4ZK-07

Enhancing Antibody Design: A Graph Perspective with p-Laplacian Regularization

Tuan Nguyen† Hirotada Honda† Takashi Sano† Shugo Nakamura†

†Toyo University, Graduate School of Information Networking for Innovation and Design

1. Introduction

Antibodies, crucial immune proteins generated in response to infections, consist of two heavy chains and two light chains. The overall structure of antibodies contains six variable regions determining specificity known as Complementarity Determining Regions (CDRs). In this study, assuming a given antigen and antibody framework, we focus on CDR design (sequence-structure co-design). However, with the vast search space for CDRs, experimental testing is impractical with traditional methods. However, with the vast search space for CDRs—where a sequence with L amino acids yields up to 20^L possible sequences— makes experimental testing impractical with traditional methods. Recently, co-designing structures and sequences is achieved through the Diffusion Model and Transformer[1], with the Transformer architecture viewed as a type of Graph Neural Network (GNN)[2]. However, GNNs have limitations in heterophilous graphs, where nodes with different characteristics have numerous neighboring nodes. In protein networks, where amino acids of different types interact, the graph is heterophilous. Addressing this limitation is crucial for advancing antibody design problems.

In the study, we explore the p-Laplacian regularization, addressing heterophily and enhancing the task of antibody design.

2. Related works

Self-Attention for Heterophily. Self-Attention, the core of Transformers, operates by updating token representations through the aggregation of information from other tokens, acts as a low-pass filter[3], emphasizing commonalities and smoothing differences in token features. Thus, this mechanism is not suited for heterophilic networks as it encourages connected nodes to be alike.

p-Laplacian regularization. In machine learning, there is work utilizes Laplacian with fidelity to combat over-smoothing in Transformer models[4] and incorporating p-Laplacian in Transformer models, showcasing advantages in Language Modeling and Image Classification[5]. However, none of previous works have tried the technique to the task of antibody design.

3. Methods

An amino acid in a protein complex is characterized by its type ($s_i \in \{\mathrm{ACDEFGHIKLMNPQRSTVWY}\}$), C_α atom coordinate ($x_i \in \mathbb{R}^3$), and orientation ($O_i \in SO(3)$). Here, $i = 1 \ldots N$, and N represents the total amino acids in the complex across all chains. Our focus is on enhancing the Transformer block to tackle heterophily in protein structures. The baseline model is inspired by Luo et al.'s work on Antibody Sequence-Structure Co-design[1], utilizing a Transformer block within a Diffusion model. Please refer to the original paper[1] for detailed insights into the Diffusion model.

In the diffusion process, Transformer blocks encode the CDR state at a time step t with the context structure $\{s_j^t, x_j^t, O_j^t\}_{j=l+1}^{l+m} \cup \{s_j^t, x_i^t, O_i^t\}_{i=\{1 \ldots N\} \setminus \{l+1 \ldots l+m\}}$. Information is generated for step $t+1$ to denoise the CDR amino acid types, positions, and orientations. Amino acid and pair embeddings use distinct MLPs. The single amino acid MLP generates $\boldsymbol{e_i}$, encoding amino acid types, torsional angles, and 3D coordinates. The pairwise MLP encodes Euclidean distances and dihedral angles between amino acids i and j into feature vectors $\boldsymbol{z_{ij}}$. After that, hidden representations $\boldsymbol{h_i}$ is derived from $\boldsymbol{e_i}$ and $\boldsymbol{z_{ij}}$, representing the amino acid and its environment. Two types of embedding, $\boldsymbol{z_{ij}}$ and $\boldsymbol{h_i}$, are then obtained.

Table 1: Evaluation of the generated antibody CDRs.

CDR	Method	AAR	RMSD
H1	DiffAb	65.75%	1.188Å
	DiffAb+Lap	68.05%	1.109Å
H2	DiffAb	49.31%	1.076Å
	DiffAb+Lap	50.02%	1.058Å
H3	DiffAb	26.78%	3.597Å
	DiffAb+Lap	28.82%	3.436Å
L1	DiffAb	56.67%	1.388Å
	DiffAb+Lap	58.22%	1.242Å
L2	DiffAb	59.32%	1.373Å
	DiffAb+Lap	58.89%	1.131Å
L3	DiffAb	46.47%	1.627Å
	DiffAb+Lap	49.84%	1.575Å

