frontiers

TCR Binding Predictors Fail to Generalize to Unseen Peptides

Filippo Grazioli [1,*], **Anja Mösch** [1], **Pierre Machart** [1], **Kai Li** [2], **Israa Alqassem** [1], **Timothy J. O'Donnell** [3], **Martin Renqiang Min** [2]

[1] *NEC Laboratories Europe, Heidelberg, Germany*
[2] *NEC Laboratories America, Princeton, NJ 08540, USA*
[3] *Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA*

Correspondence*:
Corresponding Author
filippo.grazioli@neclab.eu

ABSTRACT

Several recent studies investigate TCR-peptide/-pMHC binding prediction using machine or deep learning approaches. Many of these methods achieve impressive results on test sets which include peptide sequences that are also included in the training set. In this work, we investigate how state-of-the-art deep learning models for TCR-peptide/-pMHC binding prediction generalize to unseen peptides. We create a dataset called *TChard*, which include positive samples from IEDB, VDJdb, McPAS-TCR and the MIRA set, as well as negative samples from both randomization and 10X Genomics assays. We propose the *hard split*, a simple heuristic for training/test split, which ensures that test samples exclusively present peptides that do not belong to the training set. We investigate the effect of different training/test splitting techniques on the models' test performance, as well as the effect of training and testing the models using mismatched negative samples generated randomly, in addition to the negative samples derived from assays. Our results show that modern deep learning methods fail to generalize to unseen peptides. We provide an explanation why this happens and verify our hypothesis on the *TChard* dataset. We then conclude that robust prediction of TCR recognition is still far for being solved.

Keywords: TCR, peptide, MHC, binding prediction, interaction prediction, machine learning

1 INTRODUCTION

Studying T cell receptors (TCRs) has become an integral part of cancer immunotherapy and human infectious disease research (1, 2, 3, 4). TCRs are able to identify intra-cellular processed peptides originating from infected or aberrant cells. TCRs are heterodimers consisting of an α- and a β-chain, which bind to peptides presented on the cell surface by either major histocompatibility complex (MHC) class I or class II molecules, depending on the cell type (5, 6, 7). The binding of the TCR to the peptide-MHC (pMHC) complex occurs at the complementarity-determining region 3 (CDR3). The CDR3α consists of alleles from the V and J genes; for the CDR3β, the D gene is additionally involved (8, 9). These alleles can be recombined unboundedly, which results in a high TCR repertoire diversity, essential for a broad T cell-based immune response (10). When a naive TCR is exposed to an antigen and activated for the first time, a memory T cell population with this TCR may develop, which enables a long-lasting immune response (11, 12).

Numerous recent studies investigate TCR-peptide/-pMHC binding prediction by applying different machine or deep learning methods (13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24). Many of these studies use data from the Immune Epitope Database (IEDB) (25), VDJdb (26) and McPAS-TCR (27), which mainly contain CDR3β data and lack information on CDR3α. Such methods achieve high test performance when evaluated on test sets that belong to the same source as the training set. However, we show that these methods exhibit weak cross-dataset generalization, i.e., the models suffer from severe performance degradation when tested on a different dataset. For example, as shown in Figure S1, several machine learning models trained on McPAS-TCR perform poorly on VDJdb.

In this work, in order to evaluate the relevance of the available data for deep-learning-based TCR-peptide/-pMHC binding prediction, we aggregate binding samples obtained from IEDB, VDJdb and McPAS-TCR. We additionally consider a set of binding samples from (28); we refer to it as MIRA set. Non-binding data points are collected from IEDB, as well as from the 10X Genomics samples provided in the NetTCR-2.0 repository (22). A simple analysis of the class distribution (binding versus non-binding) of the resulting data points reveals that all TCR sequences exclusively appear in either binding or non-binding TCR-peptide/-pMHC pairs; no CDR3 sequence is observed in both positive and negative samples (see Figure 1C). Machine learning models trained naively on data with this class distribution are prone to learning undesirable inductive biases. In fact, our results in Section 4.1 suggests they tend to classify samples only as a function of the CDR3 sequences, which could be memorized.

For unbiased evaluation, we propose a new integrated dataset, which we name *TChard*. This dataset consists in the integration of the aforementioned samples. Additionally, to reduce the bias associated with CDR3 sequence memorization, we randomly recombine the TCRs and pMHCs of the available samples and create randomized mismatched negative samples, as described in Section The *TChard* dataset. To the best of our knowledge, this dataset constitutes the largest set of TCR-peptide/-pMHC samples available at the time this work is being written.

