

iScience, Volume 25

Supplemental information

The role of antigen expression in shaping the repertoire of HLA presented ligands

Heli M. Garcia Alvarez, Zeynep Koşaloğlu-Yalçın, Bjoern Peters, and Morten Nielsen

Supplementary Figures

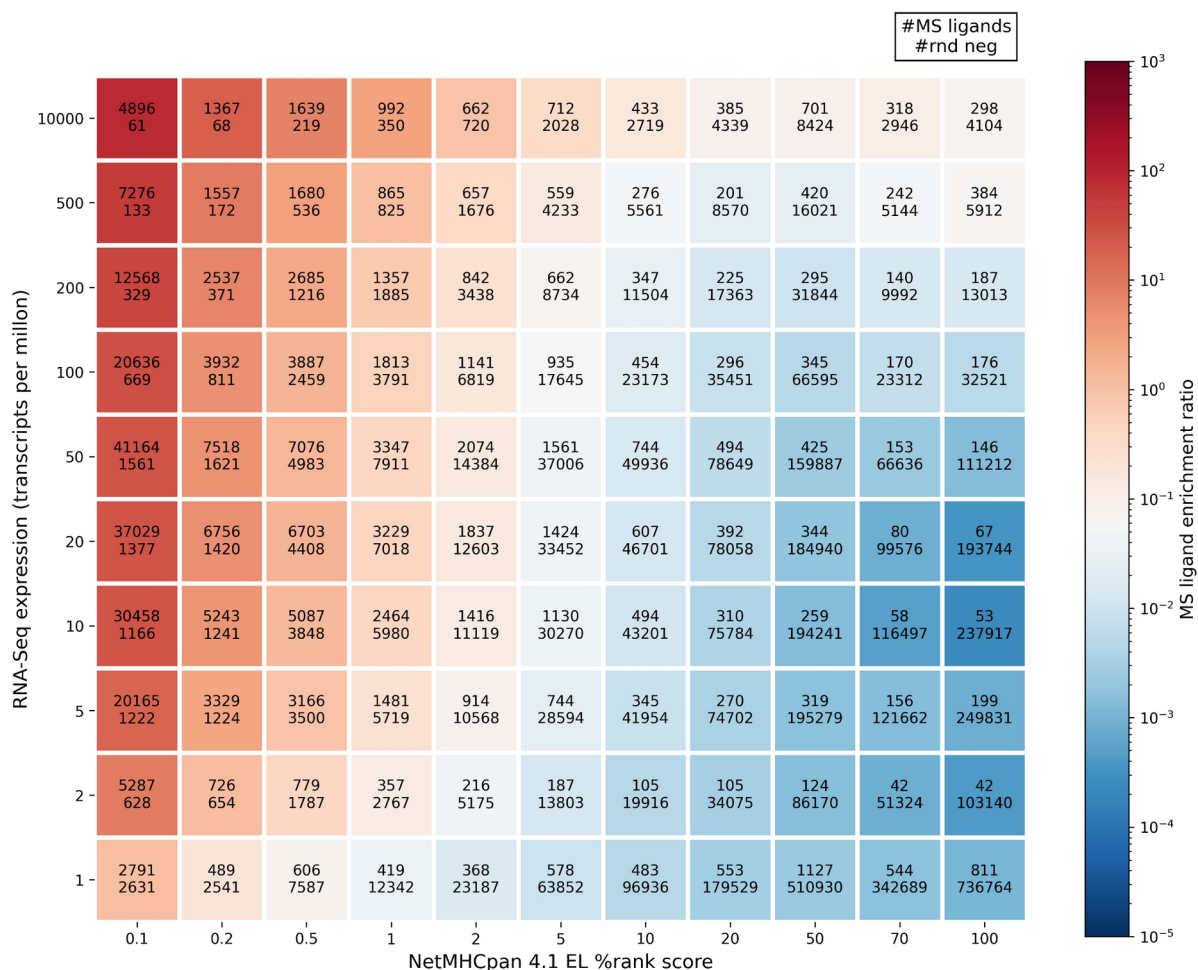


Figure S1. Relationship between predicted HLA binding scores of MS eluted ligands and artificially generated random negatives and the gene expression values of their corresponding source proteins for datasets B-D, related to Fig. 1. NetMHCpan-4.1 EL percentile rank scores and RNA-Seq expression values were binned to generate a 2-dimensional array where EL percentile rank scores are shown on the x-axis and TPM values on the y-axis. Only MS datasets from datasets B-D were used to construct this array. The numbers on both the x and y-axis represent the rightmost edge of each bin, for instance, the cell on the upper right corner contains peptides in the range (70,100] of EL %rank scores and (500, 10000] of TPM values. As an exception, the cell on the lower left corner contains peptides in the interval [0,0.1] of EL %rank scores and [0,1] of TPM values. Each cell displays the number of MS ligands (top) and the number of random natural negative peptides (bottom) that fall into it, and it is colored according to the ratio between these two quantities, referred to as the “MS ligand enrichment ratio”. The midpoint of the color scale was set to coincide with the ratio of total MS ligands to total background peptides (white cells).

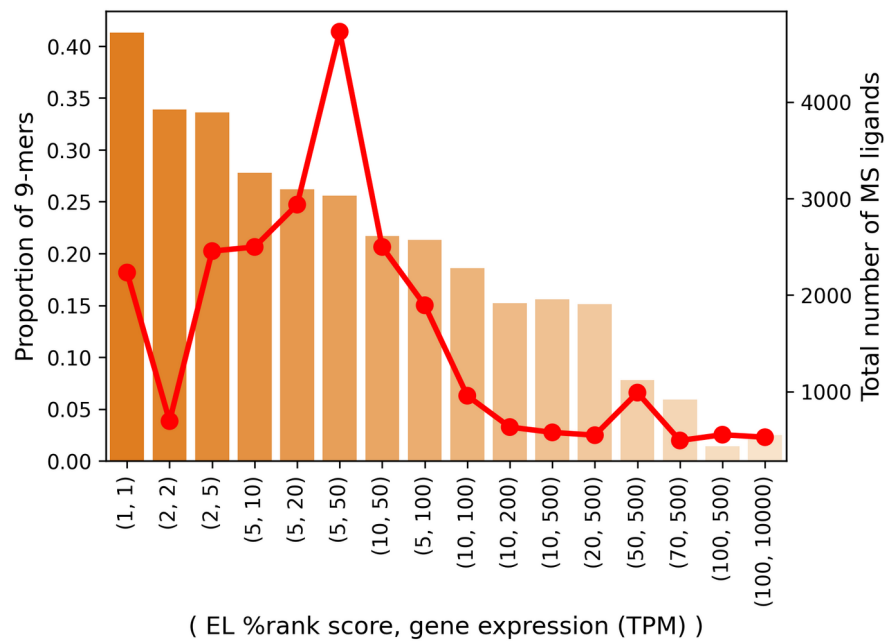


Figure S2. Proportion of 9-mers across the equivalence frontier, referenced in the array of Fig. 1. The x-axis corresponds to each of the bins in the equivalence frontier of Fig. 1 (cells shown in bright colors in Fig 1B). The y-axis shows both the proportion of 9-mers (left) and the total number of peptides (right) falling in each of the studied cells of the array.

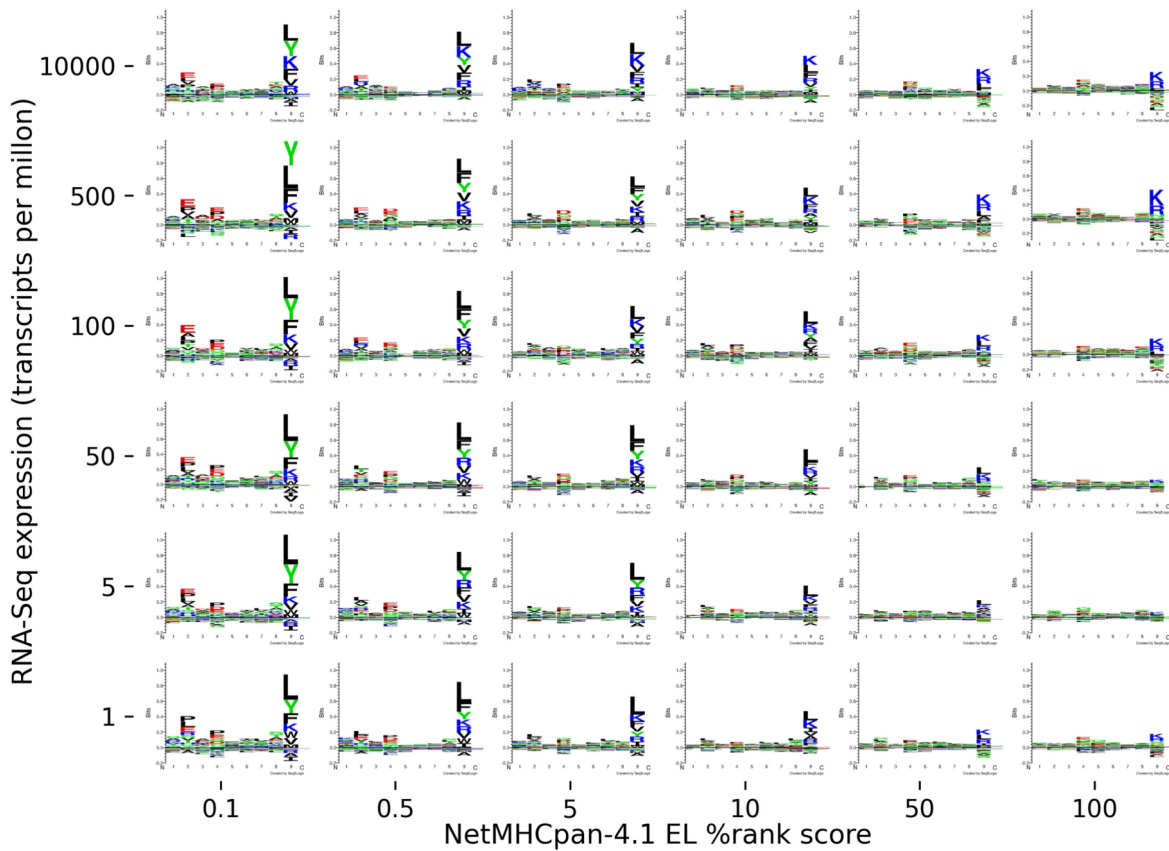


Figure S3. Binding motifs of MS eluted ligands discriminated by their predicted HLA binding score and gene expression values, related to Fig. 1. Similarly to Fig. 1, NetMHCpan-4.1 EL percentile rank scores and RNA-Seq expression values were binned to generate a 2-dimensional array where EL percentile rank scores are shown on the x-axis and TPM values on the y-axis. The numbers on both the x and y-axis represent the rightmost edge of each bin, for instance, the cell on the upper right corner contains peptides in the range (50,100] of EL %rank scores and (500, 10000] of TPM values. As an exception, the cell on the lower left corner contains peptides in the interval [0,0.1] of EL %rank scores and [0,1] of TPM values. The sequence logos show the binding preferences of positive peptides from all compiled MS datasets (datasets A-D). To construct the logos in each cell of the array, 500 peptides were randomly sampled.

