Generative De Novo Protein Design with Global Context

Cheng Tan\textsuperscript{1,2}\textsuperscript{*}, Zhangyao Gao\textsuperscript{1,2}\textsuperscript{*}, Jun Xia\textsuperscript{1,2}, Stan Z. Li\textsuperscript{1,2}

\textsuperscript{1} AI Lab, School of Engineering, Westlake University
\textsuperscript{2} Institute of Advanced Technology, Westlake Institute for Advanced Study
\{tancheng.gaozhangyang.xiajun.stan.zq.li\}@westlake.edu.cn

Abstract

The linear sequence of amino acids determines protein structure and function. Protein design, known as the inverse of protein structure prediction, aims to obtain a novel protein sequence that will fold into the defined structure. Recent works on computational protein design have studied designing sequences for the desired backbone structure with local positional information and achieved competitive performance. However, similar local environments in different backbone structures may result in different amino acids, indicating that protein structure’s global context matters. Thus, we propose the \textit{Global-Context Aware} generative de novo protein design method (GCA), consisting of \textit{local} and \textit{global} modules. While \textit{local} modules focus on relationships between neighbor amino acids, \textit{global} modules explicitly capture non-local contexts. Experimental results demonstrate that the proposed GCA method outperforms state-of-the-arts on de novo protein design. Our code and pretrained model will be released.

1 Introduction

Computational protein design, which aims to invent novel protein molecules with desired structures and functions automatically, has a wide range of applications in medical, therapeutics, and pharmacology [Huang \textit{et al.}, 2016; Langan \textit{et al.}, 2019]. Recent years have witnessed remarkable advancements in this field with increased computation power, in which many of them are led by deep learning techniques [Gao \textit{et al.}, 2020; Kryshtafovych \textit{et al.}, 2019; Yang \textit{et al.}, 2019]. While classical protein design approaches depend on composite energy functions of protein physics and sampling algorithms for exploring both sequence and structure spaces, data-driven approaches take advantage of deep neural networks to generate protein sequences with less complex prior knowledge.

Designing a protein sequence for a given structure remains challenging, as the difficulty in mapping the 3D space of structures to the vast-size sequence space. Current data-driven protein design methods [Ingraham \textit{et al.}, 2019; Jing \textit{et al.}, 2021; Strokach \textit{et al.}, 2020; Cao \textit{et al.}, 2021] agree on the assumption based on biology and physics prior knowledge that, for each amino acid, its neighborhoods have the most immediate and vital effects on itself. The majority of such methods represent protein structures as graphs with hand-crafted features and aggregate local messages in hidden layers. The computational protein design process is formulated to learn valuable features from 3D structures with the local message passing mechanism. However, \textit{the similar local environment in different proteins may correspond to different amino acids}. Local neighbors do matter, but it is not enough to obtain high-quality protein sequences.

To fully explore the non-local information, we propose the \textit{Global-Context Aware} generative de novo protein design method (GCA) with both local and global modules. While local modules are built upon graph attention networks that aggregate local messages gained from neighbors with different weights, global modules extend local graph attention to global self-attention neural networks in the form of Transformer [Vaswani \textit{et al.}, 2017] architectures. As shown in Fig. 1, the local module focuses on adjacent structure information though distant nodes can deliver information implicitly; the global module explicitly gathers information from distant nodes in a self-attention mechanism. By composing multiple blocks of local and global modules, GCA can capture high-order dependencies between protein sequences and protein structures at both neighbor-level and overall-level.

Our contributions are three folds:

\begin{itemize}
\item We analyze the current computational protein design methods that ignore the global contexts while similar local patterns correspond to different amino acids.
\item We propose a global-context aware method for generative de novo protein design that satisfies rotation and translation invariance.
\item Extensive experiments on three public datasets show the superior performance of our proposed method.
\end{itemize}

This paper is organized as follows: we introduce the related work and background of de novo protein design in section 2, and shed light on the details of GCA generative de novo protein design method in section 3. Experimental results are reported in section 4, and we conclude in section 5.

