Uncovering sequence diversity from a known protein structure

Luca Alessandro Silva[1], Barthelemy Meynard-Piganeau[1,2,3], Carlo Lucibello[1], and Christoph Feinauer[1]

[1]Department of Computing Sciences, Bocconi University, Milan, Italy
[2]Politecnico di Torino, Milan, Italy
[3]Sorbonne Université, Institut de Biologie Paris Seine, Biologie Computationnelle et Quantitative LCQB, Paris, France

Abstract

We present InvMSAFold, a method for generating a diverse set of protein sequences folding into a single structure. For a given structure it defines a probability distribution over the space of sequences. This distribution captures second-order correlations observed in Multiple Sequence Alignments (MSA) of homologous proteins. Our innovation lies in generating highly diverse protein sequences while preserving structural and functional integrity. This approach offers exciting prospects, particularly in directed evolution, by providing diverse starting points for protein design.

1 Introduction

Inverse folding aims to predict amino acid sequences that fold into a given protein structure and plays a fundamental role for example in the protein design pipeline of RFDiffusion [1]. Recent deep learning approaches such as ESM-1F [2] or ProteinMPNN [3] achieve remarkable accuracies in this task. However, instead of predicting a single ground truth sequence, it is often desirable to have a model that is able to generate a variety of different sequences, for example starting from a source sequence [4, 5] and taking different molecular environments into consideration [6]. Such approaches allow to expand the sequence design space while preserving structural consistency, allowing for a larger pool of sequence when selecting for additional properties like thermostability, solubility or toxicity,

In this work, we present a method that is able to generate diverse protein sequences given a structure, including sequences far away from the natural sequence. Our method is potentially applicable in various domains. In drug discovery, for example, it would allow for the generation of large amount of diverse candidates, enabling further selection optimized for properties like bioavailability. Similarly, in biotechnology and enzyme engineering, it could facilitate the creation of enzymes with tailored properties, such as improved stability and activity under varying conditions.

Recent architectures for inverse folding are based on encoder-decoder architectures, where a structure is encoded and a sequence decoded. In our approach, we use the decoder instead for generating the parameters of a pairwise probability distribution over amino acids in a sequence [7], motivated by evidence that such pairwise models capture the relevant sequence statistics for generating novel protein sequences [8] for a given structure. While we use MSAs of families of homologous proteins when training, during inference we only use the structure of a protein in order to generate such a pairwise model. The generated pairwise model itself is very light-weight and can be used for rapidly generate a large diversity of sequences for the input structure.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

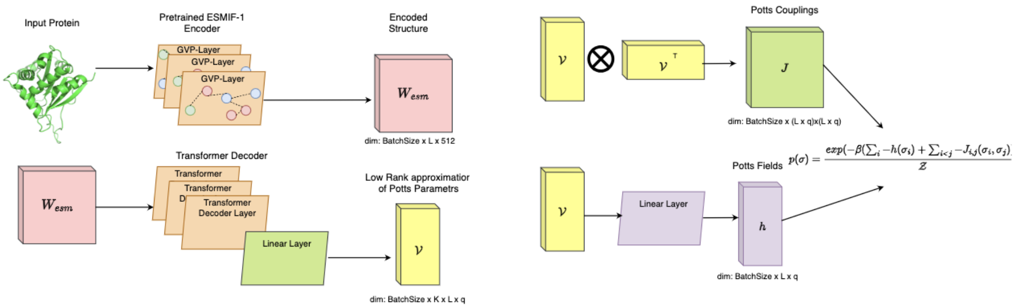

Figure 1: Left: Decoder architecture up to the outputting of the low-rank tensor $\mathcal{V}$. Right: how $\mathcal{V}$ is used to produce couplings and fields for the Potts model.

We show that the models we generate are able to create sets of sequences that capture the diversity of the protein family better than other models and are able to find sequences far away from the natural sequence that are predicted to still fold into the same structure.

