least 10 sequences (out of 2000) that satisfy pTM > 0.8, TM-to-seed > 0.8 and the given sequence identity cutoff, out of 117 total sequences in the evaluation set.

sequences. Third, we verify that each generated sequence is far away from any protein the language model has seen during training time.

We compute sequence identities with respect to the model's output alignment, but there may exist a better alignment with a higher sequence identity percentage. After running the diversification pipeline, we re-align each of the 100 generations to their seed using the pairwise aligner *needle* [Madeira et al., 2022]. The left plot in Figure 3 demonstrates close agreement between the model's output alignment and the optimal pairwise alignment.[1] We then take the same 100 generated sequences from the diversification pipeline above and compare them to all 2k sequences originally generated for that seed in the initial step of the pipeline. The middle plot in same figure demonstrates that if a generated sequence is far from the seed used to generate it, it will also be far in sequence space from any other generated sequence. Finally, we use *blastp* to search every generated sequence against Uniref90 - the model's training set - and report the maximum sequence identity for each generation to any sequence in that set [Altschul et al., 1997]. The right plot in the same figure indicates that if a generated sequence is far from the seed used to generate it, it will also be far away from any protein the model has seen during training.

5 Discussion

We introduce a novel protein language model capable of generating diversification libraries for protein sequences. Our pipeline, which combines the novel Seq2MSA model with a single sequence folding oracle, succeeds at diversifying a large number of protein sequences outside of its training set down to low sequence identity percentages. We verify that the designs are not only novel in the ensemble, but also with respect to all of the model's generations and it's entire training set, and verify that the resultant proteins have similar in silico aggregation characteristics as the seeds used to generate them. Such a model may prove useful for protein design: design trajectories that are stuck in local optimum of single site changes may benefit from a library of sequences that contain large sequence changes, both in identity and in length.

[1]Note that the optimal pairwise alignment does not mean the alignment with the greatest sequence identity percentage, so the model may learn alignments that have greater percent identities than the optimal one.

