# TCRex: a webtool for the prediction of T-cell receptor sequence epitope specificity

Sofie Gielis[1,2,3], Pieter Moris[1,3], Nicolas De Neuter[1,2,3], Wout Bittremieux[1,3,7], Benson Ogunjimi[2,4,5,6], Kris Laukens[1,2,3*], Pieter Meysman[1,2,3*]

[1] Adrem Data Lab, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
[2] AUDACIS, Antwerp Unit for Data Analysis and Computation in Immunology and Sequencing, University of Antwerp, Antwerp, Belgium
[3] Biomedical Informatics Research Network Antwerp (biomina), University of Antwerp, Antwerp, Belgium
[4] Antwerp Center for Translational Immunology and Virology (ACTIV), Vaccine and Infectious Disease Institute, University of Antwerp, Antwerp, Belgium
[5] Department of Paediatrics, Antwerp University Hospital, Antwerp, Belgium
[6] Center for Health Economics Research and Modeling Infectious Diseases (CHERMID), Vaccine and Infectious Disease Institute, University of Antwerp, Belgium
[7] Department of Genome Sciences, University of Washington, Seattle WA 98195, USA

* Corresponding authors: pieter.meysman@uantwerpen.be or kris.laukens@uantwerpen.be

Pieter Meysman and Kris Laukens contributed equally to this article.

## Abstract

Identification of T-cell receptor (TCR) repertoire epitope targets constitutes an important part of many TCR repertoire studies. To date, we are still relying on time consuming epitope binding experiments for the identification of epitope-specific TCR sequences. Recently, we showed that the prediction of epitope-TCR interaction is possible using a random forest model. We implemented this method in a webtool called TCRex. TCRex is the first tool that enables the prediction of TCR-epitope recognition. It allows users to upload TCR sequences and predict interaction with multiple known cancer or viral epitopes or train new prediction models for new epitopes. TCRex is freely available for academic use at tcrex.biodatamining.be

## 1 Introduction

T-cells form an important part of the adaptive immune system as they can recognize potentially pathogenic or aberrant peptides (epitopes), presented either on the cell surface of nucleated host cells or by professional antigen-presenting cells, and induce an immune response. The T-cell receptor (TCR) molecule is responsible for the recognition of the epitope. Each TCR protein is encoded by a genomic region that undergoes non-homologous recombination during T-cell maturation. The randomness of the recombination process induces creation of many different TCR proteins that are for the most part unique per T-cell clone and allows recognition of different epitopes. Epitope binding by the TCR is a critical step for the activation of targeted immune responses. Although multiple immunoinformatics tools exist to predict epitopes and their binding with MHC molecules (e.g. NetCTLpan (Stranzl et al., 2010), CRFMHC (Meysman et al., 2015) and Immune epitope database analysis resource (IEDB-AR) (Zhang et al., 2008)), we still lack useful tools for the prediction of epitope recognition by TCRs.

Here, we present the first tool to analyze TCR sequences and predict the likelihood that they target specific epitopes. This tool is based on the principle that similar TCR sequences often target the same epitope (Meysman et al., 2018) and that machine learning techniques can be used to learn those commonalities shared by epitope-specific TCR sequences (De Neuter et al., 2018; Dash et al., 2017; Glanville et al., 2017). It builds further on our prior work related to the feasibility of predicting TCR-epitope recognition using TCR beta sequences (De Neuter et al., 2018). In this study, we showed that a random forest classifier trained to predict TCR-epitope interactions from TCR amino acid physicochemical properties can achieve a high accuracy. We extended this work into a complete

framework trained on a large dataset containing different viral and cancer epitopes and made it freely available as a webtool called TCRex.

## 2 Methods

### *2.1 Data collection*

Epitope-specific TCR sequences were downloaded from the manually curated catalogue of pathology-associated T-cell receptor sequences (McPAS-TCR) (Tickotsky et al., 2017) and the VDJ database (VDJdb) (Shugay et al., 2018) on the 9th of April 2018. Data collection was currently restricted to TCR beta sequences from human CD8+ T-cells, as this constitutes the bulk of available data. Non-canonical TCR sequences (i.e. not starting with cysteine or not ending with phenylalanine) were removed. Control CD8+ TCR sequences were collected from (Seay et al., 2016) and (Thome et al., 2016). For both the epitope-specific and control TCR sequences, the CDR3 beta amino acid sequence and the corresponding V/J genes and families were retrieved.

