

Immunity

Machine learning methods and harmonized datasets improve immunogenic neoantigen prediction

Highlights

- ML models improve immunogenic neo-peptide ranking by up to 30% over published methods
- Presentation and T cell recognition features enhance immunogenic neo-peptide ranking
- Provision of ML classifiers enhances prioritization of both neo-peptides and mutations

Authors

Markus Müller, Florian Huber, Marion Arnaud, ..., George Coukos, Alexandre Harari, Michal Bassani-Sternberg

Correspondence

markus.muller@chuv.ch (M.M.),
michal.bassani@chuv.ch (M.B.-S.)

In brief

Müller and colleagues showcase the enhanced capability of machine learning classifiers, which were trained using consistently processed multi-cancer genomic, transcriptomic, and neoantigen immunogenicity data. Their approach improves the prioritization of immunogenic neoantigens by incorporating additional features that complement factors related to antigen presentation and expression.



Resource

Machine learning methods and harmonized datasets improve immunogenic neoantigen prediction

Markus Müller,^{1,2,3,4,*} Florian Huber,^{1,2,3} Marion Arnaud,^{1,2,3} Anne I. Kraemer,^{1,2,3} Emma Ricart Altimiras,^{1,2,3} Justine Michaux,^{1,2,3} Marie Taillandier-Coindard,^{1,2,3} Johanna Chiffelle,^{1,2,3} Baptiste Murgues,^{1,2,3} Talita Gehret,^{1,2,3} Aymeric Auger,^{1,2,3} Brian J. Stevenson,^{3,4} George Coukos,^{1,2,3,5} Alexandre Harari,^{1,2,3,5} and Michal Bassani-Sternberg^{1,2,3,5,6,*}

¹Ludwig Institute for Cancer Research, University of Lausanne, Agora Center Bugnon 25A, 1005 Lausanne, Switzerland

²Department of Oncology, Centre hospitalier universitaire vaudois (CHUV), Rue du Bugnon 46, 1005 Lausanne, Switzerland

³Agora Cancer Research Centre, 1011 Lausanne, Switzerland

⁴SIB Swiss Institute of Bioinformatics, Quartier Sorge, Bâtiment Amphipôle, 1015 Lausanne, Switzerland

⁵Center of Experimental Therapeutics, Department of Oncology, Centre hospitalier universitaire vaudois (CHUV), Rue du Bugnon 46, 1005 Lausanne, Switzerland

⁶Lead contact

*Correspondence: markus.muller@chuv.ch (M.M.), michal.bassani@chuv.ch (M.B.-S.)

<https://doi.org/10.1016/j.immuni.2023.09.002>

SUMMARY

The accurate selection of neoantigens that bind to class I human leukocyte antigen (HLA) and are recognized by autologous T cells is a crucial step in many cancer immunotherapy pipelines. We reprocessed whole-exome sequencing and RNA sequencing (RNA-seq) data from 120 cancer patients from two external large-scale neoantigen immunogenicity screening assays combined with an in-house dataset of 11 patients and identified 46,017 somatic single-nucleotide variant mutations and 1,781,445 neo-peptides, of which 212 mutations and 178 neo-peptides were immunogenic. Beyond features commonly used for neoantigen prioritization, factors such as the location of neo-peptides within protein HLA presentation hotspots, binding promiscuity, and the role of the mutated gene in oncogenicity were predictive for immunogenicity. The classifiers accurately predicted neoantigen immunogenicity across datasets and improved their ranking by up to 30%. Besides insights into machine learning methods for neoantigen ranking, we have provided homogenized datasets valuable for developing and benchmarking companion algorithms for neoantigen-based immunotherapies.

INTRODUCTION

In recent years, it has been demonstrated across tumor types in patients receiving adoptive transfer of autologous *in vitro* cultured tumor infiltrating lymphocytes (TILs) that T cells specifically recognizing mutated neoantigens play a key role in mediating effective anti-tumor responses.^{1–3} Furthermore, neoantigens are found to be implicated in the therapeutic efficacy of immune checkpoint inhibitor antibodies,^{4,5} and several studies show immune recognition following neoantigen-based vaccines,^{6,7} where patients experience no major toxicities.

Mutated proteins are processed and presented on tumor cells as human leukocyte antigen (HLA) binding peptides (HLAp) and are recognized by cognate T cell receptors (TCRs) as “non-self.” Targeting such neoantigens enables immune cells to distinguish between normal and cancerous cells, diminishing the risk of autoimmunity. Technological improvements in genomics, bioinformatics, and *in silico* HLA binding prediction tools have facilitated breakthroughs in the discovery of neoantigens encoded by somatic non-synonymous single-nucleotide variants (SNVs), insertions and deletions (InDels), and frameshifts (FSs) that arise

during the process of tumorigenesis and are not expressed by normal cells.^{8,9} Furthermore, advanced immunological screening techniques have facilitated the detection and isolation of neoantigen reactive T cells.^{10–13}

The development of innovative clinical treatment options targeting neoantigens requires the identification of neoantigens that are targeted by autologous T cells. However, only a small percentage of neoantigens are immunogenic, which makes their identification challenging.¹⁴ Various algorithms that score and rank neoantigens based on their likelihood of being presented on the patient’s HLA molecule^{15–18} and being specifically recognized by high avidity T cell clonotypes^{19–23} have been proposed. Other groups have provided pipelines for mutation detection and neoantigen prioritization.^{24,25} Despite all these efforts, a recent study shows little consensus in the neoantigen ranking performed by different laboratories,²⁶ and the performance of immunogenicity prediction methods varies between different datasets.¹⁴ As datasets with hundreds or thousands of neoantigen immunogenicity measurements become available,^{26–28} machine learning (ML) methods are able to train powerful immunogenicity prediction algorithms taking into account the multidimensional



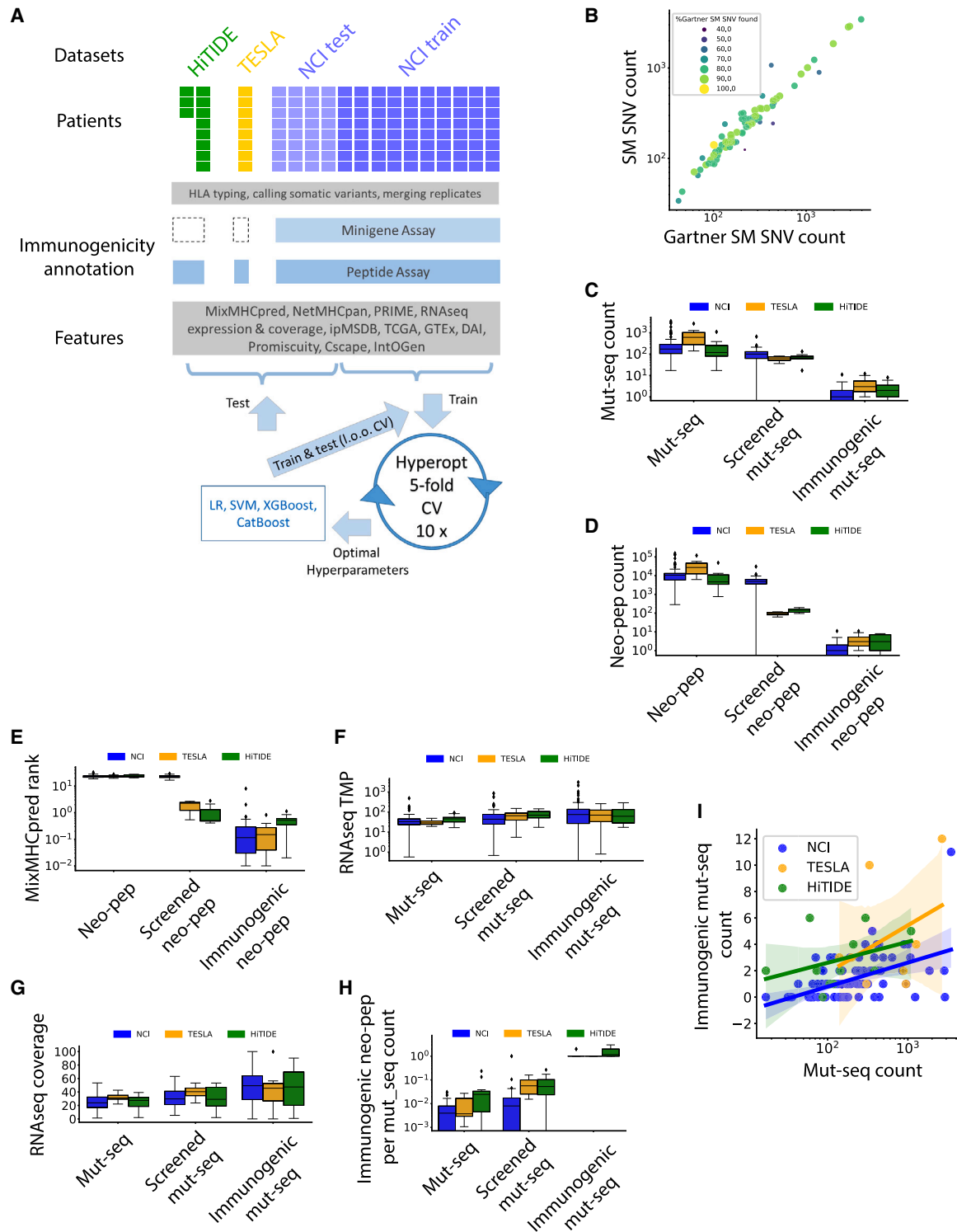


Figure 1. Statistics reveal the reproducibility of our pipeline and the bias in mutation and neo-peptide subsets

(A) Data processing workflow applied in this paper. WES and RNA-seq data were downloaded and processed. Mutations and neo-peptides were annotated with the results from the immunogenicity screens, and the feature scores and annotations were added. The NCI data matrix was split into train- and test sets, and the classifiers were trained on the subset of screened mutations or neo-peptides in NCI-train (see STAR Methods for naming rules for the data subsets) using Hyperopt parameter optimization and 5-fold cross validation (CV) in 10 replicate runs. The trained classifiers were tested on all neo-peptides or mutations in NCI-train (using leave one out CV), NCI-test, TESLA, and HITIDE cohorts.

(legend continued on next page)

structure of the data. In a recent example, the ranking based on an ML model has outperformed a ranking based on binding affinity only.²⁸ This improvement in prioritizing immunogenic neoantigens is particularly important for neoantigen or mRNA vaccines, where only a limited set of neoantigens are included.^{2,3,6,7}

Here, we studied the performance of state-of-the-art ML algorithms using two public datasets (National Cancer Institute [NCI] with 112 patients^{27,28} and Tumor Neoantigen Selection Alliance [TESLA] with 8 patients²⁶) plus an additional in-house dataset composed of 11 patients, 2 of which were already included in a previous publication.¹³ We reprocessed all whole-exome sequencing (WES) and RNA sequencing (RNA-seq) data with a uniform mutation detection pipeline and investigated the robustness of different ML algorithms and data preprocessing steps. We demonstrated that classifiers trained on the large NCI dataset can accurately predict the immunogenicity of neoantigens on each test dataset. With orthogonal features, our ML based approach outperformed previously published methods²⁸ and increased the number of immunogenic peptides ranked in the top 20 by 30%. Compared with the ranking reported in the TESLA consortium study,²⁶ our ML methods performed favorably and came first in two out of three ranking evaluation metrics. We provide classifiers and data processing methods for the improved prioritization of immunogenic neoantigens. The uniformly processed datasets are unique resources for other groups active in the field of immunogenicity prediction and in the development of innovative neoantigen-based therapies.

RESULTS

Our mutation detection is consistent with published results

Cancer cells can have several hundred somatic mutations (SMs), but only a few of them may be presented as HLA binding neo-peptides and recognized by T cells. The accurate selection of a limited number of mutations (e.g., for mRNA cancer vaccine) or neo-peptides (e.g., for multimer based sorting of neoantigen-specific T cells) that are most likely to be immunogenic is a crucial step in cancer vaccines and adoptive transfer of T cells. Here, we used two public (NCI^{27,28} and TESLA²⁶) and one in-house (the Human Integrated Tumor Immunology Discovery Engine; HiTIDE) dataset to train and test ML algorithms for the prioritization of mutations and neo-peptides (Figure 1A). The datasets consisted of WES and RNA-seq data as well as immunogenicity assay results for hundreds of mutations and/or neo-

peptides (Table 1; Data S1). The main difference between the datasets laid in the way mutations (*mut-seq*, typically 25 amino acid (aa) sequences with mutation in the center) or neo-peptides (*neo-pep*, peptides of length 8–12 including mutation) were selected for immunogenicity screening and in the screening methods used (STAR Methods). In the NCI dataset, many mutations and neo-peptides were physically screened as reported by Gartner et al.²⁸ In a cohort of 112 patients, which we defined here as *NCI_mut-seq*, for almost all the expressed mutations, minigenes encoding the mutations and 12 flanking wild-type (WT) aa on each side were transcribed *in vitro* and transfected into autologous antigens presenting cells (APCs) followed by a co-culture with TIL cultures and interferon (IFN)- γ enzyme-linked immunospot (ELISpot) immunogenicity measurement. For 80 of the 112 patients, a cohort we defined as the *NCI_neo-pep*, additional immunogenicity screens were performed to identify the optimal neo-antigenic epitopes and their HLA restrictions. The top-ranked neo-peptides predicted by NetMHCpan to span immunogenic mutations from the above mini-gene assay were pulsed on autologous APCs or APCs engineered to express the patient's HLA-I alleles, prior to co-culture with TILs and IFN- γ ELISpot readout. Neo-peptides with positive ELISpot readout were assigned as immunogenic. All other neo-peptides containing the immunogenic mutation and all neo-peptides containing screened non-immunogenic mutations were considered as non-immunogenic. In the TESLA study, immunogenicity of selected neo-peptides was determined with labeling of subject-matched TILs or peripheral blood mononuclear cells (PBMCs) with HLA-I peptide multimers.²⁶ The immunogenicity of selected neo-peptides in the HiTIDE cohort was interrogated with IFN- γ ELISpot assays following incubation of the peptides with either bulk TILs or neo-antigen enriched TILs (NeoScreen method) grown from tumor biopsies in the presence of APCs loaded with neo-peptides (Figures S1A and S1B), as previously described.¹³ Importantly, in the TESLA and HiTIDE datasets, only a selection of neo-peptides was experimentally screened, and the immunogenicity annotation of the mutations was inferred accordingly.

