MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
KYIV NATIONAL UNIVERSITY OF TECHNOLOGIES AND DESIGN
Faculty of Chemical and Biopharmaceutical Technologies
Department of Biotechnology, Leather and Fur

QUALIFICATION THESIS

on the topic **Performance evaluation of artificial intelligence methods in nanoantibody design**

First (Bachelor's) level of higher education
Specialty 162 "Biotechnology and Bioengineering"
Educational and professional program "Biotechnology"

Completed: student of group BEBT-20
Chunzi ZANG

Scientific supervisor
Tetiana SHCHERBATIUK,
Dr. Sc., Prof.

Reviewer
Iryna VOLOSHYNA,
Ph.D., As. prof.

Kyiv 2024

KYIV NATIONAL UNIVERSITY OF TECHNOLOGIES AND DESIGN

Faculty: Chemical and Biopharmaceutical Technologies
Department: Biotechnology, Leather and Fur
First (Bachelor's) level of higher education
Specialty: 162 Biotechnology and Bioengineering
Educational and professional program Biotechnology

APPROVE

Head of Department of Biotechnology,
Leather and Fur, Professor,
Doctor of Technical Science
Olena MOKROUSOVA

_____
«___»_____2024

ASSIGNMENTS

FOR THE QUALIFICATION THESIS

Chunzi ZANG

1. Thesis topic **Performance evaluation of artificial intelligence methods in nanoantibody design**

scientific supervisor Tetiana Shcherbatiuk, Dr. Sc., Prof.

approved by the order of KNUTD "__"_____2024, №____

2. Initial data for work: assignments for qualification thesis, scientific literature on the topic of qualification thesis, materials of Pre-graduation practice

3. Content of the thesis (list of questions to be developed): literature review; object, purpose, and methods of the study; experimental part; conclusions

4. Date of issuance of the assignments_____

EXECUTION SCHEDULE

№	The name of the stages of the qualification thesis	Terms of performance of stage	Note on performance
1	Introduction	From 01 April 2024 to 11 April 2024	
2	Chapter 1. Literature review	From 06 April 2024 to 20 April 2024	
3	Chapter 2. Object, purpose, and methods of the study	From 21 April 2024 to 30 April 2024	
4	Chapter 3. Experimental part	From 01 May 2024 to 10 May 2024	
5	Conclusions	From 07 May 2024 to 12 May 2024	
6	Draw up a bachelor's thesis (final version)	From 20 May 2024 to 25 May 2024	
7	Submission of qualification work to the supervisor for feedback	From 25 May 2024 to 27 May 2024	
8	Submission of bachelor's thesis to the department for review (14 days before the defense)	27 May 2024	
9	Checking the bachelor's thesis for signs of plagiarism (10 days before the defense)	31 May 2024	
10	Submission of bachelor's thesis for approval by the head of the department (from 7 days before the defense)	03 June 2024	

I am familiar with the task:

Student _____ Chunzi ZANG

Scientific supervisor _____Tetiana SHCHERBATIUK

SUMMARY

Chunzi Zang. Performance evaluation of artificial intelligence methods in nanoantibody design. – Manuscript.

Qualification thesis on the specialty 162 «Biotechnology and Bioengineering». – Kyiv National University of Technologies and Design, Kyiv, 2024.

Nanobody is an antibody found in naturally missing light chains in camel families such as alpaca and monama, as well as cartilage fish such as sharks and rays, including two constant regions, a hinge region and a heavy chain variable region. With its small size, high stability, strong affinity, low cytotoxicity, strong penetration, and simple humanization, it is widely used in disease diagnosis, treatment and novel nanodrug design. Traditional nanobody acquisition methods for animal immunization or library screening. The traditional preparation methods have the disadvantages such as cumbersome process, poor specificity, difficult protein expression, and inability to target specific epitopes. Therefore, innovative strategies are needed to transform the sequence and structure of nanobodies and design new antibodies that are not available in nature. Nowadays, through the artificial intelligence method to deep learn the complete information of the target antigen, the variable region of the antibody and the complex internal and external physical effects of the antibody, which can effectively generate the 1D sequence and 3D structure of the antibody CDR region. Nanobodies designed by artificial intelligence have strong antigen targeting characteristics, expression ability and generalization ability, and have wide application prospects. Taking diffab and AlphaPanda as an example, this paper introduces the design method of AI antibody in detail from the aspects of model building and antibody design process, and evaluates the design performance of RMSD, Seqid and ddG. The results show that the RMSD values

of CDR1 and CDR3 are greater than 2Å; CDR2 is less than 1.5Å, reaching atomic accuracy. Only the diffab designed CDR2 sequence showed good agreement, numerically over 30%.0.0067% of the CDR designed by diffab was less energetic than the natural antibody and 0.0233% of the CDR designed by AlphaPanda was lower than the natural antibody. The overall performance of the designed CDR is better in the above indicators. Based on the above data, this paper proposes improvement measures for the AI antibody design program, and puts forward new ideas and prospects for the future field of AI antibody design.

Key words: *Nanobodies; artificial intelligence; antibody design; performance evaluation*

TABLE OF CONTENTS

INTRODUCTION

As a drug, antibody has the advantages of high specificity, uniform properties, and targeted preparation for specific targets. Its application prospect has attracted much attention in the treatment of various diseases, especially in the field of tumor therapy. At present, the antibodies used in the development of targeted drugs are mainly IgG1 and IgG4. In drug development, antibody engineering modification is often needed, that is, from extending half-life, improving affinity, enhancing effector function, and reducing immunogenicity, so as to make them more suitable for fighting targets. With the deepening of antibody-drug coupling (ADC), antibody fragments (such as nanobodies) and bispecific antibodies (BsAb), there is a higher demand for antibodies in terms of affinity binding efficiency.

Nanobodies first found in camels, shark serum, natural loss of light chain, small molecular weight, containing extended CDR 3 ring, and special FR2 amino acid residues and convex antigen binding site allows them to bind to the usually blocked concave antigen region, which targeting the traditional monoclonal antibody to the target, make drug action more specific and efficient. The CDR3 ring of nanobodies is long and more variable, able to bind diverse antigens. Structural variation in CDR 1 and the expansion of CDR3 compensate for the loss of the light chain. FR2 of nanobodies is usually composed of hydrophilic amino acid residues, thus having good water solubility and high stability against high temperature, protease and pH changes. Moreover, nanobodies have better penetration and the ability to cross the BBB and even cross the damaged ones under neuropathological conditions. Considering these characteristics, nanobodies can be administered by alternative routes such as oral or intraperitoneal injection. The low immunogenicity of nanobodies makes them an ideal candidate for drug development [1]. Furthermore, diverse synthesis by humanization

and CDR randomization can further mitigate side effects. The structure of nanobodies is relatively simple and does not require common posttranslational modifications or complex eukaryotic expression systems and purification steps. Due to their small molecule and monomeric structures, nanobodies are well suited for multimerization, thus achieving multivalency, multiepitope, and multispecificity, which can increase affinity and bind multiple antigens.

Nanobodies can be produced by conventional prokaryotic expression systems [2] (e. g.*E.coli*) and eukaryotic expression systems (e.g. Saccharomyces cerevisiae) or screened by natural libraries, immune libraries, synthetic libraries, etc. These advantages make nanobodies relatively inexpensive to produce, while having a broader range of antigen recognition than conventional monoclonal antibodies.

Immune libraries were inoculated with target antigen such as alpaca or single, then purified lymphocytes from blood to extract for mRNA and then converted into cDNA. After amplification by PCR, the target sequences were screened by agarose gel electrophoresis. Finally, the VHH sequence was amplified using primers specific for the live restriction enzyme site to insert the obtained amplicon into the appropriate expression vector (usually E. coli or yeast). However, antibodies extracted in animals may cause immune responses in humans. Difficulties to obtain highly specific antibodies against rare or highly conserved antigens.

