PESI: Paratope-Epitope Set Interaction for SARS-CoV-2 Neutralization Prediction

Zhang Wan^{a}, Zhuoyi Lin^{b,†}, Shamima Rashid^{a,†}, Shaun Yue-Hao Ng^{c}, Rui Yin^{d}, J. Senthilnath^{b}, Chee-Keong Kwoh^{a,*}
^{a} School of Computer Science and Engineering, Nanyang Technological University, Singapore
^{b} Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore
^{c} School of Chemical and Biomolecular Engineering, Nanyang Technological University, Singapore
^{d} Department of Health Outcome and Biomedical Informatics, University of Florida, Gainesville, FL, USA
{zwan004,sng101}@e.ntu.edu.sg, zhuoyi.lin@outlook.com, {bshamima,asckkwoh}@ntu.edu.sg,
j_senthilnath@i2r.a-star.edu.sg, ruiyin@ufl.edu

***Abstract*—Prediction of neutralization antibodies is important for the development of effective vaccines and antibody-based therapeutics. Traditional methods rely on features based on first principles derived from the binding interface. However, they are burdened by arduous data preprocessing from a limited quantity of protein structures. In comparison, deep learning allows automatic substructure characterization and representation without hand-crafted feature engineering. In particular, large language models (LLMs) based method predicts neutralization using Fv sequences of antibody and antigen. Despite LLM's success, incorporating full-length Fv sequences suffers from: 1) inaccurate sequence-level labels in existing datasets, 2) inefficient modeling due to noisy non-contributing motifs, and 3) ignorance of non-bonded interactions that play a key role in facilitating epitope-paratope pairing. In this paper, we propose a novel approach that incorporates only the paratope and epitope for antibody-antigen neutralization prediction while adopting a novel set modeling that regards the paratope and epitope as bags of residues. Specifically, we hand-crafted a dataset containing neutralizing paratope-epitope pairs where epitopes are potentially generalizable to future unseen variants of SARS-CoV-2. Training on such a dataset enables deep learning models to predict neutralizing antibodies for prospective mutated variants of SARS-CoV-2, meanwhile addressing the problem of inaccurate sequence-level labels. A higher modeling efficiency is also achieved by disregarding non-contributing motifs. Furthermore, we also propose paratope-epitope set interaction (PESI), a set modeling model inspired by first principles that learns intra-inter non-covalent interactions through a global attention mechanism. To validate PESI, we perform a 10-fold cross-validation on our dataset. Experimental results show that PESI achieves a more balanced overall performance and a significant improvement on MCC as compared to existing architectures.**

***Index Terms*—SARS-CoV-2, paratope-epitope interaction, set modeling, deep learning, Transformer**

I. Introduction

Antibodies are a class of proteins that the immune system produces to fight invading pathogens [14], [17], [28]. The Fv region is the variable domain of antibody sequences. It accommodates paratope, the binding surface of the antibody, which primarily comprises six variable loops called complementarity-determining region (CDR). The CDR loops exhibit significant diversity, enabling antibodies to bind a wide range of targets effectively. Neutralizing antibodies could recognize and attach to intrusive viruses, and effectively prevent viral infection [7], [33], [34]. Compared to molecular drugs, antibody-based therapy presents fewer adverse effects, thus they have been used as the primary category of newly developed medications in recent years [26]. As a result, the identification of antibody-antigen neutralization is critical in the development of antibody therapies. However, it remains to be a major challenge in both immunology and antibody-based therapies [1]. To predict antibody-antigen neutralization, numerous non-covalent interactions (NCI) such as hydrogen bonds, van der Waals' forces, and electrostatic attractions between antigen and amino acid residues present at the binding site [18], [42], remain the fundamental factor. While van der Walls' forces are short-range, electrostatic interactions are long-range, which makes it more complicated to model antibody-antigen interactions.

Traditional methods focus on the interface of the antibody-antigen complex, and hand-made characteristics, such as area-based features [30], [36], [48] and interaction-based features [4], [8], [47] dependent on structural information, are the main basis of decision-making. However, such first-principle-induced methods are burdened by laborious data pre-processing such as the extraction of amino acid physico-chemical and polypeptide geometric properties from a limited quantity of structural data. As such, a bottleneck is observed due to the limited quality of structural data and the need for hand-engineering specific structure-derived features at the residue or atomic level, involving domain expertise. Besides, the intricate features are mostly based on interactions between paratope and epitope, i.e., the respective binding sites on antibody and antigen that form antibody-antigen interface, which further intensifies cumbersome pre-processing steps.

Consequently, deep learning methods are proposed to learn protein substructures or representations from existing massive

† These authors contributed equally
* To whom correspondence should be addressed

979-8-3503-3748-8/23/$31.00 ©2023 IEEE

Fv sequence data, without the need for hand-crafted feature engineering. One prominent example would be large language models (LLMs). As such, recent studies focus on pre-trained antibody language models (PALMs) [45], because language modeling in deep learning naturally aligns with modeling protein sequences since amino acid representations of proteins share the same symbols as natural language. This allows unified methodologies in tokenization and representation in language and protein modeling. Besides, scaling language models with massive amino acid sequences also derives structural knowledge based on merely sequence inputs [3], [11], [39], which also demonstrates its capacity in modeling protein sequences. For this reason, multiple PALMs [23], [32], [45] have been proposed to model sequence-related antibody tasks. As opposed to traditional methods, PALMs utilize massive antibody sequences in existing databases to learn residue imputation through masking and predicting a subpart of the sequence. This context understanding allows PALMs to capture the relations between residues and encode them into embeddings for representations. The foundation model trained on general and massive datasets is then transferred to downstream tasks such as paratope prediction [23], [45].

