Structure-guided T cell receptor and epitope interaction prediction

Alexandru Dumitrescu [1] **Emmi Jokinen** [2 3] **Dani Korpela** [1] **Harri Lähdesmäki** [1]

Abstract

T cells are adaptive immune system cells that develop in mammalian organisms and protect against foreign pathogens. Through their receptor proteins, T cells bind to infected cells and neutralize or kill them. Determining the specificity of these receptors is of crucial medical and biological importance, as it can help reveal disease pathogenesis and aid in early disease detection. In this work, we present extensive analyses of the structural information of T cell receptors (TCRs) and show how it can be used to aid their epitope specificity prediction. We determine the structures of TCRs using the state-of-the-art protein structure prediction, AlphaFold 2 and highlight the main challenges of this approach in the context of TCRs.

1. Introduction

TCRs are protein complexes that form through the recombination of the variable (V), diversity (D), and joining (J) gene segments. Along with B cells, T cells form the adaptive immune system, and they identify cells infected by foreign pathogens. In turn, infected cells will present antigen fragments (epitopes) on their surface through the major histocompatibility complex (MHC). T cells may then recognize that a cell has been infected through the interaction between the TCR and the peptide-MHC (pMHC) complex (Alberts et al., 2015). Subregions of the TCR proteins called complementarity determining regions (CDR) determine the binding affinity to certain pMHCs. CDR1 and CDR2 regions generally bind the MHC protein, while the CDR3 interacts with the epitope (Krogsgaard & Davis, 2005).

[1]Department of Computer Science, Aalto University, Espoo, 02150, Finland [2]Translational Immunology Research program, Dept. of Clinical Chemistry and Hematology, Univ. of Helsinki, Helsinki, 00290, Finland [3]Hematology Research Unit Helsinki, Helsinki University Hospital Comprehensive Cancer Center, Helsinki, 00290, Finland. Correspondence to: Alexandru Dumitrescu <alexandru.dumitresc@aalto.fi>, Harri Lähdesmäki <harri.lahdesmaki@aalto.fi>.

The 2023 ICML Workshop on Computational Biology. Honolulu, Hawaii, USA. 2023.

The total number of possible unique configurations of TCRs in a single organism is very large, with $\alpha\beta$T cells having an estimated number of 10^{15} unique TCR sequences (Davis & Bjorkman, 1988). Furthermore, the vast array of potential epitopes that need to be recognized by the adaptive immune system cells incurs the necessity that TCRs are cross-reactive, and a given pMHC complex may be bound by multiple unique TCRs with various degrees of affinity (Sewell, 2012). These facts motivated the development of a plethora of machine learning methods, in the attempt to solve the epitope specificity classification (into one or more known epitopes) (Tong et al., 2020; Jokinen et al., 2021; 2022) and the more challenging task of directly predicting the TCR-pMHC binding, for any given TCR and pMHC complexes. (Springer et al., 2021; Bradley, 2023).

TCR and pMHC binding is determined not only by their primary structure (amino acid composition) but also by the specific position and orientation of each amino acid in the two protein complexes (Rossjohn et al., 2015; Singh et al., 2017). Motivated by this, research attempting to tune the multimer version of AlphaFold 2 (Evans et al., 2022) for TCR-pMHC geometric structure prediction has been done with some degree of success (Bradley, 2023).

Using multiple sequence alignments (MSA) and structural frameworks, AlphaFold 2 (Jumper et al., 2021) demonstrated that deep learning methods can benefit from alignment methods to produce highly accurate protein structures. We use this structural information, along with the sequence information of TCRs, and analyze how much it can aid epitope specificity classification. Using properties of TCRs (which we describe in Section 2) we are able to efficiently extract these structure predictions, without the need to compute the MSAs for all of our queries, which is by far the most time-consuming step. We analyze the degree to which structures computed using AlphaFold 2 improve the epitope specificity prediction in the context of deep, contextual protein embeddings (Elnaggar et al., 2022), and highlight the main challenges of using this type of structure prediction tool in the context of TCRs.

