# PF-ABGEN: A RELIABLE AND EFFICIENT ANTIBODY GENERATOR VIA POISSON FLOW

**Chutian HUANG**[1,2] *, **Zijing LIU**[3], **Shengyuan BAI**[4], **Linwei ZHANG**[5], **Chencheng XU**[5],

**Zhe WANG**[6], **Yang XIANG**[1,2]†, **Yuanpeng XIONG**[5]‡
Department of Mathematics, The Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong SAR[1]
HKUST Shenzhen-Hong Kong Collaborative Innovation Research Institute, Futian, Shenzhen, China[2]
Department of Mathematics, Imperial College London, London, UK[3]
School of Computer Science, Dalian University of Technology, Dalian, China[4]
International Digital Economy Academy, Futian, Shenzhen, China[5]
Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR[6]
{chuangat,zwangec}@connect.ust.hk
zijing.liu@imperial.ac.uk
shengyuan_bai@mail.dlut.edu.cn
maxiang@ust.hk
{zhanglinwei,xuchencheng,xiongyuanpeng}@idea.edu.cn

## ABSTRACT

An antibody is a special type of protein in the immune system to recognize and neutralize pathogenic targets, including bacteria and viruses. Antibody design is therefore valuable for the development of new therapeutics, while experimental-based methods are generally inefficient and expensive. Despite the fruitful progress in protein design with generative neural networks, including diffusion models, they still suffer from high computational costs. In this paper, we propose **P**oisson **F**low based **Anti**B**ody **Gen**erator (PF-ABGen), a novel antibody structure and sequence designer. We adopt the protein structure representation with torsion and bond angles, which allows us to represent the conformations more elegantly, and take advantage of the efficient sampling procedure of the Poisson Flow Generative Model. Our computational experiments demonstrate that PF-ABGen can generate natural and realistic antibodies in an efficient and reliable way. Notably, PF-ABGen can also be applied to antibody design with variable lengths.

## 1 INTRODUCTION

Antibodies are a special type of Y-shaped proteins in the adaptive immune system to mount robust and specific responses to a wide range of pathogens, such as bacteria and viruses. Critical to this specificity of antibodies are the variable fragments (Fv), which are responsible for antigen binding through the Complementarity-Determining Regions (CDRs). Therefore, the rational design of effective therapeutic antibodies often depends upon a proper structure or sequence of the Fv region. However, conventional antibody design methods are generally expensive and time-consuming, as argued by Xu et al. (2019).

Computational methods can partially accelerate the discovery of novel antibodies and play an important role in developing new treatments for human diseases. Attributing to the vast development of deep learning techniques, computational protein design has shown promising perspectives (Graves et al., 2020; Gao et al., 2020). Many works leverage deep generative models to design protein

[*] Work done in International Digital Economy Academy.

[†] Corresponding author.

[‡] Corresponding author.

sequences (Ingraham et al., 2019; Hsu et al., 2022; Madani et al., 2023). These approaches significantly alleviate the computational complexities due to the large search space. However, machine learning approaches for designing the structures of the proteins are still largely left unexplored (Fischman & Ofran, 2018), which restricts their practical applications.

To design protein structures, one significant preliminary problem is the representation of protein structures. Some works interpret proteins as 3D point clouds where each atom stands for one point and apply point cloud generative models to model the Cartesian coordinates of the backbone atoms (Trippe et al., 2022; Luo et al., 2022; Eguchi et al., 2022). However, due to the translation-rotation invariant property of proteins, Wu et al. (2022) argue that these models suffer from the complex equivariant networks. Moreover, the networks' reflection-symmetric property also violates authentic structures. The usage of torsion angles and bond angles is a more natural way to describe proteins, akin to the protein folding process *in vivo*, and has been successfully applied in the computational protein structure prediction (Jumper et al., 2021).

Recently, diffusion models have become one of the most popular generative models and achieved great success in image generation (Ho et al., 2020). Diffusion models have also been employed for protein design (Wu et al., 2022; Luo et al., 2022). Despite the good performance of diffusion models, efficiency remains a challenging issue. Existing applications suffer from the low speed of Stochastic Differential Equation (SDE) samplers, while the high throughput drug discovery demands fast and high-quality designers. Xu et al. (2022) has proposed a kind of new generative model based on "Poisson flow" (PFGM), which enables 10 to 20 times acceleration while achieving similar performances to SDE approaches.

Inspired by the fruitful progress of generative models and protein design, we propose **P**oisson **F**low based **A**nti**B**ody **Gen**enator (PF-ABGen), a novel way for antibody design. To achieve fast protein generation, we adopt the Poisson Flow Generative Model (PFGM), which is more efficient compared to diffusion models based on SDE. In this work, we apply the proposed PF-ABGen model to two protein generation tasks, including structure design and sequence design. Several downstream experiments are conducted in our work to evaluate the quality of sequences and structures of the designed antibodies. The results demonstrate that PF-ABGen could serve as an efficient tool to design biologically plausible antibodies and thus potentially accelerate the development of antibody-based therapies. Moreover, most state-of-the-art works on protein design only focus on fixed-length generation, which is contradictory to the variable-length nature of proteins caused by mutations, including insertions and deletions, which frequently occur in antibodies. Our experiments show that PF-ABGen is capable of generating proteins with variable lengths.

