ImmunoInformatics 9 (2023) 100024

Contents lists available at ScienceDirect

ImmunoInformatics

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

Benchmarking solutions to the T-cell receptor epitope prediction problem: IMMREP22 workshop report

Pieter Meysman [a,b,*], Justin Barton [c,1], Barbara Bravi [d,1], Liel Cohen-Lavi [e,f,1], Vadim Karnaukhov [h,i,1], Elias Lilleskov [j,1], Alessandro Montemurro [k,1], Morten Nielsen [k,1], Thierry Mora [i,1], Paul Pereira [i,l,1], Anna Postovskaya [a,b,m,1], María Rodríguez Martínez [n,1], Jorge Fernandez-de-Cossio-Diaz [i,1], Alexandra Vujkovic [a,b,m,1], Aleksandra M. Walczak [i,1], Anna Weber [n,1], Rose Yin [o,1], Anne Eugster [g,2,**], Virag Sharma [p,2,**]

[a] AUDACIS, University of Antwerp, Antwerp, Belgium
[b] ADREM data lab, Department of Computer Science, University of Antwerp, Antwerp, Belgium
[c] Institute of Structural and Molecular Biology, University of London, London, United Kingdom
[d] Department of Mathematics, Imperial College London, London, United Kingdom
[e] National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[f] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer-Sheva, Israel
[g] Center for Regenerative Therapies Dresden, Faculty of Medicine, TU Dresden, Dresden, Germany
[h] INSERM U932, PSL University, Institut Curie, Paris 75005, France
[i] Laboratoire de physique de l'Ecole normale superieure, CNRS, PSL University, Sorbonne Universite, Universite Paris-Cité, Paris 75005, France
[j] Department of Physics, University of Washington, Seattle, WA, USA
[k] Department of Health Technology, Technical University of Denmark, Lyngby DK-2800, Denmark
[l] Sanofi R&D, Chilly-Mazarin 91380, France
[m] Clinical Virology Unit, Institute of Tropical Medicine, Antwerp, Belgium
[n] IBM Research Europe, Säumerstrasse 4, Rüschlikon 8803, Switzerland
[o] Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[p] Department of Chemical Sciences, School of Natural Sciences, University of Limerick, Limerick V94 T9PX, Ireland

A B S T R A C T

Many different solutions to predicting the cognate epitope target of a T-cell receptor (TCR) have been proposed. However several questions on the advantages and disadvantages of these different approaches remain unresolved, as most methods have only been evaluated within the context of their initial publications and data sets. Here, we report the findings of the first public TCR-epitope prediction benchmark performed on 23 prediction models in the context of the ImmRep 2022 TCR-epitope specificity workshop. This benchmark revealed that the use of paired-chain alpha-beta, as well as CDR1/2 or V/J information, when available, improves classification obtained with CDR3 data, independent of the underlying approach. In addition, we found that straight-forward distance-based approaches can achieve a respectable performance when compared to more complex machine-learning models. Finally, we highlight the need for a truly independent follow-up benchmark and provide recommendations for the design of such a next benchmark.

Introduction

A key challenge within immunoinformatics is the prediction of the target epitope for a T-cell receptor (TCR) sequence. Indeed, the recognition of an epitope by a T-cell receptor (TCR) is an essential step for the activation of T-cells and thus critical for a functioning adaptive immune system. Epitopes are short peptides presented by the major histocompatibility complex (MHC) on the surface of antigen-presenting cells, allowing cognate TCRs to bind them and to confer specificity to the T cell. TCRs consist of a heterodimer, most commonly of an alpha- and a beta chain. Each of these chains are the result of a V(D)J somatic recombination event during T-cell maturation. Due to the randomness of

* Corresponding author at: AUDACIS, University of Antwerp, Antwerp, Belgium.
** Corresponding authors.
E-mail addresses: pieter.meysman@uantwerpen.be (P. Meysman), eugster@tu-dresden.de (A. Eugster), virag.sharma@ul.ie (V. Sharma).
[1] These authors have been listed in alphabetic order.
[2] These authors should be considered as joint last authors.

https://doi.org/10.1016/j.immuno.2023.100024
Received 27 October 2022; Received in revised form 31 January 2023; Accepted 31 January 2023
Available online 3 February 2023

this recombination process, each T-cell clone expresses a potentially unique TCR and thus has a unique epitope specificity. This high TCR diversity allows the adaptive immune system to respond to the myriad of seen and unseen threats.

The advent of high-throughput adaptive immune receptor repertoire (AIRR) sequencing techniques allows to access the TCR sequences of large parts of T-cell repertoires. However, the sheer numbers of existing TCRs mean that most of the TCRs encountered in an experiment may not have been characterized before. Moreover, while the presence of specific TCRs in a population of T-cells can now often be established, their cognate epitope targets mostly remain unknown. Because epitope recognition is crucial for pathogen defense, vaccine response, tumor control and autoimmune diseases and since TCR specificity helps understanding the function of a T-cell, it is essential to learn to decipher it.

Predicting the epitope of a TCR sequence can be considered a straightforward machine learning problem. In the "seen" epitope setting, which is the focus of this study, the TCR sequence is used as the input of the model and a fixed number of epitopes are the target labels. The input features are therefore derived from the TCR sequence, however some models also include features derived from the epitope to possibly learn interactions between the two. Known TCR-epitope pairs, collected in databases such as VDJdb [1] or IEDB [2], can then be used to fit the epitope-TCR model. Within the "seen" epitope setting, only those epitopes that are contained in the epitope-TCR database are possible targets. In its simplest form, this can be a single epitope, reducing the problem to a binary classification. The task of the model is to identify which, if any, epitope is the most likely target of the input TCR.