2. Data

We use three data sets created by Jokinen et al. (2022) which contain annotated TCR sequences that interact with specific

pMHCs collected from VDJdb (Bagaev et al., 2019). The first dataset contains 1977 unique TCR sequences for 21 epitopes with confidence scores of at least 1 (we denote this small dataset as $\mathcal{D}_s$), the second contains 30503 unique TCRs for 51 epitopes with all confidence scores (denoted $\mathcal{D}_l$), and the third dataset contains both α and β chain information, resulting in 20200 sequences for 18 unique epitopes (denoted $\mathcal{D}_{\alpha\beta}$). Each TCR in the three datasets has one or more assigned epitopes (known to bind), and we consider all other epitopes negative (not binding) for that TCR. For all sequences, we utilize either one-hot residue information or residue representations extracted from ProtBERT (Elnaggar et al., 2022), a language model (LM) trained on 216 million protein sequences.

We extract structures for the $\mathcal{D}_s$ sequences using the full databases employed by AlphaFold 2 during the MSA extraction process. For $\mathcal{D}_l$ and $\mathcal{D}_{\alpha\beta}$ sequences, we use the fact that the amino acids outside the CDR3 regions are highly conserved, whereas the CDR3 region can differ even between TCRs encoded by identical gene segments. Therefore, we extract two sets of TCRs that cover all unique V and J gene pairs (covering most of the TCR diversity outside the CDR3 regions) found in $\mathcal{D}_l$ and $\mathcal{D}_{\alpha\beta}$, which contain about 600 and 1800 β and α TCR sequences, respectively. For these two TCR sets, we utilize AlphaFold 2 with the full MSA search and create reduced databases of approximately 40 thousand proteins, which we call $\mathcal{DB}_{\beta r}$ and $\mathcal{DB}_{\alpha r}$. Finally, we run the MSA search for all the remaining TCR sequences (from $\mathcal{D}_l$ and $\mathcal{D}_{\alpha\beta}$) using only $\mathcal{DB}_{\beta r}$ and $\mathcal{DB}_{\alpha r}$. The MSA feature extraction step is now computed in seconds (compared to minutes for the full databases), and the whole structure prediction for one sequence was five to six times faster.

We conduct 10-fold cross-validation experiments on all three datasets. During training, we leave out a validation subset at random (from the nine remaining training subsets) and perform early stopping based on the validation area under the receiver operating characteristic curve (AuROC). The test performance from all 10 folds is used to compare various models with and without AlphaFold 2 computed structures. We report the relative average precision (AP) improvements (on all the test folds) resulted from structural information features. The average AuROC and AP scores for all our experiments can be found in Appendix B.

3. Structure analysis

We first assess the performance of AlphaFold 2 on TCRs that have experimentally annotated structures from (Gowthaman & Pierce, 2019). For a quantitative measure of the structural prediction quality, we use the root-mean-square deviation (RMSD), defined as the square root of the average squared euclidian distances between all amino acids' αC positions of aligned predicted and experimentally verified sequences

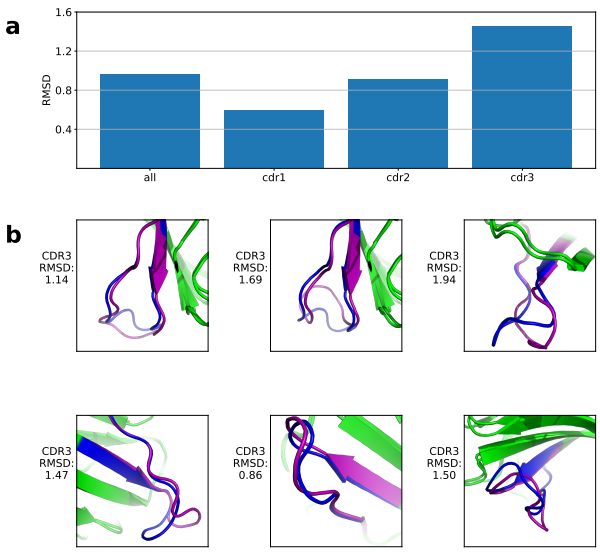

Figure 1. AlphaFold 2 structural prediction performance analysis on CDR regions: **a** The RMSD (in Å) between the experimental structures and AlphaFold predictions is plotted for various TCR regions. **b** Six pairs of aligned experimental and predicted structures are visualized, where blue and purple lines are the CDR3 subregions of the true and predicted structures, respectively, and green contains residues from the rest of the TCRs.