We perform deep learning experiments using two state-of-the-art models for TCR-peptide/-pMHC interaction prediction: ERGO II (23) and NetTCR-2.0 (22). ERGO II is a deep learning approach which adopts long short-term memory (LSTM) networks and autoencoders to compute representations of peptides and CDR3s. It can also handle additional input modalities, i.e., V and J genes, MHC and T cell type. NetTCR-2.0 employs a simple 1D CNN-based model, integrating peptide and CDR3 sequence information for the prediction of TCR-peptide specificity. Both models input peptide and CDR3s representations in the form of amino acid sequences.

We perform experiments on *TChard* and investigate the effect of different training/test splitting strategies. In contrast to previous works (23, 22), we place special emphasis on testing the models on unseen peptides. We propose the *hard split*, a splitting heuristic meant to create test sets which only contain unseen peptides, i.e. not included in the training set. In the context of neoantigen-based cancer vaccines development, neoepitopes exhibit enormous variability in their amino acids sequences; employing predictive models for TCR recognition in the development of neoantigen cancer vaccines requires robust generalization on unseen peptides. We show that evaluating the models' performance on unseen peptides leads to poor generalization.

2 THE *TCHARD* DATASET

In this section, we describe the creation of the *TChard* dataset. All samples in *TChard* include a peptide and a CDR3β sequence, associated with a binary binding label. A subset of these samples may additionally

have: (i) CDR3α sequence information, and/or (ii) allele information of the MHC (class I or II) in complex with peptides. A sample consists therefore of a tuple of molecules (from 2 to 4). When available, the V and J alleles for the α-chain and the V, D, J alleles for the β-chain are also included. We refer to the binding tuples as *positive* and to the non-binding ones as *negative*.

2.1 Dataset creation

First, we collect positive assays from the IEDB, VDJdb and McPAS-TCR databases. Additionally, we include the binding samples from the MIRA set (28), which is publicly available in the NetTCR-2.0 repository [1].

Second, we include negative assays, i.e. non-binding tuples of molecules extracted from IEDB. Additionally, a set of negative samples extracted from the NetTCR-2.0 repository is considered; this is derived from 10X Genomics assays described by Montemurro et al. (22). In this work, we refer to the negative tuples derived from negative assays as the NA set.

Third, in order to remove a small set of outliers, we operate a filtration over the length of the amino acid sequences of peptides, CDR3α and CDR3β. We only keep samples with peptide sequence length smaller than 16, CDR3α sequence length between 7 and 21 and CDR3β sequence length between 9 and 23. These filtration steps are meant to exclude a small portion of data points which present consistently longer amino acid sequences. Including them in the dataset would imply to extend the magnitude of the padding required by NetTCR-2.0 by a large margin, making computation more expensive.

Fourth, we generate negative samples via random recombination of the sequences found in the positive tuples. Building from the positive samples, we associate the peptides or pMHC complexes (when MHC allele information is available) with CDR3α and CDR3β sequences randomly sampled from the dataset, as operated in previous studies (23). We sample twice as many mismatched negative samples as there are positive ones. We discard randomly generated samples which share at least the same $(peptide, CDR3\beta)$ with any positive sample. In this work, we refer to the randomized negative tuples as the RN set.

2.2 Description of the data distributions

The full dataset - i.e. considering negative samples from both NA and RN - presents:

- 528,020 unique $(peptide, CDR3\beta)$ tuples, 385,776 of which are negative and 142,244 are positive;
- 400,397 unique $(peptide, CDR3\beta, MHC)$ tuples, 300,168 of which are negative and 100,229 are positive;
- 111,041 unique $(peptide, CDR3\beta, CDR3\alpha)$ tuples, 82,631 of which are negative and 28,410 are positive;
- 110,266 unique $(peptide, CDR3\beta, CDR3\alpha, MHC)$ tuples, 82,037 of which are negative and 28,229 are positive.

The dataset statistics considering negative samples derived from either RN or NA are presented in Table S1. Figure 1 depicts the class distribution for $(peptide, CDR3\beta, CDR3\alpha)$ samples. Analogously, Figure S2, Figure S3 and Figure S4 depict the class distribution for $(peptide, CDR3\beta)$, $(peptide, CDR3\beta, MHC)$ and $(peptide, CDR3\beta, CDR3\alpha, MHC)$ samples, respectively. Figure S5 depicts the length distribution for all sequences.