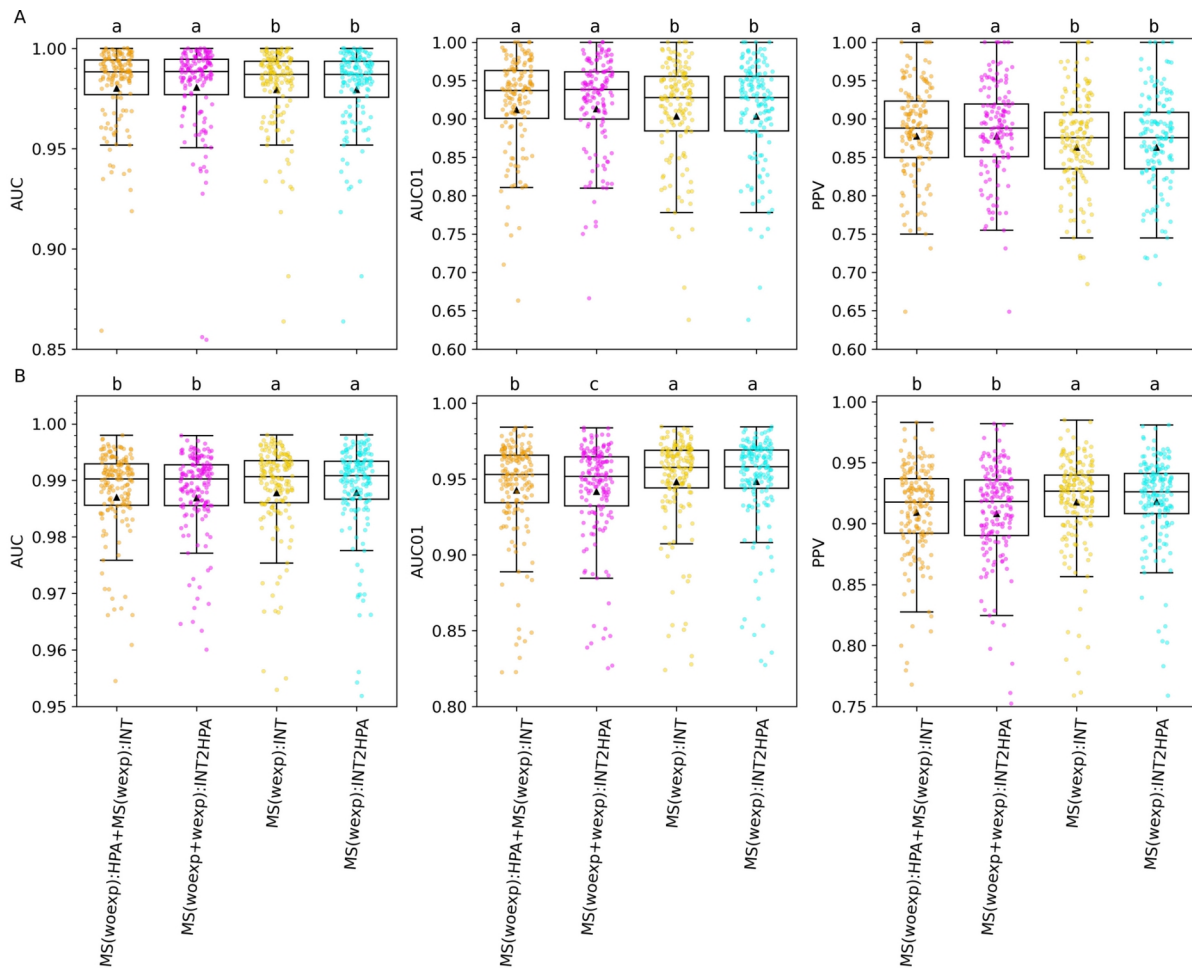


Figure S4. Performance of the models on 5-fold cross-validation for MS eluted ligands with (A) external and (B) internal RNA-Seq reference assays. Comparison of the models with an internal gene expression reference ("INT") to the ones with that same reference recalibrated to the HPA TPM value distribution ("INT2HPA"), related to Fig. 2. Predictions for the dataset A are shown in (A) while predictions for datasets B, C and D are shown in (B). For more details on the models training datasets refer to Table S2. The center line inside the box indicates the median value of the plotted metric and the triangle shows the mean. The box covers the interquartile range. The whiskers represent 1.5-fold of the interquartile range. The data points are represented using a jitter plot. The letters on top of the boxplots represent the outcome of performing all-against-all pairwise comparisons of the models' metrics using a two-tailed Binomial test, with a significance level of 5%. Apart from denoting statistical significance, the letters on top of the boxplots are assigned in alphabetical order, from the best to the worst model. That is, models with a label "a" perform at par and significantly better than models with a label "b", and so on. P-values are shown in Table S3.

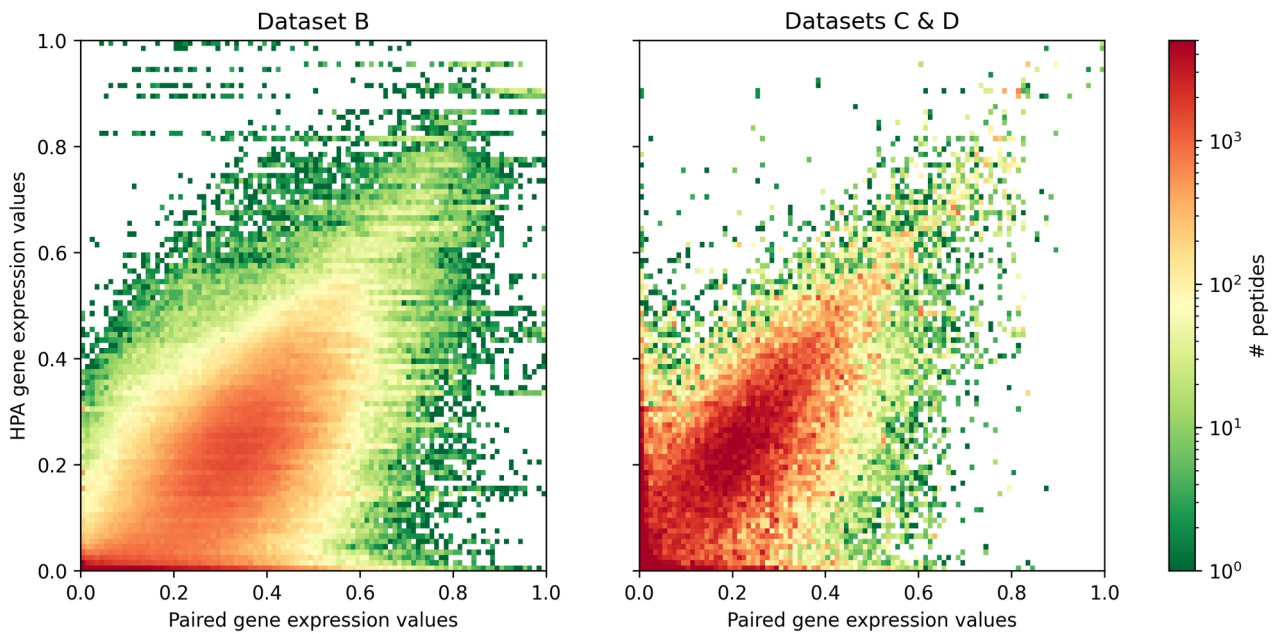


Figure S5. Joint distribution of matched and HPA gene expression values for training set peptides originating from datasets B, C and D, related to Fig. 2. The color intensity in the heatmaps represents the amount of peptides for each (x,y) pair of gene expression values. The overall Spearman Correlation Coefficient (SCC) between gene expression values for Dataset B is 0.688, while for Datasets C & D is 0.727. For all datasets, the gene expression values are normalized as specified in the Materials and Methods section of the manuscript (these are the final values used for the neural network training).

