TCR-H:
Machine Learning Prediction of T-cell Receptor Epitope Binding on Unseen Datasets

Rajitha Rajeshwar T.[1,2,3], Omar Demerdash[1,3], Jeremy C. Smith[1,2,3].

[1]UT/ORNL Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6309, United States
[2]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, Tennessee 37996-1939, United States
[3]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6309, United States

Abstract
AI/ML approaches to predicting T-cell epitope specificity achieve impressive results on test sets which include epitope sequences that are also part of the training set, but fail to generalize to test sets consisting of epitopes and TCRs that are not included in the training set i.e., that are *unseen*. We present TCR-H, a supervised classification ML model using physicochemical features trained on the largest dataset of binding and non-binding data available to date. We used only experimentally validated non-binders as negative datapoints. TCR-H exhibits an area under the curve of the receiver operator characteristic (AUC of ROC) of 0.87 for epitope *hard splitting* (i.e test sets with all epitopes unseen), 0.92 for TCR hard splitting and 0.89 for strict splitting in which neither the epitopes nor the TCRs are present in the training data. TCR-H may thus represent a significant step towards general applicability of epitope:TCR specificity prediction.

Introduction
Cytotoxic T-cells are central to the adaptive immune response. Critical to adaptive immune system activation is the specific binding of T-cell receptors (TCR) to peptide epitopes presented by the Major Histocompatibility Complex (MHC) on the surface of antigen-presenting cells. The remarkable diversity of TCRs, estimated at approximately 10^{15}-10^{61} (Dash et al., 2017), results from genetic recombination of somatic-cell DNA encoding V (variable), D (diversity), and J (joining) segments(Bradley & Thomas, 2019; Rudolph et al., 2006). Due to this high diversity arising from a combinatorial explosion of different VDJ recombination possibilities, T-cells are capable of recognizing a large variety of epitopes that can be exogenous or derived from endogenous mutated proteins (Tippalagama et al., 2023).

A significant goal in immunology is the reliable prediction of which epitopes bind to which T-cell receptors. Knowing this would aid in the design and development of vaccines and immunotherapies and would help us understand how the immune system distinguishes between self and non-self antigens. Furthermore, monitoring the presence of TCRs associated with specific epitopes can be used in disease diagnosis, prognosis, and monitoring disease progression, which is particularly important in the context of infectious diseases, where tracking T-cell responses can provide insights into the immune system's ability to recognize and control pathogens.

The vast diversity in TCR sequences and potential epitopes makes it challenging to develop a generalized computational model for TCR:epitope binding prediction. Moreover, prediction is further complicated by TCRs can exhibit cross-reactivity, recognizing multiple epitopes and that TCR recognition may change over time or be context dependent, influenced by environmental factors.

One approach in principle to solving the TCR:epitope specificity problem would be to develop accurate methods for predicting the 3D structures of TCR:epitope:MHC ternary complexes and then to use the results to predict binding strength based on the physics involved. However, predicting these 3D structures and interactions accurately is challenging. The 3D interaction between TCRs and epitopes bound to MHC is highly variable and small changes in peptide sequences can significantly impact TCR recognition. Furthermore, the conformational flexibility of TCRs and the peptides they recognize further complicates this approach.

Despite the myriad above problems we feel that TCR:epitope recognition can benefit from considering physicochemical properties determining peptide recognition, and that useful solutions to the problem may well explicitly incorporate structural and/or physicochemical concepts derived from sequence. In recent work we demonstrated how 3D structure can be incorporated into computational methods to improve prediction of peptide binding to MHC(Aranha et al., 2020), and how 3D MHC pocket similarity determines peptide binding strength. Further, we and others have shown that certain 3D interactions are commonly found in epitope binding to TCRs(Rajitha Rajeshwar & Smith, 2022).

TCRs are heterodimers consisting of two chains: the α chain generated by VJ recombination and the β chain generated by VDJ recombination. Binding primarily occurs through three complementarity-determining regions (CDRs) found on each of these chains. Although in 3D structures of TCR:epitope:MHC ternary complexes significant contacts are made outside the CDR3 region, this region most directly interacts with the peptide epitope, with CDR3β in particular, forming direct contacts with the epitopes and MHC and displaying particularly high sequence diversity (Dai et al., 2008; Garcia et al., 2009; Morris & Allen, 2012; Szeto et al., 2021). Hence, in computational approaches CDR3β has commonly been taken as an approximation to the specificity determinant TCR sequence.

With the increasing availability of TCR and epitope sequencing data from high-throughput sequencing techniques in publicly available data resources (Jokinen et al., 2021), AI and ML approaches have been able to be used to predict TCR CDR3β binding to epitopes presented by MHC class 1 (MHC-I) (Hudson et al., 2023). These tools apply methods ranging from relatively simple ML models such as Random Forest and clustering (Chronister et al., 2021; Dash et al., 2017; Jokinen et al., 2021; Pham et al., 2023) to various forms of deep learning-based AI techniques, including convolutional and recurrent neural networks (Bravi et al., 2023; Cai et al., 2022; Darmawan et al., 2023; Gao, Gao, Li, et al., 2023; Jiang et al., 2023; Jokinen et al., 2023; Myronov et al., 2023; D. Wang et al., 2023; Z. Wang & Shen, 2023; Weber et al., 2021; Zhang et al., 2023).