(measured in Å). We compute the RMSD between sequence pairs after we align them using the alignment tool from PyMOL (Schrödinger, LLC, 2015), which minimizes the RMSD. We align the sequences based on the TCR residues within the framework regions (FR) of the TCRs. More specifically, the RMSD between residues (αC atom positions) within FR regions of experimentally determined and AlphaFold 2 predicted structures is minimized.

CDR3 regions of TCRs can differ even between identical encoding gene segments (making the MSA of AlphaFold 2 less useful) and usually have multiple valid folding conformations (Reiser et al., 2003). Therefore, the CDR3 predictions are considerably more challenging than other TCR subregions, as shown in Figure 1. Still, structural similarity can be observed in some cases, and we determine to what degree these structures can still aid epitope specificity predictions throughout this work.

We consider the epitope specificity classification task, where we define $(\mathbf{x}_n, \mathbf{y}_n, \mathbf{s}_n)$ triplets of sequences, label vectors, and structural features derived from AlphaFold 2. The label vectors $\mathbf{y}_n$ are binary vectors, having $\mathbf{y}_{n,c} = 1$ if the TCR n recognizes epitope c, and 0 otherwise. We train classifiers with and without $\mathbf{s}_n$, and determine the resulted epitope-specific relative improvements.

3.1. Invariant structural features analysis

We now describe and assess the utility of various features extracted from AlphaFold 2 structural predictions. Hereafter, we refer to structural information that is invariant to translation and rotation as invariant structural features. Our first experiments use the invariant structural features through graph neural networks, and we continue with the same type of features in the context of CNN architectures.

Graphs are defined by their vertices $v_i \in \mathcal{V}$ and edge information $e_{i,j} \in \mathcal{E}$. Undirected binary graphs consider each edge as $e_{i,j} = 1$, if nodes v_i and v_j are connected, and $e_{i,j} = 0$ otherwise. We construct TCR graphs by attributing each node v_i a one-hot vector of the amino acid at position i in the protein sequence and determine the node's binary connections to all other amino acids $j \neq i$ by thresholding αC atom distances with various thresholds t.

Motivated by the previous success of structural information usage in protein function prediction (Gligorijević et al., 2021), we employ graph convolutional networks (Kipf & Welling, 2017) as our first model choice. We compare the epitope specificity performance considered from training three-layered GCN networks, with 4, 5, and 6Å distance thresholds t. In addition, all amino acids that are adjacent in a sequence are always considered connected.

In Figure 2, we illustrate the utility of the invariant structure information. Previous success in protein function prediction using protein structure-derived contact maps (Gligorijević et al., 2021) most likely relied on the very high diversity in the input contact maps. In contrast, the possible folding diversity is highly limited in CDR3 sequences, mainly due to their limited size (the average CDR3 length in $\mathcal{D}_s$ is 14, and the longest one is 22). We plot the relative improvements obtained when contact maps are derived from the structural predictions, compared to GCNs that have connections only between adjacent residues, indicating that the use of structure predictions provides no improvement.

Next, we use one of the best-performing CNN architectures for TCR epitope classification, TCRconv (Jokinen et al., 2022), and test three types of invariant structural features. For sequences of length N, we define loop distance as the distance from the first to all other CDR3 residues' αC atoms, ld $\in \mathbb{R}^{1\times N}$, the euclidian similarity between all residue pairs' αC atom coordinates, es $\in \mathbb{N} \times \mathbb{N}$, and the cosine of the angles between every residue pair, $\phi \in \mathbb{R}^{N\times N}$ (please refer to Appendix A for a more detailed description). We compare the results using invariant structures to the model using only the one-hot residue information (Figure 3). We note a consistent improvement for most epitopes when using euclidian similarity, and a larger improvement for epitopes with fewer positive TCRs in the data.

