# **Atom-Based Zero-Shot for Antigen&Antibody**

Bo Yu, Haobo Zheng

Zhejiang University

[3210105512@zju.edu.cn](mailto:3210105512@zju.edu.cn), 3210105321@zju.edu.cn

**Abstract**

*We designed an atom-based prediction method for antibody-antigen compounds. Compared to the latest works, this work uses atom position information instead of residue position to predict the binding of antibodies and antigens.*

*The general steps are to extract the protein space and sequence data from the dataset of antibody-antigen pairs; after extracting the data, encode the antibody and antigen through the CDConv network to extract features; after feature extraction, use the idea of comparative learning to divide into positive and negative samples for processing; The final model performs a zero-shot attempt to infer the antibody or antigen with the highest binding degree.*

# **Introduction**

Antibodies are an important class of protein molecules that play a key role in the immune system. The structure and function of antibodies have been extensively studied and are widely used in the fields of medicine and biology. However, traditional approaches to antibody design are often experimentally and computationally intensive and inefficient. In recent years, with the development of deep learning and artificial intelligence technology, more and more researchers have begun to explore the use of machine learning methods to design antibodies. We hope to give a new antibody as the input and find out the most possible antigen for it to dock as the output. We think this is important in some practical applications and the detection of the effectiveness of the designed antibody without expensive and time-consuming experiments is helpful for the industry.

To address this problem, we hope to predict antibody-antigen binding through all-atom-based antibody-antigen information. We design an encoder to convert the full-atom information of antibodies and antigens into amino acid information for embedding, then pass through the CDConv neural network, and subsequently adjust the feature dimension of each protein through a feature-extractor, each process multiple positive and negative samples at a time, perform comparative learning and calculate Cross-Entry-Loss to train the model. Finally, we use the trained model for Zero-Shot prediction.

# **Related Work**

**End-to-End-Full-Atom-Antibody-Design** Antibody design is an essential yet challenging task in various domains like therapeutics and biology. There are two major defects in current learning-based methods: 1) tackling only a certain subtask of the whole antibody design pipeline, making them suboptimal or resource-intensive. 2) omitting either the framework regions or side chains, thus incapable of capturing the full-atom geometry. To address these pitfalls, we propose a dynamic Multi-channel Equivariant grAph Network (dyMEAN), an end-to-end full atom model for E(3)-equivariant antibody design given the epitope and the incomplete sequence of the antibody. Specifically, we first explore structural initialization as a knowledgeable guess of the antibody structure and then propose shadow paratope to bridge the epitope-antibody connections. Both 1D sequences and 3D structures are updated via an adaptive multi-channel equivariant encoder that is able to process protein residues of variable sizes when considering full atoms. Finally, the updated antibody is docked to the epitope via the alignment of the shadow paratope. Experiments on epitope-binding CDR-H3 design, complex structure prediction, and affinity optimization demonstrate the superiority of end-to-end framework and full-atom modeling.

**CDConv-for-geometry-sequence-modeling-in-proteins** The structure of proteins involves 3D geometry of amino acid coordinates and 1D sequence of peptide chains. The 3D structure exhibits irregularity because amino acids are distributed unevenly in Euclidean space and their coordinates are continuous variables. In contrast, the 1D structure is regular because amino acids are arranged uniformly in the chains and their sequential positions (orders) are discrete variables. Moreover, geometric coordinates and sequential orders are in two types of spaces and their units of length are incompatible. These inconsistencies make it challenging to capture the 3D and 1D structures while avoiding the impact of sequence and geometry modeling on each other. This paper proposes a Continuous-Discrete Convolution (CDConv) that uses irregular and regular approaches to model the geometry and sequence structures, respectively. Specifically, CDConv employs independent learnable weights for different regular sequential displacements but directly encodes geometric displacements due to their irregularity. In this way, CDConv significantly improves protein modeling by reducing the impact of geometric irregularity on sequence modeling. Extensive experiments on a range of tasks, including protein fold classification, enzyme reaction classification, gene ontology term prediction and enzyme commission number prediction, demonstrate the effectiveness of the proposed CDConv

# **Background Knowledge**

A protein comprises one or more long chains of amino acid residues. An antibody is a Y-shaped symmetric protein with two identical sets of chains, as illustrated in Figure 1. Each set contains a heavy chain and a light chain, either of which consists of several constant domains and a variable domain. As their names suggest, the constant domains keep unchanged across different antibodies; while the variable domain varies to enable different binding specificity for different antigens, making it the main focus of antibody design. We denote the variable domains of the heavy chain and the light chain by VH and VL, respectively. The variable domain is further divided into alternating arrangements of four framework regions (FRs) and three complementarity determining regions (CDRs). The binding regions of an antigen and an antibody are called an epitope and a paratope, separately. In this paper, the paratope refers to CDR-H3 in the heavy chain following Jin et al. (2022), since it is highly variable and dominates binding (MacCallum et al., 1996).