With the continuous progress of the experimental technology, the development method is also constantly improving. Single-cell sequencing technology allows researchers to extract and analyze antibody genes from single B cells to directly access the precise sequence of antibodies. The application of this technology greatly reduces the time from antibody discovery to production and improves the development efficiency of antibody therapy. Using X-ray crystallography and cryo-electron microscopy techniques, researchers are now able to resolve the complex structure of antibodies and antigens at atomic-level resolution. This detailed structural information

enables structure-based antibody design, enhancing its binding affinity and specificity to a specific antigen by precisely modifying the structure of the antibody. Computer-aided design (CAD) technology, especially in protein modeling and simulation, provides a powerful tool for antibody design. By mimicking antibody-antigen interaction, researchers are able to predict antibody binding properties before laboratory manipulation, thus guiding antibody modification and optimization in the laboratory. Although these are high-throughput experimental methods, their flux is still insufficient relative to the sequence space of the antibodies. A combination of rational design to further reduce the screening space is needed to improve the success rate of antibody drug design [3,4].

With the continuous development of bioinformatics, computational biology and other fields, the field of antibody design has undergone revolutionary changes in recent years. Artificial intelligence technology is gradually applied to antibody design [5-7], which greatly improves the efficiency and success rate of antibody design and makes it possible to design antibodies ab initio, greatly improves the speed and accuracy of antibody discovery, and opens up a new way for the treatment of various diseases.

In recent years, the continuously developing AI technology has opened up new directions in the field of antibody design. Machine learning as a branch of artificial intelligence, is not completely rely on programming instructions, but in the data learning to find out the law, and according to the new data and feedback quickly adjustment and optimization, effectively reduce the calculation, and have good generalization ability, to some extent, make up for the manual energy function cannot capture biological macromolecular covalent interaction and the defects of dynamic change. Commonused machine learning algorithms include support vector machine, random forest, neural network and so on.

Deep learning as a subset of machine learning, using the multilayer structure of the neural network model to simulate the human brain processing information of data training, deep learning algorithm has strong feature extraction ability, using the nonlinear activation function processing protein non-covalent interactions and dynamic changes and more complex biological data. A series of artificial intelligence-based bioinformatics platforms have been established, Such as AbDiver[8] can compare the designed antibody sequences to the natural antibodies in the antibody database, So as to guide the development of antibody drugs; TAP is a bioinformatics platform to assess the physicochemical properties of antibody drugs; Camsol[9] vs. The SOLart[10] can assess the solubility of the antibody; SOLart And AGGRESCAN 3D[11] can be used to assess the aggregation of antibodies; IEDB-AR [12] is a platform to predict and analyze antibody drug immune epitopes; Hu-mAb [13] can distinguish between human and non-human Fv fragments, But the above platforms can only be optimized for a small part of the antibody design process, The practical application effect is poor. In recent years, deep learning models such as diffab, PROSEED and RFdiffusion have targeted antibodies for specific antigens based on the secondary structure of antigen antibodies, which has greatly promoted the progress of computational antibodies.

At present, the methods of artificial intelligence in antibody design mainly include: (1) antibody design methods based on machine learning: the application of machine learning algorithm in antibody design mainly includes antibody structure prediction, recognition of antigen binding site and antibody affinity optimization, etc. Commonused machine learning algorithms include support vector machine, random forest, neural network and so on (2) Antibody design method based on deep learning: Deep learning algorithm has strong feature extraction ability and can process complex biological data. In antibody design, deep learning algorithms are mainly used for antibody-based sequence-structure prediction, antigen-binding affinity prediction, and

antibody humanization. (3) Antibody design method based on the evolutionary algorithm: The evolutionary algorithm simulates the natural selection and genetic mechanism, and searches for the optimal solution in the antibody library through iterative optimization. The method has achieved good results in antibody affinity maturity and antibody humanization.

The relevance of the topic is Nanoantibody design by artificial intelligence.

The purpose of the study is to investigate the performance of AI nanobodies.

The objectives of the study is to investigate the performance of AI nanobodies.

The object of the study is to investigate performance evaluation of the AI-made nanobodies.

The subject of the study is to investigate performance evaluation of the AI-made nanobodies.

Research methods by experimental demonstration.

The scientific novelty is data analysising after designing nanobodies with AI program.

The practical significance of the results obtained is the stability of AI nanobodies is high, but the practical application ability needs to be improved.

CHAPTER 1
LITERATURE REVIEW

1.1 Artificial intelligence antibody design concept and development direction

AI antibody design is the use of AI technology to assist and accelerate the discovery and design process of antibody drugs. Artificial intelligence antibody design mainly includes the following aspects: (1) data collection and analysis: collection of known antibody sequence and structural data, as well as antibody information for binding to specific antigens.(2) Antibody sequence design: the antibody sequence is optimized by the algorithm to enhance its binding ability with the target antigen, while maintaining good biological activity and stability.(3) Structure prediction and optimization: to predict the three-dimensional structure of the antibody, and to optimize the structure through computational methods to improve the affinity and efficacy of the antibody.(4) Wet laboratory validation: the antibody designed by AI is synthesized and tested in the laboratory to verify its binding activity and biological function.

1.2 Antibody design model based on a deep learning model

1.2.1diffab model

diffab in antibody design, it mainly includes sequence-structure co-design, antibody sequence design based on antibody skeleton and antibody optimization tasks [14]. It learns the complex relationship between antibody sequence and structure by training on datasets acquired in SAbDab[15] and other databases. During training, the model will ignore antibodies with a below 4Å resolution and antibodies against non-protein antigens to improve the accuracy and generalization ability of the model. An important feature of the diffab model is that it processes the dataset using a clustering

method to identify antibody groups with similar characteristics and to manually select partial clusters as the test set. This strategy is useful for evaluating the performance of the model on different types of antibody design tasks. When evaluating the performance of diffab models, indicators such as amino acid recovery rate (AAR), C α root mean square deviation (RMSD), and binding energy (G) are usually used. These indicators enable a comprehensive assessment of the performance of the designed antibodies in terms of structure, stability, and affinity. Compared to the RosettaAntibodyDesign (RAbD) model [16], the diffab model performs better in AAR and is comparable to RAbD in RMSD and G, showing its competitiveness and potential in the field of antibody design. The diffab model similarly enables antibody design in the absence of a known antibody framework for binding to the antigen. An antibody-antigen complex was generated by removing an existing CDR-H3 and docking the antibody template to the target antigen using HDOCK [17].

1.2.2PROSEED model

PROSEED is a model [18] of co-design for sequence and structure based on context features. This model iteratively transforms the protein sequence and structure from random initialization to the expected state. This model includes a triangle-aware encoder that explains geometric constraints and interactions from context features and a rotary transition isvariant decoder that interdependently translated protein sequences and structures, iteratively converting proteins into desired states in an end-to-end and equivariant manner. By predicting the structural update of the local frame based on the invariant representation, and then using the changing base operation, the equivariant properties of the protein structure during the whole process are guaranteed. It is worth mentioning that all protein amino acids are updated once in each translation step, which greatly accelerates the inference process. PROSEED Can update the sequence and structure of all residues at once, thus enabling a more efficient inference process.

Unlike the previous approach of structuring and then generating sequence and rotating isoers, we allow the model to cross-condition the sequence and structure, with the maximum information flow between context features, sequence and structure, thus ensuring the fidelity of the generated proteins. We performed extensive experiments on the Structural Antibody Database (SAbDab) and two protein design benchmark datasets curated from CATH [19], and compared PROSEED with previous state-of-the-art methods in three modules: antigen-specific antibody CDR design, context-conditional protein design, and fixed backbone protein design.The data show that this model can generate high-fidelity proteins in sequence and structure, while being several orders of magnitude faster than sampling-based methods.

