Predicting immunogenicity by modeling the central tolerance 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

Summary.

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 processes, 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 were used to train a binary classification model, which was then applied to predict the immunogenicity of candidate neoantigens of that patient. Evaluation results on three cancer patients and one mouse cancer cell line and showed that our personalized models achieved an average accuracy of 79% and outperformed existing immunogenicity prediction tools. Furthermore, the models were able to rank neoantigens that elicited CD8+ T-cell responses within the top 15% of candidate peptides identified from the patients, thus reducing the time and costs of complicated *in vitro* validations.

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. ~20, 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 some 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] 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, certain 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[15,16].

However, the above methods fall short of accounting for the high specificity of each individual patient's 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 predict the TCR-pHLA recognition from their sequences or structures[4,17–21]. However, despite on-going intensive efforts of TCR sequencing, it is difficult to sequence the complete TCR repertoire for every patient. Thus, we need to solve the dilemma: a personalized model for predicting TCR-pHLA recognition without really sequencing the TCRs.

To address this problem, we note that the specificity of T cells are mainly shaped by the positive and negative selection processes, 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-I or HLA-II, respectively; otherwise they will die by neglect[22]. 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[22].

Here we take an alternative approach and hypothesize that the positive and negative selections can be encapsulated by the complete space of positive and negative peptides resulting from those selections. 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. The reason that we focus on HLA-I and CD8+ but not HLA-II and CD4+ is because of the data availability and the corresponding complexity of immunopeptidomics assays. Since the positive selection is not really personalized but rather HLA-dependent, we can collect all epitopes reported in positive T cell assays from the Immune Epitope Database (IEDB) that match the patient's HLA-I alleles. Using this personal dataset of positive and negative peptides, we can train a binary classification model specifically for that patient to predict his/her T cell response to any given peptide. It's worth to emphasize that, while previous studies have used immunogenic and non-immunogenic peptides to train prediction models[12,14], 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 cancer patients, as well as the performance comparison to other tools.

Results

Personalized model training to resemble the central tolerance of CD8+ T cells in an individual patient

Figures 1B describes how we prepared the training/evaluation data for each individual patient and trained a personalized model to predict the response of his/her CD8+ T cells.

Negative (non-immunogenic) training data: HLA-I self peptides from MS-based immunopeptidomics

For each individual patient, HLA-I peptides were retrieved from HLA-I immunoprecipitation assays followed by MS experiments on the patient's sample. The MS data was searched against the standard Swiss-Prot human protein database using PEAKS Xpro[11,23] with no-enzyme-specific cleavage and false discovery rate (FDR) of 1%. 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 of the patient and not recognized by the patient's T cells (i.e. non-immunogenic).

Positive (immunogenic) training data: allele-matched IEDB epitopes

We downloaded the table of T cell assays from IEDB with the following filters: linear epitopes, HLA class I, and host as human or mouse. For each individual patient, we selected 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).

Personalized model training

As illustrated in Figure 1B, the above self-peptides and allele-matched IEDB epitopes can both bind to the patient's HLA-I molecules, and the HLA binding is allele-specific. However, the former cannot be recognized by the patient's CD8+ T cells but the latter can, and the T cell

recognition is patient-specific. Thus, we can train a binary classification model using the above negative and positive data, and the trained model, named DeepImmun, 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[23,24], spectrum and retention time predictions[25–27], and other peptide prediction tasks[4,28]. More details of the model implementation and the training process can be found in the STAR Methods section.

Experimentally validated neoantigens for evaluation

To accurately evaluate the personalized model of each individual patient, it is essential to use the neoantigens that had been validated by T cell assays on that same patient. Usually there were a few dozens of candidate neoantigens that were experimentally tested per patient, with less than half a dozen showing positive T cell responses and the rest showing negative. Note that those candidate neoantigens had been carefully excluded from the patient's training data before the training process.