The encoder, with a stack of attention layers, captures amino acid relationships for denoising. Let $\boldsymbol{h_i^l}$ be the hidden output from the last layer. Attention weight computation between residues i and j is $w_{ij} = \mathrm{softmax}_{j=1}^{N}(a_{ij})$. Where $a_{ij} = \langle q(\boldsymbol{h_i}), k(\boldsymbol{h_j})\rangle + f(\boldsymbol{z_{ij}}) + g(x_j)$. Here, q, k, f, g are MLP subnetworks, and $\boldsymbol{x}$ is the location of the heavy atom. Three new embedding vectors are generated following conventional Transformer block equations:

$$\boldsymbol{h_i'} = \sum_{j=1}^{N} w_{ij} * v(\boldsymbol{h_j^l}), \qquad \boldsymbol{z_i'} = \sum_{j=1}^{N} w_{ij} * v(\boldsymbol{z_{ij}}),$$

$$\boldsymbol{x_i'} = \sum_{j=1}^{N} w_{ij} * v(\boldsymbol{x_j})$$

Here, v is another MLP subnetwork. The concatenated vector $[\boldsymbol{h_i'}, \boldsymbol{z_i'}, \boldsymbol{x_i'}]$ is processed through an additional MLP subnetwork with a residual layer, akin to the traditional Transformer block. This step yields the output $\boldsymbol{h_i^{l+1}}$ used for denoising amino acid types through an MLP, generating a 20—dimensional vector of posterior probabilities. The same procedure is applied for denoising C_α and orientations. Based on the analysis of Nguyen et al.[5], we propose a straightforward modification to the aforementioned three equations:

$$\boldsymbol{h_i'} = \sum_{j=1}^{N} \left\|\boldsymbol{h_i^l} - \boldsymbol{h_j^l}\right\|^{p-2} w_{ij} * v(\boldsymbol{h_j^l})$$

$$\boldsymbol{z_i'} = \sum_{j=1}^{N} \left\|\boldsymbol{z_{ii}} - \boldsymbol{z_{ij}}\right\|^{p-2} w_{ij} * v(\boldsymbol{z_{ij}})$$

$$\boldsymbol{x_i'} = \sum_{j=1}^{N} \left\|\boldsymbol{x_i} - \boldsymbol{x_j}\right\|^{p-2} w_{ij} * v(\boldsymbol{x_j})$$

4. Experimental Results

For evaluation, we remove the original CDR from each antibody-antigen complex within the test set and sample both the sequence and structure of the removed region. The CDR length is exactly same to the original. Metrics for assessing designed antibodies include **RMSD (Root-Mean-Square Deviation)**: Measures C_α deviation between generated and original structures with aligned antibody frameworks; **AAR (Amino Acid Recovery Rate)**: Quantifies recovery rate by sequence identity between reference and generated CDR sequences[6].

We compare our model (DiffAb+Lap) to DiffAb from Luo et al.[3]. The optimal parameter, $p = 1.8$, chosen from the set $p = \{1, 1.5, 1.8\}$. Each model produces 100 CDR samples, refined by OpenMM and Rosetta. Table 1 demonstrates that DiffAb+Lap achieves superior CDR sequence recovery (higher AAR) and lower or comparable RMSDs compared to DiffAb.

[1] Luo, S., Su, Y., Peng, X., Wang, S., Peng, J., & Ma, J. (2022). Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems, 35, 9754-9767.*

[2] Shi, H., Gao, J., Xu, H., Liang, X., Li, Z., Kong, L., Lee, S., & Kwok, J. (2022). Revisiting over-smoothing in BERT from the perspective of graph. *arXiv preprint arXiv:2202.08625.*

[3] Wang, P., Zheng, W., Chen, T., & Wang, Z. (2022). Anti-oversmoothing in deep vision transformers via the Fourier domain analysis: From theory to practice. *arXiv preprint arXiv:2203.05962.*

[4] Nguyen, T., Nguyen, T., & Baraniuk, R. (2023). Mitigating Over-smoothing in Transformers via Regularized Nonlocal Functionals. In *Thirty-seventh Conference on Neural Information Processing Systems.*

[5] Nguyen, T., Nguyen, T., Nguyen, V., & Nguyen, T. (2023). p-Laplacian Transformer. *arXiv preprint arXiv:2311.03235.*

[6] Adolf-Bryfogle, J., Kalyuzhniy, O., Kubitz, M., Weitzner, B., Hu, X., Adachi, Y., Schief, W., & Dunbrack Jr, R. (2018). RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. *PLoS computational biology, 14(4), e1006112.*