[1] https://github.com/mnielLab/NetTCR-2.0/tree/main/data

3 PREDICTING TCR RECOGNITION WITH DEEP LEARNING

We perform experiments on the *TChard* dataset with two publicly available state-of-the-art deep learning methods for TCR-peptide/-pMHC interaction prediction: ERGO II and NetTCR-2.0 [2].

We operate TCR-peptide interaction prediction considering peptide and CDR3β, as well as TCR-pMHC interaction prediction considering peptide, CDR3β, CDR3α and MHC. NetTCR-2.0 is not explicitly designed to account for MHC information; we circumvent this shortcoming by concatenating the MHC pseudo-sequence [3] to the other input amino acid sequences and perform BLOSUM50 encoding (30). We do not make distinctions between class I and II MHCs and train a single model for both types.

3.1 Random and Hard Training/Test Splits

For performance evaluation, we investigate two different strategies for training/test splits.

Random split (RS). Given a training/test ratio (80/20 in this work), this procedure consists in sampling test samples uniformly from the dataset without replacement until the desired budget is filled. The remaining samples constitute the training set. In this work, we refer to RS(RN), when the negative tuples only belong to the RN set, to RS(NA), when the negative tuples only belong to the NA set, and to RS(RN+NA), when all negative samples are considered.

The nature of TCR recognition is combinatorial. In our dataset, although a given tuple of molecules is only observed once, a given peptide can appear multiple times, paired with different CDR3β, CDR3α or MHC. Using a random training/test split ensures that test tuples are not observed at training time. However, this can lead to testing the model on peptides, MHCs, or CDR3β and CDR3α sequences that were already observed at training time in combination with different sequences. Our result show that this can lead to over-optimistic estimates of machine learning models' real-world performance. To enable neoantigen-based cancer vaccines and T-cell therapy, it is fundamental to test the model on sequences which were never observed at training time. This allows to provide rigorous, unbiased estimates of the model's performance. Neoantigens display in fact enormous variability in their amino acids sequence; to identify the most immunogenic vaccine elements, we need models that generalize to unseen sequences.

Hard split (HS). We propose a simple heuristic, which we refer to as *hard split*. Considering the whole dataset consisting in a set of tuples, we first select a *minimum* training/test ratio (85/15 in this work). Let $\mathcal{P}_{l,u}$ be the set of all peptides that are observed in at least l tuples but no more than u tuples in our dataset. We randomly sample a peptide from $\mathcal{P}_{l,u}$ without replacement. All tuples which include that peptide are assigned to the test set. If the current number of test samples is smaller than the budget defined by the training/test ratio, the sampling from $\mathcal{P}_{l,u}$ is repeated.

This heuristic ensures that the peptides which belong to the test set are not observed by the model at training time. For the $(peptide, CDR3\beta)$ tuples, we set l and u to 500 and 10000, respectively. For the $(peptide, CDR3\beta, CDR3\alpha, MHC)$ tuples, we set l and u to 100 and 5000, respectively. The l parameter is a lower bound and ensures that the selected test peptides are paired with a sufficiently broad variety of CDR3 sequences. The u parameter i an upper bound and allows to exclude test peptides that can too quickly saturate the test budged, hence reducing the variety of test peptides. We create 5 different hard splits, using 5 different random seeds for the sampling of the test peptides. For the creation of the hard

[2] ERGO II: `https://github.com/IdoSpringer/ERGO-II`; NetTCR-2.0: `https://github.com/mnielLab/NetTCR-2.0`

[3] Taken from the PUFFIN (29) repository: `https://github.com/gifford-lab/PUFFIN/blob/master/data/`

training/test splits, we consider all positive samples, as well as the negative samples from the RN set, i.e. excluding the negative samples from the negative assays. We refer to this type of split as HS(RN).

Table S2 describes the 5 HS(RN) hard splits for the $(peptide, CDR3\beta)$ samples. It presents the lists of test peptides and the number of positive and negative samples associated with each of them. Table S3 describes the 5 hard splits which we provide for the $(peptide, CDR3\beta, CDR3\alpha, MHC)$ samples.