\textsuperscript{*}Equal contribution

arXiv:2204.10673v1 [q-bio.BM] 21 Apr 2022
Figure 1: The comparison between the local module and the global module. The information flows from adjacent nodes, and distant nodes are denoted as green arrows and red arrows, respectively. The red dashed arrows in (a) indicate the implicit information flow from distant nodes.

## 2 Related Work and Background

### 2.1 Protein Sequence and Structure

The protein sequence is the sequence of amino acids in a polypeptide chain, also known as the primary structure. The protein structure refers to the tertiary structure, the three-dimensional structure created by a single polypeptide chain. A protein can be translated from the sequence to its native three-dimensional structure by the protein folding process. The close connection between protein structure and sequence leads to two critical problems in structural biology: protein structure prediction and protein design.

Protein structure prediction aims to predict the three-dimensional structure solely based on its amino acid sequence. This problem has been an essential issue throughout the last five decades until the recent data-driven method AlphaFold2 [Jumper et al., 2021] demonstrates their high accuracy competitive with experimental structures in a majority of cases. Building on blocks of various transformer neural networks [Vaswani et al., 2017] with multiple evolutionary, physical, and geometric constraints, AlphaFold2 significantly outperforms other methods and takes a major step forward in the problem of AI-assisted structural biology. At the same time, RoseTTAFold [Baek et al., 2021] proposes a three-track network architecture in which information at the primary structure level, the secondary structure level, and the tertiary structure level are successively transformed and integrated. This method is more lightweight and faster than the above-mentioned AlphaFold2, providing a convenient way for the practical applications of researchers.

Protein design predicts sequence from structure, which can be seen as the reverse process of protein structure prediction. Specifically, the novel structure is designed according to specific functions, and the sequence that can fold into this structure is in need. Functions that naturally occur in sequences of amino acids are essentially accidents of evolution [Huang et al., 2016]. Thus, modifying naturally occurring proteins such as directed evolution [Dougerty and Arnold, 2009] is relatively straightforward in this field—however, our work focus on de novo protein design, i.e., generating protein sequences for protein structures without relying on additional information.

### 2.2 Protein Engineering and Design

Protein engineering and design aim to develop valuable proteins with desired functions for scientific or medical applications. While the specific functions of a protein are determined by its sequence, protein engineering and design exploit the relationship between the structure and the sequence in order to find a decent sequence that can perform specified functions. Though modifying existing proteins is natural, the search space of a single protein is so huge that exhaustive screening on even a tiny fraction of the sequences is expensive and time-consuming.

De novo protein design that generates novel proteins from scratch based on physical and biological knowledge is a promising strategy in protein engineering and design. However, de novo protein design faces several difficulties [Brunette et al., 2015]: (1) the highly diverse structures and functions of proteins result in significant variations, (2) the success rates are remarkably lower than other protein engineering strategies. With the advancement of artificial intelligence technology [Huang et al., 2016], computational protein de novo protein design methods have been developed to address the above issues.

### 2.3 Computational De Novo Protein Design

The complete process of de novo protein design mainly involves structure generation for specific functions, sequence design for given structures, and score function evaluation for further selection. While structure generation can be well solved by assembling local structures using human intuition, and score function evaluation relies on physical principles, we recognize that sequence design that directly maps a three-dimensional structure into a one-dimensional sequence is a more complex problem.

Though the leading protein design framework Rosetta [Alford et al., 2017] that explores the connections of sequence-structure based on parametric energy functions has made some success, there is still room for improvements. To capture the implicit relationships between protein structures and protein sequences, researchers employ deep learning technologies in this field. To the best of our knowledge, SPIN [Li et al., 2014] is the first work that implements neural networks for predicting sequences for given protein structures. This method predicts sequence by integrated neural networks based on fragment-derived sequence profiles and structure-derived energy profiles. SPIN2 [O’Connell et al., 2018] improves its predecessor by applying deeper neural networks with additional structural features and obtains 4% improvements. SPROF [Chen et al., 2019] proposes predicting protein sequences from two-dimensional distance maps of pairwise residues in an image captioning frame. ProDCoNN [Zhang et al., 2020] reformulates the problem by predicting the residue type given the three-dimensional structural environment and builds a nine-layer 3D CNN for the gridded box with the atomic coordinates and types around a residue. Inspired by ProDCoNN, DenseCPD [Qi and Zhang, 2020] builds dense blocks of 3D CNNs to improve the network’s capacity further.

