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

Understanding the interaction between T Cell Receptors (TCRs) and peptide-bound Major Histocompatibility Complexes (pMHCs) is crucial for comprehending immune responses and developing targeted immunotherapies. Recent machine learning (ML) models excel at TCR-pMHC binding for training sequences. However, they underperform with peptides outside this distribution, raising concerns about their applicability in therapeutic settings.

To address this issue, we evaluate the effect of the distance between training and testing peptide distributions on ML model risk assessments, using sequence-based and 3D structure-based distance metrics. In our analysis we use two state-of-the-art models for TCR-peptide binding prediction: Attentive Variational Information Bottleneck (AVIB) and NetTCR-2.0.

Our hypothesis posited that increased similarity between test and training peptides could lead to inflated estimates of the true generalization error. However, our results indicate that 3D structural similarity metrics, rather than sequence-based metrics, are a better predictor of the model's generalization performance.

These findings highlight the importance of using the 3D structure in benchmarking the performance of TCR-pMHC binding prediction models. Specifically, it helps identify sudden generalization drops as the distance between training and test data distributions increases. Consequently, we recommend using structure-based over sequence-based distance methods for more reliable and accurate evaluations in TCR-pMHC binding studies.

Keywords: TCR, peptide, MHC, binding prediction, interaction prediction, machine learning, TCR-pMHC, 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. The immune response depends on how well the system can distinguish between healthy and infected/aberrant cells. It does so by leveraging specific molecules on the cell surface. These molecules are called Major Histocompatibility Complexes (MHC). Class I or class II MHCs are presented depending on the cell type (Alberts et al., 2017; Rowen et al., 1996; Glanville et al., 2017). The proteasome, a complex of protease enzymes, degrades proteins in the cell, creating antigenic

peptides. The MHC binds one peptide and presents it on the cell surface forming the peptide-bound MHC (pMHC) (Hewitt, 2003).

In infected cells, presented peptides may be derived from viral proteins, whereas in healthy cells, peptides are derived from housekeeping proteins (Eisenlohr et al., 2007). T cells can recognize viral peptides in the pMHC by binding it with via the T Cell Receptors (TCRs), leading to an immune response (Hewitt, 2003). 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). TCR-pMHC binding primarily occurs at the Complementarity-Determining Regions (CDRs). The peptide recognition is primarily mediated by the CDR3. The CDR3-α is derived from the alleles of the V and J genes; the D gene, in addition to the V and J, is involved in shaping the sequence and structure of the CDR3-β, (Feng et al., 2007; Rossjohn et al., 2015). These alleles can be recombined extensively, ensuring a high TCR repertoire diversity and allowing for a broad T cell-based immune response (Qi et al., 2014). The exposure of a naive T cell to an antigen leads to its activation and to to the development of a memory T cell population with the same TCR. This allows for a long-lasting immune response (Jameson and Masopust, 2018; Omilusik and Goldrath, 2019). A visual overview of the TCR-pMHC complex is presented in the Supplementary Materials.

Immunotherapies targeting cancer or viral infections use binding specificity to activate T cells, enhancing the immune system response. Binding specificity ensures that T cell binding occurs uniquely on the target, to kill say cancerous or infected cells, and avoid healthy cells. Research in immunotherapy covers two main categories: 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).

Immunotherapy design requires an in-depth understanding of the biochemical interactions between TCRs and pMHCs to accurately stimulate the immune system. The development of computational models that predict whether TCRs and pMHCs interact would allow for a faster *in silico* screening of sequences and consequent design of TCR sequences that bind specific target pMHCs. Recent advancements in machine learning (ML) lead to the development of TCR binding predictors (Springer et al., 2021; Montemurro et al., 2021; Weber et al., 2021; Grazioli et al., 2022a). The data input of these models consists of tuples of short amino acid sequences such as (peptide, CDR3-β) or (peptide, CDR3-β, CDR3-α, MHC). The task is usually formalized as a binary classification between binding or non-binding.

