Predicting immunogenicity by modeling the positive and negative selection of CD8+ T cells in individual patients

Ngoc Hieu Tran[1,2], Ming Li[1,*]

[1]Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, Canada.

[2]Bioinformatics Solutions Inc., Waterloo, Ontario, Canada.

*Corresponding author. Email: mli@uwaterloo.ca

Abstract.

Here we propose a personalized machine learning approach to predict the collective response of a patient's CD8+ T cells by modeling the positive and negative selection, i.e. the central tolerance of T cells. In particular, for each individual patient, we collected his/her HLA-I self peptides derived from mass spectrometry-based immunopeptidomics as negative selection, and allele-matched immunogenic T cell epitopes from the Immune Epitope Database as positive selection. The negative and positive peptides of each patient were used to train a personalized classification model, named DeepImmun, which was then applied to predict the immunogenicity of candidate neoantigens of that patient. We evaluated DeepImmun and three other immunogenicity prediction tools on the neoantigens of 18 cancer patients which had been validated by T cell assays in previous studies. Our personalized model outperformed the other tools and was able to rank 25% of the neoantigens within its top 10% predictions, and 75% of the neoantigens within its top 30%. The results also demonstrated that DeepImmun could be used to prioritize a shortlist of as few as 40 candidate neoantigens per patient, reducing the time and costs of complicated *in vitro* validations while still ensuring to capture the true neoantigens of interest.

Introduction

Tumor-specific human leukocyte antigen (HLA) peptides are only expressed on the surface of cancer cells and hence represent ideal targets for the immune system to distinguish cancer cells from normal cells. However, due to the high specificity of T cells, only a small fraction of those peptides can be recognized by T cells to trigger immune responses. They are referred to as neoantigens. Despite their promising role in cancer immunotherapy, the chance of finding neoantigens is remarkably low, usually less than half a dozen out of thousands of somatic mutations detected per patient[1–4]. Thus, *in silico* methods are essential to accurately predict and prioritize candidate neoantigens before *in vitro* validations, which often involve time-consuming and costly experiments. A typical approach in existing studies is to predict a limited number of candidate neoantigens, e.g. <100 per patient, and then synthesize and test them *in vitro* for immunogenicity, with the hope that 2-3 candidates show positive responses[5,6].

Majority of current *in silico* methods focus on predicting the binding affinity of class 1 and class 2 peptide-HLA (pHLA) complexes or, given the rise of recent mass spectrometry (MS) based immunopeptidomics, the entire pHLA presentation pathways[6–11]. The peptides are then ranked by their HLA binding/presentation scores, their expression levels, and other criteria, based on the assumption that if a mutated peptide is presented in large amounts on the cancer cell surface, it has a better chance to trigger T cell responses. However, such approaches only consider half of the equation and do not take into account how a pHLA complex can be recognized by T cell receptors (TCRs). Early studies on HLA class 1 (HLA-I) T cell epitopes have suggested that certain amino acid positions, e.g. 4-6, of the epitopes were in close contact with the TCRs, while the anchor positions, e.g. 1, 2, 9, were responsible for HLA-I binding[12,13]. Recently Schmidt *et al.*[14] and Gfeller *et al.*[15] proposed PRIME, a model that simultaneously predicts the HLA-I binding and the TCR recognition of a peptide by specifically assigning which of its amino acid positions are responsible for each task. In addition, some properties of amino acid residues, such as hydrophobicity, polarity, or large and aromatic side chains, were found to have statistically significant correlation with the epitope immunogenicity. Sequence similarity to known pathogen epitopes and dissimilarity to the human proteome have also been proposed for immunogenicity prediction[16,17].

However, the above methods fall short of accounting for the individuality of each patient and his/her T cells. Note that both cancer and immunotherapy are associated with a great amount of *personal* genetic variations of each individual patient. The HLA complex, the T cell population, and the tumor mutations themselves are significantly different from one patient to another. Thus,

cancer immunotherapy should strongly benefit from personalized approaches. This observation motivated us, from a machine-learning point of view, to develop personalized models for T cell immunogenicity prediction. One potential approach is to sequence the TCRs of a patient and then predict the TCR-pHLA recognition from their sequences or structures[4,18–22]. However, despite on-going intensive efforts of TCR sequencing, it is difficult to sequence the complete TCR repertoire for every patient.

To address this problem, we note that the specificity of T cells are mainly shaped by their positive and negative selection, i.e. the central tolerance that happens inside the thymus of each individual patient (Figure 1a). During the positive selection, T cells are selected by their ability to bind to pHLA complexes: they will become CD8+ or CD4+ T cells if they bind to HLA class 1 (HLA-I) or HLA class 2 (HLA-II) complexes, respectively; otherwise they will die by neglect[23]. During the negative selection, T cells are selected *against* their ability to bind to self peptides: those that have high affinity for self peptides will die by clonal deletion, preventing the risk of autoimmunity; the remaining will mature and participate in immune responses against foreign antigens. Extensive efforts, both experimental and computational, have been put into modeling complicated molecular mechanisms behind those processes[23].

