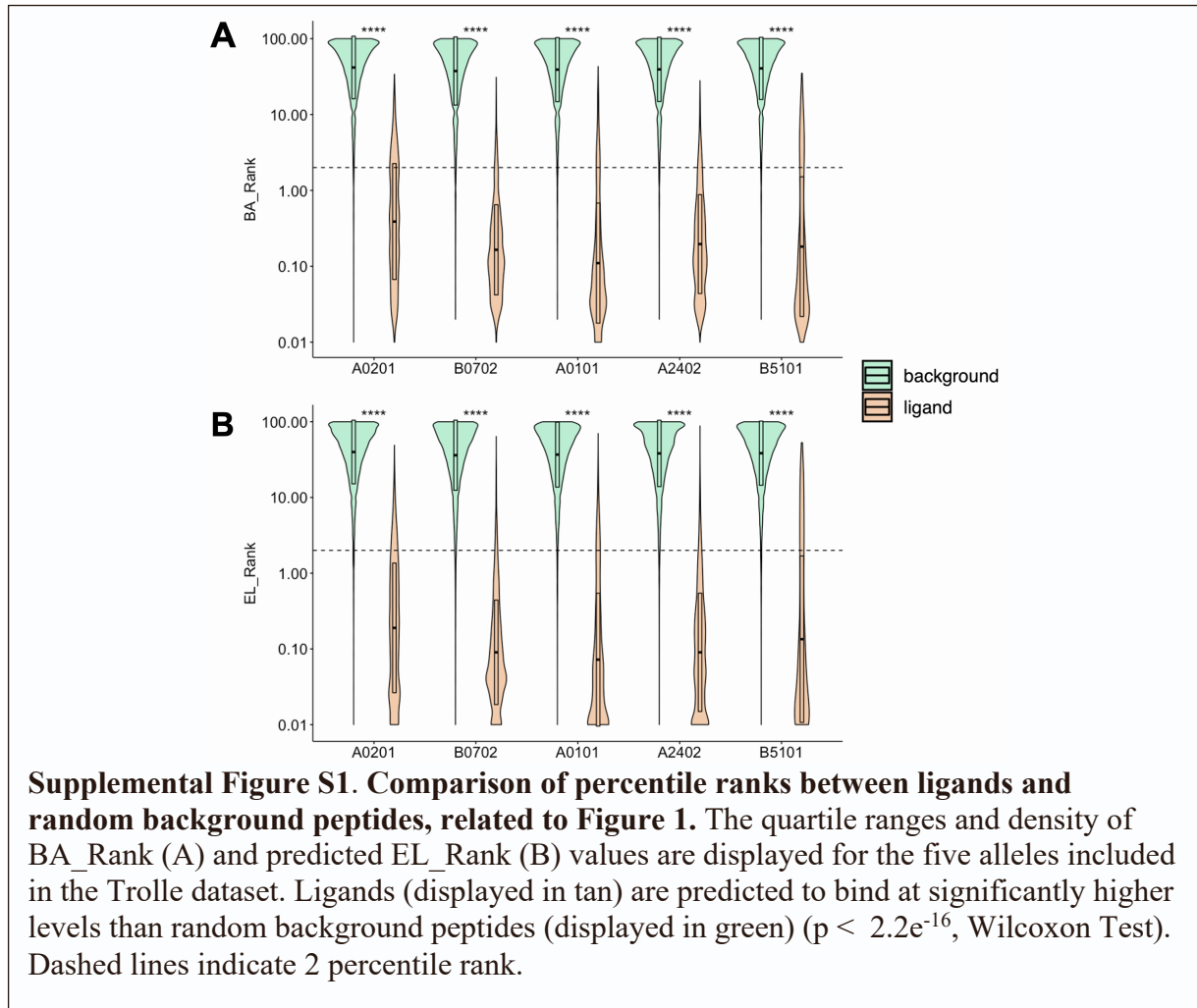## Supplemental information

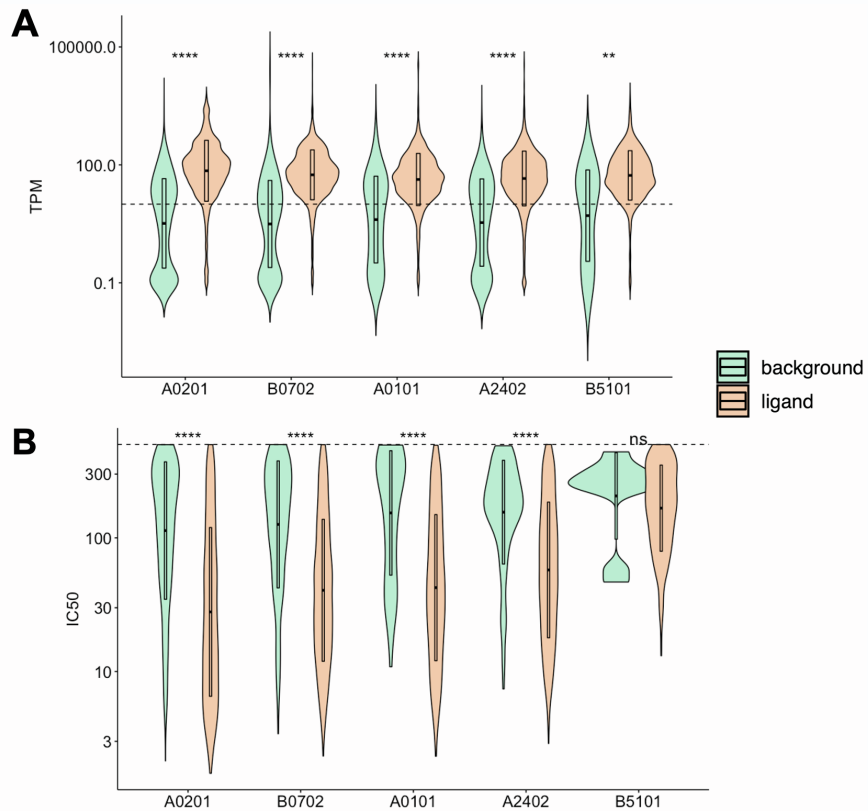## Combined assessment of MHC binding

and antigen abundance improves

T cell epitope predictions

Zeynep Koşaloğlu-Yalçın, Jenny Lee, Jason Greenbaum, Stephen P. Schoenberger, Aaron Miller, Young J. Kim, Alessandro Sette, Morten Nielsen, and Bjoern Peters