First, we uniformly processed all data, conducting HLA typing, mutation calling, RNA-seq gene expression analysis, and read coverage assessment at the specific loci of the SM. To assure capturing all relevant mutations in the NCI dataset, prior to the ML training, we assessed the extent to which we were able to reproduce the genomic analysis results published by Gartner et al.²⁸ First, for a subset of 80 of the 112 patients, for which HLA typing results from Gartner et al. were available, we

(B) Comparison of SM SNV mutation counts obtained from Gartner et al.²⁸ and our analysis for a subset of 80 patients, where each patient corresponds to a data point. The size and color of the points reflect the percentage of mutations identified by Gartner et al. that were also identified in our analysis. (C) to (H) show different statistics of the patient data for the NCI, TESLA, and HiTIDE datasets, where only SM SNVs were considered. The statistics were obtained from all mutations or neo-peptides per patient (left group of boxes), or from the subsets of screened and immunogenic mutations or neo-peptides per patient (middle and right group of boxes).

(C) Mutation counts.

(D) Neo-peptide counts. The outliers in (C) and (D) originate from NCI patients for which all non-immunogenic peptides were annotated as not-screened in Gartner et al.²⁸

(E) Average MixMHCpred %rank scores.

(F) Average RNA expression in TPM.

(G) Average RNA coverage of the mutations in %.

(H) Average number of immunogenic neo-peptides per mutation.

(I) Immunogenic mutation counts as a function of the mutation counts for each patient in the NCI, TESLA, and HiTIDE datasets.

See also Figure S1.

Table 1. Mutation and neo-peptide counts for the NCI, TESLA, and HiTIDE datasets and their subsets

	Number of patients (train or test datasets)	Immunogenicity screening method	Immunogenic	Not immunogenic	Not screened
NCI mutations	89 (train), 23 (test)	minigenes, IFN γ ELISpot	146	11,651	24,899
NCI neo-peptides	57 (train), 23 (test)	peptides, IFN γ ELISpot	103	418,872 ^a	953,486
TESLA mutations	8 (test)	<i>in silico</i>	36 ^{b,c}	461 ^b	6,231 ^b
TESLA neo-peptides	8 (test)	peptides, HLA-I peptide multimers	34 ^c	702	300,505
HiTIDE mutations	11 (test)	<i>in silico</i>	30 ^b	751 ^b	1,812 ^b
HiTIDE neo-peptides	11 (test)	peptides, IFN γ ELISpot	41	1,511	106,191

^aNon-immunogenic NCI neo-peptide dataset contains three types of neo-peptides: the screened ones with negative immunogenicity test, the inferred ones from immunogenic mutations that were not screened at the neo-peptide level, and the ones from screened but non-immunogenic mutations.

^bImmunogenicity of mutations in TESLA and HiTIDE datasets was inferred from the immunogenicity screens of the respective neo-peptides.

^cNeo-peptides are excluded if they match WT peptide or missing value cannot be imputed resulting in more mutations than neo-peptides.

compared the HLA allotypes. The HLA typing was overall consistent, and we found that, for 74 patients, all alleles were identical, for 4 patients, 1 or 2 alleles were missed by us or by Gartner et al., and in 2 patients, there were conflicting alleles but with similar sequence motifs (Data S2). In addition, the SNV SM counts we obtained correlated well with the counts reported by Gartner et al. (Figure 1B; Data S2). Overall, in the subset of 80 patients we identified 31,880 SNV SMs, including 82.2% (26,420 out of 32,148) of the SNV SMs published by Gartner et al., where 67.5% of the patients (54 out of 80) had a SNV SM overlap larger than 80% (Data S2). For a few patients there was a substantial variation in the number of mutations detected and for two patients we called less than 50% of the mutations reported by Gartner et al. Interestingly, we detected 143 of the 151 (94.7%) immunogenic SNV SMs published in Gartner et al. (Data S2; Figure S1C). Good correlations were also obtained when we compared insertion and deletion mutations with and without FSs (Figures S1D and S1E).

Immunogenicity-related feature scores highlight subtle differences between datasets

Next, we added multiple feature scores (Data S3) reflecting the propensity of a peptide to be presented, such as bulk RNA-seq gene expression of the mutated gene and its expression in the tissue-matched Cancer Genome Atlas (TCGA) (<https://www.cancer.gov/tcga>) and the tissue-matched Genotype-Tissue Expression (GTEx) atlas (<https://gtexportal.org/>), proteasomal cleavage scores,²⁹ tapasin binding,³⁰ binding affinity to HLA-I allotypes (NetMHCpan,¹⁵ MixMHCpred¹⁶), and stability ranks.³¹ Other feature scores evaluated the dissimilarity of a neo-peptide to the WT peptide counterpart (differential agretopicity index or DAI)^{19,22,32,33} and the potential of a neoantigen to bind several alleles. A notable bias toward hydrophobic aa was observed at T cell receptor contact residues within immunogenic epitopes.³⁴ We therefore employed also the PRIME predictor, that captures such hydrophobicity related molecular properties associated with TCR recognition.^{23,35} We also used our large-scale in-house immunopeptidome database (ipMSDB³⁶) of HLA-bound WT peptides identified by mass spectrometry (MS) to assess the likelihood of neo-peptides to be naturally processed and presented at the cell surface by HLA (see below and in STAR Methods).

Last, it is well established that mutations in oncogenes and tumor suppressors are enriched across cancers and specific sites are more frequently mutated. Hoyos et al. has modeled the relationship between oncogenicity and immunogenicity for tumor driver mutations, focusing on p53 mutations, and demonstrated that hotspot mutations optimally solve an evolutionary trade-off between oncogenic potential and neoantigen immunogenicity.³⁷ Therefore, we scored SNV SM based on their appearance in the population with the Integrative Onco Genomics (IntOGen) database,³⁸ and we predicted their oncogenic status (disease-driver or neutral) with the CScape tool³⁹ to assess the role of the mutation in tumorigenesis.

Comparison of basic statistics across all three datasets (Data S1) revealed that the number of SNV SMs called per patient was highest for the TESLA dataset (Figure 1C), which contained only melanoma and non-small cell lung cancer (NSCLC) samples that are known for high mutational loads. In contrast, the number of mutations per patient screened with the mini-gene approach in the NCI dataset was higher than the mutations included in neo-peptide screens in the TESLA and HiTIDE datasets. The number of immunogenic mutations per patient was higher in TESLA and HiTIDE, possibly because of differences in cancer types and the sensitivity of immunogenicity screening methods. In the NCI dataset—following the annotations provided in Gartner et al.²⁸—all neo-peptides originating from screened mutations were considered as screened, even if only the mutation, but not the neo-peptide was actually screened. Therefore, the number of neo-peptides annotated as “screened” was much higher in the NCI dataset (Figure 1D), and there was no difference in binding affinity between screened and not-screened neo-peptides (Figure 1E). In contrast, binding affinity was used as a screening criterion in the TESLA and HiTIDE datasets (Figure 1E). The RNA-seq gene expression values revealed small differences between datasets. In all datasets, mutations selected for T cell screening had higher RNA-seq gene expression, and this effect was strongest in the HiTIDE- and weakest in NCI data (Figure 1F). RNA-seq mutation coverage was consistently employed as a screening criterion in all datasets, with the TESLA dataset demonstrating the most pronounced utilization of this filter (Figure 1G). The number of immunogenic neo-peptides per mutation was higher in HiTIDE and TESLA datasets (Figure 1H). In the NCI and TESLA datasets, on average only one immunogenic

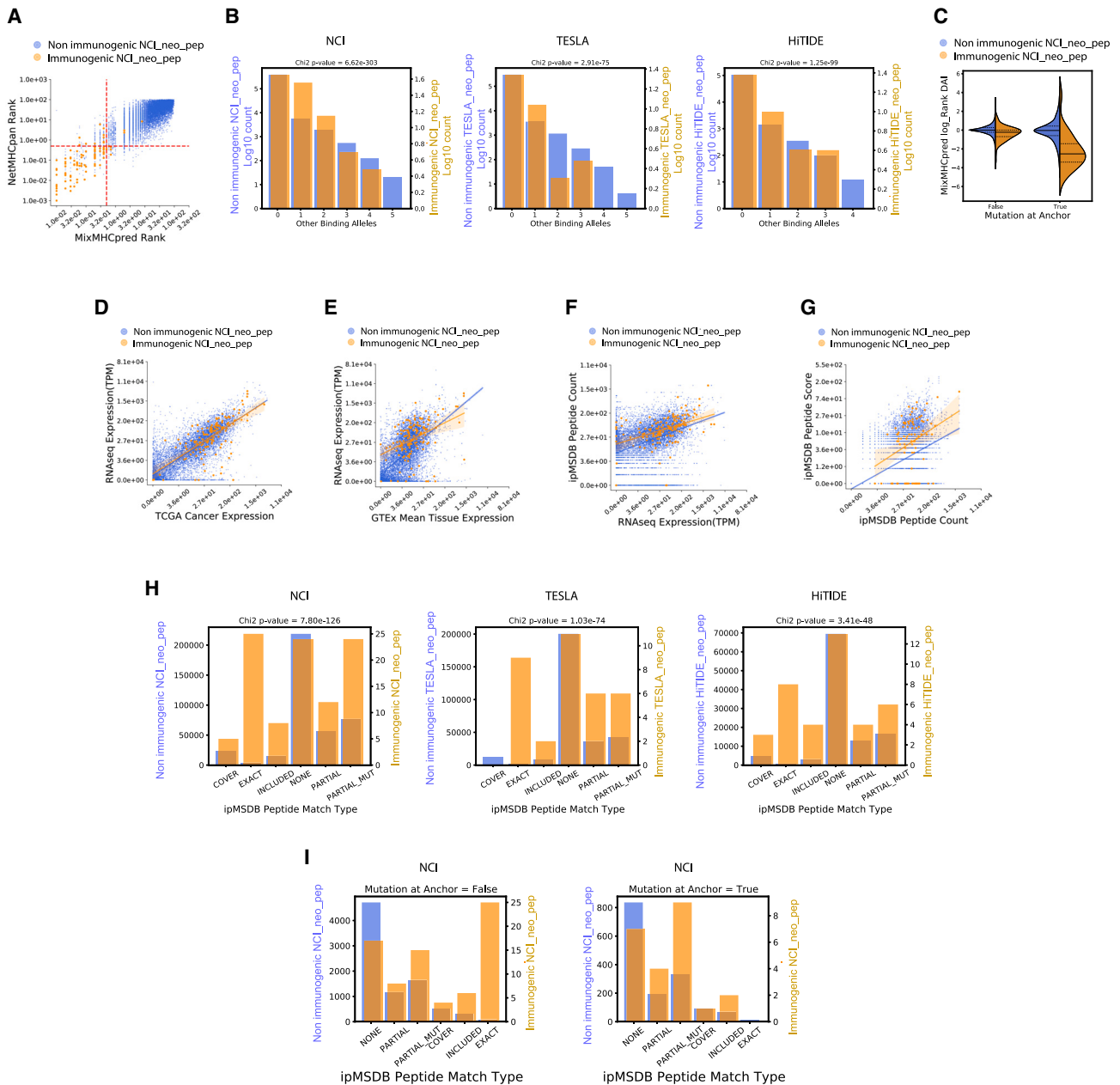


Figure 2. Exploring relationships between features and predictive value for immunogenicity

Scatterplots display the immunogenic (orange) and non-immunogenic (blue) neo-peptides or mutations with their regression lines for the screened *NCI_neo-pep/mut-seq* dataset. Only a random subsample of 10,000 points of the non-immunogenic points is shown in the scatterplots. Histograms display the feature scores of immunogenic (orange) and non-immunogenic (blue) neo-peptides for the screened *NCI_neo-pep*, *TESLA_neo-pep*, and *HITIDE_neo-pep* datasets. The scale of the immunogenic neo-peptide counts is given on the right y axis; the scale of the non-immunogenic counts is on the left y axis. The p values shown in the histogram titles evaluate the difference between immunogenic and non-immunogenic feature values and are calculated by a χ^2 test.

(A) Scatterplot of MixMHCpred and NetMHCpan %rank scores. Red dashed lines mark the 0.5% ranks.

(B) Histogram for "Number Binding Alleles" scores. Note the different log-scales for immunogenic and non-immunogenic neo-peptides counts.

(C) Violin plot of MixMHCpred log-rank DAI for neo-peptides with mutations at anchor and non-anchor positions.

(D) TCGA expression versus RNA-seq expression.

(E) GTEx expression versus RNA-seq expression.

(F) Scatterplot of *ipMSDB Peptide Count* per protein versus RNA-seq expression.

(G) *ipMSDB Peptide Count* per protein versus *ipMSDB Peptide Score*.

(legend continued on next page)

neo-peptide was detected per immunogenic mutation, whereas in the HiTIDE cohort this number was slightly higher. The number of immunogenic neo-peptides per patient correlated with the total number of SNV SMs detected in a patient (Figure 1). In summary, the NCI dataset had the highest number of screened mutations and neo-peptides with the least selection bias and is therefore most suitable for training ML models.