However, the inaccurate sequence-level label in existing datasets such as OAS [31] curtails the credibility of the trained LLMs [45], because virus binding can only be attributed to a small fraction of antibodies from SARS-CoV-2 patients. Therefore, it is difficult to obtain trustworthy machine learning (ML) models at sequence-level due to the undependable individual-level labels with noise. In addition, since only a small portion of residues (those on CDR) in the Fv region is involved in binding interaction, the noisy non-contributing residues may bring up spurious features and impair the generalization capacity of predictive models. This indicates that ML model trained on full-length Fv sequences could make predictions based on non-contributing residues. Although sometimes correct decisions can be achieved, the correlations between labels and extracted features are not causally related. Instead, the prediction should be ideally driven by features of residues involved in paratope-epitope binding (termed as contributing motifs). Furthermore, the LLMs are also incapable of capturing antibody-antigen interactions which is the key factor that directly determines binding. In [45], a sequence-level classifier to predict SARS-CoV-2 antibody is trained based on noisy individual-level labels. Highly-ranked sequences are compared with truly binding sequences in Coronavirus Antibody Database (CoV-AbDab) [37], where sequences above a pre-set identity threshold are accepted as binding antibody. This solution suffers from: 1) compositional similarity between sequences doesn't guarantee functional similarity, and 2) interactions between antibody and antigen are ignored since this is the critical element for antibody-antigen binding.

In this paper, we propose a new approach for the prediction of antibody-antigen neutralization of SARS-CoV-2. Inspired by recent advances in identifying SARS-CoV-2 epitopes which can be potentially extended to predict epitopes of other viruses [43], we propose to tackle the aforementioned problems by introducing a new dataset and a novel model. Following methodologies in [43], a dataset containing paratope-epitope pairs that are potentially generalizable to unseen mutated variants of SARS-CoV-2 is curated based on CoV-AbDab dataset [37]. The handcrafted dataset provides accurate labels as compared to noisy sequence-level labels in OAS database [31]. We directly adopt paratope-epitope pairs as input and thus eliminate noise of non-contributing motifs, achieving higher modeling efficiency. Furthermore, we notice the difference between predicting neutralization based on antibody-antigen sequences and the paratope-epitope pairs, and propose a set modeling approach with the paratope-epitope set interaction (PESI). While traditional methods focus on local regional features and sequential/positional information as extracted by Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), set modeling regards paratope and epitope as two collections (or sets) of residues with intra-inter interactions between set elements (as shown in Figure 1). Based on permutation invariant modules proposed in [22], PESI can capture non-covalent interactions through considering relations between any two residues with global attention. To further capture paratope-epitope interactions, we also introduce a co-attention module [25] to weight bilateral residue representations. Compared with the previous state-of-the-art architectures, PESI achieves a more balanced performance and greatly improves performance on MCC that is more rigorous [10]. In order to scale up PESI with massive data, we introduce masked pre-training to learn prior knowledge on paired heavy-light antibody sequences from the OAS database [31]. Pre-trained PESI achieves better performance than its non-pre-trained counterpart on our newly proposed dataset. Our ablation study in Fig. II demonstrates the contributions of set modeling and co-attention. While pre-trained PESI has a balanced performance among all metrics, a degradation is observed when replacing set modeling with traditional methods or deleting co-attention.

Our main contributions are summarized as follows:

- We propose a novel approach for antibody-antigen neutralization prediction that can: 1) benefit from accurate labels to potentially provide generalizable predictive results based on our newly proposed dataset, 2) improve modeling efficiency by disregarding non-contributing residues, 3) capture inter-intra non-bonded interactions through set modeling with attention mechanism.
- We propose to inject reasonable inductive bias using set modeling. A Set Transformer based model, PESI, is proposed to capture residue interactions. A co-attention module is also introduced to extract features for paratope-epitope interaction.
- We propose a masking-based pre-training technique that enables PESI to achieve greater performance by learning imputations on massive Fv sequence data in a self-supervised manner.

