Assessing the Generalization Capabilities of TCR Binding Predictors via Peptide Distance Analysis

Leonardo V. Castorina [1,*], **Filippo Grazioli** [2], **Pierre Machart** [2], **Anja Mösch** [2] **and Federico Errica** [2]

[1] *School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, UK*
[2] *NEC Laboratories Europe, 69115 Heidelberg, Germany*

Correspondence*:
Corresponding Author
leonardo.castorina@ed.ac.uk

ABSTRACT

Accurate TCR-peptide/pMHC binding prediction is crucial to understand immune responses and develop therapeutic interventions. In this study, we propose a novel approach to assess the generalization capabilities of machine learning models in predicting TCR-peptide binding for unseen viral peptides. Our method involves measuring the distance between training and test peptides and leveraging this metric to assess model generalization capabilities. In this study, we employ sequence-based and structure-based distances.

Through extensive experimentation and analysis, we demonstrate that the distance-based assessment provides valuable insights into the generalizability of the model and its ability to predict TCR binding to previously unobserved viral peptides.

The findings from this study shed light on the challenges and opportunities in TCR-peptide/pMHC binding prediction and contribute to the advancement of computational immunology and peptide-based therapeutics.

Keywords: TCR, peptide, MHC, binding prediction, interaction prediction, machine learning, TCR - T cell receptor, 3D structure

1 INTRODUCTION

The immune system has evolved to recognize different pathogens, such as viruses, which enter cells and exploit their resources for replication. These cells can be healthy or affected by detrimental malfunction like pre-cancerous and cancerous cells. The immune response depends on how well the system can distinguish between healthy and infected/aberrant cells. It does so by using the presence of specific molecules on the cell surface. These molecules are called Major Histocompatibility Complexes (MHC). The proteasome, a complex of protease enzymes, degrades proteins in the cell, creating antigenic peptides. The MHC binds the peptide and presents it on the cell surface forming the peptide-bound MHC (pMHC) (Hewitt, 2003).

In infected cells, some of these peptides are derived from viral proteins, in contrast to healthy cells where peptides are derived from common cellular proteins (Eisenlohr et al., 2007). T cells can recognize these viral peptides in the pMHC by binding their T cell receptors (TCRs), which subsequently leads to an immune response (Hewitt, 2003). A visual overview of the TCR-pMHC complex is presented in the

Supplementary Materials. TCRs are able to recognize many different peptides via diverse sequences at the variable regions of their α and β chains (Smith-Garvin et al., 2009). The variety of TCRs within the TCR repertoire of a human body increases the likelihood of a TCR binding to a foreign peptide.

Immunotherapies targeting cancer or viral infections use binding specificity to activate T cells, enhancing the immune system response. Specificity ensures that T cell binding occurs uniquely on the target, to kill say cancerous cells, and avoid healthy cells. Research in immunotherapies covers two categories of therapies: Adoptive T Cell Therapy and T Cell Inducing Vaccines. In Adoptive T Cell Therapy, cancer patients receive specific T cells targeting and destroying the tumour (He et al., 2019; Waldman et al., 2020). T-Cell Inducing Vaccines leverage antigenic peptides or other classes of antigens to trigger specific T-cell development against pathogens such as viruses (Gilbert, 2011; Heslop and Leen, 2013).

The design of these immunotherapies requires an in-depth understanding of biochemical interactions between TCRs and pMHC and precise stimulation of the immune system. This requires a thorough understanding of which sequences of TCR and pMHC interact, the strength of the interactions, and the strength of the immune response. The development of computational models that predict whether TCRs and pMHC interact would allow for a faster *in silico* screening of sequences and consequent design of TCR sequences that bind specific target pMHCs. However, public TCR-pMHC interaction datasets are often limited in diversity and potentially biased. For example, there can be sequence bias towards commonly studied viruses or binding/non-binding pairs bias from experiment setups. Additionally, laboratory methods that validate the TCR-pMHC interactions, like surface plasmon resonance, titration calorimetry and fluorescence anisotropy, are usually resource-intensive and time-consuming (Piepenbrink et al., 2009).

Researchers have recently achieved good performances with machine-learning-based (ML) solutions for TCR binding prediction (Springer et al., 2021; Montemurro et al., 2021; Weber et al., 2021; Grazioli et al., 2022a). The task consists of classifying input tuples of short amino acid sequences such as (peptide, CDR3-β) or (peptide, CDR3-β, CDR3-α, MHC) as binding or non-binding. The splitting of this data into training, validation and test sets can significantly affect the estimated generalization performance of ML models (Grazioli et al., 2022b; Deng et al., 2023).

However, obtaining good splitting procedures is not obvious. Grazioli et al. (2022b) investigated the effect of different splitting techniques on the TChard dataset, which combines samples from IEDB Vita et al. (2019), VDJdb Bagaev et al. (2020) and McPAS-TCR Tickotsky et al. (2017). They also investigated the effect of considering negative samples from wet lab assays and from random shuffling of the positive tuples. They observed that, when using vanilla Random Split (RS) with negative samples from wet lab assays, ML-based models achieve an estimated area under the receiver operator characteristic (AUROC) larger than 95% on the test set. In this setting, the sets of CDR3 sequences in positive and negative samples are disjoint. This allows the models to memorize whether a CDR3 sequence was observed in either the negative or positive samples at training time. Grazioli et al. (2022b) show that a simple countermeasure to this problem consists in creating negative samples by randomly shuffling the positive tuples. In this setting, when the RS is employed, models achieve an AUROC comprised between 70% and 80%. Nevertheless, using the RS, CDR3 sequences and peptides may appear in both the training and test set, leading to overoptimistic estimates of the real-world model generalization capabilities. Grazioli et al. (2022b) propose an alternative splitting method called Hard Split (HS). The HS ensures that test peptides are never observed at training time. Therefore, the peptide sequences in the training, validation and test sets are unique to their respective sets. When using the HS and negative samples are obtained via random shuffling, the model performs barely better than random guesses, with an AUROC smaller than 55%.

To ensure real-world applicability, the HS aims at evaluating model performance under the highest robustness requirements. In fact, new peptides arising from new pathogens may differ significantly from those used at training time. ML models can only be safely employed for broad real-world applications if they show sufficient generalization to unseen sequences. For example, as demonstrated by the COVID-19 pandemic, viral antigens from SARS-CoV-2 have a high mutation rate, giving rise to new strains with significant sequence variability (Hossain et al., 2021). Other than being a real threat on an international health scale, these variants present a challenge for interaction prediction models, as new sequences may belong to significantly different distributions than those in the training set. The prediction of samples significantly different from the training set is a problem known as out-of-distribution (OOD) prediction.