During the past years, several solutions to unravel and predict the specificity and the cognate epitope target of a TCR have been proposed, ranging from a simple database look-up to deep learning-based prediction models. The advantages and disadvantages of these different approaches have not yet been systematically examined, most having only been evaluated within the context of their initial publications and data sets. In addition, the annotation of TCR-epitope pairs is a complex problem: the promiscuity of TCR-epitope binding and the technical but also experimental variations underlying paired TCR-epitope data make the identification of a clear signal difficult [3]. There is thus a clear need to benchmark existing TCR-epitope prediction approaches, enabling the field to progress towards an understanding of the principles underlying T-cell specificity.

Here we report the findings of the first public TCR-epitope prediction benchmark performed in the context of the ImmRep 2022 TCR-epitope specificity workshop (https://www.pks.mpg.de/immrep22). Leading scientists in the field as well as junior researchers interested in the TCR specificity problem were invited to participate and were offered datasets to train and test a collection of existing prediction models. The aim of the workshop was to evaluate and compare the obtained outputs to classify the approaches and most importantly, to help identify an ideal dataset and optimal evaluation strategies for future follow-up efforts. Describing the outcome of the workshop, we attempt to group the selected and tested methods, both - based on the TCR feature input as well as on the underlying prediction algorithm - in an attempt to identify patterns within the performance results. We conclude the report with lessons learned on the TCR-epitope problem from this first benchmarking study and make several recommendations towards future attempts at benchmarking.

Materials and methods

Construction of the training- and test data

The benchmark data set was derived from the VDJdb database (downloaded on 23/06/2022). VDJdb is a curated database of TCRs with known antigen specificities [4]. Only those TCR-epitope pairs with paired alpha-beta chain data were collected. Any duplicated TCR-epitopes were removed, as defined by V/J gene usage as well as the CDR3 (complementarity determining region) sequence for both the alpha- and the beta chain. Only TCR-epitope pairs obtained by tetramer- or dextramer-sort were retained. Furthermore, all the dextramer-sort entries originating from the 10X technical report [5] were excluded because of the high reported cross-reactivity of these TCRs [3]. Lastly, only those 17 MHC-epitopes that had at least 50 unique TCR sequences were retained.

A full list of epitopes and the number of their associated TCR is provided in Table S1. To constitute the negative control data set, unpublished paired alpha-beta chain TCR sequences without peptide specificity information and obtained from 10x genomics sequencing of CD8+CD96+ T cells from 11 control individuals were provided by A. Eugster. The entire data was subsequently separated into "positive"/"negative" training- and test data sets as follows.

The "positive" set for each epitope under consideration was extracted from the VDJdb as described. The "negative" set for each epitope was constructed by randomly sampling a set of TCRs specific to any of the other 16 epitopes. The size of this negative set was three times larger than the number of the positive TCRs for the epitope in consideration. The negative set was further expanded by randomly sampling TCRs from the negative control dataset to obtain twice the number of positive TCRs. Thus, the final set for each epitope had a negative/positive ratio of 5:1. For instance, for the epitope "ATDALMTGF" with 132 positive TCRs, there are 660 negative TCRs, of which 396 TCRs originated from the swapping of the TCRs from other epitopes while 264 TCRs were sampled from the negative control data. The data was then split randomly into a training and test set in the ratio of 80:20.

Models applied

In total, 23 TCR-epitope prediction models were trained and tested during the course of the workshop, which can be found in Table 1. Most approaches have been previously published, or are novel variants of existing models. These variations were created specifically for the workshop to explore the added benefit of integrating specific information types for the TCR-epitope prediction problem, notably the integration of specific TCR chain data. For comparison purposes, a 'Random' model was included, which produced scores between zero and one, based on the Numpy random number generator. Data, optimization, model and evaluation (DOME) machine learning reporting criteria can be found in Table S2.

Evaluation of prediction performance

Models were trained on the available training set for each epitope, or on all available training sets where appropriate for multi-label models. To decrease information leakage, the models were only allowed to train on the available data and any pre-training steps on TCR specificity were excluded. A decision value was then assigned to each TCR in the test data set for a given epitope with a blinded ground truth with respect to negative and positive samples. The ground truth of the test set was thus not available to workshop participants during model training and application.

Two data set setups were utilized to evaluate the models. The first included a mixture of positive and negative test data for each epitope, with the target epitope being known. Thus, each model had to score the likelihood of the TCRs included in the dataset binding to the specified epitope. From these results, the area under the ROC curve (AUC) was calculated. In the second setup, all positive test data from the previous setup were merged, and each method was challenged to provide predictions for every possible epitope seen during training. The epitopes were then ranked for each TCR from the most likely to the least likely according to the prediction scores of the model, and the rank of the true epitope was enumerated. As an epitope has multiple true TCRs, an average rank was calculated across all TCRs for one epitope. Within multi-label classification, this is equivalent to the coverage, i.e. the

Table 1
List of models tested.