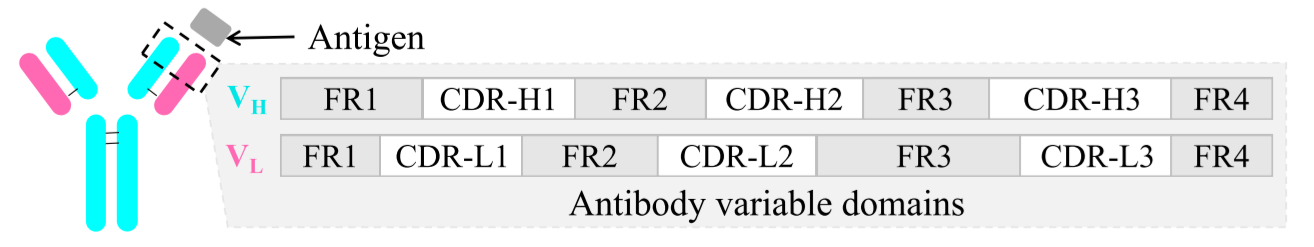


Figure 1. Variable domains in the heavy/light chain (VH / VL).

# **Atom-Based Zero-Shot**

In this section, we first briefly introduce the data format and preliminary processing method; secondly, we introduce the structure of neural network; thirdly, we review the ideas and methods of Contrastive-Learning and Zero-Shot; finally, we look at the learning and forecast of antibody&antigen in our actual model

## **Atom-based data processing method**

In our model, we read the full atomic sequence and spatial coordinates of antibodies and antigens. We first consider using the average coordinates of all atoms to replace the coordinates of amino acids into the neural network; after training, we find that the effect is not good, with the guess that the information between atoms is erased by taking the average directly; so we try to use all atomic coordinates of amino acids, and fit them through the MLP model to obtain the position coordinates of amino acids. After training, it is found that using MLP to process all-atom coordinates does have a significant hint on the model training results.

## **Network construction**

The first level transmits amino acid position coordinates, coordinates after rotation invariance processing, and amino acid sequence embedded information to the next layer; the second level contains 8 layers, and each layer contains 4 components. The first component is Identity, after the sequence information is processed, it is directly input as part of the next layer, and the other part of the input is the output after all the information is processed by the other three components: MLP, CDConv, and Linear. In this way, it is passed to the next level after passing through the layer eight times. It is worth noting that in the second level, we reduce the number of amino acids by half and expand the feature dimension of amino acids to double every time we pass through two layers. At the output of the second level, a protein is represented as an N-dimensional vector. The third level is a feature-extractor, which is actually an MLP, extracting the N-dimensional vector of each protein into a fixed-dimensional vector.

## **Contrastive-Learning and Zero-Shot**

Contrastive learning is very flexible. We only need to define positive and negative samples to train the corresponding pairing information. As shown in Figure 2, after multiplying the two matrices, a new matrix is obtained. The blue elements on the diagonal of this matrix represent positive sample pairs, and the rest are negative sample pairs. Calculate the CrossEntryLoss on the matrix and optimize the Loss.

