AutoFoldFinder: An Automated Adaptive Optimization Toolkit for De Novo Protein Fold Design

Shuhao Zhang*
Carnegie Mellon University
shuhaoz2@andrew.cmu.edu

Youjun Xu †
Peking University
yjxu@pku.edu.cn

Jianfeng Pei
Peking University
jfpei@pku.edu.cn

Luhua Lai
Peking University
lhlai@pku.edu.cn

Abstract

Although an explosive number of protein structures are revealed each year, the number of basic protein architecture - protein folds - stays stable. Because of the determining relationship between function and structure, it remains highly interesting to scientists to explore protein structure space and subsequently enrich the diversity of protein function space. Current protein structure exploration approaches either rely on sampling of natural protein fragments or require human-crafted constraints. To facilitate more emancipated structure space probing, we present an automated adaptive optimization toolkit for de novo protein fold design - AutoFoldFinder. We also further introduce CM-align to better quantify structure map similarity in the optimization process. Our results indicate a higher efficiency to produce novel yet biologically and physically meaningful folds compared with state-of-the-art methods, increasing novel fold reconstruction rate by 27.3%.

1 Introduction

Amino acid sequences are usually folded into various geometric structures as proteins to execute biological functions. However, despite the explosive number of experimentally examined protein structures powered by significant advances in structure biology equipment and techniques, the number of unique protein folds - the connectivity and arrangement characteristics of multiple secondary structures - stays relatively stable. Nonetheless, because of the close correlation between protein folds and biological functions[1], it remains a central question of high interest to explore new fold architectures unlike those in natural proteins, which can push the boundaries of protein folds and open up broader protein function space.

Early protein generative models focus on conditional generation that takes fold representation as a conditional vector to be fed into the network (e.g. gcWGAN[2] and cVAE[3]). This kind of methods depends on manual fold design and limits the exploration space. To break this restriction, AutoFoldFinder - an automated optimization toolkit for de novo protein fold design - is presented as an unconditional strategy to explore the protein space, which aims to automatically find novel protein folds and enrich the protein structure library. The main philosophy of AutoFoldFinder is to iteratively adjust the goals of the generation pipeline and enforce it to explore previously unprobed structures.

In order to diversify generated structures as much as possible, we integrate congruence coefficient map alignment[4] into our system as an alternative to Kullback-Leibler divergence (KL-divergence)

*Co-first author. Work done at Beijing Infinite Intelligence Pharma Technology Co., Ltd.
†Co-first author. Correspondence to Y. Xu and L. Lai.

Machine Learning for Structural Biology Workshop, NeurIPS 2021.

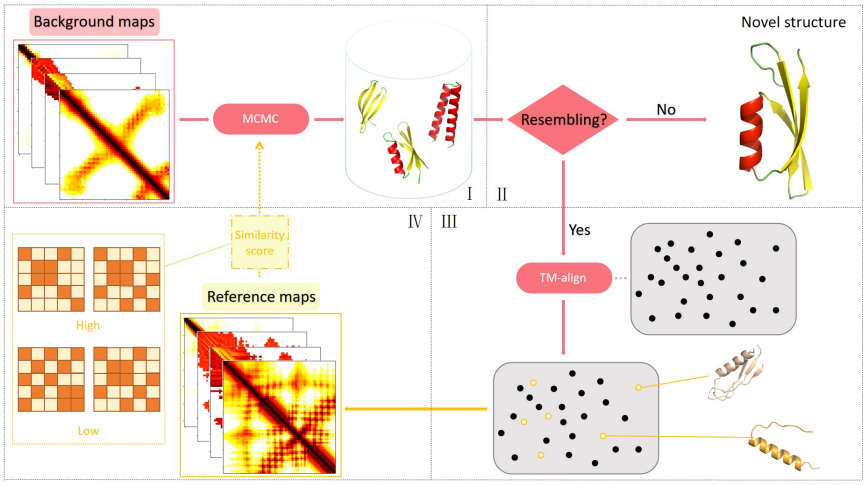

Figure 1: AutoFoldFinder pipeline sketch.

used in deep network hallucination. The motivation is to assess the distance between two maps in a structurally reasonable manner. Instead of globally comparing the whole maps, this metric is able to reflect a divergence on local features of contact maps (e.g. α-helix, β-sheet). Partial alignment information after registration can be detected through congruence coefficient map alignment, which is not perceivable by KL-divergence.