Here we take an alternative approach and hypothesize that the positive and negative selection can be encapsulated by the complete space of positive and negative peptides resulting from that selection. We propose to use MS-based immunopeptidomics to obtain HLA-I self peptides to resemble the negative selection of CD8+ T cells in an individual patient (Figure 1b). For the positive selection, we collect all epitopes reported in positive T cell assays from the Immune Epitope Database (IEDB)[24] that match the patient's HLA-I alleles. Both self peptides and T cell epitopes bind to HLA-I molecules of the patient, however, the former are non-immunogenic and the latter are immunogenic to the patient's T cells. Using this personal dataset of negative and positive peptides, we then train a binary classification model specifically for that patient to predict his/her T cell response to any given peptide. Note that in this study we only focus on CD8+ T cells and HLA-I peptides but not CD4+ T cells and HLA-II peptides due to the current availability of those data.

It is worth to emphasize that, while previous studies have used immunogenic and non-immunogenic peptides to train immunogenicity prediction models[12,14,15], the peptides were often collected from existing literatures or databases in an HLA-allele-specific manner. Our work is the first to use self peptides derived from MS-based immunopeptidomics of an individual patient as non-immunogenic for that particular patient. In the following sections, we shall describe more

details of our model, the training and validation on neoantigens of cancer patients, as well as the performance comparison to other tools.

Results

Personalized models for T cell immunogenicity prediction

To demonstrate our idea, we collected data and trained personalized immunogenicity prediction models for 18 individual patients (Table 1). The datasets were curated from ten previously published studies where tumor-specific neoantigens had been identified and validated in T cell assays[5,6,25–32]; patients with at least two neoantigens were selected for our study. Overall, the number of neoantigens was 2-6 per patient, while the total number of peptides that had been tested in T cell assays was 15-434 per patient. The neoantigens arised from different sources including missense mutations, non-coding regions, bacteria-derived, etc., and multiple cancer types such as melanoma, lung, ovarian, etc. The information of 18 patients is summarized in Table 1, their neoantigens are provided in the Supplementary Data 1.

As described earlier, the key idea of our approach is to use HLA-I self peptides derived from MS-based immunopeptidomics as negative training data for immunogenicity prediction models (Figure 1b). This would require retrieving HLA-I peptides by HLA immunoprecipitation assays followed by MS experiments on a patient's sample. Unfortunately, MS-based immunopeptidomics is relatively new and not always accessible. In particular, among the 18 patients that we considered, only four had MS-based immunopeptidomics data available[5,25–27] (Table 1). For each of those four patients, the MS data was downloaded and subsequently searched against the standard Swiss-Prot human protein database using PEAKS Xpro[11,33] with no-enzyme-specific cleavage and false discovery rate (FDR) of 1% (Methods section). The identified peptides were further checked for their HLA-I characteristic length distribution; those with lengths <8 or >14 amino acids were filtered out. The resulting peptides were considered as HLA-I self peptides of the patient and not recognized by the patient's T cells, i.e. non-immunogenic. The number of HLA-I self peptides varied from 3,630-35,551, indicating a large variation of depth between the immunopeptidomics samples. For the remaining 14 patients with no available MS data, we used the six HLA-I alleles of each patient to search the IEDB[24] for peptides eluted by MHC class 1 (MHC-I) assays in human (Methods section). Those allele-matched MHC-I eluted peptides were then used instead of the unavailable self peptides as negative training data (Table 1).

For positive training data, we downloaded the table of T cell assays from IEDB with the following filters: linear epitopes, HLA class I, and host as human (Methods section). For each individual patient, we selected T cell epitopes that matched his/her HLA-I alleles and had positive Assay Qualitative Measure, including Positive, Positive-High, Positive-Intermedia, Positive-Low. Epitopes with lengths <8 or >14 amino acids were filtered out. The resulting peptides were assumed to be recognized by the patient's T cells, i.e. immunogenic. The number of T cell epitopes varied from 133-4,785 per patient (Table 1). The training data is provided in the Supplementary Data 2.

Using the above positive and negative training data, we can train a binary classification model to predict whether a peptide is immunogenic or non-immunogenic. Since the training data was collected specifically for each individual patient based on his/her HLA-I self peptides and alleles, the trained model can be considered as patient-specific, i.e. personalized. We used a bi-directional long short-term memory (LSTM) network coupled with amino acid embedding to model the peptides as they had been successfully used previously for de novo sequencing[33,34], spectrum and retention time predictions[35–37], and other peptide prediction tasks[4,38]. For each individual patient, his/her training data was splitted into three sets: train-validation-test with ratio 80%-10%-10% and no overlapping peptides. We trained 100 ensemble models per individual patient and performed downsampling on negative peptides, as the amount of negative data was much higher than that of positive data. The ensemble models were sorted according to their performance on the validation set and the average of the top-10 models was selected as the final prediction model for that patient. More details of the model implementation and the training process can be found in the Supplementary Method.