## 2 Related Work

### Generative Models for proteins

A considerable amount of work has explored the possibilities of using generative models in protein design. Anand et al. (2019) uses the Generative Adversarial Network (GAN) to generate pairwise distances between backbone atoms only for fixed-length proteins. Wang et al. (2018) proposes a residue probability and weight network for sequence design given backbone conformations, the application of which is limited. Eguchi et al. (2022) employs a Variational Auto Encoder (VAE) to generate coordinates with a rotationally and translationally invariant loss. However, the coordinate representation demands high computational resources and can hardly generalize to large proteins.

### Diffusion models for Antibody Generation

Many recent works employ diffusion models for protein design. Conditioned on antigen information, Luo et al. (2022) models the complementarity-determining regions (CDR) on the antibodies as a series of amino acid sequences and backbone chain coordinates as well as side-chain orientations. Their model applies a diffusion process with three different distributions for amino acid type, $C_\alpha$ coordinates and $SO(3)$ orientations, respectively.

Wu et al. (2022) presents a diffusion-based model to design protein structures represented by critical angles between adjacent amino acids. They train a diffusion model where noise data is sampled from a wrapped normal distribution. However, they do not design protein sequences in their work.

### Language Model for Protein Generation

Recently, ProGen has been proposed by Madani et al. (2023) as a language model to generate protein sequences across diverse families and functions. However, this language model for proteins only generates sequences or their high-dimensional embeddings, which restricts its application.

## 3 Preliminaries

**Poisson Equation and Poisson Field**

Poisson equation is a second-order elliptic partial differential equation describing many natural phenomena, such as the relationship between electric potential $\phi$ and the source charge $\rho$. Given $\mathbf{x} \in \mathbb{R}^N$, $\rho(\mathbf{x}) \in \mathcal{C}^0 : \mathbb{R}^N \to \mathbb{R}$ a charge function and $\varphi(\mathbf{x}) : \mathbb{R}^N \to \mathbb{R}$ a potential function, the Poisson equation can be written as

$$\nabla^2 \varphi(\mathbf{x}) = -\rho(\mathbf{x}),$$

and the solution is given by a Green's function as

$$\varphi(\mathbf{x}) = \int G(\mathbf{x}, \mathbf{y}) \rho(\mathbf{y}) d\mathbf{y},$$

where $G$ is the Green's function satisfying $\nabla^2 G(\mathbf{x}, \mathbf{y}) = \delta(\mathbf{x} - \mathbf{y})$, $\delta$ is the Dirac-delta function.

The negative gradient field of $\varphi(\mathbf{x})$, referred as Poisson field of the source $\rho$, can be calculated as

$$\mathbf{E}(\mathbf{x}) = -\nabla \varphi(\mathbf{x}) = -\int \nabla_{\mathbf{x}} G(\mathbf{x}, \mathbf{y}) \rho(\mathbf{y}) d\mathbf{y},$$

and this defines a particle dynamics $\frac{d\mathbf{x}}{dt} = \mathbf{E}(\mathbf{x})$ which describes the trajectories of the Ordinary Differential Equation (ODE) as particles moving according to the Poisson field $\mathbf{E}(x)$.

**Poisson Flow Generative Model**

Based on the above knowledge, Xu et al. (2022) proposes Poisson Flow Generative Model (PFGM) to generate 2D images. PFGM views pixels as electrical particles on the plane $z = 0$ and generates data by the flow of the particles through the electric field lines. More specifically, to avoid mode collapse, they first augment the data with an additional dimension, i.e. $\tilde{\mathbf{x}} = (\mathbf{x}, z) \in \mathbb{R}^{N+1}$, and employ the above particle dynamics as the forward ODE, and the corresponding backward ODE $d\tilde{\mathbf{x}} = \mathbf{v}(\tilde{\mathbf{x}})dt$ in the augmented space, where $\mathbf{v}(\tilde{\mathbf{x}})$ is the negative normalized Poisson Field.

In the training stage, PFGM trains a network to learn the negative normalized Poisson field as the evolution of the data. During the generation stage, it first samples data on the plane $z = z_{max}$, which can be mapped to a uniform distribution on the hemisphere, then tracks their motion via the backward ODE back to the $z = 0$ plane, where they will be distributed resembling the raw data. PFGM could generate images faster than diffusion models based on SDE while maintaining the same quality.

## 4 Methods

As shown in Figure 1(a), two noteworthy parts comprising an antibody are the heavy chain and the light chain. Each chain is composed of amino acids and has variable lengths. In this paper, we only work on the design of one chain (heavy/light) and use the heavy chain as an example (it can be easily and reasonably generalized to the light chain). There are twenty canonical amino acids, all of which contain the $N - C_\alpha - C$ backbones and vary in the side chains linked to the $C_\alpha$ atom. In this section, we formulate the representation of antibodies and illustrate the framework of PF-ABGen for the antibody design in both structures and sequences.