Features beyond binding affinity and gene expression correlate with immunogenicity

Next, we investigated how the mutation or neo-peptide features correlated with immunogenicity. By examining these correlations, we sought to gain insights into the factors that contribute to immunogenicity and potentially identify key determinants of immune recognition. We found that, in agreement with published results,^{26,28,40} features describing proteasomal cleavage, transporter associated with antigen presentation (TAP) import into endoplasmic reticulum and binding stability, correlated with immunogenicity in all three datasets, and they correlated poorly with binding affinity (Figures S2A–S2C). In addition, as previously demonstrated,^{26,28} features reflecting the binding affinity between a neo-peptide and the patients' HLA-I alleles were among the strongest predictors for immunogenicity for all three datasets (Figure S2D). Although NetMHCpan and MixMHCpred prediction %rank scores correlated, they contained complementary information. For example, in the NCI dataset, ten immunogenic neo-peptides did not pass the binding threshold of %rank ≤ 0.5 with NetMHCpan, but they passed it with MixMHCpred (Figure 2A). We found that promiscuous neo-peptides that were predicted to bind to multiple patient's HLA-I alleles were more likely to be immunogenic than neo-peptides predicted to bind a single allele (Figure 2B), possibly because binding to multiple alleles increases the chance for HLA-I presentation and makes the presentation of neo-peptides more resistant to loss of specific HLA-I alleles.⁴¹ Along the same lines, mutations with a higher number of neo-peptides weakly binding to a patient's HLA-I alleles, were more likely to be immunogenic (Figure S2E). The PRIME prediction rank differences between immunogenic and non-immunogenic neo-peptides were similar to those of MixMHCpred or NetMHCpan (Figure S2F). DAI values for binding prediction log-ranks were lower for immunogenic neo-peptides (Figure S2G) in agreement with previous results.^{19,22,33} As expected, the location of mutations in an anchor position was not significant per se (Figure S2H), but it became important in combination with DAI values, which were significantly lower (t test p value 2.19×10^{-17}) for immunogenic mutations at anchor positions (Figure 2C). Based on the analyzed data, there was no obvious tendency for mutations to be placed in the middle of a neo-peptide, and the enrichment of immunogenic mutations in the middle of 10 mers reported for the TESLA dataset²⁶ could not be confirmed for the NCI and HiTIDE datasets (Figure S2I). As expected, immunogenic neo-peptides were strongly enriched in the group of 9 or 10 mer peptides, reflecting the length preferences of HLA-I alleles (Figure S2J). HLA binding-aff-

inity predictors that incorporate peptide length preferences were used to select the neo-peptides for immunogenicity testing. Hence, based on these three datasets, it is challenging to determine whether this enrichment stems from a bias in selecting neo-peptides or if it represents an intrinsic characteristic of immunogenic peptides. It has been demonstrated that gene or protein expression positively impacts HLA-I presentation^{40,42} and immunogenicity.^{26,28} In all three datasets, immunogenic mutations had higher gene expression and higher coverage of the mutation in the patient's tumor bulk RNA-seq data compared with non-immunogenic ones (Figures S3A and S3B). To investigate the possibility of substituting a patient's gene expression values with data from publicly available datasets, particularly in scenarios where the patient's tumor tissue RNA-seq data are unavailable, we included tissue-matched RNA-seq expression data from the TCGA and GTEx repositories as additional features. For both immunogenic and non-immunogenic mutations, the TCGA gene expression correlated strongly (Pearson's $R = 0.818$) with its expression in the patient's cancer tissue (Figure 2D). The gene expression in GTEx correlated to a lower extent (Pearson's $R = 0.645$), and the regression line for immunogenic mutations was shifted to higher RNA-seq values compared with the regression line for non-immunogenic ones (Figure 2E). We concluded that immunogenic mutated genes had higher gene expression in cancer tissues compared with the matched healthy tissues in GTEx, and the expression values were better captured by TCGA. Lastly, cancer cell fraction (CCF), clonality, and zygosity were not associated with immunogenicity (Data S3).

Our in-house ipMSDB database³⁶ contains WT HLA-I and -II ligands identified by MS in multiple healthy and cancerous human tissues and cell lines with various HLA allotypes. The ipMSDB version used in this work contains 547,476 unique HLA-I peptides, which we used to infer the HLA-I presentation of a neoantigen based on the coverage of the corresponding WT peptide and on the natural presentation of the source protein. We found that the number of ipMSDB peptides mapped to a protein ("*ipMSDB Peptide Count*") was significantly higher for proteins containing immunogenic mutations across all three datasets (Figure S3C). These data indicate that immunogenic peptides in the three datasets preferably belong to proteins that are naturally processed and presented, in agreement with previous findings.^{18,36,43} *ipMSDB Peptide Count* for a given protein correlated (Pearson's $R = 0.498$) with mRNA expression of the corresponding gene (Figure 2F), but this correlation could not fully explain the higher *ipMSDB Peptide Count* values for immunogenic mutations (Figure S3D), suggesting that these features are not fully redundant. In addition, the "*ipMSDB Peptide Score*" measures the overlap between the WT peptide within ipMSDB and the neo-peptides (Figure S3E). The correlation between the *ipMSDB Peptide Score* and the *ipMSDB Peptide Count* (Pearson's $R = 0.435$) reflects that proteins with overall more ipMSDB peptides had a better chance to cover a neo-peptide. However, the *ipMSDB Peptide Score* was higher for

(H) Histograms for "*ipMSDB Peptide Match Type*."

(I) Histograms for *ipMSDB Peptide Match Type* for neo-peptides with or without a mutation at an anchor position of the HLA allele with the lowest MixMHCpred % rank score.

See also Figures S2 and S3.

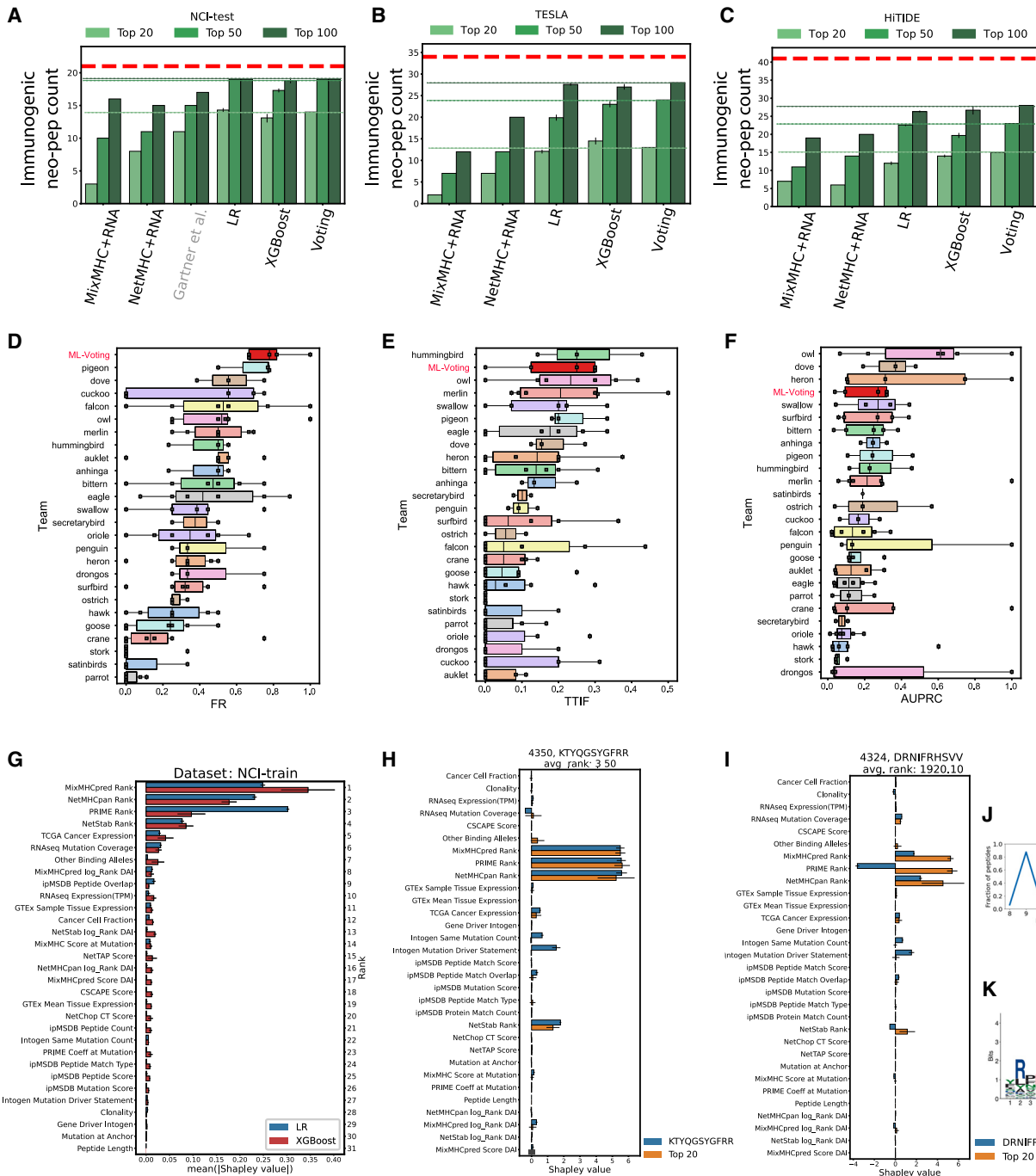


Figure 3. Assessments of the classifier's performance and feature importance

(A) Immunogenic neo-peptides were ranked per patient and the number of immunogenic neo-peptides in the top 20, 50, or 100 ranks was calculated per patient and summed up for all patients in the NCI-test dataset. The ranking was performed either by NetMHCpan and RNA expression, MixMHCpred and RNA expression as described in the text, or logistic regression (LR), XGBoost, or the voting classifier. "Gartner et al." refers to the ranking reported in Gartner et al.²⁸ The red dashed horizontal lines indicate the total number of immunogenic neo-peptides in NCI-test. The green lines mark the median performance of the voting classifier in the top 20, 50, or 100 ranks according to their respective colors.

(B) As in (A), but for the TESLA dataset.

(C) As in (A), but for the HITIDE dataset.

(D) Comparison of the fraction ranked (FR) score obtained by the voting classifier trained on NCI-train and tested on TESLA. FR scores of the TESLA participants were obtained from Wells et al.²⁶ The FR score gives us the fraction of immunogenic neo-peptides ranked in the top 100 per patient.

(E) Same as (D) but for the top-20 immunogenic fraction (TTIF) score. The TTIF score gives us the fraction of immunogenic neo-peptides among all screened neo-peptides ranked in the top 20 per patient.

(legend continued on next page)

immunogenic neo-peptides compared with non-immunogenic ones (Figure 2G), and this shift was significant in all three datasets (Figure S3F). We also found a highly significant enrichment of immunogenic neo-peptides, which either mapped exactly to the WT counterpart sequences in ipMSDB or were fully included in such sequences (Figure 2H). These results indicated that immunogenic neo-peptides were preferably found in HLA-I presentation “hotspots” and that utilizing sequence matching to ipMSDB proves to be an effective strategy for prioritizing “true” HLA-I binding neo-peptides, as long as the mutation does not occur in an anchor position (Figure 2I). When mutations arise in anchor positions, they tend to produce a predicted peptide variant that exhibits superior binding affinity compared with the original WT peptide especially for immunogenic peptides (Figure 2C). Consequently, in these scenarios, the likelihood of finding the WT peptide represented in the ipMSDB is reduced (Figure 2I).

Further, we included features that evaluate the impact of a mutation on the cellular or molecular function of the mutated protein. Although Cscape³⁹ is an oncogenicity predictor, we demonstrated that it had also a predictive value for immunogenicity (Figure S3G), possibly because oncogenic mutations often destabilize the protein structure, leading to rapid degradation of the protein and presentation on HLA-I.⁴⁴ We also included mutation annotations from the IntOGen³⁸ database, and we further found that mutations annotated as oncogenic drivers were enriched for immunogenicity in all three datasets (Figure S3H), and there was a slight immunogenicity enrichment for mutations with a lower prevalence in the population (Figure S3I).

Classifiers trained on a large unbiased dataset accurately rank neo-peptides in other datasets

Neoantigen-based personalized immunotherapy strategies rely on the selection of the most promising mutations or neo-peptides. For both mutations and neo-peptides, we trained a separate ML model, which calculates the probability that a mutation or neo-peptide can induce a spontaneous immune response, as was captured by the immunogenicity screening assays, and this probability is then used for the ranking. First, we investigated the ranking of neo-peptides. We used the Bayesian optimization framework Hyperopt⁴⁵ to train the classifiers and their hyperparameters on NCI-train (Figure 1A; see supplemental information for the details). Through leave-one-out cross-validation (CV) testing on the NCI-train dataset, we observed that the logistic regression (LR)⁴⁶ classifier’s performance showed improvement as the number of non-immunogenic neo-peptides in the training set increased (Figure S4A). Additionally, increasing the

number of Hyperopt iterations also contributed to the enhanced performance of the LR classifier (Figure S4B). These findings highlight the importance of a larger training set and extensive Hyperopt iterations in optimizing the performance of the LR and other classifiers for neo-peptide immunogenicity prediction.

Furthermore, the choice of data normalization method had an impact on the performance of the LR classifier, as demonstrated by Figure S4C. Notably, employing quantile normalization resulted in a remarkable 134.0% increase in the number of immunogenic neo-peptides ranked within the top 20, in comparison with the scenario where no normalization was applied (Figure S4D). These findings underscore the importance of implementing appropriate data normalization techniques, such as quantile normalization, to enhance the accuracy and predictive power of the LR classifier.

Furthermore, the choice of classifier algorithm had an impact on the number of immunogenic neo-peptides ranked among the top positions (Figure S4E). For NCI-train with leave-one-out CV, LR performed best, followed by XGBoost,⁴⁷ CatBoost,⁴⁸ and the SVMs.⁴⁹ The LR classifier was able to rank 49.1% of immunogenic neo-peptides in the top 20, 62.2% in the top 50, and 75.6% in the top 100 (Figure S4F; Data S4). The principal-component analysis (PCA) plot (Figure S4G) revealed that LR and XGBoost produce distinct and complementary rankings. The plot visually demonstrated that these LR and XGBoost offer diverse perspectives and capture different aspects of neo-peptide immunogenicity, indicating the potential benefit of leveraging their combined results for a more comprehensive and accurate assessment of immunogenic rankings. Therefore, we constructed a voting classifier, which averaged the immunogenic class probabilities of all ten LR and ten XGBoost classifier replicates (STAR Methods). Across the NCI-test, TESLA, and HiTIDE test datasets, the ranking of the voting classifier was always better or comparable to the rankings of the LR and XGBoost classifiers (Figures 3A–3C). We concluded that the voting classifier provides a ranking that is more robust and less dependent on the dataset.