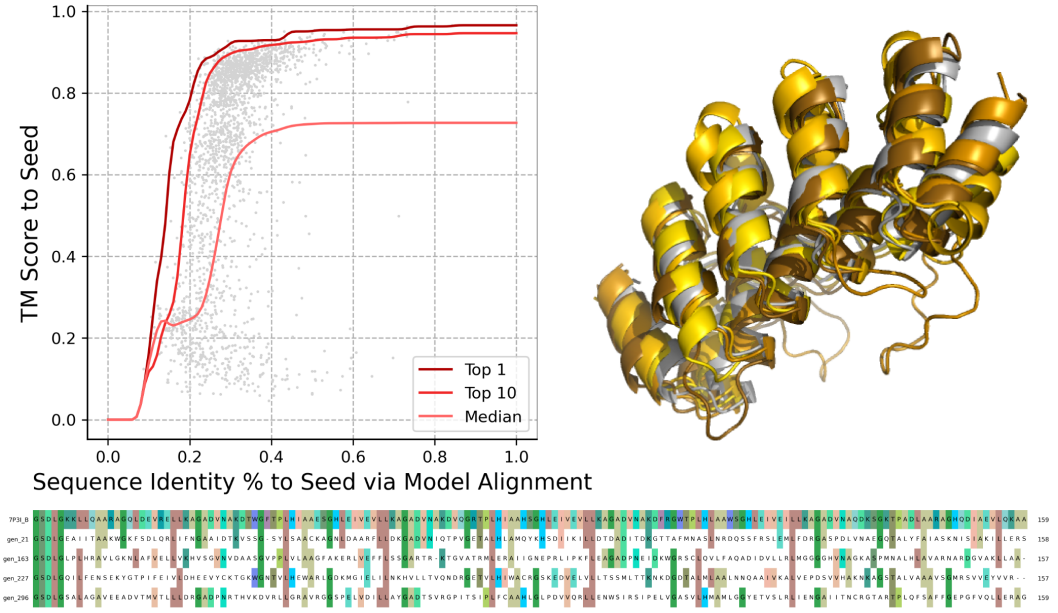

Figure 2: The tradeoff between sequence identity and structural alignment to the original seed using the Seq2MSA model. The lines indicate the TM-score of the top-1, top-10 and median generation up to the given sequence identity percentage, and the gray dots are individual generations. Note that the lines are smoothed with a Gaussian filter. The structural ensemble and the corresponding alignment show the top 4 generated proteins with sequence identity < 30% and greatest alignment score to the original protein, which appears as the gray structure and the first sequence in the alignment.

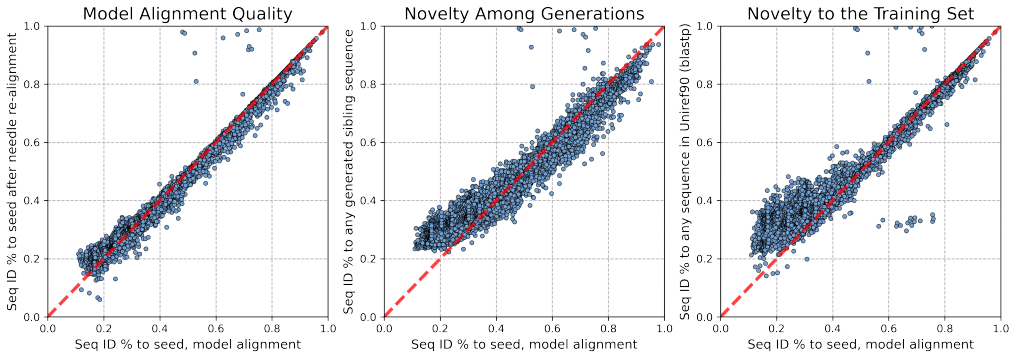

Figure 3: A comparison between the sequence identity of generations to their seed sequence with other metrics of novelty. The left plot compares sequence identities between model generated alignments and the optimal dynamic programming alignment. The middle plot shows the percent identities between any generated sibling sequences, e.g. those sequences that were generated by the same seed. The right plot shows percent identities between generations and their closest Uniref90 blastp hit. All three metrics show a clear correlation to the percent identity between generation and its original seed, implying that if a generation is far from its seed sequence it is in fact truly novel.

5.1 Limitations

Although our model's generations seem to fold well under in-silico metrics, a more complete evaluation would include *in vitro* folding metrics. A more complete in silico evaluation would also include comparisons to existing deep learning models such as those provided by Madani et al. [2021], Gligorijevic et al. [2021], Hsu et al. [2022], Dauparas et al. [2022]. Finally, we are predominantly interested in this model for the purposes of design. Further experimentation would include using the generated ensemble during directed evolution to redesign proteins for specific, functional goals.

References

Rebecca F Alford, Andrew Leaver-Fay, Jeliazko R Jeliazkov, Matthew J O'Meara, Frank P DiMaio, Hahnbeom Park, Maxim V Shapovalov, P Douglas Renfrew, Vikram K Mulligan, Kalli Kappel, et al. The rosetta all-atom energy function for macromolecular modeling and design. *Journal of chemical theory and computation*, 13(6):3031–3048, 2017.

Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

Ivan Anishchenko, Samuel J Pellock, Tamuka M Chidyausiku, Theresa A Ramelot, Sergey Ovchinnikov, Jingzhou Hao, Khushboo Bafna, Christoffer Norn, Alex Kang, Asim K Bera, et al. De novo protein design by deep network hallucination. *Nature*, 600(7889):547–552, 2021.

Adam P Arkin and Douglas C Youvan. An algorithm for protein engineering: simulations of recursive ensemble mutagenesis. *Proceedings of the National Academy of Sciences*, 89(16):7811–7815, 1992.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, 39(6):691–696, 2021.