### 4.1 Antibody Representation

To design antibodies, both the backbone structures and the side chains need to be depicted. Most of the state-of-the-art works employ 3D Cartesian coordinates to represent conformations of proteins, which is straightforward but demands $SE(3)/SO(3)$-equivariant networks. Moreover, this representation also suffers from heavy computational costs. In this work, we use the angles between the backbone chain to assemble the authentic folding process. This representation naturally

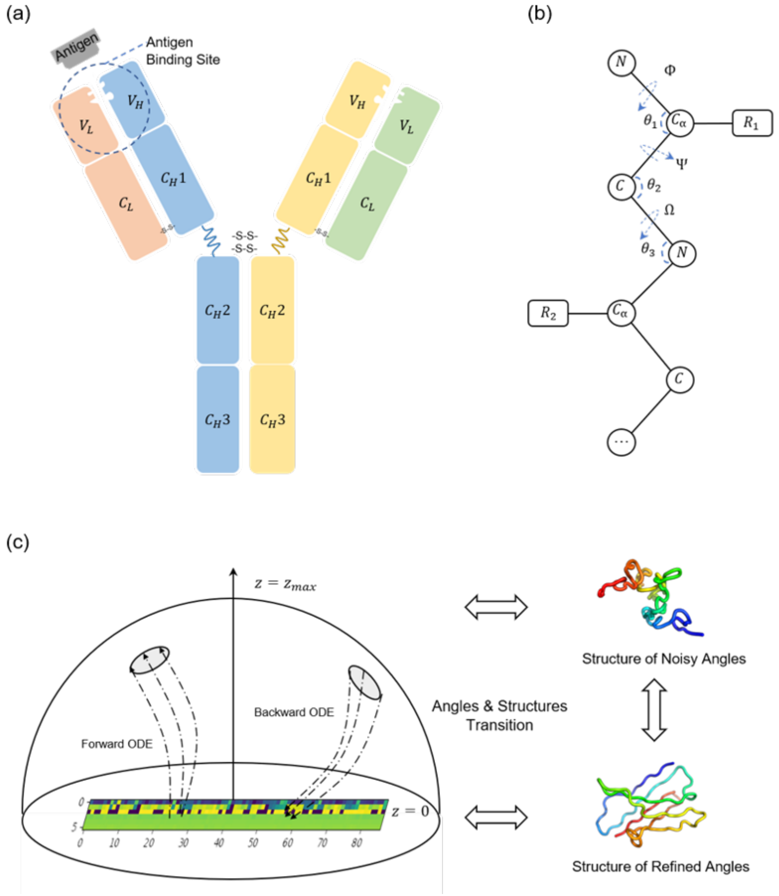


Figure 1: **(a)** An example of antibody structure. The chains colored in blue and yellow are heavy chains, and the chains colored in orange and green are light chains. Each antibody can bind to a specific antigen through the binding site, also denoted as Complementarity-Determining Regions (CDRs). **(b)** Schematic structure of two adjacent amino acids. All the amino acids share the $N - C_\alpha - C$ backbone and vary in the side chains ($R_1$ and $R_2$) attached to the $C_\alpha$ atom. The three dihedral angles $\phi, \psi, \omega$ and three bond angles $\theta_1, \theta_2, \theta_3$ are exemplified. **(c)** Overview of PF-ABGen. In the forward ODE stage, we calculate the empirical Poisson field with the given structure or sequence data and train a network to learn it. Then, we sample data at $z = z_{max}$ from a distribution equivalent to a uniform distribution on a hemisphere in the augmented space and flow it back to $z = 0$ via a backward ODE. The structure at the top corresponds to a set of noisy angles (on plane $z = z_{max}$) and the structure at the bottom corresponds to a set of refined angles (on plane $z = 0$).

incorporates equivariance and exempts us from the complex network design. Since the bond lengths between backbone atoms are consistent, we only use angles to represent the structure and neglect the distances. Three dihedral torsion angles and three bond angles can fully reconstruct the backbone chains. Therefore, the angles can be represented as $\mathbf{x} \in [-\pi, \pi)^{N\times 6}$, where $N$ is the length of the antibody sequence. The details of the angles are shown in Table 1 and visualized in Figure 1(b). These angles can be easily converted to 3D Cartesian coordinates by iteratively adding subsequent atoms to the protein backbone as described in the work by Parsons et al. (2005).

As for the side chains, the type of amino acids determines the category of the residues and thus determines the side chain structure. The one-hot encoding strategy is adopted here to represent twenty types of amino acids, which can be written as $\mathbf{s} \in \{0, 1\}^{N\times 20}$, where $N$ is the length of the antibody sequence.

For the co-design task of structure and sequence, we can simply integrate the previously defined representations for backbone and residue types. In addition, the coordinates of side chains can be solved by any protein packers (Misiura et al., 2022; Liu et al., 2022) with the given conformation of backbones and sequence of the protein.