We conclude our contributions as follows. First, we provide the first-of-its-kind novel fold generation method based on sequence optimization with distance maps, which is different from other previously published tools. Second, we employ congruence coefficient map alignment to more accurately assess the distance between contact maps, which significantly enhances the efficiency of generating novel protein structures. Third, our optimization toolkit can deal with novel fold reconstruction problem as well, significantly exceeding current state-of-the-art method.

2 Method

2.1 Optimization strategy

We followed the de novo design pipeline as described by Anishchenko et sl. [5] as a baseline. The core is to maximize the probability $P(sequence, structure)$ ($P(SQ, ST)$ for convenience) through Markov Chain Monte Carlo (MCMC) simulated annealing.

The energy defined in Monte Carlo sampling is as follows:

$$E = D(\mathbf{P}(ST|SQ) \parallel \mathbf{P}(ST) + \sum_{a=1}^{20} D(f_a \parallel f_a^{PDB}) \tag{1}$$

A baseline of hallucinated sequences is generated using the standard pipeline first (Figure 1). All of them are compared with 1233 representative PDB structures using TM-align, a universally used structure comparison tool[6]. This index reveals the structure similarity between generated proteins and natural protein representatives. To prevent sequences from resembling these representatives, distance maps of the most similar resulting sequences are set as reference maps. In subsequent optimization rounds, an additional aim is to maximize the probability of the resulting structure to be different from references. The updated energy equation becomes:

$$E' = E + \sum_{i=1}^{n} D(\mathbf{P}(ST|SQ) \parallel \mathbf{P}(SQ_i) \tag{2}$$

where SQ_i is the i-th reference structure.

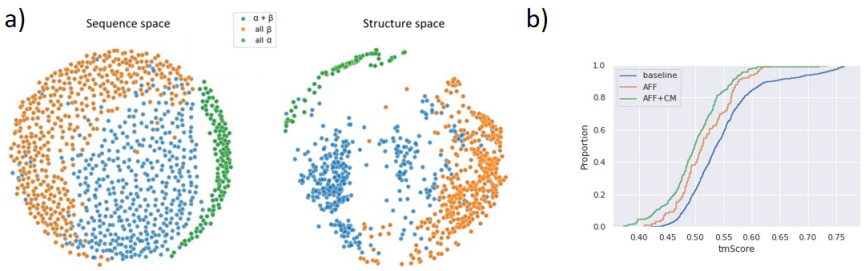

Figure 2: **a)** sequence and structure space generated by the baseline. Sequence similarity and structure similarity (TM-score) were used to measure the distance respectively, **b)** empirical cumulative distribution plot of TM-scores between hallucinated sequences and PDB representatives.

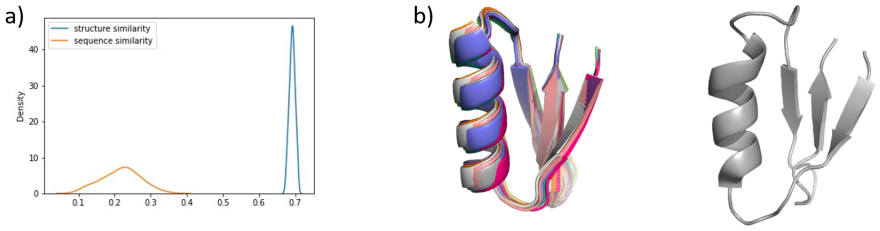

Figure 3: Similarities between AutoFoldFinder-generated proteins and original proteins. **a)** structure and sequence similarities of "EHEE" fold, **b)** structure comparison between generated proteins and original "EHEE" protein (grey).