The performance of ML ranking can vary depending on the dataset used. To investigate this further we trained and tested the LR classifier on HiTIDE with leave-one-out CV (see STAR Methods) and compared it with the LR classifier trained on the much larger NCI-train dataset. The HiTIDE-trained LR performed clearly better on HiTIDE neo-peptides, but it performed worse on the TESLA and NCI-test datasets (Figures S4H–S4J). The LR classifiers had a preference for features such as RNA-seq expression, CCF, and ipMSDB scores, which were used in the HiTIDE cohort to select neo-peptides for immunogenicity screening (Figure S4K). These findings demonstrated that ML

(F) Same as (D) but for the “area under the precision recall curve” (AUPRC) score. The AUPRC score gives us the ability of a ranking to place immunogenic neo-peptides before non-immunogenic ones.

(G) Neo-peptide feature importance calculated using Shapley values for LR and XGBoost classifiers trained on NCI-train.

(H) Shapley values of the KTYQGSYGFR neo-peptide (blue bars) from NCI-test patient 4,350 compared with average Shapley values of the top 20 ranked neo-peptides of patient 4,350 (orange bars). LR classifier trained on NCI-train ranked the neo-peptide in rank 3.5 on average. The error bars indicate the standard deviation over the ten replicate runs.

(I) Same as in (H) but for immunogenic neo-peptide DRNIFRHSV of patient 4,324 in NCI-test (blue bars), which had an average rank of 1,920.1 in the 10 replicate runs.

(J) HLAp length distribution for HLA-C06:02 allele taken from MHC Motif Atlas (<http://mhc motif atlas.org/>).

(K) Bare motif without pseudo-count correction obtained from the 75 HLA-C06:02 10 mers included in HLA Motif Atlas.

See also Figure S4.

classifiers could easily capture inherent biases related to the selection of neo-peptides for screening assays, potentially resulting in suboptimal rankings when applied to other datasets. This justifies our approach of training our classifiers on the NCI dataset, which is characterized by minimal bias, to mitigate the impact of dataset-specific biases and achieve more accurate and reliable rankings.

Next, we compared the performance of our ML ranking methods with an alternative simple approach where neo-peptides were initially sorted based on MixMHCpred or NetMHCpan %rank scores and then by RNA-seq expression to resolve the ties. We demonstrated the superior performance of the ML classifiers compared with this basic ranking strategy (Figures 3A–3C). NetMHCpan performed better than MixMHCpred on the NCI-test and TESLA datasets, where NetMHCpan was used to select neo-peptides for screening, but led to similar ranking for HiTIDE, where MixMHCpred was used for the screening selection. Finally, we compared our results with the rankings published by Gartner et al.²⁸ for the 23 patients in NCI-test. Our results demonstrated that LR, XGBoost, and the voting classifiers ranked more immunogenic neo-peptides in the top 20, 50, and 100 ranks (Figure 3A; Data S4). Compared with Gartner et al., LR placed 30.0% more neo-peptides into the top 20, 26.7% more into the top 50, and 11.8% more into the top 100.

In addition, we conducted a comparison between our ML approach for the TESLA dataset and the consortium results reported by Wells et al. for these data.²⁶ Our ML ranking achieved the best ranking among the TESLA participants when considering the three evaluation metrics introduced by Wells et al.,²⁶ with an average rank of 2 compared with the second-best average rank of 3.3 for the “owl” group. Specifically, our voting classifier obtained a median “fraction ranked (FR)” score (see STAR Methods) of 77.8% (Figure 3D), which was better than the FR scores reported by all other groups participating in the TESLA study. Our median “top-20 immunogenic fraction (TTIF)” score of 0.25, was reached by only one other group (Figure 3E), whereas our median area under the precision recall curve (AUPRC) score (0.273) ranked fourth among all participants (Figure 3F). Because the highest-ranking neo-peptides in the lists submitted by the TESLA participants were actually screened in the immunogenicity screens, we here evaluated the TESLA participants partially on their best-ranked peptides, whereas there was no such bias for our ML methods. The results clearly demonstrate that our ML classifiers, trained on the NCI-train dataset, exhibited strong generalization capabilities, and yielded highly accurate results when applied to the independent TESLA dataset.

In order to assess the significance of each feature in the LR and XGBoost rankings, we computed the Shapley values associated with each feature.^{50,51} This analysis allowed us to quantify the contribution of each feature in determining the final ranking of neo-peptides by these classifiers. Figure 3G shows that the strongest Shapley values for LR and XGBoost stemmed from MixMHCpred, NetMHCpan, and PRIME rank features, followed by stability rank, TCGA expression and RNA-seq mutation coverage, number of binding HLA alleles, MixMHCpred, DA1, and ipMSDB overlap score. For example, Figure 3H demonstrates the Shapley values for neo-peptide EKIALFQSL of patient 4,350 in NCI-test with an average rank of 48.1 in the ten LR replicates, which is much better than rank 1,641 reported by Gartner

et al. The better ranking resulted from the stronger binding affinity reported by MixMHCpred, PRIME, and NetMHCpan for the HLA-B39:01 allele compared with MHCFlurry v1.6, which was used by Gartner et al., but also IntOGen scores, binding stability, TCGA expression and “ipMSDB Peptide Match Overlap” contributed to the good rank. In contrast, the neo-peptide DRNIFRHSV from patient 4,324 in NCI-test was ranked poorly by our LR classifier (average rank 1,920.1) and by Gartner et al. (rank 24,392) because the peptide had poor %rank scores for allele HLA-C06:02 by all used binding-affinity predictors (Figure 3I), and also Gartner et al. reported a poor %rank with MHCFlurry. The HLA-C06:02 allele binds mainly 9 mers and only a few 10 mers (Figure 3J), resulting in a poor MixMHCpred %rank for 10 mers, even if the 75 10 mer ligands included in the major histocompatibility complex (MHC) Motif Atlas⁵² show a clear preference for arginine in the second, and leucine and valine in the 10th position (Figure 3K). Overall ipMSDB and IntOGen features had lower Shapley feature importance, but their contribution to the ranking of immunogenic neo-peptides was still evident (t test p value for *rank_score* increase is 2.54×10^{-8} for ipMSDB features, and 3.70×10^{-10} for IntOGen features) (Figure S4L). Excluding these features from LR prioritization reduced the number of *neo-peptides* ranked in the top 20 in NCI-test by 16.1%.

Effective ranking of immunogenic mutations requires dedicated training of classifiers

Most neoantigen-based cancer vaccination strategies use long mutated peptides (15–25 mers) or RNA mini-gene constructs encoding such sequences and rely on the selection of the most promising mutations. When prioritizing mutations, the relative importance of mutation features such as RNA-seq expression or coverage is expected to change compared with their significance in prioritization of the minimal neo-peptide sequences (see below). Therefore, instead of using the above neo-peptide classifiers to build a mutation ranking method, we trained mutation classifiers from scratch using the mutation features (Data S3). When we trained the LR and XGBoost classifiers on NCI-train, XGBoost slightly outperformed LR (Figures 4A, S4M, and S4N). For the NCI-test data, both LR and XGBoost performed better in the top 20 than the ranking published by Gartner et al. (Figure 4A; Data S4), but the difference was less pronounced than for neo-peptides. Although binding-affinity features were still most powerful (Figure 4B), the importance of non-HLA-binding-related features, such as RNA-seq coverage, TCGA expression, ipMSDB scores, and IntOGen scores features increased compared with the corresponding neo-peptide features, whereas the importance of binding-affinity ranks decreased (Figure 4C). This emphasizes that prioritizing mutations is different from prioritizing neo-peptides and requires different ML strategies. As for neo-peptides, ipMSDB and IntOGen features contributed complementary information and improved LR based mutation ranking (t test p value for *rank_score* increase is 0.0264 for ipMSDB features and 0.0844 for IntOGen features) (Figure S4O). Our approach enabled us to develop specialized classifiers dedicated to mutation and neo-peptide prioritization, thereby ensuring a more tailored and accurate assessment of their importance in the context of neoantigen immunogenicity prediction.

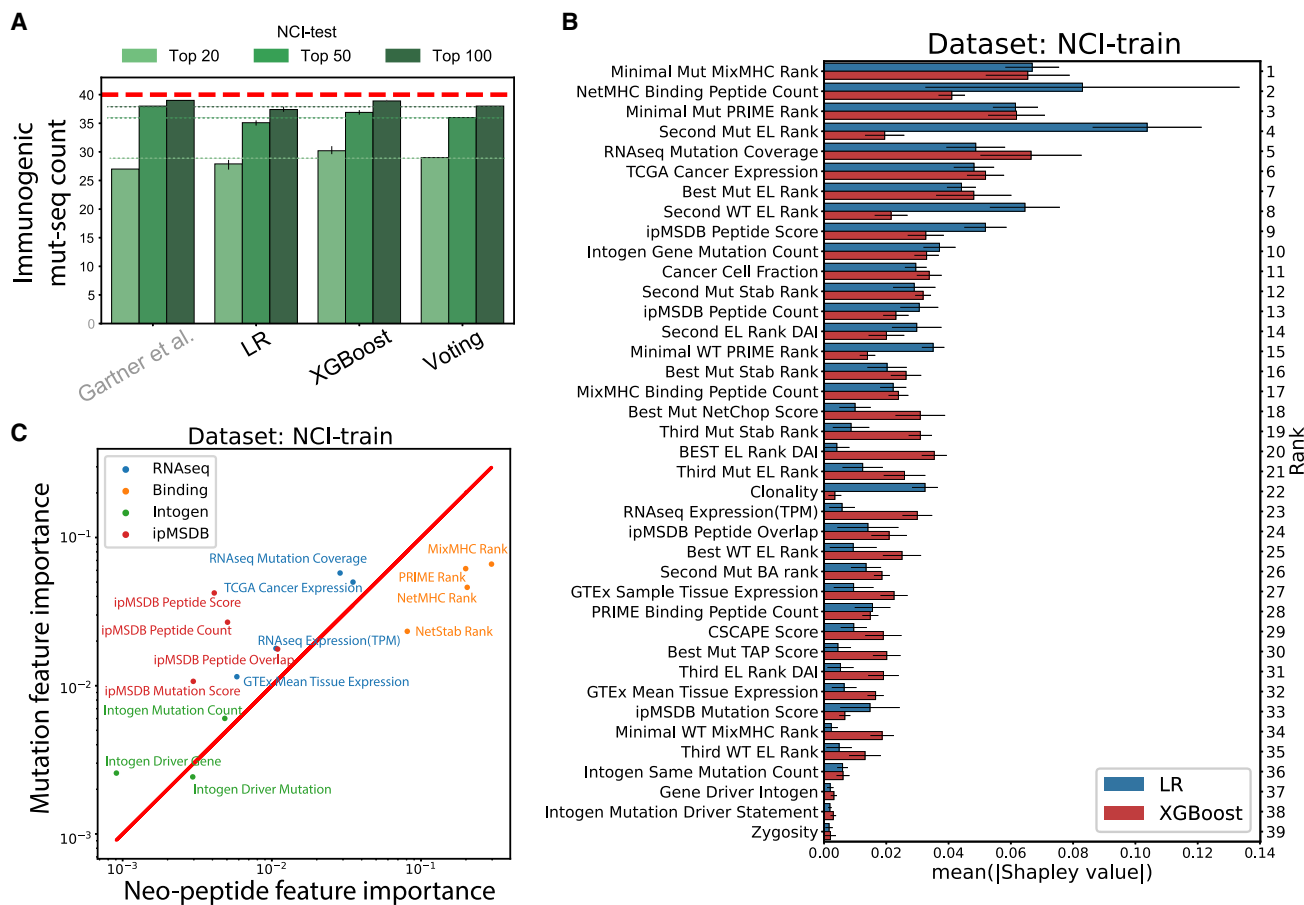


Figure 4. Effective ranking of immunogenic mutations requires dedicated training of classifiers

(A) Immunogenic mutations were ranked per patient and the number of immunogenic mutations in the top 20, 50, or 100 ranks was calculated for each patient. The y axis represents these numbers summed over all patients in the dataset. The red dashed horizontal lines indicate the total number of immunogenic mutations in a dataset. The number of top-ranking immunogenic mutations is shown for patients in NCI-test for the LR, XGBoost, and voting classifiers. Gartner et al. refers to the ranking reported in Gartner et al.²⁸ The horizontal green lines mark the mean performance of the voting classifier in the top 20, 50, or 100 ranks according to their respective colors.

(B) Mutation Shapley feature importance for the LR and XGBoost classifiers in NCI-train. The error bars indicate the standard deviation over 10 replicate runs. The features on the y axis are ordered by decreasing feature importance of both LR and XGBoost.

(C) Neo-peptide feature importance (Figure 3G) compared with mutation feature importance (B) for RNA-seq expression-, binding affinity-, IntOGen-, and ipMSDB-related features used by both neo-peptides and mutation classifiers.

See also Figure S4.

DISCUSSION

Accurate prediction and prioritization methods of patient-specific neoantigens is still an important barrier for development of effective cancer vaccines and neoantigen-based T cell therapies. Because currently the number of mutations included in a personalized cancer vaccine is in the range of about 20 mutations, the selection of mutations is rather straightforward in case of low tumor mutational burden (TMB)⁵³; however, this becomes a critical challenge in the medium to high TMB. Furthermore, the utilization of different validation assays for assessing immunogenicity in various laboratories, along with the use of diverse protocols for T cell isolation and expansion,⁵⁴ has a potential to introduce variations, underscoring the importance of harmonizing datasets and providing prediction solutions with generalized good performance across labs. Our systematic

analysis of immunogenic and non-immunogenic neoantigens, demonstrated that many feature scores reflecting processes of the antigen presentation machinery, such as binding affinity and stability, RNA expression and coverage, the presence of non-mutated counterparts of neo-peptides in immunopeptidome hotspots, binding promiscuity, and the role of the mutated gene in oncogenicity, were all predictive for immunogenicity across datasets and immunogenicity validation methods. Indeed, a neoantigen quality model incorporated similar features, such as the differential presentation and T cell cross reactivity against the neoantigen and its WT counterpart.⁵⁵ Variations of this model were applied to predict the survival of patients treated with anti-CTLA4 and anti-PD-1,⁵⁵ to predict immune editing in long term survivors of pancreatic ductal adenocarcinoma (PDAC),⁵⁶ and the induction of neoantigen-specific T cell responses following treatment with personalized mRNA vaccine.⁵³

However, the applicability of the “high-quality” model is limited to providing predictions solely for 9-mer peptides and the model does not consider the important information from RNA-seq data. The complex multidimensional structure of the feature manifold motivated the use of ML techniques, to efficiently combine these features for the prioritization of neo-peptides or mutations.