Maria Chatzou, Cedrik Magis, Jia-Ming Chang, Carsten Kemena, Giovanni Bussotti, Ionas Erb, and Cedric Notredame. Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics*, 17(6):1009–1023, 2016.

Brian Coventry. *Learning How to Make Mini-Proteins that Bind to Specific Target Proteins*. University of Washington, 2021.

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*, page eadd2187, 2022.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE transactions on pattern analysis and machine intelligence*, 2021.

Noelia Ferruz, Steffen Schmidt, and Birte Höcker. Protgpt2 is a deep unsupervised language model for protein design. *Nature communications*, 13(1):1–10, 2022.

Vladimir Gligorijevic, Daniel Berenberg, Stephen Ra, Andrew Watkins, Simon Kelow, Kyunghyun Cho, and Richard Bonneau. Function-guided protein design by deep manifold sampling. *bioRxiv*, 2021.

Brian L Hie, Duo Xu, Varun R Shanker, Theodora UJ Bruun, Payton A Weidenbacher, Shaogeng Tang, and Peter S Kim. Efficient evolution of human antibodies from general protein language models and sequence information alone. *bioRxiv*, 2022.

Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Jia Liu and T Ashton Cropp. Rational protein sequence diversification by multi-codon scanning mutagenesis. In *Enzyme Engineering*, pages 217–228. Springer, 2013.

Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

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. Deep neural language modeling enables functional protein generation across families. *bioRxiv*, 2021.

Fábio Madeira, Matt Pearce, Adrian RN Tivey, Prasad Basutkar, Joon Lee, Ossama Edbali, Nandana Madhusoodanan, Anton Kolesnikov, and Rodrigo Lopez. Search and sequence analysis tools services from embl-ebi in 2022. *Nucleic acids research*, 50(W1):W276–W279, 2022.

Alex CW May. Percent sequence identity: the need to be explicit. *Structure*, 12(5):737–738, 2004.

Jeremy Minshull, Sridhar Govindarajan, Tony Cox, Jon E Ness, and Claes Gustafsson. Engineered protein function by selective amino acid diversification. *Methods*, 32(4):416–427, 2004.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

Xavier Robin, Juergen Haas, Rafal Gumienny, Anna Smolinski, Gerardo Tauriello, and Torsten Schwede. Continuous automated model evaluation (cameo)—perspectives on the future of fully automated evaluation of structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 89(12):1977–1986, 2021.

Anna J Ruff, Alexander Dennig, and Ulrich Schwaneberg. To get what we aim for–progress in diversity generation methods. *The FEBS journal*, 280(13):2961–2978, 2013.

Amol V Shivange, Jan Marienhagen, Hemanshu Mundhada, Alexander Schenk, and Ulrich Schwaneberg. Advances in generating functional diversity for directed protein evolution. *Current opinion in chemical biology*, 13(1):19–25, 2009.

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*, 20(1):1–15, 2019.

Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

Tuck S Wong, Daria Zhurina, and Ulrich Schwaneberg. The diversity challenge in directed protein evolution. *Combinatorial chemistry & high throughput screening*, 9(4):271–288, 2006.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, 2022.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004.

A Appendix

A.1 Training Setup

Our training architecture is a 12-layer encoder-decoder sequence-to-sequence transformer [Vaswani et al., 2017] with architecture parameters taken from Lewis et al. [2019]. We trained for a total of 250k steps with a batch size of 2^{21} tokens per step. Data-parallel training was performed on a total of 128 Nvidia Volta GPUs, and took approximately one week. We used an initial learning rate of $3e^{-4}$ with a linear warmup from 0 for the first 2k steps, and then a polynomial decay to $4e^{-4}$ by the end of training.

