Lessons learned from the IMMREP23 TCR-epitope prediction challenge

Morten Nielsen, Anne Eugster, Mathias Fynbo Jensen, Manisha Goel, Andreas Tiffeau-Mayer, Aurelien Pelissier, Sebastiaan Valkiers, María Rodríguez Martínez, Barthélémy Meynard-Piganeeau, Victor Greiff, et al.

ImmunoInformatics 16 (2024) 100045

Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.journals.elsevier.com/immunoinformatics

Lessons learned from the IMMREP23 TCR-epitope prediction challenge

Morten Nielsen [a,*], Anne Eugster [b], Mathias Fynbo Jensen [a], Manisha Goel [b], Andreas Tiffeau-Mayer [c], Aurelien Pelissier [d], Sebastiaan Valkiers [i], María Rodríguez Martínez [e], Barthélémy Meynard-Piganeeau [f], Victor Greiff [g], Thierry Mora [h], Aleksandra M. Walczak [h], Giancarlo Croce [i], Dana L Moreno [i], David Gfeller [i], Pieter Meysman [j], Justin Barton [k]

[a] Department of Health Technology, Technical University of Denmark, Lyngby, DK-2800, Denmark
[b] Center for Regenerative Therapies Dresden, Faculty of Medicine, TU Dresden, Dresden, Germany
[c] Division of Infection and Immunity, University College London, United Kingdom
[d] Institute of Computational Life Sciences, ZHAW, 8400, Winterthur, Switzerland
[e] Biomedical Informatics and Data Science, Yale School of Medicine, 06510, New Haven, United States
[f] Computational and Quantitative Biology, LCQB UMR 7238, Institut de Biologie Paris Seine, CNRS, Sorbonne Universite, Paris, 75005, France
[g] Department of Immunology, University of Oslo and Oslo University Hospital, Oslo, Norway
[h] Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris Cité, F-75005, Paris, France
[i] Department of Oncology UNIL CHUV, Ludwig Institute for Cancer Research, University of Lausanne, Lausanne, Switzerland
[j] University of Antwerp, Antwerp, Belgium
[k] Institute of Structural and Molecular Biology, University of London, United Kingdom

ARTICLE INFO

Keywords:
TCR specificity
Benchmark
Prediction
Machine learning

ABSTRACT

Here, we present the findings from IMMREP23, the second benchmark competition focused on predicting the specificity of TCR-pMHC interactions.

The interaction of T cell receptors (TCR) towards their pMHC target is a cornerstone of the cellular immune system. Over the last decade, substantial progress has been made within the field of TCR specificity prediction, providing proof of concept for predicting TCR-pMHC interactions in a narrow space of "seen" pMHC targets where substantial training data is available. However, a significant challenge persists in extending the predictive capability to novel "unseen" pMHC targets. Furthermore, the performance of proposed methods is often challenged when evaluated outside the initial publication and data sets.

To address these issues, IMMREP23 challenge invited participants to predict, for a given test set of TCR-pMHC pairs, the likelihood that a pair would bind. A total of 53 teams participated, providing a total of 398 submissions.

The benchmark confirms that current methods achieve reasonable performance in the "seen" pMHC setting. However, most participating methods had close to random performance on the subset of "unseen" peptides, underlining that this prediction challenge remains essentially unsolved.

Finally, another key lesson from the benchmark is the critical issue of data leakage. Specifically, the data set construction procedure employed in IMMREP23 led to biases in the negative test data set. These biases were identified by several participating teams, and complicated the interpretation of the benchmark results. Based on these results, we put forward suggestions on how future competitions could avoid such data leakages and biases.

1. Introduction

The interaction of T cell receptors (TCR) towards their cognate pMHC target is a cornerstone of the specificity of cellular adaptive immunity. The TCR is a heterodimeric surface protein most often consisting of an α and β chain. The part of the TCR interacting with the pMHC complex is defined by six loops denoted as Complementary Determining Regions (CDRs).

Over the last decade, substantial progress has been made within the field of TCR specificity prediction [1], and current state-of-the-art methods have provided proof of concept for accurately predicting TCR-pMHC interactions in a narrow space of "seen" pMHC where

* Corresponding author.
E-mail address: morni@dtu.dk (M. Nielsen).

https://doi.org/10.1016/j.immuno.2024.100045
Received 2 August 2024; Received in revised form 17 September 2024; Accepted 24 September 2024
Available online 28 September 2024

substantial TCR data is available. These works however also underline the current major challenge for the prediction of TCR specificity, namely the limited extrapolation capability of current models to novel ("unseen") pMHC targets, imposing a significant constraint on their broader applicability [2,3,4].

Predicting TCR pMHC interactions represents a classic machine learning problem [5]. In the "seen" pMHC scenario, the TCR sequence serves as input, with a set number of pMHCs as target labels. In this setting, only pMHCs within the database are potential targets, simplifying the problem to multiclass classification. Data to train such methods are most often obtained from public databases such as VDJdb [6], McPAS-TCR [7], and IEDB [8]. However, these databases only contain data related to positive TCR-pMHC interactions, and machine learning models for classification or regression require negative data. A critical component of the development of TCR specificity prediction models thus lies in the proper definition of these negatives. The two most common approaches applied have been i) swapped negatives, where negative TCRs for a given pMHC are sampled from TCRs specific to other (dissimilar) pMHC, and ii) sampling TCRs from negative control datasets [1]. Both approaches have advantages and disadvantages, as extensively discussed in the literature. As of today, many approaches are using swapped negatives to train and test their model. However, it has been reported that swapped negatives can result in performance overestimation when used for model evaluation [9,10].

Various solutions, from simple database look-ups to deep learning-based models, have emerged to predict TCR specificity and cognate pMHC targets. However, systematic examination of these approaches' advantages and drawbacks remains scarce, with evaluations often performed on internal and limited datasets. In 2022, the first IMMREP benchmark was conducted seeking to address these issues. Here, specific datasets were defined to train and test various prediction models, and outputs were compared to classify approaches and identify ideal datasets and evaluation strategies for future efforts.

Here, we describe IMMREP23, the second benchmark of TCR-pMHC specificity prediction. The competition ran from November 1, 2023 to December 11, 2023. The challenge invited participants to model TCR-pMHC recognition as a binary classification task. For a given test set of TCR-pMHC pairs, participants were asked to use their models to predict the likelihood that a pair would bind.

