# Antibody Sequence Design With Graph-Based Deep Learning Methods

Begench Hangeldiyev[1], Anar Rzayev[1], Azamat Armanuly[1] , Luiz Felipe Vecchietti[2], Meeyoung Cha[2,1], Ho Min Kim[3,1]
[1]KAIST [2]Data Science Group, Institute for Basic Science [3]Protein Communication Group, Institute for Basic Science

## 1. Introduction

Lately, the Covid-19 pandemic highlighted the need for the fast development of therapeutics able to neutralize a target antigen. For that, a protein structure able to interact with the antigen is desired. With this designed structure, a protein sequence should be decoded to maximize the chances of effective expression and binding in wet lab experiments. This process of decoding an amino acid sequence from a protein structure is defined as a protein sequence design problem.

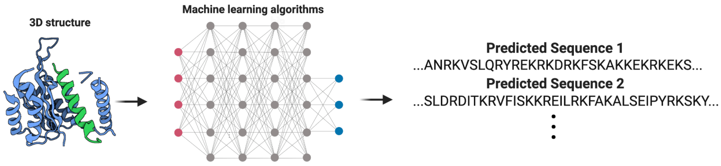
3D structure
Machine learning algorithms
Predicted Sequence 1
...ANRKVSLQRYREKRKDRKFSKAKKEKRKEKS...
Predicted Sequence 2
...SLDRDITKRVFISKKREILRKFAKALSEIPYRKSKY...
⋮

*Figure 1. Protein Sequence Design*

Recent methods that represent a protein as a proximity graph over amino acids achieved breakthrough results in protein sequence design [1-5]. Representing the protein as a graph, the features are associated to nodes, i.e. node features, and to edges, i.e. edge features. Relational reasoning over this graph structure can be performed using deep learning methods such as Graph Neural Networks (GNNs) and Geometric Vector Perceptrons (GVPs) [2], and, from the features learned, the protein sequence is decoded. These methods are highly computationally efficient being able to design sequences in a few seconds.

Sequence design of antibodies is challenging, given that only a small subset of protein datasets contain antibody structures. Also, the loop-like conformations and vast diversity of possible amino acid sequences in the antibody variable regions, e.g. the complementary-determining regions (CDRs), increases the difficulty of this problem. We propose and train an antibody-specific model, named Ing_Ab, based on the algorithm introduced by Ingraham et al [1]. For comparison, we choose the current state-of-the-art method for protein sequence design, ProteinMPNN [3]. We showed that the structures predicted for the sequences designed by the proposed antibody-specific model achieves lower root-mean-square deviation (RMSD) when compared with ProteinMPNN, a model trained using a general protein dataset.

## 2. Methodology

### *2.1 Graph-Based Protein Sequence Design(Ingraham et al, 2019)*

Ingraham et al [1] presented a relational language model for decoding a protein sequence from a graph representation of a target structure. The architecture is divided into an encoder and a decoder. The encoder is responsible to extracting features from the 3D structures using multi-head self-attention on the graph. Only the k-nearest neighbors of a node are considered. Then a decoder predicts the protein sequence autoregressively using causal self-attention, i.e. taking into account nodes that were already decoded. In [1] each node is related to an amino acid in the protein sequence. The node features includes the amino acid identity and dihedral angles of the protein backbone. The edge features are composed by the distance and orientations between two residues.

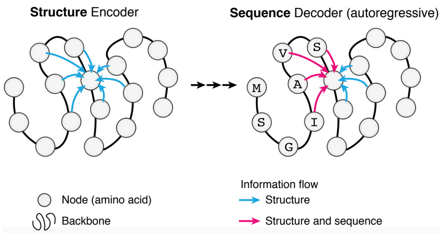

*Figure 2. Encoder and Decoder [1]*

### 2.2 *Ing_Ab*

In this work, the SAbDAb [4] is used. Here, our focus in only in generating the sequence for the variable regions of the antibody. With that in mind, the IMGT server is used to obtain parts of the antibody structure related to its variable region (heavy and light chains).The architecture for training our antibody-specific model is similar to the one proposed in Ingraham et al [1]. For training the dataset was split into train (95%), validation (2.5%), and test sets (2.5%). As Ingraham et al can only decode single chain proteins, we concatenated heavy and light chains as a single chain. After performing a preliminary hyperparameter search, we train our model for 100 epochs with the batch size set to 6000.

## 3. Results

To evaluate ProteinMPNN and Ing_Ab we prepared a reference dataset consisting of 112 antibody structures, that were added to SabDab after the training of our model. For each model, ProteinMPNN and Ing_Ab, we generated 50 amino acid sequences for each antibody structure in the evaluation set. Each of the generated sequences was added as an input to IgFold for antibody structure prediction. In this way, we have predicted 112x50 structures for each model. The main metric used for comparison is the RMSD value between the predicted structure and the native antibody structure. The structures are aligned and the distances calculated using Cealigner.

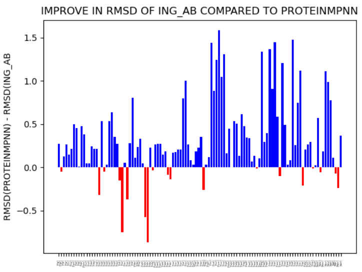

*Figure 3. Comparison between Ing_Ab and ProteinMPNN for each structure in the evaluation set (x-axis).*

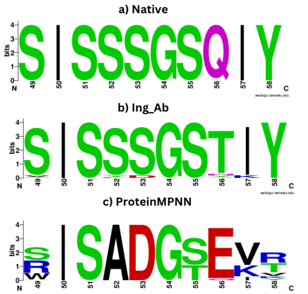

*Figure 4. Amino acid distribution distribution for the CDR-H2 of one of the antibodies in the evaluation set (PDB=7v24)*

| Model | RMSD_IgF (mean) | RMSD_IgF (min) | pRMSD |
|---|---|---|---|
| ProteinMPNN | 1.9156 | 0.7252 | 0.8333 |
| Ing_Ab (ours) | 1.5866 | 0.5624 | 0.6977 |

*Table 1. Comparison of RMSD and confidence metrics with structures predicted by IgFold*

## 4. Discussion

In general, the structures predicted by IgFold for sequences generated by Ing_Ab achieves lower mean RMSD and minimum RMSD when compared to the native structure. Additionally, the pRMSD metric, indicating the confidence of IgFold in the antibody structure prediction, is also lower for Ing_Ab when compared to ProteinMPNN. It is seen that the sequence generated by Ing_Ab is closer to the native sequence, differing with only one residue if sampling is performed in a greedy fashion. We suggest that training the network with antibody-only data may lead the generation of sequences with closer characteristics to native antibodies when compared to ProteinMPNN that is trained with general protein datasets.

## 5. References

[1] Ingraham, John, et al. "Generative models for graph-based protein design." Advances in neural information processing systems 32 (2019).
[2] Jing, Bowen, et al. "Learning from protein structure with geometric vector perceptrons." arXiv preprint arXiv:2009.01411 (2020).
[3] Dauparas, Justas, et al. "Robust deep learning–based protein sequence design using ProteinMPNN." Science (2022): eadd2187.
[4] Dunbar, James, et al. "SAbDab: the structural antibody database." Nucleic acids research 42.D1 (2014): D1140-D1146.