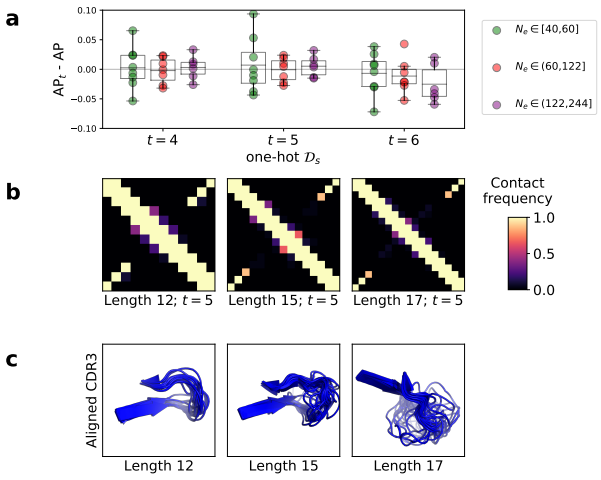

Figure 2. Invariant feature analysis: **a** Epitope-specific AP score difference between GCNs trained with contacts resulted from various thresholds t and a GCN trained using only adjacent residue contacts on $\mathcal{D}_s$. The mean performance difference is reported for three epitope categories, based on their number of positive TCRs in the data N_e. **b** Average αC contact maps (31 sequences from $\mathcal{D}_s$ for each length) derived from AlphaFold 2 predicted structures for various CDR3 lengths and a threshold of $t = 5$Å. **c** Visualization of the CDR3 3D shapes after aligning the TCRs to a reference TCR sequence (chosen randomly) using PyMOL

3.2. Coordinate analysis

We now attempt to use most of the information available in the AlphaFold 2 structural predictions, by using the 3D coordinates of four atoms in each residue within the TCRs. Doing so allows us the maximum amount of flexibility in constructing structural features for our models, and we hypothesize that both the backbone position of the TCRs, as well as the orientation of all residues in the sequences contribute to the higher performance increase resulting from coordinate information compared to the invariant features.

The main challenge when dealing with 3D coordinate structures is training a model which identifies the same sequence patterns irrespective of global translations and rotations applied to a sequence. To a large extent, we show that this issue can be mitigated in the context of TCRs, by aligning the highly conserved framework regions of all train and test sequences to those of a randomly chosen reference TCR S_r. To do this, we use the align algorithm developed by PyMOL (Schrödinger, LLC, 2015), with which we reposition all sequences s.t. the RMSD between them and S_r is minimized on the framework regions.

We analyze the epitope-specific predictive performance difference when we add the structural information in the form of three-dimensional coordinates of N in the amino group,

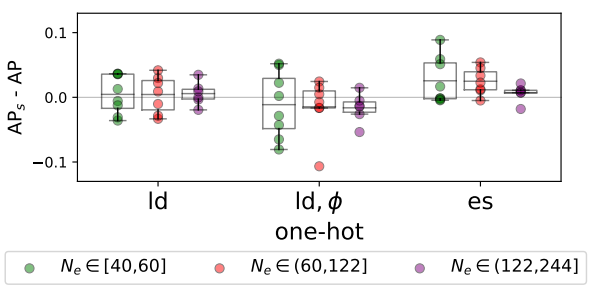

Figure 3. Loop distance (ld), loop distance and residue angles (ld, ϕ), and euclidian similarity (es) are concatenated with one-hot residue representations, and TCRconv's performance using these invariant structures (AP$_s$) is subtracted from TCRconv's performance trained only on one-hot for each epitope (AP).

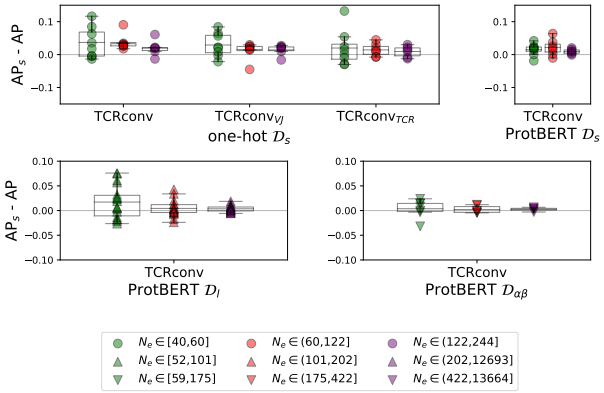

Figure 4. Performance difference between models with and without structure. For $\mathcal{D}_s$, we report both one-hot and ProtBERT-embedded amino acids results. For the one-hot experiments, we also add the V and J genes as categorical information and use all residues from the TCRs (denoted TCRconv$_{VJ}$ and TCRconv$_{TCR}$). For the $\mathcal{D}_{\alpha\beta}$ experiment, we use the structures from both chains.

the αC, and the carboxyl group C and O atoms, for each amino acid in a sequence. We pre-process the positional information using an MLP (with dimension $\mathbb{R}^{12\times12}$) and then concatenate along the d-dimensional input representation of each residue, the atoms' position representations forming $\mathbb{R}^{d+12}$ valued amino acid input features.

