Cross Attentive Antibody-Antigen Interaction Prediction with Multi-task Learning

Shion Honda [*1] **Kyohei Koyama** [*1] **Kamiya Kotaro** [*1]

Abstract

We achieved state-of-the-art performance in a paratope prediction task with a cross transformer encoder and a multi-task learning technique. The cross transformer encoder naturally takes paired data as an input so that it can incorporate antibody information with antigen information. Moreover, it outputs transformed paired data to be synergistically combined with the multi-task learning. Our model improved ROC AUC and MCC compared to one of the state-of-the-art models and should be considered as a highly advantageous method of paratope and epitope interaction prediction.

1. Introduction

1.1. Background

The production of antibodies by the human immune system serves a vital role in the body's response to pathogens and infection. The antibodies, Y-shaped proteins, act to counterpoise the spread of a foreign body by binding to the surface of a presented antigen, thereby labeling it for destruction by the immune system with high specificity (Sela-Culang et al., 2013; Nicholson, 2016). Antibody therapy is widespread, and the accurate prediction of the binding activity has enormous potential to speed up the design of vaccines and to assist in the control and spread of diseases during epidemics of novel pathogens (Kuroda et al., 2012; Norman et al., 2019; Liu et al., 2020). Binding occurs between an epitope, which can be any region of the pathogenic antigen, and a paratope, which specifically relates to six complementarity-determining regions (CDRs) of the host antibody.

Many computational methods have been developed for the discovery of binding sites or suitable CDR combinations for novel antigens (Binz et al., 2005; Liberis et al., 2018; Deac et al., 2019). Epitopes and paratopes are often modeled as short amino acid chains. However, an epitope structure in particular may be affected by its position within a longer antigenic protein sequence. Since local information has been found to be critical for protein chains' functionality prediction, many models have used convolutional neural networks (CNNs) to aggregate local amino acids to extract this information (Li et al., 2018; Zhang & Kabuka, 2020; Bileschi et al., 2019).

Although CNNs effectively aggregate amino acids local information, models that use the transformer architecture can additionally capture long-range dependencies and more global relationships (Vaswani et al., 2017; Devlin et al., 2018). Moreover, the transformer's query-key-value mechanism should allow the model to capture the dependency between antigens and antibodies, which we call cross attention. Cross attention can capture the binding sites of an antibody given an antigen or those of an antigen given an antibody. See section 1.3 for discussion of previously published methods.

We describe, for the first time, the incorporation of multi-task learning for binding prediction using this architecture. This method allows the transfer of loss, i.e. the prediction error, between learning of the epitope and learning of the paratope. In this way we achieve comparable performance to state-of-the-art models while improving ROC AUC and MCC for the standard test dataset.

1.2. Task Definition

The paratope and epitope prediction is regarded as a sequence-to-sequence transformation task from the machine learning perspective (Sutskever et al., 2014; Gehring et al., 2017). In this study, each amino acid residue of the two proteins is transformed into binary value. That means that our model's output is a prediction with only two classes: "binds" and "does not bind" for an input of antibody and antigen pair. The input is a pair of two sequences: an antigen sequence and an antibody sequence. The output is a pair of two corresponding binary sequences for the multi-task prediction. In this study, we compare our model with other single-task (paratope prediction only) models as they were not predicting both sequences.

[*]Equal contribution [1]SyntheticGestalt Ltd, London, United Kingdom. Correspondence to: Kyohei Koyama <k.koyama@syntheticgestalt.com>, Kotaro Kamiya <k.kamiya@syntheticgestalt.com>.

Presented at the ICML 2020 Workshop on Computational Biology (WCB).

1.3. Related Work

There are two representative works of paratope prediction which utilize a neural network-based approach: Parapred (Liberis et al., 2018) and AG-Fast-Parapred (Deac et al., 2019).

Parapred used convolutional layers, bidirectional recurrent neural networks, and linear layers to predict the binding site of an antibody without any antigen information.

AG-Fast-Parapred was an successive work of Parapred. This model introduced a cross-modal attention layer, which let the antibody attend the antigen. This model restricted the number of neighborhood antigen residues to 150 residues, which were then attended over by any antibody residue.

The success of these models suggests that amino acid sequences alone are often sufficient to make accurate predictions of binding. This is fortunate because structural information is not available in many cases. Notably, there are numerous pitfalls to using crystalized protein structures for drug discovery (Zheng et al., 2014), which can be partly avoided by restricting the input to sequences alone.

2. Contribution

Our main contributions are summarized as follows: 1) enabling simultaneous paratope and epitope prediction via multi-task learning and 2) proposing a cross transformer encoder for handling paired data and predicting both the paratope and epitope to be combined with multi-task learning.

1) Although there exist several studies predicting protein-protein interactions, to the best of our knowledge, our model is the first to predict binding sites on antibody-antigen pairs using multi-task learning.