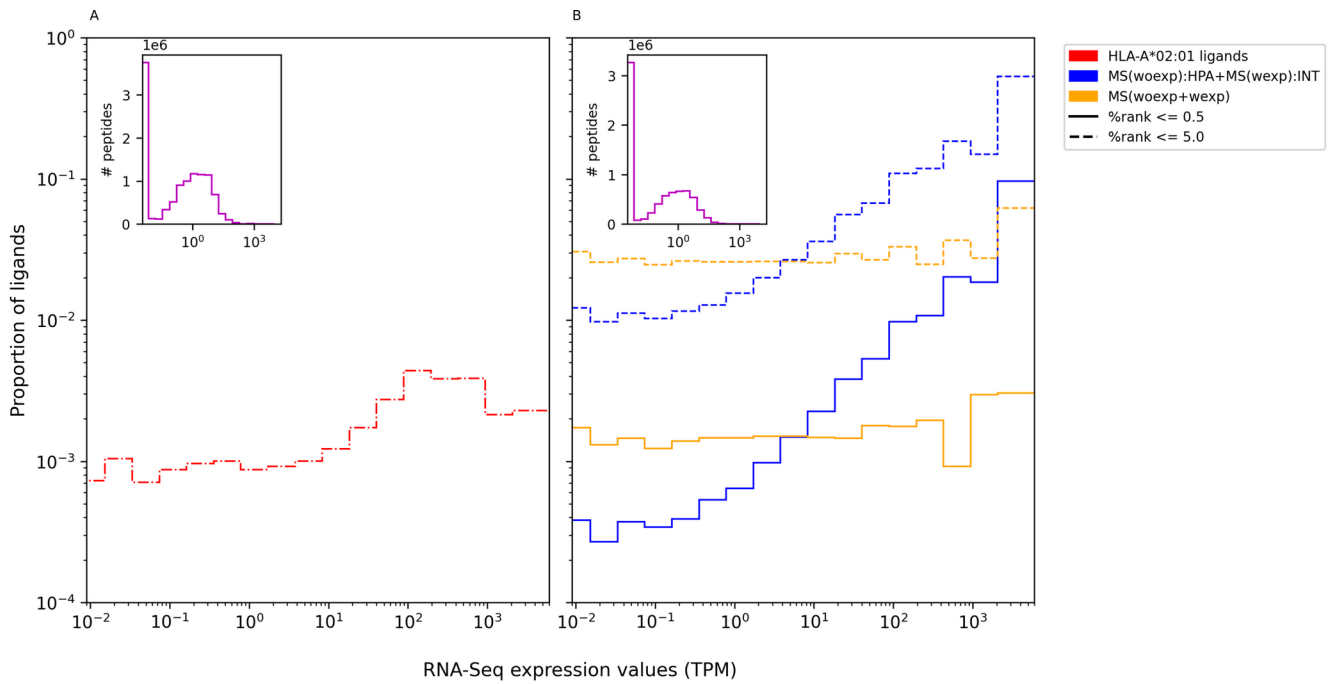


Figure S6. Distribution of the proportion of (A) observed and (B) predicted HLA-A*02:01 ligands across different TPM values by the selected method MS(woexp):HPA+MS(wexp):INT and its counterpart without gene expression values, MS(woexp+wexp), related to Fig. 3. For panel (A), training set HLA-A*02:01 ligands (positives), 8 to 11 amino acids long, were mapped back to their source proteins. For panel (B), 5000 proteins were randomly selected from the human proteome, then digested into 8 to 11-mers and finally predicted with the already mentioned models. Ligands were defined taking into account two threshold rank values: 0.5% (full line) and 5% (dashed line). The number of observed or predicted ligands in each TPM bin was normalized by the total number of background peptides falling into that same bin (shown in inset, magenta curve). Both for (A) and (B) proteins were given gene expression values according to the HPA database. The y-axis in both insets is expressed in scientific notation.

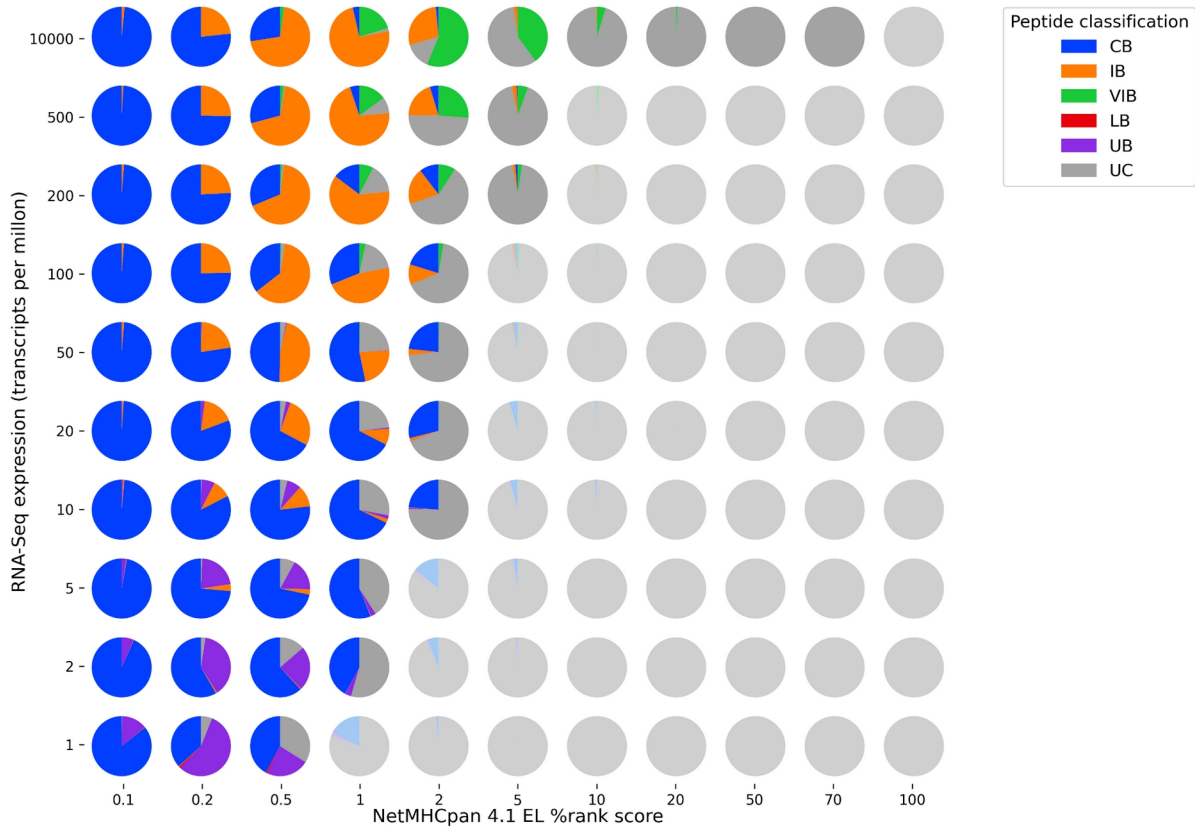


Figure S7. Comparison of the 5-fold cross-validation predictions of a model trained with and without gene expression values discriminated by the MS ligands predicted HLA binding score and their gene expression values, related to Fig. 1 and 3. MS ligands from datasets A-D were classified according to the predictions obtained by model MS(woexp):HPA+MS(wexp):INT and its counterpart trained without gene expression MS(woexp+wexp). As in Fig. 1, NetMHCpan-4.1 EL percentile rank scores and RNA-Seq expression values were binned to generate a 2-dimensional array where EL percentile rank scores are shown on the x-axis and TPM values on the y-axis. The pie charts in each cell of this grid show the distribution of ligands in each of the defined groups. The pie charts in bright colors are located to the left of the frontier of equivalence (refer to Fig. 1). CB: conserved binder, IB: improved binder, VIB: very improved binder, LB: lost binder, UB: unimproved binder and UC: unclassified peptides.