3.2 Validation Approach and Performance Evaluation

For more robust performance evaluation, we repeat the experiments for each different training/test split (i.e. 5 times). The area under the receiver operator characteristic (AUROC) curve (31, 32), the area under the precision-recall (AUPR) curve (33, 34), the F1 score (F1) (35), as well as precision, recall and classification accuracy are computed on the test sets and averaged.

We adopt the default configuration for both ERGO II and NetTCR-2.0, as proposed in their original implementations. For ERGO II, we adopt the LSTM amino acid sequences encoder. The training is performed for a maximum of 1000 epochs and, in order to avoid over-fitting, the best model is selected by saving the weights corresponding to the epoch where the AUROC is maximum on the validation set. The validation set is obtained via 80/20 stratified random split of the training set.

3.3 Training/Test Splitting Strategies in Related Works

In this section, we describe how ERGO II and NetTCR-2.0 perform training/test splitting and how this differs from our approach.

Springer et al. (23) propose four different settings. In the Single Peptide Binding (SPB) setting, it is tested whether an unknown TCR binds to a predefined target peptide; at training time, TCRs which are known to bind to that peptide are employed. In the TCR-Peptide Pairing I (TPP-I) setting, which is comparable to our RS, test peptides and TCRs can be observed at training time. In the TCR-Peptide Pairing II (TPP-II) setting, test TCRs cannot be observed at training time, but peptides can. In the TCR-Peptide Pairing III (TPP-III) setting, it is ensured that both test TCRs and test peptides are unseen, i.e. not included in training tuples. Mismatched negative samples are derived from a randomization heuristic, analogous to how we construct the RN set in this work.

Montemurro et al. (22) compute the peptide-specific Levenshtein distance among CDR3s. Using the Hobohm 1 algorithm (36), redundancies among the CDR3s are removed. Five partitions are created to allow cross-validation. Single-linkage clustering of the redundancy-reduced positive training data is performed for partitioning and negative samples from 10X Genomics and randomization are added. For evaluating the model, test data points are separated from the training data by a given Levenshtein similarity threshold, i.e. samples with similarities to the training data above this threshold were removed. In contrast to our work, Montemurro et al. (22) do not investigate generalization on unseen peptides.

4 RESULTS

Figure 2 shows test results for ERGO II and NetTCR-2.0, for the RS and HS splitting strategies, in both the peptide+CDR3β and the peptide+CDR3β+CDR3α+MHC settings. We perform experiments considering negative samples from the NA set only, from the RN set only and jointly from both the NA and RN sets.

4.1 Over-optimistic Classification Performance due to Sequence Memorization

As depicted in Figure 2, almost perfect classification is achieved when training with negative samples only from the NA set and testing using the RS(NA) split. As shown in Figure S2C and Figure S4C, when considering negative samples from the NA set only, the binding and non-binding class histograms of the CDR3 sequences are disjoint. Hence, models can learn to correctly map a large portion of test tuples to the correct label simply by memorizing the CDR3 sequences, ignoring the peptide. We believe these results are over-optimistic and should not be considered as the approximation of these models' real-world performance.

4.2 The Hard Split allows for Realistic Evaluation

Using the HS heuristic appears to make prediction on the test set consistently harder, if not impossible. This tendency is observed in the peptide+CDR3β setting (Figure 2A and Figure 2B) and in the peptide+CDR3β+CDR3α+MHC setting (Figure 2C and Figure 2D). In the peptide+CDR3β setting, when testing the models using the HS(RN) split, the predictions on the test set barely exceed random-level performance, i.e. almost no generalization to unseen peptides is occurring (AUROC ≈ 0.55). This phenomenon is observed when the models are trained using negative samples from the RN set only, as well as when using negative samples from both the RN and NA sets.

The effect of including negative samples from NA at training time does not significantly influence test performance when the HS is adopted. Conversely, when RS is performed, significant differences are caused by the utilization of the negative samples from NA. This reinforces our claims regarding sequence memorization. ERGO II, in the peptide+CDR3β setting (Figure 2A), achieves over-optimistic performance when the negative samples come from both NA and RN and testing is operated using RS(RN+NA). The same phenomenon is observed in Figure 2B for ERGO II in the peptide+CDR3β+CDR3α+MHC setting and in Figure 2D for NetTCR-2.0 in the peptide+CDR3β+CDR3α+MHC setting.

Figure S6 depicts NetTCR-2.0 results on the $(peptide, CDR3\beta, CDR3\alpha, MHC)$ samples, but ignoring the MHC; we report these results for fairness, as NetTCR-2.0 is not originally designed to handle MHC pseudo-sequences.