The rise of graph neural networks has also caused a revolution in the field of computational protein design. Struct-
Trans [Ingraham et al., 2019] represents the protein structure as a graph, and employs message passing neural networks or transformer neural networks to predict protein sequence. This work provides a new paradigm for structure-based de novo protein design. GVP [Jing et al., 2021] proposes geometric vector perceptrons extending standard dense layers to operate on collections of Euclidean vectors, thus augmenting graph neural networks with expressive geometric reasoning. ProteinSolver [Strokach et al., 2020] formulates the problem as a constrained satisfaction problem to assign labels to residues such that forces interactions between amino acids are compatible. Fold2Seq [Cao et al., 2021] utilizes a novel fold representation in 3D voxels to provide additional knowledge. Our work focuses on representing protein as a graph and predicting protein sequences based only on structures.

3 Proposed Method

3.1 Preliminaries

Protein primary structure is the linear sequence of amino acids, typically noted as a string of letters. A protein sequence \( S^N = \{a_i\}^N \mid a_i \in \{A, R, N, ..., V\} \) has \( N \) amino acids while each of them is represented by a letter of twenty possible letters such as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, and V. A protein sequence will fold into a protein tertiary structure \( X^N = \{x_i\} \in \mathbb{R}^3: 1 \leq i \leq N, \omega \in \{\text{Co}, \text{C}, \text{N}, \text{O}\} \), where \( N \) is the number of amino acids and \( \omega \) indicates the chain in protein, Co, C, N, O are four types of amino acids.

As shown in Fig. 2, protein design task predicts the protein sequence of a given protein structure, while structure prediction task is the opposite. The 3D visualization is created by Mol* Viewer [Sehnal et al., 2021].

![Figure 2: The comparison of two important tasks in protein modeling: structure prediction and protein design.](image)

3.2 Represent Protein as A Graph

The structure of a protein is represented as a graph \( G = (\mathcal{V}, \mathcal{E}) \) where node feature \( v_i \in \mathcal{V} \) corresponds to an amino acid while edge feature \( E = \{e_{ij}\}_{j \in N_i} \) suggests the rotation-invariant and translation-invariant relationships between each pair of nodes \( v_i \) and \( v_j \). In particular, \( N_i \) denotes the \( K \)-nearest neighbors of node \( i \) calculated by Euclidean distances of the backbone.

![Figure 3: The illustration about dihedral angles of the protein.](image)

![Figure 4: A general view of how the local coordinate system is built.](image)

For node features, we construct three dihedral angles \( \{\phi_i, \psi_i, \omega_i\} \) of the protein backbone from \( C_{i-1}, N_i, C_{i+1}, C_i \) and \( N_{i+1} \) as shown in Fig. 3. Then these dihedral angels are embedded on the 3-torus as:

\[
v_i = \{\sin, \cos\} \times \{\phi_i, \psi_i, \omega_i\}.
\]

For edge features, we focus on describing relative spatial relationships between amino acids that satisfy rotation-invariant and translation-invariant properties. To simplify the computation, we only consider the position \( x_i^\alpha \) of the alpha carbon Co as it’s the central carbon atom in each amino acid. The distance \( ||x_i^\alpha - x_j^\alpha||_2 \), \( \forall i \neq j \) is encoded by Gaussian radial basis functions \( r(\cdot) \).

Then, as shown in Fig. 4, the direction is encoded by \( O_i^T \frac{x_i^\alpha - x_i^{\alpha-1}}{||x_i^\alpha - x_i^{\alpha-1}||} \) while \( O_i = [b_i \quad n_i \quad b_i \times n_i] \) defines a local coordinate system for each amino acid by:

\[
\begin{align*}
    u_i &= \frac{x_i^\alpha - x_i^{\alpha-1}}{||x_i^\alpha - x_i^{\alpha-1}||}, \\
    u_i' &= \frac{x_i^\alpha - x_i^{\alpha-1}}{||x_i^\alpha - x_i^{\alpha-1}||}, \\
    b_i &= \frac{u_i - u_i'}{||u_i - u_i'||}, \\
    n_i &= \frac{u_i \times u_i'}{||u_i \times u_i'||}.
\end{align*}
\]