Supplemental Figures

Supplemental Figure S1. **Comparison of percentile ranks between ligands and random background peptides, related to Figure 1.** The quartile ranges and density of BA_Rank (A) and predicted EL_Rank (B) values are displayed for the five alleles included in the Trolle dataset. Ligands (displayed in tan) are predicted to bind at significantly higher levels than random background peptides (displayed in green) ($p < 2.2e^{-16}$, Wilcoxon Test). Dashed lines indicate 2 percentile rank.
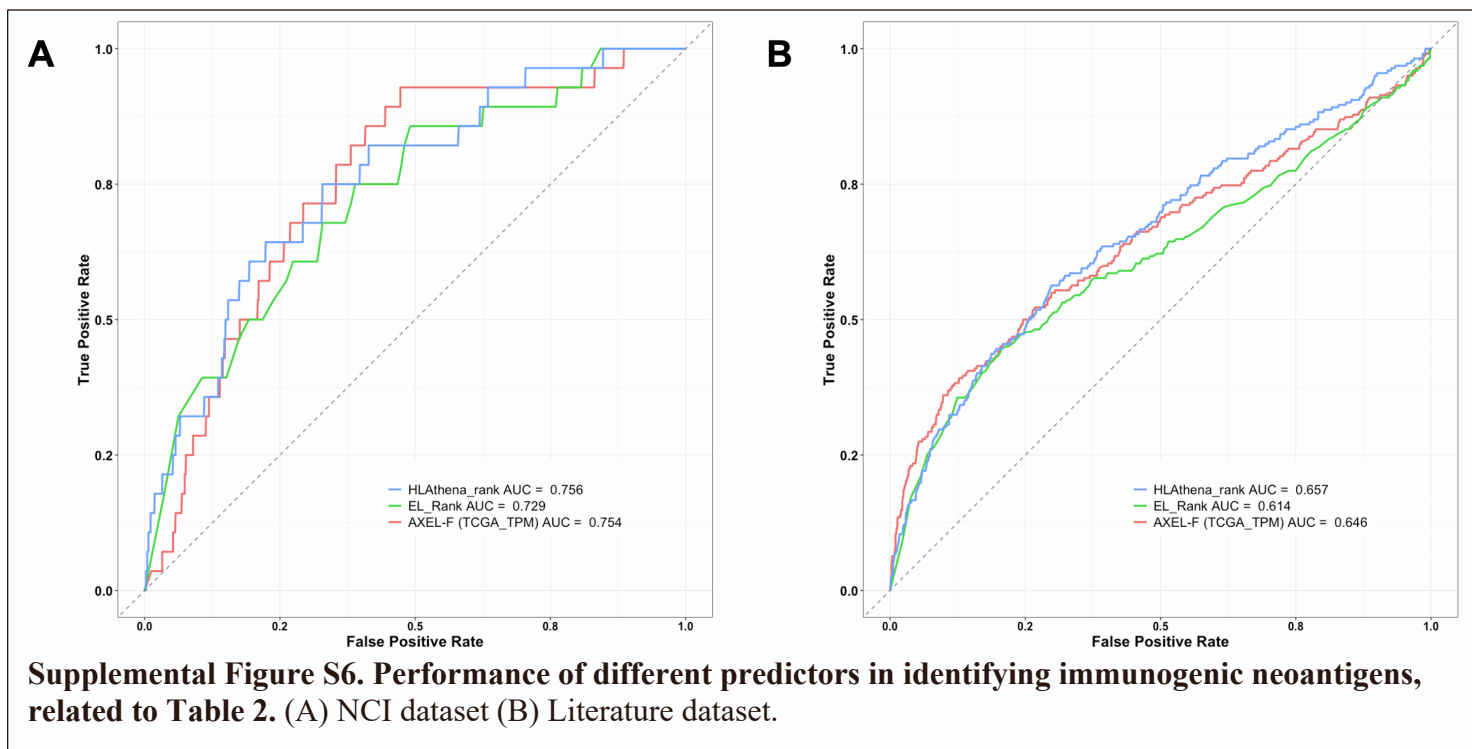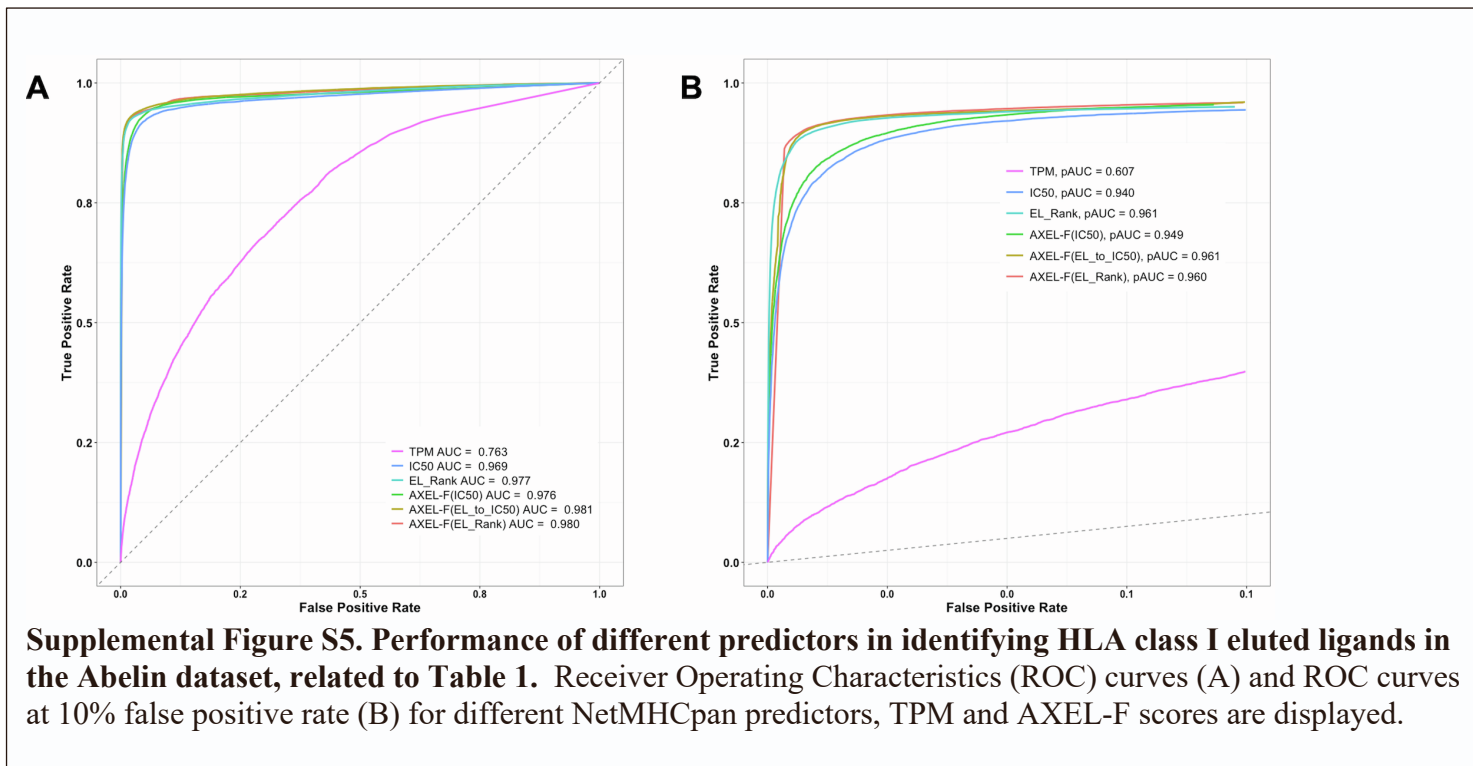