Performance evaluation of immunogenicity prediction tools

The personalized model of each individual patient was evaluated on his/her own neoantigens that had been identified and validated by T cell assays. As mentioned earlier, the number of neoantigens was 2-6 per patient, while the total number of peptides that had been tested in T cell assays was 15-434 per patient (Table 1). However, it should be noted that, before T cell assay validation, those peptides had actually been selected from an even larger pool of candidates[5,6,25–32]. Thus, we expanded the evaluation set of each patient by adding negative peptides randomly drawn from that patient's self peptides (or allele-matched MHC-I eluted peptides from IEDB when self peptides are not available). The ratio between neoantigens to negative peptides in the evaluation set of each patient was set at 1:100, i.e. an immunogenicity prediction tool needs to prioritize 1 neoantigen out of every 100 negative peptides. We found

that the ratio 1:100 was in line with the number of neoantigens and the number of candidates identified from genomics/proteogenomics pipelines in most existing neoantigen studies[5,6,25–32]. Any common peptides found in both the evaluation set and the training data were excluded from the training data. The evaluations sets of 18 patients are provided in the Supplementary Data 3.

We compared our personalized approach, named DeepImmun, to three other leading tools, including PRIME[14,15] (version 2.0), NetMHCpan[7] (version 4.1), and IEDB immunogenicity predictor[12]. NetMHCpan is one of the earliest and most common tools for HLA-I binding prediction. IEDB immunogenicity predictor is one of the earliest HLA-I immunogenicity prediction tools that showed the association between immunogenicity and amino acid positions 4-6, plus other properties of amino acid residues, such as hydrophobicity, polarity, or large and aromatic side chains. PRIME is a recent immunogenicity prediction tool that simultaneously models HLA-I binding and TCR recognition. The four prediction tools were tested on the evaluation sets of 18 individual patients described above. Their prediction scores are provided in the Supplementary Data 3. Two criteria were used for the performance comparison. The first criterion is the area under the receiver operating characteristic curve (ROC-AUC) calculated from the prediction scores and the peptide classification in each evaluation set. The second criterion is the ranks of neoantigens versus negative peptides in each evaluation set based on the prediction scores. More details of those criteria and their calculation can be found in the Supplementary Method.

The performance of four immunogenicity prediction tools are summarized in Figure 2. Figure 2a shows their prediction ROC curves and AUCs on the combined evaluation set of 18 patients. DeepImmun achieved an AUC of 0.69 and outperformed the other three tools, demonstrating the advantage of our personalized approach. Figure 2b further shows the ability of the four prediction tools to rank neoantigens on top of non-immunogenic peptides. As explained earlier, the main purpose of immunogenicity prediction is to prioritize candidate neoantigens to reduce the time and costs of *in vitro* validations. Thus, our evaluation sets contained neoantigens and non-immunogenic peptides at a ratio of 1:100 and the prediction tools need to prioritize 1 neoantigen out of every 100 non-immunogenic peptides (similar to genomics/proteogenomics pipelines in most existing neoantigen studies[5,6,25–32]). Figure 2b shows that DeepImmun was able to rank 25% of the neoantigens within its top 10% predictions, and nearly 75% of the neoantigens within its top 30% (Q1 and Q3 quartiles of the boxplot, respectively). In other words, for an average evaluation set of an individual patient that included 4 neoantigens and 400 non-immunogenic peptides, DeepImmun's top 40-120 predicted candidates would contain

1-3 neoantigens of interest. Figure 2b also shows that DeepImmun again outperformed the other tools in terms of the ranking ability.

As we focused on a personalized approach, we also looked into the prediction performance on each individual patient. The prediction AUCs in Figure 2c shows that DeepImmun outperformed the other tools on 12 out of 18 patients. DeepImmun did not perform well on two particular patients, *CTE0015* from Bobisse et al.[30] and *Ott3* from Ott et al.[32]. One of the best prediction results of DeepImmun was achieved on patient *Mel-15* (AUC=0.85), a melanoma patient dataset first published by Bassani-Sternberg et al.[5]. This dataset is well-known as the native immunopeptidomics sample with the highest coverage to date, where more than 35K self-peptides had been identified by MS-based immunopeptidomics (Table 1). Bassani-Sternberg et al.[5] and Wilhelm et al.[35] identified 28 mutated HLA-I peptides from *Mel-15's* MS and RNA-seq data, and among them, five neoantigens were validated in T cell assays with positive responses. Figure 2d shows the predicted ranks of those five neoantigens versus other non-immunogenic peptides identified from the same patient. DeepImmun produced the best ranks across all five neoantigens, and more importantly, three of them were ranked within its top 3% predictions. Another patient *Mel-51* from Kalaora et al.[26] with more than 23K self-peptides identified by MS-based immunopeptidomics also shows excellent prediction results by DeepImmun (AUC=0.87, Figure 2c). The prediction results of all 18 patients are provided in the Supplementary Data 3.

Discussion

In this study we proposed DeepImmun, a personalized model for immunogenicity prediction by resembling the negative and positive selection of CD8+ T cells in an individual patient. The model used HLA-I self peptides derived from MS-based immunopeptidomics of an individual patient as negative training data and allele-matched positive T cell epitopes from the IEDB as positive training data. We evaluated DeepImmun and three other immunogenicity prediction tools on the neoantigens of 18 cancer patients that had been validated by T cell assays in previous studies. The results showed that DeepImmun achieved an average AUC of 69% and outperformed the other tools on 12 of 18 patients. More importantly, our model was able to rank 25% of the neoantigens within its top 10% predictions, and nearly 75% of the neoantigens within its top 30%. On average, DeepImmun's top 40-120 predicted candidates contained 1-3 neoantigens per patient, thus reducing the time and costs for further *in vitro* validations.