Table 1: Angle Representation

| **Angle** | **Description** |
|---|---|
| $\psi$ | Dihedral torsion angle about $N_i - C_{\alpha_i} - C_i - N_{i+1}$ |
| $\omega$ | Dihedral torsion angle about $C_{\alpha_i} - C_i - N_{i+1} - C_{\alpha_{i+1}}$ |
| $\phi$ | Dihedral torsion angle about $C_i - N_{i+1} - C_{\alpha_{i+1}} - C_{i+1}$ |
| $\theta_1$ | Bond angle between $N_i - C_{\alpha_i} - C_i$ |
| $\theta_2$ | Bond angle between $C_{\alpha_i} - C_i - N_{i+1}$ |
| $\theta_3$ | Bond angle between $C_i - N_{i+1} - C_{\alpha_{i+1}}$ |

### 4.2 Model

Figure 1(c) illustrates the overview of our PF-ABGen model, and the details are as followings.

We first perturb the original data $\mathbf{x}$ (angles/sequences) into an augmented space, i.e.

$$\tilde{\mathbf{x}} = (\mathbf{y}, z), \tag{1}$$

where

$$\mathbf{y} = \mathbf{x} + \|\epsilon_{\mathbf{x}}\|\,(1+\tau)^m \mathbf{u}, \quad z = |\epsilon_z|\,(1+\tau)^m, \tag{2}$$

and $\epsilon = (\epsilon_{\mathbf{x}}, \epsilon_z) \sim \mathcal{N}\left(0, \sigma^2 I_{N+1,N+1}\right)$, whose norm serves as a small perturbation scalar, $\mathbf{u} \sim \mathcal{U}\left(S_N(1)\right)$ and $m \sim \mathcal{U}[0, M]$. $M$ is a hyper-parameter.

In the training procedure, we use the perturbed data $\tilde{\mathbf{x}}$ to calculate the empirical negative normalized Poisson field

$$\mathbf{v}_{\mathcal{B}}(\tilde{\mathbf{x}}) = -\sqrt{N}\hat{\mathbf{E}}_{\mathcal{B}}(\tilde{\mathbf{x}}) / \left\|\hat{\mathbf{E}}_{\mathcal{B}}(\tilde{\mathbf{x}})\right\|_2, \tag{3}$$

where $\hat{\mathbf{E}}_{\mathcal{B}}(\tilde{\mathbf{x}}) = c(\tilde{\mathbf{x}}) \sum_{i=1}^{|\mathcal{B}|} \frac{\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i}{\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}_i\|^{N+1}}$, $\mathcal{B}$ is a large batch from dataset $\mathcal{D}$, $\tilde{\mathbf{x}}_i = (\mathbf{x}_i, 0)$, $c(\tilde{\mathbf{x}})$ is a multiplier for numerical stability.

A neural network $f_\theta$ is then trained to learn the empirical field by minimizing the mean square loss:

$$\mathcal{L}(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \|f_\theta\left(\tilde{\mathbf{y}}_i\right) - \mathbf{v}_{\mathcal{B}}\left(\tilde{\mathbf{y}}_i\right)\|_2^2 \tag{4}$$

The network $f_\theta$ inherits from the main architectures of deep Noise Conditional Score Network++ (NCSN++) by Song et al. (2020) while making several adaptive adjustments.

During the sampling process, the initial value $\tilde{\mathbf{x}} = (\mathbf{x}, z)$ is sampled on a maximum plane $z = z_{max}$, which can be mapped from a uniform distribution on a hemisphere $S_{N+1}^{+}(z_{max}) = \{\tilde{\mathbf{x}} \in \mathbb{R}^{N+1}, ||\tilde{\mathbf{x}}||_2 = z_{max}, \tilde{\mathbf{x}}_{N+1} > 0\}$. Let $\tilde{\mathbf{x}} = (\mathbf{x}, z)$ flow by the backward ODE $d\tilde{\mathbf{x}} = \mathbf{v}(\tilde{\mathbf{x}})dt$, where

$\mathbf{v}(\tilde{\mathbf{x}})$ is the negative normalized Poisson Field trained before. The ode sampler terminates until $z$ reaches the plane $z = 0$. In practice, the ODE can be written as

$$d(\mathbf{x}, z) = \left(\frac{d\mathbf{x}}{dt}\frac{dt}{dz}dz, dz\right) = \left(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1}, 1\right) dz = \left(\mathbf{v}(\tilde{\mathbf{x}})_{\mathbf{x}}\mathbf{v}(\tilde{\mathbf{x}})_z^{-1} z, z\right) dt', \tag{5}$$

where $z = e^{t'}, dz = zdt'$. Such a change of variable allows for a faster sampling procedure because $z$ decays exponentially as $t$ decays linearly. The right-hand side of Figure 1(c) shows the evolvements of the angles' corresponding structures.

The ODE process of PFGM achieves $10\times \sim 20\times$ faster sampling speeds than the SDE samplers while maintaining a similar quality, as demonstrated by Xu et al. (2022).