2 Methods

In this Section, we describe the components of our architecture, InvMSAFold, and its training procedure. InvMSAFold solves the inverse protein folding problem by mapping a structure X, given as residue positions, into a probability distribution over sequences, $p_\theta(s \mid X)$. Briefly, InvMSAFold is a deep neural network composed of graph convolutions and attention layers that outputs the couplings $J_\theta(X)$ and fields $h_\theta(X)$ of a Potts model. Sequences are then modeled by the Potts pairwise distribution in (1), so that $p_\theta(\sigma \mid X) = p(\sigma \mid J_\theta(X), h_\theta(X))$.

The Potts model for proteins. Consider a multiple sequence alignment (MSA) of length L, with sequences $\sigma = (\sigma_1, \ldots, \sigma_L)$ having components $\sigma_i \in \{1, \ldots, q\}$, where $q = 21$ is the size of the alphabet (number of aminoacid types + blank). the Potts model is a simple model with pairwise interaction for the statistic of sequences. It can be derived from a maximum entropy principle when imposing as the only constraints that single-site frequencies and two-site correlations are reproduced. Despite its simplicity, it has proven to be highly effective in capturing the full statistical structure of MSA families, thanks to the co-evolutionary principles motivating it. The model is defined by a vector (called field) of q components for each site, $h_i(a)$, and by a symmetric matrix (called coupling) of $q \times q$ elements, $J_{ij}(a, b)$, for each pair of sites (i, j). We will denote with h and J the tensors of size $L \times q$ and $L \times L \times q \times q$ containing all fields and all couplings. The sequence probability according to the Potts model is then given by

$$p(\sigma|J, h) = \frac{1}{Z_{J,h}} e^{\sum_{i<j} J_{ij}(\sigma_i,\sigma_j)+\sum_i h_i(\sigma_i)}, \tag{1}$$

with $Z_{J,h}$ a normalization factor. Couplings and fields are typically fitted on a protein family using Boltzmann sampling [9], pseudolikelihood maximization ([10], see later), or mean-field methods. In this work, instead of fitting a separate model for each different family, we will train a neural network to produce, conditional on the target structure, coupling, and fields modeling a generic family.

The InvMSAFold architecture. InvMSAFold is a feedforward neural network composed of two parts, the structure encoder and couplings/fields decoder.

For the encoder, we select the one from the ESM-IF1 model [2] which follows the GVP-GNN architecture of Ref. [11]. Crucially, the encoder gives a rotationally invariant representation Y of the input X. Thanks to this invariance, the decoder part of the architecture doesn't have to worry about symmetry consistency in producing couplings and fields.

The decoder takes as input batches of encoded structures. Different lengths in the batch are accommodated by zero-padding up to a common length L. The sequence of L tokens of length D-dimensional ($D = 512$ in our experiments) is fed to 6 transformed layers with 8-heads self-attention. Finally, we apply position-wise a two-layer MLP that we reshape into a $K \times q$ matrices, with q the alphabet size and K the arbitrarily chosen rank of our coupling representations, smaller than the typical length L.

Calling v the $L \times K \times q$ tensor output of previous operations, we use it to obtain a low-rank representation of the coupling matrices as follows:

$$J_{ij}(a,b) = \sum_{k=1}^{K} v_i^k(a) v_j^k(b). \tag{2}$$

Our architecture is therefore able to infer the Potts parameters from a structure of arbitrary length L. In a standard setting where a Potts model is fitted to a single MSA, low-rank decompositons of J have been shown to be as effective as the full-rank counterparts [12]. We then have to train our model by maximizing the log likelihood of the infered Potts model on the complete MSA of sequences homologous with the native one of the structure.