Many of the AI/ML methods hitherto developed achieve impressive results on 'random split' test sets that include TCR and epitope sequences that are also included in the training set. However, for general applicability a model must be able to predict binding for a "strict split" with *unseen* TCRs and epitopes i.e., in which neither the TCRs nor the epitopes tested are in the training set. Success in this endeavor would perhaps indicate an algorithm has been discovered that has learned the general principles of TCR:epitope recognition and would be a major step towards a holy grail of immunology - the accurate computational mapping of TCRs to their cognate epitopes.

The strict split goal can be approached stepwise. In a first step datasets can be derived with an epitope "hard split" i.e., in which the test set contains unseen epitopes or a TCR hard split in which the test set contains unseen CDR3βs. When using an epitope hard split current methods mostly fail, with scores mostly falling to almost random (AUC of the ROC ~0.5; Dens et al., 2023;

Grazioli et al., 2022). TCR hard splitting has been addressed in (Cai et al., 2022) with AUC values of 0.77 or below. Strict splitting attempts have also been published with AUC values ranging between 0.5 and 0.7 (Korpela et al., 2023; Weber et al., 2021)

Given the above considerations, we attempted to address the unseen epitopes and TCRs problem by developing an ML supervised binary classification model incorporating sequence-based physicochemical descriptors of CDR3β and epitopes using the largest dataset of binding and non-binding data available to date. Our model possesses two original characteristics: the use of only experimentally validated non-binders as negative datapoints in the training and testing sets and the use of a highly diverse set of physicochemical features calculated over entire peptide sequences rather than single amino acid residues.

Existing machine and deep learning methods vary in terms of the datasets used, mostly relying on randomly generated data as the negative/non-binding dataset. However, there is the possibility that any given randomly generated sequence assumed to not bind might actually bind. Furthermore, it has been shown that the use of randomly generated sequences for negative binders leads to overestimation of model accuracy (Dens et al., 2023; Grazioli et al., 2022). Therefore, we include in our dataset only negatives that have been experimentally validated as non-binders.

The choice of supervised ML over deep learning was motivated in part by the ability of the model to be based on explicit features posited to be important for binding, which in turn may aid future model interpretation, explainability and rationality. Furthermore, it has been shown that the use of features that lend themselves to ML models leads to better performance as well (Pham et al., 2023). Most ML models have been sequence based, encoding sequences using either the BLOSUM substitution matrix (Cai et al., 2022; Darmawan et al., 2023; Gao, Gao, Li, et al., 2023; Pham et al., 2023) and/or Atchley factors(Atchley et al., 2005). The BLOSUM matrix applies a score for amino-acid substitutions, with positive values indicating conservative substitutions and negative values indicating non-conservative ones. Atchley factors are multidimensional, composite features for each amino acid, including polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge, derived using unsupervised ML on primarily physicochemical features (Jiang et al., 2023; Lu et al., 2021; Moris et al., 2021). Here, we use physicochemical features calculated over entire epitope and CDR3β sequences i.e., not just individual amino-acid residues.

Various ML methods are tested: Random Forest (RF), Gradient Boosting trees (GBT), eXtreme Gradient Boosting (xGBT) and Support Vector Machines (SVM). The best performing, an SVM model, which we name TCR-H, exhibited AUC of the ROC metrics of 0.87, 0.92 and 0.88 for the epitope hard split, TCR hard split and the epitope/TCR strict split, respectively. TCR-H thus represents a significant step towards general applicability of computational prediction of epitope:TCR specificity in biology and medicine.

Results and Discussion

Our binary classification was of binding or non-binding. The performance of the models was evaluated in terms of the sensitivity, specificity, accuracy and F1 scores and the area under the receiver-operator characteristic (ROC) curve (AUC) (see SI for definitions). The metrics of the ROC curve illustrate the trade-off between sensitivity (true positive rate) and specificity (true negative rate). The AUC quantifies the ability of the model to distinguish between the two classes, where a classifier with perfect accuracy yields an AUC of 1.0, while a random classifier would yield an AUC of 0.5 or less.

We first present ML performance results on epitope hard split data, i.e. in which the test set contains only unseen epitopes. The resulting performance metrics on the test set for RF, GBT, xGBT and SVM are presented in Table 1. Taking all the performance metrics into consideration, the ensemble models RF, GBT and xGBT do not perform particularly well, especially for AUC and specificity. Only in terms of recall and F1-score do these models perform somewhat decently; however, it should be noted that the F1, a composite metric that is a function of both precision and recall, is elevated here only by the recall, with precision being relatively poor. In contrast, for the model trained with SVM, performance is much better across the board, with the AUC of ROC, accuracy, and precision—all poor in the ensemble tree models—improving dramatically. The AUC of ROC is 0.80 (Figure S1), a significant improvement on previously published efforts.

In an attempt to further improve the SVM performance, we calculated correlations between pairs of features. We found that by removing one of the features of any pair that had a correlation > 0.8 improved the model's performance to an AUC of ROC of 0.87 (Fig. 1A) as well as improving most other statistical metrics (we selected the feature to be removed from the correlated pair at random). The AUC of ROC for the epitope hard split with this model (which we dub "TCR-HE") is close to values obtained for random splits in other models. The model has very high sensitivity i.e., a large fraction of the true positives are predicted positives. The specificity (true negative rate) is also high, although perhaps not quite as impressive.