In practice, RS and HS represent two extremes of a much wider spectrum of data splitting options, which would be useful to assess the robustness of ML models as the difference in distributions between sets increases, thus making the task of prediction harder. Knowing how well ML predictors generalize to an increasing shift in distributions between training and test sets would provide a tangible metric to measure when it is appropriate and to which extent we can employ these methods in the real world.

For this reason, in this work we introduce a new dataset splitting algorithm called *Distance Split (DS)*. Analogously to the HS, peptides placed in the training set are absent from the test set and vice versa, but the distances between sequences in the training and in the test set are controlled. We increase such distance to test performance on peptides that are "further apart" from the ones seen by the model at training time, making the task increasingly harder. We explore different distance metrics, i.e., sequence metrics such as LocalDS and GlobalDS, respectively derived from local and global sequence alignments (Cock et al., 2009), as well as shape metrics such as Root Mean Squared Deviation (*RMSD*) (Schrödinger, LLC, 2015), leveraging recent advances in protein folding to obtain the 3D structure of peptides.

2 METHOD

2.1 Dataset Creation

With the goal of vaccine development against viruses, we selected a viral subset of the VDJdb dataset, focusing on human host and MHC class I (Shugay et al., 2018). We omitted MHC class II due to low samples available. The dataset includes peptides from SARS-CoV-2, Influenza A, Human Immunodeficiency Virus (HIV), Hepatitis C Virus (HCV) and Cytomegalovirus (CMV). We discarded data points which do not include both the CDR3-β and the peptide. The dataset contains 52 unique MHC A and 1 MHC B alleles, 16,504 unique CDR3-α sequences, 28,831 unique CDR3-β sequences and 757 unique peptides, for a total of 34,415 binding samples. To generate non-binding (i.e., negative) samples, we randomly paired sequences of CDR3-β and peptides. We do this under the assumption that it is unlikely that random pairs will bind, as also outlined by Grazioli et al. (2022a). To create a balanced dataset, we randomly generated 36,641 samples of non-binding combinations of CDR3-β and peptide sequences, increasing the total number of data points to 65,946. In this study, we restricted our attention to (peptide, CDR3-β) samples.

2.1.1 Obtaining 3D Structures of Peptides

The VDJdb dataset contains information on the primary structure of proteins, i.e., the sequence of amino acids. In reality, these sequences exist as 3D shapes and as part of a complex (Shugay et al., 2018). The 3D shapes of peptides is however unavailable in the VDJdb dataset. We used ESMFold, a light-weight sequence-to-shape language model (Lin et al., 2022) and its online API [1] to generate 3D structures for

[1] https://esmatlas.com/resources?action=fold

the peptide sequences. For structures that gave errors, we used OmegaFold (Wu et al., 2022), a similar sequence-to-shape model. The shapes are then used to obtain a 3D distance between peptides.

2.2 The Distance Split

The goal of the Distance Split (DS) algorithm is to create data splits of the available (peptide, CDR3-β) samples such that the peptides in the test split are at a pre-specified median distance away from the training ones. In this work we used three percentile intervals over the distance distribution: $0-33$, $33-66$ and $66-100$. We hypothesized that, as the distance increases, the generalization performance of the models should decrease, as the test peptides are more different from the training set.

To divide peptides into different splits, we used three distance metrics: LocalDS, GlobalDS and RMSD. LocalDS (resp. GlobalDS) is a local (resp. global) alignment algorithm from BioPython that produces a distance score between two peptides represented as a sequence (Cock et al., 2009). We used the BLOSUM62 matrix as a scoring dictionary (Henikoff and Henikoff, 1992). BLOSUM62 is an amino acid substitution matrix representing the log-likelihood of all possible substitutions of amino acids. We used gap open penalty $D = -10$ and gap extension penalty $S = -1$.

We calculated the RMSD shape distance using PyMol (Schrödinger, LLC, 2015). PyMol aligns any pairs of peptide 3D structures by structural similarity (as opposed to sequence similarity) and computes the RMSD. Once aligned, PyMol calculates the RMSD of the distance between Cα in the two structures.

For each of metric, we computed the distances between all possible pairs of 757 peptides creating a symmetric distance matrix $\mathbf{M} \in \mathbb{R}^{757\times757}$, obtaining $\mathbf{M}_{GlobalDS}$, $\mathbf{M}_{LocalDS}$ and $\mathbf{M}_{RMSD}$. We then calculated the median for each peptide (i.e., row-wise median) $\mathbf{m}_{med} \in \mathbb{R}^{757}$. Each value of $\mathbf{m}_{med}$ corresponds to the median distance of the peptide against all of the other peptides. For each of the three percentile intervals $0-33$, $33-66$ and $66-100$, we calculate the lowest and highest thresholds of distances and filter out all the peptides falling outside the range. The data consists of pairs of (peptide, CDR3-β) and a binding/non-binding label. These pairs appear with different frequencies in the dataset, so for each split (training, validation, and test) we define a target budget of pairs based on the split ratio and the number of occurrences of peptides.

We started by selecting the percentage distance range for the test set. The distance constraints are applied to the matrix $\mathbf{M}$, filtering out peptides outside the distance range. The remaining peptides are picked at random, and the related pair (peptide, CDR3-β) is added to the test set. The process was repeated until the budget was filled. The unselected peptides are then recycled back with the filtered out peptides and then distributed to the validation split. The remaining peptides are used for the training split.

Finally, for the validation and test set, we further imposed maximum and minimum pairs counts. This was to avoid having too great of a diversity (i.e. many peptides with low pair counts), or too little of a diversity (i.e. few peptides with high pair counts). In this work we used a maximum count of 5,000 and minimum count of 5. The complete algorithm for the DS is shown in the supplementary materials.

2.3 Training Splits for TCR-Peptide Interaction

We generated training, test and validation splits of (peptide, CDR3-β) samples employing RS, HS and DS for GlobalDS, LocalDS and RMSD. We used the 80-10-10 split ratio for train-test-validation. As our TCR-Peptide Interaction ML model, we use the recent Attentive Variational Information Bottleneck (AVIB) model (Grazioli et al., 2022b). We first performed a hyperparameter optimization on the HS. The resulting

optimal hyperparameters were then used for all experiments. These were: 150-dimensional latent space, 1 attention head, 0.1 dropout rate, 0.001 learning rate and 1.899×10^{-5} weight decay.

3 RESULTS

We computed the distance metrics LocalDS, GlobalDS and RMSD for the viral VDJdb dataset and compared them against peptides count in Figure 1 and 2 respectively. We found that, due to the short length of the peptides, GlobalDS and LocalDS were identical to each other in our dataset. Hence, from here on, we will refer uniquely to LocalDS. Summary statistics of the metrics are summarized in Table 1. The peptides count distribution is heavily skewewd towards a count of less than 10 instances, with several outliers as shown by the high standard deviation. The LocalDS distribution is roughly normally distributed. Finally the RMSD has a bimodal distribution with peaks at roughly 0.25 and 0.63 Å.