Name	Chain	CDR usage	Distance/ Machine Learning	Reference
diffrbm_alpha*	alpha	cdr3	ML	[6]
diffrbm_beta*	beta	cdr3	ML	[6]
netTCR_CDR123_ab*	alpha-beta	cdr123	ML	[7]
netTCR_CDR3_ab*	alpha-beta	cdr3	ML	[7]
netTCR_CDR3_b*	beta	cdr3	ML	[7]
pMTnet	beta	cdr3	ML	[8]
Random	-	-	-	Numpy random number generator
SETE	beta	cdr3	ML	[9]
sonia_a*	alpha	cdr3	ML	[10]
sonia_b*	beta	cdr3	ML	[10]
sonia_ab*	alpha-beta	cdr3	ML	[10]
TCRbase_CDR123_ab*	alpha-beta	cdr123	Distance	https://services.healthtech.dtu.dk/service.php?TCRbase-1.0
TCRbase_CDR3_ab*	alpha-beta	cdr3	Distance	https://services.healthtech.dtu.dk/service.php?TCRbase-1.0
TCRbase_CDR3_b*	beta	cdr3	Distance	https://services.healthtech.dtu.dk/service.php?TCRbase-1.0
TCR-BERT	beta	cdr3	ML	[11]
TCRAI	alpha-beta	vjcdr3	ML	[3]
tcrdist3_a*	alpha	cdr123	Distance	[12]
tcrdist3_b*	beta	cdr123	Distance	[12]
tcrdist3_ab*	alpha-beta	cdr123	Distance	[12]
tcrex_a*	alpha	vjcdr3	ML	[13]
tcrex_b*	beta	vjcdr3	ML	[13]
tcrex_ab*	alpha-beta	vjcdr3	ML	[13]
TCRGP	alpha-beta	cdr123	ML	[14]
TITAN*	beta	cdr3	ML	[15]

* These models have been trained and applied by one of the original model authors, and thus can be considered having an advantage.

average number of labels to include to avoid missing the ground truth label [16]. All prediction scores were then collected and analyzed with the same evaluation script, which calculated the AUC / the average rank for each epitope, as well as the average over all epitopes.

Github and data repository

The data sets and evaluation scripts can be found at https://github.com/viragbioinfo/IMMREP_2022_TCRSpecificity.

Results

No relation between size of the training data and model performance

When benchmarked on the 17 MHC-epitope test data, all methods reached a non-random performance (AUC $>$ 0.5) for most epitopes, as can be seen in Fig. S1. This demonstrates that independent of the method used, it is possible to classify unseen TCRs for a given epitope within this dataset. Only the SARS-CoV-2-derived epitope NQKLIANQF consistently scored poorly or even randomly across all methods (AUC 0.539 mean $\pm$ 0.065 s.d.). In contrast, the easiest to predict epitope, NYNYLYRLF, also a known SARS-CoV-2 epitope, featured a near-perfect classification for most methods (AUC 0.956 mean $\pm$ 0.042 s.d.). TCRs for both the NQK and NYN epitopes were derived from the same MHC-dextramer study [17], thus there is no experimental difference in how the TCRs were collected or from which individuals. Furthermore, both epitopes had very similar training data sizes, namely 112 and 88 respectively. An overall analysis did show a strong relationship the performance and the similarity between TCR sequences in the training set (quantified as the average Levenshtein distance between the CDR3 sequences), as can be seen in Fig. 1. The strongly performing NYN epitope had an average Levenshtein distance of 12.6 within its training CDR3 sequences, while the NQK epitope featured an average distance of 17.7. This relationship was found to be consistent across most tested methods, as can be seen in Fig. S4. Therefore, those epitopes that had many highly similar TCR sequences seemed to be easier to classify within the presented held-out data set-up.

Distance-based methods provide a good baseline prediction

Methods that annotate unseen TCRs for binding a specific seen epitope can be broadly divided into two categories, namely distance-based methods and feature-based classification methods. Distance-based methods, such as TCRbase, mainly use a single distance metric to calculate the similarity between unseen TCRs and seen TCRs in the training data, independent of the epitope. In its simplest iteration, this can be the amount of amino acid mismatches between the two CDR3 sequences (i.e. the Hamming distance). If the distance is below a given threshold, an unseen TCR can be judged sufficiently similar to the training set TCR to be annotated with the same epitope. The distance metric itself can then be considered a confidence estimate of the method, where larger distances are considered less reliable annotations. More commonly, these distance-based methods rely on a k-nearest neighbor approach (k-NN), as is the case for all distance-based methods used in this benchmark. Within these k-NNs, it is the label of the k most similar TCRs that is used to predict the target epitope.

Feature-based methods are defined here as those that try to identify common patterns underlying the training set TCRs that bind a given epitope. These patterns then form the basis of predicting the binding preference of unseen TCRs. The underlying model is always a supervised machine learning method, where the known TCRs binding each epitope are provided as training data. The result is therefore a fitted model, which can then be applied to any unseen TCR. These are distinct from the distance-based methods as they try to learn which features are important for each epitope. However, as can be seen in Fig. 2, some distance-based approaches have a performance that is very close to those of the best performing feature-based methods. This supports the use of distance-based methods as a comparative baseline, as any new, more complex methodologies claiming to learn TCR-epitope patterns should be required to outperform basic matching algorithms.

An additional distinction can be made between those machine-learning approaches that train one model for each epitope separately (peptide-specific) or one model for all epitopes simultaneously (pan-specific). However, research has shown that even in the latter case, most pan-specific methods internally act as having a model per epitope [15]. This is because the generalisation of epitope-TCR pairing across different epitopes is still currently challenging due to the so far rather limited number of peptides characterized by TCR data [18]. This also matches with the found results in this study, as there seemed to be no consistent difference between these two approaches, as can be seen in Table S3.