In order to facilitate the variable-length generation, we add masks to the loss function as follows:

$$\begin{aligned} \mathcal{L}_i(\theta) &= ||f_\theta(\tilde{\mathbf{y}}_i) - \mathbf{v}_{\mathcal{B}_L}(\tilde{\mathbf{y}}_i)||_2^2 * \mathbf{M}_i, \\ \mathcal{L}(\theta) &= \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\mathcal{L}_i(\theta) \end{aligned} \tag{6}$$

where $\mathbf{M}_i$ is the mask tensor corresponding to data $\mathbf{x}_i$.

# 5 Results

## 5.1 Dataset

To train and evaluate our model, we collect all antibodies, which are deposited before June 11 2022 and with resolutions better than 3Å, from the antibody structure database (SABDAB) (Dunbar et al., 2014). Moreover, we curate two structural datasets including antibodies with a fixed number and variable numbers of amino acids to exhibit the capability of our model of generating proteins with variable lengths. In practice, we further convert the Cartesian coordinates of the three heavy atoms in each amino acid into the six angles described in Section 4.1 and employ a one-hot representation for the type of amino acids.

For the first dataset, we unify the sequence length to 86, which is the minimum length of Fv segments in the heavy chains of all the collected antibodies. For these chains longer than 86, we cut the sequence to multiple sections of 86 amino acids without superposition (except for the last section). If the sequence length is larger than 86 and not divisible by 86, then the last section is chosen as the last 86 amino acids of the antibody (Figure 2). In the end, we obtain 10718 samples in this dataset.

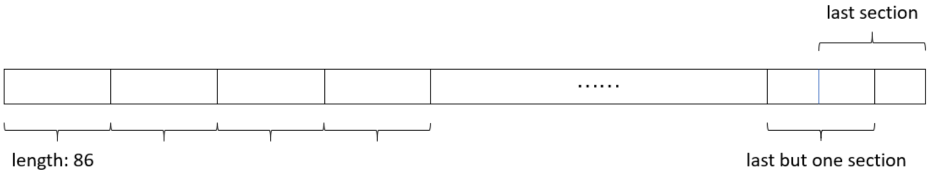


Figure 2: Augmentation strategy of data. Given the sequence with length $l$, we cut it into multiple sections of 86 amino acids without superposition (except for the last section). If $l > 86$ and $86 \nmid l$, then the last section is chosen as the last 86 amino acids of the antibody.

The length and conformational diversity of CDRs mainly comes from six highly variable loops, denoted as H1, H2, H3, L1, L2, and L3. Among them, H3, the most diverse loop, determines the binding specificity against antigens (Figure 3(a)). For the second dataset, we focus on the CDR H3 region. As shown in Figure 3(b), the lengths of H3 loops among different antibodies obey a long-tailed distribution, and we choose 25 as the cutting length. More specifically, we cut the first 25 amino acids for H3 loops longer than 25, and pad with 0 to length 25 for loopers shorter than 25. In this way, we obtain 3973 samples.

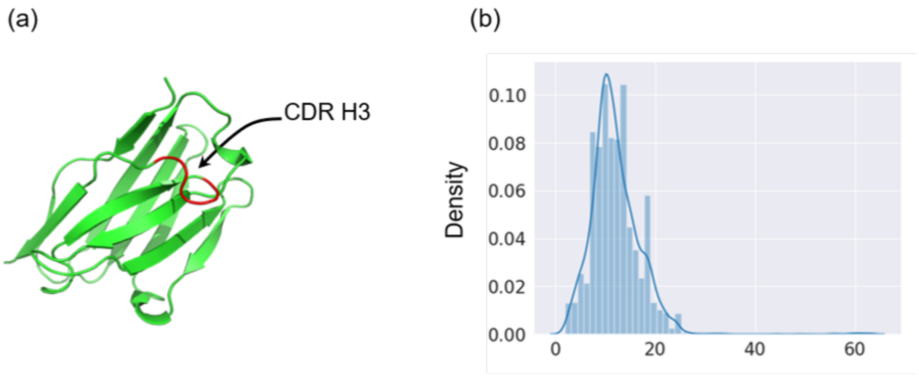
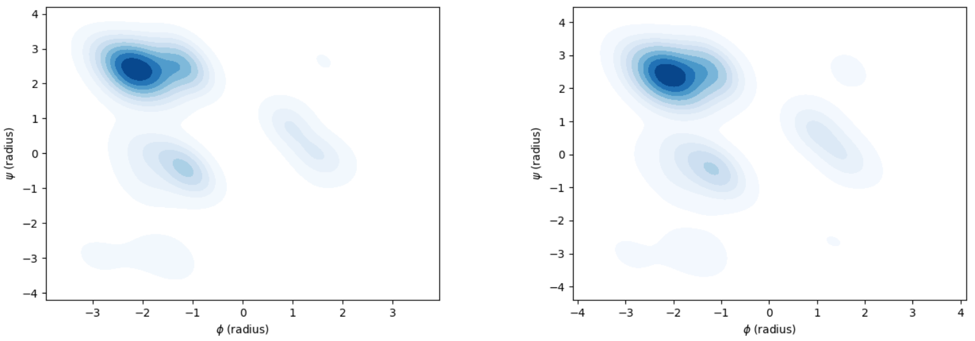


Figure 3: Structure design for H3 loop. **(a)** CDR H3 loop in the antibody 4J1U. **(b)** The histogram of the length of H3 loop. Due to the long-tailed distribution of sequence lengths, we choose 25 as the maximum length for generating H3 structures.