Next, we trained a model using SVM and the same set of features that are uncorrelated with each other using a hard split in the TCR CDR3βs (a model that may be referred to as TCR-Hβ. The results on the independent test set, provided in Table 2, show that again the model performs accurately with an AUC of ROC of 0.92 and other metrics also being high. Finally, we report metrics on the strict split, which gives an AUC of 0.89 and again the other metrics are good.

Table 1. Comparison of ML models for epitope hard split test set. TP=number of true positives; TN=number of true negatives; FP=number of false positives; FN=number of false negatives. TCR-HE has correlations removed.

Model	**AUC of ROC**	**TP**	**TN**	**FP**	**FN**	**Accuracy**	**Precision**	**Recall**	**Specificity**	**F1**
XGB	0.51	31859	654	21817	104	0.597	0.593	0.996	0.029	0.744
GBT	0.54	31248	2329	20142	715	0.617	0.608	0.977	0.104	0.749
RF	0.5	31963	0	22471	0	0.587	0.587	1.00	0.0	0.739
SVM	0.80	30403	14593	7878	1560	0.826	0.794	0.951	0.649	0.865
TCR-HE	0.869	29567	18297	4174	2396	0.879	0.876	0.925	0.814	0.9

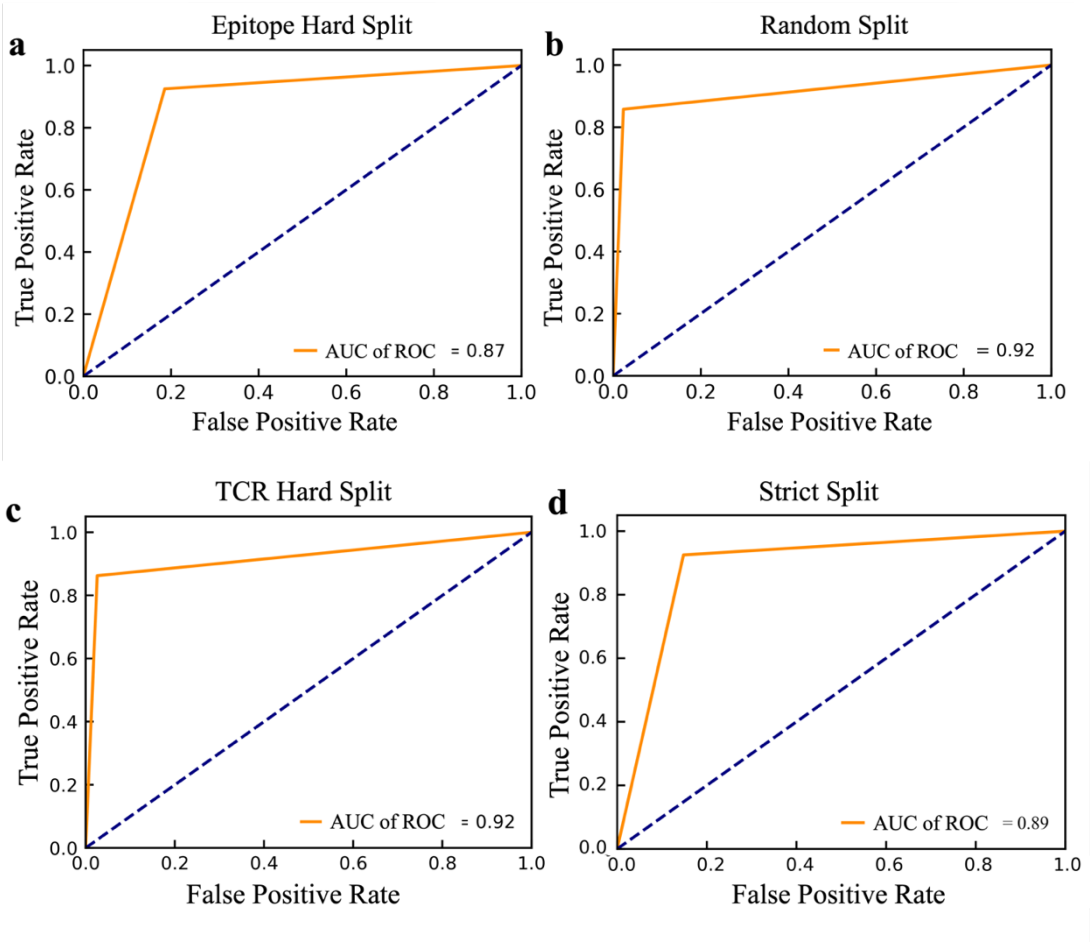

Figure 1. AUC of ROC on the independent test set for the TCR-H model with uncorrelated features trained with a) epitope hard split for the training and independent test set, and b) random split. C) hard split of the TCR CDR3β D) strict split

For completeness, and to compare against what has been typically done in prior predictive model development studies, we also performed a benchmark of TCR-H using the commonly employed random split. For this, the AUC of the ROC of TCR-H model is 0.91 (Table 2) which is comparable to existing methods. In particular, it is equal to the performance of an SVM model built using sequences encoded using BLOSUM62 for which the resulting matrices were transformed into feature vectors through flattening and concatenation (Pham et al., 2023).

Table 2. Performance metrics of TCR-H trained and tested on TCR hard split (TCR-Hβ□□□strict split (TCR-HβE) and random split (TCR-HRS).