5 DISCUSSION

In this work, we aim to test the reliability of state-of-the-art deep learning methods on TCR-peptide/-pMHC binding prediction for unseen peptides. To this purpose, we introduce the *TChard* dataset, which integrates TCR-peptide/-pMHC samples from the IEDB, McPAS-TCR and VDJdb databases, as well as the MIRA set and the 10X Genomics samples from the NetTCR-2.0 repository.

We perform experiments with two state-of-the-art deep learning models for TCR-peptide/-pMHC interaction prediction, ERGO II and NetTCR-2.0. We study the peptide+CDR3β and the peptide+CDR3β+CDR3α+MHC settings. We compare the effect of different training/test splitting strategies, RS and HS. RS is a naive random split, while HS allows to test the models on unseen peptides. We investigate the effect of training and testing the models using mismatched negative samples generated randomly (RN), in addition to the negative samples derived from assays (NA).

As shown in our experiments, when the HS is performed, the two models do not generalize to unseen peptides; this appears to be in contrast to the TPP-III results presented by Springer et al. (23). Conversely, when a simple RS is employed and negative samples only belong to NA, almost perfect classification is

achieved. We believe this phenomenon is due to the class distribution of the CDR3 sequences, and the related sequence memorization. As shown in Figure 1C, when considering negative samples from NA only, the positive and the negative samples are completely disjoint. Hence, a given CDR3 sequence is only presented in either binding or non-binding samples. This leads to learning an inductive bias which classifies tuples as binding or non-binding exclusively based on the CDR3 sequence, without considering which peptide they are paired with; this appears to be confirmed also by the findings of Weber et al. (24).

In order to make progress towards robust TCR-peptide/-pMHC interaction prediction, machine learning models should achieve satisfactory test performance on the hard training/test split (HS), which we propose in this work. Only then, such models will be applicable for real-world applications, e.g. personalized cancer immunotherapy and T cell engineering. Possible strategies to achieve this goal might require exploring different feature representations, e.g. SMILES (37) encodings as proposed in TITAN (24). Further possible methods might rely on physics-based simulations for the generation of large-scale datasets. Additionally, transfer learning techniques (38) might allow to leverage knowledge from large databases of protein-ligand binding affinity, e.g. BindingDB (39), which includes more than 1M labeled samples.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

FG pre-processed the data, created the dataset, performed the machine learning experiments and drafted the manuscript. All other authors contributed to the conceptualization of the work and revised the manuscript. In particular, AM supported the data pre-processing and provided immuno-oncological guidance.

FUNDING

This work is funded by the NEC Corporation.

DATA AVAILABILITY STATEMENT

The dataset created for this study can be found in the following repository: **(ToDo - Zenodo link)**. The code used to create the dataset and to run the machine learning experiments can be found in **(ToDo - GitHub link)**, as well as in **(ToDo - Zenodo link)**.

REFERENCES

1 .Kalos M, June C. enAdoptive T Cell Transfer for Cancer Immunotherapy in the Era of Synthetic Biology. *Immunity* **39** (2013) 49–60. doi:10.1016/j.immuni.2013.07.002. Number: 1.

2 .Woodsworth DJ, Castellarin M, Holt RA. enSequence analysis of T-cell repertoires in health and disease. *Genome Medicine* **5** (2013) 98. doi:10.1186/gm502. Number: 10.

3 .Maus MV, Fraietta JA, Levine BL, Kalos M, Zhao Y, June CH. enAdoptive Immunotherapy for Cancer or Viruses. *Annual Review of Immunology* **32** (2014) 189–225. doi:10.1146/annurev-immunol-032713-120136. Number: 1.

4 .Kunert A, van Brakel M, van Steenbergen-Langeveld S, da Silva M, Coulie PG, Lamers C, et al. enMAGE-C2–Specific TCRs Combined with Epigenetic Drug-Enhanced Antigenicity Yield Robust and Tumor-Selective T Cell Responses. *The Journal of Immunology* **197** (2016) 2541–2552. doi:10.4049/jimmunol.1502024. Number: 6.

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

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

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

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

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

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

11 .Jameson SC, Masopust D. enUnderstanding Subset Diversity in T Cell Memory. *Immunity* **48** (2018) 214–226. doi:10.1016/j.immuni.2018.02.010.