Combining alpha and beta chain improves epitope prediction

Historically, the majority of methods to address the TCR-epitope specificity problem focus on the CDR3 sequence of the beta chain only for predictions, as historical TCR sequencing efforts have mainly only characterized the TCR beta chain. However, the TCR heterodimer complex consists of both alpha and beta chains, and both are known to

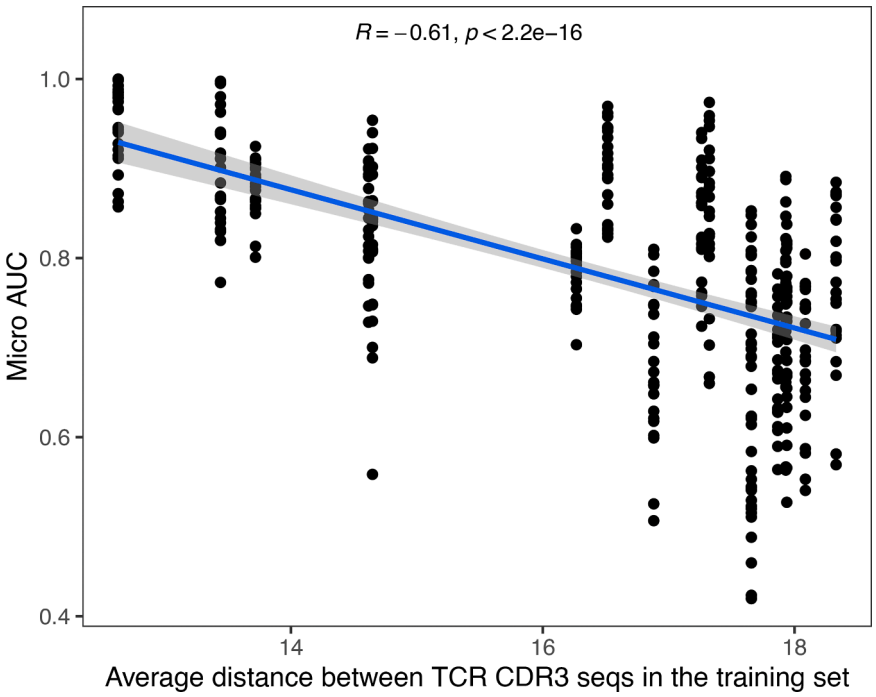

Fig. 1. Relationship between the prediction performance and similarity of TCR sequences in the training set. The x-axis shows the average Levenshtein distance between the CDR3 sequences in the training data set. The y-axis shows the micro AUC for each epitope and each method.

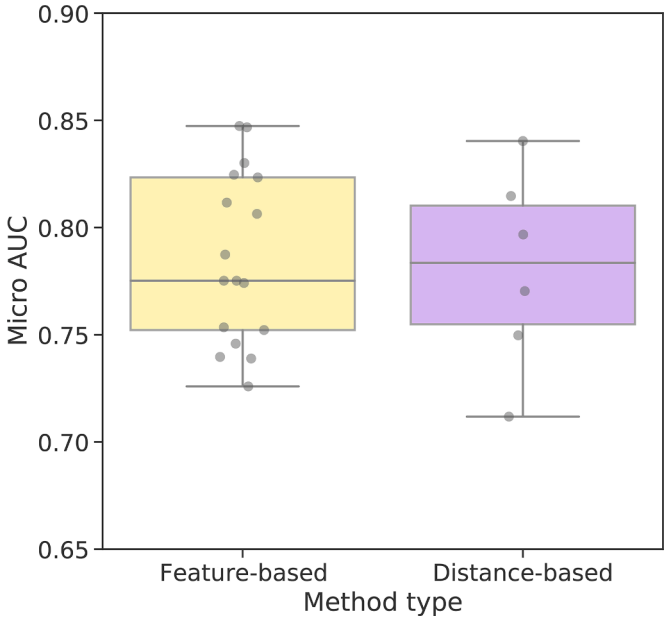

Fig. 2. Average microAUC by approach. Distance-based methods to the left annotate TCRs based on the similarity to known epitope-specific TCRs. Feature-based approaches use machine learning to learn associated features specific to an epitope, which is then used for classification.

make contacts with the epitope [19]. Paired alpha-beta chain information is still rare as it requires TCR sequencing at the single cell level, but prior studies have suggested that prediction performance can be increased by using information from both chains, alpha and beta [20]. The current benchmark data set only included the TCR-epitope entries with paired alpha- and beta chain information. This allowed a direct comparison between methods only using one of the two chains, by only using the relevant part of the input data. In this manner, the difference in prediction performance using either alpha, beta, or both chains could be assessed on the same TCR-epitope pair training- and test data. As can be seen in Fig. 3, methods that use both chains (alpha and beta) consistently outperform methods that only use a single chain (alpha or beta). The same trend can be seen at the level of individual AUCs, as seen in Table S3 and Fig. S2, where the most performant method per epitope is usually one that uses both chains. Furthermore, when clustering the individual performances, the use of single or both chains can be seen as the most prominent grouping factor, as can be seen in Fig. S3.

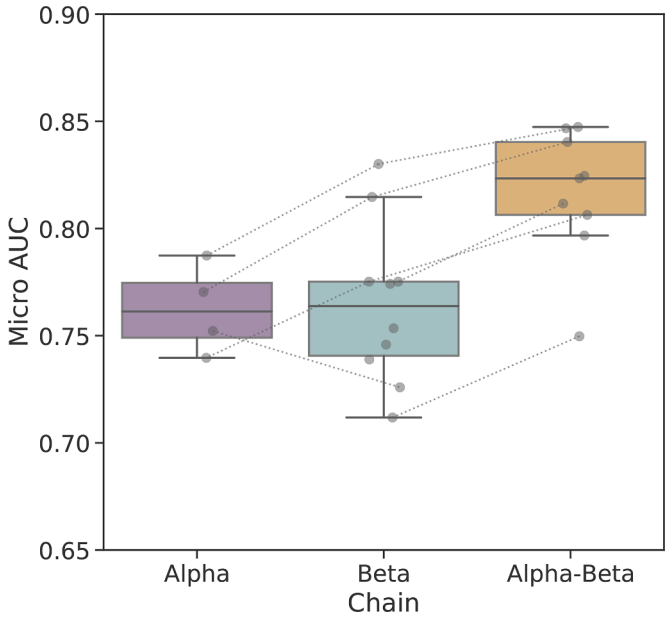

Fig. 3. Comparison of average microAUC of methods considering both TCR chains (Alpha-Beta) and methods that only consider the beta chain (Beta) or the alpha chain (Alpha). The lines denote those methods that use the same architecture but with different input.