2) To predict protein functionalities, it is necessary to utilize global information on an entire sequence to capture the long-range dependencies. Here, the attention mechanism is more suitable than using only convolutional layers. Additionally, in order to reflect the phenomenon that binding is strongly related to a pair of two amino acids, we introduce a cross transformer encoder which mingles two inputs and makes the paratope prediction dependent on the antigen sequence and vice versa.

3. Methodology

To predict binding sites, the two input protein sequences are converted to the binary values by feature extractor and final layer. The feature extractor is composed of three types of embedding and three types of encoders. The final layer is composed of a position-wise linear layer and the sigmoid function. Our model architecture is illustrated in Figure 1.

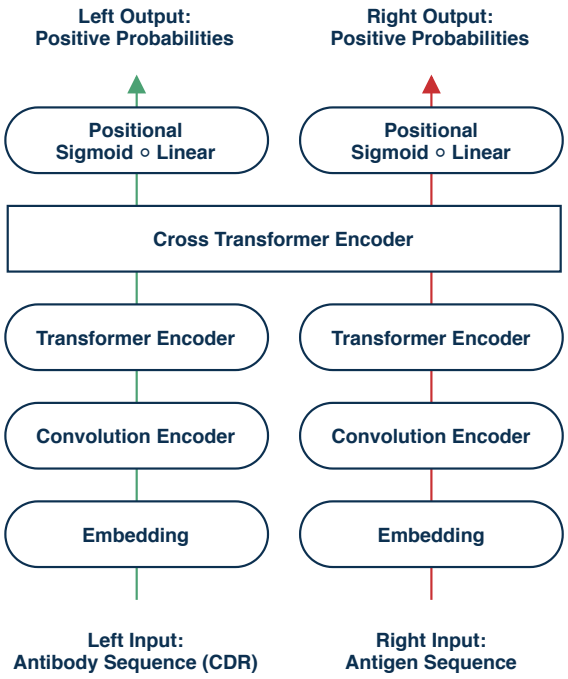

Figure 1. Entire model architecture.

3.1. Three Types of Embedding

By following the BERT (Devlin et al., 2018) approach, we summed up three types of embeddings in stead of one-hot vectors: 1) an amino acid token embedding, 2) an absolute positional embedding, and 3) a CDR number embedding.

The omission of the seven additional features representing physico-chemical properties from the feature candidates was performed because the properties used in pre-studies can be considered equivalent to the amino acids tokens when considered as categories. That is, there is a one-to-one relationship between a token and a property because both represent the residues, meaning there is no reduction in performance when amino acid properties are not explicitly considered.

3.2. Three Types of Encoders

We utilized three types of encoders: 1) the convolution encoder, 2) the transformer encoder and 3) the cross transformer encoder. The convolution encoder and the transformer encoder aim to aggregate self-information locally and globally, respectively. The cross transformer encoder utilizes a combination of self-information and counterpart-information.

The two transformer encoders (Vaswani et al., 2017; Devlin et al., 2018) separately take the two outputs from the con-

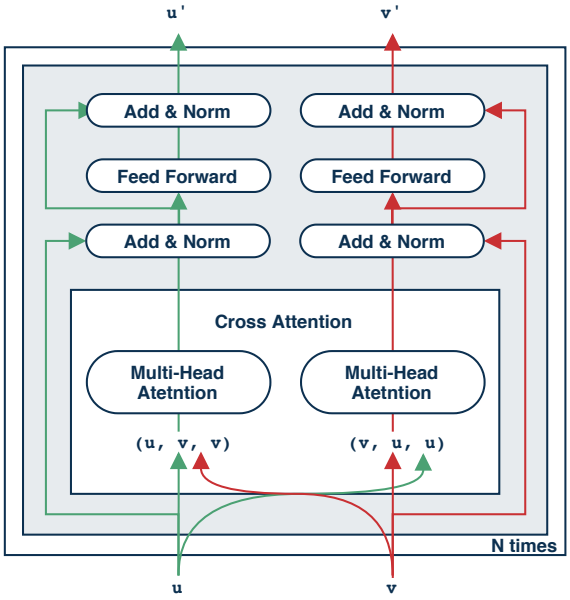

Figure 2. Cross transformer encoder. Norm indicates the layer normalization. Feed Forward indicates the position-wise linear layer.

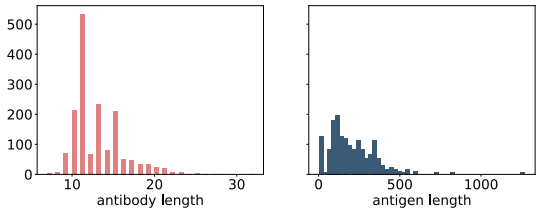

Figure 3. Histograms for antibody and antigen length.