12 .Omilusik KD, Goldrath AW. enRemembering to remember: T cell memory maintenance and plasticity. *Current Opinion in Immunology* **58** (2019) 89–97. doi:10.1016/j.coi.2019.04.009.

13 .Jurtz VI, Jessen LE, Bentzen AK, Jespersen MC, Mahajan S, Vita R, et al. Nettcr: sequence-based prediction of tcr binding to peptide-mhc complexes using convolutional neural networks. *BioRxiv* (2018) 433706.

14 .De Neuter N, Bittremieux W, Beirnaert C, Cuypers B, Mrzic A, Moris P, et al. On the feasibility of mining cd8+ t cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics* **70** (2018) 159–168.

15 .Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. Determining epitope specificity of t cell receptors with tcrgp. *BioRxiv* (2019) 542332.

16 .[Dataset] Wong E, Gold M, Meermeier E, Xulu B, Khuzwayo S, Sullivan Z, et al. Trav1-2 cd8 t-cells including oligoconal expansions of mait cells are enriched in the airways in human tuberculosis. commun biol 2: 203 (2019).

17 .Moris P, De Pauw J, Postovskaya A, Ogunjimi B, Laukens K, Meysman P. Treating biomolecular interaction as an image classification problem–a case study on t-cell receptor-epitope recognition prediction. *bioRxiv* (2019).

18 .Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, Laukens K, et al. Detection of enriched t cell epitope specificity in full t cell receptor sequence repertoires. *Frontiers in immunology* **10** (2019) 2820.

19 .Tong Y, Wang J, Zheng T, Zhang X, Xiao X, Zhu X, et al. Sete: Sequence-based ensemble learning approach for tcr epitope binding prediction. *Computational Biology and Chemistry* **87** (2020) 107281.

20 .Springer I, Besser H, Tickotsky-Moskovitz N, Dvorkin S, Louzoun Y. Prediction of specific tcr-peptide binding from large dictionaries of tcr-peptide pairs. *Frontiers in immunology* **11** (2020) 1803.

21 .Fischer DS, Wu Y, Schubert B, Theis FJ. Predicting antigen specificity of single t cells based on tcr cdr 3 regions. *Molecular systems biology* **16** (2020) e9416.

22 .Montemurro A, Schuster V, Povlsen HR, Bentzen AK, Jurtz V, Chronister WD, et al. Nettcr-2.0 enables accurate prediction of tcr-peptide binding by using paired tcrα and β sequence data. *Communications biology* **4** (2021) 1–13.

23 .Springer I, Tickotsky N, 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).

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

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

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

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

28 .Klinger M, Pepin F, Wilkins J, Asbury T, Wittkop T, Zheng J, et al. Multiplex identification of antigen-specific t cell receptors using a combination of immune assays and immune receptor sequencing. *PLoS One* **10** (2015) e0141561.

29 .Zeng H, Gifford DK. Quantification of uncertainty in peptide-mhc binding prediction improves high-affinity peptide selection for therapeutic design. *Cell systems* **9** (2019) 159–166.

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

31 .Davis J, Goadrich M. The relationship between precision-recall and roc curves. *Proceedings of the 23rd international conference on Machine learning* (2006), 233–240.

32 .Fawcett T. An introduction to roc analysis. *Pattern recognition letters* **27** (2006) 861–874.

33 .Manning C, Schutze H. *Foundations of statistical natural language processing* (MIT press) (1999).

34 .Saito T, Rehmsmeier M. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one* **10** (2015) e0118432.

35 .Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. *European conference on information retrieval* (Springer) (2005), 345–359.

36 .Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. *Protein Science* **1** (1992) 409–417.

37 .Weininger D, Weininger A, Weininger JL. Smiles. 2. algorithm for generation of unique smiles notation. *Journal of chemical information and computer sciences* **29** (1989) 97–101.

38 .Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *Journal of Big data* **3** (2016) 1–40.

39 .Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. Bindingdb in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic acids research* **44** (2016) D1045–D1053.