### 5.2 Structure Only Generation

After training our PFGM model, we first assess its capability to recover the secondary structures in the generated backbones. In particular, we compare our randomly generated structures with natural antibody structures by the Ramachandran plot (Ramachandran & Sasisekharan, 1968), which is commonly utilized to visualize the energetically allowed regions for backbone through illustrating the co-occurrence frequency between dihedral angles $\psi$ and $\phi$. Generally, the Ramachandran plot of natural proteins should converge into three clusters, which represent three natural conformations (Figure 4(a)). More specifically, the left bottom cluster, the right cluster, and the left top cluster stand for the left-handed $\alpha$ helix, the right-handed $\alpha$ helix, and the $\beta$ sheet, respectively. As shown in Figure 4(b), our generated angles can recover the three natural secondary conformations. Moreover, the left-handed $\alpha$ helix region is more concentrated compared to the other two regions, which is similar to the plot of natural proteins and can be verified by previous works (Cintas, 2002; Wu et al., 2022). Parts of our generated structures with the angle representation can be found in Appendix Figure A1.



(a) Natural Structure

(b) Generated Structure

Figure 4: Ramachandran plot. **(a)** Natural distribution of torsion angle $\phi$ against $\psi$. **(b)** The generated distribution of $\phi$ and $\psi$. The generated distribution is analogous to the ground truth distribution and can retrieve the three natural secondary conformations.

### 5.3 Sequence Only Generation

Despite the highly variable regions in the variable fragments, there also exist several conserved regions which form the very basic skeleton of antibodies (Morrison et al., 1984; Oda et al., 2003). Obviously, a reliable antibody generator should be capable of recovering most of these constant regions. To validate the generated sequence, we align our generated sequences to the natural sequence database. As a result, the number of matched amino acids varies between $30 \sim 55$ among $86$. Figure 5(a) and 5(b) are two randomly selected examples from the sequence alignment results. The top lines show our generated sequences and the bottom lines show the true sequences.

```
Q-VQLL-ESGGGV-VQ-PGGSLRLSCAV-SGFNIPR----YG-MG-WVRQAPGKGLEWMGT---ITPGDKRK-A---YSP----SF-KGKA--TITADKSS--NTAYMQV--S------SL
  || | |||||  |  ||||||||||  |||        |  |  ||||||||||||      |        |   ||     |  ||    ||     |  | |      |      ||
-EVQ-LVESGGG-LV-KPGGSLRLSCA-ASGF----TFSSY-SM-NWVRQAPGKGLEW---VSSI-------SASSSYS-DYADS-AKG--RFTI-----SRDN-A----KTSLFLQMNSL
```

(a) Score=51

```
EVQLVESGGGLVQA-GGSLK-LSCAG-SGFT-SDY-YGFHWTRQL-SWF-RQAPGKGLEWVANITR-------YA--DSVR-GRFTISRDNAKNTA--YLQMS-SL
|||||||||||||  ||||  ||||  |||| |   |   |     ||  ||||||||||||||         |   |||  ||||||||||||    ||||  ||
EVQLVESGGGLVQ-PGGSL-RLSCA-ASGFTFS--SY---W----MSW-VRQAPGKGLEWVANI--KPDGSEKY-YVDSV-KGRFTISRDNAKN--SVYLQM-NSL
```

(b) Score=66

Figure 5: Two randomly selected examples from the sequence alignment results. The score represents the number of matched amino acids. Top line shows generated sequence while bottom line shows true sequence.

However, solely comparing the sequence similarity does not capture the fact that these sequences contain realistic sequence motifs of antibodies. The Antigen Receptor Numbering and Receptor Classification (ANARCI) tool, which employs a set of hidden Markov models for residue renumbering, is commonly used to annotate the CDR loops of antibodies (Dunbar & Deane, 2016). To validate whether our model deciphers the underlying patterns of the CDR loops, We feed all of our generated sequences to ANACRI. As a result, most of our generated sequences can be identified and annotated as species “human” (see Appendix Table A). This concretely demonstrates that our sequence design strategy does not simply generate “similar” sequences, but also provides valuable insights.

### 5.4 Variable-length H3 Structure Design

We also conducted an experiment on designing protein conformation with variable sizes. More specifically, we train a PFGM model to generate the CDR H3 loop (Figure 3(a)), which is the most flexible region that involves insertion, deletion and variation. As shown in Figure 3(b), the lengths of H3 loops in the training dataset vary from five to sixty, which is accidentally ignored by other studies for protein design (Luo et al., 2022; Wu et al., 2022). By applying masks to the padding length, we successfully generate variable length H3 structures, as shown in Figure 6. The generated angles of the last several amino acids are almost 0 and can be easily filtered. This result shows that our model is also powerful for a more general scene in antibody-based therapies.