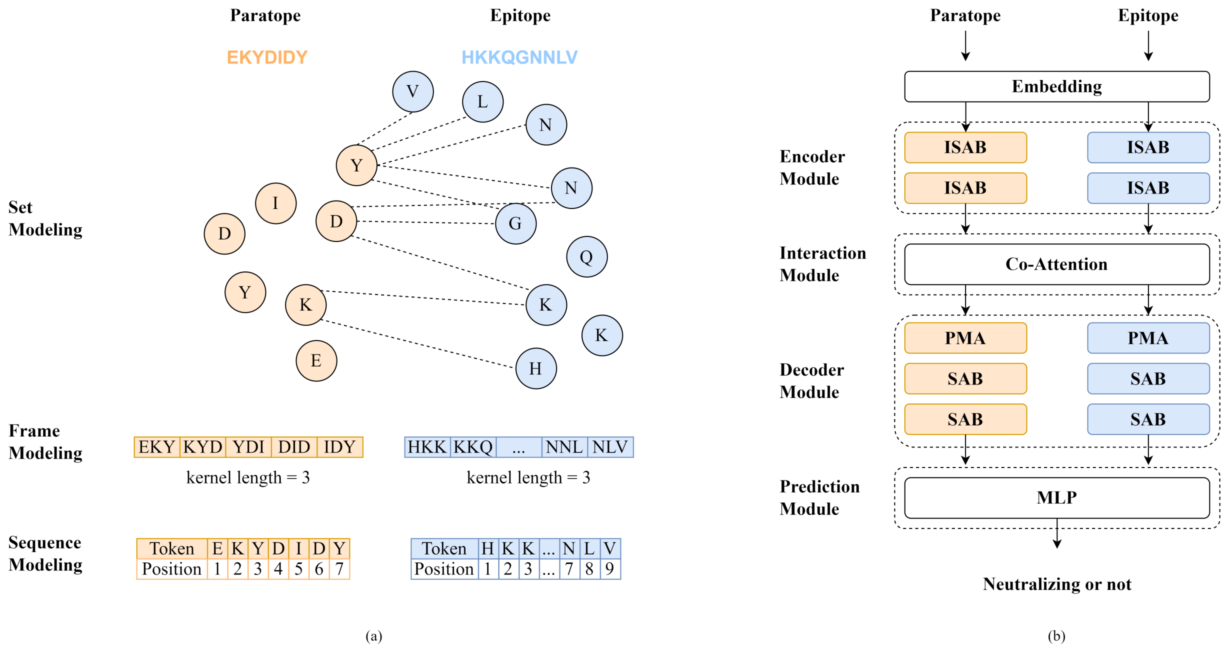

Fig. 1. Modeling and Network Architecture of Paratope-Epitope Set Interaction (PESI). **(a)** Set modeling approach in comparison with traditional frame modeling and sequence modeling. Set modeling regards paratope and epitope as two collections of residue sets, and considers non-bonded interactions as inter-intra interactions between set elements. **(b)** The Network Architecture of PESI. A pseudo-Siamese encoder-decoder architecture is designed. Permutation invariant modules proposed in [22] capture intra-relations within two residue sets. A co-attention module is proposed to capture inter-relations for paratope-epitope interactions. ISAB, Co-Attention, PMA, SAB are proposed attention modules to constitute PESI, which we shall introduce in Section II-D.

II. Materials and Methods

A. Paratope-Epitope Interactions

The prediction of antibody-antigen binding is pivotal in immunology. Antibody specificity is regulated by a repertoire of paratope-epitope non-covalent interactions, forming a motif vocabulary that plays a key role in this process [1]. In [1], a group of these motifs, which form a commonly shared vocabulary of paratope-epitope interactions, has been identified in a set of non-redundant antibody-antigen structures. The vocabulary is compact, consisting of fewer than 10^4 motifs, and different from non-immune protein-protein interactions. Furthermore, it also mediates oligo- and polyreactive interactions between paratope-epitope pairs.

Inspired by [1], we suppose that current studies employing the whole Fv region for antibody-antigen binding prediction may suffer from noisy non-interacting residues, which results in spurious correlation. Paratopes have frequently been observed to be situated mainly in the CDRs, although not completely limited to them [20], and therefore a common approach is to include the entire Fv region as input for the prediction of binding. However, this may introduce a spurious correlation as only a small part of amino acids, i.e., interaction motifs, contribute to binding specificity. As a result, ML models could make decisions based on background noise introduced by non-contributing residues rather than motif interaction features. Such erroneous associations can be severely observed particularly when the training set is small, as the ML model fails to capture the interaction motifs accurately. Therefore, we propose including only the paratopes and epitopes for the prediction of binding in order to purify the inputs, because the essential scheme of binding is dependent on interaction motifs.

Recent advances in understanding SARS-CoV-2 also motivate us to employ paratope-epitope based model for antibody-antigen interaction. In [43], the authors have identified nine epitope sites on the surface of the SARS-CoV-2 Spike protein (S protein) that are accessible to antibodies and unencumbered by the glycan shield. The study combines atomistic modeling, multi-microsecond molecular dynamics (MD) simulations, and various bioinformatics analysis methods to predict epitope sites. The methodology is validated through recovering five epitopes that overlap with experimentally characterized epitopes, and the identified epitopes could serve as starting points for developing stable immunogenic constructs that can elicit antibody production. It can be extended to predict epitopes for other viral proteins, with the goal of producing a vaccine that guarantees broad protection against multiple members of this virus family. Therefore, a manually curated paratope-epitope dataset was derived from the CoV-AbDab [37] using the delineated epitope sites of [43].

B. Dataset and Problem Formulation

Based on CoV-AbDab [37] (accessed on 14 June 2022), we filtered out 328 structures with 4-letter PDB codes in order to identify epitope-paratope pairs. Comprehensive epitope maps are constructed for the SARS-CoV-2 [21], [43], and nine epitope candidate sites were identified by Sikora et al., [43] by constructing a detailed map of the S protein using MD studies and glycan pattern characterization. Following the same numbering and delineated epitope sites as in [43], we mapped the epitope residues onto the full 3D structure of the SARS-CoV2 S protein (PDB code 7VXF) [46]. We align the epitope sites against the reference strain sequence hCoV - 19/Wuhan/W IV 04/2019 (WI04) provided by the GISAID database [38] with an accession number EPI_ISL_402124. (WI04). Table S1 in Supplementary Material [*] shows the epitope sites' residue positions and their corresponding sequence in WI04. We curated our dataset based on the visualization of antibody-antigen complexes using PyMOL [40]. We directly concatenate the CDR fragments for the paratope residues and the nine epitope sites for the epitope residues In total, 310 paratope-epitope pairs are obtained, among which 228 sequence pairs are neutralizing (labeled as 1) while the remaining 82 samples are non-neutralizing (labeled as 0).