FIGURE CAPTIONS

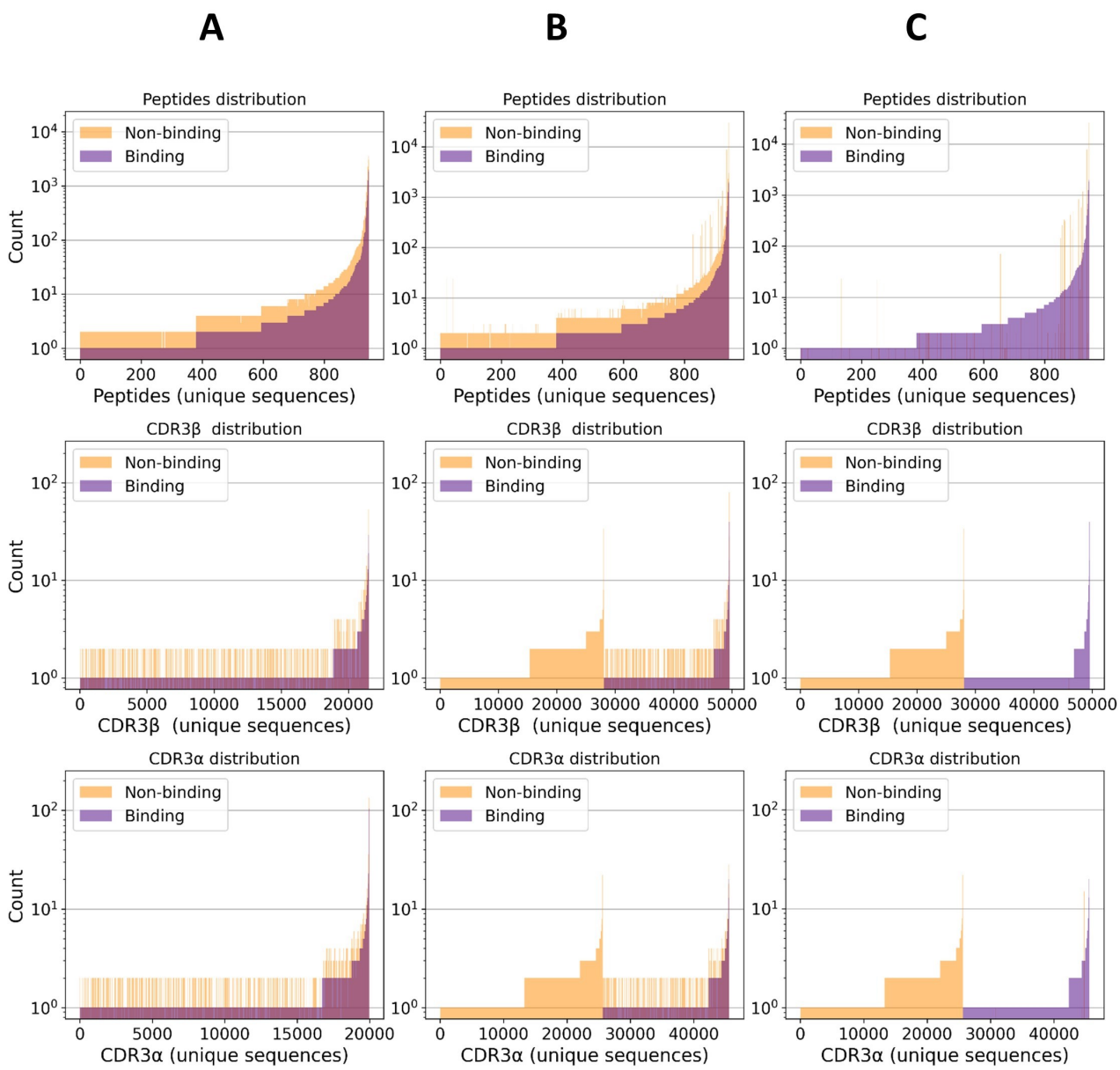

Figure 1. Class distribution of $(peptide, CDR3\beta, CDR3\alpha)$ samples. **(A)** Negative samples only include randomized negative samples (i.e. no negative assays). **(B)** Negative samples include negative assays and randomized negative samples. **(C)** Negative samples only include negative assays.