Supplemental Figure S2. **Comparison of expression and predicted binding between ligands and random background peptides for the peptide subset of predicted binders, related to Figure 1.** The quartile ranges and density of RNA expression in TPM (A) and predicted IC50 values (B) are displayed for the five alleles included in the Trolle dataset, only considering peptides with predicted binding affinities of smaller than 500nM. Ligands (displayed in tan) are predicted to bind at significantly higher levels than random background peptides (displayed in green) ($p < 2.2e^{-16}$, Wilcoxon Test). Dashed lines indicate TPM of 10 and IC50 of 500nM.
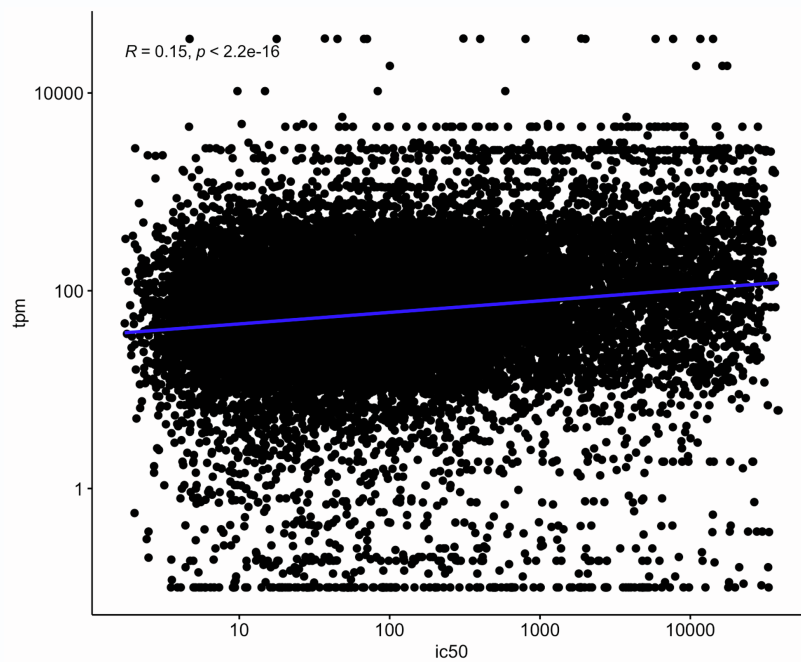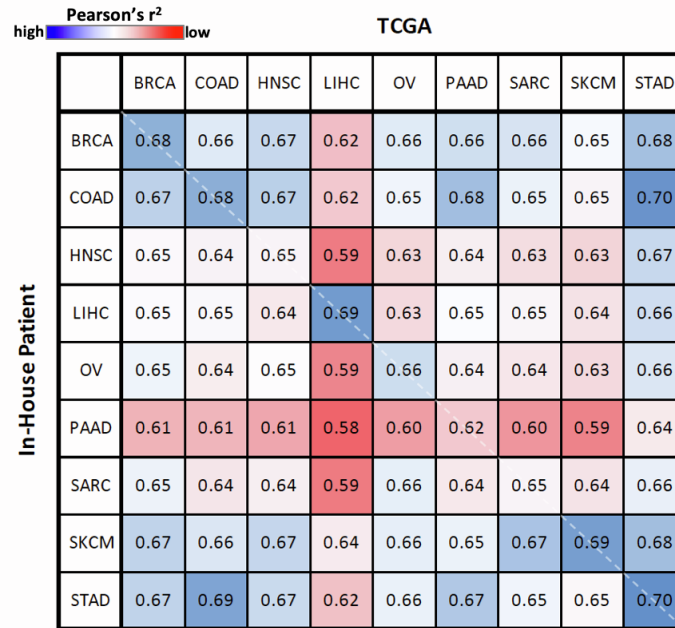
Supplemental Figure S3. **Correlation between expression and binding, related to Figure 1.** Expression values (TPM) are plotted against the predicted IC50 values (nM) for the ligands of the Trolle dataset. Expression and binding are not correlated (Pearson's correlation).
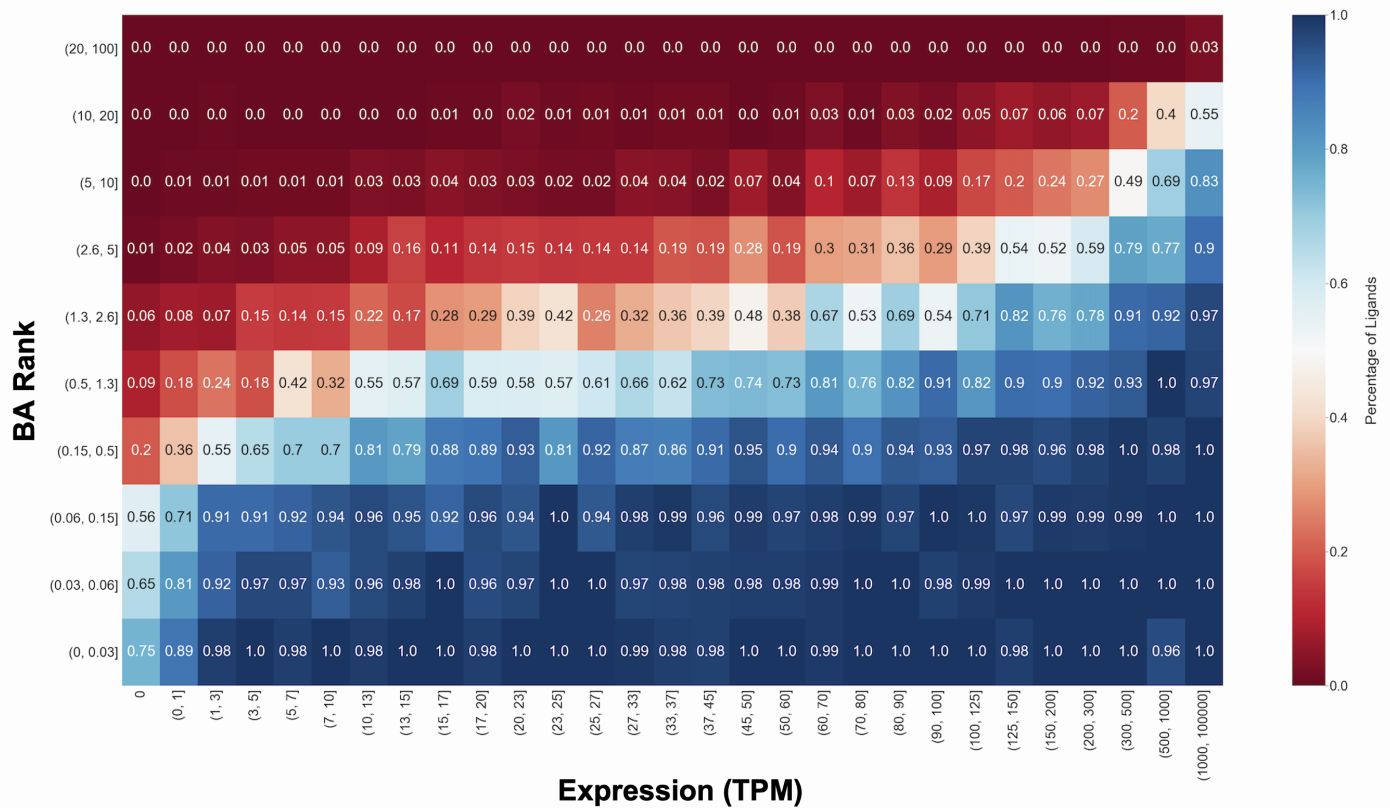
Supplemental Figure S4. **Interplay between HLA binding and abundance of source proteins of HLA class I eluted ligands when percentile ranks are used, related to Figure1.** The predicted binding ranks and TPM values were separated into ranges to create a 2-dimensional matrix with the TPM on the x-axis and the BA_Rank on the y-axis. Each peptide was assigned to a cell in this matrix according to its BA_Rank and TPM values. For each cell, the percentage of ligands among all peptides that fall into the corresponding BA_Rank and TPM ranges was determined (top number in each cell) and the cell was colored accordingly. Second number in each cell describes the number of ligands and the third number the number of background peptides.