1.2.3RFdiffusion Model

RFdiffusion[20] Is a model designed by the David Baker team based on the Generative Adversarial Network (GAN) [21]. GAN trains a generative model to generate new data that are almost indistinguishable from real data, while training discriminant models to distinguish between real data and generated data. RFdiffusion On the basis of protein folding model Rose TTAFold based on the structure of denoising fine tuning, through adding three-dimensional Gaussian noise and simulated Brownian motion in protein structure add translation, rotating noise, training model in reverse noise reduction to minimize the prediction and the real structure, mean variance can in random initialization structure by denoising new protein backbone structure , in the protein monomer design, protein binding agent design, symmetrical oligomers design and metal binding protein design achieved excellent performance. The cryo-EM structure of the designed binder bound to influenza hemagglutinin was almost identical to the designed model, confirming the accuracy of RFdiffusion. RFdiffusion Make it possible to design simple molecules to design diverse functional proteins, bringing new possibilities to the field of protein design.

1.2.4AlphaPanda model

AlphaPanda [22] is an antibody design model independently developed by our team, inspired by AlphaFold2, integrating Transformer model and 3DCNN model. This model uses the Transformer model to capture global information, capture the local structural features of the antibody-antigen complex, and then the diffusion model to generate the sequence and structure of the antibody. While considering protein overall and local information, consider the pair and non-pair contacts, avoiding the defects of the generation process of the autoregressive model and self-consistent iterative model. AlphaPanda Has the common advantages of other advanced protein or antibody design software. It uses the isovariable neural network to process the coordinates in the 3D space and explicitly consider the antigenic structure, realizing the simultaneous diffusion generation of the sequence and structure of the CDR region. Transformer Model usually require large amounts of data to obtain considerable results, which is not necessary for CNN models, and introducing CNN to learn antibody structure effectively reduces the operation. Combining the above advantages, AlphaPanda achieves good performance and can be applied not only in antibody design, but also more widely in various fields of other protein design.

1.3 Application of the deep learning model

Deep learning is a learning method based on artificial neural networks to identify complex patterns and data by mimicking the processing of the human brain. In the field of antibody design, deep learning is able to process and analyze large bioinformatic datasets, including protein sequence, structural and functional data, thus identifying key parameters of antibody design. Using deep learning models predicts the effect of antibody variation on its affinity and specificity, thus guiding more effective design of antibody sequences. Deep learning models have unique advantages

in antibody structure prediction, antigen-antibody binding affinity prediction, antigen-antibody interaction site prediction, and antibody sequence design.

Conclusions to chapter 1

1. AI can assist in antibody design.
2. Artificial intelligence has broad prospects in the field of antibody design.

CHAPTER 2

OBJECT, PURPOSE, AND METHODS OF THE STUDY

2.1Object

Design of a new CDR using AI software.

2.2Purpose

AI software design to design new antibody CDR after learning antibodies in the PDB database.

2.3Methods of the study

2.3.1 Different application approaches for deep learning software

1、Antibody structure prediction of deep learning models

Antibody structure prediction is a key step in antibody design, and deep learning models achieve remarkable results in this regard.The main methods include:

Prediction method based on convolutional neural network (CNN): Convolutional neural network has been widely used in the fields of image classification, object detection and semantic segmentation. Convolutional neural network is essentially using a simple activation function to fit to a complex function between the network layers, which is something in a pure digital sense. With the advent of the information age with the progress of The Times and the explosion of big data, the convolutional neural network also faces great challenges and opportunities. In fact, more and more scenes can be adapted to the convolutional neural network. CNN has advantages in the field of image processing and can extract local features in the antibody structure. The researchers use CNN to predict the structure by transforming the 3D structure of the antibody into the image form.(2) Prediction of the antibody structure. By representing

the antibody structure as a graph, the GNN is used to learn relationships between nodes and thus improve prediction accuracy.

2、Antigen-antibody affinity prediction of the deep learning model

Antigen-antibody binding affinity is an important indicator to evaluate antibody performance. The deep learning model has the following methods for affinity prediction: (1) Method based on convolutional neural network: Using CNN for affinity prediction by extracting local features of antigen and antibody binding regions.(2) Method based on recurrent neural network: the sequence information of antigen and antibody is encoded, and the RNN is used to learn the correlation between sequences, so as to predict the affinity.(3) Methods based on the attention mechanism: The attention mechanism can enable the model to focus on the key areas of antigen-antibody binding and improve the accuracy of affinity prediction.

3、Antigenic antibody interaction site prediction

The binding properties between the protein molecules can be affected by their neighboring residues. The traditional sliding window strategy simply directly splicing the feature vectors of the target residues and adjacent residues as input to the prediction model, and does not distinguish the effects of adjacent residues at different positions, so the trained model cannot achieve the desired expected performance. In order to solve the above problems, Lu Shuai et al. proposed a new sequence feature representation method (SW-ATT) [23]. Combining SW-ATT with convolutional neural networks creates a method called SW-ATTCNN residue attribute prediction, which first updates the feature expression of the input sequence using a sliding window strategy integrated into the attention mechanism. Then, the convolutional neural

network (CNN) is used to extract the deep features of the local environment of the target residues. Finally, the classification prediction of residue properties is realized with the help of a fully connected neural network.

4、Antibody sequence design of the deep learning model

Antibody sequence design is the core task of antibody engineering. The research of deep learning model in this aspect mainly includes: (1) method based on generating antagonistic network (GAN): GAN generates antibody sequences with specific functions by learning the distribution of antibody sequences.(2) Method based on the variational autoencoder (VAE): the VAE encodes the antibody sequence into a low-dimensional vector, and the new antibody sequence is generated through the decoder.(3) Methods based on reinforcement learning: Reinforcement learning optimizes the generation process of antibody sequence to make the generated antibodies have better performance.

2.3.2Operation of the AI antibody design program

Artificial intelligence is the research and development can simulate, extend and extend the human intelligence theory, method, technology and application system of a new technology science, its itself is a comprehensive frontier and highly cross interdisciplinary disciplines, research, investigate the category wide and extremely complex, its development needs and computer science, mathematics, cognitive science, neuroscience and social science discipline depth fusion. With the improvement of experimental biology and artificial intelligence computing capabilities, the interdisciplinary discipline of computational biology has flourished, making it possible to turn biological mechanisms into computational models.

Machine learning is the core of artificial intelligence technology by studying how computers simulate or realize human learning behavior to acquire new knowledge or

skills, reorganize the existing knowledge structure and continuously improve their own performance. Data-based machine learning is one of the important methods in modern intelligent technology. It studies finding laws from the observed data (samples), and uses these laws to predict the future data or unobservable data.

The AI antibody design program is based on different machine learning models, and the antibody data in the antibody library is learned to assist and accelerate the design process of antibody molecules. As a subset of machine learning, deep learning mainly uses neural networks to learn the data. By learning the internal laws and representation levels of sample data, the machine can have the ability to analyze and learn like a human, and can recognize data such as text, image and sound. Through multi-layer processing, the initial "low-level" feature representation is gradually transformed into the "high-level" feature representation, and the learning tasks such as complex classification can be completed with the "simple model". Compared to traditional machine learning methods, deep learning models do not require manual feature extraction, and they can automatically learn from useful features as well as hierarchical structures in the raw data.

Based on deep learning, there are many different types of design models, such as diffusion model, autoregressive model, etc. The common ground of such models is all based on the common design of structural sequence of the secondary structure during the interaction of antigen antibodies to generate antibodies specifically against a specific antigen structure.

2.3.3The AI antibody design model

1、Diffusion model

The structure of protein is the basis of its function, and protein structure prediction is the core of antibody design. Accurate protein structure prediction models can predict not only amino acid positions but also amino acid orientation. The orientation of amino acids determines the orientation of their side chain extension and is therefore essential for the reconstruction of the all-atom structure.

In the process of protein structure prediction, the deep learning model takes protein sequence and multiple sequence alignment (MSAs) as input, and transforms it into 3 D structure [24,25,26].