2.2 CM-align

To overcome the drawback of KL-divergence that it can only measure similarity on a whole-map level, we introduce congruence coefficient map alignment [4] - we refer to as CM-align in this paper - as an alternative to KL-divergence. Congruence coefficient is a measure of matrix similarity.

Definition1. Let $X, Y \in \mathbb{R}^{m \times n}$ be two real matrices. The congruence coefficient between X, Y is defined as:

$$r_c(X, Y) = \frac{tr(XY^T)}{\sqrt{tr(XX^T)tr(YY^T)}} \tag{3}$$

where $r_c = 1$ means the highest similarity and $r_c = -1$ means the least. To facilitate local feature similarity assessment, an alignment is required between distance maps. The alignment is achieved by a contact map alignment method called *map_align*[7]. It uses dynamic programming to optimize map overlap. The main idea is to maximize the summation of similarity between each two sequence separation of aligned contacts. Details about the algorithm can be found in [7].

3 Results

3.1 Novel folds exploration

A baseline of 978 hallucinated sequences of length 50 were generated using the standard pipeline. Dimensional scaling was performed to visualize these sequences in a 2D space. Figure 2a presents the generated protein sequence and structure space including all-α, all-β, and mixed α+β structures, which is conformed with [5]. All structures were compared with representative structures from PDB using TM-align and 155 of them exhibited structural similarity with documented structures (TM-score > 0.6). TM-score below 0.5 is considered to indicate potential novelty of protein folds[2] and 217 structures fall under this criteria, taking up 22.2% of total generations.

Then we generated 366 sequences using our AutoFoldFinder pipeline. Compared with the baseline, AutoFoldFinder generated sequences that are structurally more dissimilar to PDB representatives. 113

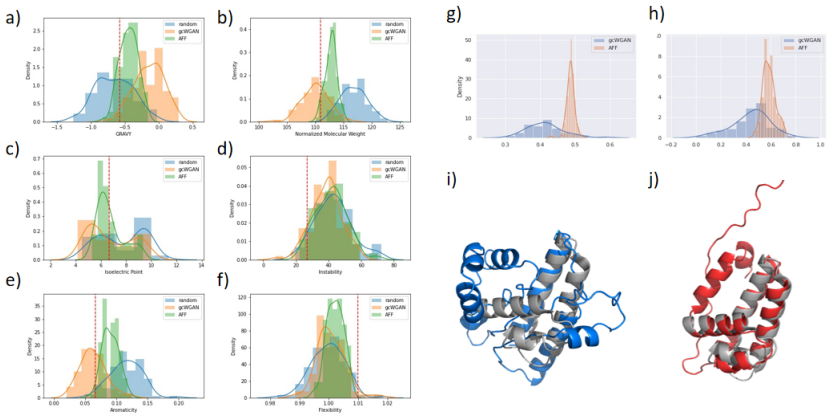

Figure 4: Biophysical properties and identities comparison between random, gcWGAN and AutoFoldFinder. Vertical red lines indicate the value of the original natural protein. **a)** GRAVY, **b)** normalized molecular weight, **c)** isoelectric point, **d)** instability, **e)** aromaticity, **f)** average flexibility, **g)** structure similarity to original protein, **h)** secondary structure ratio, **i)** representative structure generated by gcWGAN, **j)** representative structure generated by AutoFoldFinder. Structures rendered gray indicate original protein.

designed sequences have TM-score under 0.5, increasing novel generation efficiency from 22.2% to 30.9%. To further increase the efficiency of our model, we introduced CM-align into the optimization pipeline. The alignment score, instead of KL-divergence, was used to evaluate the similarity between two distance maps. The involvement of CM-align significantly intensified the inclination of diverging from known structures, resulting nearly all 250 generated structures to have no evident structure analogues in PDB representatives. AutoFoldFinder with CM-align further increased novel generation efficiency to 51.2%.