Model	AUC of ROC	TP	TN	FP	FN	Accuracy	Precision	Recall	Specificity	F1
TCR-Hβ	0.918	18760	28780	766	2994	0.93	0.96	0.86	0.97	0.91
TCR-HβE	0.888	29570	19143	3328	2393	0.89	0.898	0.92	0.85	0.91
TCR-HRS	0.917	18382	28771	691	3045	0.93	0.96	0.86	0.98	0.91

Comparison with previous models.

The field of computational epitope-TCR binding prediction is very active, and it is incumbent on us to compare the performance of TCR-H with models reported in the literature (Cai et al., 2022; Gao, Gao, Fan, et al., 2023a; Gao, Gao, Li, et al., 2023; Jiang et al., 2023; Lu et al., 2021; Montemurro et al., 2021; Moris et al., 2021; Pham et al., 2023). Specifically, we compared the performance of TCR-H against that of reported epitope hard-split AUC of ROC metrics (Figure 2). Although training and test datasets differ across methods, all adhere to the hard-split procedure. TCR-H outperforms these models in the most rigorous evaluation of model transferability, i.e., the hard split that ensures none of the epitopes in the independent test set are present in the training set.

Some methods, including Pan-Peptide and epiTCR, have in some epitope hard-split tests shown significant improvement over random, with AUC of ROC scores of ~0.75. Notwithstanding, TCR-H performed significantly better than these test results, with an AUC of ROC of 0.87. Furthermore, the pitfalls of negative data bias in the TCR epitope specificity challenge have recently been further emphasized, with the performance of Pan-peptide reduced to random (AUC of ROC = 0.49) when a true hard split was applied (Dens et al., 2023).

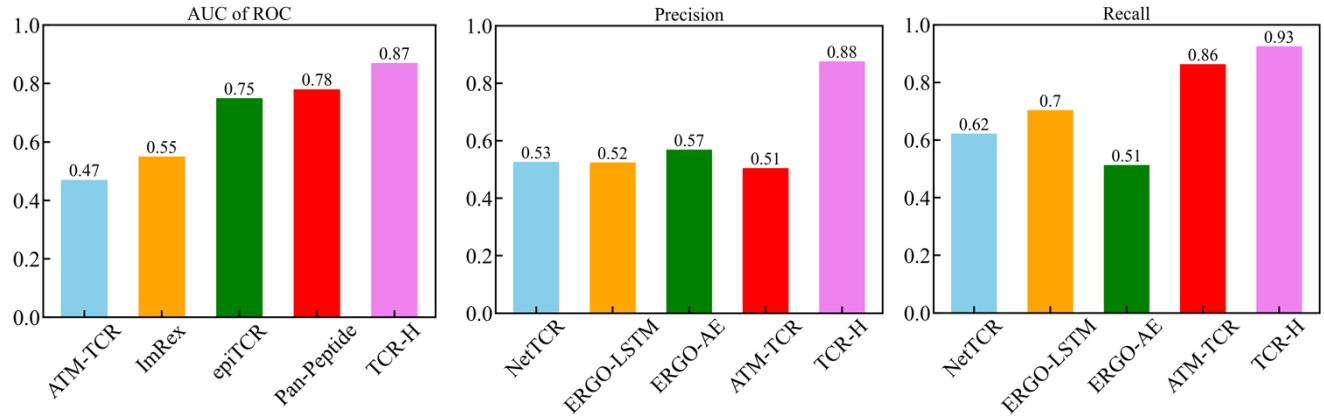

Figure 2. AUC of ROC, Precision and Recall of TCR-HE compared with that of some previously reported epitope hard split models (Cai et al. (2022); Moris et al., 2021; Pham et al., 2023 ;Gao, Gao, Fan, et al., 2023).

Metrics other than the AUC of ROC are important in potential applications of these methodologies but are not always calculated. In particular, precision (which measures the ratio of true binders to the total number predicted to be binders) and recall (the fraction of true binders correctly identified as such by the model) are of interest. Fig. 2 also shows Precision and Recall metrics for NetTCR, ERGO-LSTM, ERGO-AE and ATM-TCR, taken from Cai et al. (2022), and evaluated on the same independent epitope hard split data set which was generated with a hard split. It is clear from the figure that TCR-HE performs well also on hard-split precision and recall.

Conclusion.

Building on our previous work on 3D structure and physicochemical approaches we present here an ML model, TCR-H, capable of predicting unseen TCR:epitope binding *i.e.,* in for which neither the epitope nor the TCR have been included in set of data used to train the model. This contrasts with most previous efforts that used random splits. TCR-H exhibits an AUC of ROC of 0.87 for epitope hard splitting, 0.92 for TCR hard splitting and 0.89 for splitting in which neither the epitopes nor the TCRs are present in the training data. Other metrics such as precision and recall are also excellent.

AI/ML methods can be notoriously fickle and depend on the data sets used for training and testing. For example, Fig. S3 shows epitope hard split AUC of ROC data with TCR-H and various previously-reported models. Testing of a given model on different specific epitope hard split data sets, leading to significantly different AUC values. We believe that the approach taken here, as for most models, will need continual further testing and refining.

A key asset of our approach, however, is the quality of the datapoints used in training and testing, namely, the fact that we only consider experimentally validated data, as opposed to using negative data that is randomly generated. However, as stated in (Montemurro et al., 2021) it is nevertheless important to recognize that there is further room for improvement in the accuracy of the negative data that can be generated.