Beyond the selection of the descriptive features, we evaluated several data normalization methods and found that they had a strong impact on the outcome. In addition, we applied the Hyperopt⁴⁵ framework to find the optimal classifier hyperparameters, a technical step that is important for the overall performance of ML tools. Several classifier algorithms were then trained on the large NCI^{27,28} cohort, which was the least biased and most comprehensive of the three datasets. We observed that the LR⁴⁶ and XGBoost⁴⁷ classifiers outperformed the others and that their results were to some extent complementary, motivating the use of a voting classifier, which combined the LR and XGBoost probabilities and uniformly provided more robust results. Importantly, the LR and XGBoost classifiers trained on NCI-train resulted in accurate immunogenicity rankings for neoantigens in the TESLA²⁶ and in-house HiTIDE datasets, which have different HLA restrictions, originate in different tumor types, which were obtained from different laboratories and screened with different immunogenicity assays. Our ML ranking achieved the highest position among the TESLA participants when considering all three evaluation metrics. Additionally, our classifiers surpassed the performance of the classifier reported by Gartner et al. for the NCI-test dataset²⁸ in which our approach resulted in a remarkable 30% increase in the number of immunogenic neo-peptides ranked within the top 20.

In order to assess the significance of features in the classification task, we used Shapley values.^{50,51} For prioritization of mutations, features describing both the mutations (e.g., RNA expression and ipMSDB) and their neo-peptides (e.g., binding affinity) had high importance. In contrast, for prioritization of neo-peptides, binding affinity and stability features dominated. Overall, the performance of HLA binding prediction tools has greatly improved over the last years, especially due to the availability of high-scale accurate MS data of eluted HLA peptides and the implementation of advanced ML approaches. However, our analysis showed that for some peptides that failed to be placed in the top ranks, the limiting factor was, to our surprise, the still suboptimal accuracy of the HLA binding affinity prediction. Nevertheless, we demonstrated that many other features are positively associated with the ranking. This was particularly visible when we excluded ipMSDB and IntOGen features from the features set used for the classification, leading to a decrease in performance.

Our classifiers perform well for datasets with different immunogenicity validation methods, providing an advantage that allows them to be utilized by diverse groups, irrespective of their chosen validation methods. Our results will contribute to immunogenicity prediction in two scenarios. First, users can reproduce all the features we included in our work and apply our trained classifiers directly for antigen prioritization on their data or combine our classifiers with classifiers trained on their own data. Second, our harmonized datasets can serve as a basis. The available features can be edited, and additional features

can be included. Users can train and benchmark their own classifiers and ML methods with these datasets. To conclude, together with our ML classifiers and ML methods, we provide easily accessible data for method development and benchmarking with the aim to improve the selection of immunogenic neo-peptides and mutations for the development of effective personalized immunotherapy treatments.

Limitations of the study

Of note, some potential limitations should be considered. The datasets may contain some false-negative neo-peptides because only a subset was screened for immunogenicity. It is equally important to note that the assessment of neoantigen-specific responses may underestimate their true potential due to the possibility of T cell exhaustion, which can result in limited expansion or diminished reactivity during *in vitro* culture.⁵⁷ This situation could be improved by screening more neo-peptides per mutation or by applying semi-supervised learning methods, which use a combination of clustering and classification algorithms to correct the labels of some wrongly assigned data points. In this study we exclusively considered SNV SMs, but it is known that peptides mapped to insertions, deletions, and out-of-frame and gene fusion events have a high immunogenicity potential due to their increased dissimilarity to WT HLA-bound peptides. However, the amount of immunogenicity data available for non-SNV genomic alterations is limited. To circumvent this limitation, one could leverage the predictors built on SNV mutations and neo-peptides to predict the immunogenicity for non-SNV mutations too. In addition, HLA loss of heterozygosity and expression silencing frequently occurs in cancers^{41,58} and such silenced HLA alleles may be excluded from HLA binding and stability predictions. Furthermore, once enough data for CD4⁺ T cell recognition of neo-peptides will be available, predictors for neoantigens bound to HLA-II complexes may apply a similar approach.^{7,59–63}

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
 - Lead Contact
 - Materials Availability
 - Data and Code Availability
- METHOD DETAILS
 - Datasets

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at <https://doi.org/10.1016/j.immuni.2023.09.002>.

ACKNOWLEDGMENTS

This study was supported by the Ludwig Institute for Cancer Research, by grant KFS-4680-02-2019 from the Swiss Cancer Research Foundation (M.B.-S.), and by the Swiss National Science Foundation, PRIMA grant PR00P3_193079 (M.B.-S.). This work was also supported by grants from

Cancera, Mats Paulssons, and by a gift from the Biltema Foundation that was administered by the ISREC Foundation, Lausanne, Switzerland. We thank David Gfeller for very insightful discussions and critical remarks.

AUTHOR CONTRIBUTIONS

M.M. and M.B.-S. designed the study and drafted the manuscript. F.H., A.I.K., and B.J.S. designed, implemented, and executed the analysis of the WES and RNA-seq data. F.H., M.M., and E.R.A. wrote the software to calculate feature scores. M.M. designed and implemented ML methods, ML data analysis and visualization. J.M. and M.T.-C. performed sample preparation for WES and RNA-seq. M.A., J.C., and A.H. designed and analyzed the immunogenicity assays. B.M., T.G., and A.A. performed the immunogenicity assays. G.C. oversees clinical phase I trials, provided access to patient material, and helped with data interpretation. All authors read and helped revise the paper.

DECLARATION OF INTERESTS

The research presented in this paper is associated with the pending patent application PCT/EP2022/082845. The inventors listed on the patent application are M.B.-S., F.H., and M.M.

INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: March 27, 2023

Revised: June 26, 2023

Accepted: September 5, 2023

Published: October 9, 2023

REFERENCES

- Tran, E., Ahmadzadeh, M., Lu, Y.C., Gros, A., Turcotte, S., Robbins, P.F., Gartner, J.J., Zheng, Z., Li, Y.F., Ray, S., et al. (2015). Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* 350, 1387–1390. <https://doi.org/10.1126/science.aad1253>.
- Tran, E., Robbins, P.F., Lu, Y.C., Prickett, T.D., Gartner, J.J., Jia, L., Pasetto, A., Zheng, Z., Ray, S., Groh, E.M., et al. (2016). T-cell transfer therapy targeting mutant KRAS in cancer. *N. Engl. J. Med.* 375, 2255–2262. <https://doi.org/10.1056/NEJMoa1609279>.
- Chen, F., Zou, Z., Du, J., Su, S., Shao, J., Meng, F., Yang, J., Xu, Q., Ding, N., Yang, Y., et al. (2019). Neoantigen identification strategies enable personalized immunotherapy in refractory solid tumors. *J. Clin. Invest.* 129, 2056–2070. <https://doi.org/10.1172/JCI99538>.
- Rizvi, N.A., Hellmann, M.D., Snyder, A., Kvistborg, P., Makarov, V., Havel, J.J., Lee, W., Yuan, J., Wong, P., Ho, T.S., et al. (2015). Cancer immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* 348, 124–128. <https://doi.org/10.1126/science.aaa1348>.
- Snyder, A., Makarov, V., Merghoub, T., Yuan, J., Zaretsky, J.M., Desrichard, A., Walsh, L.A., Postow, M.A., Wong, P., Ho, T.S., et al. (2014). Genetic basis for clinical response to CTLA-4 blockade in melanoma. *N. Engl. J. Med.* 371, 2189–2199. <https://doi.org/10.1056/NEJMoa1406498>.
- Sahin, U., Derhovanessian, E., Miller, M., Kloke, B.P., Simon, P., Löwer, M., Bukur, V., Tadmor, A.D., Luxemburger, U., Schrörs, B., et al. (2017). Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* 547, 222–226. <https://doi.org/10.1038/nature23003>.
- Ott, P.A., Hu, Z., Keskin, D.B., Shukla, S.A., Sun, J., Bozym, D.J., Zhang, W., Luoma, A., Giobbie-Hurder, A., Peter, L., et al. (2017). An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* 547, 217–221. <https://doi.org/10.1038/nature22991>.
- Schumacher, T.N., and Schreiber, R.D. (2015). Neoantigens in cancer immunotherapy. *Science* 348, 69–74. <https://doi.org/10.1126/science.aaa4971>.
- Yarchoan, M., Johnson, B.A., Lutz, E.R., Laheru, D.A., and Jaffee, E.M. (2017). Targeting neoantigens to augment antitumor immunity. *Nat. Rev. Cancer* 17, 209–222. <https://doi.org/10.1038/nrc.2016.154>.
- Hadrup, S.R., Bakker, A.H., Shu, C.J., Andersen, R.S., van Veluw, J., Hombink, P., Castermans, E., Thor Straten, P., Blank, C., Haanen, J.B., et al. (2009). Parallel detection of antigen-specific T-cell responses by multidimensional encoding of MHC multimers. *Nat. Methods* 6, 520–526. <https://doi.org/10.1038/nmeth.1345>.
- Bentzen, A.K., Marquard, A.M., Lyngaa, R., Saini, S.K., Ramskov, S., Donia, M., Such, L., Furness, A.J.S., McGranahan, N., Rosenthal, R., et al. (2016). Large-scale detection of antigen-specific T cells using peptide-MHC-I multimers labeled with DNA barcodes. *Nat. Biotechnol.* 34, 1037–1045. <https://doi.org/10.1038/nbt.3662>.
- Bentzen, A.K., and Hadrup, S.R. (2017). Evolution of MHC-based technologies used for detection of antigen-responsive T cells. *Cancer Immunol. Immunother.* 66, 657–666. <https://doi.org/10.1007/s00262-017-1971-5>.
- Arnaud, M., Chiffelle, J., Genolet, R., Navarro Rodrigo, B., Perez, M.A.S., Huber, F., Magnin, M., Nguyen-Ngoc, T., Guillaume, P., Baumgaertner, P., et al. (2022). Sensitive identification of neoantigens and cognate TCRs in human solid tumors. *Nat. Biotechnol.* 40, 656–660. <https://doi.org/10.1038/s41587-021-01072-6>.
- Buckley, P.R., Lee, C.H., Ma, R., Woodhouse, I., Woo, J., Tsvetkov, V.O., Shcherbinin, D.S., Antanaviciute, A., Shughay, M., Rei, M., et al. (2022). Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens. *Brief. Bioinform.* 23, bbac141. <https://doi.org/10.1093/bib/bbac141>.
- Reynisson, B., Alvarez, B., Paul, S., Peters, B., and Nielsen, M. (2020). 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. <https://doi.org/10.1093/nar/gkaa379>.
- Bassani-Sternberg, M., Chong, C., Guillaume, P., Solleder, M., Pak, H., Gannon, P.O., Kandalaf, L.E., Coukos, G., and Gfeller, D. (2017). Deciphering HLA-I motifs across HLA peptidomes improves neoantigen predictions and identifies allosteric regulating HLA specificity. *PLoS Comput. Biol.* 13, e1005725. <https://doi.org/10.1371/journal.pcbi.1005725>.
- O'Donnell, T.J., Rubinsteyn, A., Bonsack, M., Riemer, A.B., Laserson, U., and Hammerbacher, J. (2018). MHCflurry: open-source class I MHC binding affinity prediction. *Cell Syst.* 7, 129–132.e4. <https://doi.org/10.1016/j.cels.2018.05.014>.
- Pyke, R.M., Mellacheruvu, D., Dea, S., Abbott, C.W., Zhang, S.V., Phillips, N.A., Harris, J., Bartha, G., Desai, S., McClory, R., et al. (2021). Precision neoantigen discovery using large-scale immunopeptidomes and composite modeling of MHC peptide presentation. *Mol. Cell. Proteomics* 20, 100111. <https://doi.org/10.1016/j.mcpro.2021.100111>.
- Duan, F., Duitama, J., Al Seesi, S., Ayres, C.M., Corcelli, S.A., Pawashe, A.P., Blanchard, T., McMahon, D., Sidney, J., Sette, A., et al. (2014). Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J. Exp. Med.* 211, 2231–2248. <https://doi.org/10.1084/jem.20141308>.
- Koşaloğlu-Yalçın, Z., Lanka, M., Frentzen, A., Logandha Ramamoorthy Premlal, A.L.R., Sidney, J., Vaughan, K., Greenbaum, J., Robbins, P., Gartner, J., Sette, A., et al. (2018). Predicting T cell recognition of MHC class I restricted neoepitopes. *Oncol Immunology* 7, e1492508. <https://doi.org/10.1080/2162402X.2018.1492508>.
- Bulik-Sullivan, B., Busby, J., Palmer, C.D., Davis, M.J., Murphy, T., Clark, A., Busby, M., Duke, F., Yang, A., Young, L., et al. (2018). Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nat. Biotechnol.* 37, 55–63. <https://doi.org/10.1038/nbt.4313>.
- Richman, L.P., Vonderheide, R.H., and Rech, A.J. (2019). Neoantigen dissimilarity to the self-proteome predicts immunogenicity and response to immune checkpoint blockade. *Cell Syst.* 9, 375–382.e4. <https://doi.org/10.1016/j.cels.2019.08.009>.