Furthermore, in Figure 3 we compare RMSD against LocalDS. There is a correlation of 0.03 (p = 1.67×10^{-112}) between RMSD and LocalDS. Thus, both metrics were used as distance metrics for the DS.

Metric	Average	Median	Stdev
Epitopes (counts)	45	2	576
LocalDS	2.0	2.31	0.46
GlobalDS	2.0	2.31	0.46
RMSD (Å)	0.46	0.36	0.27

Table 1. Summary statistics of peptides counts, LocalDS Distance and RMSD for the viral VDJdb dataset.

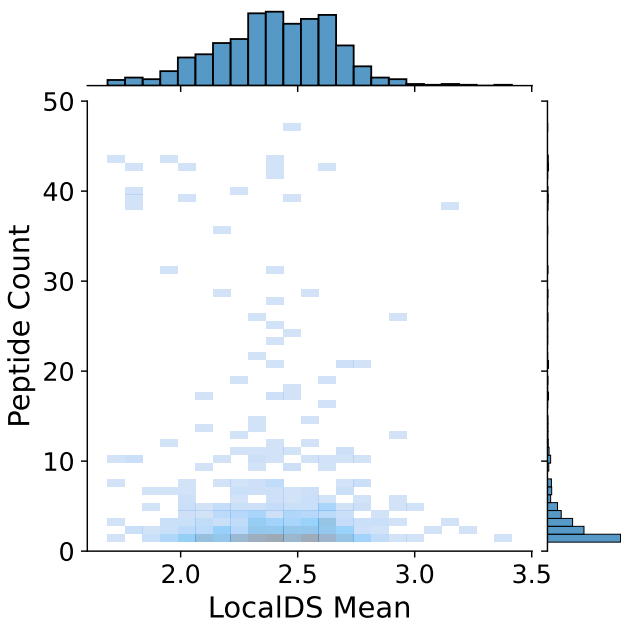


Figure 1. Epitope count against LocalDS (sequence) distance

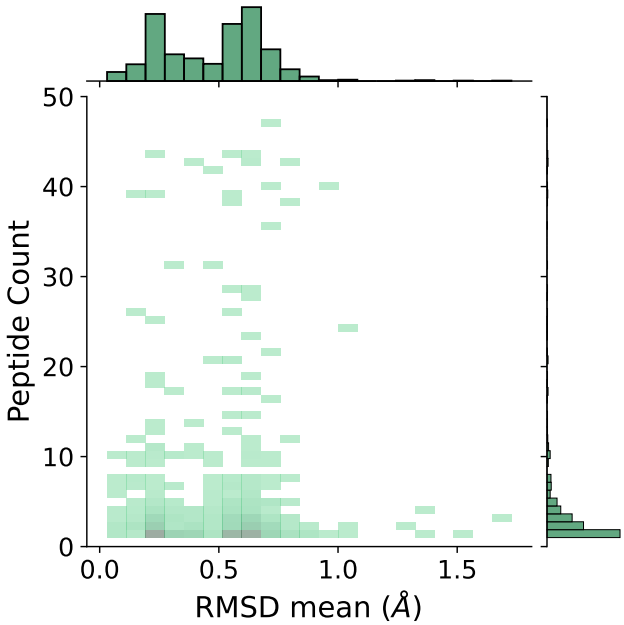


Figure 2. Epitope count against RMSD (shape) distance

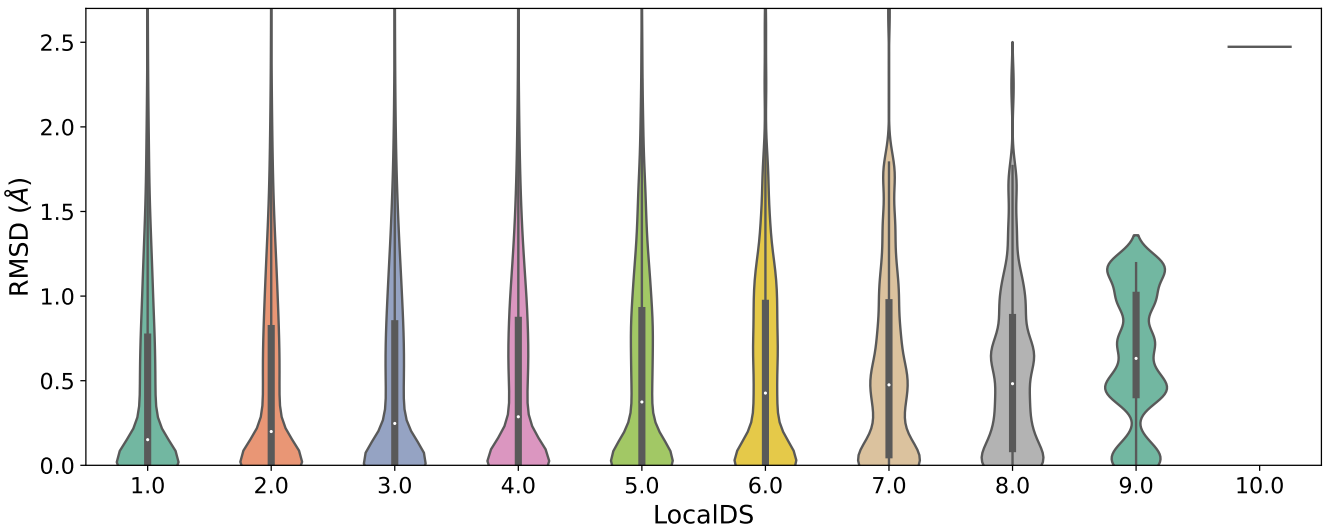


Figure 3. Distribution of shape distance metric RMSD against LocalDS for the viral VDJdb dataset.

We used AVIB, a current state-of-the-art model for TCR-peptide interaction prediction, to evaluate how the performance is affected by training on different data splits (Grazioli et al., 2022a). We trained AVIB on the RS, HS and DS using metrics LocalDS and RMSD at thresholds $0-33$, $33-66$ and $66-100$ percentiles for both LocalDS and RMSD. Each split was repeated with 5 different random seeds and the averaged results are shown in Figure 4. Overall, RS achieves the highest scores as expected, which in turn means the proposed algorithm produces non-trivial data-splits. For LocalDS, as the training-test set distance threshold is increased, values are within standard deviation of each others, meaning LocalDS does not seem an effective distance metric to use for this dataset. On the other hand, the RMSD split shows a much more expected behavior, with the performances steadily decreasing as the training-test set distance threshold is increased. Therefore, the use of the 3D information appears to be helpful in evaluating the generalization performance of AVIB, while controlling the difficulty of the task and it likely deserves further investigation in the future on a broader scope of datasets and methods.