The orientation is encoded by the common-used quaternion representation of rotation matrix \( q(O_i^T O_j) \). Thus, the edge feature \( e_{ij} \) is the concatenation of the distance, direction and orientation encodings as:

\[
e_{ij} = \left(r(||x_j^\alpha - x_i^\alpha||), O_i^T \frac{x_j^\alpha - x_i^\alpha}{||x_j^\alpha - x_i^\alpha||}, q(O_i^T O_j)\right).
\]
3.3 Network Architecture

Local module

The local module is a graph neural network (GNN) that aggregates both node embeddings and local edge embeddings and updates the node embedding for further sequence generations. Considering a L-layer GNN, the key operations aggregating and updating can be formulated as follows:

\[ h^{(l)}_{N_i} = \text{aggregating}^{(l)}\left(\{h^{(l-1)}_i, h^{(l-1)}_j, e_{ij} : j \in N_i\}\right), \]
\[ h^{(l)}_i = \text{updating}(h^{(l-1)}_i, h^{(l)}_{N_i}), \]

where \( h^{(l)}_i \in \mathbb{R}^D \) denotes the embedding of node \( i \) on the \( l \)-th layer, \( h^{(l)}_{N_i} \in \mathbb{R}^{K \times D} \) denotes the local edge embedding of node \( i \)'s neighbors on the \( l \)-th layer, \( K \) is the number of local neighbors, and \( D \) is the dimensions of the embedding.

In particular, \( h^{(0)}_i \in \mathbb{R}^D \) is the embedding of \( v_i \), and \( e_{ij} \in \mathbb{R}^E \) is the embedding of the edge feature \( e_{ij} \). The local edge information flows into node embeddings at each layer, while distant edge information flows through high-level layers.

In order to capture the relationships in local neighborhoods, we generalize graph attention scheme that take advantage of attention coefficients \( \alpha \in \mathbb{R}^{N \times K} \) as strong relational inductive bias. Specifically, the attention coefficients are calculated as follows:

\[ \alpha_{ij} = \frac{\exp(c_{ij})}{\sum_{k \in N(i)} \exp(c_{ik})}, \forall j \in N_i, \]

where \( c_{ij} \) is expressed as:

\[ c_{ij} = \sigma^T [W h^{(l-1)}_i || W h^{(l-1)}_j || W e_{ij}], \forall j \in N_i, \]

and \( W \in \mathbb{R}^{D \times D}, a \in \mathbb{R}^{3D} \) are learnable parameters, \( \sigma \) is the activation function, \( || \) is the concatenation operation.

Thus, the aggregating operation is adopted as:

\[ h^{(l)}_{N_i} = \sum_{j \in N_i} \alpha_{ij} W_r [h^{(l-1)}_i || h^{(l-1)}_j || h_{e_{ij}}], \]

where \( W_r \in \mathbb{R}^{D \times 3D} \) encodes the relation between \( i \) and \( j \). The updating operation is simply renovating hidden layers by their local neighbors: \( h^{(l)}_i = h^{(l)}_{N_i} \).

Global module

The global module is the fully self-attention network that generalizes Transformer [Vaswani et al., 2017] to protein graph. Specifically, the attention coefficients are calculated as follows:

\[ \alpha_{ij} = \frac{\exp(c_{ij})}{\sum_{k \in V} \exp(c_{ik})}, \]

where \( c_{ij} \) is expressed as:

\[ c_{ij} = \frac{1}{\sqrt{d}} \left( W_q h^{(l-1)}_i \right)^T \left( W_k h^{(l-1)}_j || h^{(l-1)}_j || h_{e_{ij}} \right), \]

where \( W_q \in \mathbb{R}^{D \times D}, W_k \in \mathbb{R}^{D \times 3D} \) are parameter matrices for the query and key, and \( d \) is a scale factor.