In contrast to IMMREP22, this competition was conducted on a dataset compiled of novel unpublished paired TCR data with annotated specificity to 20 pHLA (covering 6 distinct HLA molecules). No specific training data was defined for the competition, and participating methods could thus be (and were) trained on any data available. The challenge was hosted on the Kaggle competition platform at https://kaggle.com/competitions/tcr-specificity-prediction-challenge.

Here, we describe the main insights gained from this benchmarking study and recommend strategies for future benchmarking efforts in the TCR-pMHC domain.

It is critical to underline that we have not in any way assessed the accuracy of the test data used for the performance assessment. A known major problem in TCR sequencing of epitope specific cells is the contamination by non-specific TCRs (11). We have opted not to go into a further investigation of this, and as a result false positives/negatives annotations are almost for sure present in the data. We however not believe such mis-annotations will influence the result of the benchmark, since the effect will be identical for all methods.

2. Materials and methods

2.1. Data generation

Data for IMMREP23 were generated by 4 different groups. Below is included a brief description of the 4 experimental setups (for further details refer to supplementary material).

2.1.1. Data set 1: dextramer and plate-based scRNAseq (Eugster and goel, tu dresden)

PBMCs were isolated from healthy adults and stained with dextramers (see supplementary table S1) and standard CD8 T cell identification markers. Single CD8+ cells were isolated and sorted into 96-well plates, and single-cell sequenced using the Smart-seq2 protocol . A total of 40 unique TCRs were identified across three peptides (see supplementary table S2).

2.1.2. Data set 2: 10x (Sine R hadrup, DTU denmark)

PBMCs were stained using barcoded pMHC multimers and standard CD8 T cell identification markers. pMHC multimer positive T cells were sorted and mixed across samples and loaded onto a Chromium Controller. We utilize the 10x Genomics 5′ v2 chemistry that allows the cell barcode to be appended at the 5′-end of transcripts, which is for capturing all V(D)J-, pMHC-, and hashing- associated barcodes as described previously [12]. The downstream processing was conducted according to the manufacturer's instruction (10x Genomics), and the different products (GEXs, TCRs, and barcodes) were sequenced on a NovaSeq running a 150 paired-end program. A total of 48 pMHC multimers were included in the study, 17 of which were included in the final IMMREP test data set (see below).

Gene expression, hashing-associated reads, and pMHC-associated reads were processed as described in supplementary materials. A total of 245 unique TCRs were identified across the 17 peptides.

2.1.3. Data set 3: immudex

A human PBMC cell sample was stained with a panel of dCODE Dextramer® (RiO) reagents and then with the Immune Discovery Panel (IDP) containing 30 BD® AbSeq antibodies. dCODE Dextramer®-positive cells (PE+) were sorted and subjected to the BD Rhapsody™ Single-Cell Analysis System for full-length TCR/BCR VDJ sequencing, on an Illumina NextSeq 500.

The sequencing data was processed through Seven Bridges, BD Rhapsody™ Sequence Analysis Pipeline, and subsequently analyzed using the BD SeqGeq™ software package. 9 pMHC multimers were included in the study, 6 of which were included in the final IMMREP. A total of 89 TCRs were identified across the 6 peptides.

2.1.4. Data set 4: immunoscape

This data set was generated by Florian Schmidt et al. [13]. 15 pMHC multimers were included in the study, 10 of which were included in the final IMMREP test data set. A total of 265 TCRs were identified across the 10 peptides.

2.2. Test data set

2.2.1. Positives

The positive data was constructed from paired-chain data described above. This data consisted of the V-gene, J-gene, and CDR3 of both the alpha and beta chains. In cases where the V- or J-gene allele was not specified, the allele was set to the most common allele (most often *01).

To construct the full TCR sequences, the CDR3 sequences and V/J genes were submitted to the Thimble script in Stitchr [14] with the species set to human (e.g. "-s HUMAN"). In cases where multiple V- or J-gene alleles were listed for a given entry, all combinations were applied, and the set of TCRs recorded. In the case of the current data, all such cases resulted in duplicated TCRs, and a single entry was kept (randomly selecting the V/J gene from the multiple options).

Only entries where both TCR chains could be processed by Stitchr were kept. These entries were then submitted to ANARCI [15] to annotate the individual CDRs from the full sequence. Here, CDR1 was defined as positions 27–38, CDR2 as positions 56–65, and CDR3 as positions 105–117 in the alignment. Furthermore, extended CDR3 sequences were also defined as positions 104–118 in the alignment, as some models relied on the inclusion of the conserved C- and N-terminus

Table 1
Peptides sequence, HLA restrictions and number of positive TCRs for the IMMREP23 data set.

Peptide	HLA	# TCR
GILGFVFTL	A*02:01	103
RAKFKQLL	B*08:01	62
VSDGGPNLY	A*01:01	58
EPLPQGQLTAY	B*35:01	48
NLVPMVATV	A*02:01	43
YVLDHLIVV	A*02:01	34
TDLGQNLLY	A*01:01	33
VTEHDTLLY	A*01:01	25
GLCTLVAML	A*02:01	24
VLEETSVML	A*02:01	21
RPHERNGFTVL	B*07:02	19
SALPTNADLY	A*01:01	18
QIKVRVDMV	B*08:01	16
IPSINVHHY	B*35:01	16
RPPIFIRRL	B*07:02	15
IVTDFSVIK	A*11:01	14
YLQPRTFLL	A*02:01	13
TPRVTGGGAM	B*07:02	13
FTDALGIDEY	A*01:01	12
TSDACMMTMY	A*01:01	11

of the CDR3 sequences.

After annotating the CDR sequences, duplicates in terms of peptide+CDR1+CDR2+CDR3 sequences were removed. Furthermore, 3 cross-reactive TCRs from the ImmunoScape dataset, which were reactive to both IVTDFSVIK and NLVPMVATV, were discarded to keep the data problem binary.