4 DISCUSSION

The TCR-peptide/pMHC interaction prediction problem presents evaluation challenges (Grazioli et al., 2022b). When data is randomly split, models can achieve over-optimistic test performance, as test sequences can be observed at training time.

As a solution, Grazioli et al. (2022b) proposed the Hard Split (HS), which consists in randomly sampling test peptides and moving all instances of that peptide to the test set. The HS ensures that test peptides are never seen at training time, thus simulating what happens in the real world when new peptides arise from new pathogens.

This raises the question of whether it is possible to infer how well a model will generalize to an unseen peptide, by knowing how different the novel peptide is from those seen at training time.

We divided the peptides into different training and test sets using a sequence metric (LocalDS) and a shape metric (RMSD). To obtain the RMSD, we generated the 3D structures of the peptides using ESMFold and OmegaFold and then used PyMol to align and calculate the Root Mean Squared Distance (RMSD)

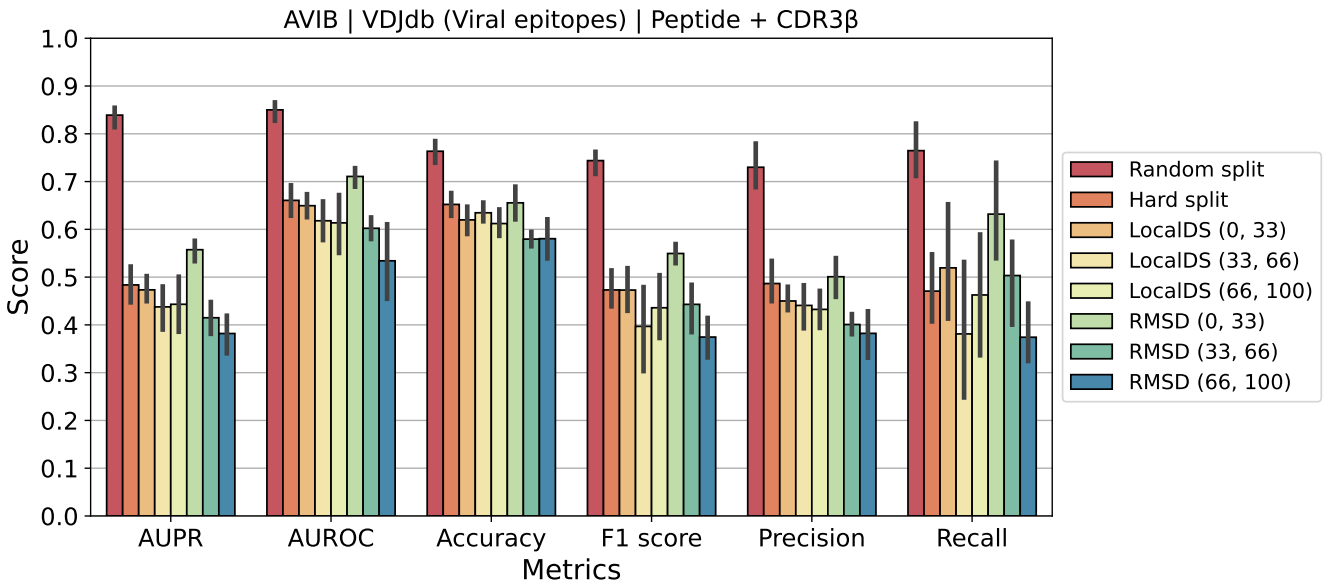


Figure 4. AVIB performance scores after training on different splits of Random Split, Hard Split and Distance Split with metrics LocalDS and RMSD. Mean and standard deviation were calculated for each of the 5 random seeds.

between the Cα atoms in the protein backbone. The LocalDS metric is a score obtained from the local alignment between two peptides sequences.

To simulate increased in difficulty, each test split was chosen to be at an increasing distance from the train/validation sets. Specifically, the test splits were chosen to cover an increasing range of percentiles: on average, $0 - 33$, $33 - 66$ and $66 - 100$.

After filling the test split, a small percentage of peptides that met the selected threshold remained unused. These peptides could be either outside the minumum/maximum pair count or unselected peptides because the budget was filled. To maintain budget ratios of the other splits and preserve sequence diversity, we recycle these peptides in the validation and train set.

Therefore, the increase in difficulty as the threshold increases can be attributed to two key factors. Firstly, the threshold selects test peptides that are located at a specific distance from the training and validation set. Secondly, due to this sampling bias, the training and validation sets become more similar because the test set contains the more distant peptides. The remaining peptides for train and validation therefore have a reduced sequence diversity.

Generally, we observe that sequence metrics tend to be similar as distance is increased. The structure metric RMSD, on the other hand, seems to be more sensitive to increases in distance, meaning that as distance is increased, the performance of AVIB is lower.

Coupled with the information in Figure 3, we see that even as LocalDS increases, the RMSD increases very slightly. This is further confirmed by the low but significant correlation coefficient.

Several factors could explain this. The 3D structure, as opposed to the amino acid sequence, is more relevant to describe protein-protein interactions. For example, certain angles in structures could make specific areas interact more than others. Related to this, some small differences in sequence metrics may result in larger structural differences, for example if two oppositely charged amino acids are placed next

to each other. The 3D structure and RMSD metric can therefore provide a stronger signal for binding prediction models and may be used for testing models performance in OOD settings or for the development of new models.

Obtaining a 3D structure of a protein often requires lengthy experiments to crystallize a protein which may take days, months, or years (Ballone et al., 2020). With the advent of sequence-to-shape models like AlphaFold (Jumper et al., 2021), OmegaFold (Wu et al., 2022) and ESMFold (Lin et al., 2022), this lengthy process can often be reduced to minutes or hours and makes it possible to obtain a fairly accurate 3D shapes of proteins. Structural validation in the lab of these models is currently underway, with some researchers reporting partial success, even when solving for previously unknown structures (Van Breugel et al., 2022). Successful applications of sequence-to-shape models in protein research include vaccine design (Goswami et al., 2022), binding affinity ranking (Chang and Perez, 2023), protein sequence design (Castorina et al., 2022; Watson et al., 2022) and benchmarking (Castorina et al., 2023).

The structures produced by sequence-to-shape models like the ones mentioned above, may be biased towards the training sequences in the Protein Data Bank (PDB), which is composed of larger structures compared to peptides, which are made up of just a few amino acids in length (Berman et al., 2003). Nevertheless, in this study, we use the shape as a representation of the real shape and assume that differences between two shapes would be consistent. This analysis could be extended to the full VDJdb dataset to test how the findings transfer to non-viral peptides.

Future work can explore the differences in structural distance between models such as AlphaFold, ESMFold and OmegaFold in the context of peptides, to further evaluate biases and divergences of structures generated by different sequence-to-shape models.