Then, the aggregating operation is formulated as:

\[ h^{(l)}_{N_i} = \sum_{j \in V} \alpha_{ij} W_r h^{(l-1)}_j, \]

the updating operation is defined by employing layer normalization (LayerNorm), dropout (DropOut) and fully connected networks (FFN):

\[ h^{(l)}_i = \text{LayerNorm}(h^{(l)}_{N_i} + \text{DropOut}(\text{FFN}(h^{(l)}_{N_i}))). \]

The overall architecture with stacked local modules and global modules is shown in Fig. 5.

3.4 Rotation and Translation Invariance

To ensure the rotation and translation invariance of our proposed model, we should consider it from two aspects: invariance of features and invariance of operations.

Invariance of features

We define the feature transformation for nodes as:

\[ F_v : \mathbb{R}^{4 \times N \times 3} \rightarrow \mathbb{R}^{N \times 6} \]

that transforms three-dimensional coordinates of four chains of amino acids into 3-torus of dihedral angles \( v \) in Eq. 1. The feature transformation for edges is defined as:

\[ F_e : \mathbb{R}^{4 \times N \times 3} \rightarrow \mathbb{R}^{N' \times N' \times 24} \]

that construct edge features \( e \) in Eq. 3, where \( N' = N \) in global modules and \( N' = K \) in \( K \)-nearest graph in local modules.

The rotation invariance of features is defined as:

\[
\begin{align*}
F_{v'}(QX) &= F_v(X) \\
F_{e'}(QX) &= F_e(X)
\end{align*}
\]

for any orthogonal matrix \( Q \in \mathbb{R}^{3 \times 3} \). And the translation invariance of features is defined as:

\[
\begin{align*}
F_{v'}(X + g) &= F_v(X) \\
F_{e'}(X + g) &= F_e(X)
\end{align*}
\]

for any translation vector \( g \in \mathbb{R}^3 \).

As mentioned in Sec. 3.2, we guarantee the invariance of the feature transformations by constructing rotation-invariant and translation-invariant features that are expressive to represent the protein structure and can be uniquely identified.

Invariance of operations

We define the operation of the GCA model for nodes as:

\[ O_v : \mathbb{R}^{N \times 6} \rightarrow \mathbb{R}^{N \times D} \]

that learns from handcrafted node features \( v \) to the \( D \)-dimensional hidden space. The operation of the GCA model for edges is defined as:

\[ O_e : \mathbb{R}^{N \times N' \times 23} \rightarrow \mathbb{R}^{N \times N' \times D} \]

that maps edge features \( e \) to the \( D \)-dimensional hidden space.
Analogous to the invariance of features, we define the rotation invariance of operations as:

\[
\begin{align*}
O_V \circ F_V(AXQ) &= O_V \circ F_V(X) \\
O_E \circ F_E(AXQ) &= O_E \circ F_E(X),
\end{align*}
\]  

(19)

and the translation invariance of operations:

\[
\begin{align*}
O_V \circ F_V(X + g) &= O_V \circ F_V(X) \\
O_E \circ F_E(X + g) &= O_E \circ F_E(X).
\end{align*}
\]  

(20)

The node embedding naturally obeys the rotation and translation invariance as it only relies on a single node without any connection to other nodes. For the edge embeddings, we sort the edges for each node according to the distance of the central carbon atoms \(\|x_i^{C\alpha} - x_j^{C\alpha}\|^2\). The sorting process can be seen as a mapping conditioned on the distance, thus the input \(x\) is changed into \(S(x; \|x_i - x_j\|_{i\neq j}^2)\), and here we use \(x_i\) to represent \(x_i^{C\alpha}\) for notation simplicity. Since

\[
\|x_i^g - [x_j + g]\|^2 = \|x_i - x_j\|^2,
\]  

(21)

\[
\|x_iQ - x_jQ\|^2 = (x_i - x_j)QQ^T(x_i - x_j)^T = (x_i - x_j)I(x_i - x_j)^T = \|x_i - x_j\|^2,
\]  

(22)

the mapping \(S\) satisfies rotation and translation invariance to \(X\) as it is arbitrary mapping of \(\|x_i - x_j\|^2\).

In summary, our proposed method can be formulated as: \(O_V \circ F_V(X)\) for node features, and \(O_E \circ F_E(S(X))\) for edge features.