Finally, only peptides with at least 10 TCR pairs were kept for the test dataset, resulting in 598 TCR pairs across 2o peptides (see Table 1)

2.2.2. Negatives

Negatives were generated by swapping TCRs from one peptide with TCRs binding to other peptides with a Levenshtein distance (between the peptides) greater than 3. This value is guided by earlier findings suggesting that cross-reactive peptides most often share two or less mutations [16]. Here, 3 negatives were generated for each positive observation by swapping TCRs with the peptides that had at least 10 positive TCR pairs after all filtering (e.g. the positive peptides in the test data). Furthermore, 2 additional negatives were generated for each positive observation, except that the TCRs were sampled from the peptides that had <10 positive TCR pairs (e.g. those left out from the test dataset). This resulted in a positive-to-negative ratio of 1:5, except for GILGFVFTL (1:4.11), RAKFKQLL (1:4.84), and VSDGGPNLY (1:4.97), because there were not sufficient TCR pairs in the left-out data to generate 2 negatives per positive.

The final data set contained 3484 entries (598 positive, 2886 negatives) covering 20 pMHCs, and 6 MHCs.

2.2.3. Seen and unseen peptide subsets

The peptides in the test data vary widely in the number of TCRs available in the public domain, with some having thousands of unique TCRs, while others are completely uncharacterized. Based on these observations, we labeled 3 peptides (SALPTNADLY, TSDACMMTMY, and FTDALGIDEY) absent from the VDJdb and IEDB as unseen.

2.2.4. Definition of public and private test data set

For the competition, the test data was split into public (7.4 %) and private (92.6 %) subsets. The split into a public and private data subset is a feature of the Kaggle competition setup, and thus has no relation to the concept of public and private T cell receptors. During the competition the performance on the public data set is reported back to the participants. In contrast, the target values for private data set are kept secret until after the competition deadline and is used as the official leaderboard for determining the final ranking of the participating methods. In

Table 2
Participating models. Details on the modeling pipeline and data used for training for the subset of participating methods were this information was provided by the participants. The last column defines if a method used the structure of the test (defined in detail below) to boost performance. This information was provided by the authors of each method.

Model Name	Model Type	Training Data	Details	Uses test set structure
IMW Detect	Custom TCR-ML model	IMWdb	Version 1 - clean prediction Version 2 - iterative retraining	V1: no V2: yes
QImmuno	Bayesian nearest-neighbor association	IEDB, VDJdb, and curated data	Both methods employ TCRdist to compare test set TCRs to curated databases of paired chain and single chain TCRs with annotated pMHC specificity. Leaked TCRs from the Immrep23 dataset (annotated to a single pMHC) were included as additional training samples. TCRdist scores are rescaled to probabilities using nonlinear logistic regression to combine information from single and paired chain near matches with priors (number of possible pMHCs) in a Bayesian manner. Version 1 - prediction made per TCR-pMHC pair, uses pMHC multiplicity as a Bayesian prior, uses similarity of TCRs to other peptides presented on the same HLA where pMHC data is scarce Version 2 - explicitly performs multiclass prediction for each TCR among possible pMHCs. Corrects neighbor distances by local density estimates to account for non-uniform background probabilities of generation	Yes
NetTCR	CNN	IEDB and VDJdb	Handling of data with incomplete TCR annotation, data imbalance (few pMHC with large numbers of TCR and many with few), and integration of	No

(continued on next page)

Table 2 (*continued*)

Model Name	Model Type	Training Data	Details	Uses test set structure
			TCRbase rescaling. The different models are (with reference to the NetTCR-2.2 architecture) M1:NetTCR-2.3 ensemble trained directly on a mixed-chain dataset. M2:NetTCR-2.3 ensemble trained directly on a mixed-chain dataset with potential outliers removed. M3:NetTCR-2.3 ensemble trained directly on a mixed-chain dataset with potential outliers removed and scaled by TCRbase M4:NetTCR-2.3 ensemble consisting of alpha-, beta- and paired-chain models trained separately on each type of chain data (except mixed) with potential outliers removed and scaled by TCRbase M5: NetTCR-2.3 ensemble consisting of mixed, alpha-, beta- and paired-chain models trained separately on each type of chain data with potential outliers removed and scaled by TCRbase	
MixTCRpred	Deep network	IEDB, VDJdb, literature curation, in house generated data	MixTCRpred_s1: Trained on publicly available data for 15 seen peptides. MixTCRpred_s2: Trained on data from s1 + new data generated in house during the IMMREP benchmark for 4 additional peptides (SALPTNADLY, TDLGQNLLY, TSDACMMTMY, VSDGGPNLY) and for one peptide already present in s1 (VTEHDTLLY), reaching a total of 19 epitopes with training data.	V1/V2: no V3: yes

Table 2 (*continued*)

Model Name	Model Type	Training Data	Details	Uses test set structure
			MixTCRpred_s3: Trained on data from s2 + inferred positives and negatives based on the structure of the test set (e.g., TCRs occurring only once are by design positives; TCRs with good scores for one epitope with reliable predictions are likely negatives in all other instances).	
ESM shallow	A 3-layer perceptron accepting protein embeddings from ESM2 as input.	Only training data provided by the organizer.	We used ESM2 with 650 million parameters. The embeddings of the alpha and beta chains were concatenated or tested independently.	No
TULIPv2	Unsupervised Encoder-Decoder based transformer	IEDB MCPAS VDJDB	The model is trained to predict the next aminoacid of the epitope, given its interacting TCR. This defines a conditional probability distribution over the epitope space. We use the epitope probability to rank TCRs. Does not use negative TCRs for training.	yes
Koi			a collection of small peptide-specific models	yes
NN distance baseline				yes
TCRbase	Sequence similarity based model	IEDB and VDJdb	Method described in (9)	no

IMMREP23, the public subset was sampled only from the two peptides GILGFVFTL and RAKFKQLL covered with the largest number of positive TCR, sampling 30 % for each maintaining the positive:negative class ratio. The remaining data formed the private dataset. The data set with annotated target values (these were not available for the competition) is available at https://github.com/justin-barton/IMMREP23/. During the IMMREP23 competition, performance evaluations on the public data were reported, while the performance on the private data was released only after the benchmark was completed.