The good performance of TCR-H suggests that a supervised SVM-based model, applied to a feature set consisting of physicochemical features derived from whole sequences of TCR CDR3βs and epitopes, may be able to capture complex feature dependencies and interdependencies underlying accurate binding classification, It is interesting that ensemble tree-based methods did not perform as well as SVM, suggesting that to achieve accurate binding prediction, a complex function of the full set of features is advantageous, as opposed to an ensemble of simple, nonparametric "functions", as represented by the simple individual trees.

For general applicability a TCR:epitope recognition model must be able to predict binding for TCRs and epitopes for any given TCR:epitope pair. This should include cases in which the TCRs nor the epitopes queried are neither in the training set nor close to members in the training set i.e., the strict split scenario tested here. Success in this endeavor may indicate that an algorithm is being approached that has learned the general principles of TCR:epitope recognition. TCR-H is a clear step in this direction, and towards making the vast universe of potential epitope and TCR sequences amenable to computational prediction of functional TCR/epitope pairs.

Methodology

Dataset:

The dataset used comprises positive binding data curated for human MHC class 1 from IEDB, VDJdb, and McPAS-TCR databases, and negative binding data from IEDB and 10X Genomics assays, sourced from TChard (Grazioli et al., 2022). The dataset includes CDR3β sequence lengths ranging from 9 to 23 residues, while the epitope sequences have lengths shorter than 16. The total dataset consists of 147,069 negative and 107,376 positive datapoints and was divided into training and test data sets of 80 and 20 percent, respectively.

Two different data splits were 1) Hard Split 2) Random Split. In the random split, the entire data was split randomly into training and test sets in the ratio of 80:20. Because the dataset contains instances where a TCR or epitope can be represented multiple times, a random split of the data can lead to the same epitope or TCR being in both the training and test sets. In the hard epitope split the test set consists of only unseen epitopes. The epitope hard split training data set (total: 200,011) consists of 124,598 negative datapoints and 75,413 positive data whereas the test dataset (total: 54,434) consists of 22,471 negative data and 31,963 positive data with 65 unseen epitopes. For the TCR hard split, again the training and test data were split approximately in the ratio of 80:20. In the test dataset with unseen TCR CDR3βs, out of 51,300 data points there were 21,754 positive data and 29,546 negative data. In the Strict split, the training data consisted of 65826 positive data and 83697 negative data whereas test dataset consisted of 22,471 negative data and 31,963 positive data. The training and test data sets are provided in the supporting information (Table in SI).

Feature Set:

All sequence-based properties of the CDR3β loops and the epitopes were calculated using the Peptides Python package [https://github.com/althonos/peptides.py] (Osorio et al., 2015), resulting in 96 features each for the CDR3β and epitope sequences. The feature set includes quantitative structure-activity (QSAR) descriptors as well as sequence profiles representing hydrophobicity and physicochemical properties. A complete list of the features used is given in Table 4. In the feature set used to train the ML models, some features differ substantially in magnitude from one another. In training ML models whose optimization is gradient-based, the presence of features differing by an order of magnitude or more can pose a numerical challenge and yield erroneous results. To avoid this problem, features were normalized, or scaled, so that they were all within the same order of magnitude. To achieve this, the *preprocessing.scale* function in Scikit Learn was used, which centers features to the mean and normalizes them to unit variance. The feature set is listed in Table 4.

Table 4: Sequence-based properties used as features (96 each for epitope and CDR3b) in the present study, the numbers in the parentheses represents the number of features corresponding to given property.

BLOSUM indices	10
Cruciani properties	8
FASGAI vectors	6
Kidera factors	10
MS-WHIM scores	3
PCP descriptors	5
physical_descriptors	2
ProtFP descriptors	8
Sneath vectors	4
svger_descriptors	10
ST-scales	8
T-scales	5
VHSE-scales	8
Z-scales	5
Boman	1
Charge	1
Hydrophobic Moment	1
Hydrophobicity	1
Isoelectric Point	1
Molecular Weight	1
m/z(mass/charge)	1

Machine Learning Methods:

We examined four ML methods: Random Forest (RF), Gradient Boosting trees (GBT), eXtreme Gradient Boosting and Support Vector Machines (SVM). All models were trained using the Scikit Learn application programming interface.

Random Forest (RF) is an ensemble learning method, *i.e.,* in which the final model is a composite of multiple individual models. RF operates by constructing multiple simple decision trees during training. Each tree in the forest is constructed using a random subset of the features. The final prediction is based on the aggregated votes of individual trees. Here, the RF model was built using default hyperparameters and scaled features. Gradient-Boosting Trees (GBT) is another ensemble method we tested. In contrast to RF, whose trees are constructed independently of one another during model training, GBT successively augments the model in each iteration with a tree such that it minimizes a loss function representing the discrepancy between the training targets and the corresponding model predictions. eXtreme Gradient Boosting (XGB) is similar to GBT in that it is based on minimizing a loss function with each tree that is successively added to the ensemble

of trees, but instead of optimizing using just the gradient, it makes use of the Hessian as well, in effect optimizing using a Newton-Raphson type of approach.