Several computational studies on TCR binding prediction employ data from the Immune Epitope Database (IEDB) (Vita et al., 2019), VDJdb (Bagaev et al., 2020) and McPAS-TCR (Tickotsky et al., 2017). These databases mainly contain CDR3-β sequences, but often lack information on CDR3-α. Additionally, public TCR-pMHC interaction datasets are often limited in diversity and size. They also present sequence bias towards commonly studied viruses or binding/non-binding pairs bias from experiment setups. Furthermore, laboratory methods that validate the TCR-pMHC interactions, such as surface plasmon resonance, titration calorimetry and fluorescence anisotropy, are usually resource-intensive and time-consuming, hindering the creation of larger datasets (Piepenbrink et al., 2009).

Generally, machine learning (ML) TCR binding predictors achieve high test performance when evaluated on test sets originating from the same source as the training set. However, various studies (Grazioli et al., 2022b; Deng et al., 2023) have shown that these methods exhibit weak cross-dataset generalization, meaning, models performance are significantly lower when training and test samples come from different distributions. Furthermore, Grazioli et al. (2022b) investigated the effect of different splitting techniques on

the TChard dataset, which combines samples from IEDB, VDJdb and McPAS-TCR. 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) then showed 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 score between 70% and 80%. Nevertheless, using the RS, peptides and CDR3 sequences may appear in both the training and test sets, leading to inflated estimates of the real-world model generalization capabilities.

Hence, to approximately test models performance for unseen peptides, they propose an alternative splitting method, named 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 proposed by Grazioli et al. (2022b) aims at evaluating models' generalization abilities under the toughest setup possible. 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.

In practice, however, RS and HS represent two extremes of a wider spectrum of data splitting options. Assessing other splitting options can provide further insights on the robustness of ML models.

Our work is driven by the following research question: can we estimate the performance of ML models given an unseen test peptide and its distance from the training ones? We hypothesize that, as the difference in distributions between training and test sets increases, the prediction task becomes gradually 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. However, given a distance metric, we control the distances between sequences in the training and in the test sets. 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.

We show that smaller distance between the training and test peptides indeed correlates with the ability of a ML-based binding prediction model to generalize. This allows researchers to determine which kind of structures it is safe to apply the ML model to. Furthermore, our experiments highlight that 3D structure distances act as better predictor of performance than sequence-based distance metrics.

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 the small number of available samples. 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 shuffled the available (peptide, CDR3-β) pairs. This process is commonly employed in the literature (Jurtz et al., 2018; Grazioli et al., 2022a) and leverages the hypothesis that random pairs will most likely not bind. 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, the data consists of pairs of (peptide, CDR3-β) and a binding/non-binding label.

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 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, as described in the following section.

2.2 The Distance Split

Given a distance metric, the goal of the Distance Split (DS) algorithm is to create data splits of the available (peptide, CDR3-β) samples, enforcing a specified median distance between training and test (and validation) peptides. We hypothesized that, as the training-test distance increases, the generalization performance of the models worsens, as the test peptides are increasingly more different from the training set. In the following, the DS algorithm is described.

As distance metrics, we employed LocalDS, GlobalDS and RMSD. LocalDS is a local alignment algorithm from BioPython that produces a distance score between two peptides represented as a sequence (Cock et al., 2009). Similarly, GlobalDS computes a distance score based on global alignemnt. 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 between Cα of the two structures..

For each metric, we computed the pairwise distances between all possible pairs of 757 peptides, creating a symmetric distance matrix $\mathbf{M} \in \mathbb{R}^{757\times757}$. This resulted in $\mathbf{M}_{GlobalDS}$, $\mathbf{M}_{LocalDS}$ and $\mathbf{M}_{RMSD}$. For each