In Figure 4, we show the relative improvements when adding coordinate information for all three datasets. For $\mathcal{D}_s$, we report the improvements for both one-hot and ProtBERT encoded residues. We also add information about the rest of the TCR in the model by adding the corresponding encoding genes as categorical features ("TCRconv$_{VJ}$") and alternatively using the whole TCRs ("TCRconv$_{TCR}$"). This ensures the improvements are not only artifacts of the V and J gene information being encoded within the structures we input to the model.

Notably, when TCRs are encoded with one-hot vectors, the relative improvement added by structural information is larger compared to ProtBERT residue-encoded embeddings. This is in accordance with (Vig et al., 2021), which showed that protein LMs intrinsically learn protein structure, as their attention weights are correlated with contact maps. Although slightly larger, the improvements on D_s, when using one-hot and coordinate information (Figure 4), are similar to those obtained using euclidian similarity and one-hot information (Figure 3), which indicates that the most predictive information contained in AlphaFold 2-predicted structures is the pairwise proximity between residues, and that is likely already captured by ProtBERT representations.

Furthermore, the structural performance prediction of AlphaFold 2 significantly deteriorates in the absence of MSAs (Lin et al., 2022) and the MSA corresponding to the CDR3 subregion is significantly less useful, as the V(D)J recombination produces CDR3 sequences in a quasi-random manner. Therefore, the ProtBERT-encoded residue improvements in Figure 4, rely on the ability of AlphaFold 2 to predict structural information which is both missing from ProtBERT, and is accurate enough, considering that the produced MSAs for the most crucial region (CDR3) will not be useful.

Similar consistent but small improvements are shown for $\mathcal{D}_l$ and $\mathcal{D}_{\alpha,\beta}$ datasets, as the TCRs are ProtBERT-encoded in these cases for the results shown in Figure 4.

4. Conclusion

We provided an extensive analysis of TCR structural information derived from AlphaFold 2. The structural templates and MSAs can provide valuable input features, but in the context of highly variable immune cell configurations, structure prediction becomes more challenging. Although the diversity of CDR3 stable configurations (for a fixed sequence) makes it difficult to obtain an exact number quantifying the structural prediction accuracy, our experiments indicate the necessity of further improvements in structural prediction methods in this context. We postulate that the enormous TCR and antibody diversity will always impair accurate structural prediction by extrapolation of similar (genetically related organism) proteins. This would either require very large amounts of annotated TCR and antibody structures or modeling the physicochemical properties driving protein folding.

References

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., ..., and Hunt, T. *Molecular Biology of The*

Cell, sixth edition, chapter 24. Garland Science, Taylor and Francis Group, LLC, 711 Third Avenue New York NY 10017, US, 2015.

Bagaev, D. V., Vroomans, R. M. A., Samir, J., Stervbo, U., Rius, C., Dolton, G., ..., and Shugay, M. VDJdb in 2019: database extension, new analysis infrastructure and a T-cell receptor motif compendium. *Nucleic Acids Research*, 48(D1):D1057–D1062, 10 2019. ISSN 0305-1048. doi: 10.1093/nar/gkz874.

Bradley, P. Structure-based prediction of T cell receptor:peptide-MHC interactions. *eLife*, 12:e82813, jan 2023. ISSN 2050-084X. doi: 10.7554/eLife.82813.

Davis, M. M. and Bjorkman, P. J. T-cell antigen receptor genes and T-cell recognition. *Nature*, 334(6181):395–402, Aug 1988. ISSN 1476-4687. doi: 10.1038/334395a0.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., ..., and Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., ..., and Hassabis, D. Protein complex prediction with alphafold-multimer. *bioRxiv*, 2022. doi: 10.1101/2021.10.04.463034.