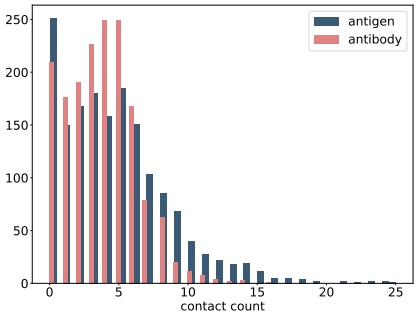

Figure 4. Histogram for the contact count of antibodies and antigens.

volutional layers. Then, the two resultants (u, v) are further encoded to (u', v') by the cross transformer encoder. We employed two combined multi-head attention layers, namely a cross attention layer, to realize a mutual reference with the following source-target manner: we set $(q, k, v) = (u, v, v)$ for a left input and $(q, k, v) = (v, u, u)$ for a right input (see Figure 2). The (u, v) and (u', v') have the same shapes so that the cross transformer encoder can be used as a composable part. In particular, this architecture should be powerful when the data are paired to make use of the attention mechanism on the both sides.

3.3. Multi-task Learning

We implemented multi-task learning by using two outputs from the model and a total loss $L = L_{\text{antibody}} + L_{\text{antigen}}$. Corresponding to the two outputs, these two base losses where calculated on the antibody side (L_{antibody}) and the antigen side (L_{antigen}), respectably. The base loss is the average of the binary cross entropy losses over the sequence of the positive probabilities.

3.4. Model Parameters

The embedding dimension for the model parameters is 128. The three convolutional layers have the different kernel sizes: 3, 15, and 31. Both the transformer encoder and the cross transformer encoder have a single layer with 16 heads. The final position-wise linear layer has 64 nodes. The dropout rate is 0.3. We use the RAdam optimizer (Liu et al., 2019) with a learning rate of 0.001.

4. Experiment

4.1. Dataset of Antibody-Antigen Pairs

We used the same dataset as the previous studies: Parapred (Liberis et al., 2018) and AG-Fast-Parapred (Deac et al., 2019). The dataset is a subset of the Structural Antibody Database (SAbDab) (Dunbar et al., 2014), which consists of crystal structures of antibody-antigen complexes. This dataset provides us with 1,662 antibody-antigen sequence pairs (277 antibody-antigen complexes, each with 6 CDRs with two extra residues on the both sides (Al-Lazikani et al., 1997)). Hence, each sequence is a sequence of amino acid residues (in the form ...YCQHHYGTP...). The model input is a pair comprising a CDR and an entire antigen. We also created binary sequences for both antibody and antigen based on the distance, i.e. whether or not each amino acid has at least one atom within 4.5 Å of any counterpart's atom, resulting in binary sequences for CDRs and entire antigens.

The average and median lengths of the antibody CDR sequences are 12.8 and 11.0, while those of the antigen sequences are 211 and 179, respectively (see Figure 3). The average and median count of the contacts in the antibody sequences are 3.65 and 4.00, respectively, while those in the

Table 1. Mean and standard deviation of the metrics for the **antibody** sequences of the test datasets. We used a threshold obtained by maximizing Youden's index to calculate MCC.

METHOD	ROC AUC	MCC
Ours with multi-task	**0.8881 ± 0.0019**	**0.6053 ± 0.0045**
Ours with single-task	0.8868 ± 0.0033	0.5996 ± 0.0091
AG-Fast-Parapred	0.8862 ± 0.0009	0.6045 ± 0.0053
Parapred	0.8791 ± 0.0026	0.5900 ± 0.0063

Table 2. Mean and standard deviation of the metrics for the **antigen** sequences of the test datasets.

METHOD	ROC AUC	MCC
Ours with multi-task	0.6359 ± 0.0092	0.1168 ± 0.0042

antigen sequences are 4.58 and 4.00 (see Figure 4). Notably, the dataset is imbalanced in terms of labels: the positive ratio is 0.284 in antibodies and 0.0217 in antigens.

4.2. Dataset Split

We randomly chose 330 complexes as a test dataset out of total 1,662 complexes. The remaining 1,332 interactions were split into subsets for 10-fold cross-validation. Each subset was used as a validation set and nine subsets were used as the training set. Thus, during test evaluation, we measured the mean and standard deviation of each metric with 10 trained models.

4.3. Metrics

We used ROC AUC and Matthews correlation coefficient (MCC) as evaluation metrics. Since every amino acid residue in a sequence is predicted as a binary value, the metrics are computed for each antibody/antigen sequence. Then, the final metrics are the average over the number of sequences. However, to compare our results with existing single-task models, we report the ROC AUC and MCC on the antigen during the experiment.