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

distance matrix, we then calculated the row-wise median $\mathbf{m}_{med} \in \mathbb{R}^{757}$. Each value of $\mathbf{m}_{med}$ corresponds to the median distance between a given peptide and all other peptides. We then selected three intervals over the cumulative distribution of the median distances of $\mathbf{m}_{med}$, defined by a lower and an upper bound (b_l, b_u), i.e., $(0, 33)$, $(33, 66)$ and $(66, 100)$. Given the median vector $\mathbf{m}_{med}$, for each considered interval, we computed the distances corresponding to the lower and upper percentiles, i.e., d_l and d_u, for b_l and b_u, respectively. We then filtered out all the peptides whose median distance falls outside the interval of interest. The sets of peptides that fall inside this percentile interval are then sampled for test and validation, as described in the following. For each training-validation-test split, we used the 80-10-10 ratio. Since the DS algorithm employs iterative sampling of the validation and test samples, based on the split ratio and the number of available data points, we computed a test and validation budget, i.e., the expected number of data points in the test and validation splits, respectively.

The sampling procedure of the test samples was operated considering the peptides that belong to the selected percentile interval. Iteratively, a peptide was randomly sampled from this set and all the related (peptide, CDR3-β) pairs that present the sampled peptide were added to the test set. The process was repeated until the test budget was filled. The peptides that were not sampled were then recycled back with the filtered-out peptides and then employed for the validation split, following the same iterative procedure. The remaining peptides were used for training.

During the sampling of the validation and test sets, after sampling a peptide, the related (peptide, CDR3-β) pairs were only added to the validation or test set if their respective pair count was in between a lower and an upper bound. This was meant to constrain the peptide diversity. 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.

Employing the DS, as well as the RS and HS, we generated training, test and validation splits of the available data points for all distance metrics, $\mathbf{M}_{GlobalDS}$, $\mathbf{M}_{LocalDS}$ and $\mathbf{M}_{RMSD}$.

2.3 TCR-Peptide Interaction Prediction Model

As our TCR-peptide interaction prediction model, we use two state-of-the-art models: NetTCR-2.0 Jurtz et al. (2018) and the Attentive Variational Information Bottleneck (AVIB) (Grazioli et al., 2022b), This model presents probabilistic encoders for the BLOSUM-encoded peptide and CDR3 sequences, implemented as 1D convolutional networks. The encoders estimate Gaussian posteriors over a latent space. The various posteriors are then combined using an attention mechanism, named Attention of Experts, that estimates a single multi-sequence posterior distribution, which is then employed by a decoder to estimate the binding probability. Although the model supports both the CDR3-α and CDR3-β sequences, we only employed the CDR3-β for our experiments. We performed a hyperparameter optimization on the HS. The resulting 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 scores were identical to each other in our dataset. Hence, from here on, we will exclusively consider LocalDS. Descriptive 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 approximately normally distributed. Finally, the RMSD has a bimodal distribution with peaks at ~ 0.25 and ~ 0.63 Å.

In Figure 3, we compare RMSD against LocalDS. We observe no correlation between the two metrics (PearsonR = 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
Peptide (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. Descriptive statistics of peptides counts, LocalDS and RMSD for the viral VDJdb dataset. LocalDS and GlobalDS are sequence distance metrics based on local and global alignment scores, respectively. RMSD refers to the average 3D distance between peptides.