### *2.2 Model training and performance evaluation*

Training of epitope-specific prediction models occurred in the same manner as presented in (De Neuter et al., 2018). In brief, the amino acid sequences of the CDR3 regions of the TCR beta chains were converted into physiochemical features and the V/J genes and families were one-hot encoded. For each epitope, for which we retrieved at least 30 epitope-specific sequences, a random forest classifier was trained with 100 individual decision trees. Given a TCR beta sequence, these models give the probability of binding a specific epitope. All models were evaluated using a repeated subsampling strategy to obtain the receiver operating characteristic curve (ROC), precision-recall (PR) curve, the accuracy, the mean area-under-the-receiver-operating-characteristic-curve (AUC) value and the mean PR value. Models that had poor AUC (< 0.7) or PR (< 0.35) values were excluded from the final webtool. In total, 43 prediction models with a sufficiently high performance were retained for the initial release. For each of these models, classification thresholds were determined for 1%, 0.1% and 0.01% false positive rates (FPR). These threshold values were calculated by predicting the class probability scores for 100 000 randomly selected control TCR beta sequences. The 1%, 0.1% and 0.01% FPR thresholds were set respectively to the 1000th, the 100th and the 10th highest class probability score.

## 3 Webtool

TCRex provides a user-friendly web interface to predict epitope binding for human TCR beta sequences. An overview of the general workflow is presented in figure 1. To start a prediction analysis, users can upload a TCR data file. Different file formats are supported by TCRex, including the immunoSEQ (https://www.adaptivebiotech.com/immunoseq) and MiXCR (Bolotin et al., 2015) output formats. In addition, we propose a very simple tab delimited format that includes the CDR3 amino acid sequences and the V/J genes for all TCR beta sequences following the international ImMunoGeneTics information system (IMGT) notation (Lefranc et al., 2015). In case no V/J gene information is available, users are advised to provide the corresponding V/J families.

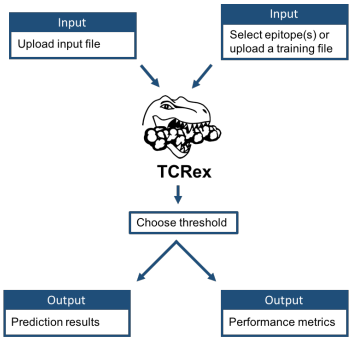


Fig. 1. Overview of the TCRex workflow. TCR-epitope interaction predictions start after uploading a TCR data file and selecting the epitopes of interest. If the latter are not available in the database, users can train new prediction models by uploading epitope-specific TCR sequences. After choosing the classification threshold, the prediction results can be downloaded as a CSV file together with the performance metrics of the prediction model in case a new classifier was trained.

After uploading the input file, one or more epitopes can be selected from the database. In this first release, prediction models for 43 different epitopes are available, including 38 viral and 5 cancer

epitopes. With the development and improvement of new TCR sequencing techniques and the rising interest in epitope-specific TCR repertoire analysis, we expect this number to grow rapidly in the near future. The database will be updated on a half-year schedule with new epitope data made available in the scientific literature. In addition, it is also possible to make predictions for epitopes that are not available in the database. To this end, the user can upload their own dataset containing epitope-specific TCR sequences, which is then used to train new prediction models for their own use. These custom models are hidden to other users and are removed when the task is finished.

When all the required information is submitted, the webtool redirects the user to a webpage that gives an overview of all the steps in the prediction process and the current status of the analysis. Once the analysis is finished, a webpage with the prediction results is returned. This interactive results summary allows the user to select the desired classification threshold. By default, a threshold with 0.01% FPR is used, but this can be easily changed to any user-defined class probability threshold.

After the analysis, the prediction results can be downloaded as a CSV file. For every TCR sequence provided by the user and every selected epitope, this CSV file contains a score that represents the probability of epitope-TCR recognition. All scores with a value greater than or equal to the chosen classification threshold are considered to bind the epitope of interest. These epitope-TCR pairs are indicated in the output file with an asterisk. In the case where the user trained a new prediction model, the results are presented with a summary of the performance metrics along with the ROC and PR curves and a visualization of the important features. All results are kept available for 48 hours.

## 4 Conclusion

TCRex is the first toolbox enabling the prediction of epitope-specific T-cell receptor sequences. Predictions can be made for various cancer and viral epitopes present in the TCRex database or for new epitopes by training new prediction models. TCRex is freely available for academic use as a user-friendly webtool and can be used for the identification of T-cell receptor repertoire targets.