Sequence-to-shape models could also be used to predict full TCR complexes for binding prediction. In their current state, however, tools for complex modelling are computationally expensive and, although they may be comparable in speed with molecular dynamics simulations, it is currently unfeasible to test all possible combinations of TCR, peptides and pMHC for the binding prediction task.

The results suggest that the use of 3D shapes in the context of TCR-peptide/pMHC interaction prediction could help reduce the uncertainty about the generalization capabilities of ML models to unseen sequence. Given enough computational resources, 3D shapes could be predicted for the whole TCR structure, as well as for the MHCs presenting the peptide. This would enable the design of models that, given the individual structures of each input sequence, will predict more accurate binding interactions.

CONFLICT OF INTEREST STATEMENT

Authors AM, PM and FE are employees by NEC Laboratories Europe GmbH. LVC and FG were employed by NEC Laboratories Europe GmbH during the project.

AUTHOR CONTRIBUTIONS

All authors were involved in discussions, planning, brainstorming, and writing of the paper. FG and LVC both wrote and analysed the distance split algorithm. FG also trained the AVIB models on the new datasets. LVC produced the 3D structures of epitopes.

FUNDING

Funding for this project was provided by NEC Labs Europe GmbH.

ACKNOWLEDGMENTS

We would like to thank Timo Sztyler (NEC Labs Europe) and Martin Renqiang Min (NEC Labs America) for helpful discussions, especially at the inception of the project.

DATA AVAILABILITY STATEMENT

The VDJdb dataset used for this study can be found in https://vdjdb.cdr3.net. The AVIB model is available at: https://github.com/nec-research/vibtcr. The Distance Split code is available at: https://github.com/nec-research/peptide-distance-analysis

REFERENCES

Bagaev, D. V., Vroomans, R. M., Samir, J., Stervbo, U., Rius, C., Dolton, G., et al. (2020). Vdjdb in 2019: database extension, new analysis infrastructure and a t-cell receptor motif compendium. *Nucleic Acids Research* 48, D1057–D1062

Ballone, A., Lau, R. A., Zweipfenning, F. P. A., and Ottmann, C. (2020). A new soaking procedure for X-ray crystallographic structural determination of protein–peptide complexes. *Acta Crystallogr F Struct Biol Commun* 76, 501–507. doi:10.1107/S2053230X2001122X

Berman, H., Henrick, K., and Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. *Nat Struct Mol Biol* 10, 980–980. doi:10.1038/nsb1203-980

Castorina, L., Subr, K., and Wood, C. (2022). Timed-design: Efficient protein sequence design with deep learning. *Zenodo* 2022, 6997495

Castorina, L. V., Petrenas, R., Subr, K., and Wood, C. W. (2023). PDBench: Evaluating computational methods for protein-sequence design. *Bioinformatics* 39, btad027. doi:10.1093/bioinformatics/btad027

Chang, L. and Perez, A. (2023). Ranking Peptide Binders by Affinity with AlphaFold**. *Angewandte Chemie* 135. doi:10.1002/ange.202213362

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi:10.1093/bioinformatics/btp163

Deng, L., Ly, C., Abdollahi, S., Zhao, Y., Prinz, I., and Bonn, S. (2023). Performance comparison of TCR-pMHC prediction tools reveals a strong data dependency. *Frontiers in Immunology* 14. doi:10.3389/fimmu.2023.1128326

Eisenlohr, L. C., Huang, L., and Golovina, T. N. (2007). Rethinking peptide supply to MHC class i molecules. *Nature Reviews Immunology* 7, 403–410. doi:10.1038/nri2077

Gilbert, S. C. (2011). T-cell-inducing vaccines - what's the future. *Immunology* 135, 19–26. doi:10.1111/j.1365-2567.2011.03517.x

Goswami, A., Kumar, S. M., and Gore, M. M. (2022). De Novo *Design of Anti-Variant COVID Vaccine with T-cell Memory*. Preprint, Bioinformatics. doi:10.1101/2022.10.20.513049

Grazioli, F., Machart, P., Mösch, A., Li, K., Castorina, L., Pfeifer, N., et al. (2022a). Attentive Variational Information Bottleneck for TCR-peptide Interaction Prediction. *Bioinformatics* , btac820doi:10.1093/bioinformatics/btac820

Grazioli, F., Mösch, A., Machart, P., Li, K., Alqassem, I., O'Donnell, T. J., et al. (2022b). On TCR binding predictors failing to generalize to unseen peptides. *Front. Immunol.* 13, 1014256. doi:10.3389/fimmu.2022.1014256

He, Q., Jiang, X., Zhou, X., and Weng, J. (2019). Targeting cancers through TCR-peptide/MHC interactions. *Journal of Hematology and Oncology* 12. doi:10.1186/s13045-019-0812-8

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* 89, 10915–10919

Heslop, H. E. and Leen, A. M. (2013). T-cell therapy for viral infections. *Hematology* 2013, 342–347. doi:10.1182/asheducation-2013.1.342

Hewitt, E. W. (2003). The MHC class i antigen presentation pathway: strategies for viral immune evasion. *Immunology* 110, 163–169. doi:10.1046/j.1365-2567.2003.01738.x

Hossain, M. K., Hassanzadeganroudsari, M., and Apostolopoulos, V. (2021). The emergence of new strains of SARS-CoV-2. What does it mean for COVID-19 vaccines? *Expert Review of Vaccines* 20, 635–638. doi:10.1080/14760584.2021.1915140

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589. doi:10.1038/s41586-021-03819-2

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., et al. (2022). *Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction*. Preprint, Synthetic Biology. doi:10.1101/2022.07.20.500902

Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., et al. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology* 4. doi:10.1038/s42003-021-02610-3

Piepenbrink, K. H., Gloor, B. E., Armstrong, K. M., and Baker, B. M. (2009). Methods for Quantifying T cell Receptor Binding Affinities and Thermodynamics. In *Methods in Enzymology* (Elsevier), vol. 466. 359–381. doi:10.1016/S0076-6879(09)66015-8

Schrödinger, LLC (2015). The PyMOL molecular graphics system, version 1.8. PyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC.