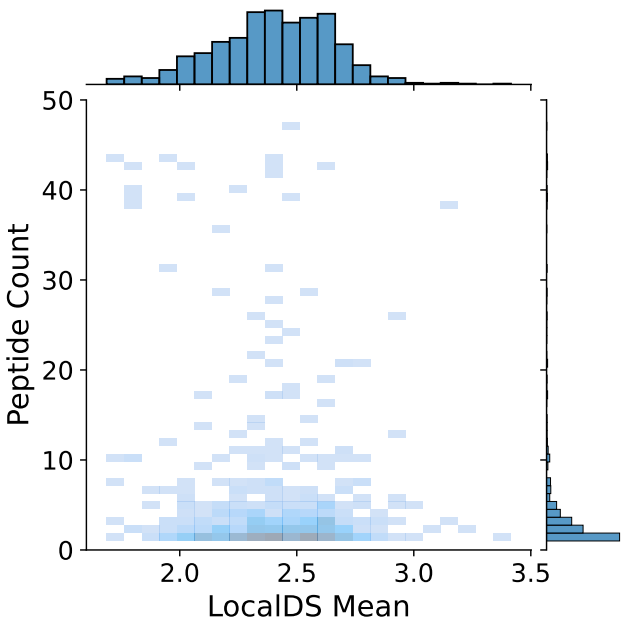

Figure 1. Distribution of peptide count against LocalDS (sequence) distance

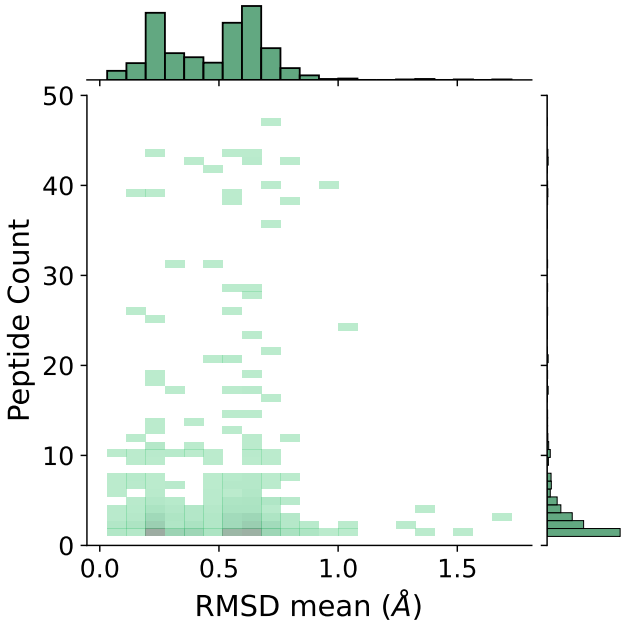

Figure 2. Distribution of peptide count against RMSD (3D shape) distance

For our experiments, two state-of-the-art deep learning models for TCR-peptide interaction prediction: NetTCR-2.0 Jurtz et al. (2018) and AVIB Grazioli et al. (2022a). We evaluated how their performance is affected by training and testing on different data splits. We trained and tested AVIB adopting the RS, HS and DS. For the DS, we used the LocalDS and RMSD distance metrics and the following lower and upper bound pairs over the cumulative distribution of the median distances: $(0, 33)$, $(33, 66)$ and $(66, 100)$. Each split was repeated with 5 different random seeds and the averaged results are shown in Figure 4.

Overall, as shown in 4 and Supplementary Figure 2, both models perform similarly in the splits, though AVIB achieves slightly higher scores. As expected, models perform best in the RS, as peptides in the train