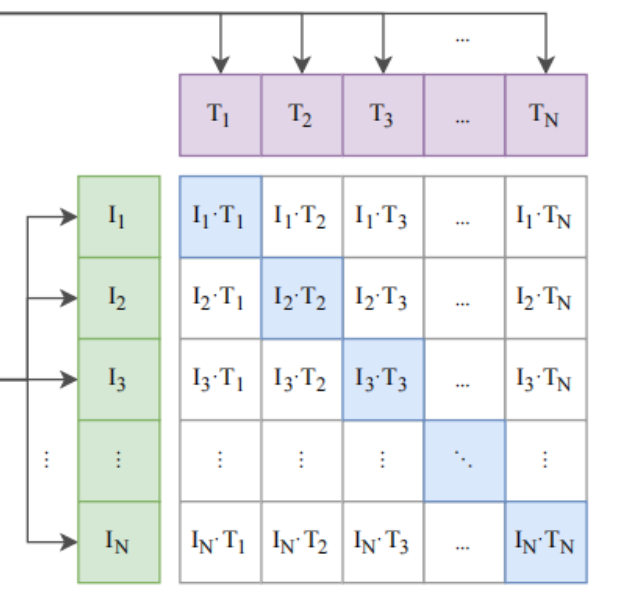


Figure 2. Result Matrix of Contrastive-Learning

On this basis, Zero-Shot uses the model trained by contrastive learning to make predictions. As shown in figure 3, after a vector is multiplied by multiple vectors, the element with the largest value corresponds to the positive sample.

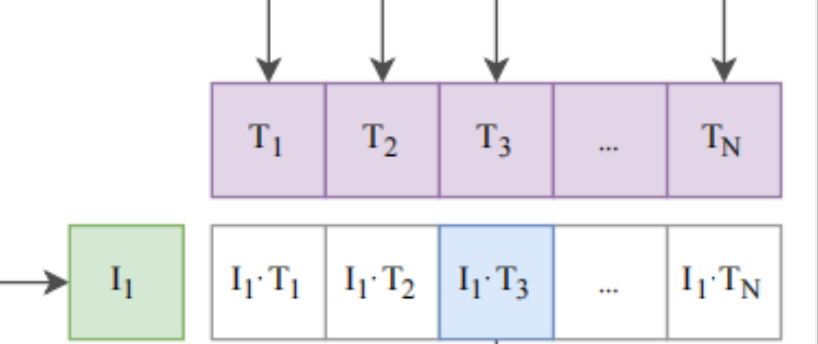


Figure 3. Zero-Shot vector

## **Actual combination comparison**

Using the method mentioned in Section 4.3, multiply the vectors of antibodies and antigens output by the neural network in batches, and calculate the CrossEntryLoss after obtaining the matrix to train the model. In this way, a pair of combined antibodies and antigens can be used as positive samples, and other antibodies and antigens can be used as negative samples. After training, we can use the model for Zero-Shot method prediction. Multiply the eigenvector of an antigen by the eigenvectors of multiple antibodies, and the antibody corresponding to the maximum value is the pair of antigen-antibody with the highest binding degree.

# **Experiments**

## **Valid-Loss and Batch-Size**

In theory, the larger the batch\_size, the larger the number of positive and negative samples, the better the training effect, and the smaller the loss

But because there are fewer epochs, the loss should become larger in the early stage because there are more negative samples, and the lower the probability of correct prediction, the greater the loss, shown in Figure 3.