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., ..., and Bonneau, R. Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23303-9.

Gowthaman, R. and Pierce, B. G. TCR3d: The T cell receptor structural repertoire database. *Bioinformatics*, 35(24):5323–5325, December 2019.

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., and Lähdesmäki, H. Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Comput Biol*, 17(3):e1008814, March 2021.

Jokinen, E., Dumitrescu, A., Huuhtanen, J., Gligorijević, V., Mustjoki, S., Bonneau, R., ..., and Lähdesmäki, H. TCRconv: predicting recognition between T cell receptors and epitopes using contextualized motifs. *Bioinformatics*, 39(1), 12 2022. ISSN 1367-4811. doi: 10.1093/bioinformatics/btac788. btac788.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ..., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2.

Kipf, T. N. and Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Krogsgaard, M. and Davis, M. M. How T cells' see'antigen. *Nature immunology*, 6(3):239–245, 2005.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., ..., and Rives, A. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022. doi: 10.1101/2022.07.20.500902.

Reiser, J.-B., Darnault, C., Grégoire, C., Mosser, T., Mazza, G., Kearney, A., ..., and Malissen, B. CDR3 loop flexibility contributes to the degeneracy of TCR recognition. *Nature Immunology*, 4(3):241–247, Mar 2003. ISSN 1529-2916. doi: 10.1038/ni891.

Rossjohn, J., Gras, S., Miles, J. J., Turner, S. J., Godfrey, D. I., and McCluskey, J. T Cell Antigen Receptor Recognition of Antigen-Presenting Molecules. *Annual Review of Immunology*, 33(1):169–200, 2015. doi: 10.1146/annurev-immunol-032414-112334. PMID: 25493333.

Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. November 2015.

Sewell, A. K. Why must T cells be cross-reactive? *Nature Reviews Immunology*, 12(9):669–677, Sep 2012. ISSN 1474-1741. doi: 10.1038/nri3279.

Singh, N. K., Riley, T. P., Baker, S. C. B., Borrman, T., Weng, Z., and Baker, B. M. Emerging Concepts in TCR Specificity: Rationalizing and (Maybe) Predicting Outcomes. *The Journal of Immunology*, 199(7):2203–2213, 10 2017. ISSN 0022-1767. doi: 10.4049/jimmunol.1700744.

Springer, I., Tickotsky, N., and Louzoun, Y. Contribution of T Cell Receptor Alpha and Beta CDR3, MHC typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology*, 12, 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.664514.

Tong, Y., Wang, J., Zheng, T., Zhang, X., Xiao, X., Zhu, X., ..., and Liu, X. SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction. *Comput Biol Chem*, 87:107281, June 2020.

Vig, J., Madani, A., Varshney, L. R., Xiong, C., Socher, R., and Rajani, N. BERTology Meets Biology: Interpreting Attention in Protein Language Models. In *International Conference on Learning Representations*, 2021.

A. Invariant structural features

For an n-length CDR sequence, we extract various relative distances and similarities between all residue pairs $i, j \in \{1, 2, \ldots, n\}$, forming $A \in \mathbb{R}^{n \times n}$. Since the convolutional architectures will require a fixed-dimensional input, we concatenate the resulting invariant feature information matrices to form $A' \in \mathbb{R}^{N \times N_b}$, with N being the global maximum sequence length (s.t. the channel dimension always has the same length), and N_b is the current minibatch maximum sequence length.

Given the three-dimensional coordinates of the N in the amino group, the αC, and the carboxyl group C atoms, we form the vectors $\overrightarrow{\mathrm{C}\alpha\mathrm{C}}$ and $\overrightarrow{\mathrm{CN}}$. For two amino acids with vector pairs ($\overrightarrow{\mathrm{C}\alpha\mathrm{C}}_i$,$\overrightarrow{\mathrm{CN}}_i$) and ($\overrightarrow{\mathrm{C}\alpha\mathrm{C}}_j$, $\overrightarrow{\mathrm{CN}}_j$), we form their plane equations $A_i x + B_i y + C_i z + D_i = 0$, and $A_j x + B_j y + C_j z + D_j = 0$ and extract the cosine of the angles between them:

$$\phi = A_{ij} = \frac{|A_i A_j + B_i B_j + C_i C_j|}{\sqrt{A_i^2 + B_i^2 + C_i^2}\sqrt{A_j^2 + B_j^2 + C_j^2}} = \cos(\mathbf{r}_i, \mathbf{r}_j) \tag{1}$$

The main drawback of this approach is that there is no sensible padding, which makes this analysis imperfect. The predictive influence of residue angles measured by this analysis might, therefore, be underestimated.