23. Schmidt, J., Smith, A.R., Magnin, M., Racle, J., Devlin, J.R., Bobisse, S., Cesbron, J., Bonnet, V., Carmona, S.J., Huber, F., et al. (2021). Prediction of neo-epitope immunogenicity reveals TCR recognition determinants and provides insight into immunoediting. *Cell Rep. Med.* 2, 100194. <https://doi.org/10.1016/j.xcrm.2021.100194>.
24. Bjerregaard, A.M., Nielsen, M., Hadrup, S.R., Szallasi, Z., and Eklund, A.C. (2017). MuPeXI: prediction of neo-epitopes from tumor sequencing data. *Cancer Immunol. Immunother.* 66, 1123–1130. <https://doi.org/10.1007/s00262-017-2001-3>.
25. Hundal, J., Kiwala, S., McMichael, J., Miller, C.A., Xia, H., Wollam, A.T., Liu, C.J., Zhao, S., Feng, Y.Y., Graubert, A.P., et al. (2020). pVACtools: A computational toolkit to identify and visualize cancer neoantigens. *Cancer Immunol. Res.* 8, 409–420. <https://doi.org/10.1158/2326-6066.CIR-19-0401>.
26. Wells, D.K., van Buuren, M.M., Dang, K.K., Hubbard-Lucey, V.M., Sheehan, K.C.F., Campbell, K.M., Lamb, A., Ward, J.P., Sidney, J., Blazquez, A.B., et al. (2020). Key parameters of tumor epitope immunogenicity revealed through a consortium approach improve neoantigen prediction. *Cell* 183, 818–834.e13. <https://doi.org/10.1016/j.cell.2020.09.015>.
27. Parkhurst, M.R., Robbins, P.F., Tran, E., Prickett, T.D., Gartner, J.J., Jia, L., Ivey, G., Li, Y.F., El-Gamil, M., Lalani, A., et al. (2019). Unique neoantigens arise from somatic mutations in patients with gastrointestinal cancers. *Cancer Discov.* 9, 1022–1035. <https://doi.org/10.1158/2159-8290.CD-18-1494>.
28. Gartner, J.J., Parkhurst, M.R., Gros, A., Tran, E., Jafferji, M.S., Copeland, A., Hanada, K.I., Zacharakis, N., Lalani, A., Krishna, S., et al. (2021). A machine learning model for ranking candidate HLA class I neoantigens based on known neopeptides from multiple human tumor types. *Nat. Cancer* 2, 563–574. <https://doi.org/10.1038/s43018-021-00197-6>.
29. Nielsen, M., Lundegaard, C., Lund, O., and Keşmir, C. (2005). The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage. *Immunogenetics* 57, 33–41. <https://doi.org/10.1007/s00251-005-0781-7>.
30. Larsen, M.V., Lundegaard, C., Lambeth, K., Buus, S., Brunak, S., Lund, O., and Nielsen, M. (2005). An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions. *Eur. J. Immunol.* 35, 2295–2303. <https://doi.org/10.1002/eji.200425811>.
31. Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I.D., Sørensen, M., Nielsen, M., and Buus, S. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity. *Eur. J. Immunol.* 42, 1405–1416. <https://doi.org/10.1002/eji.201141774>.
32. Ghorani, E., Rosenthal, R., McGranahan, N., Reading, J.L., Lynch, M., Peggs, K.S., Swanton, C., and Quezada, S.A. (2018). Differential binding affinity of mutated peptides for MHC class I is a predictor of survival in advanced lung cancer and melanoma. *Ann. Oncol.* 29, 271–279. <https://doi.org/10.1093/annonc/mdx687>.
33. Capietto, A.H., Jhunjunwala, S., Pollock, S.B., Lupardus, P., Wong, J., Hänsch, L., Cevallos, J., Chestnut, Y., Fernandez, A., Lounsbury, N., et al. (2020). Mutation position is an important determinant for predicting cancer neoantigens. *J. Exp. Med.* 217. <https://doi.org/10.1084/jem.20190179>.
34. Chowell, D., Krishna, S., Becker, P.D., Cocita, C., Shu, J., Tan, X., Greenberg, P.D., Klavinskis, L.S., Blattman, J.N., and Anderson, K.S. (2015). TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. USA* 112, E1754–E1762. <https://doi.org/10.1073/pnas.1500973112>.
35. Gfeller, D., Schmidt, J., Croce, G., Guillaume, P., Bobisse, S., Genolet, R., Queiroz, L., Cesbron, J., Racle, J., and Harari, A. (2023). Improved predictions of antigen presentation and TCR recognition with MixMHCpred2.2 and PRIME2.0 reveal potent SARS-CoV-2 CD8+ T-cell epitopes. *Cell Syst.* 14, 72–83.e5. <https://doi.org/10.1016/j.cels.2022.12.002>.
36. Müller, M., Gfeller, D., Coukos, G., and Bassani-Sternberg, M. (2017). ‘Hotspots’ of antigen presentation revealed by human leukocyte antigen ligandomics for neoantigen prioritization. *Front. Immunol.* 8, 1367. <https://doi.org/10.3389/fimmu.2017.01367>.
37. Hoyos, D., Zappasodi, R., Schulze, I., Sethna, Z., de Andrade, K.C., Bajorin, D.F., Bandlamudi, C., Callahan, M.K., Funt, S.A., Hadrup, S.R., et al. (2022). Fundamental immune–oncogenicity trade-offs define driver mutation fitness. *Nature* 606, 172–179. <https://doi.org/10.1038/s41586-022-04696-z>.
38. Martínez-Jiménez, F., Muñíos, F., Sentís, I., Deu-Pons, J., Reyes-Salazar, I., Arnedo-Pac, C., Mularoni, L., Pich, O., Bonet, J., Kranas, H., et al. (2020). A compendium of mutational cancer driver genes. *Nat. Rev. Cancer* 20, 555–572. <https://doi.org/10.1038/s41568-020-0290-x>.
39. Rogers, M.F., Shihab, H.A., Gaunt, T.R., and Campbell, C. (2017). CScape: a tool for predicting oncogenic single-point mutations in the cancer genome. *Sci. Rep.* 7, 11597. <https://doi.org/10.1038/s41598-017-11746-4>.
40. Abelin, J.G., Keskin, D.B., Sarkizova, S., Hartigan, C.R., Zhang, W., Sidney, J., Stevens, J., Lane, W., Zhang, G.L., Eisenhaure, T.M., et al. (2017). Mass spectrometry profiling of HLA-associated peptidomes in mono-allelic cells enables more accurate epitope prediction. *Immunity* 46, 315–326. <https://doi.org/10.1016/j.immuni.2017.02.007>.
41. McGranahan, N., and Swanton, C. (2019). Neoantigen quality, not quantity. *Sci. Transl. Med.* 11. <https://doi.org/10.1126/scitranslmed.aax7918>.
42. Bassani-Sternberg, M., Pletscher-Frankild, S., Jensen, L.J., and Mann, M. (2015). Mass spectrometry of human leukocyte antigen class I peptidomes reveals strong effects of protein abundance and turnover on antigen presentation. *Mol. Cell. Proteomics* 14, 658–673. <https://doi.org/10.1074/mcp.M114.042812>.
43. Pearson, H., Daouda, T., Granados, D.P., Durette, C., Bonnell, E., Courcelles, M., Rodenbrock, A., Laverdure, J.P., Côté, C., Mader, S., et al. (2016). MHC class I-associated peptides derive from selective regions of the human genome. *J. Clin. Invest.* 126, 4690–4701. <https://doi.org/10.1172/JCI88590>.
44. Yewdell, J.W., Dersh, D., and Fähræus, R. (2019). Peptide channeling: the key to MHC class I immunosurveillance? *Trends Cell Biol.* 29, 929–939. <https://doi.org/10.1016/j.tcb.2019.09.004>.
45. Bergstra, J., Komer, B., Eliasmith, C., Yamins, D., and Cox, D.D. (2015). Hyperopt: a Python library for model selection and hyperparameter optimization. *Comput. Sci. Disc.* 8, 014008. <https://doi.org/10.1088/1749-4699/8/1/014008>.
46. Cox, D.R. (1958). The regression analysis of binary sequences. *J. R. Stat. Soc. B* 20, 215–232.
47. Chen, T., and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD '16 (Association for Computing Machinery)*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
48. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., and Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems (Curran Associates, Inc.)*.
49. Boser, B.E., Guyon, I.M., and Vapnik, V.N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory COLT '92 (ACM)*, pp. 144–152. <https://doi.org/10.1145/130385.130401>.
50. Shapley, L.S. (1953). A value for n-person games. In *Contributions to the Theory of Games II (Princeton University Press)*.
51. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.)*.
52. Tadros, D.M., Eggenschwiler, S., Racle, J., and Gfeller, D. (2023). The MHC Motif Atlas: a database of MHC binding specificities and ligands. *Nucleic Acids Res.* 51, D428–D437. <https://doi.org/10.1093/nar/gkac965>.

53. Rojas, L.A., Sethna, Z., Soares, K.C., Olcese, C., Pang, N., Patterson, E., Lihm, J., Ceglia, N., Guasp, P., Chu, A., et al. (2023). Personalized RNA neoantigen vaccines stimulate T cells in pancreatic cancer. *Nature* 618, 144–150. <https://doi.org/10.1038/s41586-023-06063-y>.
54. Lim, K.P., and Zainal, N.S. (2021). Monitoring T cells responses mounted by therapeutic cancer vaccines. *Front. Mol. Biosci.* 8, 623475.
55. Łuksza, M., Riaz, N., Makarov, V., Balachandran, V.P., Hellmann, M.D., Solovyov, A., Rizvi, N.A., Merghoub, T., Levine, A.J., Chan, T.A., et al. (2017). A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* 551, 517–520. <https://doi.org/10.1038/nature24473>.
56. Łuksza, M., Sethna, Z.M., Rojas, L.A., Lihm, J., Bravi, B., Elhanati, Y., Soares, K., Amisaki, M., Dobrin, A., Hoyos, D., et al. (2022). Neoantigen quality predicts immunoeediting in survivors of pancreatic cancer. *Nature* 606, 389–395. <https://doi.org/10.1038/s41586-022-04735-9>.
57. Wherry, E.J., and Kurachi, M. (2015). Molecular and cellular insights into T cell exhaustion. *Nat. Rev. Immunol.* 15, 486–499. <https://doi.org/10.1038/nri3862>.
58. Rosenthal, R., Cadieux, E.L., Salgado, R., Bakir, M.A., Moore, D.A., Hiley, C.T., Lund, T., Tanić, M., Reading, J.L., Joshi, K., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. *Nature* 567, 479–485. <https://doi.org/10.1038/s41586-019-1032-7>.
59. Abelin, J.G., Harjanto, D., Malloy, M., Suri, P., Colson, T., Goulding, S.P., Creech, A.L., Serrano, L.R., Nasir, G., Nasrullah, Y., et al. (2019). Defining HLA-II ligand processing and binding rules with mass spectrometry enhances cancer epitope prediction. *Immunity* 51, 766–779.e17. <https://doi.org/10.1016/j.immuni.2019.08.012>.
60. Alspach, E., Lussier, D.M., Miceli, A.P., Kizhvatov, I., DuPage, M., Luoma, A.M., Meng, W., Lichti, C.F., Esaulova, E., Vomund, A.N., et al. (2019). MHC-II neoantigens shape tumour immunity and response to immunotherapy. *Nature* 574, 696–701. <https://doi.org/10.1038/s41586-019-1671-8>.
61. Kreiter, S., Vormehr, M., van de Roemer, N., Diken, M., Löwer, M., Diekmann, J., Boegel, S., Schrörs, B., Vascotto, F., Castle, J.C., et al. (2015). Mutant MHC class II epitopes drive therapeutic immune responses to cancer. *Nature* 520, 692–696. <https://doi.org/10.1038/nature14426>.
62. Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. *Lancet Oncol.* 18, 1009–1021. [https://doi.org/10.1016/S1470-2045\(17\)30516-8](https://doi.org/10.1016/S1470-2045(17)30516-8).
63. Smith, C.C., Selitsky, S.R., Chai, S., Armistead, P.M., Vincent, B.G., and Serody, J.S. (2019). Alternative tumour-specific antigens. *Nat. Rev. Cancer* 19, 465–478. <https://doi.org/10.1038/s41568-019-0162-4>.
64. Kawaguchi, S., and Matsuda, F. (2020). High-definition genomic analysis of HLA genes via comprehensive HLA allele genotyping. *Methods Mol. Biol.* 2131, 31–38. https://doi.org/10.1007/978-1-0716-0389-5_3.
65. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21. <https://doi.org/10.1093/bioinformatics/bts635>.
66. Van der Auwera, G.A., Carneiro, M.O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., et al. (2013). From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* 43, 11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>.
67. Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. *Ann. Oncol.* 26, 64–70. <https://doi.org/10.1093/annonc/mdu479>.
68. Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* 31, 213–219. <https://doi.org/10.1038/nbt.2514>.
69. Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* 22, 568–576. <https://doi.org/10.1101/gr.129684.111>.
70. Barras, D., Ghisoni, E., Chiffelle, J., Orcurto, A., Dagher, J., Fahr, N., Benedetti, F., Crespo, I., Zimmermann, S., Duran, R., et al. (2022). Tumor microenvironment cellular crosstalk predicts response to adoptive TIL therapy in melanoma. Preprint at bioRxiv. <https://doi.org/10.1101/2022.12.23.519261>.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD8-PB	BD Biosciences	Cat# 558207
CD4-FITC	BioLegend	Cat# 317408
CD3-APC Fire 750	BioLegend	Cat# 344840
CD137-PE	Miltenyi Biotec	Cat# 130-119-885
Aqua L/D	ThermoFisher Scientific	Cat# L34966
Chemicals, peptides, and recombinant proteins		
Peptides	Peptide & Tetramer Core Facility, CHUV, Lausanne	https://www.unil.ch/dof/en/home/menuinst/research-platforms/ptcf.html
Peptides	Covalab, Lyon	https://www.covalab.com
Peptides	ThermoFisher Scientific	https://www.thermofisher.com/ch/en/home/life-science/protein-biology/peptides-proteins/custom-peptide-synthesis-services.html
Critical commercial assays		
ELISpot Pro: Human IFN- γ (ALP)	Mabtech	Cat# 3420-2APT-10
Deposited data		
NCI WES, RNASeq	Gartner et al. ²⁸	https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001003.v2.p1
TESLA WES, RNASeq	Wells et al. ²⁶	https://www.synapse.org/#!Synapse:syn21048999
HiTIDE WES, RNAseq	This paper	https://ega-archive.org/studies/EGAS00001007101
NCI training mut-seq immunogenicity table	Gartner et al. ²⁸	https://doi.org/10.35092/yhjc.c.4792338.v2
NCI testing mut-seq immunogenicity table	Gartner et al. ²⁸	https://doi.org/10.35092/yhjc.c.4792338.v2
NCI training neo-pep immunogenicity table	Gartner et al. ²⁸	https://doi.org/10.35092/yhjc.c.4792338.v2
NCI testing neo-pep immunogenicity table	Gartner et al. ²⁸	https://doi.org/10.35092/yhjc.c.4792338.v2
TESLA neo-pep immunogenicity table	Wells et al. ²⁶	https://www.cell.com/cms/10.1016/j.cell.2020.09.015/attachment/34faddbd-0e41-4761-99fa-eba5af540ee2/mmc4.xlsx
TESLA neo-pep immunogenicity table	Wells et al. ²⁶	https://www.cell.com/cms/10.1016/j.cell.2020.09.015/attachment/9b558895-8657-4167-ab23-8cf4472d395b/mmc7.xlsx
HiTIDE neo-pep immunogenicity table	This paper	https://ega-archive.org/studies/EGAS00001007101
Combined data matrix neo-pep	This paper	https://figshare.com/s/147e67dde683fb769908
Combined data matrix mut-seq	This paper	https://figshare.com/s/2462b62bb6630fe2d257
NCBI GRCh37 v37	Genome Reference Consortium	https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/
GENECODE Release 38	The GENECODE Project	https://www.gencodegenes.org/human/
GTEX v8	The GTEx Consortium	https://www.gtexportal.org/
TCGA September 2021	The Cancer Genome Atlas Program	https://www.cancer.gov/tcga
Software and algorithms		
Logistic regression scikit-learn v1.0.2	Cox ⁴⁶	https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
Support vector machine scikit-learn v1.0.2	Boser et al. ⁴⁹	https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