The key idea of our personalized approach is to use HLA-I self peptides derived from MS-based immunopeptidomics of an individual patient to resemble the negative selection of that patient's T cells. However, MS-based immunopeptidomics may not always be available for every patient. In such cases, one could use allele-matched MHC-I eluted peptides from public databases, i.e. IEDB, to replace the unavailable self peptides as negative training data. In this study, we have validated our approach in both scenarios: 4 patients with self peptides derived from MS-based immunopeptidomics and 14 patients with allele-matched MHC-I eluted peptides from IEDB. Thus, our personalized model can be applied to any patient, regardless of whether MS-based immunopeptidomics data is available or not. Nevertheless, we strongly believe that our study has demonstrated the validity and the benefits of personalized approaches for neoantigen prediction and it shall propel the application of MS-based immmunopeptidomics alongside existing genomics pipelines for cancer immunotherapy.

Methods

Preparation of training and evaluation data for each of 18 cancer patients

To obtain the HLA-I self peptides of 4 patients *Mel-15*, *Mel-0D5P*, *Mel-51*, and *iSPC2_IFN*, we analyzed their MS immunopeptidomics data published in the original studies[5,25–27]. The MS data was searched against the standard Swiss-Prot human protein database (version June 15, 2020) using PEAKS Xpro[11,33] (version 10.6) with no-enzyme-specific cleavage and false discovery rate of 1%. Precursor mass and fragment ion mass tolerances were set as 15.0 ppm and 0.05 Da, respectively; M(Oxidation) and NQ(Deamidation) were set as variable modifications. The identified peptides were subsequently checked for their HLA-I characteristic length distribution; those with lengths <8 or >14 amino acids were filtered out. The resulting peptides were considered as HLA-I self peptides and were used as negative training data for those 4 patients.

For the remaining 14 patients with no MS data available, we first downloaded human MHC-I eluted peptides from the IEDB (Supplementary Figure S1, Supplementary Data 2). For each patient, we used his/her alleles to select the corresponding peptides and used them as negative training data.

The table of T cell epitopes was downloaded from IEDB with the following filters: linear epitopes, MHC class I, and host as human (Supplementary Figure S2, Supplementary Data 2). For each patient, epitopes were selected if they matched the patient's HLA-I alleles and had positive Assay Qualitative Measure (including Positive, Positive-High, Positive-Intermedia, Positive-

Low). Epitopes with lengths <8 or >14 amino acids were filtered out. The resulting peptides formed the positive training data.

For evaluation data, we obtained the tumor-specific neoantigens of those 18 patients which had been validated as immunogenic by T cell assays in the original studies[5,6,25–32]. For each patient, we further expanded the evaluation set by adding non-immunogenic (negative) peptides randomly drawn from that patient's self peptides (or allele-matched MHC-I eluted peptides from IEDB when self peptides are not available). The ratio between neoantigens to negative peptides in the evaluation set of each patient was set at 1:100, i.e. an immunogenicity prediction tool needs to prioritize 1 neoantigen out of every 100 negative peptides. Any common peptides found in both the evaluation set and the training data were excluded from the training data. The expanded evaluations sets of 18 patients are provided in the Supplementary Data 3.

Bidirectional LSTM network and model training

The model implementation was done in the TensorFlow[39] framework (version 2.6). Amino acid letters were converted to integer indices and peptide sequences were padded with 0 till a maximum length of 15. We used a Keras Sequential model that consists of an embedding layer of 8 neural units, a bi-directional LSTM layer of 8 units, a fully-connected layer with L2 regularizer, and a sigmoid activation layer. The model was trained using Adam optimizer and binary cross entropy loss for 100 epochs, only model weights with the best validation loss were saved. The Python implementation is available on GitHub.

For each individual patient, the training data was splitted into three sets train-validation-test with ratio 80%-10%-10% and no overlapping peptides. As the number of negative peptides was often several times higher than the number of positive peptides, we performed downsampling on negative peptides and trained 100 ensemble models per individual patient. The ensemble models were sorted according to their performance on the validation set and the average of the top-10 models was selected as the final prediction model for that patient. The pre-trained models of 18 cancer patients are available on GitHub. For prediction, DeepImmun takes an input peptide sequence of a patient and outputs a score from 0 to 1, a higher score indicates that the input peptide is more likely to be recognized by CD8+ T cells of that patient. Since each model is personalized, its prediction should be interpreted only in the context of the patient whose personal data was used to train the model.

Evaluation of immunogenicity prediction tools

DeepImmun, PRIME[14] (version 2.0), NetMHCpan[7] (version 4.1), and IEDB immunogenicity predictor[12] were tested on the neoantigen evaluation datasets of 18 cancer patients described above. Their predictions scores are provided in Supplementary Data 3. The four prediction tools were evaluated based on two criteria. The first criterion is the area under the receiver operating characteristic curve (ROC-AUC) of their predictions. DeepImmun score is a value from 0 to 1 (explained above). For PRIME and NetMHCpan, we used their %Rank scores, which represent the rank of the predicted score compared to a set of random natural peptides (as suggested in the tools' documentation). IEDB immunogenicity predictor's score is mainly a log enrichment score, often from -1 to 1; a higher score indicates a higher chance that the peptide is immunogenic. The tools' predicted scores were used to calculate the ROC-AUC with scikit-learn[41]. The ROC-AUC was calculated on the evaluation set of each patient and on the combined evaluation set of all 18 patients (Figures 2c and 2a, respectively)

The second criterion is the tools' ability to rank neoantigens on top of non-immunogenic peptides in the evaluation set of each individual patient. The peptides in each evaluation set were sorted according to the tools' predicted scores and their percentiles were calculated. The percentiles of the neoantigens were then used to plot the boxplots in Figure 2b.