Pseudolikelihood inference The log-likelihood of (1) is a concave function in the parameters. As a consequence, the global maximum is attainable by simple optimization methods such as gradient ascent. Gradient existimation is expensive though and requires Markov Chain Monte Carlo sampling from the model itself. An alternative objective function is given by the log-pseudolikelihood, which involves the minimization of

$$\mathcal{L}(J,h) = \mathbb{E}_{\sigma \sim D} \frac{1}{L} \sum_{i=1}^{L} \log p(\sigma_i \,|\, \sigma_{\setminus i}, J, h). \tag{3}$$

where D is a training set of sequences. The pseudo-likelihood estimator is a consistent and efficient estimator [10]: if sequences in the MSA are samples drawn from a Potts distribution, its parameters are recovered at an optimal rate in the limit of large training set size. In our setting, we train the network by computing the gradients $\nabla_\theta \mathcal{L}$, that is backpropagating the gradient through J and h to the parameters of InvMSAFold. To encourage diversity in the generating sequences, at training time we augment each example (X, σ) with a set of homologous sequences that we use to compute the loss (3). To avoid overfitting, we also add an L2 penalty on the couplings and fields in (3).

Alternative Pairwise Modeling: ArDCA. The Potts model, while being effective in reproducing the statistics of MSAs, requires running MCMC for sampling at inference time. As an alternative inferred model that allows autoregressive sampling and exact likelihood evaluation, we consider ArDCA from Ref. [13]. The parameters defining ArDCA, still takes the form of coupling J and field h, but now used to produce conditional probabilities in the form $p(s_i|s_{<i}, J_{i,<i}, h_i)$. The couplings in ArDCA are slightly different than in DCA, and cannot be interpreted as *direct couplings*. Moreover, the further away we go in the sequence the more couplings are to be summed with the site fields. This causes the strength of the coupling to decrease as we increase the position in the sequence. This scaling property makes the coupling matrix full rank, and therefore hard to approximate with our model. To counter this effect, in the ArDCA implementation, after generating J according (2), we rescale it as $J_{ij} \leftarrow \frac{J_{ij}}{max(i,j)}$. The choice of a scaling inversely proportional with the position, comes from observation ArDCA from [14].

3 Results

3.1 Data Collection and Preprocessing

To train and evaluate models, as common in the Inverse Folding literature, we rely on the curated CATH [15] 4.2 40$\%$ non-redundant data set, in which structures are grouped by their CATH(class, architecture, topology/fold and homologous superfamily) classification. Such a hierarchical classification of protein domains based on their folding patterns has proven to be fundamental in many fields. In particular we leverage it to evaluate the ability of our models to generalize across different protein folds. We split the CATH dataset into a train dataset and three testing dataset, which we label respectively *sequence, structure and superfamily*. 80$\%$ of the sequences are divided into train and sequence testing data set: the latter contains those structures of which we have seen an homologous during the training. The structure test data set, which contains 10$\%$ of sequences, has structures not seen during training, but of which we have seen a structure belonging to the same superfamily.

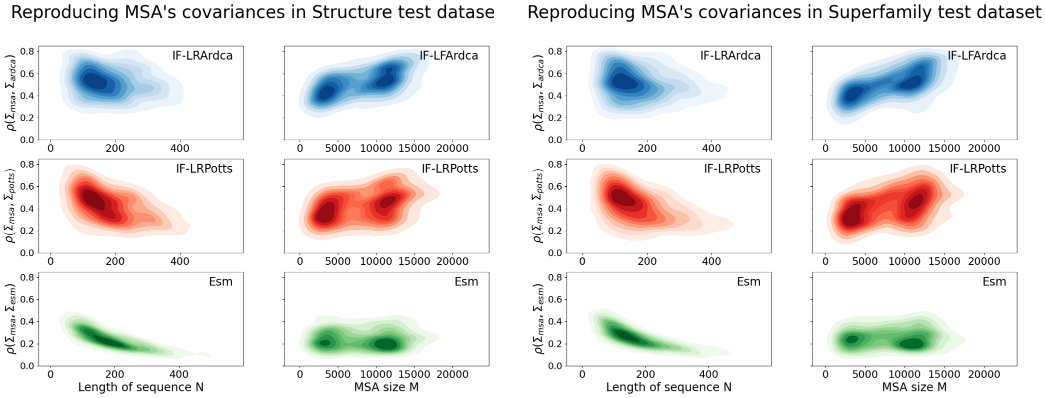

Figure 2: Comparing the correlations between generated and true samples for ArDCA(blue), potts(red) and ESM-IF1(green). On the left we have the results for the sequence in the superfamily test dataset, on the right of the structure test dataset. For each dataset, the plot on the left has on the x-axis the length of the sequence N, while the one on the right has the MSA size M

Finally, in the superfamily test dataset we have sequences of which we have not even seen an element belonging to the same superfamily during training: these sequences hence have very different folds from the one seen during training.