A.2 Datasets

We create a large training dataset of MSAs by searching every protein sequence in the 2020 release of UniRef50 [Suzek et al., 2007] against every sequence in (the 2020 release of) Uniref90 using HHBlits [Steinegger et al., 2019] with the default parameters. From each MSA we randomly selected a maximum of 1024 sequences to be part of the training set due to disk space constraints.

As an evaluation set, we use the proteins from the 04/01/22-06/25/22 release of the CAMEO [Robin et al., 2021]. The structures for these proteins are held-out from our structure prediction oracle. We predict the structure of each protein and discard those proteins with predicted TM-score [Zhang and Skolnick, 2004] (pTM) < 0.8, which has generally been found to be indicative of good performance for single sequence folding models [Lin et al., 2022, Wu et al., 2022]. This leaves an evaluation set of 117 sequences from both easy and medium difficulties.

A.3 Generation Set Up

Figure 4 compares the effects of training a model with alignment information vs. without, as well as prompting the model at decoding time with the first five residues of the seed sequences vs. sampling with no prompting. As expected, adding alignment information into the model generally improves the predicted TM-score of the sequences under our folding oracle, which is a proxy for how well the sequence folds in vivo Lin et al. [2022]. Sequence alignment also helps the generated sequences better align structurally to their original seed. Fixing the first five residues at decoding time also increases both structural metrics, at the cost of increasing the percent sequence identity to the seed. Since our model was trained on very diverse MSAs (mean sequence identity to the seed 20% since many of the sequences found in natural MSAs are fragments that only match portions of the original seed), prompting is an ad hoc fix to keep generations slightly closer to the seed. A more complete solution would re-train our original model on MSAs that are closer to their original seed sequence.

A.4 Full Counts on the Evaluation Set

Table 1 and Figure 5 shows a more complete picture of our diversification efforts, on each of the proteins in the evaluation set. The figure shows an interesting and sharp tail off: for some sequences, we are able to generate many (> 100) sequences that have high structural confidence under a folding oracle and low sequence identity percentage to the original seed sequence. For many others, none of the 2000 generated proteins per seed fulfils this criteria. It is not immediately obvious why certain proteins are "harder" than others to diversify - the success and failure cases both contain a variety of structures (helix only, sheet only, and mixed helix/sheet), as well as a variety of lengths.

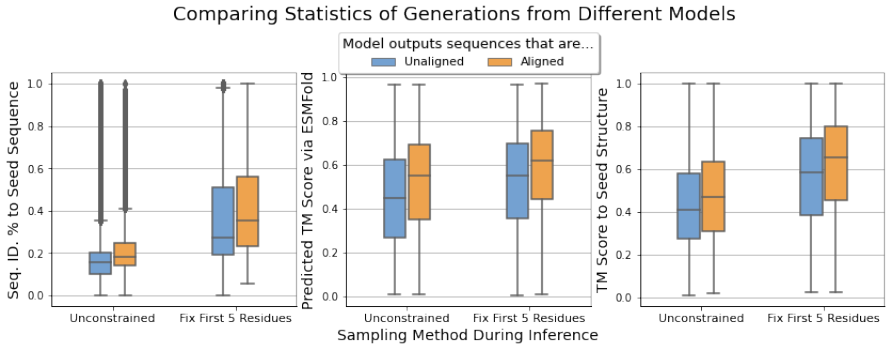

Figure 4: A comparison of different generation methods, and different ways of training the original Seq2MSA model. Adding alignment information and fixing the first five residues at decoding time increase confidence in the predicted structures of the generated sequences, and the structural alignment between the generated sequences and the seed used to generate them.