Next, we use euclidian similarity, defined as:

$$\mathrm{es} = \frac{1}{1 + ||\mathbf{r}_i - \mathbf{r}_j||_2}, \tag{2}$$

where $\mathbf{r}_i$ and $\mathbf{r}_j$ are the three-dimensional coordinates of αC atoms of residues i and j. Compared to relative residue angle information, this method has sensible padding (0 padding corresponds to 0 similarity between a given residue and a padding token). This is likely the main reason why euclidian similarity is the best-performing invariant structural feature.

Loop distance ld is defined as the distance between a reference residue (chosen as the first residue in the CDR3 sequence), to all other residues, forming a $\mathbb{R}^{1 \times N_b}$ feature vector (with N_b being the sequence length). The motivation behind this simple feature information is based on the assumption that the tip of the CDR3 loops should most often come in contact with the epitope, and therefore the highest ld values should indicate the most likely contacting residues.

B. Epitope prediction performance

Table 1. Performance analysis on $\mathcal{D}_s$, for one-hot encoded sequences. Performance reported for various epitope frequencies N_e.

Structure type	additional features	AuROC $N_e \in (40, 60]$	AuROC $N_e \in (60, 122]$	AuROC $N_e \in (122, 244]$	AP $N_e \in (40, 60]$	AP $N_e \in (60, 122]$	AP $N_e \in (122, 244]$
no struct	-	0.77 ± 0.15	0.8 ± 0.13	0.82 ± 0.07	0.38 ± 0.22	0.44 ± 0.26	0.61 ± 0.14
coord. values	-	$\mathbf{0.81 \pm 0.13}$	$\mathbf{0.82 \pm 0.12}$	$\mathbf{0.84 \pm 0.06}$	$\mathbf{0.42 \pm 0.23}$	$\mathbf{0.47 \pm 0.26}$	$\mathbf{0.63 \pm 0.13}$
ld	-	0.79 ± 0.13	0.8 ± 0.13	0.83 ± 0.07	0.39 ± 0.22	0.44 ± 0.26	0.61 ± 0.12
ld, ϕ	-	0.78 ± 0.15	0.8 ± 0.12	0.82 ± 0.07	0.37 ± 0.23	0.42 ± 0.24	0.59 ± 0.14
es	-	0.8 ± 0.14	0.82 ± 0.12	0.83 ± 0.07	0.41 ± 0.23	0.46 ± 0.25	0.62 ± 0.13
no struct	VJ	0.8 ± 0.13	$\mathbf{0.83 \pm 0.11}$	0.83 ± 0.07	0.42 ± 0.23	0.5 ± 0.26	0.62 ± 0.14
coord. values	VJ	$\mathbf{0.82 \pm 0.13}$	$\mathbf{0.83 \pm 0.13}$	$\mathbf{0.84 \pm 0.07}$	$\mathbf{0.45 \pm 0.22}$	$\mathbf{0.51 \pm 0.27}$	$\mathbf{0.63 \pm 0.13}$
no struct	TCR	$\mathbf{0.85 \pm 0.11}$	$\mathbf{0.85 \pm 0.12}$	0.86 ± 0.07	0.45 ± 0.23	0.54 ± 0.25	0.65 ± 0.12
coord. values	TCR	$\mathbf{0.85 \pm 0.11}$	$\mathbf{0.85 \pm 0.13}$	$\mathbf{0.87 \pm 0.07}$	$\mathbf{0.47 \pm 0.22}$	$\mathbf{0.55 \pm 0.25}$	$\mathbf{0.66 \pm 0.12}$

Table 2. Performance analysis on $\mathcal{D}_s$, $\mathcal{D}_l$, and $\mathcal{D}_{ab}$ for ProtBERT encoded sequences. Performance reported for various epitope frequencies N_e.