During training, a noise plan was used to break the protein frame on a certain number of "time steps" (T) to a distribution that is indistinguishable from a random distribution (C α coordinates are destroyed by three-dimensional Gaussian noise and destroyed by Brownian motion on so3). During training, a PDB structure and a random time step (t) are sampled, and the t-noise step is applied to the structure. Data were generated by denoising the samples with a prior distribution.

Diffusion is divided into forward diffusion and backward diffusion. The forward diffusion process gradually adds noise to the data until the data distribution approximately reaches the prior distribution. The generative diffusion process starts with the prior distribution and iteratively transforms it into the desired distribution. An iterative perturbative denoising scheme has been empirically formulated for learning and generating amino acid orientations represented by SO (3) elements. The ward diffusion is the inference process of denoising.

Finally, a neural network is used to parameterize the multilayer perceptron (MLPs) to generate single and paired amino acids.

The dataset used to train the model was obtained from the SAbDab database. Structures with resolution below 4Å were first removed and antibodies against

nonprotein antigens were discarded. Then the sequence and structure co-design based on the CDR-H3 sequence with 50% sequence identity in the database.

To optimize the antibody, the CDR sequence and structure were first perturbed using a forward diffusion process. Then, the sequence starting from the first step (Tt) (remaining t) of the generation diffusion process is denoised to obtain a set of optimized antibodies.

2、Transformer model

The Transformer model introduces the self-attention mechanism [27], and by its extensive use, the model is able to weigh the importance of different positions in the input sequence when generating the output. The model can simultaneously consider the contextual connections of all amino acids in the peptide chain, rather than processing them in context as in traditional recurrent neural networks (RNNs). The Transformer model is able to generate highly diverse antibody sequences, which improves through the advantages of self-attention mechanism and parallel computing, and improves the efficiency of training and inference of the model. Various variants of Transformer have been proposed based on Transformer models against the characteristics of different antibodies. TransPHLA[27] is a pan-specific method that can be applied to rare and invisible human leukocyte antigen alleles. The core idea of the TransPHLA model is to apply self-attention to construct and optimize the model, The model consists of four main submodules: (1) embedded blocks (including the coding of amino acids in the sequence and the position information of the sequence); (2) Encoder block (apply multiple self-attention to follow the different components of the sequence, And shield the filling position of the sequence to prevent misleading the model); (3) Feature optimization block (using the full connection layer of the gyro channel with the first rising and then descending, Processing the features obtained

from the previous self-attention block, To achieve a better feature representation); (4) Projection blocks (using multiple fully connected layers to predict the final binding fraction). TransPHLA Not only achieve better performance with higher efficiency, but also addresses the limitations of many methods for HLA alleles and variable length peptides.

3、3DCNN model

The 3DCNN model can capture pairwise interactions in antibody-antigen complexes, and unpaired interactions in antibody-antigen complexes, and requires less training sample data, while avoiding the defects of the autoregressive model and self-consistent iterative model on the generation process. However, the 3 DCNN method does not consider the global information and does not adopt an efficient generation method.

2.3.4Design the combination of the models

The diffab model is based on the diffusion model, and the AlphaPanda model is combined with the diffusion model, the Transformer model, the 3DCNN model.These two models were used to learn antibodies in the PDB, each designing a new CDR.

Conclusions to chapter 2

1. Various design models are combined to form a model that can fully design antibodies.

2. The AI antibody design model is CDR designed based on the amino acid structure.

CHAPTER 3

EXPERIMENTAL PART

3.1 Construction of the diffusion model —— takes the diffab model as an example

The relationship between CDR sequences and the secondary structure of antigen antibody interactions and how to spatially distribute CDR of newly generated antibodies are key to antibody design. The model should be explicitly based on the 3 D structure of the antigen and generate a CDR suitable for antigenic structures in 3D space. This is essential for the generalization of the model to new antigens. Second, the interactions between amino acids are mainly determined by side chains, atomic groups that extend from the protein backbone. Thus, the model should be able to consider both the amino acid position and orientation. Third, the model should continuously optimize antibodies to enhance their ability to bind to antigen in different environments. To address these questions, the diffab model jointly samples the antibody CDR sequences and structures, making the joint distribution of the CDR sequences and their structures directly dependent on the antigen structure.The CDR is first initialized with an arbitrary sequence, position, and orientation, and the diffusion model first aggregates information from the antigen and antibody frames.Then, the amino acid type, position and orientation of each amino acid on the cdr. In the final step, the model is based on the predicted orientation using the side chain filling algorithm at the atomic level. The diffab model is the first to propose the common design of the antibody sequence and structure through the three-dimensional structure of the antigen. While designing the protein sequence and coordinates, the design of amino acids has successfully achieved antibody design at atomic resolution.

Traditional model calculation method is mainly based on sampling algorithm manual and statistical energy function and iterative modification of protein sequence

and structure, these methods are inefficient, and due to the rough energy environment, easy to fall into the local optimal state, base, in the sequence method although more effective, but only according to previously observed new antibodies, but it is difficult to produce antibodies for specific antigen structure. Deep learning methods generate protein sequences by using the language model and model the 3D structure of antigens, considering not only the backbone atomic coordinates but also the orientation of amino acids.

3.2 Antibody design of diffusion model —— Example of diffab model

3.2.1 Experimental objectives

The CDR was designed by learning from diffab, using 8a67 from the PDB database as a template.

3.2.2 Installation and startup of the design program

Conda is an open source package management system and environment management system used to install different versions of software packages and their dependencies, and can easily switch between them. It works for Linux, OSX and Windows and is created for Python programs, but can be packaged and distributed with any software.

The diffab program in this experiment was used in a conda system-based Linux environment.

FinalShell Is an SSH client tool with multi-platform support for Windows, Mac OS X, and Linux. Mainly used for all-in-one server management.

Start diffab with conda in finalshell.Figure 3-1 The diffab of the program was initiated

Figure 3.1 – Start the diffab environment

3.2.3 Antibody design and analysis

1、Conduct the antibody design, and the input:

python design_pdb.py /home/data/t030413/diffab-origin/diffab-main/8a67/8a67.pdb --heavy H --light L --config ./configs/test/codesign_single06.yml -d cpu

2、Analyze the design structure and input:

Python design_eval_single.py--root=/home/data/t030413/diffab-origin/diffab-main/results/codesign_single06/7xjf.pdb_2023_10_08__10_37_06 --pfx='‘

3.3 Antibody design of diffusion model —— Example of AlphaPanda model

3.3.1Experimental target

Using 8a67 in the PDB database as a template, let AlphaPanda learn from it and design a new CDR.

3.3.2Installation and startup of the design program

The software and procedures used for antibody design and analysis are the same as diffab.

3.4 Design antibody performance analysis

3.4.1 Main evaluation indicators

1、 RMSD:RMSD is the root mean square squared difference of C α between the generated CDR and the antibody real CDR. In the protein structure analysis, and molecular dynamics simulations, the RMSD value is used to measure the extent of the deviation of each part from the average position, that is, the magnitude of the motion of each atom. The lower RMSD represented less differences compared to the native CDR, suggesting a better CDR structure to bind antigen. The threshold value of RMSD is usually limited within a certain range to determine the similarity of the structure, the value of RMSD is usually 1-2Å. When the RMSD is less than or equal to 1A, the two structures are considered to be very similar; when the RMSD is greater than 2Å, the two structures are considered to be quite different, and the protein of about 1.5Å can reach the atomic accuracy.

2、Seqid: Seqid usually represents sequence similarity, namely the degree of similarity between two protein sequences, which is usually expressed in percentage form. It measures the proportion of identical amino acid residues in two sequences. Higher sequence similarity means that the two sequences may be more structurally and functionally similar because they share more amino acid residues. The higher Seqid values indicate that the designed antibody sequence shows a high similarity to the target sequence. It is generally believed that designed antibodies with Seqid values around 30% have good homology with the native structure.

3、ddG:The ddG is the difference between the energy of the designed antibody (dG-gen) and the natural antibody energy (dG-ref). The smaller the dG value, the more stable the antibody is.