All calculations of the performance evaluation are provided in the Python scripts on GitHub.

Author contributions

M.L. designed and supervised the research. N.H.T implemented the model and performed the data analysis. N.H.T and M.L. wrote the manuscript.

Acknowledgements

This work was funded by the NSERC grant OGP0046506 and the Canada Research Chair programme for M.L. We thank D. Bu, S.W. Sun, B. Shan, L. Xin, R. Qiao for discussions on the relevant topics.

Declaration of interests

N.H.T is an employee of Bioinformatics Solutions Inc.

Data and code availability

The training and evaluationdatasets of 18 cancer patients are provided in the Supplementary Data 1-3. The Python scripts and pre-trained models are publicly available on GitHub via the following link: https://github.com/nh2tran/DeepImmun

References

1. Hu, Z., Ott, P. A. & Wu, C. J. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat. Rev. Immunol.* **18**, 168–182 (2018).
2. The problem with neoantigen prediction. *Nat. Biotechnol.* **35**, 97 (2017).
3. Vitiello, A. & Zanetti, M. Neoantigen prediction and the need for validation. *Nat. Biotechnol.* **35**, 815–817 (2017).
4. Tran, N. H., Xu, J. & Li, M. A tale of solving two computational challenges in protein science: neoantigen prediction and protein structure prediction. *Brief. Bioinform.* **23**, (2022).
5. Bassani-Sternberg, M. *et al.* Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat. Commun.* **7**, 13404 (2016).
6. Bulik-Sullivan, B. *et al.* Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* (2018) doi:10.1038/nbt.4313.
7. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
8. Bassani-Sternberg, M. *et al.* Deciphering HLA-I motifs across HLA peptidomes improves neo-antigen predictions and identifies allostery regulating HLA specificity. *PLoS Comput. Biol.* **13**, e1005725 (2017).
9. Sarkizova, S. *et al.* A large peptidome dataset improves HLA class I epitope prediction across most of the human population. *Nat. Biotechnol.* **38**, 199–209 (2020).
10. Racle, J. *et al.* Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat. Biotechnol.* **37**, 1283–1286 (2019).

11. Xin, L. *et al.* A streamlined platform for analyzing tera-scale DDA and DIA mass spectrometry data enables highly sensitive immunopeptidomics. *Nat. Commun.* **13**, 1–9 (2022).
12. Calis, J. J. A. *et al.* Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput. Biol.* **9**, e1003266 (2013).
13. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1754–62 (2015).
14. Schmidt, J. *et al.* Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep Med* **2**, 100194 (2021).
15. Gfeller, D. *et al.* Predictions of immunogenicity reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *bioRxiv* 2022.05.23.492800 (2022) doi:10.1101/2022.05.23.492800.
16. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **183**, 818–834.e13 (2020).
17. Richman, L. P., Vonderheide, R. H. & Rech, A. J. Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Syst* **9**, 375–382.e4 (2019).
18. Springer, I., Besser, H., Tickotsky-Moskovitz, N., Dvorkin, S. & Louzoun, Y. Prediction of Specific TCR-Peptide Binding From Large Dictionaries of TCR-Peptide Pairs. *Front. Immunol.* **11**, 1803 (2020).
19. Montemurro, A. *et al.* NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Commun Biol* **4**, 1060 (2021).
20. Fischer, D. S., Wu, Y., Schubert, B. & Theis, F. J. Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Mol. Syst. Biol.* **16**, e9416 (2020).
21. Dhusia, K., Su, Z. & Wu, Y. A structural-based machine learning method to classify binding affinities between TCR and peptide-MHC complexes. *Mol. Immunol.* **139**, 76–86 (2021).
22. Morel, P. A., Faeder, J. R., Hawse, W. F. & Miskov-Zivanov, N. Modeling the T cell immune

response: a fascinating challenge. *J. Pharmacokinet. Pharmacodyn.* **41**, 401–413 (2014).

23. Hogquist, K. A., Baldwin, T. A. & Jameson, S. C. Central tolerance: learning self-control in the thymus. *Nat. Rev. Immunol.* **5**, 772–782 (2005).
24. Vita, R. *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res.* **47**, D339–D343 (2019).
25. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
26. Kalaora, S. *et al.* Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature* **592**, 138–143 (2021).
27. Apavaloaei, A. *et al.* Induced pluripotent stem cells display a distinct set of MHC I-associated peptides shared by human cancers. *Cell Rep.* **40**, 111241 (2022).
28. Bentzen, A. K. *et al.* Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* **34**, 1037–1045 (2016).
29. McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
30. Bobisse, S. *et al.* Sensitive and frequent identification of high avidity neo-epitope specific CD8 + T cells in immunotherapy-naive ovarian cancer. *Nat. Commun.* **9**, 1092 (2018).
31. Rajasagi, M. *et al.* Systematic identification of personal tumor-specific neoantigens in chronic lymphocytic leukemia. *Blood* **124**, 453–462 (2014).
32. Ott, P. A. *et al.* An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217–221 (2017).
33. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**, 63–66 (2019).
34. Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 8247–8252 (2017).

35. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* **12**, 3346 (2021).
36. Zhou, X.-X. *et al.* pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **89**, 12690–12697 (2017).
37. Ma, C. *et al.* Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **90**, 10881–10888 (2018).
38. Tran, N. H., Zhang, X. & Li, M. Deep Omics. *Proteomics* **18**, (2018).
39. Developers, T. *TensorFlow*. (2022). doi:10.5281/zenodo.6574269.
40. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
41. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
42. Tran, N. H. *et al.* Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence* **2**, 764–771 (2020).

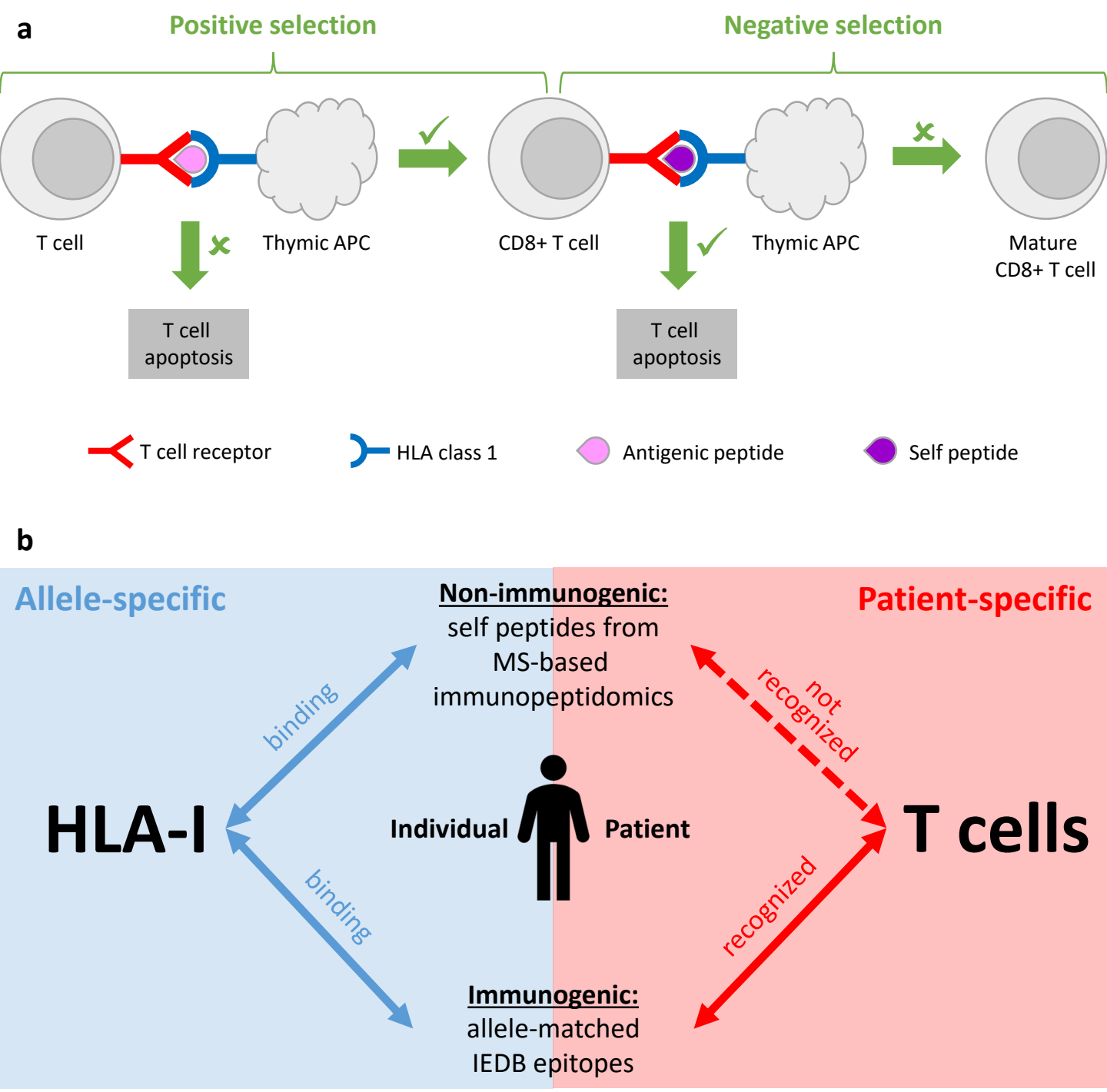
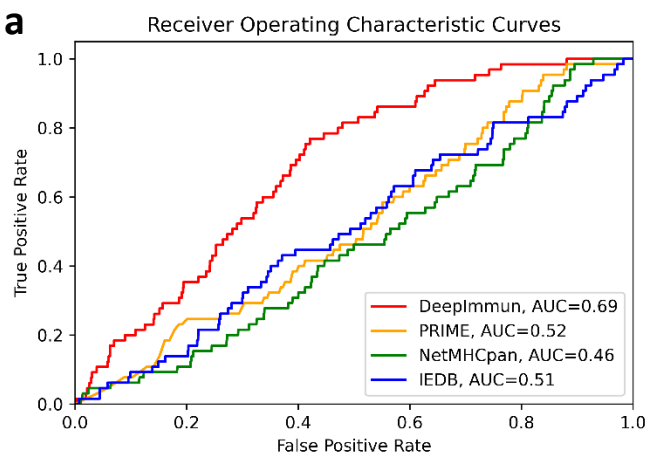