## 6 Conclusion

In this work, we propose Poisson Flow based AntiBody Genenator (PF-ABGen), a novel way for antibody design. To evaluate the performance of our model, we conduct several downstream experiments. More specifically, our generated angles can recover natural secondary conformations, and our sequence design strategy provides valuable insights rather than simply generating “similar” sequences. Moreover, our model is capable of generating proteins with variable lengths. These results demonstrate that PF-ABGen could serve as an efficient tool to design biologically plausible antibodies and thus potentially accelerate the development of antibody-based therapies.

Nevertheless, there still exist several promising perspectives in the future. First, although our model exempts us from complex equivariant networks, the angle representation might accumulate errors when recovering angles to coordinates. In addition, the structure-sequence co-design and antigen-

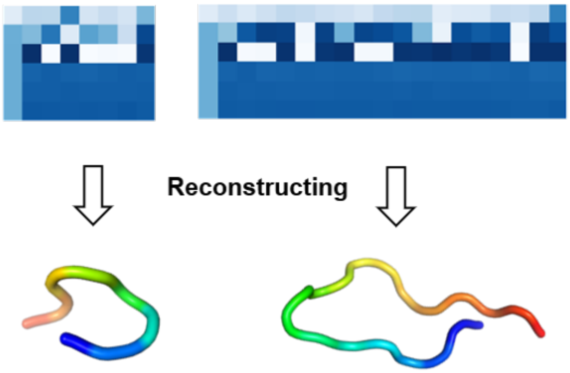


Figure 6: Two examples of our generated H3 loop angles and their corresponding structure. The remaining amino acids, where the bonding and torsion angles tend to zero, have been removed. With a masked loss, our generated results have variable lengths.

conditioned structure/sequence design are still challenging. We believe the arising research interest in drug discovery will shed light on these challenging issues in the future.

### Acknowledgements

The work of YX was supported by the Project of Hetao Shenzhen-HKUST Innovation Cooperation Zone HZQB-KCZYB-2020083.

## References

Namrata Anand, Raphael Eguchi, and Po-Ssu Huang. Fully differentiable full-atom protein backbone generation.(2019). In *ICLR 2019 Workshop DeepGenStruct*, 2019.

Pedro Cintas. Chirality of living systems: a helping hand from crystals and oligopeptides. *Angewandte Chemie International Edition*, 41(7):1139–1145, 2002.

James Dunbar and Charlotte M Deane. Anarci: antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

James Dunbar, Konrad Krawczyk, Jinwoo Leem, Terry Baker, Angelika Fuchs, Guy Georges, Jiye Shi, and Charlotte M Deane. Sabdab: the structural antibody database. *Nucleic acids research*, 42(D1):D1140–D1146, 2014.

Raphael R Eguchi, Christian A Choe, and Po-Ssu Huang. Ig-vae: Generative modeling of protein structure by direct 3d coordinate generation. *PLoS computational biology*, 18(6):e1010271, 2022.

Sharon Fischman and Yanay Ofran. Computational design of antibodies. *Current opinion in structural biology*, 51:156–162, 2018.

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

Jordan Graves, Jacob Byerly, Eduardo Priego, Naren Makkapati, S Vince Parish, Brenda Medellin, and Monica Berrondo. A review of deep learning methods for antibodies. *Antibodies*, 9(2):12, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022.

John Ingraham, Vikas Garg, Regina Barzilay, and Tommi Jaakkola. Generative models for graph-based protein design. *Advances in neural information processing systems*, 32, 2019.

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.

Jiale Liu, Changsheng Zhang, and Luhua Lai. Geopacker: A novel deep learning framework for protein side-chain modeling. *Protein Science*, 31(12):e4484, 2022.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models. *bioRxiv*, pp. 2022–07, 2022.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, Jan 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01618-2. URL `https://doi.org/10.1038/s41587-022-01618-2`.

Mikita Misiura, Raghav Shroff, Ross Thyer, and Anatoly B Kolomeisky. Dlpacker: deep learning for prediction of amino acid side chain conformations in proteins. *Proteins: Structure, Function, and Bioinformatics*, 90(6):1278–1290, 2022.

Sherie L Morrison, M Jacqueline Johnson, Leonard A Herzenberg, and Vernon T Oi. Chimeric human antibody molecules: mouse antigen-binding domains with human constant region domains. *Proceedings of the National Academy of Sciences*, 81(21):6851–6855, 1984.

Masayuki Oda, Haruo Kozono, Hisayuki Morii, and Takachika Azuma. Evidence of allosteric conformational changes in the antibody constant region upon antigen binding. *International immunology*, 15(3):417–426, 2003.

Jerod Parsons, J Bradley Holmes, J Maurice Rojas, Jerry Tsai, and Charlie EM Strauss. Practical conversion from torsion space to cartesian space for in silico protein synthesis. *Journal of computational chemistry*, 26(10):1063–1068, 2005.