3.4.2 CDR designed by diffab model

100 CDR 1,100 CDR 2 and 100 CDR 3 were designed by diffab, and the CDR with the lowest RMSD was selected as a representative to map their 3 D structure.

Table 3.1 –The three CDRs designed by diffab

CDR	filename	rmsd	seqid	dG_gen	dG_ref	ddG
H_CDR1	0053.pdb	2.96743	7.692308	2570.3850	1532.00403	1038.38098
H_CDR2	0088.pdb	0.22689	40	1537.231079	1531.873535	5.357544
H_CDR3	0045.pdb	2.477394	23.529412	4116.40332	1532.004028	2584.399292

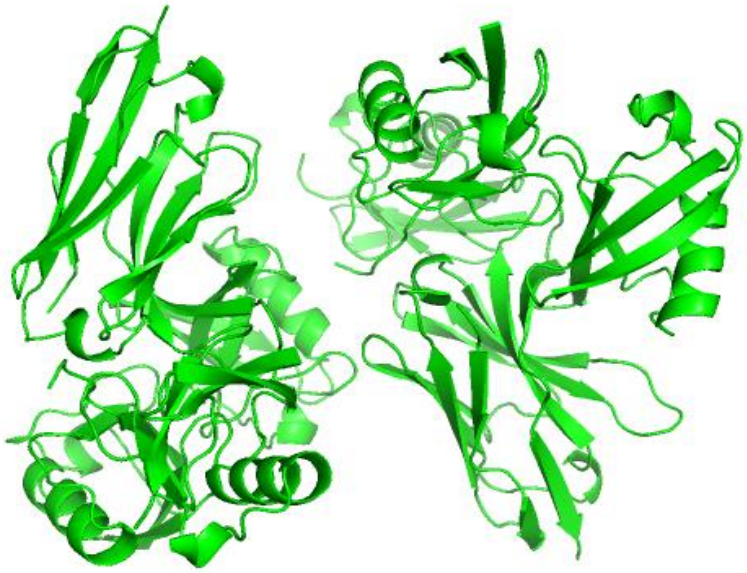

Figure 3.2 – The graphic model of 8a67

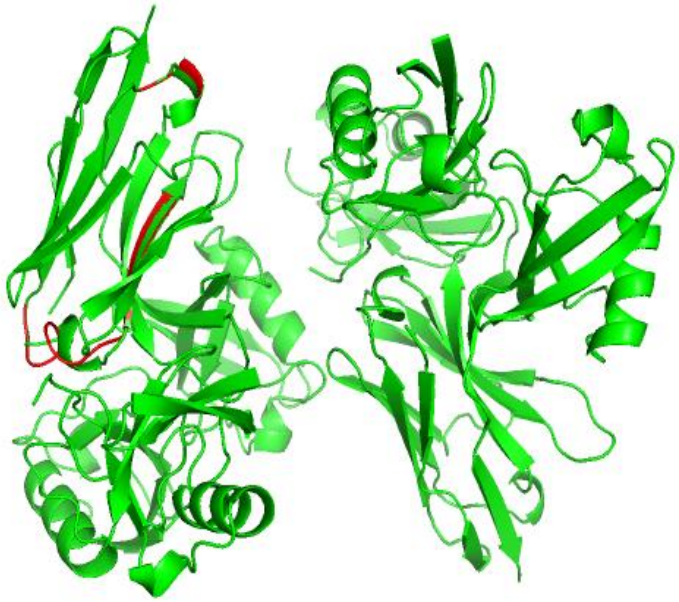

Figure 3.3 – CDR1 0053

Figure 3.4 – CDR2 0088

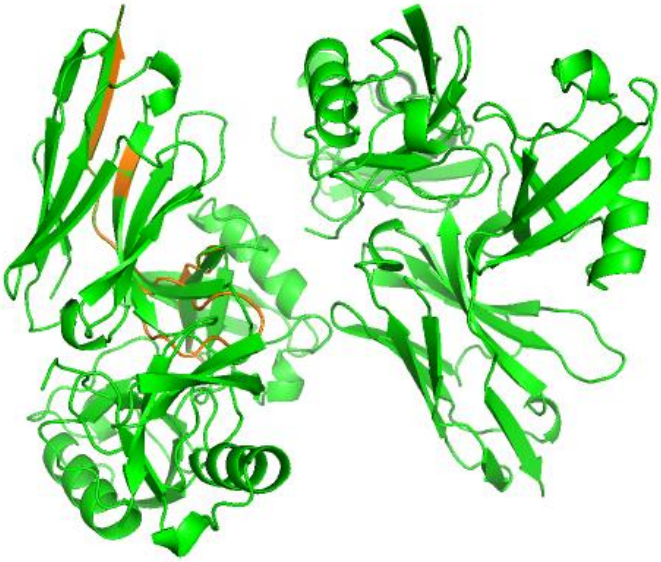

Figure 3.5 – CDR3 0045

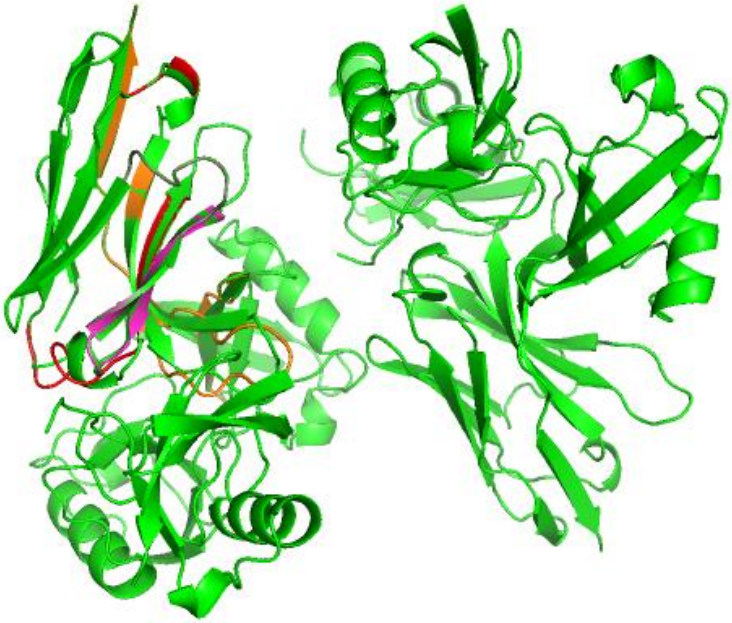

Figure 3.6 – three CDRs designed by diffab

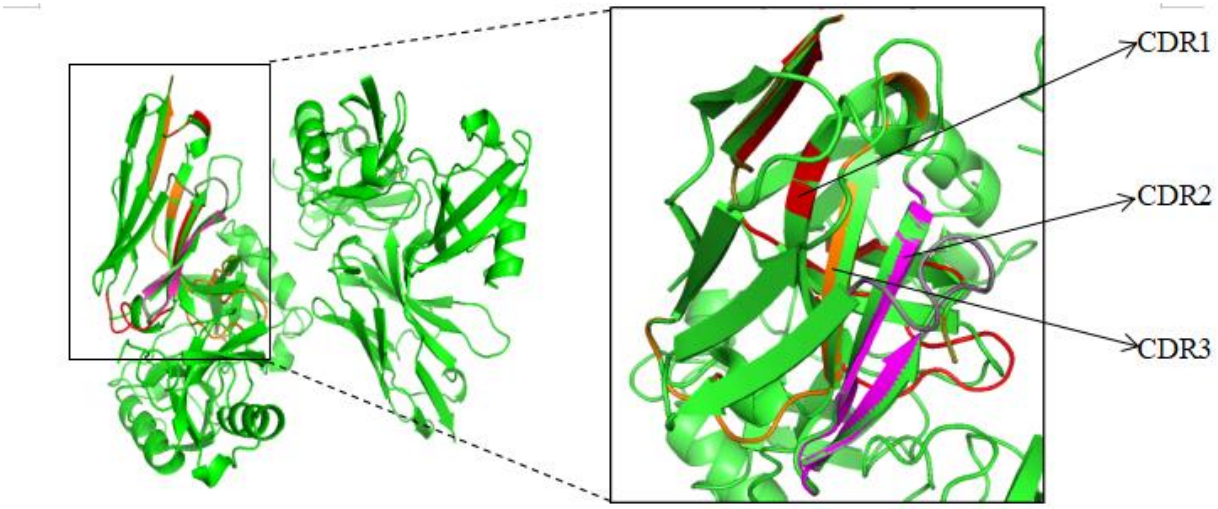

Figure 3.7 – diffab The three CDRs designed

3.4.3 CDR designed by AlphaPanda model

200 CDR 1,200 CDR 2 and 200 CDR 3 were designed by AlphaPanda, and the CDR with the lowest RMSD was selected as the representative to map their 3 D structure.