Our approach of delineating epitope-paratope pairs is a merger of linear and conformational epitopes and can be considered as interpolated epitope and paratope pairs, respectively. An advantage of our approach is that it detects neutralization from epitope or paratope sequences alone. Thereby, it could be suitable for practical applications in rapid, high-throughput in-silico identification from a large list of candidate epitopes or paratopes, such as based on motif-scans of antibody repertoires in the initial stages. It can be complementary with later pipelines of extensive modeling to assess antibody binding affinity such as in those required by scaffolding onto framework regions, structure re-construction, etc.

Neutralizing antibody prediction is formulated as a binary classification in our proposed dataset containing paratope-epitope pairs $\{p^{(i)}, e^{(i)}, y^{(i)}\}_{i=1}^{N}$ where $p^{(i)}$ and $e^{(i)}$ indicate the $i-th$ paired paratope and epitope residue strings, respectively. The embedding layer projects the two sets of residues to $R_{para}, R_{epi} \in \mathbb{R}^{N \times L \times d}$, where N is the number of samples, L is the maximum number of residues, and d is the dimension of the representation of the residues. We aim at predicting the label $Y = \{y^{(i)}\}_{i=1}^{N}$ where $y^{(i)}$ indicates whether the paratope $p^{(i)}$ can neutralize epitope $e^{(i)}$.

C. Set Modeling

Paratope and epitope inputs are strings of amino acids. Traditional methods usually adopt two intuitive approaches for string feature extraction which we term as frame modeling and sequence modeling. Frame modeling treats a string as numerous overlapped windows (or frames) where each window contains several contiguous tokens. An example is the n-gram model [6] that separates a sentence by overlapping its n-length sub-parts [†]. In comparison, sequence modeling regards a string as a permutation of tokens and takes the order and position into account. We consider RNN-based or Transformer-based models as two different instances of sequence modeling.

The set of assumptions that ML models make to deal with unseen inputs in order to predict outputs are inductive bias. Setting different inductive biases for different scenarios has spawned a variety of ML models. Although frame modeling and sequence modeling correspond to two different perspectives of inductive bias towards sequences, we argue that both modeling approaches are biased to be applied to the paratope-epitope dataset.

For frame modeling, it is reasonable to extract features on local and contiguous amino acid frames. The assumption made is that neighboring amino acids may together form interaction motifs that provide strong binding specificity. CNNs are often utilized as a result. However, it only considers local relations through convolutional kernels while omits global information. Frame modeling focuses on string-based local features but neglects structural information. An example is two sequentially distant but spatially neighboring residues synergistically contribute to specificity, but no correlation feature is calculated based on frame modeling.

For sequence modeling, it is intuitive to accept sequential tokens as inputs in a recurrent node and extract position features. However, we notice that this may suffer from spurious correlation especially when data is limited, because the order of residue segments is not the determining factor for paratope-epitope binding. Therefore, although the learned patterns for relative orders of residues can give out correct predictions, it doesn't make decisions (purely) based on capturing correct interaction motif features as no mechanism forces it to do so. Thus, it's limited by insufficient inductive bias.

Therefore, we propose set modeling as a more natural fit where limited inductive biases are disregarded. Specifically, paratope and epitope are modeled as two collections of residues with inter-intra interactions. The binding specificity is determined by interaction motifs composed of residues as set elements. To learn residue interactions, the global relation feature is first captured within the two sets. Then, the interactions between paratope and epitope are explicitly considered by focusing on every pair of elements in two sets. In this way, both intra-peptide relations (within epitope or paratope, respectively) and inter-peptide relations (i.e. between epitope and paratope) are attended. Besides, spurious correlation is less likely to dominate decision basis, because ML model is forced to predict neutralization by learning the interaction features from motifs located in epitope-paratope peptide pairs alone.

D. PESI

In this section, we introduce the modules employed to constitute PESI whose overall architecture is shown in Figure 1. We start from introducing the set encoding modules proposed in [22], i.e., ISAB, PMA, and SAB. Then, we further introduce the co-attention module [25] to facilitate paratope-epitope interactions.

Attention: We first introduce dot product attention and Multi-Head Attention (MHA) proposed in [44]. Given input queries $Q \in \mathbb{R}^{n \times d_q}$, keys $K \in \mathbb{R}^{n \times d_k}$, and values $V \in \mathbb{R}^{n \times d_v}$ where n is number of tokens and d_q, d_k are the dimension of keys and values, the Attention and Multi-Head Attention are defined as in Equation 1 and 2, respectively,

[*] https://github.com/ZhangWan-byte/PESI

[†] For instance, "AGCTG" is $[AGC, GCT, CTG]$ in 3-gram.

TABLE I
PERFORMANCE OF PESI AND BASELINES

Model	Acc	F1	ROC-AUC	G-mean	MCC
BiLSTM	0.7301 ± 0.0106	0.8382 ± 0.0080	0.5054 ± 0.0565	0.0294 ± 0.0416	-0.0015 ± 0.0347
TextCNN	0.7269 ± 0.0076	0.8379 ± 0.0044	0.5889 ± 0.0522	0.0136 ± 0.0192	0.0011 ± 0.0277
MasonsCNN	0.7323 ± 0.0046	**0.8410 ± 0.0026**	0.5693 ± 0.0339	0.0000 ± 0.0000	-0.0046 ± 0.0065
Ag-Fast-Parapred	0.7204 ± 0.0445	0.8190 ± 0.0329	0.6259 ± 0.0679	**0.2824 ± 0.1304**	0.0815 ± 0.1490
PIPR	**0.7355 ± 0.0000**	**0.8414 ± 0.0020**	**0.6962 ± 0.0587**	0.0140 ± 0.0198	0.0016 ± 0.0022
ResPPI	0.7258 ± 0.0240	0.8174 ± 0.0315	0.6082 ± 0.0695	0.1562 ± 0.1629	0.0621 ± 0.1353
PESI	0.7172 ± 0.0320	0.8246 ± 0.0230	0.6276 ± 0.0693	**0.2347 ± 0.1931**	**0.1166 ± 0.1027**
PESI+	**0.7419 ± 0.0174**	0.8404 ± 0.0117	**0.6314 ± 0.0692**	0.2154 ± 0.1200	**0.1244 ± 0.0946**