Supplemental Figure S5. Performance of different predictors in identifying HLA class I eluted ligands in the Abelin dataset, related to Table 1. Receiver Operating Characteristics (ROC) curves (A) and ROC curves at 10% false positive rate (B) for different NetMHCpan predictors, TPM and AXEL-F scores are displayed.

Supplemental Figure S6. Performance of different predictors in identifying immunogenic neoantigens, related to Table 2. (A) NCI dataset (B) Literature dataset.

Supplemental Figure S7. Pearson's correlation coefficients between cancer type specific median TPM values from TCGA and in-house patient-specific TPM values, related to Table 1 and STAR Methods.
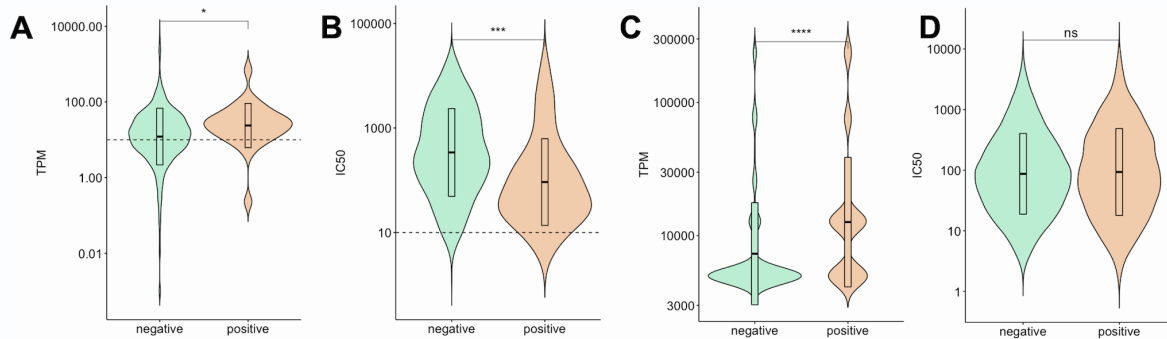
Supplemental Figure S8. **Comparison of expression and binding of epitopes and non-epitopes, Related to STAR Methods.** The quartile ranges and density of mRNA expression in TPM and predicted binding affinity in IC50 are shown for epitopes and non-epitopes from the NCI dataset (A) and (B), and for the SARS-CoV-2 dataset (C) and (D). Epitopes (displayed in tan) are expressed at significantly higher levels and are predicted to bind stronger than non-epitopes (displayed in green) (p < 2.2e^{-16}, Wilcoxon Test).

Supplemental Table S2. AUC values of predictors for Trolle set using naive approach to combine HLA binding and RNA expression, related to STAR Methods.

	AUC	AUC	pAUC	pAUC
Zero TPM penalty	-	+	-	+
TPM	0.812	n/a	0.629	n/a
IC50	0.99	0.984	0.954	0.957
BA_Rank	0.991	0.985	0.957	0.96
EL_Score	0.993	0.986	0.969	0.969
EL_Rank	0.993	0.986	0.968	0.968

Supplemental Table S5. Allele-specific AUC values of NetMHCpan and AXEL-F predictions for the Abelin dataset, related to STAR Methods. Only alleles highlighted in gray and bold were included in the training dataset of AXEL-F. Red values highlight the best performing method for the corresponding allele.

Abelin Dataset				NCI Dataset			
Alele	IC50	EL_Rank	AXEL-F		IC50	EL_Rank	AXEL-F
HLA-A*01:01	**0.995**	**0.996**	**0.996**	**HLA-A01:01**	**0.98**	**0.902**	**0.922**
HLA-A*02:01	**0.987**	**0.99**	**0.991**	**HLA-A02:01**	**0.443**	**0.514**	**0.786**
HLA-A*02:03	0.992	0.994	0.994	HLA-A03:01	0.755	0.69	0.833
HLA-A*02:04	0.991	0.994	0.994	HLA-A68:01	0.833	0.786	0.929
HLA-A*02:07	0.977	0.985	0.987	**HLA-B07:02**	**0.836**	**0.82**	**0.902**
HLA-A*03:01	0.978	0.982	0.983	HLA-B08:01	0.57	0.38	0.715
HLA-A*24:02	**0.994**	**0.994**	**0.994**	HLA-B18:01	0.684	0.526	0.614
HLA-A*29:02	0.985	0.987	0.991	HLA-B40:01	0.861	1	0.889
HLA-A*31:01	0.964	0.948	0.961	HLA-B44:03	0.988	1	1
HLA-A*68:02	0.859	0.878	0.894	HLA-C07:01	0.98	0.96	0.86
HLA-B*35:01	0.923	0.934	0.959	HLA-C07:02	0.98	1	0.99
HLA-B*44:02	0.99	0.989	0.989	HLA-C08:02	0.909	0.955	0.909
HLA-B*44:03	0.986	0.987	0.989	HLA-C12:03	0.639	0.667	0.75
HLA-B*51:01	**0.963**	**0.973**	**0.983**	# allele for which performance is best	3	4	7
HLA-B*54:01	0.963	0.962	0.968				
HLA-B*57:01	0.919	0.938	0.959				
# allele for which performance is best	3	4	14				