## Acknowledgements

We gratefully acknowledge Sascha Degenhardt for his programming contributions and the developers of the VDJdb and McPAS-TCR for the collection of TCR data in easily accessible databases.

## Funding

This research was supported by the University of Antwerp with an University Research Fund (BOF Concerted Research Action) and an Industrial Research Fund (IOF), by the Research Foundation Flanders (FWO) [Personal PhD grants to NDN (1S29816N) and PMo (1141217N), postdoctoral grant to WB (12W0418N) and research project grant (G067118N)] and by the Belgian American Educational Foundation (BAEF) [postdoctoral grant to WB].

## References

Bolotin, D. A., Poslavsky, S., Mitrophanov, I., Shugay, M., Mamedov, I. Z., Putintseva, E. V., and Chudakov, D. M. (2015). MiXCR: Software for comprehensive adaptive immunity profiling. *Nature Methods*, 12(5), 380–381.

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.

De Neuter, N., Bittremieux, W., Beirnaert, C., Cuypers, B., Mrzic, A., Moris, P., Suls, A., Van Tendeloo, V., Ogunjimi, B., Laukens, K., and Meysman, P. (2018). On the feasibility of mining CD8+ T cell receptor patterns underlying immunogenic peptide recognition. *Immunogenetics*, 70(3), 159–168.

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., Ji, X., Han, A., Krams, S. M., Pettus, C., Haas, N., Lindestam Arlehamn, C. S., Sette, A., Boyd, S. D., Scriba, T. J., Martinez, O. M., and Davis, M. M. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature*, 547(7661), 94–98.

Lefranc, M. P., Giudicelli, V., Duroux, P., Jabado-Michaloud, J., Folch, G., Aouinti, S., Carillon, E., Duvergey, H., Houles, A., Paysan-Lafosse, T., Hadi-Saljoqi, S., Sasorith, S., Lefranc, G., and Kossida, S. (2015). IMGT®, the international ImMunoGeneTics information system® 25 years on. *Nucleic Acids Research*, 43(Database issue), D413–D422.

Meysman, P., Ogunjimi, B., Naulaerts, S., Beutels, P., Van Tendeloo, V., and Laukens, K. (2015). Varicella-Zoster Virus-Derived Major Histocompatibility Complex Class I-Restricted Peptide Affinity Is a Determining Factor in the HLA Risk Profile for the Development of Postherpetic Neuralgia. *Journal of Virology*, 89(2), 962–969.

Meysman, P., De Neuter, N., Gielis, S., Bui Thi, D., Ogunjimi, B., and Laukens, K. (2018). The workings and failings of clustering T-cell receptor beta-chain sequences without a known epitope preference. *bioRxiv*.

Seay, H. R., Yusko, E., Rothweiler, S. J., Zhang, L., Posgai, A. L., Campbell-Thompson, M., Vignali, M., Emerson, R. O., Kaddis, J. S., Ko, D., Nakayama, M., Smith, M. J., Cambier, J. C., Pugliese, A., Atkinson, M. A., Robins, H. S., and Brusko, T. M. (2016). Tissue distribution and clonal diversity of the T and B cell repertoire in type 1 diabetes. *JCI Insight*, 1(20).

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., Douek, D. C., Luciani, F., van Baarle, D., Kedzierska, K., Kesmir, C., Thomas, P. G., Price, D. A., Sewell, A. K., and Chudakov, D. M. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, 46(Database issue), D419–D427.

Stranzl, T., Larsen, M. V., Lundegaard, C., and Nielsen, M. (2010). NetCTLpan: Pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, 62(6), 357–368.

Thome, J. J. C., Grinshpun, B., Kumar, B. V., Kubota, M., Ohmura, Y., Lerner, H., Sempowski, G. D., Shen, Y., and Donna, L. (2016). Longterm maintenance of human naive T cells through in situ homeostasis in lymphoid tissue sites. *Sci Immunol*, 1(6).

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., and Friedman, N. (2017). McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics*, 33(18), 2924–2929.

Zhang, Q., Wang, P., Kim, Y., Haste-Andersen, P., Beaver, J., Bourne, P. E., Bui, H. H., Buus, S., Frankild, S., Greenbaum, J., Lund, O., Lundegaard, C., Nielsen, M., Ponomarenko, J., Sette, A., Zhu, Z., and Peters, B. (2008). Immune epitope database analysis resource (IEDB-AR). *Nucleic acids research*, 36(Web Server issue), W513–W518.