Table 3.2–The three CDRs designed by AlphaPanda

CDR	filename	RMSD	Seqid	dG_gen	dG_ref	ddG
H_CDR1	0124.pdb	2.601117	7.692308	4083.1604	1531.873535	2551.286865
H_CDR2	0123.pdb	0.662668	0	1824.910278	1531.970581	292.939697
H_CDR3	0107.pdb	3.960965	20	11081.05469	1531.605957	9549.44873

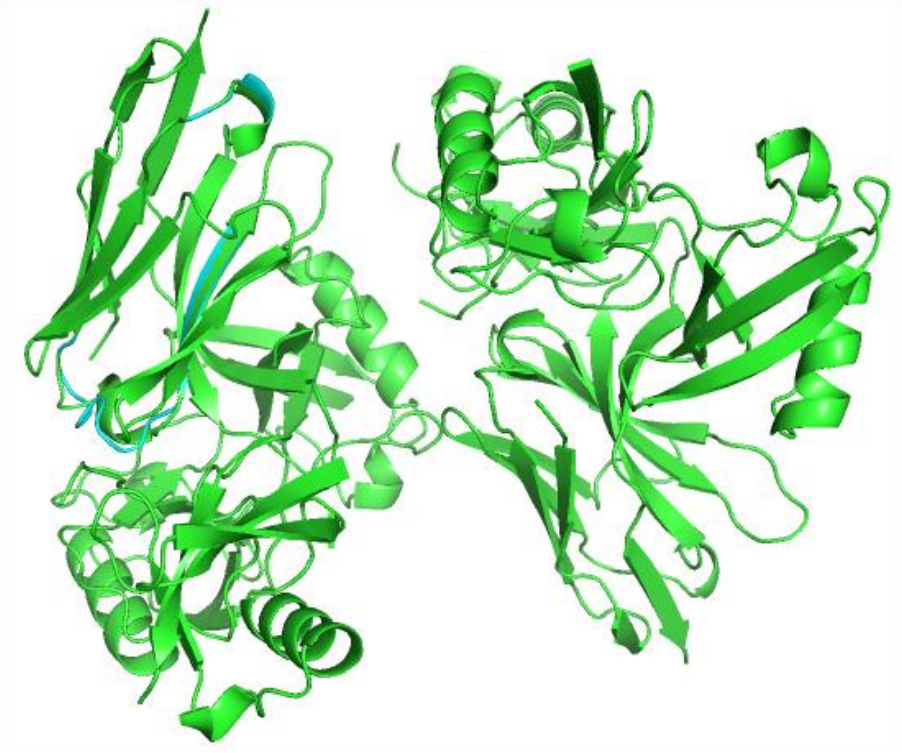

Figure 3.8 – AlphaPanda Design of the H_CDR1 0124

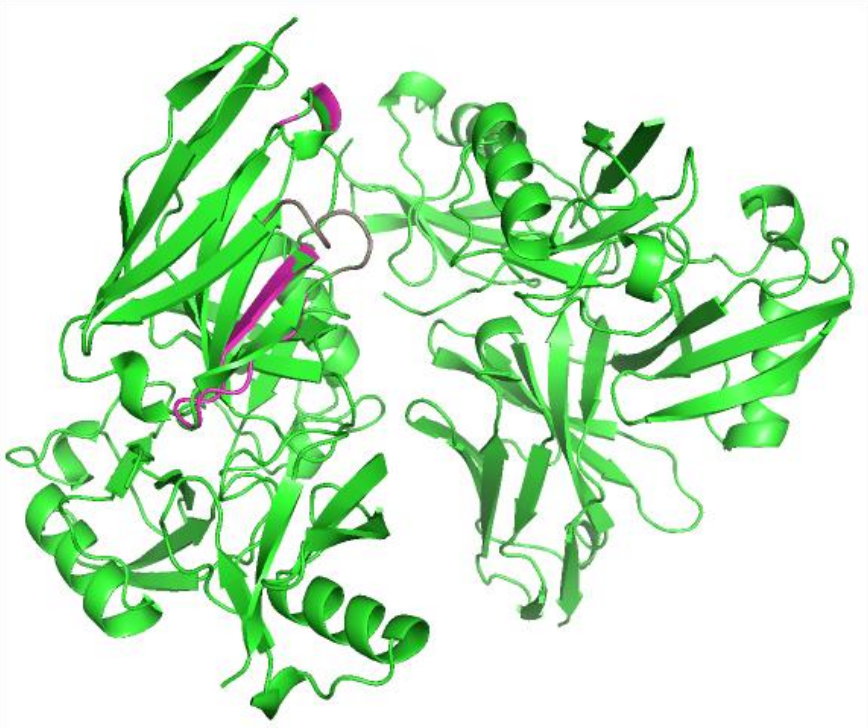

Figure 3.9 – AlphaPanda Design of the H_CDR2 0123

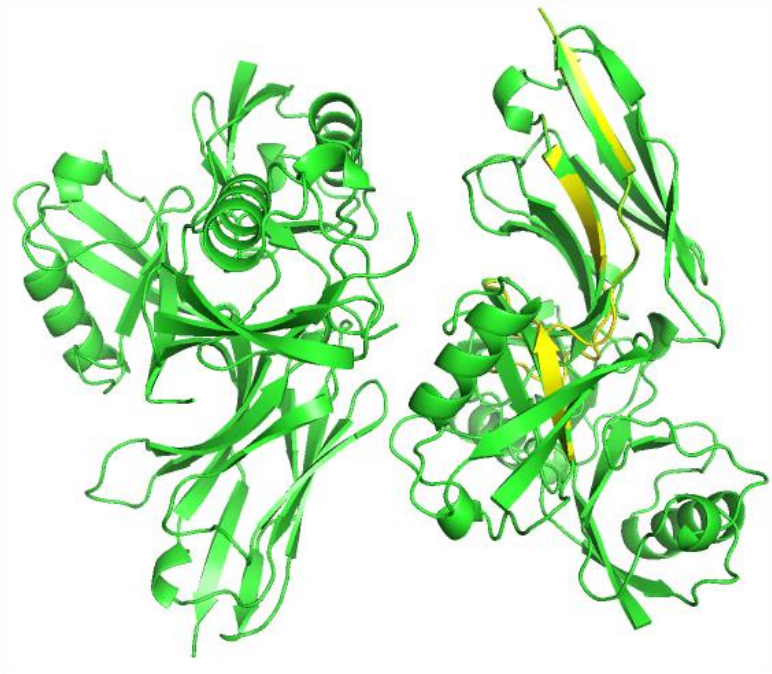

Figure 3.10 – AlphaPanda Design of the H_CDR3 0107

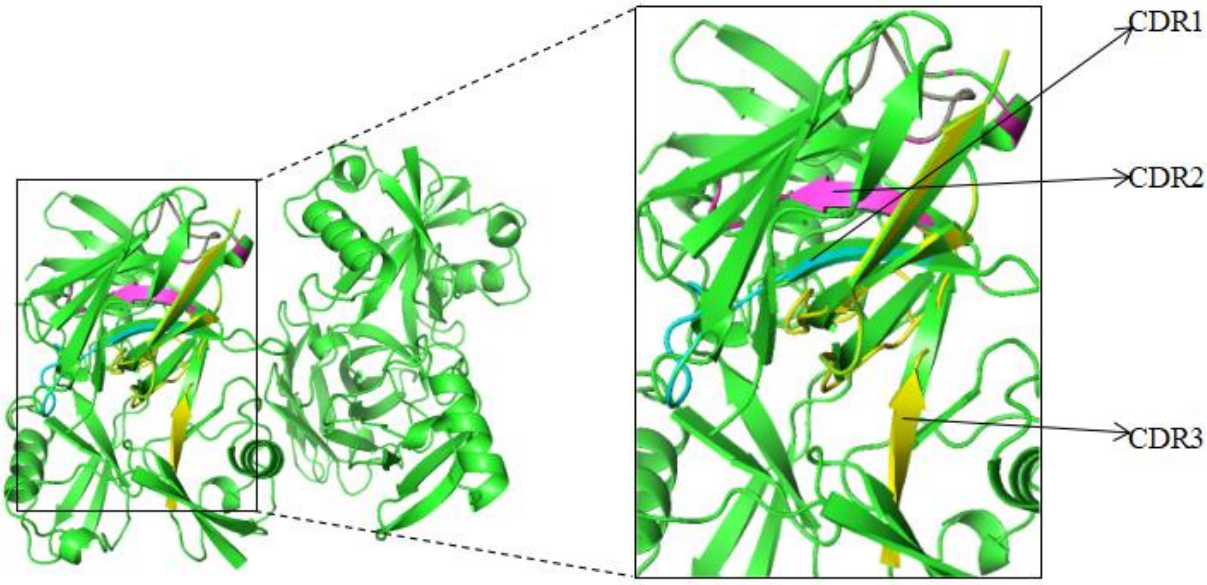

Figure 3.11 – AlphaPanda The three CDRs designed

Table 3.3–diffab and AlphaPanda designed the antibodies

		Group statistics			Independent samples' t-test			
	group	The number of cases	average value	standard deviation	F	conspicuousness	free degree	The sig (double-tailed)
RMSD (CDR1)	diffab	100	3.479473	0.199845	1.768	0.185	298	0
	AlphaPanda	200	3.096392	0.2112423				
RMSD (CDR2)	diffab	100	0.597853	0.1971638	11.106	0.001	285.5	0
	AlphaPanda	200	1.225389	0.3205259				
RMSD (CDR3)	diffab	100	5.365680	1.780742131	4.917	0.027	153.752	0
	AlphaPanda	200	6.594885	1.304545138				
Seqid (CDR1)	diffab	100	15.033722	10.65926507	172.363	0	119.21	0
	AlphaPanda	200	8.899085	4.768450991				
Seqid (CDR2)	diffab	100	35.873016	9.561413591	21.36	0	247.986	0
	AlphaPanda	200	24.706349	12.38648082				
Seqid (CDR3)	diffab	100	20.397758	8.754151246	6.917	0.009	161.433	0.222
	AlphaPanda	200	19.170635	6.853498616				
ddG (CDR1)	diffab	100	931.263474	557.2611156	38.468	0	294.998	0
	AlphaPanda	200	3462.876782	1007.219385				
ddG	diffab	100	166.638082	288.0456505	2.054	0.153	298	0.469

(CDR2)	AlphaPanda	200	195.055208	334.6308831				
ddG	diffab	100	8311.244534	4184.65				
					0.228	0.633	298	0
(CDR3)	AlphaPanda	200	12502.66387	3981.934659				

3.5diffab Design performance comparison with AlphaPanda

From the data available in Table 3-3:

1. RMSD metric:

(1) CDR 1: The mean (3.479) was higher than the mean of AlphaPanda (3.096). Independent sample t-test showed significant difference between diffab and AlphaPanda (P <0.05). t> 0 indicates that the RMSD value of diffab was generally greater than AlphaPanda, that is, the structural similarity of CDR 1 designed by diffab was slightly lower than that of CDR 1 designed by AlphaPanda.

(2) CDR 2: The mean (0.598) of diffab group was higher than that of AlphaPanda group (1.225). Independent sample t-test showed significant difference between group and AlphaPanda group diffab (P <0.05). <0 of t indicates that the RMSD value of diffab was generally less than AlphaPanda, meaning the structural similarity of CDR 2 designed by diffab was slightly higher than CDR 2 designed by AlphaPanda.

(3) The value of CDR 3: the mean (5.366) of diffab group was higher than the mean of AlphaPanda group (6.595). Independent sample t-test showed a significant difference between diffab group and AlphaPanda group (P <0.05). The t <0 indicated that the RMSD value of diffab was generally less than AlphaPanda, and the structural similarity of CDR 3 designed by diffab was slightly higher than the CDR 3 designed by AlphaPanda.