In classification, SVM aims to find the hyperplane that best separates the data into different classes, maximizing the margin between the classes. SVM can be trained on the original set of features (corresponding to a linear SVM model) or a projection of the original set of features into a non-linear space using different kernel functions. We used the Radial Basis Function (RBF) kernel, also known as the Gaussian kernel, with the default hyperparameters. The prediction model was built with scaled features.

Acknowledgements

This research used resources of the Compute and Data Environment for Science (CADES) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

References

Aranha, M. P., Jewel, Y. S. M., Beckman, R. A., Weiner, L. M., Mitchell, J. C., Parks, J. M., & Smith, J. C. (2020). Combining Three-Dimensional Modeling with Artificial Intelligence to Increase Specificity and Precision in Peptide–MHC Binding Predictions. *The Journal of Immunology*, *205*(7), 1962–1977. https://doi.org/10.4049/jimmunol.1900918

Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), 6395–6400. https://doi.org/10.1073/pnas.0408677102

Bradley, P., & Thomas, P. G. (2019). *Using T Cell Receptor Repertoires to Understand the Principles of Adaptive Immune Recognition*. https://doi.org/10.1146/annurev-immunol-042718

Bravi, B., Di Gioacchino, A., Fernandez-de-Cossio-Diaz, J., Walczak, A. M., Mora, T., Cocco, S., & Monasson, R. (2023). A transfer-learning approach to predict antigen immunogenicity and T-cell receptor specificity. *ELife*, *12*, 1–35. https://doi.org/10.7554/elife.85126

Cai, M., Bang, S., Zhang, P., & Lee, H. (2022). ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model. *Frontiers in Immunology*, *13*(July), 1–12. https://doi.org/10.3389/fimmu.2022.893247

Chronister, W. D., Crinklaw, A., Mahajan, S., Vita, R., Koşaloğlu-Yalçın, Z., Yan, Z., Greenbaum, J. A., Jessen, L. E., Nielsen, M., Christley, S., Cowell, L. G., Sette, A., & Peters, B. (2021). TCRMatch: Predicting T-Cell Receptor Specificity Based on Sequence Similarity to Previously Characterized Receptors. *Frontiers in Immunology*, *12*. https://doi.org/10.3389/fimmu.2021.640725

Dai, S., Huseby, E. S., Rubtsova, K., Scott-Browne, J., Crawford, F., Macdonald, W. A., Marrack, P., & Kappler, J. W. (2008). Crossreactive T Cells Spotlight the Germline Rules for αβ T Cell-Receptor Interactions with MHC Molecules. *Immunity*, *28*(3), 324–334. https://doi.org/10.1016/j.immuni.2008.01.008

Darmawan, J. T., Leu, J. S., Avian, C., & Ratnasari, N. R. P. (2023). MITNet: a fusion transformer and convolutional neural network architecture approach for T-cell epitope prediction. *Briefings in Bioinformatics*, *24*(4), 1–17. https://doi.org/10.1093/bib/bbad202

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., Crawford, J. C., Clemens, E. B., Nguyen, T. H. O., Kedzierska, K., La Gruta, N. L., Bradley, P., & Thomas, P. G. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, *547*(7661), 89–93. https://doi.org/10.1038/nature22383

Dens, C., Laukens, K., Bittremieux, W., & Meysman, P. (2023). The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nature Machine Intelligence*, *5*(10), 1060–1062. https://doi.org/10.1038/s42256-023-00727-0

Gao, Y., Gao, Y., Fan, Y., Zhu, C., Wei, Z., Zhou, C., Chuai, G., Chen, Q., Zhang, H., & Liu, Q. (2023a). Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, *5*(3), 236–249. https://doi.org/10.1038/s42256-023-00619-3

Gao, Y., Gao, Y., Fan, Y., Zhu, C., Wei, Z., Zhou, C., Chuai, G., Chen, Q., Zhang, H., & Liu, Q. (2023b). Pan-Peptide Meta Learning for T-cell receptor–antigen binding recognition. *Nature Machine Intelligence*, *5*(3), 236–249. https://doi.org/10.1038/s42256-023-00619-3

Gao, Y., Gao, Y., Li, W., Wu, S., Xing, F., Zhou, C., & Fu, S. (2023). *Neo-epitope identification by weakly-supervised peptide-TCR binding prediction*. *1*.

Garcia, K. C., Adams, J. J., Feng, D., & Ely, L. K. (2009). The molecular basis of TCR germline bias for MHC is surprisingly simple. *Nature Immunology*, *10*(2), 143–147. https://doi.org/10.1038/ni.f.219

Grazioli, F., Mösch, A., Machart, P., Li, K., Alqassem, I., O'Donnell, T. J., & Min, M. R. (2022). On TCR binding predictors failing to generalize to unseen peptides. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.1014256

Hudson, D., Fernandes, R. A., Basham, M., Ogg, G., & Koohy, H. (2023). Can we predict T cell specificity with digital biology and machine learning? *Nature Reviews Immunology*. https://doi.org/10.1038/s41577-023-00835-3

Jiang, M., Yu, Z., & Lan, X. (n.d.). *VitTCR: A deep learning method for peptide recognition*. https://doi.org/10.1101/2023.06.02.543411

Jiang, M., Yu, Z., & Lan, X. (2023). *VitTCR: A deep learning method for peptide recognition*. https://doi.org/10.1101/2023.06.02.543411