Integrating V/J gene usage or CDR1/CDR2 improves epitope prediction

Most methods focus only on the CDR3 region of the beta (or alpha chain) of the TCR, as this region is the most variable and responsible for the majority of contacts with the epitope residues. Even though the CDR1 and CDR2 of a TCR are wholly determined by the V gene usage, they add a degree of variability to the chains and facilitate the crucial contacts between the TCR and the epitope-MHC complex. To investigate the complementary information contained in V/J genes and CDR1/CDR2 segments, we can divide methods into three categories based on their inputs: (i) Only requiring CDR3 amino acid sequence (CDR3) as the input, (ii) Requiring the CDR1/CDR2/CDR3 amino acid sequence (CDR123) as the input, and (iii) Requiring the CDR3 amino acid sequence as well as the V/J gene usage (VJCDR3). From the average performance results, as seen in Fig. 4, we observed that methods that consider the CDR1/CDR2 regions, either directly (CDR123) or indirectly (VJCDR3), outperform those that only consider the CDR3 amino acid sequence. However, considering the amino acid sequence of the CDR1 and CDR2 regions does not, in this setting, seem to have any improvement over simply considering the V gene annotation itself. Some methods, such as tcrdist3, do utilize a prior alignment of these regions to calculate their difference (as well as an additional variable loop between CDR2 and CDR3 which is given the name “CDR2.5”), and thus already formulate the concept that some CDR1/CDR2 are more similar than others.

Epitope ranking mostly, but not always follows binary classification performance

Prior sections have focused on method performance as calculated by the AUC of the ROC curve, thus the trade-off between false negatives and false positives for the binary classification problem. However, epitope-TCR pairing is not a true binary classification problem, as the target epitope is commonly unknown. Thus each TCR needs to be matched with any epitope, and is thus a multi-label classification problem. For this reason, we considered the epitope rank as an alternative metric, where the prediction scores for each of the 17 epitopes is compared for every TCR in the test set, and the rank of the ground truth is averaged across all TCRs for a given epitope. In this instance, a lower score is considered better and a score of 1 signifies a perfect prediction for an epitope. As this required applying the model for each epitope on the same data set, this could not be accomplished for every method previously tested. Some methods were not able to create models for each epitope, others did not provide the option to predict per epitope. As can be seen in Fig. 5, the relative performance across models between AUC and ranked epitope has remained relatively stable. The most performant method with regards to AUC remains the most performant method with regards to the epitope rank. This is not unsurprising as the AUC already captures the difference in prediction scores between a positive pair and a negative pair. The key difference is that the epitope rank focuses solely on the positive pairs. Indeed, the overall ranking between some methods does change, as can be seen in supplemental Table S4. Most notably, basic clustering approaches, such as TCRbase, score relatively better on average epitope rank than AUC. This can be attributed to the fact that these methods do not hold the concept of a negative background, as they are searching for TCRs similar to those TCRs known to bind an epitope. Thus, it can be expected that these methods are less able to distinguish negative samples, despite being proficient in the annotation of positive samples with the right epitope.

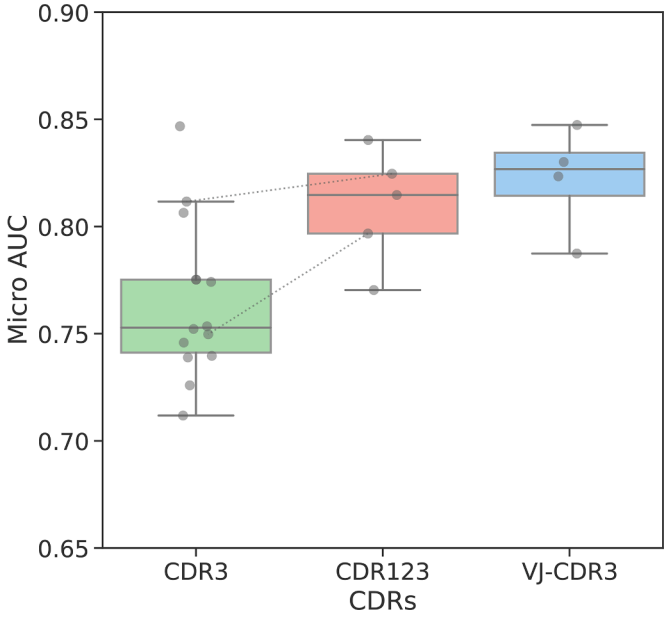

Fig. 4. Average performance by CDR region usage. From left to right, methods using only CDR3 amino acid sequence, methods that use CDR3 and V/J genes, and methods that use CDR1,CDR2,CDR3 amino acid sequences as input. The lines denote those methods that use the same architecture but with different input.

Discussion

Limitations of the benchmark

While many methods were included into this benchmark, our effort was not exhaustive. The wealth of currently existing methods for TCR-epitope prediction makes it impossible to compare them all in a single effort. In addition, many methods that have been described in the literature unfortunately lack a publicly accessible code/interface.