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., et al. (2018). VDJdb: A curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research* 46, D419–D427. doi:10.1093/nar/gkx760

Smith-Garvin, J. E., Koretzky, G. A., and Jordan, M. S. (2009). T cell activation. *Annual Review of Immunology* 27, 591–619. doi:10.1146/annurev.immunol.021908.132706. PMID: 19132916

Springer, I., Tickotsky, N., and Louzoun, Y. (2021). Contribution of t cell receptor alpha and beta CDR3, MHC typing, v and j genes to peptide binding prediction. *Frontiers in Immunology* 12. doi:10.3389/fimmu.2021.664514

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. (2017). Mcpas-tcr: a manually curated catalogue of pathology-associated t cell receptor sequences. *Bioinformatics* 33, 2924–2929

Van Breugel, M., Rosa E Silva, I., and Andreeva, A. (2022). Structural validation and assessment of AlphaFold2 predictions for centrosomal and centriolar proteins and their complexes. *Commun Biol* 5, 312. doi:10.1038/s42003-022-03269-0

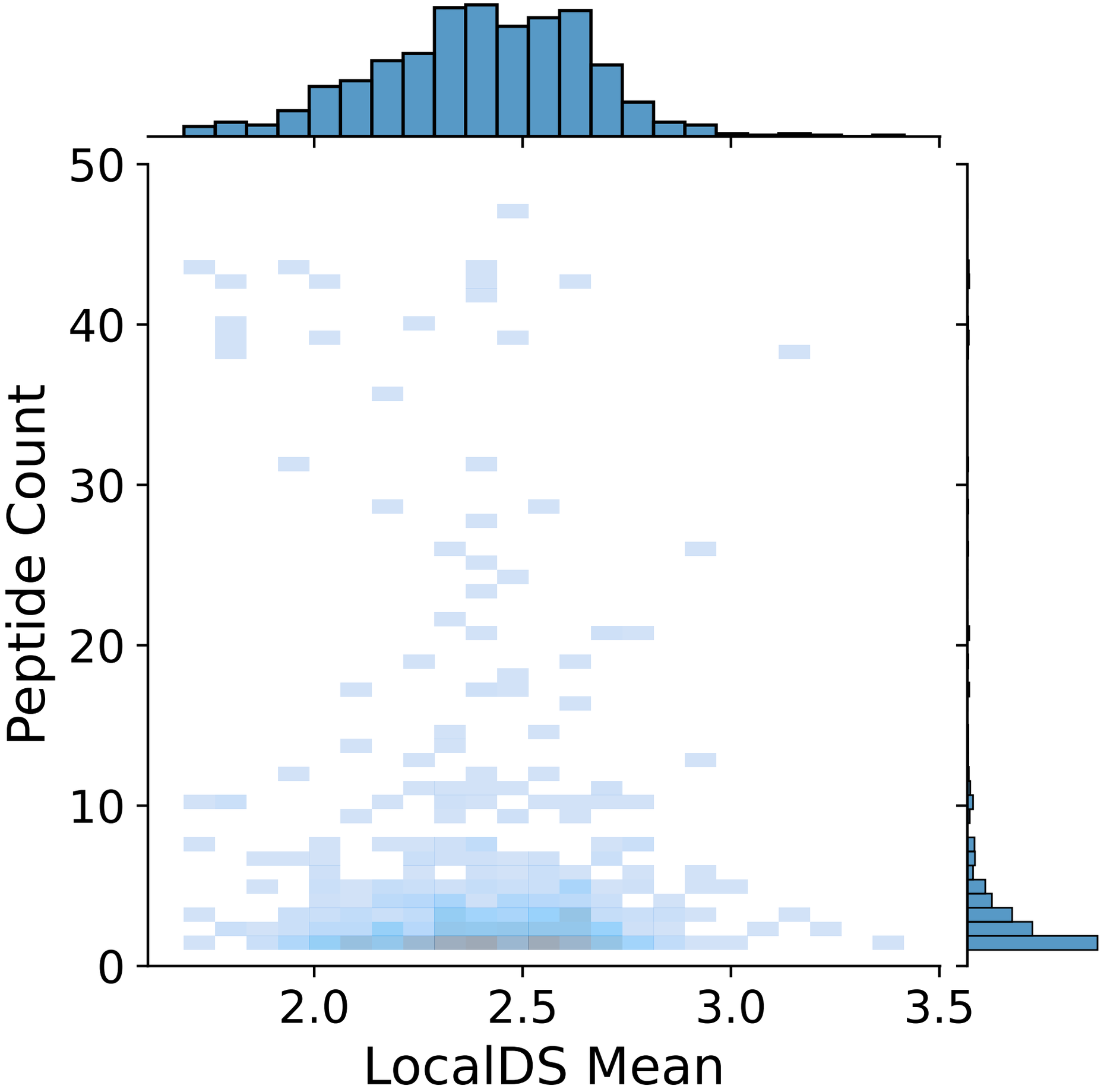
Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., et al. (2019). The immune epitope database (iedb): 2018 update. *Nucleic acids research* 47, D339–D343

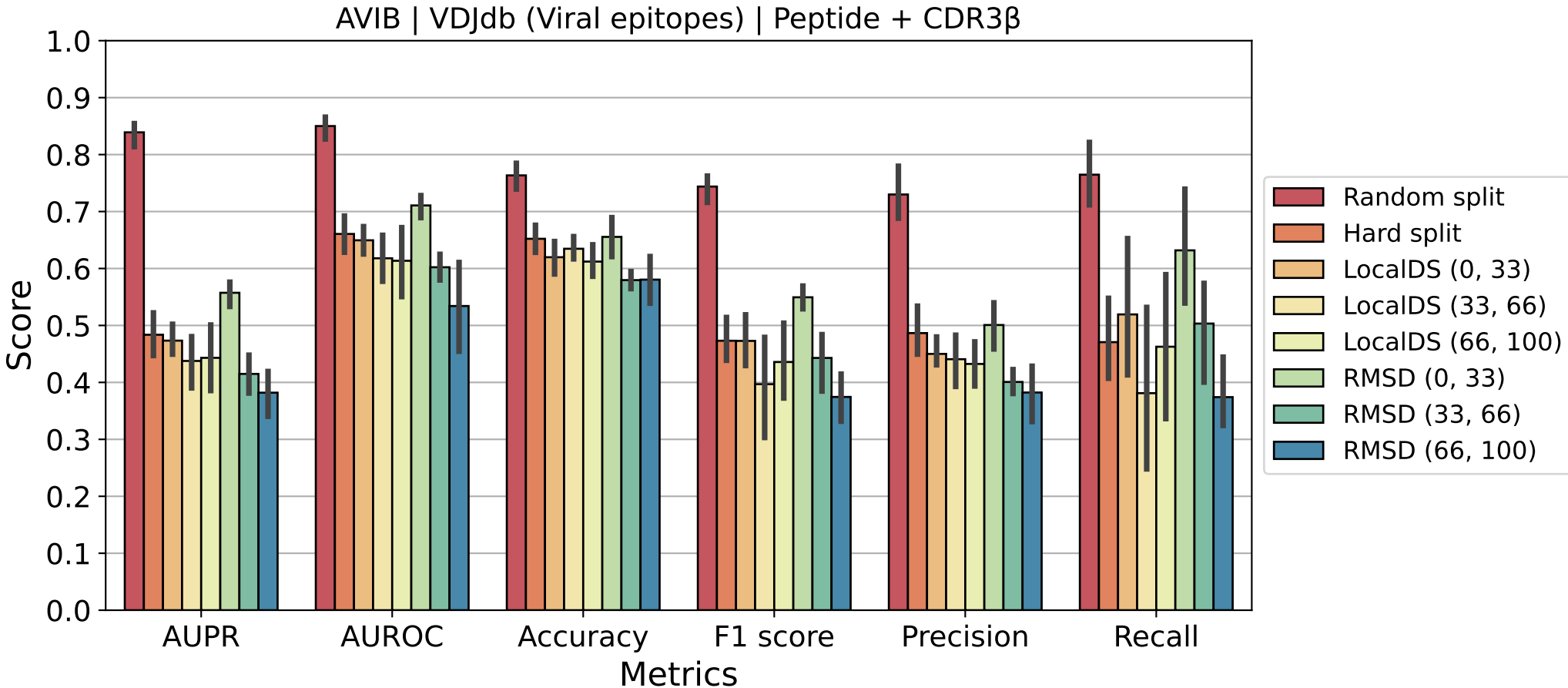
Waldman, A. D., Fritz, J. M., and Lenardo, M. J. (2020). A guide to cancer immunotherapy: From T cell basic science to clinical practice. *Nat Rev Immunol* 20, 651–668. doi:10.1038/s41577-020-0306-5