Once we have divided the structures of CATH as detailed above, for each structure we generate a MSA searching for homologous sequences inside the comprehensive Uniprot50 data set. To obtain the MSAs we leverage the MMseqs2 (Many-against-Many sequence searching) software, which allows for accurate and fast generation of many MSAs.

3.2 Second Order Reconstruction

We tested the three different models in their ability to reproduce the covariance matrix of the MSA for input sequences belonging to the structure and super-family test data sets. The experimental procedure for both datasets is the following: 1) For every sequence whose MSA had at least 2000 samples we obtained the generative model for the three different models. 2) To generate the samples from Potts we leverage the efficient library bmDCA, for ESM-IF1 we used the built in sampler and for ArDCA we built our own sampler. Since getting samples from ESM-IF1 was much slower, for every sequence is ESM-IF1 we get 900 samples, while for the latter two we get as many samples as sequences in the MSA. We don't think the result are sensitive to this fact. 3) For ESM-IF1, given the samples, we re-aligned the samples using the full MSA to get a fair comparison. 4) Given the samples, we compute the covariance matrices of the generated samples and the one of the true MSA. We then compute the Pearson correlation between the flattening of the two.

Once we have these correlation values for the two test datasets, we compute the kernel density estimation of them against the sequence length N and the MSA size M. As we can see from Figure 2, ArDCA and Potts do both significantly better than ESM-IF1. ArDCA seems also to dominate potts; especially its performance does not seem to deteriorate with the lenght of the input sequence, while both potts and ESM-IF1 exhibit this feature. Both ArDCA and potts do better as the original MSA size is larger, while ESM-IF1(as expected) is not influenced.

3.3 Iso-Structure Exploration

In our pursuit of understanding the relationship between protein sequence variations and structural deformations, we designed a controlled experiment that systematically explores the tolerance of a protein structure to increasing sequence dissimilarity. This experiment provides a unique perspective on protein structural robustness and represents a novel application of state-of-the-art computational tools.

For a given protein structure of interest, we initiated the experiment by generating a spectrum of sequence variants. We systematically increased the hamming distance of these variants from the

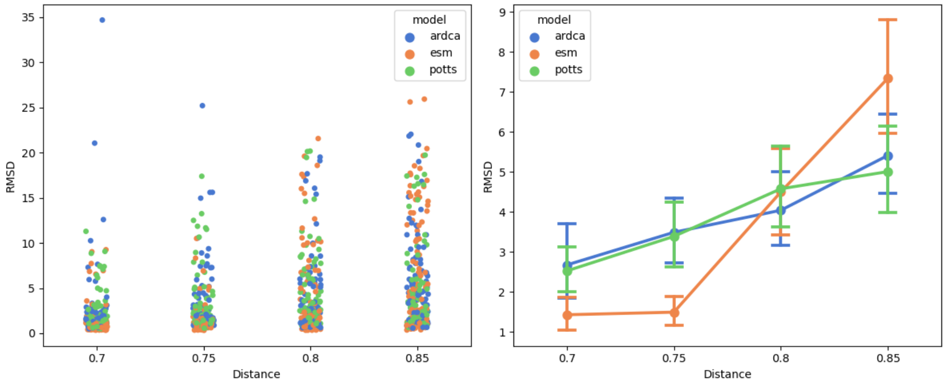

Figure 3: comparison of the quality of the generated sequence under the constraint of increasing hamming distance (normalized by protein length) from the native sequence. We started aggregated the score of 9 pdbs from the superfamily set, and sampled from ESM-IF1 and our methods. We then refold the sequence with alphafold and compare the refolded structure with the original one using the RMSD. We observe the ESM-IF1 is not able to generate good sequence far away fro the native one.

original native sequence. This divergence was achieved by progressively altering amino acids, thereby introducing increasing levels of sequence dissimilarity while preserving the protein's overall fold.