3.2 Predefined novel folds reconstruction

As a showcase of the generalizability of our optimization scheme, we also investigated the ability of AutoFoldFinder to generate novel sequences within a predetermined fold. We first investigated a basic fold "EHEE" (PDB: 5UP5) examined in [8]. This fold has not been found in nature within its target size. Distance maps and torsion angle maps were calculated and set as the reference to be closed to during the optimization. The optimization is performed by maximizing:

$$E' = E - D(\mathbf{P}(ST|SQ) \parallel \mathbf{P}(SQ_i) \tag{4}$$

In total 100 sequences were generated for "EHEE" fold. All sequences resembled target structure as shown by TM-score of 0.69 (Figure 3). Meanwhile, the sequence similarities only ranged from 0 to 0.4, suggesting generated sequences were hardly sequence homologues to the original sequence. It means we broadly opened up the sequence space of this target fold while sufficiently exploiting its structure.

To further assess the ability of AutoFoldFinder, we compared it with gcWGAN, a state-of-the-art method for novel fold design[2]. We picked a novel fold in this work (PDB: 6H5H) and performed optimization based on distance maps and torsion angle maps of this protein. Then we compared several biophysical properties of 100 produced sequences by our design with gcWGAN designs. Summary of results are provided in Figure 4. AutoFoldFinder produced sequences with GRAVY, isoelectric point, and flexibility closer to natural sequences. We should note here that AutoFoldFinder considers no information of mimicking natural sequences during the optimization process except amino acid composition. Optimization is built solely on structure construction. It is surprising to find AutoFoldFinder performs even better on reproducing some of these biophysical features.

Regarding structure similarity, AutoFoldFinder can generate sequences with better similarity to the original structure compared with gcWGAN in general, increasing average TM-score from 0.41 to 0.49 (Figure 4g). Since average TM-score of Non-homologous protein structures is documented to

be 0.1512[6], our model increases the structure similarity to target fold by (0.49-0.15)/(0.41-0.15)-1=27.3%. We also examined secondary structure ratio of two sets of designs because it measures how orderly a protein structure is[9]. Random loops are not favored in de novo protein design. In general, AutoFoldFinder designs have a higher composition of secondary structure elements than gcWGAN (Figure 4h), which is also reflected in representative structures produced by two methods (Figure 4i and 4j).

References

[1] Jingtong Hou, Gregory E. Sims, Chao Zhang, and Sung-Hou Kim. A global representation of the protein fold space. *Proceedings of the National Academy of Sciences*, 100(5):2386–2390, 2003.

[2] Mostafa Karimi, Shaowen Zhu, Yue Cao, and Yang Shen. De novo protein design for novel folds using guided conditional wasserstein generative adversarial networks. *Journal of Chemical Information and Modeling*, 60(12):5667–5681, Dec 2020.

[3] Joe G. Greener, Lewis Moffat, and David T. Jones. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, 8(1):16189, Nov 2018.

[4] Pietro Di Lena and Pierre Baldi. Fold recognition by scoring protein maps using the congruence coefficient. *Bioinformatics*, 37(4):506–513, 09 2020.

[5] Ivan Anishchenko, Tamuka M. Chidyausiku, Sergey Ovchinnikov, Samuel J. Pellock, and David Baker. De novo protein design by deep network hallucination. *bioRxiv*, 2020.

[6] Jinrui Xu and Yang Zhang. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, 26(7):889–895, 02 2010.

[7] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298, 2017.

[8] Gabriel J. Rocklin, Tamuka M. Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K. Mulligan, Aaron Chevalier, Cheryl H. Arrowsmith, and David Baker. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357(6347):168–175, 2017.

[9] Wouter G. Touw, Coos Baakman, Jon Black, Tim A. H. te Beek, E. Krieger, Robbie P. Joosten, and Gert Vriend. A series of PDB-related databanks for everyday needs. *Nucleic Acids Research*, 43(D1):D364–D368, 10 2014.