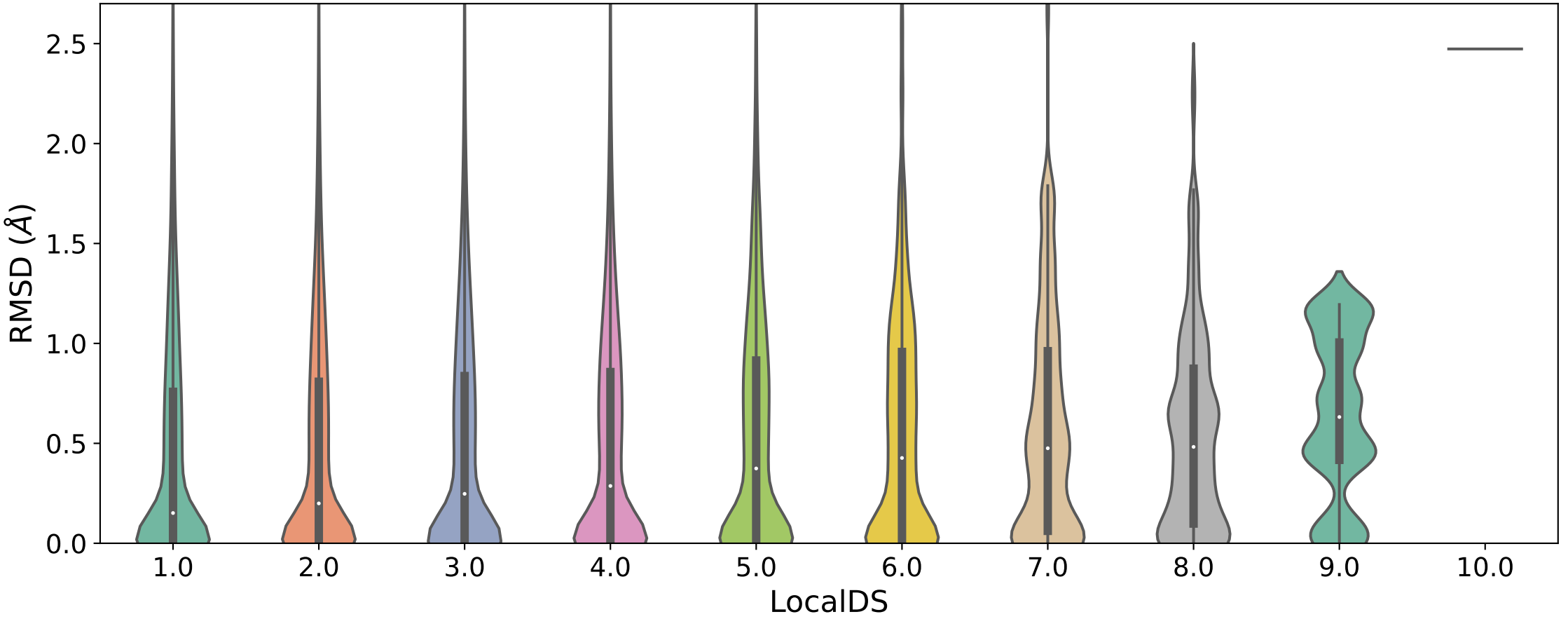
Watson, J. L., Juergens, D., Bennett, N. R., Trippe, B. L., Yim, J., Eisenach, H. E., et al. (2022). Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* doi:10.1101/2022.12.09.519842

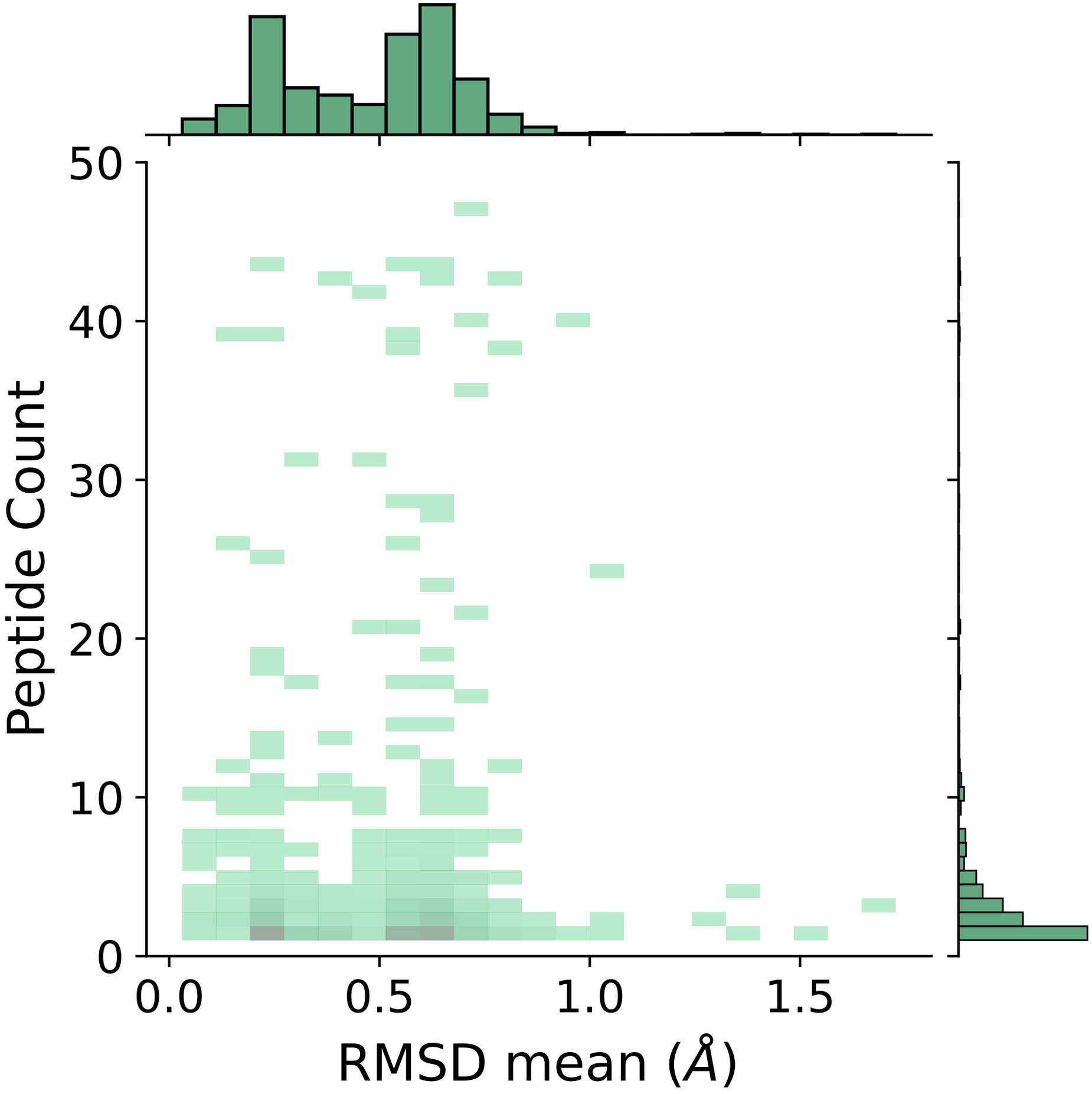
Weber, A., Born, J., and Martínez, M. R. (2021). TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* 37, i237–i244. doi:10.1093/bioinformatics/btab294