In addition, ground-truth true negative data does not exist within the TCR-epitope context. Two strategies to circumvent this problem were combined for this benchmark: swapped TCR-epitope pairs and unrelated repertoires from healthy individuals without epitope knowledge as background sequences. Both have their advantages and disadvantages. Swapping TCR-epitope pairs by considering the TCRs positive for one epitope as being negative for another, means that methods can in theory learn the TCR patterns associated with the negative epitopes in training and test data. Utilizing unrelated, healthy repertoires has the disadvantage that there may be an experimental and/or biological bias in the data. This lends itself to the possibility that any model may learn the distinction between positive and negative based on such biases alone. In addition, healthy repertoire data is not free of potential misannotations, due to the presence of T cells specific for common dominant epitopes also present in the positive data. Simulated data with a known ground truth could not be used here as we wished to evaluate the contribution of CDR1/2, alpha/beta, which is a required assumption prior to simulation. Using the simulated datasets as negatives only, would have created more opportunity to bias. As highlighted by the difference in performance ranking based on AUC and epitope ranking, classifiers potentially do have a tendency to learn patterns within the negative data that changes how they will consider the problem. This benchmark is far too limited to provide a solid answer to what negative strategy is most appropriate, and this remains an important topic for future research.

Another key choice involved the lack of removal of similar TCRs from the test set compared to the training data set. This would have required a strict definition of what constitutes a sufficiently similar TCR. There does not currently exist a common consensus of how different a TCR can be before it can be expected to no longer bind the same epitope. Furthermore, whether to include or exclude similar TCRs from the datasets is a decision that depends fully on the downstream application. Mostly, the end goal of these methods is to identify epitope-specific TCR pairs among a large TCR sequence repertoire extracted independently. It is already known that public clonotypes do occur across different

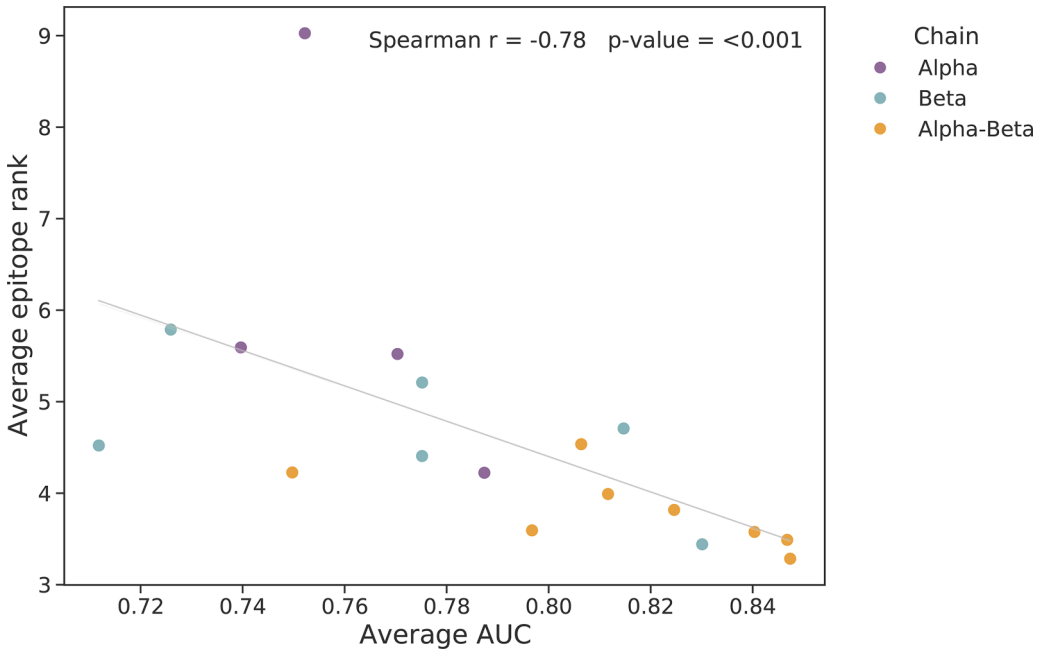

Fig. 5. Average performance of methods based on microAUC in the binary classification problem (x-axis) plotted against the epitope rank performance for multi-label prediction (y-axis). Each dot in the plot represents a single method.

individuals, and that small subsets of TCRs can be successfully annotated with their epitopes just by matching their TCR sequences to annotated TCR sequences of a database [21–23]. However, public clonotypes are the exception rather than the rule. Removal of the highly similar TCRs from the test set does allow evaluation to what degree these methods are capable of linking distant TCRs to the correct epitope. This results in a far more limited positive test set, with more focus on the classification of the negative samples.

Finally, many methods are hampered in their success by the current limited training data set size. Thus their measured performance is not representative of their theoretical potential with an unrestricted training data set. For example, several methods have shown improvements by utilizing transfer learning or pre-training steps, which are not possible with this limited benchmark [8]. In addition, methods using only beta-chain data can typically rely on a larger dataset than methods with both alpha and beta chains, which may explain their poorer performance in this instance. The performing data set curation from paired chain data also had an impact on the size of the training and the test set of a few epitopes. For instance, two TCRs would be considered for evaluation if their alpha chains are different, implying that they are overall different but they may collapse to a single TCR if their beta chain is identical. Furthermore, the dataset used here and split into test- and training data sets was derived from the same experiments. This was necessary given the current scarcity of available TCR-epitope specific data. Ideally, the train-test split would be derived from independent experiments to avoid any experimental information bleed, however this can only be done for a very limited number of epitopes.

The shape and scope of future benchmarks

The goal of this study was to evaluate the possibility of a benchmark between methods and highlight important lessons learned from this pilot study. Due to the aforementioned limitations, no strong conclusions should be made about the superiority of one approach to another at this point. Regardless of the limitations, an independent benchmark has become necessary due to a rapid surge in the number of the TCR-epitope prediction tools. It is no longer feasible or reasonable to expect a single study to compare to all other existing methods. Furthermore, as highlighted in this study, the choice of data set on which to evaluate can have a large impact on the performance. Related is the commonly accepted phenomenon that a method will always score best when applied by its own authors and on the data set in the paper where it is introduced.