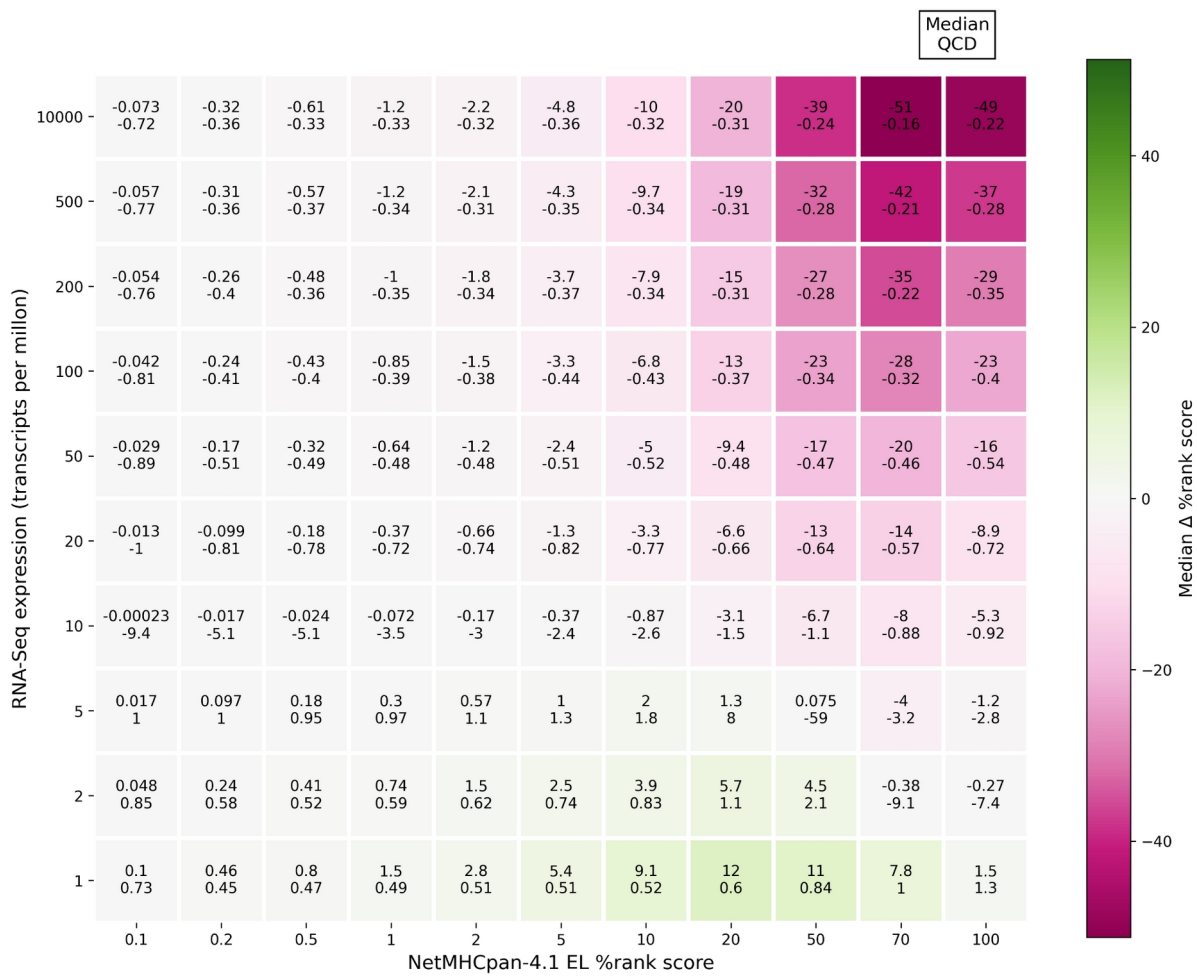


Figure S8. Difference in the EL %rank score predictions of a model trained with and without gene expression values, related to Fig. 1 and 3. As in Fig. 1, NetMHCpan-4.1 EL percentile rank scores and RNA-Seq expression values were binned to generate a 2-dimensional array where EL percentile rank scores are shown on the x-axis and TPM values on the y-axis. For all MS ligands (sources A-D), the delta EL %rank score prediction was calculated, which is defined as the prediction value obtained with model MS(woexp):HPA+MS(wexp):INT minus the prediction value obtained with model MS(woexp+wexp). Each cell of the array displays the median delta EL %rank score (top) and its QCD, or Quartile Coefficient of Dispersion (bottom), for all its corresponding MS ligands. The QCD is computed as $(Q_3 - Q_1)/(Q_1 + Q_3)$, where Q_1 and Q_3 stand for the first and third quartiles of a dataset, respectively.

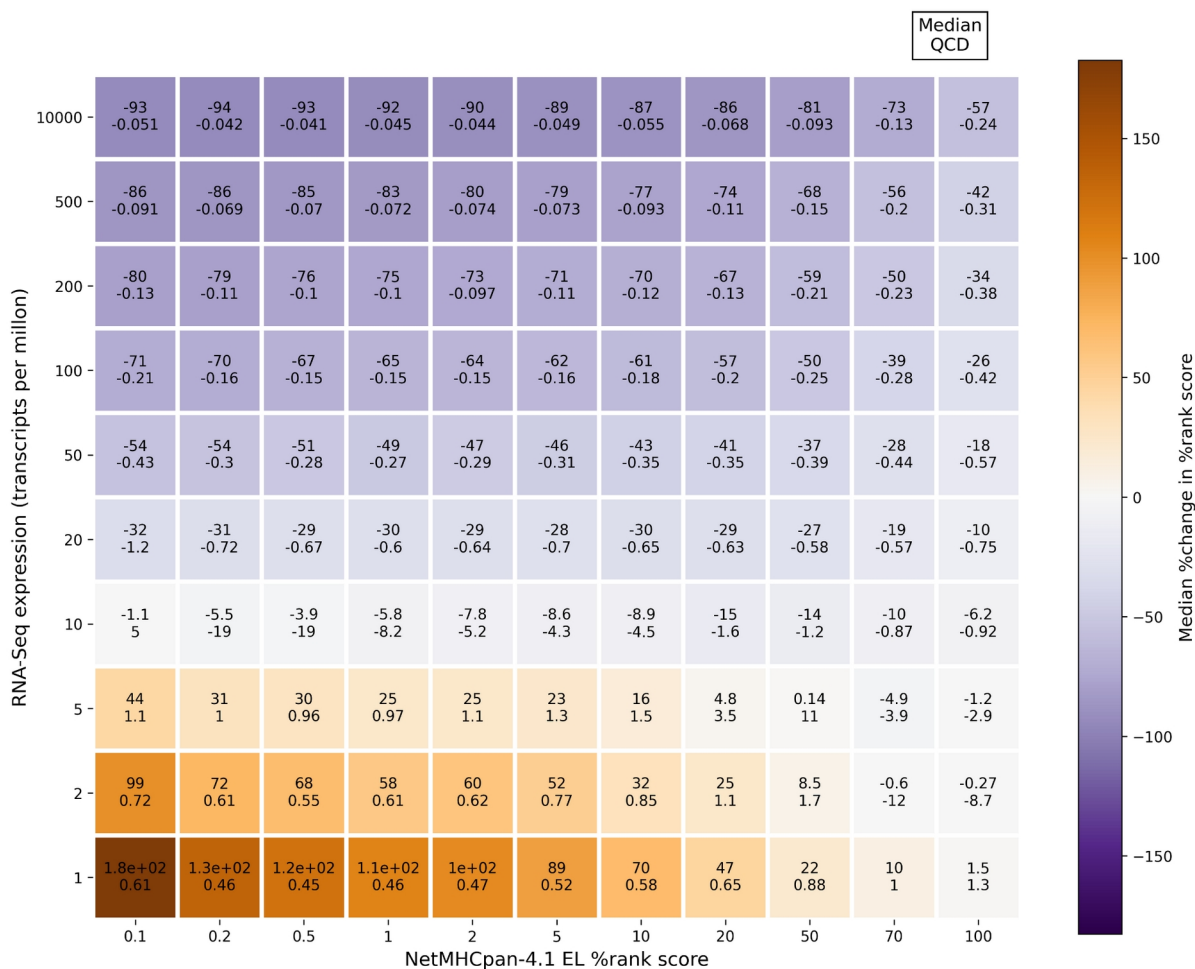


Figure S9. Percentage change in the EL %rank score predictions of a model trained with and without gene expression values, related to Fig. 1 and 3. As in Fig. 1, NetMHCpan-4.1 EL percentile rank scores and RNA-Seq expression values were binned to generate a 2-dimensional array where EL percentile rank scores are shown on the x-axis and TPM values on the y-axis. For all MS ligands (datasets A-D), the percentage change in the EL %rank score prediction was calculated, which is defined as the delta EL %rank score (MS(woexp):HPA+MS(wexp):INT EL %rank score - MS(woexp+wexp) EL %rank score, as in SFig. 7) divided by the EL %rank score obtained with the model MS(woexp+wexp) multiplied by 100. Each cell of the array displays the median percentage change in the EL %rank score (top) and its QCD, or Quartile Coefficient of Dispersion (bottom), for all its corresponding MS ligands. The QCD is computed as $(Q_3 - Q_1)/(Q_1 + Q_3)$, where Q_1 and Q_3 stand for the first and third quartiles of a dataset, respectively.

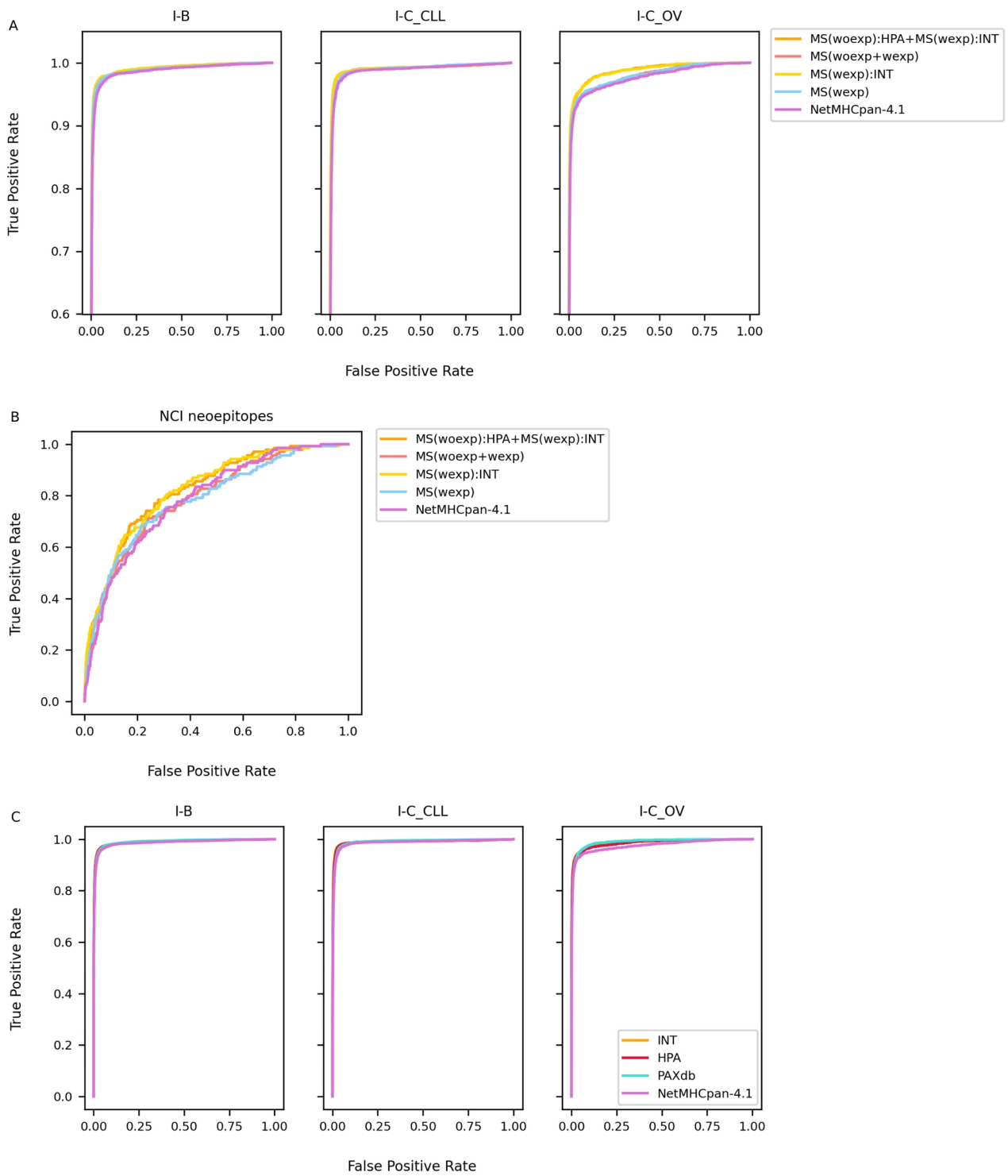


Figure S10. AUC-ROC curves corresponding to the performance of the methods evaluated on the different datasets of the independent benchmark shown in Fig. 4. Panel A corresponds to the results displayed on Figure 4A, and the same applies for the results in panel B and C.