2.Seqid metric:

(1) CDR 1: the mean of the diffab group (15.034) is lower than the mean of the AlphaPanda group (8.899), The results of the independent samples t-test showed that, A significant difference between the diffab and AlphaPanda groups (P <0.05), The t> 0 indicates that the Seqid value of diffab is overall greater than the AlphaPanda, That is, the sequence similarity of CDR 1 designed by diffab is higher than that of CDR 1 designed by AlphaPanda, However, the mean value of both groups was less than 30%, Therefore, the design of both models of CDR 1 is unsatisfactory in terms of sequence similarity.

(2) CDR 2: the mean value of the diffab group (35.87) is higher than the mean value of the AlphaPanda group (24.706), The results of the independent samples t-test showed that, A significant difference between the diffab and AlphaPanda groups (P <0.05), The t> 0 indicates that the Seqid value of diffab is overall greater than the AlphaPanda, That ely, the CDR 2 sequence similarity designed by diffab is higher than the CDR 2 designed by AlphaPanda, The diffab group had a mean value above 30%, The AlphaPanda group approached 30%, diffab A high level of design in the sequence similarity of CDR 2.

(3) CDR 3: the mean value (20.398) is higher than the mean value of the AlphaPanda group (19.171), The results of the independent samples t-test showed that, There was no significant difference between the diffab and AlphaPanda groups (P> 0.05), The t> 0 indicates that the Seqid value of diffab is overall greater than the AlphaPanda, Namely, the CDR 3 sequence designed by diffab is more similar than the CDR 3 designed by AlphaPanda, However, the mean value of both groups was less than 30%, Design results showed little difference between the two groups.

3.ddG metric:

(1) CDR 1: diffab 3 (931.263) was lower than the mean of AlphaPanda group (3462.877). Independent sample t-test showed significant difference between diffab and AlphaPanda group (P <0.05), t <0 indicates that ddG value of diffab is less than AlphaPanda, or CDR 1 designed by diffab is more stable than CDR 1 designed by AlphaPanda.

(2) The mean value of CDR 2: diffab group (166.638) was lower than the mean value of AlphaPanda group (195.005). The independent sample t-test showed that there was no significant difference between diffab group and AlphaPanda group (P> 0.05). The value of t. <0 indicates that the dG value of diffab is generally less than AlphaPanda, that is, the CDR 2 designed by diffab is slightly more stable than the CDR 2 designed by AlphaPanda, but the difference was not obvious.

(3) The mean of CDR 3: diffab group (8311.246) is lower than the mean of AlphaPanda group (12502.664). Independent sample t-test showed significant difference between diffab group and AlphaPanda group (P <0.05), t <0 indicates that the ddG value of diffab is less than AlphaPanda, that CDR 3 designed by diffab is more stable than CDR 3 designed by AlphaPanda.