(Continued on next page)

REAGENT or RESOURCE	SOURCE	IDENTIFIER
StratifiedKfold scikit-learn v1.0.2	scikit-learn	https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html
CatBoost v1.0.4	Prokhorenkova et al. ⁴⁸	https://github.com/catboost/catboost
XGBoost v1.5.1	Chen and Guestrin ⁴⁷	https://github.com/dmlc/xgboost
Hyperopt v0.2.7	Bergstra et al. ⁴⁵	http://hyperopt.github.io/hyperopt/
SHAP v0.41.0	Lundberg and Lee ⁵¹	https://github.com/slundberg/shap
HLA-HD v1.4.0	Kawaguchi et al. ⁶⁴	https://www.genome.med.kyoto-u.ac.jp/HLA-HD/
STAR v2.7.8a	Dobin et al. ⁶⁵	https://github.com/alexdobin/STAR/releases
GATK v4.2.0.0	Van der Auwera et al. ⁶⁶	https://gatk.broadinstitute.org/
Sequenza v3.0.0	Favero et al. ⁶⁷	https://cran.r-project.org/web/packages/sequenza/vignettes/sequenza.html
HaplotypeCaller v4.2.0.0	Van der Auwera et al. ⁶⁶	https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller
Mutect1 v1.1.5	Cibulskis et al. ⁶⁸	https://github.com/broadinstitute/mutect/releases/tag/1.1.5
Mutect2 v4.2.0.0	Van der Auwera et al. ⁶⁶	https://gatk.broadinstitute.org/hc/en-us/articles/360056969692-Mutect2
VarScan2 v2.4.4	Koboldt et al. ⁶⁹	https://github.com/dkoboldt/varscan
MixMHCpred v2.1	Bassani-Sternberg et al. ¹⁶	https://github.com/GfellerLab/MixMHCpred
NetMHCpan v4.1	Reynisson et al. ¹⁵	https://services.healthtech.dtu.dk/service.php?NetMHCpan-4.1
PRIME v1.0.1	Schmidt et al. ²³	https://github.com/GfellerLab/PRIME
NetChop v3.1	Nielsen et al. ²⁹	https://services.healthtech.dtu.dk/service.php?NetChop-3.1
NetMHCstabpan v1.0a	Harndahl et al. ³¹	https://services.healthtech.dtu.dk/service.php?NetMHCstabpan-1.0
NetCTLpan v1.1	Larsen et al. ³⁰	https://services.healthtech.dtu.dk/service.php?NetCTLpan-1.1
Cscape July 2017	Rogers et al. ³⁹	http://cscscape.biocompute.org.uk/

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact Michal Bassani-Sternberg (michal.bassani@chuv.ch).

Materials Availability

This study did not generate new unique reagents.

Data and Code Availability

WES and RNASeq raw data files for HiTIDE Patient1-Patient11 have been deposited to the secure EGA repository (<https://ega-archive.org>) in standard fastq file format under EGA ID EGAS00001007101.

Python code can be downloaded from the github repository: <https://github.com/bassanilab/NeoRanking.git>

Datasets for mutations containing the feature values and immunogenicity annotations can be downloaded from: <https://figshare.com/s/2462b62bb6630fe2d257>

Datasets for neo-peptides containing the feature values and immunogenicity annotations can be downloaded from: <https://figshare.com/s/147e67dde683fb769908>

ipMSDB peptides can be downloaded from: <https://figshare.com/s/4f551e68e44d9cbf9ccd>

HLA allotypes for all patients can be downloaded from the figshare repository: <https://figshare.com/s/35361871fdad4d1754d7>

Neo-peptide LR and XGBoost classifiers trained on NCI-train can be downloaded from: <https://figshare.com/s/a000b0990465ab3e9d33>

Mutation LR and XGBoost classifiers trained on NCI-train can be downloaded from: <https://figshare.com/s/3c27fa3b705a74bdfa10>

METHOD DETAILS

Datasets

This study includes three cohort datasets consisting of whole exome (WES) and bulk RNA (RNAseq) sequencing data from healthy and matched cancerous tissues or cell lines, as well as information about the immunogenicity of somatic mutations. The datasets were re-analyzed using a uniform pipeline as indicated below. Because some elements are associated with the mutations and some with predicted neoantigens, throughout the text we used the following naming conventions: mutations (*mut-seq*) refers to 25mer sequences with a mutation in the center, neo-peptides (*neo-pep*) refers to 8-12 (if not otherwise indicated) amino acids (AA) long subsequences of the 25mers containing the mutation, neoantigen if the statement applies to both mutations and neo-peptides. For the different subsets of our data, we used the following naming convention: *DATASET_PEPTIDETYPE*. The *DATASET* is equal to either NCI, NCI-train, NCI-test, TESLA, HiTIDE, or empty if all datasets are addressed (more information about these datasets is indicated below). The *PEPTIDETYPE* is equal to either *mut-seq*, *wt-seq* (wild type (WT) version of *mut-seq*), *neo-pep*, or *wt-pep* (WT version of *neo-pep*). If *PEPTIDETYPE* is omitted, both *mut-seq* and *neo-peptide* are referred to. In the text and figures we referred to different peptide subsets of the data. Datasets consist of not-screened and screened (either screened experimentally or inferred) *neo-peps* or *mut-seqs*. Screened *neo-peps* or *mut-seqs*, can be immunogenic or non-immunogenic (either screened experimentally or not) (Table 1). For example, immunogenic *TESLA_neo-pep* denotes all immunogenic neo-peptides of the TESLA dataset, immunogenic *TESLA* denotes either immunogenic mutations or neo-peptides of the TESLA dataset.

NCI cohort

The largest dataset is a compilation of published datasets from the Rosenberg lab at the Surgery Branch of the National Cancer Institute (NCI).^{1,27,28} It was downloaded from the dbGap repository (<https://dbgap.ncbi.nlm.nih.gov>) under accession number phs001003.v1.p1. The NCI dataset contains mainly skin cutaneous melanoma, colon and rectum adenocarcinoma, lung adenocarcinoma, and breast invasive carcinoma (Data S1). Immunogenicity assay information was obtained from Gartner et al. At the time of download (December 2021), for 112 patients, a cohort defined here as *NCI_mut-seq*, we were able to retrieve matched WES and RNAseq data files as well as results from immunogenicity screens of somatic genomic mutations (non-synonymous single nucleotide variants (SNV), InDels and FSs). Filters based on RNAseq data were generally applied to prioritize mutations prior to the immunogenicity screening.²⁸ In these screens, minigenes encoding the mutations and 12 flanking WT AA on each side were transcribed in vitro and transfected into autologous APCs followed by a co-culture with TIL cultures and IFN- γ ELISPOT immunogenicity measurement. For 80 of the 112 patients, a cohort we defined as *NCI_neo-pep*, other immunogenicity screens were performed to identify the optimal neo-antigenic epitopes and their HLA restrictions. For the mutations, which tested positive in the minigene immunogenicity assay, the top-ranked neo-peptides predicted by NetMHCpan were submitted to immunogenicity assays. Autologous APCs or APCs engineered to express the patient's HLA-I alleles were pulsed with the selected neo-peptides, prior to co-culture with TILs and IFN γ ELISpot readout. Neo-peptides with positive ELISpot readout are called immunogenic. We used the same neo-peptide annotations as provided by Gartner et al.: of the mutations screened by minigenes (positive and negative), the neo-peptides that were not found to be immunogenic in the neo-peptide screens or not screened were considered not immunogenic or negative. All neo-peptides derived from mutations that were not screened by minigenes are annotated as not screened. Following Gartner et al., we divided the *NCI_mut-seq* and *NCI_neo-pep* cohorts into a training set (89 patients for *NCI_mut-seq*, 57 patients for *NCI_neo-pep*) and a test set (23 patients each) (Data S1). The lower number of patients compared to Gartner et al. (70 patients versus 57 for training and 26 patients versus 23 for testing) is due to the missing RNAseq data on dbgap. A description of the pipeline used by Gartner et al. to process the WES and RNAseq data can be found in the supplemental information of Parkhurst et al.²⁷

TESLA cohort

The TESLA consortium²⁶ shared tumor and normal WES and tumor RNAseq data of nine patients with 25 different scientific groups working in the field. The participants used their proprietary software pipelines to call the somatic mutations (non-synonymous SNV or short InDels) and rank the epitopes according to their immunogenicity potential. The TESLA consortium collected these ranked lists and compiled a list of highly or reproducibly ranked neo-peptides for immunogenicity screening, where HLA-I peptide multimers were incubated with subject-matched TILs or peripheral blood mononuclear cells (PBMCs).²⁶ For the first batch consisting of six patients (three melanoma and three NSCLC), 608 neo-peptides (8-14 mers) were screened and 37 of them were found to be immunogenic. For the second batch of another three melanoma patients, a compilation of 310 neo-peptides (9-11 mers) was screened, resulting in four immunogenic epitopes. In total, datasets of eight patients (five with skin cutaneous melanoma, and three with NSCLC) were downloaded and processed as indicated below. We inferred annotations for the mutations from annotations of the neo-peptides, where a mutation is called immunogenic when at least one of its neo-peptides was reported as immunogenic, non-immunogenic when at least one of its neo-peptides was screened but none was found to be not immunogenic, and not screened otherwise. The data was downloaded from the Synapse repository (<https://www.synapse.org/>) under accession number Synapse:syn21048999. We were not able to download the WES and RNAseq data files for one patient. More information can be found in Data S1.

HiTIDE cohort

Samples used in this study were from individuals enrolled in a clinical trial and approved by the institutional regulatory committee at Lausanne University Hospital (Ethics Committee, University Hospital of Lausanne-CHUV). All patients provided informed consent. The HiTIDE in-house dataset consists of WES and RNAseq data and immunogenicity screening results for 11 patients with metastatic melanoma, lung, kidney, and stomach cancers (Data S1), enrolled in phase I clinical trials of TIL ACT (NCT03475134⁷⁰ & NCT04643574). Data included in this study is comprised of both published¹³ and unpublished data.

DNA was extracted with the commercially available DNeasy Blood & Tissue Kit (Qiagen) according to the manufacturers' protocols. RNA was extracted using the Total RNA Isolation RNeasy Mini Kit (Qiagen) according to the manufacturer's protocol (including DNase I (Qiagen) on-column digestion). WES and RNAseq measurements were performed at the Lausanne Genomic Technologies Facility (GTF) (2 patients), at the Health 2030 Genome Center (1 patient) or at Microsynth (8 patients). WES libraries were prepared using the Agilent SureSelect XT Human All exome V5 (GTF) or V7 (Microsynth) kits or Twist Human Core Exome kit in combination with the Twist Human Refseq (Genome Center). RNAseq libraries were prepared using the Illumina Truseq stranded mRNA reagents. WES and RNAseq samples were sequenced on the Illumina HiSeq 2500 (GTF), NextSeq 500/550 (Microsynth) or HiSeq 4000 (Genome Center) systems.

Variant calling and RNAseq analysis was performed as indicated below and neo-peptides were ranked based on the MixMHC and PRIME ranks, RNAseq gene expression and coverage, cancer cell fraction, and ipMSDB scores to select for each patient a set of neo-peptides that were then screened for immunogenicity by interrogating both in-house generated NeoScreen TILs and TIL clinical products (NCT03475134, NCT04643574), using IFN γ Enzyme-Linked ImmunoSpot (ELISpot), as previously described.¹³ Briefly, TILs were challenged with 1 μ g/mL neo-peptides in pre-coated 96-well ELISpot plates (Mabtech). Following 16 to 20hrs co-culture, cells were gently harvested from ELISpot plates, which were then developed according to the manufacturer's instructions and counted with a Bioreader-6000-E (BioSys). Phorbol 12-myristate 13-acetate ionomycin (PMA-iono) (Thermo Fisher Scientific) was used to stimulate TILs as positive control. Positive conditions were defined as those with an average number of spots higher than the counts of the negative control (TILs alone) plus 3 times the standard deviation of the negative. To confirm the recognition of immunogenic neo-peptides by CD8+ T cells, TILs were retrieved from plates, centrifuged and stained to assess the up-regulation of the activation marker 4-1BB on CD8+ or CD4+ T-cells by flow cytometry (Figure S1). For five immunogenic neo-peptides, non-detectable (ND) flow cytometry responses for CD8+ and CD4+ were obtained. In these five instances, the annotation of neo-peptides as immunogenic CD8+ targets was determined by assessing their predicted binding affinity to HLA-I alleles and the absence of CD4+ responses. As for the TESLA dataset, we inferred the mutation immunogenicity annotation from the annotations of the neo-peptides containing the mutation. HiTIDE WES and RNAseq data can be downloaded from the European Genome-phenome Archive (EGA, <https://ega-archive.org>), and can be accessed with the EFA ID EGAS00001007101. Immunogenicity annotations can be found on figshare <https://figshare.com/s/147e67dde683fb769908>.