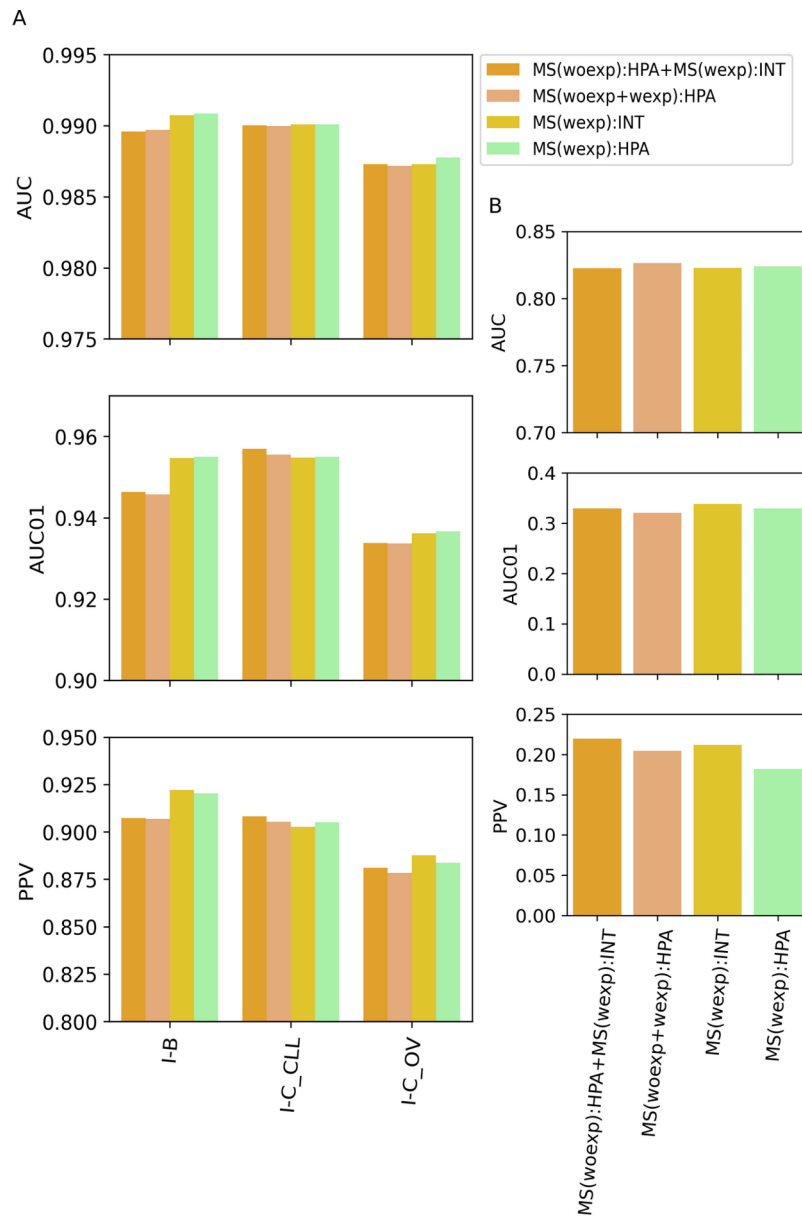
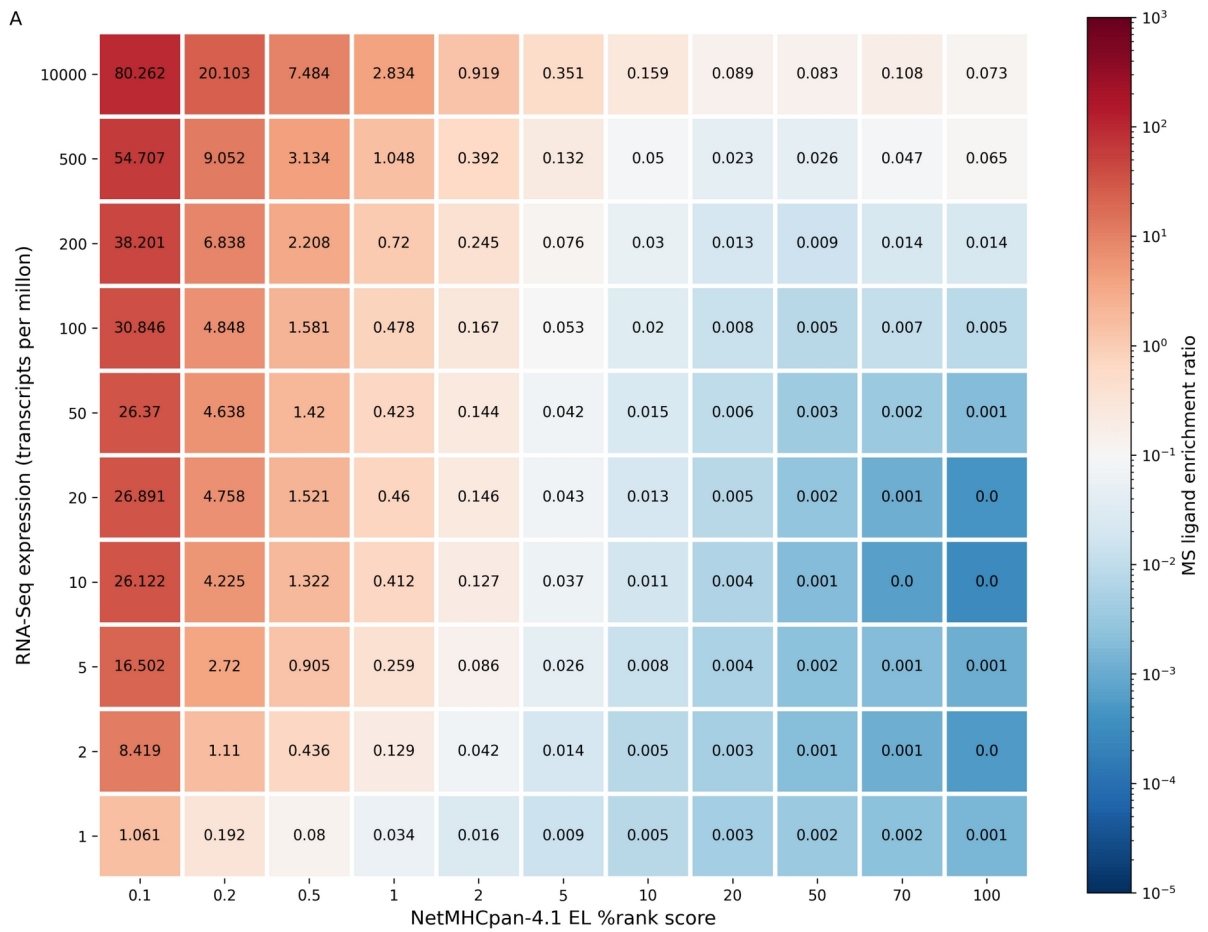


Figure S11. Independent evaluation of the models employing internal (“INT”) and external (HPA) references to assign gene expression values to the peptides on their training sets, related to Fig. 4. (A) illustrates the performance of the trained models on the independent datasets of MS eluted ligands (I-B and I-C) and (B) shows the performance of the models on the I-NCI neoepitope dataset. On both datasets, models MS(woexp):HPA+MS(wexp):INT and MS(wexp):INT perform on par with their equivalents that employ the HPA gene expression reference, MS(woexp+wexp):HPA and MS(wexp):HPA, respectively (p-values not shown).