Note: PESI+ is mask pre-trained PESI on OAS.

$$Attn(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}}V) \tag{1}$$

$$MHA(Q,K,V) = Concat(h_1, h_2, ..., h_n)W^O \tag{2}$$

where $h_i = Attn(QW_i^Q, KW_i^K, VW_i^V)$ and W_i^Q, W_i^K, W_i^V are learnable matrices projecting original inputs to another vector space for stronger representation capacity.

MAB: Given input matrices $X, Y \in \mathbb{R}^{n\times d}$ of two d-dimensional vector sets, the Multi-Head Block (MAB) is defined in Equation 3.

$$MAB(X,Y) = LayerNorm(H + rFF(H)) \tag{3}$$

where $H = LayerNorm(X + MHA(X,Y,Y))$. The rFF is a row-wise feed-forward neural network that operates independently and identically on each instance. The $LayerNorm$ is layer normalization [2].

SAB: Based on MAB, the Set Attention Block (SAB) is defined as in Equation 4.

$$SAB(X) = MAB(X,X) \tag{4}$$

ISAB: To reduce quadratic computational complexity, an Induced Set Attention Block (ISAB) is proposed, as defined in Equation 5. An m-dimensional vector $I \in \mathbb{R}^{m\times d}$ is defined as a learnable parameter termed as inducing points.

$$ISAB_m(X) = MAB(X,H) \in \mathbb{R}^{n\times d} \tag{5}$$

where $H = MAB(I,X) \in \mathbb{R}^{m\times d}$

PMA: Set modeling requires permutation invariance. It has been proven for MAB, SAB, and ISAB to fit this property in [22]. However, a remaining issue is reducing set representation for final predictions, which incurs the necessity for a pooling function. Suppose $Z \in \mathbb{R}^{n\times k}$ is the output matrix of the set feature. Based on Multi-Head Attention, a learnable seed center matrix $S \in \mathbb{R}^{k\times d}$ with k seeds is proposed in [22]. The Pooling by Multi-Head Attention (PMA) is defined in Equation 6. The output of PMA contains k elements and therefore enables it to extract internal interactions.

$$PMA_k(Z) = MAB(S, rFF(Z)) \tag{6}$$

Co-Attention: The interactions between contributing motifs are utterly important for the prediction of paratope-epitope binding. As such, we introduce a co-attention module proposed in [25] to weigh two sets of feature vectors extracted from the paratope and epitope. Specifically, given two feature vectors $X = \{x_1, x_2, ..., x_m\}$ and $Y = \{y_1, y_2, ..., y_n\}$, where m and n are the number of residues in the paratope and epitope, respectively. Then, the residue-level scores between every residue pair on paratope and epitope, respectively, are computed with the co-attention module, as defined in Equation 7.

$$s_{ij} = f(x_i)Mf(y_j) \tag{7}$$

where the scoring matrix $S = [s_{ij}]$ indicates how well each residue in paratope and epitope interact with each other. $M \in \mathbb{R}^{d\times d}$ is the learnable transformation matrix. $x_i, y_j \in R^d$ are representations of $i-th$ residue of paratope and $j-th$ residue or epitope, respectively. $f(.)$ is an MLP layer to project the original residue representation into another space for transformation. Each row/column of matrix S indicates the matching score between paratope/epitope residue and all epitope/paratope residues, the weighted representations of paratope/epitope representations can be computed by the matching strength, as defined in Equation 8 and 9.

$$x_i = \frac{1}{n}\sum_j s_{ij} * x_i \tag{8}$$

$$y_j = \frac{1}{m}\sum_i s_{ij} * y_j \tag{9}$$

E. Pre-training

To further utilize existing data, we further propose to introduce a mask pre-training approach to improve PESI performance. The Observed Antibody Space (OAS) [31] collects antibody sequences from over 80 different studies with over 1 billion sequences. Although it contains noisy sequence-level labels, we contend that PESI can benefit from learning prior composition set elements for antibody. As such, we curated a dataset containing 1,572,406 full-length sequences where antibody heavy and light chains are natively paired from OAS on 15 June 2023.

The target is to learn residue imputation based on massive sequences, behind which the intuition of context learning allows to capture the statistical patterns and dependencies in order to encode residue into continuous vector representations. Specifically, we randomly mask 15% of the residues [13], among which 80% of them are replaced by the mask token (such as `[MASK]`), 10% of them are randomly assigned with a random token, and the rest 10% remain unchanged. The training objective is to predict the masked residue given the 75% original residues. A random 10% of OAS paired sequence dataset is split as a test set. The best model on the test set is chosen for next-stage fine-tuning where the pre-trained weights are utilized as initialization for k-fold cross-validation on our hand-crafted CoV-AbDab dataset. For PESI, we initialize weights of the embedding layer, paratope encoder, paratope decoder and co-attention module using pre-trained weights, and use random initialization for other modules.

F. Baseline Methods

BiLSTM [16], [41] and TextCNN [19] are representative baselines for sequence and frame modeling. MasonsCNN [29] is a simple yet experimentally proven architecture in PPI scenarios through web lab experiments. Ag-Fast-Parapred [12] is proposed to predict binding sites based on antibody-antigen pairs. PIPR [9] uses a Siamese architecture with a deep residual recurrent convolutional neural network to capture local features and contextualized information that is important for understanding how protein sequences interact with each other. ResPPI [27] employs ResNet [15] for feature extraction.