Features describing mutations and neo-peptides

All datasets were processed with a uniform pipeline, which is an assembly of standard tools to perform variant calling, RNAseq expression and coverage analysis, and HLA typing. A consensus HLA typing was created for each patient from WES and RNAseq data with HLA-HD v1.4.0.⁶⁴ RNAseq reads were aligned to GRCh37 and counted with STAR v 2.7.8a,⁶⁵ where GENCODE v38 annotation was used to define genomic regions for reads counting. If several tumor samples per patient were available, the maximal RNAseq expression and read coverage scores were taken. Prior to variant calling, WES reads were processed following GATK v4.2.0.0⁶⁶ best practices workflow for somatic short variant discovery. Tumor content and copy numbers were estimated by Sequenza v3.0.0⁶⁷ and were used for the calculation of the cancer-cell fractions and clonality. SNVs, InDels, and FS identified by the four variant calling algorithms (HaplotypeCaller,⁶⁶ Mutect2,⁶⁶ Mutect1,⁶⁸ and VarScan2⁶⁹) were merged together and high confidence variants identified by a minimum of two algorithms were selected for downstream analysis. High confidence somatic variants affecting protein-coding genes, following GENCODE v38 annotation, were used to generate tumor-specific mutations (25mers) and class-I neo-peptides (8-12mers). Neo-peptides that also match a WT sequence are discarded. Neo-peptides were then processed by the following tools: MixMHCpred¹⁶ v2.1 and NetMHCpan v4.1¹⁵ for HLA class I binding affinity prediction, PRIME²³ v1.0.1 for antigen presentation and T-cell receptor (TCR) recognition, NetMHCstabpan³¹ v1.0a for HLA class I binding stability, NetChop²⁹ v3.1 for C-terminal proteasomal cleavage, and NetCTLpan³⁰ v1.1 for recognition by the TAP transporter complex. For binding affinity and stability %rank scores were used and differential agretopicity indexes (DAI)^{19,22} were calculated as the $\log(\text{neo-pep \%rank}) - \log(\text{wt-pep \%rank})$. HLA-binding anchor positions of the peptides were calculated based on MixMHCpred sequence motifs. Patient-specific HLA haplotypes were used as input for binding affinity and stability prediction. The oncogenic status of SNVs was predicted by CScape,³⁹ which is a ML tool trained on data from the COSMIC database (<http://cancer.sanger.ac.uk/cosmic/help/gene/analysis>) that predicts the oncogenicity of a mutation based on sequence conservation at the mutation site, as well as genomic, proteomic and structural features. SNV cancer driver status annotations were obtained from the Integrative Onco Genomics (IntOGen) database.³⁸ GTEx v8 (<https://www.gtexportal.org/>) and TCGA (<https://www.cancer.gov/tcga>) databases containing tissue-specific gene expression data, were used for the annotation of mutated gene expression. Sample-specific RNAseq data were additionally used to obtain mutated gene expression and mutation read coverage. ipMSDB³⁶ is an in-house database containing HLA-I and -II binding peptides identified by mass spectrometry from various healthy and cancerous tissues and cell lines with different HLA-I and -II alleles. For this work we used an ipMSDB version with 547,476 unique HLA-I binding peptides, where samples from HiTIDE cohort were excluded. We inferred the presentation of a mutation or neo-peptide based on ipMSDB information available for the corresponding *wt-seq* or *wt-pep*. We classified *wt-pep* matches to ipMSDB into the following subgroups: *EXACT* if *wt-pep* matches exactly a peptide found in an ipMSDB, *INCLUDED* if *wt-pep* is fully included in a longer ipMSDB peptide, *PARTIAL (PARTIAL_MUT)* if the match is partial (including the position of the mutation), and *COVER* if *wt-pep* fully covers a shorter ipMSDB peptide. To infer the presentation of the source protein, we assigned the '*ipMSDB Peptide Count*' score, that represents the number unique peptides for a given protein found in ipMSDB, whereas the '*ipMSDB Peptide Score*' counts the AAs of the unique peptides in ipMSDB that overlap with a query

wt-pep. ‘ipMSDB Mutation Score’ counts the AAs of the unique peptides in ipMSDB that overlap with a query *wt-pep* specifically at the mutation position, and the ‘ipMSDB Peptide Overlap’ calculates the fraction of the query peptide covered by peptides in ipMSDB (0: no overlap, 1: full coverage).

For each patient, we produced both a mutation and a neo-peptide table, each containing one row per mutation and per neo-peptide with the feature scores and annotations as columns. More detailed descriptions of feature scores can be found below and in [Data S3](#). Download links for all available tools used in this study can be found in the [key resources table](#). All the scores in the tables can be calculated using publicly accessible tools that are free to use in academics. We also provide a public version of our ipMSDB database on figshare (<https://figshare.com/s/4f551e68e44d9cbf9ccd>).

Data normalization

Some numerical feature values may have patient or dataset dependent biases, and data normalization was required in order to harmonize them. We experimented with different data transformers from the python scikit-learn preprocessing toolkit: *QuantileTransformer*, *PowerTransformer*, *MinMaxScaler* and *StandardScaler*. The *QuantileTransformer* mapped all values to the interval [0, 1] such that the values are evenly distributed. The *PowerTransformer* applied a transformation to make the values Gaussian-like with zero mean and standard deviation of one. The *MinMaxScaler* transformed the values to the interval [0, 1] by an affine transformation. The distribution of the values remained the same apart from a constant shift and change of scale. The *StandardScaler* transformed the values to their z-score by subtracting the mean and dividing by the standard deviation. Each transformation was applied to all the numerical features of each patient after MV imputation (see *DataTransformer* class).

Conversion of categorical values to numerical

Categorical features can only take on a few predefined values, which cannot be compared or ordered in a natural way. Categorical features were turned into numerical ones by a process called *target encoding*. For each categorical feature i the numerical value of its category c_{ij} was set to the rate r_{ij} of immunogenic mutations or neo-peptides in this category in the training data: $r_{ij} = E(y_k | X_{ki} = c_{ij}) = \frac{\sum_{k=1}^n \mathbb{1}_{X_{ki} = c_{ij}} y_k}{\sum_{k=1}^n \mathbb{1}_{X_{ki} = c_{ij}}}$, where $y_k = 1$ if k is an immunogenic mutation or neo-peptide and 0 otherwise, X is the $n \times m$ (n mutations or neo-peptides, m features) training data matrix i.e. X_{ki} is the value of feature i for mutation or neo-peptide k , and $\mathbb{1}[\cdot]$ is a function that maps a boolean to an integer: $\mathbb{1}[true] = 1$ and $\mathbb{1}[false] = 0$ (see *CatEncoder* class)

Missing value imputation

Missing values (MV) were treated differently for numerical and categorical features. For numerical features, missing values were set either to the maximum or minimum value of that feature in the training data. For example, if the feature represents the TPM values from RNAseq, a missing value means that no reads were detected for this gene and the missing value was replaced by 0 (minimum). On the other hand, if the feature represents the rank of MixMHC binding prediction, a missing value means that none of the patient's allele was predicted to bind to the neo-peptide by the respective binding affinity tool, therefore the missing value was replaced by the maximum binding rank (100). The rules whether to choose minimum or maximum values were set manually for each feature (class *GlobalParameters* in python code). For RNAseq mutation coverage we made the imputed values dependent on the RNAseq percentile value: if RNAseq of a gene is smaller than the 50% percentile for a patient, we imputed a value of 0%, if RNAseq is larger than 50% percentile and smaller than 75% for a patient, we imputed 11%, and for RNAseq values larger than 75% percentile with imputed 23%. For categorical features no imputation was performed but the peptide was assigned to the category for MVs (see *DataTransformer* class).

Subsampling of neo-peptides for training

In order to limit computation time during Hyperopt training on neo-peptides, the size of *NCI-train_neo-pep* was limited by randomly sampling 100,000 non-immunogenic neo-peptides from *NCI-train_neo-pep*, while all immunogenic neo-peptides in *NCI-train_neo-pep* were retained. For the test datasets, no limitation on the size of the data matrices was enforced. Subsampling of non-immunogenic neo-peptides was applied each time when running the Hyperopt optimization. For training on the smaller mutation data *NCI-train_mut-seq*, no limit on the number of non-immunogenic mutations was used.

Rank score

The rank score is defined as: $rank_score = \sum_{k=1}^n e^{-\alpha \cdot (r_k - 1)}$, where r_k is the rank of an immunogenic mutation or neo-peptide k determined by a classifier's *predict_proba* function among all other mutations or neo-peptides in a data set or patient (best rank is 1). The higher the predicted probability of an immunogenic mutation or neo-peptide the lower its rank, and the higher its contribution to *rank_score*. The factor α determines how much weight we gave to low ranking mutations or neo-peptides compared to high ranking ones. For training neo-peptides on screened *NCI-train* we used $\alpha = 0.005$, whereas $\alpha = 0.05$ was used for mutations. $\alpha = 0.02$ was used when calculating the rank scores per patient for both mutations and neo-peptides. We defined the rank score vector as: $rank_score_vec = (e^{-\alpha \cdot (r_1 - 1)}, e^{-\alpha \cdot (r_2 - 1)}, \dots, e^{-\alpha \cdot (r_n - 1)})$, for n immunogenic mutations or neo-peptides.

Classifiers

Logistic regression⁴⁶ (LR) is frequently used to estimate the probabilities of binary responses. It assumes that the log-odds (logarithm of the class probability ratio) is a linear function of the feature values plus an offset resulting in a linear class boundary. LR is a fast and scalable method that is robust to outliers or mislabeled data vectors.

The Support Vector Machine (SVM) classifier was developed by Vapnik and collaborators.⁴⁹ In its basic form, it fits a linear class boundary that maximizes the margin separating the two classes while minimizing the hinge loss of misclassified data vectors, which yields robust classification results. It can easily be extended to fitting nonlinear class boundaries by replacing the linear kernel with

non-linear ones, which makes it a very flexible classifier. Here we use the SVM with linear (SVM-Linear) and radial basis function kernel (SVM-RBF).

XGBoost⁴⁷ is a gradient tree boosting method, where the class labels are approximated by a sum of regression trees. Overfitting is avoided by penalizing trees with many leaves and large values within the leaf nodes. Further, feature and row sampling and shrinkage of regression tree leaf values can be applied. It works on sparse data and allows missing value imputation. The method is made scalable to very large datasets by a series of algorithmic improvements and parallelization as well as GPU usage.

CatBoost⁴⁸ is another recent gradient-boosting classifier. It handles categorical features and their interactions efficiently during training. It prevents overfitting especially for smaller datasets by estimating the gradients on different data than the ones used to estimate the trees by a technique called ordered boosting. Efficient implementation of the training procedure and an implementation tailored for GPU's make it a fast and scalable algorithm.

Classifier training with Hyperopt

Every classifier algorithm has hyper-parameters, which need to be set prior to training. Since these hyper-parameter settings can have a drastic effect on the classification performance it was important to select the best hyperparameters for each classifier. We used the Hyperopt framework, which implements sequential model-based optimization (also known as Bayesian optimization). In this iterative approach, a new guess of hyper-parameter values was calculated based on the results of previously tested values, which calculates for each hyper-parameter the value with the highest expected improvement. The user can define a loss function to be minimized and the range and the prior distributions of possible hyper-parameter values. Hyperparameters of all classifiers can be found on the project's github repository in the OptimizationParams class.

Hyperopt optimization was performed on the screened mutations or neo-peptides of NCI-train if not otherwise stated (when using HiTIDe for training we included both screened and non-screened neo-peptides). In each Hyperopt iteration we used 5-fold cross-validation. Since our datasets are unbalanced (many more non-immunogenic mutations or neo-peptides than immunogenic ones) we used stratified CV (scikit-learn class StratifiedKFold) to make sure that we have a similar proportion of immunogenic mutations or neo-peptides in both training and validation sets. A classifier was trained on the training sets and the performance was evaluated using the *rank_score* of the immunogenic mutations or neo-peptides in the validation set. A trained classifier C ranked an immunogenic mutation or neo-peptide among all non-immunogenic mutations or neo-peptides of the validation set by sorting the $C.predict_proba(x_i)$ probability values (x_i is the feature vector of mutation or neo-peptide i) in decreasing order. Once the optimal hyperparameters were obtained, the classifier C with the optimal hyperparameters was retrained on NCI-train using the $C.fit$ method, and these refitted classifiers were stored in binary files. See python class ClassifierManager for more details. The number of Hyperopt iterations was set to 200, if not otherwise stated. Since Hyperopt training is not deterministic it was always repeated 10 times to estimate the variability.

Classifier evaluation on test datasets

The performance of a classifier C trained on NCI-train was evaluated on the independent test datasets NCI-test, TESLA and HiTIDe. For each patient p in a test dataset, all (screened and not-screened) mutations or neo-peptides i are sorted using the $C.predict_proba(x_i)$ values and the ranks of the immunogenic mutations or neo-peptides were calculated.

Training and testing on NCI-train with leave-one-out CV

When we evaluated the classifier performance on the training sets (NCI-train, HiTIDe) themselves, we used leave-one-out CV. We excluded one patient from the training set in the Hyperopt optimization and used this patient as an independent test set. This procedure was repeated for all patients in the training set.

Voting classifier

The voting classifier simply added the probability $p = predict_proba(x_k)$ of all base classifiers and then performed the ranking. The weighted voting classifier includes a weight w for the probabilities p of classifiers c in two groups $G_{1,2}$: $p = (1 - w) \sum_{c \in G_1} p_c + w \sum_{c \in G_2} p_c$ (see ClassifierManager class on github).

Feature importance

For a given set of m features $F = \{f_1, f_2, \dots, f_m\}$, a feature vector x , and dataset X , the Shapley value⁴¹ $\varphi_f(x)$ tells us the contribution of a feature value x_f to the difference between the actual prediction $g_F(x)$ and the mean prediction $E[g_F(x)|x \in X]$, where the mean is taken over the dataset X : $g_F(x) = E[g_F(x)|x \in X] + \sum_{i=1}^m \varphi_{f_i}(x)$. The calculation of the Shapley value involves a summation over all feature subsets, which is computationally very time consuming, and the python package SHAP⁵¹ implements different sampling approximations of the exact Shapley values (so called Explainers).