2.3. Evaluation metrics

The primary evaluation metric used during the competition was the average per peptide AUC0.1:

$$average\ AUC0.1 = \frac{1}{N}\sum_{1}^{N} AUC0.1(p)$$

Where N is the number of peptides in the test set and $AUC0.1(n)$ is

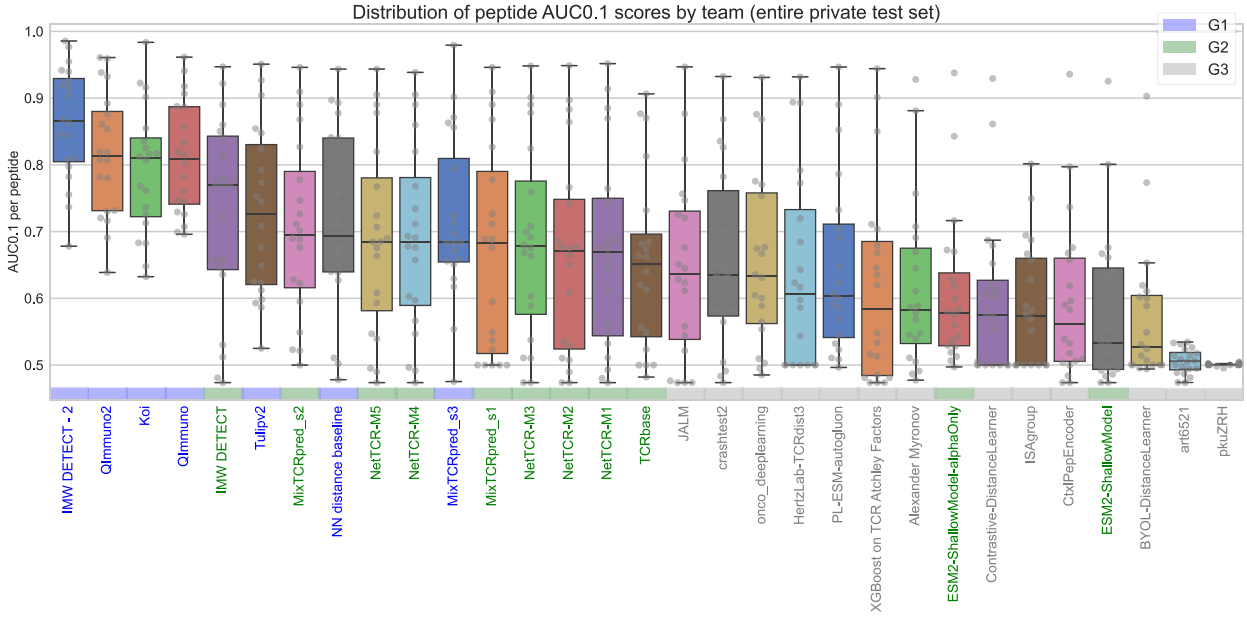

Fig. 1. Predictive power as measured in terms of the AUC0.1 per peptide on the entire private test data set. Each box plot shows the distribution of the AUC0.1 values over the 20 test set peptides. The participating methods are split into the three sub-groups defined in the main text; G1 methods (blue) G2 methods (green) and G3 (grey).

the AUC0.1 for each peptide p, defined as the partial area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curve up to a false positive rate (FPR) of 0.1. As proposed in [17], McClish standardization was also applied.

2.4. Baseline method

The TCRbase method was included as baseline in the benchmark. In short, TCRbase assigns a score for an element in a test set, as the highest similarity to all entries in a given database of positive TCR. The similarity is calculated per CDR as the kernel-similarity of BLOSUM62-encoded k-mers ranging from size 1–30 between the two sets of CDRs that are compared. The weighting for the CDRs was set to 1,1,3,1,1,3 for CDR1α-, CDR2α-, CDR3α-, CDR1β-, CDR2β-, and CDR3β, respectively, in line with earlier recommendations (for further details refer to [9]).

3. Results

3.1. Participants

A total of 53 teams participated in the benchmark, providing a total of 398 submissions. A subset of teams provided details on the modeling pipeline and data used for training, which can be found in Table 2.

3.2. Overall performance

The overall benchmark performance of each team on the entire private data set is shown in Fig. 1. Here, the participating methods are split into three sub-groups; G1 methods which are confirmed by the authors of the methods to have used the test set structure for predictions (see Table 2), G2 methods which are confirmed by the authors of the methods not to have used the test set structure in predictions, and the rest forming G3.

From this plot, one can appreciate that the best performing methods for most parts belong to G1, followed by the methods in G2, and the poorest performing methods all are from the G3. The G1 is for most parts (with the exception of Tulip [18]) formed by novel and/or unpublished methods within the TCR specificity prediction space. These methods all demonstrate a predictive power much beyond that of the conventional state-of-the-art methods such as MixTCRpred [4] and NetTCR [2]. Most methods in the G2 group have comparable performance with the exception IMW DETECT that shows an substantial predictive advantage. Finally, the G3 is formed by methods with a relatively poor predictive power, and a performance that is lower than the sequence-based TCRbase baseline method.

3.3. Variation of performance across epitopes

Fig. 2 displays the performance of the different methods for the individual peptides in more detail. This figure demonstrates a very high difference in predictive performance not only between the individual methods (as also shown in Fig. 1) but also between the different peptides. For instance, the performance for GIL and GLC is high (AUC0.1 $>$ 0.7) across almost all methods. This is a reflection of the high number of accurate TCR data available in the public domain mapped towards these peptides. Also, the high performance of the G1 methods, as defined in Figure, 1 across most of the peptides is apparent. Likewise, only the G1 methods display predictive power across the complete set of unseen peptides (SAL, TSD, and FTD). Further, it is interesting to observe the cases where selected teams had very good performance on specific peptides, while all others failed. A few such examples include IPS for “IMW DETECT”, and VTE for “IMW DETECT” and “NN distance baseline”. As we do not have access to details regarding these methods, we cannot disentangle whether these high-performance values should be contributed to training data or specific machine learning methodologies.