Structure type	**Dataset**	**AuROC**			**AP**		
		$N_e \in [40, 60]$	$N_e \in (60, 122]$	$N_e \in (122, 244]$	$N_e \in [40, 60]$	$N_e \in (60, 122]$	$N_e \in (122, 244]$
no struct	$\mathcal{D}_s$	0.86 ± 0.12	0.85 ± 0.13	$\mathbf{0.86 \pm 0.07}$	0.5 ± 0.22	0.55 ± 0.25	0.66 ± 0.12
coord. values	$\mathcal{D}_s$	$\mathbf{0.87 \pm 0.11}$	$\mathbf{0.86 \pm 0.12}$	$\mathbf{0.86 \pm 0.07}$	$\mathbf{0.51 \pm 0.22}$	$\mathbf{0.57 \pm 0.25}$	$\mathbf{0.67 \pm 0.11}$
		$N_e \in (52, 101]$	$N_e \in (101, 202]$	$N_e \in (202, 12693]$	$N_e \in (40, 60]$	$N_e \in (60, 122]$	$N_e \in (122, 244]$
no struct	$\mathcal{D}_l$	0.78 ± 0.13	0.77 ± 0.11	0.72 ± 0.13	0.23 ± 0.2	$\mathbf{0.24 \pm 0.2}$	$\mathbf{0.3 \pm 0.28}$
coord. values	$\mathcal{D}_l$	$\mathbf{0.79 \pm 0.13}$	$\mathbf{0.78 \pm 0.11}$	$\mathbf{0.73 \pm 0.13}$	$\mathbf{0.25 \pm 0.21}$	$\mathbf{0.24 \pm 0.2}$	$\mathbf{0.3 \pm 0.27}$
		$N_e \in (40, 60]$	$N_e \in (60, 122]$	$N_e \in (122, 244]$	$N_e \in (40, 60]$	$N_e \in (60, 122]$	$N_e \in (122, 244]$
no struct	$\mathcal{D}_{ab}$	$\mathbf{0.79 \pm 0.11}$	0.82 ± 0.13	$\mathbf{0.78 \pm 0.13}$	0.21 ± 0.19	$\mathbf{0.47 \pm 0.25}$	$\mathbf{0.6 \pm 0.28}$
coord. values	$\mathcal{D}_{ab}$	$\mathbf{0.79 \pm 0.12}$	$\mathbf{0.83 \pm 0.13}$	$\mathbf{0.78 \pm 0.12}$	$\mathbf{0.22 \pm 0.18}$	$\mathbf{0.47 \pm 0.26}$	$\mathbf{0.6 \pm 0.28}$

Table 3. Performance analysis on $\mathcal{D}_s$ for GCN models using adjacency matrices based on AlphaFold 2 structural information (contacts determined based on thresholds $t = 4, 5, 6$Å). Here, "no struct" refers to a GCN trained using adjacency matrices that only consider amino acids directly adjacent in the sequence as being in contact (no structural information is used).

Structure type	**AuROC**			**AP**		
	$N_e \leq 60$	$N_e \in (60, 122]$	$N_e > 122$	$N_e \leq 60$	$N_e \in (60, 122]$	$N_e > 122$
no struct	0.74 ± 0.13	0.74 ± 0.12	0.77 ± 0.09	0.26 ± 0.2	$\mathbf{0.32 \pm 0.21}$	0.47 ± 0.16
$t = 4$	0.74 ± 0.14	$\mathbf{0.76 \pm 0.11}$	0.77 ± 0.08	0.26 ± 0.2	0.31 ± 0.2	$\mathbf{0.48 \pm 0.15}$
$t = 5$	$\mathbf{0.75 \pm 0.14}$	$\mathbf{0.76 \pm 0.11}$	$\mathbf{0.77 \pm 0.08}$	$\mathbf{0.27 \pm 0.21}$	0.31 ± 0.2	$\mathbf{0.48 \pm 0.16}$
$t = 6$	0.71 ± 0.16	0.74 ± 0.13	0.76 ± 0.09	0.26 ± 0.19	0.3 ± 0.2	0.45 ± 0.16