G. Evaluation Metrics

We compare predictive performance using accuracy, F1 score, area under the receiver operating characteristic (ROC-AUC), geometric mean (G-mean) and Matthews correlation coefficient (MCC), implemented by Scikit-learn [35] and Imbalanced-learn [24]. Concerning the imbalanced nature of our proposed dataset, we highlight the importance of MCC [10] as imbalance measurements which are defined below:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)\cdot(TP+FN)\cdot(TN+FP)\cdot(TN+FN)}} \tag{10}$$

The MCC ranges between -1 and +1, where +1 indicates a perfect classifier and -1 indicates a completely incorrect classifier. It generates a high score only when the prediction yields ideal outcomes across all four entries of the confusion matrix, i.e., True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), and is therefore robust to class imbalances. MCC produces a more informative and truthful score in evaluating binary classifications than the precision and F1 score [10], which has been widely used as an evaluation metric in bioinformatics [5]. Therefore, we give preference to MCC over other metrics.

III. Results

A. Performance Analysis

We conduct 10-fold cross-validation three times on our proposed dataset. The mean and standard deviation of the metrics are reported in Table I where the highest scores are highlighted in bold red and the second highest are bolded. [‡] Pre-trained PESI achieves the highest scores on accuracy and MCC, and reaches the second highest score on ROC-AUC, indicating a more balanced performance as compared to baselines. PESI without pre-training achieves the second highest G-mean and MCC As compared to PESI, all the other baselines suffer from lower MCC scores. This is due to the lower learning efficiency of imperfect inductive bias in baselines which can be especially impactful when training data is quantitatively limited. The highest MCC holder aside from PESI is Ag-Fast-Parapred which has also shown the best result in terms of G-mean. We believe it benefits from the dilated convolution for receptive field expansion and the specifically designed attention mechanism for information fusion. Besides, PIPR is another conspicuous architecture that achieves the highest F1 score and ROC-AUC while maintaining the second place in accuracy. We believe PIPR has a relatively reasonable inductive bias due to its Recurrent CNN (RCNN) module, as it first extracts low-level frame features with convolutional kernels and then considers positional relations between extracted regional features through RNN. Since our paratope and epitope are concatenated by several segments of contiguous amino acids, it is reasonable to consider higher-level positional information of regional features. Although both Ag-Fast-Parapred and PIPR achieved high MCC scores, they are great examples of employing reasonable inductive bias based on the observations towards task-specific data.

B. Ablation Study

We further conducted an ablation study to evaluate our proposed setting by comparing it with its counterpart that utilizes a full-length Fv region. Samples without original Fv sequences are deleted. Leaving all hyper-parameters unchanged, we further conduct three times of 10-fold cross-validation whose results are shown in Table II. The best scores in the traditional sequence setting and our proposed paratope-epitope setting are bolded, respectively. Ag-Fast-Parapred also achieves the highest MCC and G-mean in sequence. However, its accuracy and F1 score are prominently lower than others, which we attribute to the class imbalance of the dataset. In contrast, all the other sequence setting models show good results in these two metrics but low performance measured by G-mean and MCC, we think they are all affected by the class imbalance issue. In the paratope-epitope setting, the best model is pre-trained PESI which achieves the highest scores in all metrics except for G-mean where the best model is non-pre-trained PESI.

Comparing the best scores in the two settings, we find that scores in the sequence setting are all higher than our setting, aside from MCC. However, there is no single model that achieves high scores in all metrics. With Set modeling on paratope-epitope, PESI achieves relatively front-ranking and balanced performance, particularly on G-mean and MCC. Besides, most sequence setting models have much lower G-mean and MCC scores than PESI in the proposed setting.

[‡] Codes, dataset and model configuration files are publicly available at https://github.com/ZhangWan-byte/PESI.

TABLE II

Ablation Study on sequence setting and co-attention.

Setting	Model	Acc	F1	ROC-AUC	G-mean	MCC
Sequence	BiLSTM	0.7462 ± 0.0000	0.8516 ± 0.0000	0.4659 ± 0.0804	0.0000 ± 0.0000	0.0000 ± 0.0000
	TextCNN	**0.7474 ± 0.0018**	**0.8523 ± 0.0009**	0.6992 ± 0.0751	0.0118 ± 0.0167	0.0100 ± 0.0141
	MasonsCNN	0.7462 ± 0.0000	0.8516 ± 0.0000	0.6661 ± 0.0570	0.0000 ± 0.0000	0.0000 ± 0.0000
	Ag-Fast-Parapred	0.6846 ± 0.0669	0.7574 ± 0.0921	0.6506 ± 0.0629	**0.3333 ± 0.2290**	**0.1190 ± 0.1593**
	PIPR	0.7462 ± 0.0000	0.8494 ± 0.0022	**0.7791 ± 0.0596**	0.0240 ± 0.0183	0.0022 ± 0.0017
	ResPPI	0.7449 ± 0.0091	0.8502 ± 0.0057	0.6906 ± 0.0567	0.0262 ± 0.0371	0.0123 ± 0.0481
Para-Epi	PESI w/o CA	0.7183 ± 0.0360	0.8260 ± 0.0227	0.6143 ± 0.0807	0.2308 ± 0.1307	0.0928 ± 0.1200
	PESI+ w/o CA	0.7183 ± 0.0126	0.8307 ± 0.0095	0.5953 ± 0.0536	0.1049 ± 0.0764	0.0316 ± 0.0379
	PESI	0.7172 ± 0.0320	0.8246 ± 0.0230	0.6276 ± 0.0693	**0.2347 ± 0.1931**	0.1166 ± 0.1027
	PESI+	**0.7419 ± 0.0174**	**0.8404 ± 0.0117**	**0.6314 ± 0.0692**	0.2154 ± 0.1200	**0.1244 ± 0.0946**

Note: Para-Epi indicates set modeling on paratope-epitope pairs.