4 Experiments

4.1 Dataset

We use the CATH 4.2 dataset collected by [Ingraham et al., 2019] to evaluate the ability of our method to generalize across different protein folds. This dataset obtains full chains up to length 500, and structures have been partitioned with 40% non-redundancy by their CATH (Class, Architecture, Topology, Homologous) for all domains. As the evaluation set and the testing set have minor similarities to the training set, we consider this dataset is approximate to the real-world scenarios that require the design of novel structures. With no CAT overlap between sets, there are 18,024 chains in the training set, 608 chains in the validation set, and 1,120 chains in the testing set, respectively. Two subsets of the entire testing set are evaluated simultaneously: a 'Short' subset containing chains up to length 100 and a 'Single chain' subset for comparing with baselines that only use the single chain.

Moreover, we collect a protein dataset of humans from Alphafold [Jumper et al., 2021] Protein Structure Database (Alphafold DB). While the raw dataset includes 12,399 chains, we filter out chains with more than 500 amino acids to keep the experiment consistent to that on CATH 4.2 dataset. The resulting dataset contains 6,901 chains, and there are 5,520 chains in the training set, 100 chains in the validation set, and 1,281 chains in the testing set.

We also consider a smaller dataset TS50, which is the standard benchmark introduced by [Li et al., 2014]. The model is still trained on the CATH 4.2 dataset, and we filter the training and validation sets to ensure there is no overlap with TS50.

4.2 Measurement

Perplexity

Following [Ingraham et al., 2019], we define the perplexity that evaluates the predicted protein sequences from natural language perspective:

\[
\text{PERP}(S^N, X^N) = \exp \left(-\frac{1}{N} \sum_{i=1}^{N} S_i^N \log p(S_i^N | X_i^N) \right),
\]  

(23)

where \((S^N, X^N)\) is the sequence-structure pair of a protein with \(N\) amino acids. \(S_i^N, X_i^N\) denote the \(i\)-th amino acid in sequence and structure respectively. \(p(S_i^N | X_i^N)\) is the output probability from the model.

Recovery

To evaluate the predicting accuracy of the protein sequence at per-residue level, we consider the recovery:

\[
\text{REC(D)} = \frac{1}{|D|} \sum_{(X^N, S^N) \in D} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[S_i^N = \text{arg max}_i p(S_i^N | X_i^N)],
\]  

(24)
GCA outperforms other structure-based models in both PERP and REC metrics. The performance suggests the strong capacity and generalization ability of GCA.

Table 4: Performance of different methods on TS50 dataset assessed by PERP and REC.

| Methods     | PERP  | REC   |
|-------------|-------|-------|
| Rosetta     | 30.0  |       |
| SPIN        | 30.3  |       |
| ProteinSolver | 30.8 |       |
| SPIN2       | 33.6  |       |
| StructTrans | 36.1  |       |
| StructGNN   | 38.0  |       |
| SPROF       | 39.2  |       |
| ProDCoNN    | 40.7  |       |
| GCA         | **43.0** |       |

Table 4: Performance of different methods on TS50 dataset assessed by PERP and REC.

5 Conclusion and Discussion

We introduce the consideration of global information and propose the global-context aware generative de novo protein design method, consisting of local and global modules. While keeping rotation and translation invariances, the local module propagates neighborhood messages across layers, and the global module emphasizes long-term dependencies. Experimental results show that GCA outperforms state-of-the-art methods on three public benchmark datasets. In 'Short' and 'Single chain' sets of CATH, the global-context aware mechanism significantly improves the performance, indicating the potentials to promote structure-based protein design. Moreover, the outstanding performance on AlphaFold-Human and TS50 datasets suggest the strong capacity and generalization ability of GCA. We look forward to applying GCA to practical wet lab experiments.
R. Dustin Schaeffer et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Bae et al., 2021 Minkyung Bae, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, and R. Dustin Schaeffer et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

Brunette et al., 2015 TJ Brunette, Fabio Parmeggiani, Po-Ssu Huang, Gira Bhabha, Damian C Ekiert, Susan E Tsutakawa, Greg L Hura, John A Tainer, and David Baker. Exploring the repeat protein universe through computational protein design. *Nature*, 528(7583):580–584, 2015.