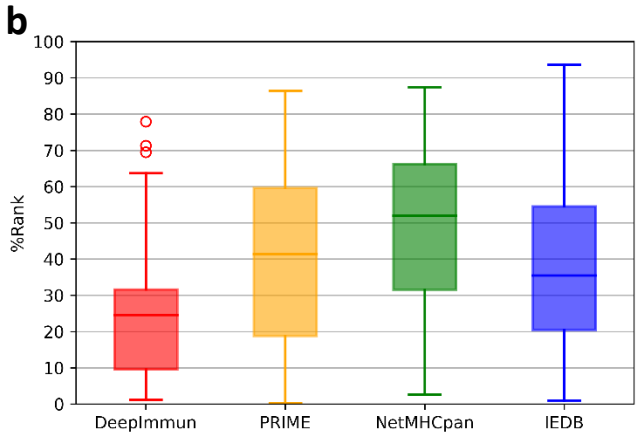
Figure 1. Predicting the immunogenicity of candidate neoantigens by modeling the central tolerance of CD8+ T cells in each individual patient.

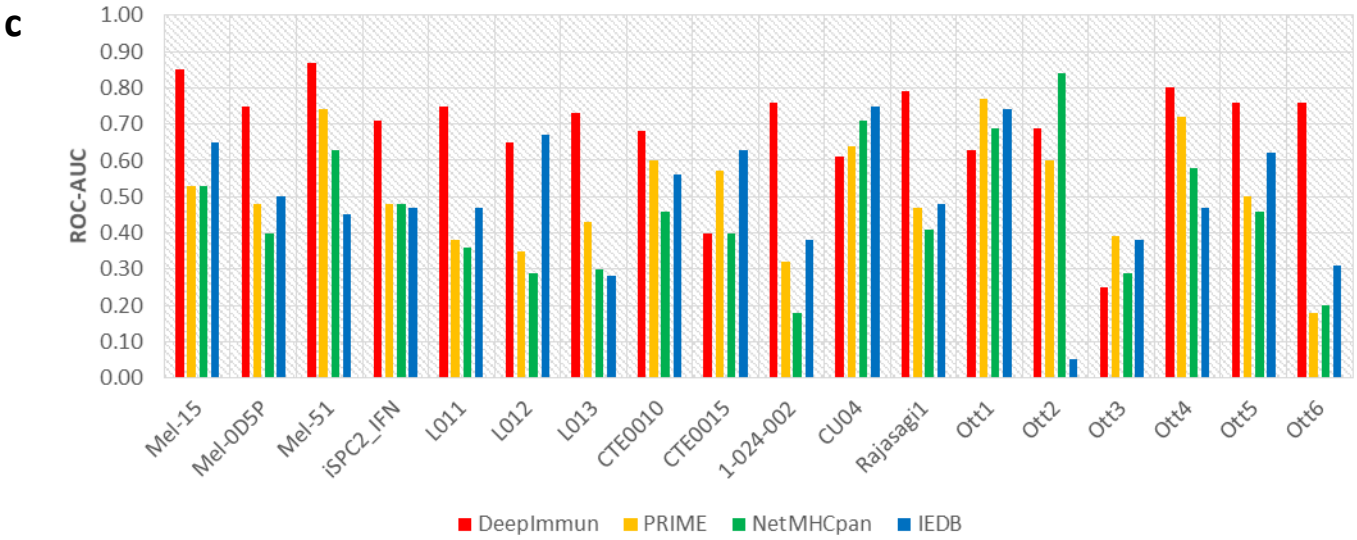
a. The positive and negative selection processes, i.e. the central tolerance of CD8+ T cells.
During the positive selection, T cells are selected by their ability to bind to peptide-HLA complexes: they will become CD8+ T cells if they bind to HLA-I complexes, otherwise they will die by apoptosis. During the negative selection, CD8+ T cells are further selected ***against*** their ability to bind to self peptides: those that have high affinity for self peptides will die by apoptosis, preventing the risk of autoimmunity; the remaining will become mature CD8+ T cells.

b. Personalized immunogenicity prediction model to resemble the central tolerance of CD8+ T cells in an individual patient.
We use MS-based immunopeptidomics to obtain HLA-I self peptides to resemble the negative selection of CD8+ T cells in an individual patient. For the positive selection, we collect all epitopes reported in positive T cell assays from IEDB that match the patient's HLA-I alleles. Using this personal dataset of positive and negative peptides, we then train a binary classification model specifically for that patient to predict his/her T cell response to any given peptide.