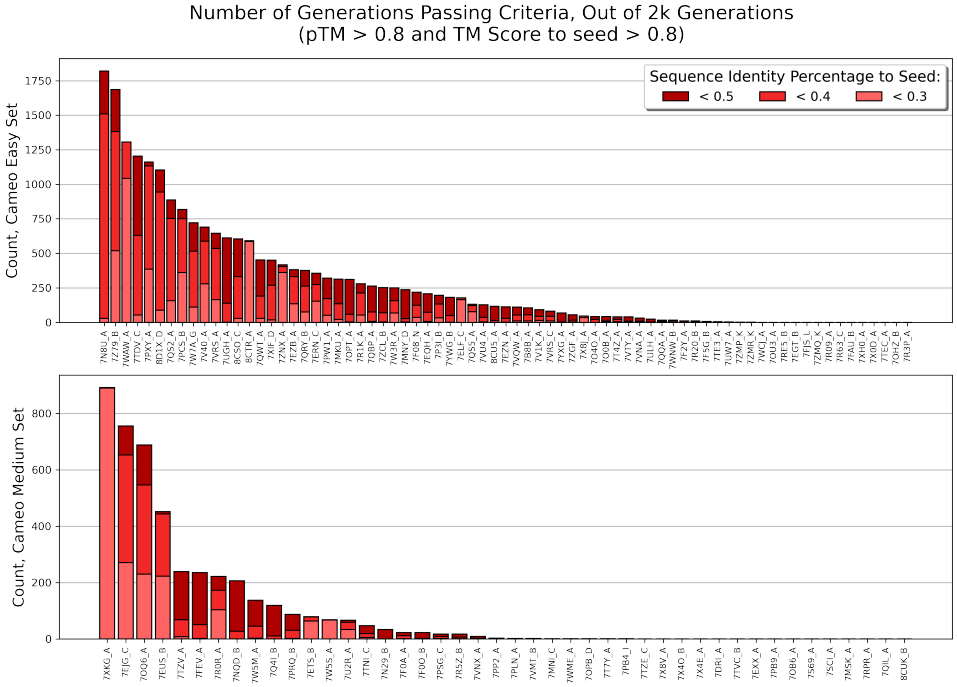

Figure 5: The number of generated sequences passing specified diversification criterion, out of 2000 generations per seed sequence. The seed sequences in the evaluation seed have PDB ids that are listed (ID and chain) on the x axis. Our model successfully generates 100s of structurally similar structures with diverse sequences for some proteins, and fails to generate any for others.

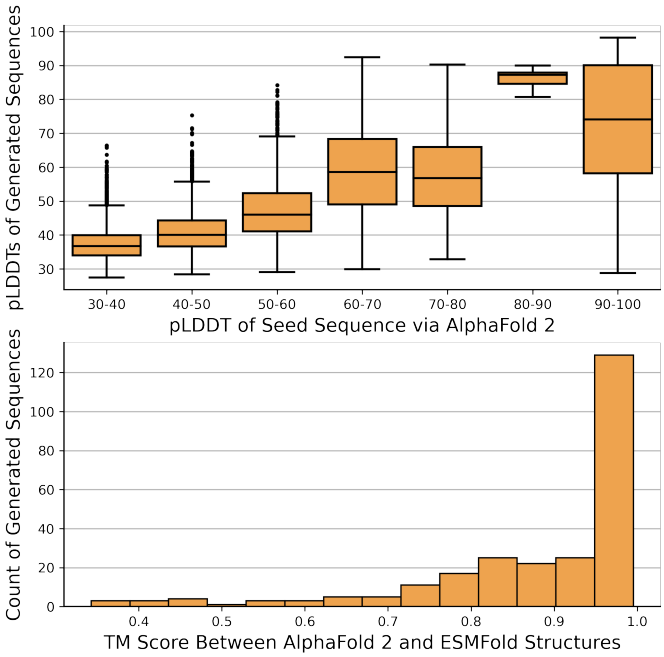

Figure 6: Validating our single-sequencing folding oracle by comparing it against AlphaFold 2 in single sequence mode. It is possible that our sequences simply "look" like natural sequences in a family because they have similar distributions and placements of residues to the MSAs they were trained on, and that our single sequence folding oracle is predicting their structure solely based on these family features rather than what the sequence would actually fold to in vitro. However, we find that not only can AlphaFold 2 in single sequence mode fold many of our generations just as well as it folds the natural seed protein used to generate them, we also find that the structures between AF2 and the single sequence oracle have high agreement when AlphaFold 2 has pLDDT > 80. This gives confidence that our sequence structures are in fact stable and structurally similar to their original seeds.