As such, we contend that traditional modeling approaches in traditional data organization are prone to suffer from the class imbalance problem. Such an issue can be alleviated by including more reasonable inductive bias as in Ag-Fast-Parapred and PESI.

Furthermore, we compare PESI and its counterpart without co-attention module ("w/o CA" in Figure II) where the drop in scores is observed. In particular, PESI has higher ROC-AUC, G-mean and MCC than PESI w/o CA, and vice versa for accuracy and F1 score. We reckon co-attention learning paratope-epitope interaction shows the importance of dealing with the class imbalance problem. However, PESI+ surpasses PESI+ w/o CA on all metrics, and PESI w/o CA outperforms PESI+ w/o CA, which indicates the significant role of co-attention during both transfer learning and pre-training, since it is intuitive that paratope and epitope information should not be learned independently but jointly, and that learning solid features on the paratope side can help learn its corresponding features on epitope side [§].

IV. Conclusion

In this paper, we propose a novel approach for neutralization antibody prediction. A dataset containing paratope-epitope pairs where epitopes are potentially generalizable to unseen variants of SARS-CoV-2 is handcrafted based on the CoV-AbDab database [37] using the latest advance in epitope identification [21], [43], [49]. We further observe the limitations of existing modeling approaches and propose to model paratope and epitope as two collections of interactive residues, i.e., set modeling, as compared to previous frame modeling and sequence modeling. Based on set modeling, we design PESI, a Set Transformer [22] based pseudo-Siamese encoder-decoder model for neutralization prediction. A co-attention module is also proposed to capture residue inter-relations between paratope and epitope. We further introduce a mask pre-training strategy to boost model performance using massive paired sequences from OAS database [31]. The 10-fold cross-validation is conducted three times on the proposed dataset. Experimental results show that PESI achieves a more balanced performance compared to previous prevailing architectures.

V. Future Work

Although set modeling indicates a new direction to model paratope-epitope interactions, we anticipate designing models by further looking into more reasonable inductive bias. The co-attention module can be further studied to fully capture residue interactions, such as under the framework of graph neural networks which also demonstrates permutation invariance. New challenges may be encountered and tackled such as over-squashing and graph re-wiring. Furthermore, pre-training a pseudo-Siamese encoder-decoder model to learn integrated paratope-epitope relations, potentially from multi-source datasets, remains an open problem. We look forward to future endeavors for better pre-training techniques to deal with the heterogeneity in existing datasets.

Acknowledgment

This research is supported by the MOE Academic Research Fund Tier 2 (Grant No: MOE2019-T2-2-175) under the project ”Host-pathogen protein-protein interaction approaches for predicting virulence”.

References

[1] Rahmad Akbar, Philippe A Robert, Milena Pavlović, Jeliazko R Jeliazkov, Igor Snapkov, Andrei Slabodkin, Cédric R Weber, Lonneke Scheffer, Enkelejda Miho, Ingrid Hobæk Haff, et al. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Reports*, 34(11), 2021.

[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[3] Tristan Bepler and Bonnie Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

[4] Daria Beshnova, Yan Fang, Mingjian Du, Yehui Sun, Fenghe Du, Jianfeng Ye, Zhijian James Chen, and Bo Li. Computational approach for binding prediction of sars-cov-2 with neutralizing antibodies. *Computational and Structural Biotechnology Journal*, 20:2212–2222, 2022.

[5] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PloS one*, 12(6):e0177678, 2017.

[6] Andrei Z Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig. Syntactic clustering of the web. *Computer networks and ISDN systems*, 29(8-13):1157–1166, 1997.

[7] Dennis R Burton. Antibodies, viruses and vaccines. *Nature Reviews Immunology*, 2(9):706–713, 2002.

[§] We didn't initialize epitope encoder and decoder using pre-trained weights.

[8] Chen Chen, Veda Sheersh Boorla, Deepro Banerjee, Ratul Chowdhury, Victoria S Cavener, Ruth H Nissly, Abhinay Gontu, Nina R Boyle, Kurt Vandegrift, Meera Surendran Nair, et al. Computational prediction of the effect of amino acid changes on the binding affinity between sars-cov-2 spike rbd and human ace2. *Proceedings of the National Academy of Sciences*, 118(42):e2106480118, 2021.

[9] Muhao Chen, Chelsea J-T Ju, Guangyu Zhou, Xuelu Chen, Tianran Zhang, Kai-Wei Chang, Carlo Zaniolo, and Wei Wang. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14):i305–i314, 2019.

[10] Davide Chicco and Giuseppe Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

[11] Ratul Chowdhury, Nazim Bouatta, Surojit Biswas, Christina Floristean, Anant Kharkar, Koushik Roy, Charlotte Rochereau, Gustaf Ahdritz, Joanna Zhang, George M Church, et al. Single-sequence protein structure prediction using a language model and deep learning. *Nature Biotechnology*, 40(11):1617–1623, 2022.

[12] Andreea Deac, Petar VeliČković, and Pietro Sormanni. Attentive cross-modal paratope prediction. *Journal of Computational Biology*, 26(6):536–545, 2019.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

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

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[17] Alissa M Hummer, Brennan Abanades, and Charlotte M Deane. Advances in computational structure-based antibody design. *Current Opinion in Structural Biology*, 74:102379, 2022.