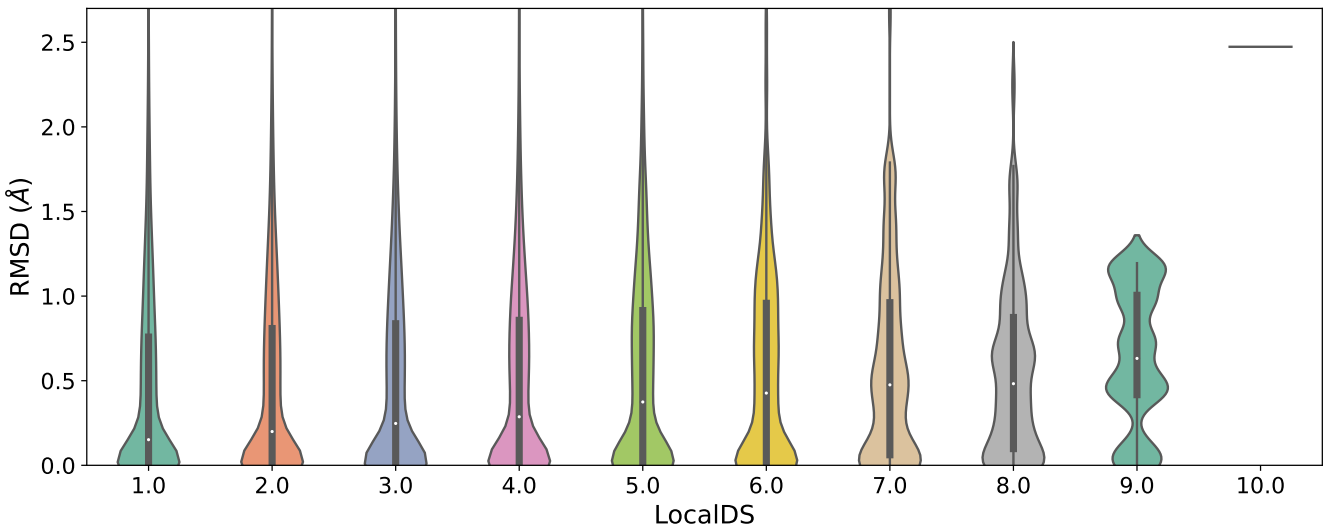

Figure 3. Distribution of RMSD (3D shape) distance metric against LocalDS (sequence) distance for the viral VDJdb dataset.

set also appear in the test set. On the other hand, performance on the HS is significantly lower compared to the RS.

Regarding the DS, for LocalDS, as the training-test median distance is increased, test scores do not show significance differences for both models, suggesting that the LocalDS does not seem an effective distance metric to use for this dataset. On the other hand, for the RMSD-based DS the performances steadily decrease as the median training-test set distance 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.

For the RMSD DS, both models perform worse than the HS for the more distant ranges of DS $(33, 66)$ and $(66, 100)$). In the case of the HS, the test set may contain epitopes similar to the training set, thus making the task slightly easier. On the other hand, in the extreme range of $(66, 100)$), the DS algorithm ensures that the test set is at a consistent distance.

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, e.g., 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