Jokinen, E., Dumitrescu, A., Huuhtanen, J., Gligorijević, V., Mustjoki, S., Bonneau, R., Heinonen, M., & Lähdesmäki, H. (2023). TCRconv: predicting recognition between T cell receptors and epitopes using contextualized motifs. *Bioinformatics (Oxford, England)*, *39*(1). https://doi.org/10.1093/bioinformatics/btac788

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., & Lähdesmäki, H. (2021). Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Computational Biology*, *17*(3), 1–27. https://doi.org/10.1371/JOURNAL.PCBI.1008814

Korpela, D., Jokinen, E., & Dumitrescu, A. (2023). EPIC-TRACE: predicting TCR binding to unseen epitopes using attention and contextualized embeddings. *BioRxiv Immunology*, 1–18. https://www.biorxiv.org/content/10.1101/2023.06.26.546489v1?rss=1&utm_source=researcher_app&utm_medium=referral&utm_campaign=RESR_MRKT_Researcher_inbound

Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., Bernatchez, C., Heymach, J. V., Gibbons, D. L., Wang, J., Xu, L., Reuben, A., & Wang, T. (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence*, *3*(10), 864–875. https://doi.org/10.1038/s42256-021-00383-2

Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., Crinklaw, A., Hadrup, S. R., Winther, O., Peters, B., Jessen, L. E., & Nielsen, M. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology*, *4*(1), 1–13. https://doi.org/10.1038/s42003-021-02610-3

Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., Ogunjimi, B., Laukens, K., & Meysman, P. (2021). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, *22*(4), 1–12. https://doi.org/10.1093/bib/bbaa318

Morris, G. P., & Allen, P. M. (2012). How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature Immunology*, *13*(2), 121–128. https://doi.org/10.1038/ni.2190

Myronov, A., Mazzocco, G., Król, P., & Plewczynski, D. (2023). BERTrand-peptide:TCR binding prediction using Bidirectional Encoder Representations from Transformers augmented with random TCR pairing. *Bioinformatics*, *39*(8), 1–8. https://doi.org/10.1093/bioinformatics/btad468

Osorio, D., Rondón-Villarreal, P., & Torres, R. (2015). Peptides: A package for data mining of antimicrobial peptides. *R Journal*, *7*(1), 4–14. https://doi.org/10.32614/rj-2015-001

Pham, M. D. N., Nguyen, T. N., Tran, L. S., Nguyen, Q. T. B., Nguyen, T. P. H., Pham, T. M. Q., Nguyen, H. N., Giang, H., Phan, M. D., & Nguyen, V. (2023). epiTCR: a highly sensitive predictor for TCR–peptide binding. *Bioinformatics*, *39*(5). https://doi.org/10.1093/bioinformatics/btad284

Rajitha Rajeshwar, T., & Smith, J. C. (2022). Structural patterns in class 1 major histocompatibility complex-restricted nonamer peptide binding to T-cell receptors. *Proteins: Structure, Function and Bioinformatics*, *90*(9), 1645–1654. https://doi.org/10.1002/prot.26343

Rudolph, M. G., Stanfield, R. L., & Wilson, I. A. (2006). HOW TCRS BIND MHCS, PEPTIDES, AND CORECEPTORS. *Annual Review of Immunology*, *24*(1), 419–466. https://doi.org/10.1146/annurev.immunol.23.021704.115658

Szeto, C., Lobos, C. A., Nguyen, A. T., & Gras, S. (2021). *Szeto et al. - 2021 - TCR Recognition of Peptide–MHC-I Rule Makers and Breakers - ijms-22-00068-v2.pdf*.

Tippalagama, R., Chihab, L. Y., Kearns, K., Lewis, S., Panda, S., Willemsen, L., Burel, J. G., & Lindestam Arlehamn, C. S. (2023). Antigen-specificity measurements are the key to understanding T cell responses. In *Frontiers in Immunology* (Vol. 14). Frontiers Media S.A. https://doi.org/10.3389/fimmu.2023.1127470

Wang, D., He, F., Yu, Y., & Xu, D. (2023). Meta-learning for T cell receptor binding specificity and beyond. *Nature Machine Intelligence*, *5*(4), 337–339. https://doi.org/10.1038/s42256-023-00641-5

Wang, Z., & Shen, Y. (2023). *SABRE : Self-Attention Based model for predicting T-cell Receptor Epitope Specificity*.

Weber, A., Born, J., & Rodriguez Martínez, M. (2021). TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics*, *37*, I237–I244. https://doi.org/10.1093/bioinformatics/btab294

Zhang, Y., Jian, X., Xu, L., Zhao, J., Lu, M., Lin, Y., & Xie, L. (2023). iTCep: a deep learning framework for identification of T cell epitopes by harnessing fusion features. *Frontiers in Genetics*, *14*(May), 1–11. https://doi.org/10.3389/fgene.2023.1141535

Supporting Information.

In Table S1 we present the categorized numbers of data in the various datasets used.

Table S1: Categorized dataset numbers.