[18] Charles A Janeway Jr. Immunobiology the immune system in health and disease. *Artes Medicas*, 1997.

[19] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[20] Vered Kunik, Bjoern Peters, and Yanay Ofran. Structural consensus among antibodies defines the antigen binding site. *PLoS computational biology*, 8(2):e1002388, 2012.

[21] Jason T Ladner, Sierra N Henson, Annalee S Boyle, Anna L Engelbrektson, Zane W Fink, Fatima Rahee, Jonathan D'ambrozio, Kurt E Schaecher, Mars Stone, Wenjuan Dong, et al. Epitope-resolved profiling of the sars-cov-2 antibody response identifies cross-reactivity with endemic human coronaviruses. *Cell Reports Medicine*, 2(1), 2021.

[22] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019.

[23] Jinwoo Leem, Laura S Mitchell, James HR Farmery, Justin Barton, and Jacob D Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7), 2022.

[24] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[25] Zhuoyi Lin, Sheng Zang, Rundong Wang, Zhu Sun, J Senthilnath, Chi Xu, and Chee Keong Kwoh. Attention over self-attention: Intention-aware re-ranking with dynamic transformer encoders for recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[26] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, 27(1):1–30, 2020.

[27] Shengyu Lu, Qingqi Hong, Beizhan Wang, and Hongji Wang. Efficient resnet model to predict protein-protein interactions with gpu computing. *IEEE Access*, 8:127834–127844, 2020.

[28] Robert M MacCallum, Andrew CR Martin, and Janet M Thornton. Antibody-antigen interactions: contact analysis and binding site topography. *Journal of molecular biology*, 262(5):732–745, 1996.

[29] Derek M Mason, Simon Friedensohn, Cédric R Weber, Christian Jordi, Bastian Wagner, Simon M Meng, Roy A Ehling, Lucia Bonati, Jan Dahinden, Pablo Gainza, et al. Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering*, 5(6):600–612, 2021.

[30] Yoochan Myung, Douglas EV Pires, and David B Ascher. Csm-ab: Graph-based antibody–antigen binding affinity prediction and docking scoring function. *Bioinformatics*, 38(4):1141–1143, 2022.

[31] Tobias H Olsen, Fergus Boyles, and Charlotte M Deane. Observed antibody space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022.

[32] Tobias H Olsen, Iain H Moal, and Charlotte M Deane. Ablang: an antibody language model for completing antibody sequences. *Bioinformatics Advances*, 2(1):vbac046, 2022.

[33] Paul WHI Parren and Dennis R Burton. The antiviral activity of antibodies in vitro and in vivo. 2001.

[34] Susan Payne. *Viruses: from understanding to investigation*. Elsevier, 2022.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[36] Douglas EV Pires and David B Ascher. mcsm-ab: a web server for predicting antibody–antigen affinity changes upon mutation with graph-based signatures. *Nucleic acids research*, 44(W1):W469–W473, 2016.

[37] Matthew IJ Raybould, Aleksandr Kovaltsuk, Claire Marks, and Charlotte M Deane. Cov-abdab: the coronavirus antibody database. *Bioinformatics*, 37(5):734–735, 2021.

[38] Org Re3data. Global Initiative on Sharing Avian Influenza Data, 2012. Freunde von GISAID e.V. Germany. Registry of Research Data Depositories (re3data.org). Global Initiative on Sharing Avian Influenza Data, 2012. https://gisaid.org/. Accessed Feb 2022.

[39] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. bioRxiv 10.1101/622803.

[40] LLC Schrödinger. The pymol molecular graphics system.

[41] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[42] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. The structural basis of antibody-antigen recognition. *Frontiers in immunology*, 4:302, 2013.

[43] Mateusz Sikora, Sören von Bülow, Florian EC Blanc, Michael Gecht, Roberto Covino, and Gerhard Hummer. Computational epitope map of sars-cov-2 spike protein. *PLoS computational biology*, 17(4):e1008790, 2021.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[45] Danqing Wang, Fei Ye, and Hao Zhou. On pre-trained language models for antibody. *bioRxiv*, pages 2023–01, 2023.

[46] Yifan Wang, Cong Xu, Yanxing Wang, Qin Hong, Chao Zhang, Zuyang Li, Shiqi Xu, Qinyu Zuo, Caixuan Liu, Zhong Huang, et al. Conformational dynamics of the beta and kappa sars-cov-2 spike proteins and their complexes with ace2 receptor revealed by cryo-em. *Nature communications*, 12(1):7345, 2021.

[47] Alexander H Williams and Chang-Guo Zhan. Fast prediction of binding affinities of sars-cov-2 spike protein and its mutants with antibodies through intermolecular interaction modeling-based machine learning. *The Journal of Physical Chemistry B*, 126(28):5194–5206, 2022.

[48] Yong Xiao Yang, Pan Wang, and Bao Ting Zhu. Binding affinity prediction for antibody–protein antigen complexes: A machine learning analysis based on interface and surface areas. *Journal of Molecular Graphics and Modelling*, 118:108364, 2023.

[49] Ying Zhang, Kuan Liu, Jiaming Huang, Xinxin Xia, Jiupeng Cao, Guangming Zhao, Patrick WK Fong, Ye Zhu, Feng Yan, Yang Yang, et al. Graded bulk-heterojunction enables 17% binary organic solar cells via nonhalogenated open air coating. *Nature communications*, 12(1):4815, 2021.