Performance evaluation of immunogenicity prediction tools

As described above, our personalized approach required the personal data of each individual patient for training and evaluation. In particular, HLA-I self peptides from MS-based immunopeptidomics are needed as negative training data and experimentally validated neoantigens are needed for evaluation; both types of data have to come from the same patient. This kind of individual patient's data is limited and we were able to find four such datasets from the literature[5,25,29–31], including three cancer patients and one mouse cancer cell line: Mel-15, Mel-0D5P, Mel-51, and Mouse-EL4. A summary of those datasets is presented in Figure 2A. The number of HLA-I self peptides varied from 746-35548 per patient, while the number of immunogenic epitopes varied from 304-2417. The number of candidate neoantigens tested in T cell assays varied from 7-152 and the number of immunogenic neoantigens varied from 2-5 per patient. The full lists of peptides for training and evaluation were provided in Supplementary Tables S1, S2.

Mel-15 is a melanoma patient dataset first published by Bassani-Sternberg et al.[5]. Bassani-Sternberg et al.[5] and Wilhelm et al.[25] identified 28 mutated HLA-I peptides from the MS and RNA-seq data of native tumor tissues of the patient. The peptides were subsequently tested against the patient's own T cells and five of them showed positive responses (Figure 2A).

Mel-5D0P is a melanoma patient dataset first published by Chong et al.[29]. The authors used an MS-based proteogenomic approach to identify non-canonical HLA peptides from tumor samples. They identified 152 HLA-I peptides from long non-coding genes, transposable elements, alternative open reading frames, and tumor-associated antigens; four of them showed positive responses against the patient's own T cells. Mel-51 dataset was obtained from Kalaora et al.[30], who investigated bacteria-derived HLA peptides from melanoma samples. 15 recurrent bacteria-derived HLA-I peptides were identified from patient Mel-51, four of them showed positive T cell responses. The murine cancer cell line Mouse-EL4 dataset was obtained from Laumont et al.[31], who used an MS-based proteogenomic approach to investigate non-coding regions to identify tumor-specific antigens. The authors identified seven tumor-specific peptides, including five aberrantly expressed and two mutated; two of them showed positive T cell responses.

For performance evaluation, we compared our method to three leading tools, including PRIME[14], NetMHCpan[7], 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. It should be noted that NetMHCpan is designed for predicting HLA-I binding rather than immunogenicity, but it is widely used in many neoantigen prediction workflows. We included NetMHCpan here for reference, but evaluating it for immunogenicity prediction may be unfair.

We evaluated the four prediction tools based on two criteria. The first criterion is the area under the receiver operating characteristic curve (ROC-AUC) of their predictions on candidate neoantigens that had been tested in T cell assays of each individual patient. Figure 2B shows the AUC of the four prediction tools. DeepImmun outperformed the other tools on two patients Mel-15 and Mel-0D5P. On patient Mel-51, DeepImmun, PRIME, and NetMHCpan showed comparable performance. For Mouse-EL4, IEDB predictor achieved the highest AUC of 90%, followed by DeepImmun and NetMHCpan (PRIME does not support mouse data). The unusually higher performance of the IEDB predictor on mouse data than on human data may be because Mouse-EL4 contains two well-studied alleles, H2-Db and H2-Kb, which were included in IEDB training data[12]. Overall, DeepImmun achieved an average AUC of 79% per dataset. The

relative performance between the four predictions tools also reflected the characteristics of their underlying models: DeepImmun is a personalized model; PRIME is an allele-specific model; IEDB predictor is a general model trained on a limited dataset; and NetMHCpan is a HLA-binding prediction model.

The main purpose of immunogenicity prediction is to prioritize and select candidate neoantigens to reduce the time and costs of *in vitro* validations. Thus, the second evaluation criterion is the ability to rank neoantigens among potentially mutated peptides identified from an individual patient. Mutated peptides, including neoantigens, cannot be found in the standard Swiss-Prot human protein database. Thus, a de novo sequencing workflow was first applied to the MS data[32] of the patient to identify new peptides that were not in the database. The four prediction tools were then applied to rank those candidate peptides, including mutated peptides and the neoantigens. They were expected to rank the neoantigens at the top of the list of candidate peptides. It's worth noting that our MS-based approach only considered candidate peptides identified from MS data, while other genomic approaches[1] often considered candidate peptides translated from genomic sequences.