A.5 Agreement Between Folding Methods

In order to increase the confidence in our sequences folding to their predicted structure, we take the top 100 generated sequences for each seed sequence (that pass pTM > 0.8 and TM-to-seed > 0.8 under our single sequence model) and fold it using AlphaFold 2 [Jumper et al., 2021] in single sequence mode: no templates and no MSA. AlphaFold in single sequence mode leverages far less evolutionary information than our single sequence model, since AlphaFold gets that information from MSAs while our single sequence model memorizes evolutionary information during language model pre-training [Lin et al., 2022]. The top of Figure 6 shows the results. In general, if AlphaFold 2 in single sequence mode folds the seed sequence well, it tends to also have high pLDDT of the generations, which adds confidence that our generated sequences really do fold to their predicted structures and are not adversarial examples leveraging MSA information.

For the sequences generated by our Seq2MSA model that satisfy both pLDDT under AlphaFold > 80 and pTM under our single sequence model > 0.8, we compare the TM-score between structures predicted by AlphaFold in single sequence mode and structures predicted by our single sequence model. The result is in the bottom of Figure 6. The histogram shows that the majority of high-confidence predictions by both models also have high structural agreement, further raising confidence in the predicted structures of each generated sequence.

A.6 Hallucinations

In order to test whether our pipeline could generalize to de novo sequences, we download the hallucinated sequences from Anishchenko et al. [2021]. We select only those sequences that were

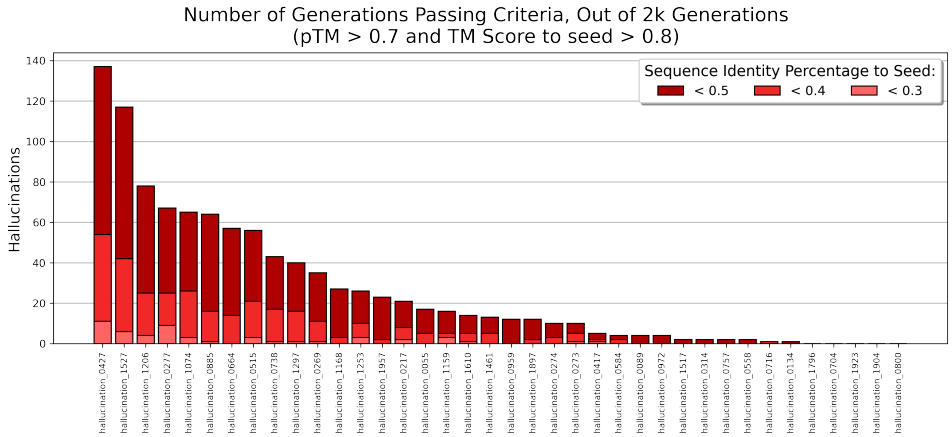

Figure 7: The number of generated sequences passing diversification criterion, when applying the Seq2MSA model on completely de novo proteins. Remarkably, even though our model was trained only on natural proteins, it manages to succesfully generate diverse ensembles for many of the de novo designs.

experimentally characterized by size-exclusion chromatography, and further filter by folding oracle pTM > 0.7. We reduce the cutoff from 0.8 since only two hallucinations passed pTM > 0.8. We then run the same pipeline from the main text on each denovo protein. The results are in Figure 7. The plot clearly shows that de novo proteins are harder than natural proteins - there are far fewer generations per sequence satisfying the diversification criterion. However, the plot also shows that for many de novo sequences, we are able to generate diverse ensembles at the < 0.5 and even < 0.4 sequence identity level.