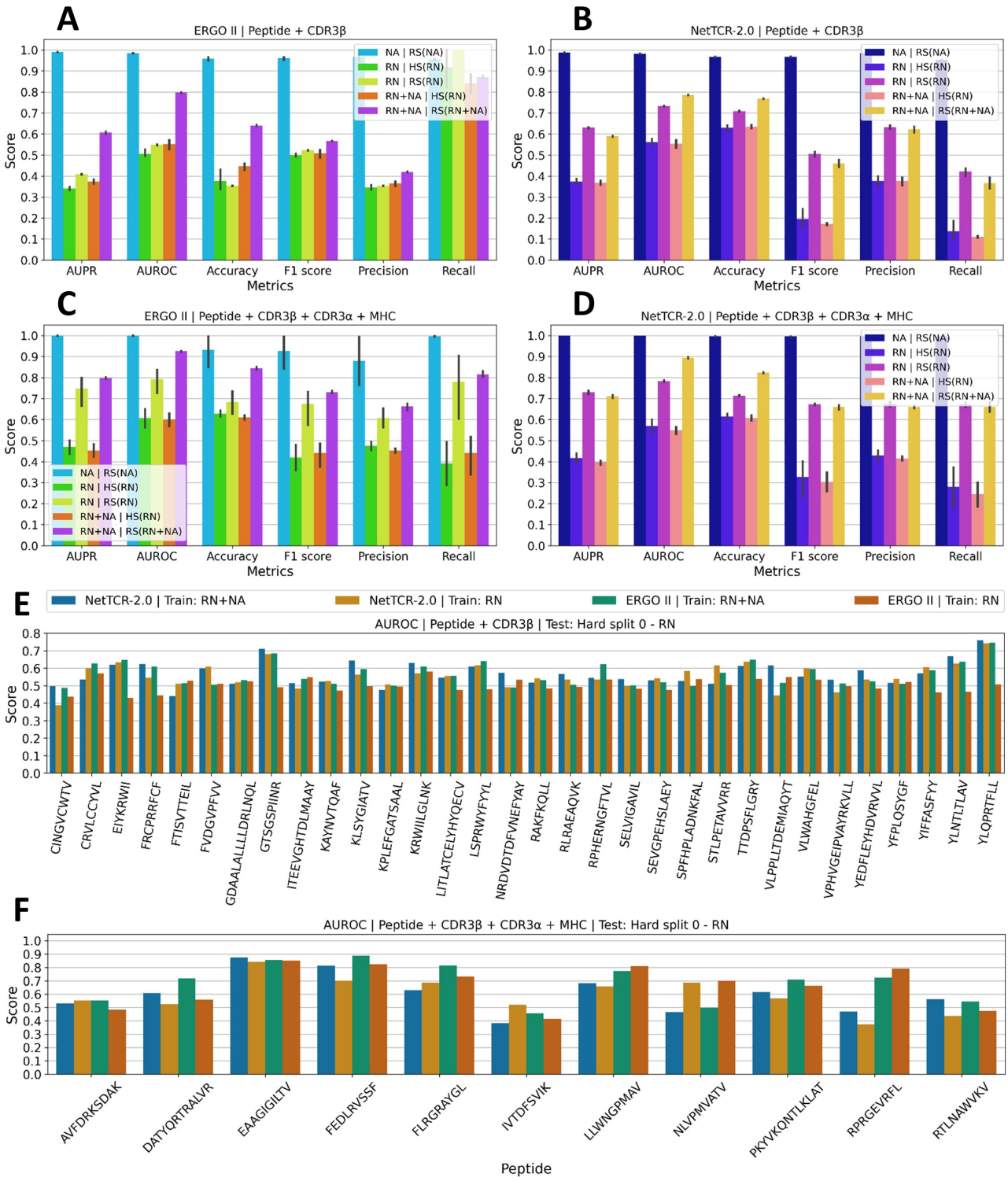

Figure 2. Test results for ERGO II and NetTCR-2.0 for TCR-peptide/-pMHC interaction prediction trained and tested on *TChard*. AUPR: area under the precision-recall curve. AUROC: area under the receiver operator characteristic curve. NA: negative samples from negative assays. RN: negative samples from random mismatching. RS(·): random split. HS(·): hard split. Confidence intervals are standard deviation over 5 experiments with independent training/test splits. **(A, B, C, D)** ERGO II and NetTCR-2.0 results on $(peptide, CDR3\beta)$ and $(peptide, CDR3\beta, CDR3\alpha, MHC)$ samples. Legend: *Source of training negatives* | *Training/test split*. **(E)** Peptide-specific AUROC computed on the $(peptide, CDR3\beta)$ test set obtained with hard split 0 (see Table S2). **(F)** Peptide-specific AUROC computed on the $(peptide, CDR3\beta, CDR3\alpha, MHC)$ test set obtained with hard split 0 (see Table S3).