A



B

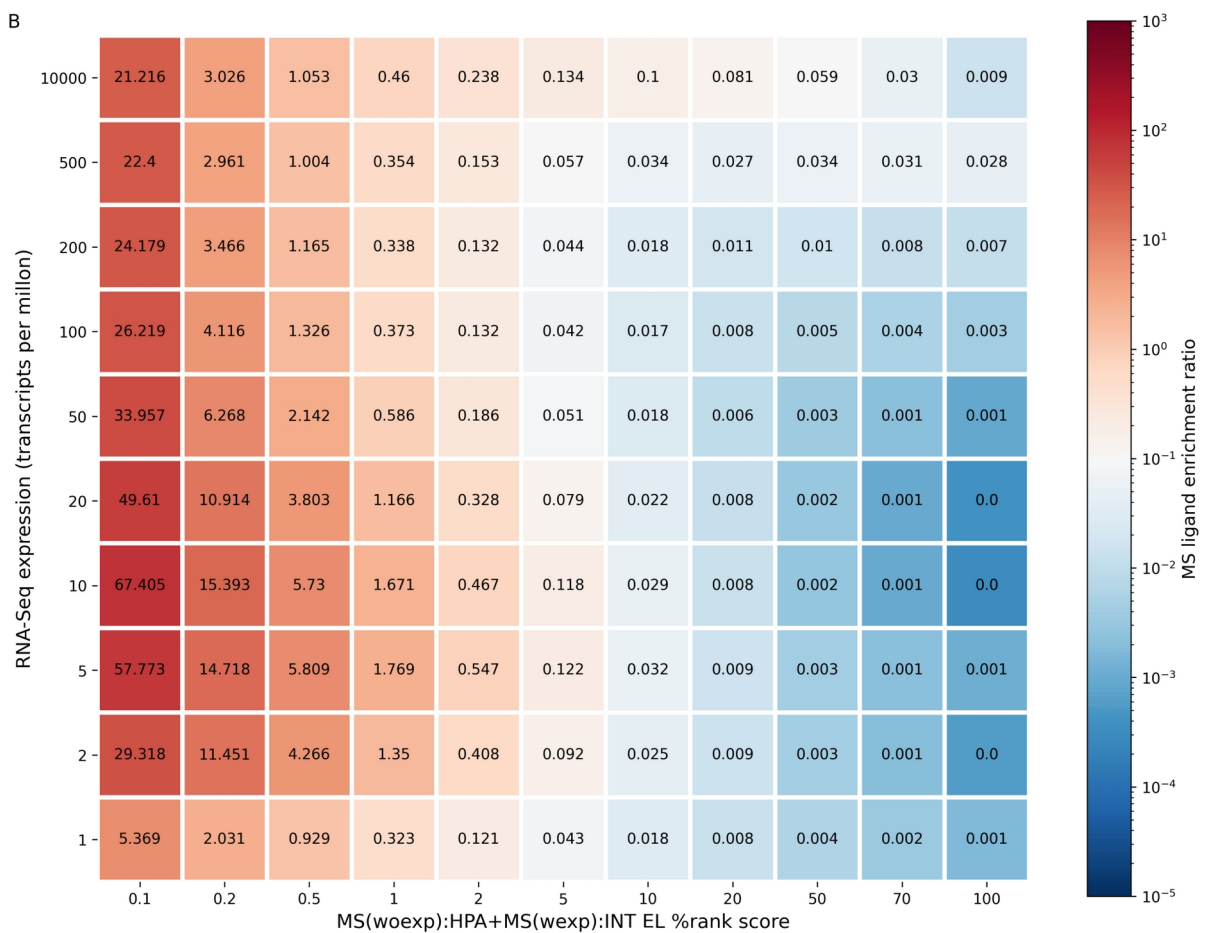


Figure S12. Relationship between predicted HLA binding scores of MS eluted ligands (and artificially generated random negatives) and the gene expression values of their corresponding source proteins, related to Fig. 1. EL percentile rank scores and RNA-Seq expression values were binned to generate a 2-dimensional array where EL percentile rank and TPM values on the y-axis. All compiled MS datasets (datasets A-D) were used to construct this array. The numbers on both the x and y-axis represent the rightmost edge of each bin, for instance, the cell on the upper right corner contains peptides in the range (70,100] of EL %rank scores and (500, 10000] of TPM values. As an exception, the cell on the lower left corner contains peptides in the interval [0,0.1] of EL %rank scores and [0,1] of TPM values. Each cell displays the ratio between the number of MS ligands and the number of random natural negative peptides that fall into it, and it is colored according to this magnitude, referred to as the “MS ligand enrichment ratio”. The midpoint of the color scale was set to coincide with the ratio of total MS ligands to total background peptides (white cells). **(A)** Equivalent to Figure 1 in the manuscript, displaying NetMHCpan-4.1 predictions on the x-axis. **(B)** Similar to A, but displaying NetMHCpanExp-1.0 predictions on the x-axis.

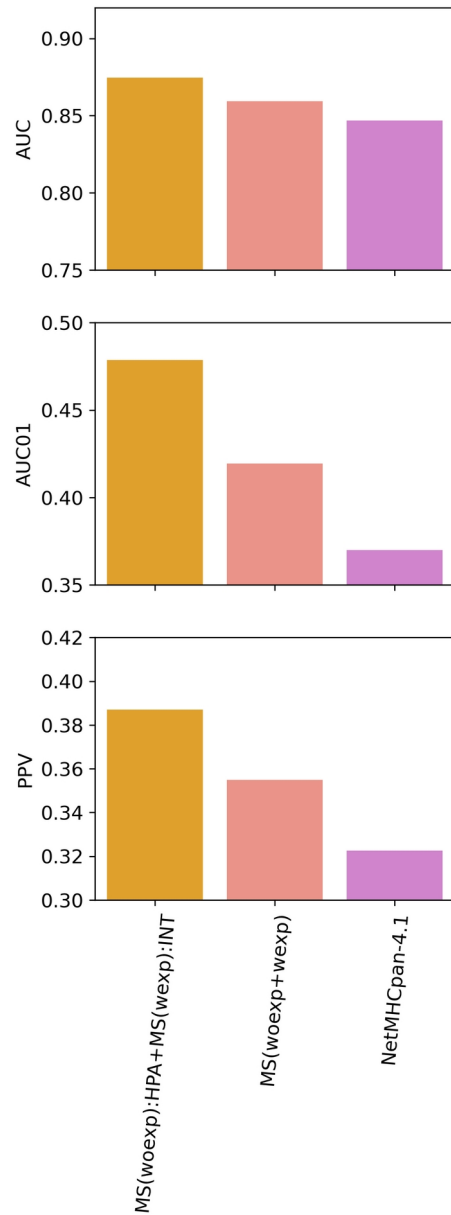


Figure S13. Performance of the trained models and NetMHCpan-4.1 on the TESLA neoepitope dataset, expanding on the independent benchmark shown in Fig. 4. The plot shows the performance of NetMHCpanExp (or MS(woexp):HPA+MS(wexp):INT), its equivalent trained without gene expression data (MS(woexp+wexp)) and NetMHCpan-4.1 in terms of the AUC, AUC01 and PPV (p-values not shown).

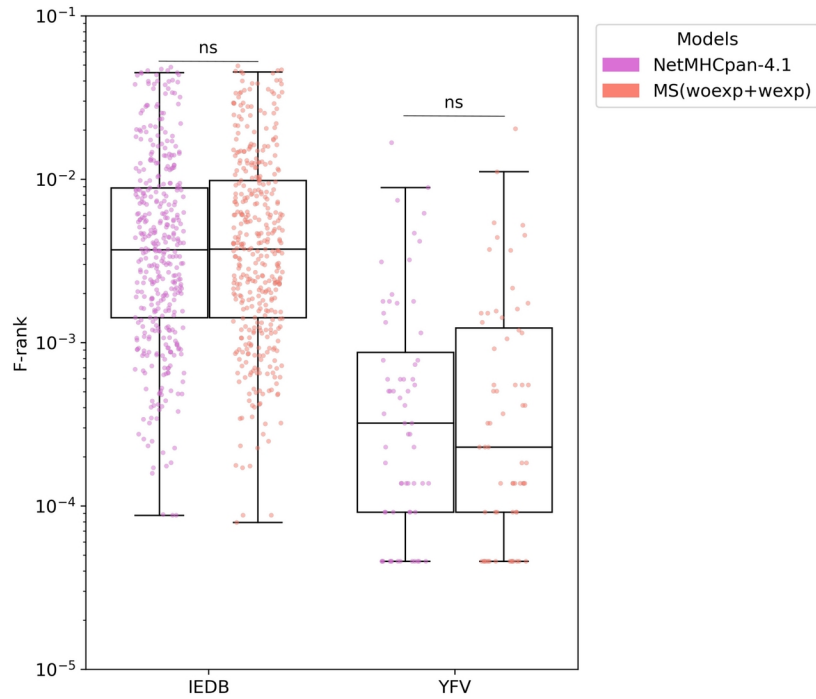


Figure S14. Comparison of the predictive power of NetMHCpan-4.1 and MS(woexp+wexp), the newly developed method NetMHCpanExp without gene expression values, on two sets of CD8+ epitopes, the first one extracted from the IEDB and the second one from the Yellow Fever Virus (YFV), expanding on the independent benchmark shown in Fig. 4. For each of the epitopes in both sets an F-rank value was calculated (scatter plots). The lower the F-rank (closer to 0) the better the prediction score for a given epitope. CD8+ epitopes with an F-rank > 0.05 for any of the two evaluated methods were excluded from the present analysis. The IEDB dataset consists of 393 epitopes (36 positive instances were excluded due to the mentioned F-rank threshold) and the YFV dataset consists of 64 epitopes. The boxplots illustrate the distribution of F-ranks. The center line inside the box indicates the median value of the plotted metric. The box covers the interquartile range. The whiskers represent 1.5-fold of the interquartile range. The difference in F-rank values between the studied methods for the two benchmarks was calculated using a two-tailed Binomial test. P-values are shown in Table S7.