3.4. Data leakage and biases

During the IMMREP23 competition, a manuscript containing part of the test data was published [13]. Likewise, for a small number of data entries, the CDR3b (extended) annotations were found to miss the C-terminal residue. To investigate to what degree this early release and incomplete CDR3b annotations impacted the predictive performance of the different methods, we evaluated the performance on the subset of the private data set excluding the data from this source (supplementary figure S1), and excluding both the data from this source and the data

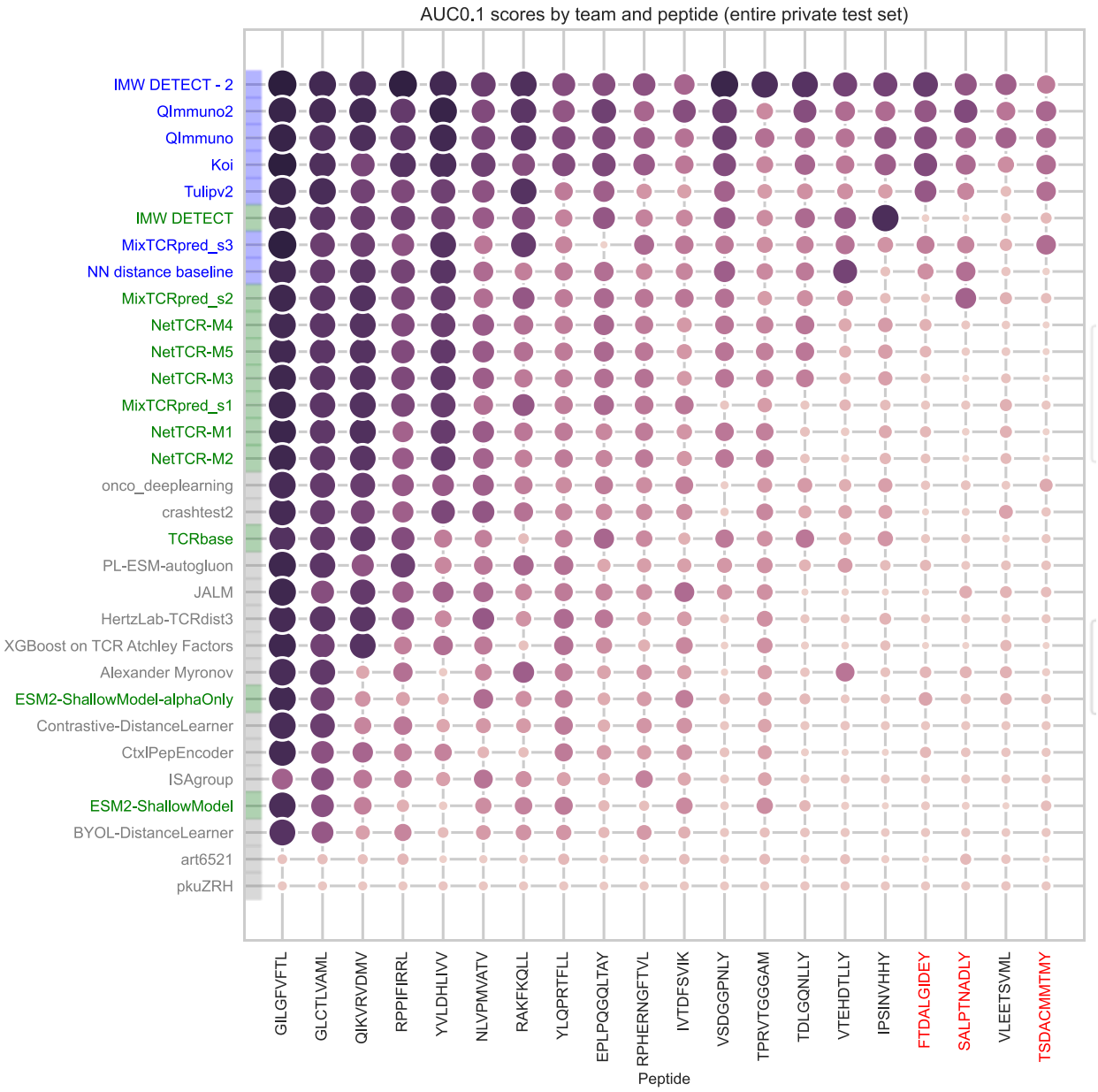

Fig. 2. AUC0.1 scores by method and peptide. The AUC0.1 performance values for each method and peptide were calculated from the entire private test data set. Circle size and color reflect the AUC0.1 value. Peptides in red are unseen peptides. G1 models in blue are those that are confirmed by the authors of the given method to use the test set structure for predictions. G2 models in green are those that are confirmed not to use the test set structure in predictions.

with incomplete CDR3b annotations (supplementary figure S1). Both figures confirm that these data issues had minor impacts on the predictive performance and associated ranking of the different methods. At the top of the ranking, IMW Detect 2 (the only method admitted to have made use of the published Immunoscape test data) lost its advantage in accuracy and dropped to second place behind the Qimmuno-2 entry. Given this minor impact, we for clarity and transparency used the full IMMREP23 test data for model evaluation.

The test data of the IMMREP23 benchmark were collected by the organizers with the implicit assumption that models' predictions for a given TCR-epitope pair would be made independently of the other TCR-epitope pairs in the dataset. As a consequence, consideration was not given to the fact that the method of simulating non-binding examples (described in material and methods) introduced target leakage into the test data set. This leakage manifests in TCRs binding to peptides with few positive TCRs being sampled as negatives multiple times for other peptides. This results in the number of times that a TCR appears in the test set being inversely correlated with its probability of being a positive. By way of example, the 20 most frequent TCRs in the test data were all negatives sampled from the subset of pMHC excluded from the benchmark, and all the TCRs present only a single time in the test set were all positives.

The correlation between frequency in the test set and the target label could be exploited to improve the predictive power on the test data. Indeed, all of the top four models and six of the top 8 models in the competition disclosed using prediction approaches that could take advantage of the structure of the test dataset (Table 2).