Cao et al., 2019 Yue Cao, Payel Das, Vijil Chenthamarakshan, Pin-Yu Chen, Igor Melnyk, and Yang Shen. Fold2seq: A joint sequence(1d)-fold(3d) embedding-based generative model for protein design. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1261–1271. PMLR, 18–24 Jul 2021.

Chen et al., 2019 Sheng Chen, Zhe Sun, Lilhua Lin, Zifeng Liu, Xun Liu, Yutian Chong, Yutong Lu, Huiyiing Zhao, and Yuedong Yang. To improve protein sequence profile prediction through image captioning on pairwise residue distance map. *Journal of chemical information and modeling*, 60(1):391–399, 2019.

Dougherty and Arnold, 2009 Michael J Dougherty and Frances H Arnold. Directed evolution: new parts and optimized function. *Current opinion in biotechnology*, 20(4):486–491, 2009.

Gao et al., 2020 Wenhao Gao, Sai Pooja Mahajan, Jeremias Sulam, and Jeffrey J. Gray. Deep learning in protein structural modeling and design. *Patterns*, 1(9):100142, 2020.

Huang et al., 2016 Po-Ssu Huang, Scott E Boyken, and David Baker. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, 2016.

Ingraham et al., 2019 John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. In *Advances in Neural Information Processing Systems*, volume 32, pages 15820–15831. Curran Associates, Inc., 2019.

Jing et al., 2021 Bowen Jing, Stephan Eismann, Patricia Suriana, Raphael John Lamarre Townshend, and Ron Dror. Learning from protein structure with geometric vector perceptrons. In *ICLR*, 2021.

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

Kryshtafovych et al., 2019 Andriy Kryshtafovych, Torsten Schwede, Maya Topf, Krzysztof Fidelis, and John Moult. Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020, 2019.

Langan et al., 2019 Robert A Langan, Scott E Boyken, Andrew H Ng, Jennifer A Samson, Galen Dods, Alexandra M Westbrook, Taylor H Nguyen, Marc J Lajoie, Zibo Chen, Stephanie Berger, et al. De novo design of bioactive protein switches. *Nature*, 572(7768):205–210, 2019.

Li et al., 2014 Zhixiu Li, Yuedong Yang, Eshel Faraggi, Jian Zhan, and Yaoqi Zhou. Direct prediction of profiles of sequences compatible with a protein structure by neural networks with fragment-based local and energy-based nonlocal profiles. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2565–2573, 2014.

O’Connell et al., 2018 James O’Connell, Zhixiu Li, Jack Hanson, Rhys Hefferman, James Lyons, Kuldip Paliwal, Abdollah Dehzhangi, Yuedong Yang, and Yaoqi Zhou. Spin2: Predicting sequence profiles from protein structures using deep neural networks. *Proteins: Structure, Function, and Bioinformatics*, 86(6):629–633, 2018.

Qi and Zhang, 2020 Yifei Qi and John ZH Zhang. Densecpd: improving the accuracy of neural-network-based computational protein sequence design with densenet. *Journal of chemical information and modeling*, 60(3):1245–1252, 2020.

Sehnal et al., 2021 David Sehnal, Sebastian Bittrich, Mandar Deshpande, Radka Svobodová, Karel Berka, Václav Bazgier, Sameer Velankar, Stephen K Burley, Jaroslav Koča, and Alexander S Rose. Mol* viewer: modern web app for 3d visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, 2021.

Strokach et al., 2020 Alexey Strokach, David Becerra, Carles Corbi-Verge, Albert Perez-Riba, and Philip M. Kim. Fast and flexible protein design using deep graph neural networks. *Cell Systems*, 11(4):402–411.e4, 2020.

Vaswani et al., 2017 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

Yang et al., 2019 Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

Zhang et al., 2020 Yuan Zhang, Yang Chen, Chenran Wang, Chun-Chao Lo, Xiwen Liu, Wei Wu, and Jinfeng Zhang. Prodconn: Protein design using a convolutional neural network. *Proteins: Structure, Function, and Bioinformatics*, 88(7):819–829, 2020.