Supplementary Tables

Table S1

Dataset	Metric	Model 1	Model 2	P-value
A	AUC	MS(woexp+wexp):HPA	MS(wexp)	4.195E-29
A	AUC	MS(wexp):INT	MS(wexp)	5.172E-27
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	4.603E-25
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	6.953E-24
A	AUC	MS(wexp):HPA	MS(wexp)	2.275E-22
A	AUC	MS(woexp+wexp):HPA	MS(woexp+wexp)	4.006E-22
A	AUC	MS(woexp+wexp)	MS(wexp):INT	2.197E-12
A	AUC	MS(woexp+wexp)	MS(wexp):HPA	2.197E-12
A	AUC	MS(woexp+wexp):HPA	MS(wexp):HPA	1.737E-09
A	AUC	MS(woexp+wexp)	MS(wexp)	1.409E-08
A	AUC	MS(woexp+wexp):HPA	MS(wexp):INT	5.564E-08
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1.446E-07
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):HPA	2.543E-07
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):HPA	2.852E-01
A	AUC	MS(wexp):INT	MS(wexp):HPA	6.812E-01
B+C+D	AUC	MS(woexp+wexp)	MS(wexp):HPA	4.598E-44
B+C+D	AUC	MS(woexp+wexp)	MS(wexp):INT	2.982E-42
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	8.205E-41
B+C+D	AUC	MS(wexp):HPA	MS(wexp)	1.656E-34
B+C+D	AUC	MS(woexp+wexp):HPA	MS(woexp+wexp)	2.672E-32
B+C+D	AUC	MS(wexp):INT	MS(wexp)	3.124E-30
B+C+D	AUC	MS(woexp+wexp):HPA	MS(wexp):HPA	1.419E-25
B+C+D	AUC	MS(woexp+wexp):HPA	MS(wexp)	2.450E-19
B+C+D	AUC	MS(woexp+wexp)	MS(wexp)	1.204E-18
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	5.677E-18
B+C+D	AUC	MS(woexp+wexp):HPA	MS(wexp):INT	1.022E-13
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	6.641E-10
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):HPA	6.272E-07
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):HPA	2.872E-04
B+C+D	AUC	MS(wexp):INT	MS(wexp):HPA	6.455E-02
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	4.865E-28
A	AUC01	MS(wexp):INT	MS(wexp)	4.865E-28
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	8.953E-28
A	AUC01	MS(woexp+wexp):HPA	MS(woexp+wexp)	9.448E-27
A	AUC01	MS(woexp+wexp):HPA	MS(wexp)	5.068E-26
A	AUC01	MS(wexp):HPA	MS(wexp)	4.603E-25
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	2.749E-16
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):HPA	9.106E-15
A	AUC01	MS(woexp+wexp)	MS(wexp)	1.535E-13
A	AUC01	MS(woexp+wexp):HPA	MS(wexp):INT	5.799E-11
A	AUC01	MS(woexp+wexp):HPA	MS(wexp):HPA	5.799E-11
A	AUC01	MS(woexp+wexp)	MS(wexp):INT	5.794E-10
A	AUC01	MS(woexp+wexp)	MS(wexp):HPA	8.747E-10
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):HPA	1.269E-03
A	AUC01	MS(wexp):INT	MS(wexp):HPA	8.397E-02

Table S1. P-values corresponding to the statistical analysis performed on the results shown in Fig. 2. P-values were computed by applying a two-tailed Binomial test to compare model performances in Figure 2.

Table S1 (continued)

Dataset	Metric	Model 1	Model 2	P-value
B+C+D	AUC01	MS(woexp+wexp)	MS(wexp):INT	2.673E-51
B+C+D	AUC01	MS(woexp+wexp)	MS(wexp):HPA	2.673E-51
B+C+D	AUC01	MS(woexp+wexp):HPA	MS(woexp+wexp)	4.544E-49
B+C+D	AUC01	MS(wexp):HPA	MS(wexp)	4.544E-49
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	8.980E-44
B+C+D	AUC01	MS(wexp):INT	MS(wexp)	8.980E-44
B+C+D	AUC01	MS(woexp+wexp):HPA	MS(wexp):INT	3.124E-30
B+C+D	AUC01	MS(woexp+wexp):HPA	MS(wexp):HPA	2.749E-28
B+C+D	AUC01	MS(woexp+wexp):HPA	MS(wexp)	6.893E-24
B+C+D	AUC01	MS(woexp+wexp)	MS(wexp)	2.730E-22
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	1.598E-21
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1.464E-19
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):HPA	3.984E-11
B+C+D	AUC01	MS(wexp):INT	MS(wexp):HPA	3.345E-03
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):HPA	1.717E-02
A	PPV	MS(wexp):HPA	MS(wexp)	8.953E-28
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	5.172E-27
A	PPV	MS(wexp):INT	MS(wexp)	5.172E-27
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	3.891E-24
A	PPV	MS(woexp+wexp):HPA	MS(wexp)	3.891E-24
A	PPV	MS(woexp+wexp):HPA	MS(woexp+wexp)	6.953E-24
A	PPV	MS(woexp+wexp)	MS(wexp)	9.498E-17
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	2.364E-14
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):HPA	3.730E-13
A	PPV	MS(woexp+wexp):HPA	MS(wexp):INT	5.924E-13
A	PPV	MS(woexp+wexp):HPA	MS(wexp):HPA	7.839E-12
A	PPV	MS(woexp+wexp)	MS(wexp):INT	5.564E-08
A	PPV	MS(woexp+wexp)	MS(wexp):HPA	1.446E-07
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):HPA	7.744E-06
A	PPV	MS(wexp):INT	MS(wexp):HPA	1.621E-01
B+C+D	PPV	MS(woexp+wexp)	MS(wexp):INT	2.673E-51
B+C+D	PPV	MS(woexp+wexp)	MS(wexp):HPA	2.285E-49
B+C+D	PPV	MS(woexp+wexp):HPA	MS(woexp+wexp)	4.544E-49
B+C+D	PPV	MS(wexp):HPA	MS(wexp)	3.840E-47
B+C+D	PPV	MS(wexp):INT	MS(wexp)	3.920E-38
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	3.727E-37
B+C+D	PPV	MS(woexp+wexp):HPA	MS(wexp):INT	3.000E-31
B+C+D	PPV	MS(woexp+wexp):HPA	MS(wexp):HPA	6.893E-24
B+C+D	PPV	MS(woexp+wexp)	MS(wexp)	3.440E-18
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1.571E-17
B+C+D	PPV	MS(woexp+wexp):HPA	MS(wexp)	1.908E-15
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	7.922E-13
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):HPA	3.253E-08
B+C+D	PPV	MS(wexp):INT	MS(wexp):HPA	4.186E-05
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):HPA	8.181E-01

Table S1. P-values corresponding to the statistical analysis performed on the results shown in Fig. 2. P-values were computed by applying a two-tailed Binomial test to compare model performances in Figure 2.

Table S2

Model	Gene expression values		Color in Fig. S4
	Dataset A	Datasets B-D	
MS(woexp):HPA+MS(wexp):INT	HPA	Internal	orange
MS(woexp+wexp):INT2HPA	HPA	Internal2HPA	magenta
MS(wexp):INT	-	Internal	yellow
MS(wexp):INT2HPA	-	Internal2HPA	cyan

Table S2. Models trained on 5-fold cross-validation shown in Figure S4 (related to Fig. 2). Model nomenclature is related to the subset of the data used for training and its associated gene expression values.

Table S3

Dataset	Metric	Model 1	Model 2	P-value
A	AUC	MS(woexp+wexp):INT2HPA	MS(wexp):INT	5.799E-11
A	AUC	MS(woexp+wexp):INT2HPA	MS(wexp):INT2HPA	8.747E-10
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT2HPA	5.032E-09
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1.446E-07
A	AUC	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):INT2HPA	1.401E-01
A	AUC	MS(wexp):INT	MS(wexp):INT2HPA	1.898E-01
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT2HPA	2.450E-19
B+C+D	AUC	MS(woexp+wexp):INT2HPA	MS(wexp):INT2HPA	1.225E-12
B+C+D	AUC	MS(woexp+wexp):INT2HPA	MS(wexp):INT	1.195E-10
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	6.641E-10
B+C+D	AUC	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):INT2HPA	3.095E-02
B+C+D	AUC	MS(wexp):INT	MS(wexp):INT2HPA	4.901E-01
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT2HPA	9.498E-17
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	2.749E-16
A	AUC01	MS(woexp+wexp):INT2HPA	MS(wexp):INT	2.749E-16
A	AUC01	MS(woexp+wexp):INT2HPA	MS(wexp):INT2HPA	3.730E-13
A	AUC01	MS(wexp):INT	MS(wexp):INT2HPA	3.659E-01
A	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):INT2HPA	8.699E-01
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT2HPA	8.939E-21
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1.464E-19
B+C+D	AUC01	MS(woexp+wexp):INT2HPA	MS(wexp):INT2HPA	1.908E-15
B+C+D	AUC01	MS(woexp+wexp):INT2HPA	MS(wexp):INT	2.807E-14
B+C+D	AUC01	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):INT2HPA	4.403E-03
B+C+D	AUC01	MS(wexp):INT	MS(wexp):INT2HPA	2.815E-01
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT2HPA	5.593E-15
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	2.364E-14
A	PPV	MS(woexp+wexp):INT2HPA	MS(wexp):INT	3.818E-14
A	PPV	MS(woexp+wexp):INT2HPA	MS(wexp):INT2HPA	3.730E-13
A	PPV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):INT2HPA	8.397E-02
A	PPV	MS(wexp):INT	MS(wexp):INT2HPA	6.232E-01
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT2HPA	4.446E-23
B+C+D	PPV	MS(woexp+wexp):INT2HPA	MS(wexp):INT2HPA	9.356E-22
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1.571E-17
B+C+D	PPV	MS(woexp+wexp):INT2HPA	MS(wexp):INT	2.918E-16
B+C+D	PPV	MS(wexp):INT	MS(wexp):INT2HPA	4.518E-02
B+C+D	PPV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp):INT2HPA	1.660E-01

Table S3. P-values corresponding to the statistical analysis performed on the results shown in Fig. S4 (related to Fig. 2). P-values were computed by applying a two-tailed Binomial test to compare model performances in Figure S4.