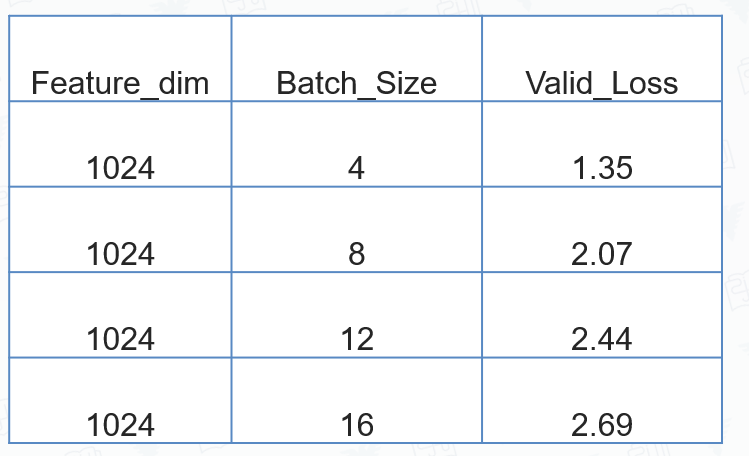


Table 1. Valid-Loss and Batch-Size

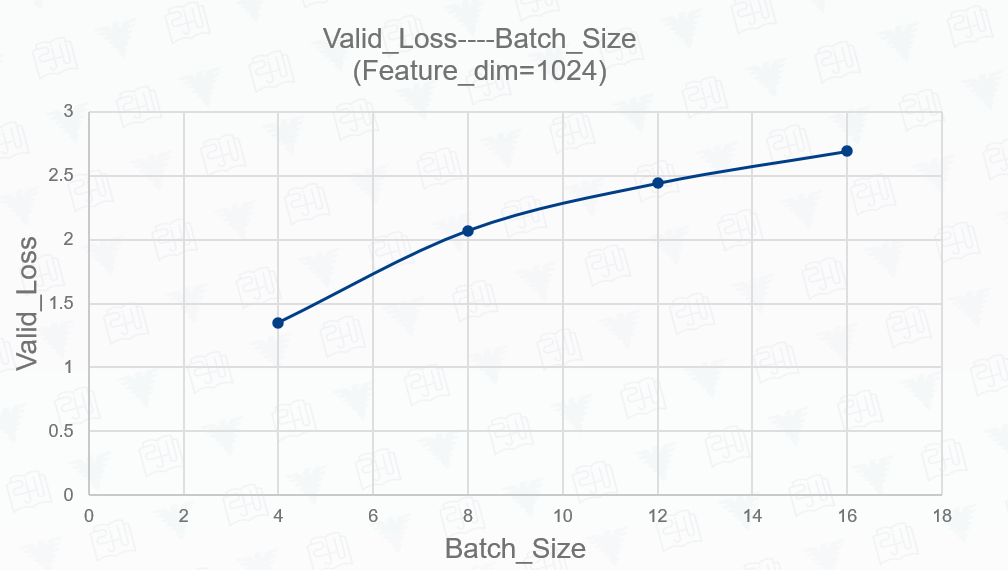


Figure 3. Valid-Loss and Batch-Size

## **Valid-Loss and Feature-dimension**

In theory, loss should first decrease and then increase with Feature\_dim, and there is a dim that can minimize loss.

But in the same way, because the epoch is less and the amount of data is less, the loss does not show a trend of first falling and then rising, shown in Figure 4

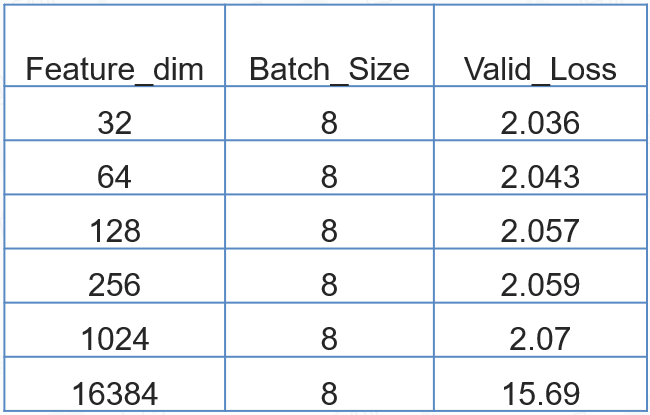


Table 2. Valid-Loss and Feature-dimension

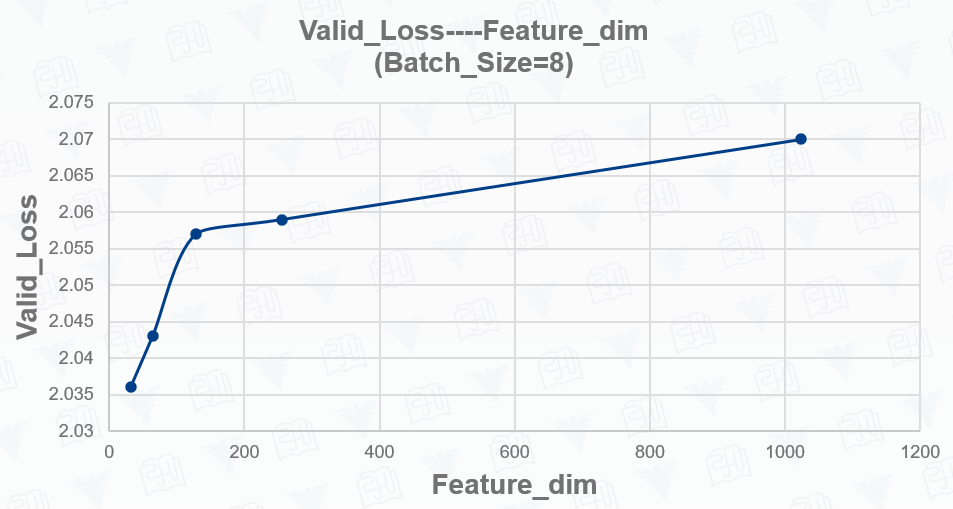


Figure 4. Valid-Loss and Feature-dimension

## **MLP or MEAN**

MEAN inputs the average coordinates of atoms in each amino acid into the model as amino acid coordinates: this method ignores the spatial relationship and interactions between different atoms. If the spatial arrangement between atoms is of high importance for describing the antibody antigen, then the average coordinates may not be informative enough.