Epitope Hard Split:

Training dataset			**Test dataset**		
Positive data	**Negative data**	**Unique peptides**	**Positive data**	**Negative data**	**Unique peptides**
75413	124598	790	31963	22471	65

TCR Hard Split:

Training dataset			**Test dataset**		
Positive data	**Negative data**	**Unique CDR3βs**	**Positive data**	**Negative data**	**Unique CDR3βs**
85621	117523	124098	21754	29546	31025

Strict Split:

Training dataset				**Test dataset**			
Positive data	**Negative data**	**Unique CDR3βs**	**Unique peptides**	**Positive data**	**Negative data**	**Unique CDR3βs**	**Unique peptides**
65826	83697	105011	739	31963	22471	50112	65

The performance model metrics are defined as follows:

Accuracy (ACC) is the proportion of correctly classified instances:

$$\text{Accuracy}=\frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

Precision (also called Positive Predictive Value) is the proportion of true positives among instances predicted as positive:

$$\text{Precision}=\frac{True\ Positives}{True\ Positives+False\ Positives} \quad (2)$$

Recall (also called Sensitivity or True Positive Rate) is the proportion of true positives among actual positive instances:

$$\text{Recall}=\frac{True\ Positives}{True\ Positives+False\ Negatives} \quad (3)$$

F1 Score is the harmonic mean of precision and recall:

$$\text{F1 Score}=2\times\frac{Precision\times Recall}{Precision+Recall} \quad (4)$$

Specificity (True Negative Rate) is the proportion of true negatives among actual negative instances:

$$\text{Specificity}=\frac{False\ Positives}{True\ Negatives+False\ Positives} \quad (5)$$

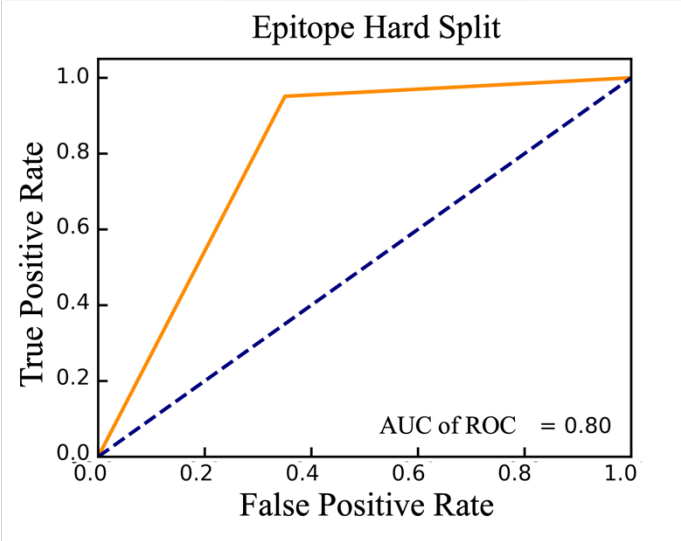

Figure S1. AUC of ROC on the independent test set for the SVM model trained with the hard split for the training and independent test set. This SVM model was trained on the set of all features, without removing correlated features.

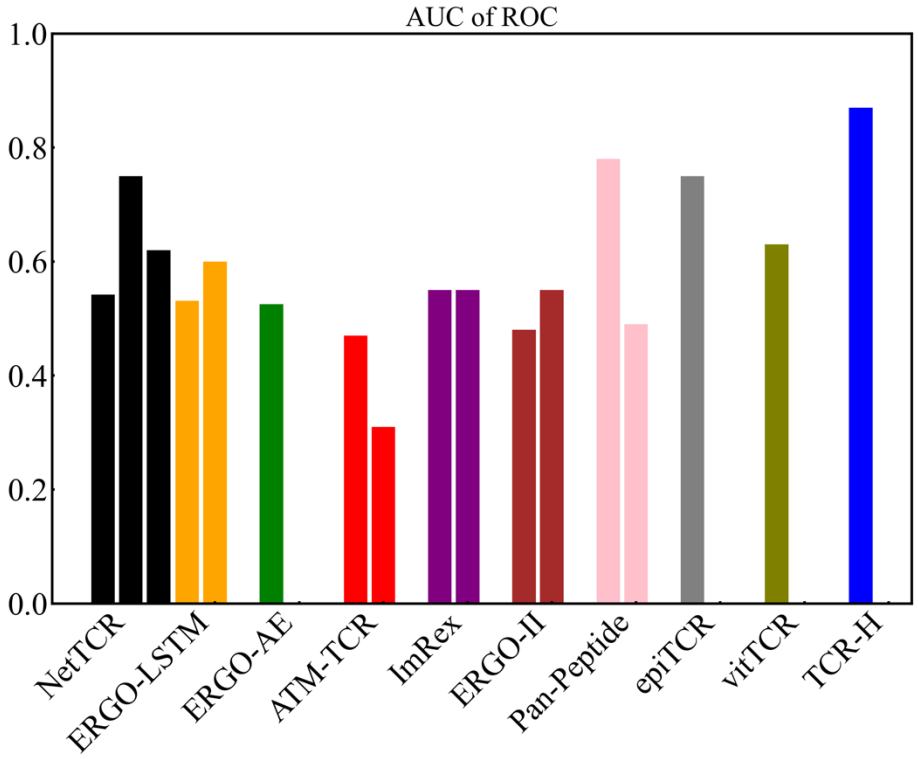

Figure S3. Epitope Hard Split AUC of ROC data with TCR-HE and various previously-reported models. Each bar represents the given model with the same training set but tested on a specific hard split data set.

Data taken from (Cai et al., 2022; Gao, Gao, Fan, et al., 2023b; Grazioli et al., 2022; Jiang et al., n.d.; Pham et al., 2023)