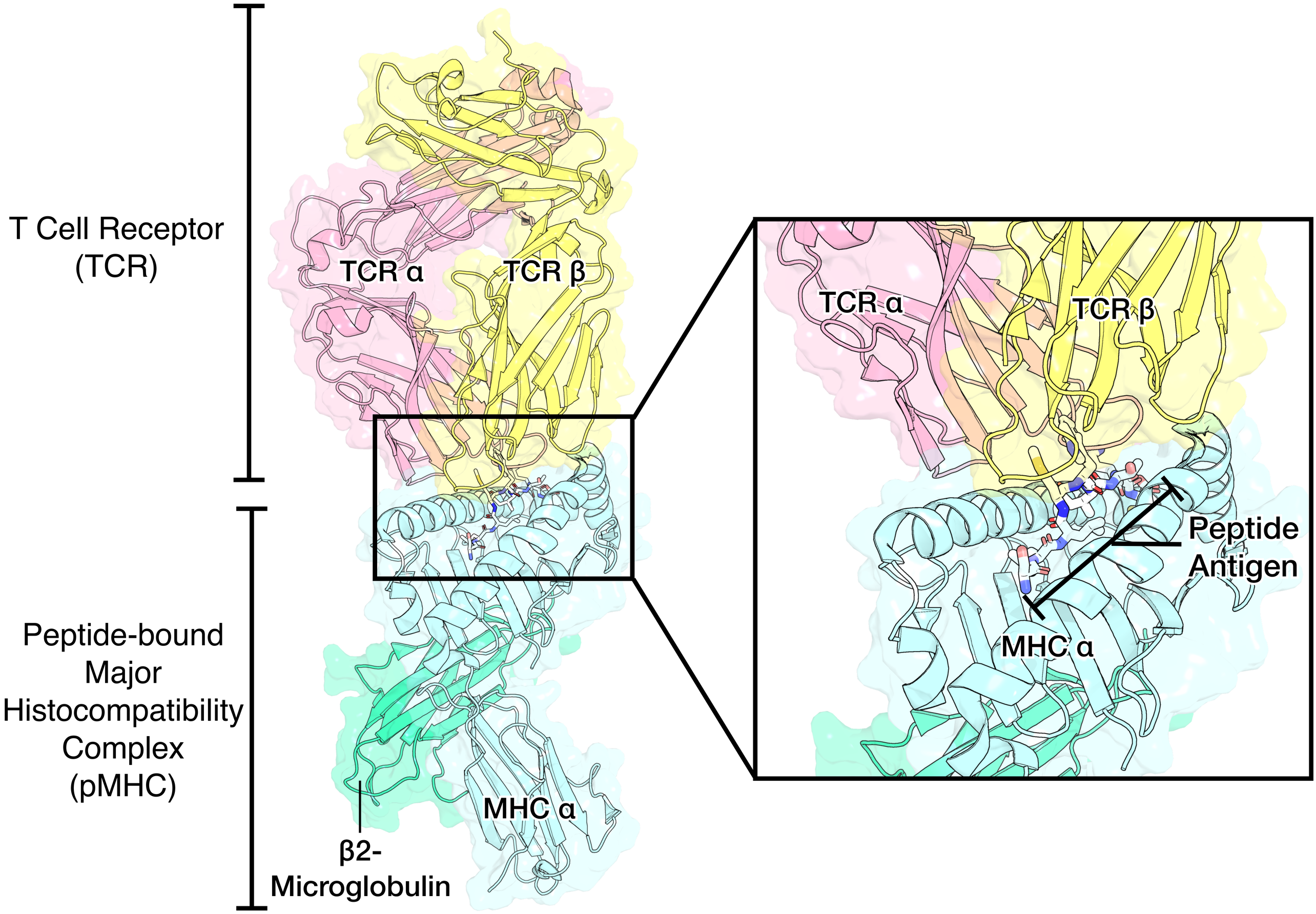
Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., et al. (2022). *High-Resolution de Novo Structure Prediction from Primary Sequence*. Preprint, Bioinformatics. doi:10.1101/2022.07.21.500999


Peptide Count
LocalDS Mean
0
10
20
30
40
50
2.0
2.5
3.0
3.5


AVIB | VDJdb (Viral epitopes) | Peptide + CDR3β
Score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
AUPR
AUROC
Accuracy
F1 score
Precision
Recall
Metrics
Random split
Hard split
LocalDS (0, 33)
LocalDS (33, 66)
LocalDS (66, 100)
RMSD (0, 33)
RMSD (33, 66)
RMSD (66, 100)


RMSD (Å)
2.5
2.0
1.5
1.0
0.5
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
10.0
LocalDS


Peptide Count
RMSD mean (Å)
0
10
20
30
40
50
0.0
0.5
1.0
1.5


T Cell Receptor
(TCR)
Peptide-bound
Major
Histocompatibility
Complex
(pMHC)
TCR α
TCR β
MHC α
β2-
Microglobulin
TCR α
TCR β
Peptide
Antigen
MHC α