(APC: Antigen-Presenting Cell; HLA: Human Leukocyte Antigen; MS: Mass Spectrometry; IEDB: Immune Epitope Database)







d

Neoantigens of patient Mel-15	Predicted ranks (%)			
	DeepImmun	PRIME	NetMHCpan	IEDB
K**L**ILWRGLK	1.2	53.5	57.4	4.4
GRIAF**F**LKY	2.0	3.8	10.1	19.8
RLF**L**GLAIK	2.6	18.8	24.8	22.2
RTYSL**S**SALR	26.5	63.4	48.1	98.4
SQ**I**ILRQH	46.1	98.4	93.7	30.3

Figure 2. Performance evaluation of four immunogenicity prediction tools on 18 cancer patients.
a. Receiver operating characteristic curves on the combined evaluation set of 18 patients.
b. Predicted ranks of immunogenic neoantigens versus non-immunogenic peptides (lower % is better).
c. Areas under the receiver operating characteristic curves (ROC-AUC) of the prediction tools on each individual patient.
d. An example of predicted ranks of patient Mel-15's neoantigens.
(IEDB refers to the immunogenicity prediction tool on IEDB website developed by Calis et al., 2013. Bold underlined letters in panel d indicate mutated amino acids).

Table 1. Summary of training data and evaluation neoantigens of 18 individual patients.

Negative (self) refers to self peptides identified from MS-based immunopeptidomics.

Negative (allele) refers to allele-matched MHC-I eluted peptides found in the Immune Epitope Database.

Patient ID	Training peptides			Evaluation neoantigens		Study	HLA alleles
	Positive	Negative (self)	Negative (allele)	Positive	Negative		
Mel-15	568	35551	42785	5	23	Bassani-Sternberg et al., 2016	HLA-A*03:01;HLA-A*68:01;HLA-B*27:05;HLA-B*35:03;HLA-C*02:02;HLA-C*04:01
Mel-0D5P	1111	10453	64628	4	148	Chong et al., 2020	HLA-A*01:01;HLA-A*23:01;HLA-B*07:02;HLA-B*15:01;HLA-C*12:03;HLA-C*14:02
Mel-51	4458	23716	98622	4	11	Kalaora et al., 2021	HLA-A*01:01;HLA-A*02:01;HLA-B*14:02;HLA-B*15:01;HLA-C*03:04;HLA-C*08:02
iSPC2_IFN	4343	3847	81463	6	10	Apavaloaei et al., 2022	HLA-A*01:01;HLA-A*02:01;HLA-B*15:03;HLA-B*53:01;HLA-C*02:10;HLA-C*04:01
L011	1442	not available	39949	6	282	Bentzen et al., 2016	HLA-A*11:01;HLA-A*24:02;HLA-B*35:01;HLA-B*49:01
L012	1602	not available	52945	2	353	McGranahan et al., 2016	HLA-A*11:01;HLA-A*24:02;HLA-B*07:02
L013	4210	not available	83835	3	349	Bentzen et al., 2016	HLA-A*02:01;HLA-A*32:01;HLA-B*15:01;HLA-B*44:02
CTE0010	4176	not available	82040	4	430	Bobisse et al., 2018	HLA-A*02:01;HLA-A*31:01;HLA-B*15:01;HLA-B*35:02;HLA-C*03:03;HLA-C*04:01
CTE0015	1143	not available	61968	6	221	Bobisse et al., 2018	HLA-A*11:01;HLA-A*02:14;HLA-B*07:02;HLA-B*27:05;HLA-C*02:02;HLA-C*07:02
1-024-002	152	not available	29154	3	16	Bulik-Sullivan et al., 2018	HLA-A*68:01;HLA-A*68:01;HLA-B*40:02;HLA-B*40:27;HLA-C*03:04;HLA-C*03:04
CU04	133	not available	12945	3	17	Bulik-Sullivan et al., 2018	HLA-A*24:26;HLA-A*26:01;HLA-B*18:01;HLA-B*38:01;HLA-C*12:03;HLA-C*12:03
Rajasagi1	407	not available	8873	2	23	Rajasagi et al., 2014	HLA-A*33:01;HLA-A*68:12;HLA-B*35:01;HLA-B*14:01
Ott1	4785	not available	95696	2	21	Ott et al., 2017	HLA-A*02:01;HLA-A*24:02;HLA-B*44:02;HLA-B*15:01
Ott2	404	not available	21744	2	16	Ott et al., 2017	HLA-A*01:01;HLA-A*01:01;HLA-B*38:01;HLA-B*56:01
Ott3	4306	not available	86229	3	23	Ott et al., 2017	HLA-A*02:01;HLA-A*03:01;HLA-B*47:01;HLA-B*27:05
Ott4	3984	not available	67230	2	26	Ott et al., 2017	HLA-A*02:01;HLA-A*25:01;HLA-B*18:01;HLA-B*27:02
Ott5	404	not available	12019	6	28	Ott et al., 2017	HLA-A*66:01;HLA-A*23:01;HLA-B*41:02;HLA-B*35:01
Ott6	295	not available	10578	2	30	Ott et al., 2017	HLA-A*66:01;HLA-A*01:03;HLA-B*08:01;HLA-B*08:01