Figure 2C shows how the four prediction tools ranked five neoantigens among 3638 candidate peptides identified by de novo sequencing on the MS data of patient Mel-15's tumor tissues. Notably, the first two neoantigens, K**L**ILWRGLK and RLF**L**GLAIK, were ranked within the top 1% by DeepImmun, while the next two, GRIAF**F**LKY and RTYSL**S**SALR, were ranked within the top 1.5% and 15%, respectively. The last neoantigen, SQ**I**ILRQH, should be interpreted with caution: it was identified and tested by Wilhelm et al.[25] as positive; however, its NetMHCpan %Rank is 11.1%, indicating that it may be not likely to be presented by HLA molecules. Overall, DeepImmun outperformed the other tools in this ranking evaluation. The list of peptides, their scores and ranks are provided in Supplementary Table S3.

Discussion

In this study we proposed DeepImmun, a personalized model for immunogenicity prediction by resembling the negative and positive selections of CD8+ T cells in an individual patient. The model used HLA-I self peptides derived from MS-based immunopeptidomics of the patient as negative training data and allele-matched positive T cell epitopes from the IEDB as positive training data. We showed that DeepImmun achieved an average accuracy of 79% on four individual patients and outperformed existing tools. More importantly, DeepImmun was able to

rank immunogenic neoantigens within the top 15% of candidate de novo peptides identified from the MS data of the patient tumor tissues, thus reducing the time and costs for further *in vitro* validations.

Our approach requires HLA-I self peptides derived from MS-based immunopeptidomics of an individual patient for training and experimentally validated neoantigens of the same patient for evaluation. Since this kind of individual patient's data is limited, we were able to find only four such datasets for the analysis in this study. Nevertheless, we strongly believe that our results have demonstrated the validity and the benefits of our personalized approach to model the negative and positive selections of CD8+ T cells for immunogenicity prediction. While our approach is based on MS immunopeptidomics, which currently may be not widely available as other genomic approaches, our study shall propel the application of MS-based immmunopeptidomics for the development of personalized cancer immunotherapy.

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.

Figure legends

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.

(B) Personalized model training to resemble the central tolerance of CD8+ T cells in an individual patient.

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

Figure 2. Performance evaluation of four immunogenicity prediction tools on three individual patients and one mouse cancer cell line.

(A) Summary of HLA alleles, training and evaluation data of four patients.

(B) Area Under the ROC Curve (AUC) of the prediction tools on experiment-validated neoantigens.

(C) Predicted ranks of immunogenic neoantigens among candidate peptides identified by de novo sequencing of MS data of patient Mel-15.

(PRIME does not support mouse data. IEDB refers to the immunogenicity prediction tool on IEDB website developed by Calis et al., 2013. Bold underlined letters in panel C indicate mutated amino acids. ROC: Receiver Operating Characteristic; MS: Mass Spectrometry).

STAR Methods

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Ming Li (mli@uwaterloo.ca).

Materials availability

This study did not generate new materials.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- Source code and model parameters are available on GitHub via the following link: https://github.com/nh2tran/DeepImmun
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Preparation of training data for each patient

HLA-I peptides were retrieved from HLA-I immunoprecipitation assays followed by MS experiments on the patient's sample. In this paper, we re-analyzed MS data from previous studies and did not perform the HLA or MS experiments ourselves. The MS data was searched against the standard Swiss-Prot human protein database (version June 15, 2020) using PEAKS Xpro[11,23] (version 10.6) with no-enzyme-specific cleavage and FDR 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 formed the negative training data. Note that, ideally, those HLA-I self peptides should be identified from a normal sample, but for a tumor sample that contains both normal and tumor cells, this approach still works because the database search uses standard human proteins without any tumor mutations. There might be

an elevated number of peptides coming from tumor-associated proteins, but they are still considered as self peptides.

Table of T cell assays was downloaded from IEDB with the following filters: linear epitopes, HLA class I, and host as human or mouse. For each individual 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.