Conclusions to chapter 3

1. Based on the above analysis, based on the performance of RMSD, seqid and ddG, it can be preliminarily inferred that the CDR designed by diffab is generally better on these indicators. In terms of RMSD, the average RMSD of CDR 1 and CDR3 designed by diffab and AlphaPanda is higher than 1.5Å, and the RMSD of CDR 2 is lower than 1.5Å. Therefore, both models need improvement for CDR 1 and CDR 3. In terms of Seqid, the Seqid value of CDR 2 designed by the diffab group only reaches 30%, so

there is much room for improvement in both models for sequence consistency. In terms of energy, for CDR 1 and CDR 3, the diffab group had significantly lower mean CDR; for CDR 2, the mean value and the difference was not significant, only 0.0067% CDR in diffab group was lower than the natural CDR, and only 0.0233% CDR in AlphaPanda group was lower than the natural CDR.

CONCLUSIONS

The ability of AI models to predict their 3D structure based on known protein sequences is essential for understanding how antibodies bind to their specific antigen. By analyzing a large amount of antibody-antigen interaction data, neural networks can discover design principles that improve antibody affinity. AI technology is able to quickly analyze and process large amounts of data, significantly shortening the cycle of antibody discovery and design. AI can not only optimize the existing antibody design, but also discover new design concepts and goals, and expand the boundaries of research.

The antibodies designed by AI can develop new treatment options for disease targets that are difficult to overcome by traditional methods, providing new hope for the treatment of difficult diseases. The application of AI technology in antibody design also heralds the transformation of pharmaceutical research and development mode. In the future, the design and development of antibodies may no longer rely on large laboratory facilities and complicated experimental processes, but rather proceed quickly and effectively on the basis of computer simulation and prediction. This will not only significantly reduce the cost of research and development, but also accelerate the marketing process of new drugs to better meet the treatment needs of patients worldwide.

How the utility and practicability of antibody design using artificial intelligence program is an important problem in the field of antibody drug design. Through performance evaluation, the bottleneck and deficiency of the program can be identified, so as to carry out targeted improvement and optimization. By comparing the performance of different design programs, we can understand the effect of AI in biological applications and develop corresponding optimization strategies.

The high stability of nanoantibodies designed by diffab enables the model to complete a wide range of antibody design tasks and can achieve competitive performance. AlphaPanda The combination of models effectively reduces the amount of operation, and shows superior performance in the design of CDR 2. There is room for improvement in the design of the two groups of models in the design of CDR, especially in improving sequence consistency and reducing RMSD value. At the same time, the advantages and disadvantages of the design also need to be combined with practical application scenarios and cost-effectiveness and other factors. The main limitations of antibodies are that the design of new antibodies relies on secondary structures bound to the target antigen, and it is unclear whether the antibodies produced can be produced in wet laboratories and actually bind to the target, and more effort is needed to design a biologically effective antibody. In the future, with the high throughput of protein production and test method development of protein design application scenarios will continue to expand, for example, under the cell-free protein synthesis technology will have stronger controllability, lower economic and time cost, higher safety [28,29], greatly accelerated design, synthesis, inspection, analysis of optimization process, make the protein design project to rapid advance. In addition, the training of AI models requires a large amount of accurately labeled high-quality data, but in the field of molecular biology, such data is expensive and scarce, which is the main "bottleneck" that limits the application of protein design. How to develop new high-throughput detection methods, how to optimize and develop new design models, and how to design new antibodies from scratch according to the physical information of antibodies are all the problems to be solved.

LIST OF REFERENCES

1. Zhang Xin, Deng Ruixin & Liu Jianqing. (2024). Structural properties of nanobodies and their research progress in the treatment of infectious diseases. Chinese Journal of Public Health (02), 252-256.

2. SUN Baihe, WU Yue, ZHAO Rui, LOU Yuxin, LI Wanting, LI Yanfei & MA Linlin. (2023). Research progress on therapeutic nanobodies with different expression systems. Chinese Journal of Bioengineering (11), 43-55.

3. Tiller, K.E. and P.M. Tessier, Advances in Antibody Design. Annu Rev Biomed Eng, 2015. 17: p. 191-216.

4. Kuroda, D., et al., Computer-aided antibody design. Protein Eng Des Sel, 2012. 25(10): p. 507-21.

5. CHEN Kaixian. (2024). Biomedical Research and Artificial Intelligence Technology in China. Shanghai Medical Corp (01), 1-7.

6. Miao Hongjiang, Dong Zekai, Xiang Qiuru & Xue Guirong. (2024). Research progress of artificial intelligence protein design technology and its application and challenges in biomedical innovation and development. Shanghai Pharmaceutical Co Ltd (07), 1-9+55.

7. Cai Yonglian & Peng Haoshu. (2023). The disciplinary diversity of artificial intelligence technology and its integration and development with the field of biomedicine. Journal of University of Shanghai for Science and Technology(Social Sciences) (03), 330-336.

8. Młokosiewicz, J., et al., AbDiver: a tool to explore the natural antibody landscape to aid therapeutic design. Bioinformatics, 2022. 38(9): p. 2628-2630.

9. Sormanni, P. and M. Vendruscolo, Protein Solubility Predictions Using the CamSol Method in the Study of Protein Homeostasis. Cold Spring Harb Perspect Biol, 2019. 11(12).

10. Hebditch, M., et al., Protein-Sol: a web tool for predicting protein solubility from sequence. Bioinformatics, 2017. 33(19): p. 3098-3100.
11. Zambrano, R., et al., AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. Nucleic Acids Res, 2015. 43(W1): p. W306-13.
12. Dhanda, S.K., et al., IEDB-AR: immune epitope database-analysis resource in 2019. Nucleic Acids Res, 2019. 47(W1): p. W502-w506.
13. Marks, C., et al., Humanization of antibodies using a machine learning approach on large-scale repertoire data. Bioinformatics, 2021. 37(22): p. 4041-4047.
14. Luo, S., et al., Antigen-Specific Antibody Design and Optimization with Diffusion-Based Generative Models. 2022.
15. Schneider, C., M.I.J. Raybould, and C.M. Deane, SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. Nucleic Acids Res, 2022. 50(D1): p. D1368-d1372.
16. Adolf-Bryfogle, J., et al., RosettaAntibodyDesign (RAbD): A general framework for computational antibody design. PLoS Comput Biol, 2018. 14(4): p. e1006112.
17. Yan, Y., et al., HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res, 2017. 45(W1): p. W365-w373.
18. Shi, C., et al., Protein Sequence and Structure Co-Design with Equivariant Translation. ArXiv, 2022. abs/2210.08761.
19. Orengo, C.A., et al., CATH--a hierarchic classification of protein domain structures. Structure, 1997. 5(8): p. 1093-108.
20. Bennett, N.R., et al., Atomically accurate de novo design of single-domain antibodies. bioRxiv, 2024.
21. Krichen, M. Generative Adversarial Networks. in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2023.

22. Hu, Y., et al., Combining transformer and 3DCNN models to achieve co-design of structures and sequences of antibodies in a diffusional manner. bioRxiv, 2024: p. 2024.04.25.587828.
23. LU Shuai. (2022). Research on Prediction of Antigen-Antibody Interaction Sites Based on Deep Learning Ph.D. (Dissertation, Zhengzhou University). Ph.D.
24. Krichen, M. Generative Adversarial Networks. in 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2023.
25. Jumper, J., et al., Highly accurate protein structure prediction with AlphaFold. Nature, 2021. 596(7873): p. 583-589.
26. Baek, M., et al., Accurate prediction of protein structures and interactions using a three-track neural network. Science, 2021. 373(6557): p. 871-876.
27. Chu, Y., et al., A transformer-based model to predict peptide–HLA class I binding and optimize mutated peptides for vaccine design. Nature Machine Intelligence, 2022. 4(3): p. 300-311.
28. Silverman, A.D., A.S. Karim, and M.C. Jewett, Cell-free gene expression: an expanded repertoire of applications. Nat Rev Genet, 2020. 21(3): p. 151-170.
29. Dondapati, S.K., et al., Cell-Free Protein Synthesis: A Promising Option for Future Drug Development. BioDrugs, 2020. 34(3): p. 327-348.