4.4. Baseline Models for Comparison

We implemented and compared four models. The first model (ours with multi-task) is our proposed architecture with the multi-task learning setting. The second one (ours with single-task) is also using our architecture, but only with the paratope prediction. The third and fourth models are the existing methods: AG-Fast-Parapred and Parapred.

4.5. Result and Discussion

The results of our study are summarized in Table 1. The best performance among the four models was achieved by our multi-task model. We also report the performance of the epitope prediction with our multi-task model (see Table 2).

By comparing the models with and without the multi-task learning, it is evident that using information from the both sides through multi-task learning is beneficial as intended. The one-sided Welch's t-test between our multi-task model and AG-Fast-Parapred showed the p-values of 0.0182 and 0.7059 for ROC AUC and MCC, respectively. Hence, the result for ROC AUC is statistically significant at the 5% significance level. Although no significant difference is observed over the MCC, our model showed a better score than the state-of the-art-model.

Even without the multi-task learning, the model's ROC AUC is better than that of AG-Fast-Parapred. This suggests that our cross transformer encoder model performs more accurately than the cross-modal attention layer of AG-Fast-Parapred. We consider that this is due to the use of the cross attention and the better abstraction of amino acid sequences with no explicit amino acid properties. By including features defined over multiple residues (e.g. chains, domains, or secondary structures), we believe it likely that the model can be improved further. Such improvements have been left to future work.

Comparing Table 2 and Table 1, the metrics of antigens in our best model are not as high as those of antibodies. The prediction of antigens is more difficult than antibodies due to differences in the positive and negative ratio between antibodies and antigens and the much longer sequence length of antigens.

5. Conclusion

In this paper, we have presented an advanced model for antibody-antigen interaction prediction. We leveraged a cross transformer encoder for a pair of two sequences together with multi-task learning which allows our model to globally capture the mutual relationship between the antibody and antigen. Our experimental results indicated that our model with multi-task learning outperformed an existing state-of-the-art method in terms of ROC AUC and MCC metrics. The model architecture outlined here is well suited to tasks using paired data and could be applicable in general protein-protein interaction prediction. Therefore, it offers a new angle of attack for research problems of this type.

References

Al-Lazikani, B., Lesk, A. M., and Chothia, C. Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology*, 273(4):927–948, 1997.

Bileschi, M. L., Belanger, D., Bryant, D., Sanderson, T., Carter, B., Sculley, D., DePristo, M. A., and Colwell, L. J. Using deep learning to annotate the protein universe. *bioRxiv*, 2019.

Binz, H. K., Amstutz, P., and Plückthun, A. Engineering novel binding proteins from nonimmunoglobulin domains. *Nature Biotechnology*, 23(10):1257–1268, 2005.

Deac, A., Veličković, P., and Sormanni, P. Attentive cross-modal paratope prediction. *Journal of Computational Biology*, 26(6):536–545, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M. Sabdab: the structural antibody database. *Nucleic Acids Research*, 42 (D1):D1140–D1146, 2014.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. Convolutional sequence to sequence learning. pp. 1243–1252, 2017.

Kuroda, D., Shirai, H., Jacobson, M. P., and Nakamura, H. Computer-aided antibody design. *Protein engineering, design & selection*, 25(10):507–522, 2012.

Li, Y., Wang, S., Umarov, R., Xie, B., Fan, M., Li, L., and Gao, X. Deepre: sequence-based enzyme ec number prediction by deep learning. *Bioinformatics*, 34(5):760–769, 2018.

Liberis, E., Veličković, P., Sormanni, P., Vendruscolo, M., and Liò, P. Parapred: antibody paratope prediction using convolutional and recurrent neural networks. *Bioinformatics*, 34(17):2944–2950, 2018.

Liu, C., Zhou, Q., Li, Y., Garner, L. V., Watkins, S. P., Carter, L. J., Smoot, J., Gregg, A. C., Daniels, A. D., Jervey, S., et al. Research and development on therapeutic agents and vaccines for covid-19 and related human coronavirus diseases. *ACS Cent. Sci*, pp. 315–331, 2020.

Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

Nicholson, L. B. The immune system. *Essays in Biochemistry*, 60(3):275–301, 2016.

Norman, R. A., Ambrosetti, F., Bonvin, A. M. J. J., Colwell, L. J., Kelm, S., Kumar, S., and Krawczyk, K. Computational approaches to therapeutic antibody design: established methods and emerging trends. *Briefings in Bioinformatics*, 2019.

Sela-Culang, I., Kunik, V., and Ofran, Y. The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, 4:302, 2013.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, pp. 3104–3112, 2014.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Zhang, D. and Kabuka, M. Protein family classification from scratch: A CNN based deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2020.

Zheng, H., Hou, J., Zimmerman, M. D., Wlodawer, A., and Minor, W. The future of crystallography in drug discovery. *Expert Opinion On Drug Discovery*, 9(2):125–137, 2014.