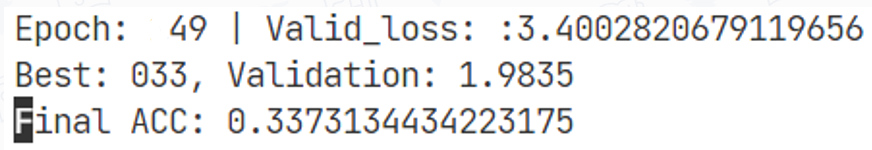


Figure 5. MEAN for coordinates

MLP outputs the position coordinates of 14 atoms in each amino acid as amino acid information: this method can make full use of the specific coordinates of the 14 atoms in each amino acid to describe the spatial structure of amino acids more comprehensively. By using MLP models, the spatial relationships and interactions between atoms can be captured, providing richer feature representations.

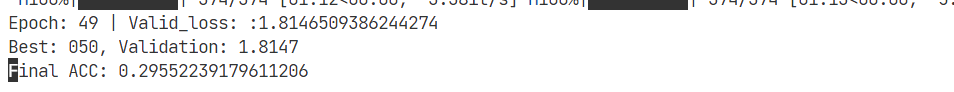


Figure 6. MLP for coordinates

Naturally, the way the MLP handles the positions of atoms performed better in the training of our model, shown in Table 3.



Table 3. MLP or MEAN

# **Highlights and Innovations**

## **topic selection**

Compared with the protein classification problem, we choose the more concrete and challenging problem of antibody-antigen binding. Since the abstract function of antibody is to specifically bind with an antigen, we hope to dig more deeply, thus yielding the topic for a more concrete antibody function: which antigen does it bind with. Although being exposed to Pytorch from zero exactly added difficulties to the problem, fortunately, we have finally done a good job from our perspective through some hard work.

## **New method**

We use the idea of Contrastive-Learning and add Zero-Shot to make combined predictions.

We adopt two encoders for antibody and antigen training. The parameters of the two encoders would be different, and the two networks can learn the difference between antibodies and antigens.

## **All Atom-Based**

Different from the general practice of only taking amino acid coordinates, we use the coordinate information of all atoms to average or MLP and make innovative improvements to its model

# **Challenges and Speculation**

Contrastive learning not only requires a large amount of data but hundreds of millions of data also requires a large amount of computing power.

This makes it difficult to visualize our training results, and the experimental data is not completely consistent with the theoretical analysis.

In addition, we use the information of the heavy chain and light chain for the antibody, and the binding site and surrounding information for the antigen. We speculate that whether the full sequence is used will have a certain impact on the prediction of the binding degree, which needs to be verified in the future.

# **Conclusion**

Due to the limited time and computational resources, our conclusion is the combination of hypothesis and incomplete experimental results.

In brief, there are 3 conclusions:

1. MLP will leverage atom-based coordinates information in the field of antibody design.
2. Increasing batch size can solve the overfitting problem to some extent
3. Modifying the feature dimensions for antibody and antigen can influence the model theoretically.

In addition, we raise a new question: Contrastive learning needs millions of data, so it’s difficult for us to train the model and figure out whether the empirical knowledge that CDRH3 is the deterministic part of the antibody. Intuitively, we think that using only the paratope part for training will make the model behave better compared with using the full sequence. But the hypothesis still needs experiments to justify.

# **References**

1. Kong X, Huang W, Liu Y. End-to-End Full-Atom Antibody Design. *ICML*, pp. 1-21, 2023.
2. Fan, H., Wang, Z., Yang, Y., & Kankanhalli, M, Continuous-Discrete Convolution for Geometry-Sequence Modeling in Proteins. *The Eleventh International Conference on Learning Representations*, pp. 1-26, 2022.