To provide a random guess baseline taking into account the problem of data leakage, we compared these methods to a naive approach that makes predictions solely based on TCR counts. This method ranked TCRs for a given pMHC by the inverse of the number of times a given TCR was present in the test data set. This TCR count baseline had an average

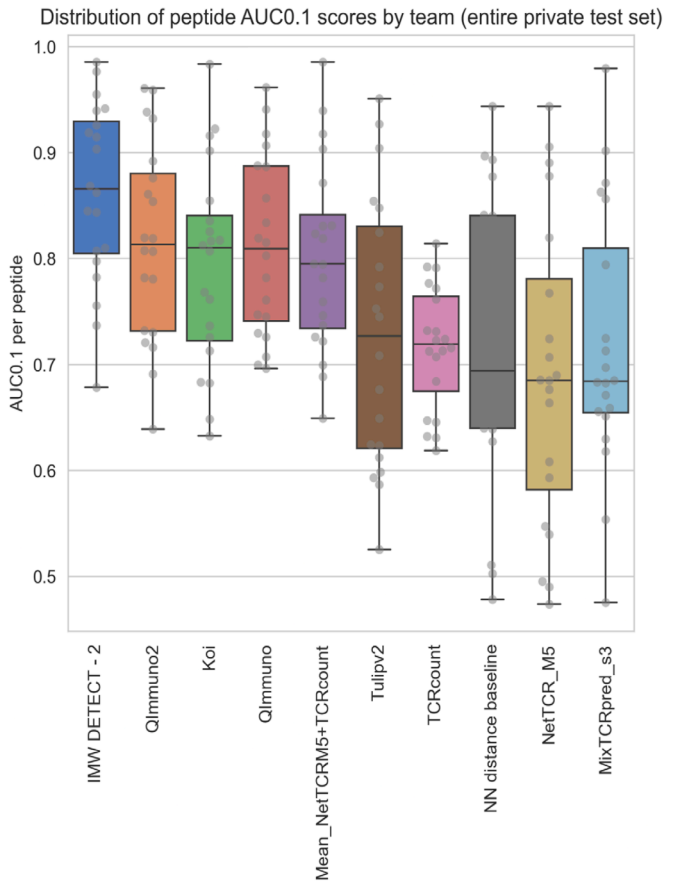

Fig. 3. Imprints on the predictive power of test set data biases. Predictive power of the G1 methods, the G2 method NetTCRB_M5, a model driven solely on the test data bias (TCRcount, defined from the 1/#TCR, and a combination of NetTCRB_M5 and TCRcount (Mean_NetTCRM5+TCRcount) on the entire private test data set.

AUC0.1 above 0.7. This prediction accuracy is lower than that of the top-ranked models demonstrating their additional prediction power, but higher than any of the methods in the G2 group (see Fig. 3). The performance gain obtained from exploiting the target leakage in the test set is particularly apparent when evaluating the performance of the different methods on the unseen peptides (see methods for details on these peptides) (see Fig. 4).

The data leakage and bias complicate the comparison of performance across models either making use of the dataset structure or not. In an initial attempt to provide insights into how different models might fare in a head-to-head comparison, we combined one of the top-ranked models from the G2 group with a TCRcount prior. Specifically, we calculated the unweighted mean of the NetTCR_M5 and TCRcount score, and we found that this combination made its AUC0.1 score comparable to the tools in the G1 group (see Fig. 3 and Fig. 4. Mean_NetTCRM5 + TCRcount).

This figure demonstrates a very high overlap between the methods with a non-random performance on the unseen peptides and the G1 methods defined from Table 2. Note also that the third-best performing method on these unseen peptides is TCRcount, introduced above. This method has no general predictive power when it comes to TCR specificity but only makes predictions from the TCRcount distribution of the test data.

Taken together these results suggest that the high AUC0.1 scores achieved by the G1 models might not be representative of their expected performance on independent data. Future benchmarking on more carefully designed prediction tasks is needed to fairly compare their performance against models.

3.5. The IMMREP23 evaluation

Due to the confounding effect of the target leakage in the test data set, we can only make concrete statements about the performance of methods that have not benefited from this leakage, i.e. the G2 and G3 methods. We are aware that we in this manner do not give the G1 methods a fair evaluation, but we cannot evaluate how they would have performed had they been trained without including these biases.

Based on this subset of methods, the "IMW DETECT" was the best-performing method, followed by MixTCRpred and the different variants of NetTCR (see Fig. 5).

Further, referring to Fig. 4, and focusing on the G2 methods the results confirm that specificity prediction for "unseen" pMHCs remains an unsolved problem, with maximal predictive performance values capped at 0.62 (for ESM2). Even in this case, only one of the three novel peptides was predicted substantially better than random. Further, in one case (MixTCRpred_s2) a team was able to generate, during the course of the competition, experimental data for two of these unseen epitopes (SAL and TSDA) and using these data to train models for these epitopes (see Table 2). This novel data, likely explain the high performance of MixTCRpred_s2 on the SAL "unseen" peptide.

4. Discussion

Here, we have described the results from the IMMREP23 competition.

Lessons learned in terms of benchmark metric, and data biases

In contrast to the earlier IMMREP22 competition, the main objective was to evaluate the current state of the field. Therefore the format of the competition was different from that of IMMREP22 in that no restraints were imposed in terms of training data. This means that we in this competition could not make any straightforward comparisons and draw conclusions regarding the impact of model architectures, training strategies, and training data.

In addition, the presence of target leakage in the test set, introduced by the method chosen by the organizers for simulating negative data, complicates the otherwise straightforward ranking of methods. While it is clear that some methods benefited from information about the target variable that would not naturally be present when making predictions on clinical data, it is unclear how these methods would perform in the absence of this leakage.

Given these observations, for future TCR-epitope benchmarks, we would suggest that the benchmark is defined in two formats both conducted on novel TCR data

1. Predict with any model trained on any available data (the IMMREP23 format)
2. Provide specific training data and apply the trained model to the test set (the IMMREP22 format)

This setup will allow the benchmark to address both central issues of evaluating the current state of the field, and optimal training and machine learning modeling strategies.

In addition, the reported data bias for IMMREP23 can be avoided by defining a test data that pairs all available TCRs to all peptides. In this setting, one could next evaluate the performance both in the context of individual pHLAs (as done here in IMMREP23) and in the context of individual TCRs, i.e. solving the multi-class problem of predicting the peptide target of a given TCR.