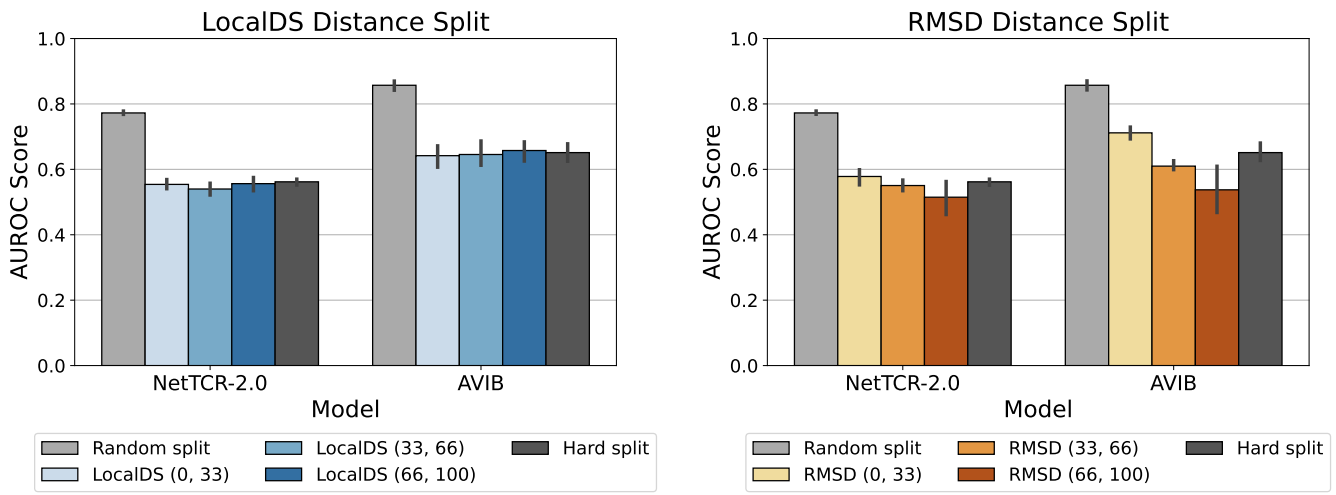

Figure 4. AUROC Scores for TCR-pMHC binding prediction of the AVIB and NetTCR-2.0 trained and tested using various splitting techniques. The commonly used *Random Split* allows test peptides to be observed also during training. In the *Hard Split*, peptides are exclusively allocated to either the training or test set. In the *Distance Split*, we enforce specific median distances between the training and test peptides, using either sequence distance (LocalDS) or shape distance (RMSD). The training-test distance between peptides is controlled by selecting three percentile intervals over the cumulative median peptide-peptide distance distribution: $(0, 33)$, $(33, 66)$ and $(66, 100)$. Mean and standard deviation are calculated over 5 repeated experiments with different random seeds. Full set of metrics for each model is available in Supplementary Figure 2

and OmegaFold and then used PyMol to align and calculate the Root Mean Squared Distance (RMSD) 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 generalization difficulty in a controlled way, we proposed the *Distance Split* (DS). Using DS, each test split was chosen in a way that guarantees a given median distance between the training and test peptides. Specifically, the test peptides were chosen by sampling from a subset of peptides, whose median distance from the other peptides is between a lower and an upper bound. In our experiments, we considered the following intervals over the cumulative median distance distribution: $(0, 33)$, $(33, 66)$ and $(66, 100)$.

Generally, we observe that increasing training-test LocalDS does not lead to worse generalization performance. The structure metric RMSD, on the other hand, seems to be a more promising metric for this task. In fact, as the training-test median distance is increased, the performance of both models tends to degrade.

As shown in Figure 3, we observe no correlation between RMSD and LocalDS. 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 out-of-distribution settings, or for the development of new models. Interestingly, we observed that AVIB's performance is slightly lower for DS in the ranges $(33, 66)$ and $(66, 100)$ compared to that of the HS. The HS ensures that both test and train peptides are *unique*.

However, even with this uniqueness, test and train peptides might still possess a certain degree of similarity. This similarity provides the model with additional chemical information about the peptide. In contrast, DS enforces a threshold distance between the train and test sets, thereby increasing the difficulty.

Obtaining a 3D structure of a protein often requires lengthy experiments to crystallize it, 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.

Our 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 sequences. 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 and trained the AVIB models on the new datasets. LVC produced the 3D structures of epitopes and trained NetTCR-2.0.

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 NetTCR-2.0 is available at: . The Distance Split code is available at: https://github.com/nec-research/peptide-distance-analysis

REFERENCES

Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., et al. (2017). *Molecular biology of the cell* (WW Norton & Company)

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

Feng, D., Bond, C. J., Ely, L. K., Maynard, J., and Garcia, K. C. (2007). Structural evidence for a germline-encoded t cell receptor–major histocompatibility complex interaction'codon'. *Nature immunology* 8, 975–983

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

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., et al. (2017). Identifying specificity groups in the t cell receptor repertoire. *Nature* 547, 94–98

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

Jameson, S. C. and Masopust, D. (2018). Understanding Subset Diversity in T Cell Memory. *Immunity* 48, 214–226. doi:10.1016/j.immuni.2018.02.010

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

Jurtz, V. I., Jessen, L. E., Bentzen, A. K., Jespersen, M. C., Mahajan, S., Vita, R., et al. (2018). Nettcr: sequence-based prediction of tcr binding to peptide-mhc complexes using convolutional neural networks. *BioRxiv* , 433706

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

Omilusik, K. D. and Goldrath, A. W. (2019). Remembering to remember: T cell memory maintenance and plasticity. *Current Opinion in Immunology* 58, 89–97. doi:10.1016/j.coi.2019.04.009

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

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., et al. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences* 111, 13139–13144. doi:10.1073/pnas.1409155111. Number: 36

Rossjohn, J., Gras, S., Miles, J. J., Turner, S. J., Godfrey, D. I., and McCluskey, J. (2015). T cell antigen receptor recognition of antigen-presenting molecules. *Annual review of immunology* 33, 169–200

Rowen, L., Koop, B. F., and Hood, L. (1996). The complete 685-kilobase dna sequence of the human β t cell receptor locus. *Science* 272, 1755–1762

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