The training data is provided in Supplementary Table S1.

Bidirectional LSTM network and model training

The model implementation was done in the TensorFlow[33] 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.

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 trained model is named as DeepImmun. For an input peptide sequence, DeepImmun 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 a particular patient. Since the 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 1.0), NetMHCpan[7] (version 4.1), and IEDB immunogenicity predictor[12] were evaluated on the neoantigens that had been validated by T cell assays on three cancer patients and one mouse cancer cell line[5,25,29–31]. The evaluation data is provided in Supplementary Table S2. For each individual patient, the evaluation data had been carefully excluded from the training data before the training process.

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 on candidate neoantigens tested in T cell assays. 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 by the tools' authors). 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 and the T cell responses of tested neoantigens were used to calculate the ROC-AUC with scikit-learn[34].

The second criterion is the tools' ranking of neoantigens among potentially mutated peptides identified from each individual patient. Mutated peptides, including neoantigens, cannot be found in the standard Swiss-Prot human protein database. Thus, a de novo sequencing workflow was first applied to the MS data[32] to identify new peptides that were not in the database. 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. A second-round database search using PEAKS Xpro was performed to select de novo peptides with FDR 1%. The identified peptides were further filtered by length and HLA-I binding, only those with 8-14 amino acids and NetMHCpan %Rank < 2% were selected. Those filters were to ensure high-confidence identifications and proper HLA characteristics.The four prediction tools were then applied to rank those candidate peptides, including the neoantigens; they were expected to rank the neoantigens at the top of the list of candidate peptides for each individual patient.

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. 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).
16. 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).
17. 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).
18. 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).
19. 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).
20. 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).
21. 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).

22. Hogquist, K. A., Baldwin, T. A. & Jameson, S. C. Central tolerance: learning self-control in the thymus. *Nat. Rev. Immunol.* **5**, 772–782 (2005).
23. Tran, N. H. *et al.* Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nat. Methods* **16**, 63–66 (2019).
24. 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).
25. Wilhelm, M. *et al.* Deep learning boosts sensitivity of mass spectrometry-based immunopeptidomics. *Nat. Commun.* **12**, 3346 (2021).
26. Zhou, X.-X. *et al.* pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Anal. Chem.* **89**, 12690–12697 (2017).
27. Ma, C. *et al.* Improved Peptide Retention Time Prediction in Liquid Chromatography through Deep Learning. *Anal. Chem.* **90**, 10881–10888 (2018).
28. Tran, N. H., Zhang, X. & Li, M. Deep Omics. *Proteomics* **18**, (2018).
29. Chong, C. *et al.* Integrated proteogenomic deep sequencing and analytics accurately identify non-canonical peptides in tumor immunopeptidomes. *Nat. Commun.* **11**, 1293 (2020).
30. Kalaora, S. *et al.* Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature* **592**, 138–143 (2021).
31. Laumont, C. M. *et al.* Noncoding regions are the main source of targetable tumor-specific antigens. *Sci. Transl. Med.* **10**, (2018).
32. Tran, N. H. *et al.* Personalized deep learning of individual immunopeptidomes to identify neoantigens for cancer vaccines. *Nature Machine Intelligence* **2**, 764–771 (2020).
33. Developers, T. *TensorFlow*. (2022). doi:10.5281/zenodo.6574269.
34. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**,

2825–2830 (2011).