Other prediction-focused fields have embraced the idea of an open competition where methods are benchmarked on the same never-before-seen data set. This is a way to independently evaluate the wide variety of methods that are available, but also to drive the field forward and enable it to reach new heights. As a conclusion of the ImmRep TCR-epitope specificity workshop, we strongly encourage the organization of similar competitions focused on the TCR-epitope prediction problem. The most essential choice will be the origin of the test data as many options are available, from a stratified approach as highlighted here, to the integration of simulated datasets with a known ground truth [24]. The ideal data set for such a challenge would unequivocally be an unpublished independent data set with both TCRs with known epitope-specificity, as well as “negative” TCRs that do not bind these epitopes. As highlighted in this study, it would ideally involve paired alpha-beta TCR sequence data, and would therefore likely be derived from a single cell sequencing experiment. In addition, the use of oligo-tagged multimers would enable both identification of those TCRs that are specific for an epitope, along with those that are likely not. Furthermore, this data set should contain multiple previously studied epitopes, so as to compare the epitope rank beyond the straight-forward classification problem. The technology to create such a dataset is currently available, and therefore only requires the willingness of either funders, institutes or companies to provide it.

Conclusions

This study contains an initial large-scale benchmark of epitope-TCR prediction methods. It was not meant to be exhaustive, nor should the results be overly interpreted as the data set was limited and the evaluation superficial. Several important observations could nevertheless be established. The use of paired-chain alpha-beta data, as well as CDR1/2 or V/J information, improves classification when this data is available, independent of the underlying approach. Straight-forward clustering approaches can achieve a respectable performance and should be used as a valid benchmark for future studies. Finally, there is a large need for a true independent benchmark on the myriad of methods within the

field.

CRediT authorship contribution statement

Pieter Meysman: Formal analysis, Data curation, Visualization, Writing – original draft. **Justin Barton:** Formal analysis, Data curation, Writing – review & editing. **Barbara Bravi:** Formal analysis, Data curation, Formal analysis, Writing – review & editing. **Liel Cohen-Lavi:** Formal analysis, Data curation, Visualization, Writing – review & editing. **Vadim Karnaukhov:** Formal analysis, Data curation, Visualization, Writing – review & editing. **Elias Lilleskov:** Formal analysis, Data curation, Writing – review & editing. **Alessandro Montemurro:** Formal analysis, Data curation, Writing – review & editing. **Morten Nielsen:** Formal analysis, Data curation, Writing – review & editing. **Thierry Mora:** Formal analysis, Data curation, Writing – review & editing. **Paul Pereira:** Formal analysis, Data curation, Writing – review & editing. **Anna Postovskaya:** Formal analysis, Data curation, Writing – review & editing. **María Rodríguez Martínez:** Formal analysis, Data curation, Writing – review & editing. **Jorge Fernandez-de-Cossio-Diaz:** Formal analysis, Data curation, Writing – review & editing. **Alexandra Vujkovic:** Formal analysis, Data curation, Writing – review & editing. **Aleksandra M. Walczak:** Formal analysis, Data curation, Writing – review & editing. **Anna Weber:** Formal analysis, Data curation, Writing – review & editing. **Rose Yin:** Formal analysis, Data curation, Writing – review & editing. **Anne Eugster:** Conceptualization, Formal analysis, Data curation, Writing – review & editing. **Virag Sharma:** Conceptualization, Formal analysis, Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: This manuscript concerns a meeting report of the IMMREP2022 workshop on TCR-epitope specificity. Many of the authors involved have been involved in the creation of the tools being benchmarked, this has been clearly listed in the report (see Table 1).

Data availability

The data sets and evaluation scripts can be found at https://github.com/viragbioinfo/IMMREP_2022_TCRSpecificity.

Acknowledgments

We wish to thank the Max Planck Institute for the Physics of Complex Systems in Dresden and the Max Planck Society for their generous financial and physical support enabling the ImmRep 2022 TCR-epitope specificity workshop. We wish to thank the VDJdb resource for providing the dataset that formed the basis of this paper. In addition, we wish to thank the valuable input from Prof. Victor Greiff on the manuscript.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.immuno.2023.100024.

References

[1] Shugay M, Bagaev DV, Zvyagin IV, Vroomans RM, Crawford JC, et al. VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res 2018;46(D1):D419–27. Available, http://academic.oup.com/nar/article/doi/10.1093/nar/gkx760/4101254/VDJdb-a-curated-database-of-Tcell-receptor. Accessed 6 November 2017.

[2] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, et al. The immune epitope database 2.0. Nucleic Acids Res 2010;38:D854–62. Available, http://nar.oxfordjournals.org/content/38/suppl_1/D854. Accessed 16 July 2014.

[3] Zhang W, Hawkins PG, He J, Gupta NT, Liu J, et al. A framework for highly multiplexed dextramer mapping and prediction of T-cell receptor sequences to antigen specificity. Sci Adv 2021;7:5835–49. Available, https://www.science.org/doi/10.1126/sciadv.abf5835. Accessed 15 December 2022.

[4] Goncharov M, Bagaev D, Shcherbinin D, Zvyagin I, Bolotin D, et al. VDJdb in the pandemic era: a compendium of T-cell receptors specific for SARS-CoV-2. Nat Methods 2022;19:1017–9. 20229 19Available, https://www.nature.com/articles/s41592-022-01578-0. Accessed 15 December 2022.

[5] 10x Genomics. A new way of exploring immunity: linking highly multiplexed antigen recognition to immune repertoire and phenotype. 10x Genomics; 2019.