Conclusions about the state of the TCR-epitope prediction field

Despite the shortcomings outlined above for the IMMREP23 competition, several important conclusions can nonetheless be drawn. First and foremost, we can observe that several methods which did not take advantage of the test set target leakage displayed clear predictive

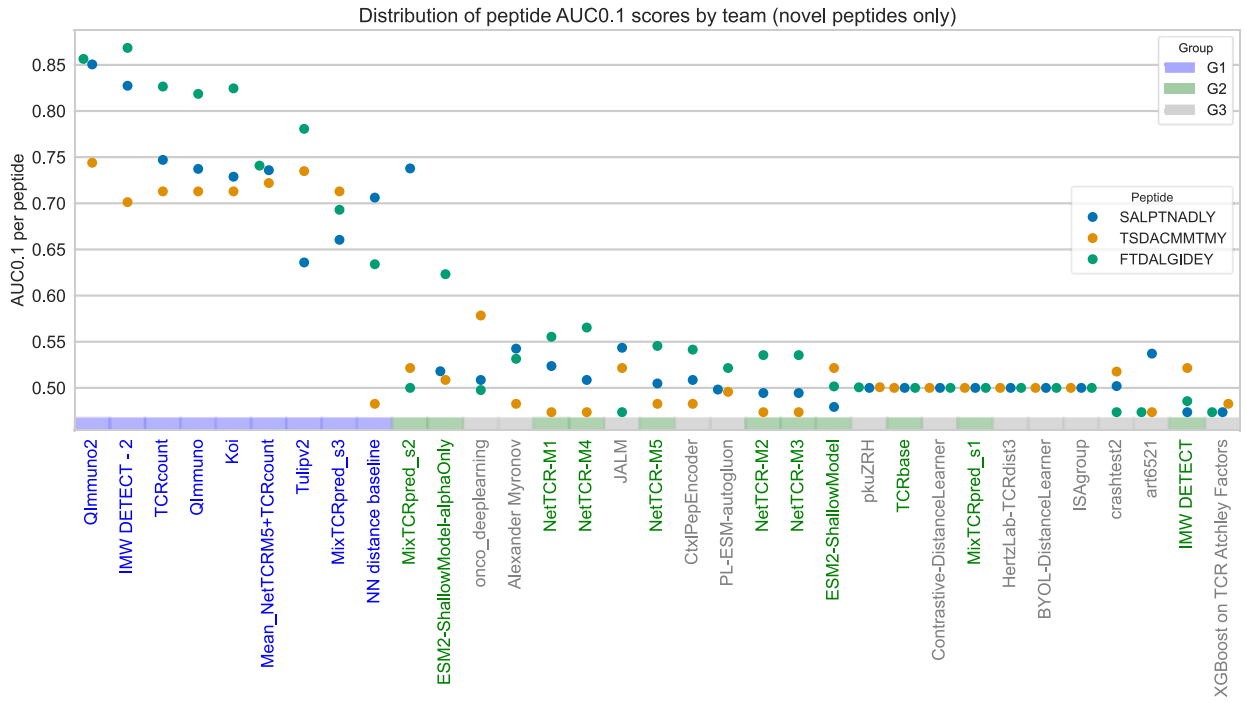

Fig. 4. Predictive performance of the different methods on the subset of unseen peptides. Methods include the two additional methods, TCRcount, and mean_NetTCRM5+TCRcount) defined in Fig. 3. G1 models in blue are those that are confirmed to use the test set structure for predictions. G2 models in green are those that are confirmed not to use the test set structure in predictions. G3 models for which there is no information about test set structure use are in grey.

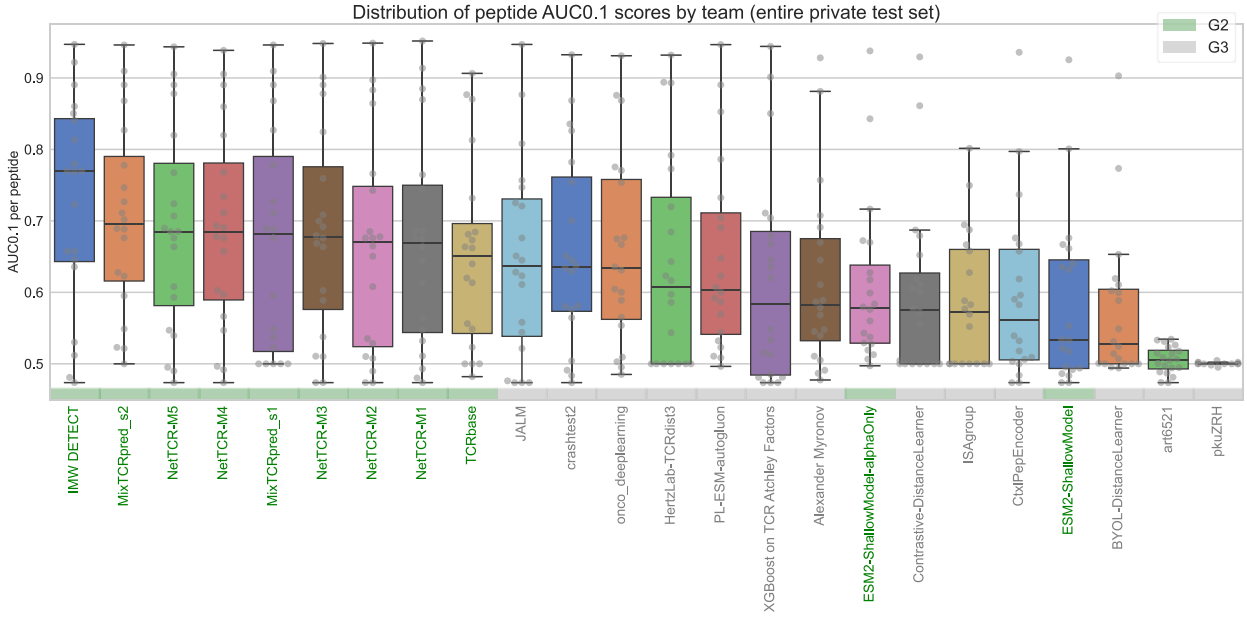

Fig. 5. Predictive power of the G2 and G3 methods was evaluated in terms of AUC0.1 for each of the 20 peptides in the entire private test data set.

power on the "seen" pHLA setting, notably "IMW DETECT", MixTCRpred, and NetTCR. Secondly, it is apparent that the unseen epitope prediction remains unsolved, and all participating methods that made blind predictions on the test data achieved very close to random performance on the subset of unseen peptides.

It is not possible to evaluate to what degree the high performance of "IMW DETECT" is due to an improved model architecture and/or an availability of proprietary in-house training data. However, descriptions of both the modeling architectures and training data are available for the MixTCRpred [4] and NetTCR [2] methods. Both models are based on the CDR1a, CDR2a, CDR3a, CDR1b, CDR2b, and CDR3b, and their performance ranking thus align with the conclusion from IMMREP22 [1] on the importance of including both TCR chains when constructing methods for TCR specificity predictions.