GN t Ramachandran and V Sasisekharan. Conformation of polypeptides and proteins. *Advances in protein chemistry*, 23:283–437, 1968.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

Jingxue Wang, Huali Cao, John ZH Zhang, and Yifei Qi. Computational protein design with deep learning neural networks. *Scientific reports*, 8(1):1–9, 2018.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022.

Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. *arXiv preprint arXiv:2209.11178*, 2022.

Yingda Xu, Dongdong Wang, Bruce Mason, Tony Rossomando, Ning Li, Dingjiang Liu, Jason K Cheung, Wei Xu, Smita Raghava, Amit Katiyar, et al. Structure, heterogeneity and developability assessment of therapeutic antibodies. In *MAbs*, volume 11, pp. 239–264. Taylor & Francis, 2019.

# A Appendix

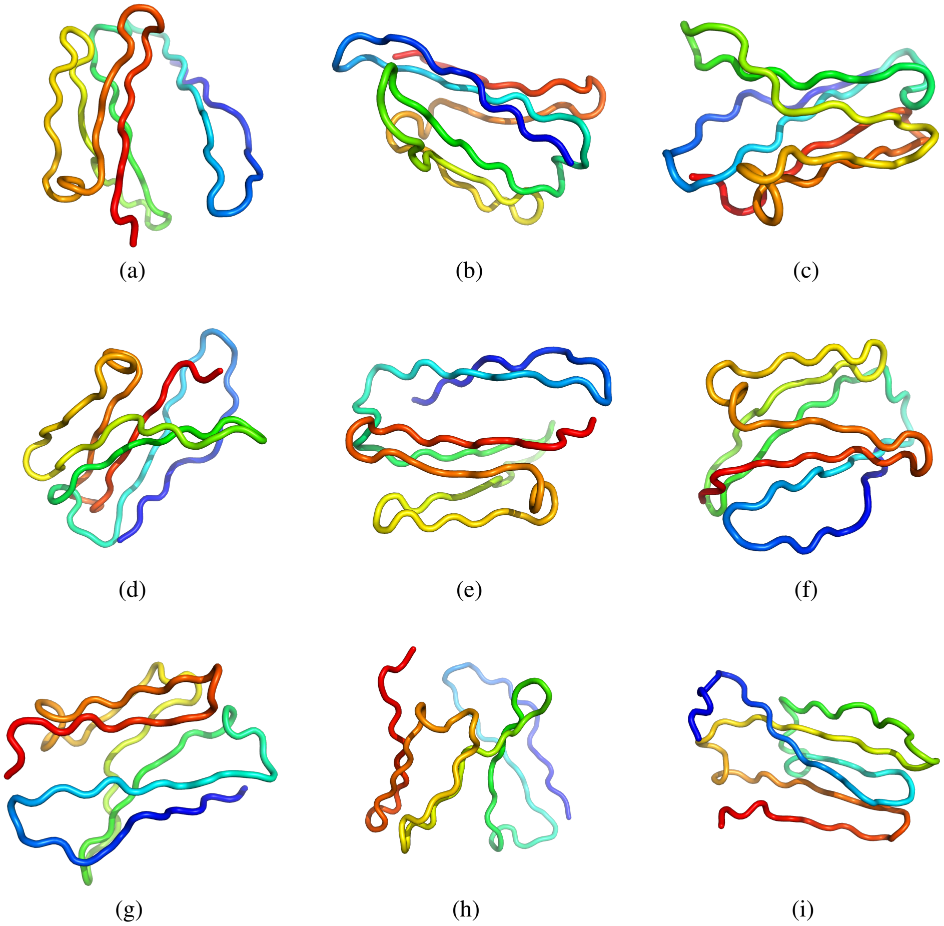

(a) (b) (c)
(d) (e) (f)
(g) (h) (i)

Figure A1: Generated Structure – 86 length

| Species | Chain Type | e-value | Score | Seqstart Index | Seqend Index |
|---|---|---|---|---|---|
| human | H | 1.1e-35 | 112.8 | 0 | 85 |
| human | H | 5.1e-31 | 97.8 | 0 | 85 |
| human | H | 1.3e-28 | 90.0 | 0 | 85 |
| human | H | 2e-35 | 112.1 | 0 | 85 |
| human | H | 2e-39 | 125.0 | 0 | 85 |
| human | H | 2.8e-40 | 127.7 | 0 | 85 |
| human | H | 1.1e-38 | 122.5 | 0 | 85 |
| human | H | 1e-37 | 119.4 | 0 | 85 |
| human | H | 6.2e-35 | 110.4 | 0 | 85 |
| human | H | 8.8e-32 | 100.2 | 0 | 85 |

Table A1: The results of ANARCI Analysis. The e-value and score in the table stand for the e-value and bit-score of the alignment to the most significant HMM, respectively. A smaller E-value means a better match. A higher bit-score represents better sequence similarity. More details can be found in Dunbar & Deane (2016).