[6] Bravi B, Gioacchino ADi, Fernandez-De-Cossio-Diaz J, Walczak AM, Mora T, et al. Learning the differences: a transfer-learning approach to predict antigen immunogenicity and T-cell receptor specificity. Biorxiv 2022. 2022.12.06.519259. Available, https://www.biorxiv.org/content/10.1101/2022.12.06.519259v1. Accessed 15 December 2022.

[7] Montemurro A, Schuster V, Povlsen HR, Bentzen AK, Jurtz V, et al. NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. Commun Biol 2021;4:1–13. 2021 4:1Available, https://www.nature.com/articles/s42003-021-02610-3. Accessed 15 December 2022.

[8] Lu T, Zhang Z, Zhu J, Wang Y, Jiang P, et al. Deep learning-based prediction of the T-cell receptor–antigen binding specificity. Nat Mach Intell 2021;3:864–75. 2021 3:10Available, https://www.nature.com/articles/s42256-021-00383-2. Accessed 15 December 2022.

[9] Tong Y, Wang J, Zheng T, Zhang X, Xiao X, et al. SETE: sequence-based ensemble learning approach for TCR epitope binding prediction. Comput Biol Chem 2020;87:107281. https://doi.org/10.1016/J.COMPBIOLCHEM.2020.107281.

[10] Sethna Z, Isacchin G, Dupic T, Mora T, Walczak AM, et al. Population variability in the generation and selection of T-cell repertoires. PlOS Comput Biol 2020;16:e1008394. Available, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008394. Accessed 15 December 2022.

[11] Wu K, Yost KE, Daniel B, Belk JA, Xia Y, et al. TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. Biorxiv 2021.2021.11.18.469186Available, https://www.biorxiv.org/content/10.1101/2021.11.18.469186v1. Accessed 15 December 2022.

[12] Mayer-Blackwell K, Schattgen S, Cohen-Lavi L, Crawford JC, Souquette A, et al. TCR meta-clonotypes for biomarker discovery with TCRdist3 enabled identification of public, hla-restricted clusters of SARS-COV-2 TCRS. Elife 2021;10. https://doi.org/10.7554/ELIFE.68605.

[13] Gielis S, Moris P, Bittremieux W, De Neuter N, Ogunjimi B, et al. Detection of enriched T-cell epitope specificity in full T-cell receptor sequence repertoires. Front Immunol 2019;10:2820. Available, https://www.frontiersin.org/article/10.3389/fimmu.2019.02820/full. Accessed 6 August 2020.

[14] Jokinen E, Huuhtanen J, Mustjoki S, Heinonen M, Lähdesmäki H. Predicting recognition between T cell receptors and epitopes with TCRGP. PlOS Comput Biol 2021;17:e1008814. Available, https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1008814. Accessed 15 December 2022.

[15] Weber A, Born J, Rodriguez Martínez M. TITAN: T-cell receptor specificity prediction with bimodal attention networks. Bioinformatics 2021;37:i237–44. Available, https://academic.oup.com/bioinformatics/article/37/Supplement_1/i237/6319659. Accessed 15 December 2022.

[16] Wu XZ, Zhou ZH. A unified view of multi-label performance measures. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017,PMLR. 8; 2017. p. 5778–91. Available, https://proceedings.mlr.press/v70/wu17a.html. Accessed 15 December 2022.

[17] Minervina AA, Pogorelyy MV, Kirk AM, Crawford JC, Allen EK, et al. SARS-CoV-2 antigen exposure history shapes phenotypes and specificity of memory CD8+ T-cells. Nat Immunol 2022;23:781–90. 2022 23:5Available, https://www.nature.com/articles/s41590-022-01184-4. Accessed 15 December 2022.

[18] Moris P, De Pauw J, Postovskaya A, Ogunjimi B, Laukens K, et al. Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. Brief Bioinform 2020. https://doi.org/10.1093/bib/bbaa318. AvailableAccessed 3 November 2020.

[19] Danska JS, Livingstone AM, Paragas V, Ishihara T, Garrison Fathman C. The presumptive CDR3 regions of both T cell receptor alpha and beta chains determine T-cell specificity for myoglobin peptides. J Exp Med 1990;172:27–33. Available, http://rupress.org/jem/article-pdf/172/1/27/1100603/27.pdf. Accessed 15 December 2022.

[20] 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. Front Immunol 2021;12:1436. https://doi.org/10.3389/FIMMU.2021.664514/BIBTEX.

[21] Simnica D, Schultheiß C, Mohme M, Paschold L, Willscher E, et al. Landscape of T-cell repertoires with public COVID-19-associated T-cell receptors in pre-pandemic risk cohorts. Clin Transl Immunol 2021;10:e1340. Available, https://onlinelibrary.wiley.com/doi/full/10.1002/cti2.1340. Accessed 15 December 2022.

[22] Kedzierska K, Day EB, Pi J, Heard SB, Doherty PC, et al. Quantification of repertoire diversity of influenza-specific epitopes with predominant public or private TCR usage. J Immunol 2006;177:6705–12. Available, https://journals.aai.

org/jimmunol/article/177/10/6705/74621/Quantification-of-Repertoire-Diversity-of. Accessed 15 December 2022.

[23] Benati D, Galperin M, Lambotte O, Gras S, Lim A, et al. Public T-cell receptors confer high-avidity CD4 responses to HIV controllers. J Clin Investig 2016;126: 2093–108. https://doi.org/10.1172/JCI83792.

[24] Weber CR, Akbar R, Yermanos A, Pavlović M, Snapkov I, et al. immuneSIM: tunable multi-feature simulation of B and T-cell receptor repertoires for immunoinformatics benchmarking. Bioinformatics 2020;36:3594–6. Available, https://academic.oup.com/bioinformatics/article/36/11/3594/5802461. Accessed 15 December 2022.