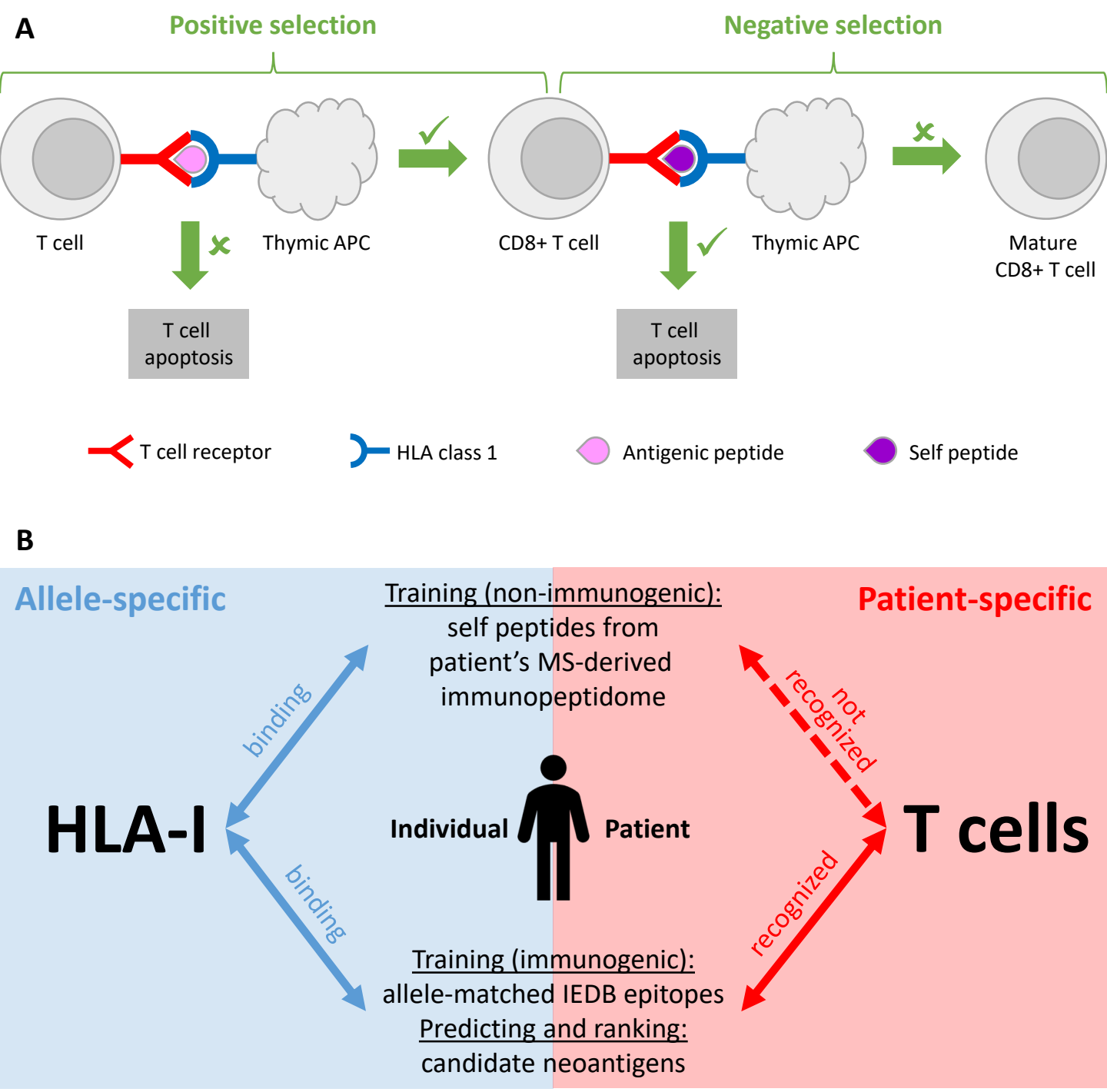
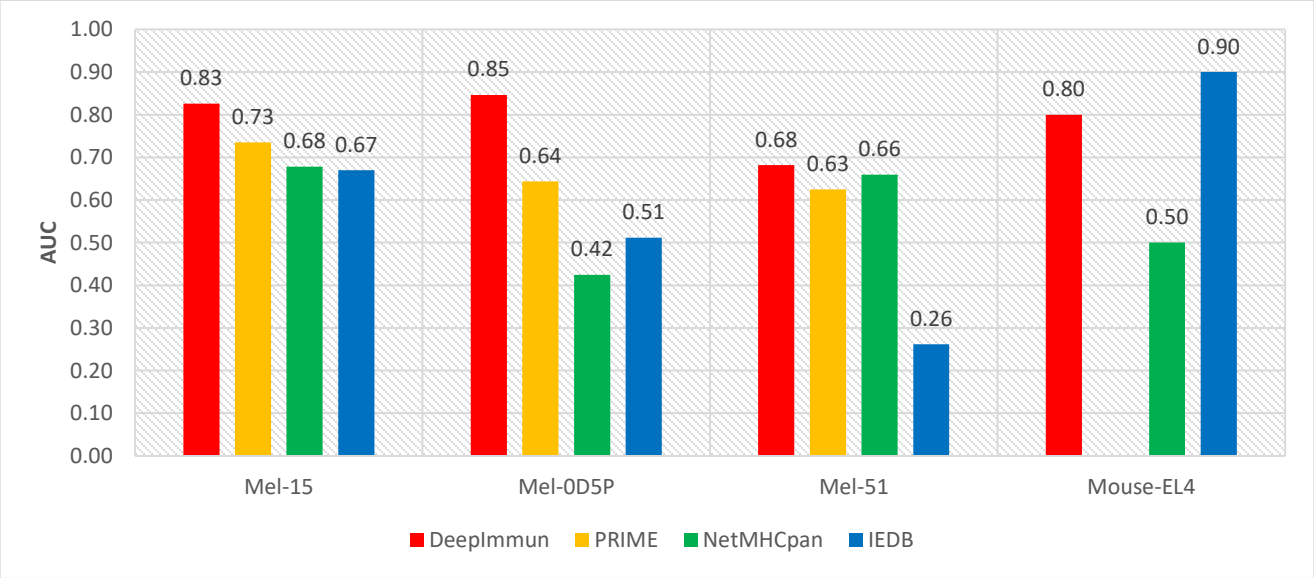