To assess the impact of these sequence variations on the protein structure, we employed a temperature-based exploration approach. Starting from sequences with minimal dissimilarity from the native sequence, we gradually elevated the temperature. This step allowed us to reach sequences that were further and further removed from the native sequence, effectively increasing the structural diversity under consideration.

Following the generation of these sequence variants, we utilized the AlphaFold protein folding model [16, 17] to refold the sequences. AlphaFold has demonstrated exceptional capabilities in predicting protein structures accurately. We leveraged its predictive power to refold the diverse set of sequences generated at different levels of sequence divergence.

To evaluate the structural fidelity of the refolded sequences, we conducted a comparative analysis with the ESM-IF1 model, a prominent computational tool in structural bioinformatics. Specifically, we measured the structural similarity between the refolded sequences and the native structure using the RMSD 3. Alphafold was used with no templates, and mmseq2 was used for the msa.

4 Discussion

Our findings reveal that our approach consistently outperforms the ESM-IF1 model in preserving the native structure of the protein for sequences that are far removed from the original native sequence. The RMSD score and pLDDT metrics consistently demonstrate a higher level of structural fidelity in the refolded sequences generated using our approach.

This experimental setup and comparative analysis underscore the effectiveness of our methodology in exploring the structural consequences of sequence diversity, emphasizing its potential in applications ranging from protein engineering to understanding the adaptation of proteins in diverse environments.

Our study marks a significant shift in the domain of inverse folding, where the conventional focus was on decoding the original sequence from a known structure. In contrast, our research addresses a more realistic scenario where we possess the native sequence but seek to generate alternative sequences. This novel approach, aimed at expanding sequence diversity while preserving the fold, carries profound implications for a range of applications. By systematically generating alternative sequences for a given protein structure, we open new avenues for filtering and selecting sequences based on desired properties, such as improved thermostability, altered substrate specificity, or reduced toxicity.

References

[1] 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, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

[2] 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.

[3] 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.

[4] Pascal Sturmfels, Roshan Rao, Robert Verkuil, Zeming Lin, Ori Kabeli, Tom Sercu, Adam Lerer, and Alexander Rives. Seq2msa: A language model for protein sequence diversification.

[5] 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.

[6] Lucien Krapp, Fernado Meireles, Luciano Abriata, and Matteo Dal Peraro. Context-aware geometric deep learning for protein sequence design. *bioRxiv*, pages 2023–06, 2023.

[7] Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018.

[8] William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502):440–445, 2020.

[9] Matteo Figliuzzi, Pierre Barrat-Charlaix, and Martin Weigt. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Molecular Biology and Evolution*, 35(4):1018–1027, 01 2018.

[10] Magnus Ekeberg, Cecilia Lövkvist, Yueheng Lan, Martin Weigt, and Erik Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.

[11] Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael JL Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. *arXiv preprint arXiv:2009.01411*, 2020.

[12] Simona Cocco, Rémi Monasson, and Martin Weigt. Inference of hopfield-potts patterns from covariation in protein families: calculation and statistical error bars. In *Journal of Physics: Conference Series*, volume 473, page 012010. IOP Publishing, 2013.

[13] Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature communications*, 12(1):1–11, 2021.

[14] Simone Ciarella, Jeanne Trinquier, Martin Weigt, and Francesco Zamponi. Machine-learning-assisted monte carlo fails at sampling computationally hard problems. *Machine Learning: Science and Technology*, 4(1):010501, 2023.

[15] Ian Sillitoe, Nicola Bordin, Natalie Dawson, Vaishali P Waman, Paul Ashford, Harry M Scholes, Camilla SM Pang, Laurel Woodridge, Clemens Rauer, Neeladri Sen, et al. Cath: increased structural coverage of functional space. *Nucleic acids research*, 49(D1):D266–D273, 2021.

[16] 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.

[17] Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature methods*, 19(6):679–682, 2022.