Of note, some participants employed embeddings from protein language models such as ESM2 [19] to predict TCR specificity, achieving partial success in predicting binding to unseen epitopes. Fine-tuning these models to fully exploit their potential might be a valuable

direction to explore in further work. We also observe that the methods participating in IMMREP23 all were sequence-based, and hence did not in any way incorporate structural information of TCR or TCR-pHLA complex into their prediction model. Given the substantial number of recent publications suggesting an important contribution of structural models and their associated features in particular for the prediction TCR specificity to "unseen" pMHCs (examples include [20–22]), it would be highly interesting to see how such models perform in blind benchmarks like the IMMREP.

CRediT authorship contribution statement

Morten Nielsen: Writing – review & editing, Writing – original draft, Project administration, Methodology, Formal analysis, Data curation, Conceptualization. **Anne Eugster:** Writing – review & editing, Data curation, Conceptualization. **Mathias Fynbo Jensen:** Writing – review & editing, Data curation. **Manisha Goel:** Writing – review & editing, Data curation. **Andreas Tiffeau-Mayer:** Writing – review & editing. **Aurelien Pelissier:** Writing – review & editing. **Sebastiaan Valkiers:** Writing – review & editing. **María Rodríguez Martínez:** Writing – review & editing. **Barthélémy Meynard-Piganeeau:** Writing – review & editing. **Victor Greiff:** Writing – review & editing. **Thierry Mora:** Writing – review & editing. **Aleksandra M. Walczak:** Writing – review & editing. **Giancarlo Croce:** Writing – review & editing. **Dana L Moreno:** Writing – review & editing. **David Gfeller:** Writing – review & editing. **Pieter Meysman:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **Justin Barton:** Writing – review & editing, Writing – original draft, Visualization, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data is made availlable at GitHub as described in the manuscript.

Acknowledgements

The IMMREP team would like to thank ImmuDex, ImmunoScape, Sine Reker Hadrup, Anne Eugster/Manisha Goel, and VDJdb for their contribution to the generation of the IMMREP23 test data set.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.immuno.2024.100045.

References

[1] Meysman P, Barton J, Bravi B, Cohen-Lavi L, Karnaukhov V, Lilleskov E, et al. Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report. ImmunoInformatics 2023;9:100024.
[2] Jensen MF, Nielsen M. Enhancing TCR specificity predictions by combined pan- and peptide-specific training, loss-scaling, and sequence similarity integration. Elife. 2024;12:RP93934.
[3] Grazioli F, Mösch A, Machart P, Li K, Alqassem I, O'Donnell TJ, et al. On TCR binding predictors failing to generalize to unseen peptides. Front Immunol 2022;13:1014256.
[4] Croce G, Bobisse S, Moreno DL, Schmidt J, Guillame P, Harari A, et al. Deep learning predictions of TCR-epitope interactions reveal epitope-specific chains in dual alpha T cells. Nat Commun 2024;15(1):3211.
[5] Greiff V, Yaari G, Cowell LG. Mining adaptive immune receptor repertoires for biological and clinical information using machine learning. Current Opinion in Systems Biology 2020;24:109–19.
[6] Goncharov M, Bagaev D, Shcherbinin D, Zvyagin I, Bolotin D, Thomas PG, et al. VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. Nat Methods 2022;19(9):1017–9.
[7] Tickotsky N, Sagiv T, Prilusky J, Shifrut E, Friedman N. McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. Bioinformatics 2017;33(18):2924–9.
[8] Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res 2019;47(D1):D339–43.
[9] Montemurro A, Jessen LE, Nielsen M. NetTCR-2.1: lessons and guidance on how to develop models for TCR specificity predictions. Front Immunol 2022;13:1055151.
[10] Dens C, Laukens K, Bittremieux W, Meysman P. The pitfalls of negative data bias for the T-cell epitope specificity challenge. Nat Mach Intell 2023;5(10):1060–2.
[11] Montemurro A, Povlsen HR, Jessen LE, Nielsen M. Benchmarking data-driven filtering for denoising of TCRpMHC single-cell data. Sci Rep 2023;13:16147.
[12] Povlsen HR, Bentzen AK, Kadivar M, Jessen LE, Hadrup SR, Nielsen M. Improved T cell receptor antigen pairing through data-driven filtering of sequencing information from single cells. Elife. 2023;12:e81810.
[13] Schmidt F, Fields HF, Purwanti Y, Milojkovic A, Salim S, Wu KX, et al. In-depth analysis of human virus-specific CD8+ T cells delineates unique phenotypic signatures for T cell specificity prediction. Cell Rep 2023;42(10):113250.
[14] Heather JM, Spindler MJ, Alonso MH, Shui YI, Millar DG, Johnson DS, et al. Stitchr: stitching coding TCR nucleotide sequences from V/J/CDR3 information. Nucleic Acids Res 2022;50(12):e68.
[15] Dunbar J, Deane CM. ANARCI: antigen receptor numbering and receptor classification. Bioinformatics 2016;32(2):298–300.
[16] Petrova G, Ferrante A, Gorski J. Cross-reactivity of T cells and its role in the immune system. Crit Rev Immunol 2012;32(4):349–72.
[17] McClish DK. Analyzing a Portion of the ROC Curve. Med Decis Making 1989;9(3):190–5.
[18] Meynard-Piganeau B, Feinauer C, Weigt M, Walczak AM, Mora T. TULIP: a transformer-based unsupervised language model for interacting peptides and T cell receptors that generalizes to unseen epitopes. Proc Natl Acad Sci U S A. 2024;121(24):e2316401121.
[19] Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118(15):e2016239118.
[20] Zhao Y, He B, Xu F, Li C, Xu Z, Su X, et al. DeepAIR: a deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. Sci Adv. 2023;9(32):eabo5128.
[21] Ji H, Wang XX, Zhang Q, Zhang C, Zhang HM. Predicting TCR sequences for unseen antigen epitopes using structural and sequence features. Brief Bioinform 2024;25(3):bbae210.
[22] Karnaukhov VK, Shcherbinin DS, Chugunov AO, Chudakov DM, Efremov RG, Zvyagin IV, et al. Structure-based prediction of T cell receptor recognition of unseen epitopes using TCRen. Nat Comput Sci 2024;4(7):510–21.