Table S4

Figure 4A				
Dataset	Model	Metric		
		AUC	AUC01	PPV
I-B	MS(woexp):HPA+MS(wexp):INT	0.9896	0.9463	0.9072
I-B	MS(woexp+wexp)	0.9867	0.9342	0.8898
I-B	MS(wexp):INT	0.9907	0.9547	0.9220
I-B	MS(wexp)	0.9885	0.9455	0.9062
I-B	NetMHCpan-4.1	0.9860	0.9297	0.8850
I-C_CLL	MS(woexp):HPA+MS(wexp):INT	0.9900	0.9569	0.9081
I-C_CLL	MS(woexp+wexp)	0.9876	0.9349	0.8724
I-C_CLL	MS(wexp):INT	0.9901	0.9548	0.9026
I-C_CLL	MS(wexp)	0.9879	0.9336	0.8697
I-C_CLL	NetMHCpan-4.1	0.9865	0.9282	0.8640
I-C_OV	MS(woexp):HPA+MS(wexp):INT	0.9873	0.9338	0.8811
I-C_OV	MS(woexp+wexp)	0.9788	0.9155	0.8456
I-C_OV	MS(wexp):INT	0.9873	0.9361	0.8878
I-C_OV	MS(wexp)	0.9792	0.9178	0.8545
I-C_OV	NetMHCpan-4.1	0.9755	0.9071	0.8310

Figure 4B				
Dataset	Model	Metric		
		AUC	AUC01	PPV
NCI	MS(woexp):HPA+MS(wexp):INT	0.8227	0.3294	0.2197
NCI	MS(woexp+wexp)	0.7898	0.2949	0.1212
NCI	MS(wexp):INT	0.8230	0.3379	0.2121
NCI	MS(wexp)	0.7866	0.3112	0.1515
NCI	NetMHCpan-4.1	0.7891	0.2721	0.0985

Figure 4C				
Dataset	Gene expr.	Metric		
		AUC	AUC01	PPV
I-B	INT	0.9896	0.9463	0.9072
I-B	HPA	0.9895	0.9456	0.9017
I-B	PAXdb	0.9900	0.9390	0.8927
I-B	NetMHCpan-4.1	0.9860	0.9297	0.8850
I-C_CLL	INT	0.9900	0.9569	0.9081
I-C_CLL	HPA	0.9898	0.9525	0.8995
I-C_CLL	PAXdb	0.9903	0.9381	0.8739
I-C_CLL	NetMHCpan-4.1	0.9865	0.9282	0.8640
I-C_OV	INT	0.9873	0.9338	0.8811
I-C_OV	HPA	0.9862	0.9296	0.8660
I-C_OV	PAXdb	0.9892	0.9246	0.8310
I-C_OV	NetMHCpan-4.1	0.9755	0.9071	0.8310

Table S4. Resultant metrics displayed as barplots in Figure 4.

Table S5

Figure 4A					
Dataset	Model 1	Model 2	P-value		
			AUC	AUC01	PPV
I-B	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	0	0	0
I-B	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	1	1	1
I-B	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	0.001	0.197	0.317
I-B	MS(woexp):HPA+MS(wexp):INT	NetMHCpan-4.1	0	0	0
I-B	MS(woexp+wexp)	MS(wexp):INT	1	1	1
I-B	MS(woexp+wexp)	MS(wexp)	1	1	1
I-B	MS(woexp+wexp)	NetMHCpan-4.1	0	0	0.01
I-B	MS(wexp):INT	MS(wexp)	0	0	0
I-B	MS(wexp):INT	NetMHCpan-4.1	0	0	0
I-B	MS(wexp)	NetMHCpan-4.1	0	0	0
I-C_OV	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	0	0	0
I-C_OV	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	0.505	0.999	0.998
I-C_OV	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	0	0	0
I-C_OV	MS(woexp):HPA+MS(wexp):INT	NetMHCpan-4.1	0	0	0
I-C_OV	MS(woexp+wexp)	MS(wexp):INT	1	1	1
I-C_OV	MS(woexp+wexp)	MS(wexp)	0.948	1	0.994
I-C_OV	MS(woexp+wexp)	NetMHCpan-4.1	0	0	0
I-C_OV	MS(wexp):INT	MS(wexp)	0	0	0
I-C_OV	MS(wexp):INT	NetMHCpan-4.1	0	0	0
I-C_OV	MS(wexp)	NetMHCpan-4.1	0	0	0
I-C_CLL	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	0	0	0
I-C_CLL	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	0.68	0	0.007
I-C_CLL	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	0	0	0
I-C_CLL	MS(woexp):HPA+MS(wexp):INT	NetMHCpan-4.1	0	0	0
I-C_CLL	MS(woexp+wexp)	MS(wexp):INT	1	1	1
I-C_CLL	MS(woexp+wexp)	MS(wexp)	0.993	0.024	0.121
I-C_CLL	MS(woexp+wexp)	NetMHCpan-4.1	0	0	0
I-C_CLL	MS(wexp):INT	MS(wexp)	0	0	0
I-C_CLL	MS(wexp):INT	NetMHCpan-4.1	0	0	0
I-C_CLL	MS(wexp)	NetMHCpan-4.1	0	0	0.032

Table S5. P-values corresponding to the statistical analysis performed on the results shown in Fig. 4. P-values were computed by applying a one-tailed Binomial test (bootstrap method) to compare performances in Figure 4. For more details regarding this statistical test, please refer to the Materials and Methods section of the manuscript.

Table S5 (continued)

Figure 4B					
Dataset	Model 1	Model 2	P-value		
			AUC	AUC01	PPV
I-NCI	MS(woexp):HPA+MS(wexp):INT	MS(woexp+wexp)	0.007	0.147	0.013
I-NCI	MS(woexp):HPA+MS(wexp):INT	MS(wexp):INT	0.533	0.718	0.668
I-NCI	MS(woexp):HPA+MS(wexp):INT	MS(wexp)	0.008	0.275	0.069
I-NCI	MS(woexp):HPA+MS(wexp):INT	NetMHCpan-4.1	0.003	0.063	0.007
I-NCI	MS(woexp+wexp)	MS(wexp):INT	0.984	0.879	0.989
I-NCI	MS(woexp+wexp)	MS(wexp)	0.368	0.842	0.919
I-NCI	MS(woexp+wexp)	NetMHCpan-4.1	0.456	0.11	0.208
I-NCI	MS(wexp):INT	MS(wexp)	0.002	0.201	0.076
I-NCI	MS(wexp):INT	NetMHCpan-4.1	0.013	0.034	0.003
I-NCI	MS(wexp)	NetMHCpan-4.1	0.581	0.029	0.03

Figure 4C (*)					
Dataset	Gene expr. Ref. 1	Gene expr. Ref. 2	P-value		
			AUC	AUC01	PPV
I-B	INT	HPA	0.306	0.194	0.005
I-B	INT	PAXdb	0.906	0	0
I-B	INT	NetMHCpan-4.1	0	0	0
I-B	HPA	PAXdb	0.973	0	0
I-B	HPA	NetMHCpan-4.1	0	0	0
I-B	PAXdb	NetMHCpan-4.1	0	0	0.005
I-C_OV	INT	HPA	0.002	0	0
I-C_OV	INT	PAXdb	0.998	0	0
I-C_OV	INT	NetMHCpan-4.1	0	0	0
I-C_OV	HPA	PAXdb	1	0.006	0
I-C_OV	HPA	NetMHCpan-4.1	0	0	0
I-C_OV	PAXdb	NetMHCpan-4.1	0	0	0.383
I-C_CLL	INT	HPA	0.013	0	0
I-C_CLL	INT	PAXdb	0.689	0	0
I-C_CLL	INT	NetMHCpan-4.1	0	0	0
I-C_CLL	HPA	PAXdb	0.867	0	0
I-C_CLL	HPA	NetMHCpan-4.1	0	0	0
I-C_CLL	PAXdb	NetMHCpan-4.1	0	0	0.003

Table S5. P-values corresponding to the statistical analysis performed on the results shown in Fig. 4.

P-values were computed by applying a one-tailed Binomial test (bootstrap method) to compare performances in Figure 4. For more details regarding this statistical test, please refer to the Materials and Methods section of the manuscript. (*) For Fig. 4C, results refer to the performance of model MS(wexp):HPA+MS(woexp):INT (and NetMHCpan-4.1) on 3 datasets annotated with different gene expression references. Recall that NetMHCpan-4.1 does not accept gene expression values. Thus, the predictions of this model are taken as a baseline.

Table S6

Figure 4A					
Dataset	MS(woexp):HPA		MS(wexp):INT	MS(wexp)	NetMHCpan-4.1
	+MS(wexp):INT	MS(woexp+wexp)			
I-B	0.906	0.879	0.93	0.904	0.865
I-C_CLL	0.922	0.869	0.919	0.865	0.849
I-C_OV	0.895	0.876	0.906	0.879	0.858
Figure 4B					
Dataset	MS(woexp):HPA		MS(wexp):INT	MS(wexp)	NetMHCpan-4.1
	+MS(wexp):INT	MS(woexp+wexp)			
I-NCI	0.19	0.11	0.2	0.13	0.08
Figure 4C					
Dataset	INT	HPA	PAXdb	NetMHCpan-4.1	
I-B	0.906	0.901	0.878	0.865	
I-C_CLL	0.922	0.909	0.866	0.849	
I-C_OV	0.895	0.89	0.854	0.858	

Table S6. Sensitivities (or TPRs) of the different methods at a FPR value of 1%, related to Fig. 4. The FPR values were extracted from the AUC-ROC curves corresponding to the evaluation of the different methods on the external datasets shown in Figure 4 and Figure S10.

Table S7

Benchmark	Model 1	Model 2	P-value
CD8+_epitopes_IEDB	NetMHCpan-4.1	MS(woexp+wexp)	5.83E-02
CD8+_epitopes_YFV	NetMHCpan-4.1	MS(woexp+wexp)	1.00E+00

Table S7. P-values corresponding to the statistical analysis performed on the results shown in Fig. S14 (expanding on the results shown in Fig. 4). P-values were computed by applying a two-tailed Binomial test to compare model performances in Figure S14.

Table S8

Cancer type	Sample ID	Clinical ID	HLA-A	HLA-A	HLA-B	HLA-B	HLA-C	HLA-C
CLL	CLL A	DFCI-5341	A0301	A3101(*)	B1402	B3502	C0401	C0802
	CLL B	DFCI-5328	A0206	A2402	B0801	B51	C0702	C1402
	CLL C	DFCI-5283	A0101	A0201	B0702	B0801	C0701	C0702
Ovarian (OV)	OV1	CP-594_v1	A0201	A2402	B3503	B4402	C0501	C1203

Table S8. Clinical IDs associated with the independent datasets I-C_OV and I-C_CLL, extracted from Sarkizova *et al.*, 2020. Related to Fig. 4. (*) As this HLA-A allele was originally annotated as “A310102”, which we suspect that corresponds to an ambiguous determination (“A3101/02”), we only computed predictions for HLA-A3101 in this case.