Figure 1. Predicting the immunogenicity of candidate neoantigens by modeling the central tolerance of CD8+ T cells in each individual patient.
(A) The positive and negative selection processes, i.e. the central tolerance of CD8+ T cells.
(B) Personalized model training to resemble the central tolerance of CD8+ T cells in an individual patient.
(APC: Antigen-Presenting Cell; HLA: Human Leukocyte Antigen; MS: Mass Spectrometry; IEDB: Immune Epitope Database)

A

Patient	Training peptides		Testing neoantigens		HLA alleles
	Positive	Negative	Positive	Negative	
Mel-15	304	35548	5	23	HLA-A03:01;HLA-A68:01;HLA-B27:05 HLA-B35:03;HLA-C02:02;HLA-C04:01
Mel-0D5P	497	10447	4	148	HLA-A01:01;HLA-A23:01;HLA-B07:02 HLA-B15:01;HLA-C12:03;HLA-C14:02
Mel-51	2417	23691	4	11	HLA-A01:01;HLA-A02:01;HLA-B14:02 HLA-B15:01;HLA-C03:04;HLA-C08:02
Mouse-EL4	1907	746	2	5	H2-Db;H2-Kb

B



C

Neoantigens	Predicted ranks			
	DeepImmun	PRIME	NetMHCpan	IEDB
Mel-15 (5 neoantigens/ 3638 candidates)				
K**L**ILWRGLK	33	1325	2140	43
RLF**L**GLAIK	36	818	906	401
GRIAF**F**LKY	53	125	355	349
RTYSL**S**SALR	545	1896	1782	3520
SQ**I**ILRQH	1902	3635	3634	661

Figure 2. Performance evaluation of four immunogenicity prediction tools on three individual patients and one mouse cancer cell line.

(A) Summary of HLA alleles, training and evaluation data of four patients.

(B) Area Under the ROC Curve (AUC) of the prediction tools on experiment-validated neoantigens.

(C) Predicted ranks of immunogenic neoantigens among candidate peptides identified by de novo sequencing of MS data of patient Mel-15.

(PRIME does not support mouse data. IEDB refers to the immunogenicity prediction tool on IEDB website developed by Calis et al., 2013. Bold underlined letters in panel C indicate mutated amino acids. ROC: Receiver Operating Characteristic; MS: Mass Spectrometry).