Evaluating performance of existing computational models in predicting CD8+ T cell pathogenic epitopes and cancer neoantigens

Paul R. Buckley[1,2,], Ruichong Ma[1,3], Chloe H. Lee[1,2], Isaac Woodhouse, Jeongmin Woo[1,2], Vasily O Tsvetkov[5,6], Dmitrii S Shcherbinin[5,6], Agne Antanaviciute[1,2], Mikhail Shugay[5,6,7], Margarida Rei[1], Alison Simmons[1], Hashem Koohy[1,2,+]*

[1] MRC Human Immunology Unit, Medical Research Council (MRC) Human Immunology Unit, MRC Weatherall Institute of Molecular Medicine (WIMM), John Radcliffe Hospital, University of Oxford, Oxford, UK, Oxford, UK

[2] MRC WIMM Centre for Computational Biology, Medical Research Council (MRC) Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom

[3]Department of Neurosurgery, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, University of Oxford, Oxford, UK, Oxford, UK

[5] Department of Genomics of Adaptive Immunity, Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Moscow, 117997, Russia
[6] Institute of Translational Medicine, Pirogov Russian National Research Medical University, Moscow, 117997, Russia
[7] Center of Life Sciences, Skoltech, Moscow, 143026, Russia
[*] Contributed equally
[+] Corresponding author

Abstract

T cell recognition of a cognate peptide-MHC complex (pMHC) presented on the surface of infected or malignant cells are of utmost importance for mediating robust and long-term immune responses. Accurate predictions of cognate pMHC targets for T Cell Receptors (TCR) would greatly facilitate identification of vaccine targets for both pathogenic diseases as well as personalized cancer immunotherapies. Predicting immunogenic epitopes therefore has been at the centre of intensive research for the past decades but has proven challenging. Although numerous models have been proposed, performance of these models has not been systematically evaluated and their success rate in predicting epitopes for humans has not been measured and compared.

In this study, we curate and present a 'standard-dataset' of class I MHC epitopes for evaluating the performance of immunogenicity classifiers. Using this data, we present a systematic evaluation of performance of several publicly available models which are commonly used to predict $CD8^+$ T cell epitopes in the context of both pathogens and cancers. The benchmarking is performed in two different settings: a pan-HLA setting in which all models are trained and evaluated by a pool of peptides from multiple HLA types, and a HLA-specific setting in which models are trained and evaluated by peptides from HLA-A02:01. In the pan-HLA setting, we observe that the best performing model achieves 75.6% accuracy suggesting considerable room for improvement. In the HLA-specific setting we observe a substantial reduction in

performance with a mean of -11.9%. Our work shows that existing models exhibit suboptimal performance for predicting immunogenic cancer neoantigens. We then further interrogate the underlying problems of model predictions and highlight HLA-bias as a main source of variation amongst other issues associated with the design of models and/or to their training data.

Introduction

An efficient antigen-specific T cell response to external pathogens or endogenous threats relies on tightly regulated processing and presentation of antigenic peptides by class I and II Major Histocompatability Complexes (MHC) and subsequent recognition of the peptide-MHC (pMHC) complex by their cognate CD8+ and/or CD4+ T cells [1,2]. Therefore, immunogenic peptides encompass attributes associated to two sets of features known as peptide-presentation and TCR recognition features [3]. Among these, the features attributed to MHC presentation have shown to be more prominent compared to those attributed to TCR recognition. Examples include heavily conserved anchor positions [4] and enriched motifs associated to distinct HLA types [5]. Indeed, recent cutting-edge models in predicting MHC presentation have shown impressive performance, as exemplified by the widely used netMHCpan[6] and other recently published models [7,8].

The recognition features of peptides on the other hand are highly degenerate due to promiscuity of TCRs imposed by positive and negative thymus selection, to avoid immune blind spots [9]. In addition to peptides' sequence-based recognition features, numerous other factors such as co-stimulation, proliferation, cytotoxicity and cytokine production are evidently associated with immunogenicity [1]. These factors collectively define the magnitude and shape of the immune response and determine whether a response is elicited [10]. Consequently, identification of 'immunogenic' peptides has proven to be more challenging than identification of peptides presented by MHC molecules[11]. To further add to this complexity, the concept of immunogenicity is ambiguously used in the existing literature to represent TCR recognition of cognate pMHC, T cell activation and T cell killing interchangeably. This complex landscape leads to nuances in the various experimental assays evaluating a T cell response, producing noisy data. The lack of a 'true negative' pool of peptides likely contributes to this complexity, as a negative observation only means a peptide failed to elicit a T cell response in a given experiment, perhaps due to the absence of cognate TCRs. This does not mean a peptide is objectively non-immunogenic. In the current study, we refer to immunogenic peptides as those which possess a 'positive' label in data repositories such as the IEDB. This label refers to peptides where a T cell response has been observed by e.g an IFNγ ELISpot assay, although other techniques are commonly used. Further challenges - e.g limited numbers of known TCR-pMHC pairs - in identifying T cell antigens have been recently reviewed by Joglekar et al[12].

Despite these challenges, over the past decade several models have been offered to predict immunogenic peptides, leveraging different correlates of immunogenicity with variable levels of success. As we recently detailed[9] a number of these studies have utilised sequence-based characteristics including amino acid features[13–16], similarity to viral peptides[17], sequence dissimilarity to self [18] and association between peptide immunogenicity and their biophysical properties such as their structural and energy features[19]. Some recent immunogenicity models have begun to put more weight on peptide recognition features by explicitly including TCR information into the model. An example includes *Repitope* by Ogishi et al where they incorporated TCR-peptide contact potential profiling[13]. Each of these models have helped

improve understanding of mechanisms underpinning peptide immunogenicity and hence T cell recognition of pathogens. However, there is currently no systematic evaluation of these models to indicate appropriateness of these tools to various tasks - or their reliability.

In this study, we evaluate and compare the performance of several publicly available models for predicting immunogenicity. Due to the differences in underlying mechanisms that invoke a T cell response, we illustrate the performance of these models in predicting human pathogenic epitopes and cancer neoantigens separately. We thus present a framework in which investigators can decide which immunogenicity classifier is more applicable to their data and research questions. Where model performance shows room for improvement, we aim to shed light on potential problems so these can be mitigated in future models.

In what follows, due to the noise and variability in currently available training datasets, we first present curated data to assist future modelling and benchmarking. We then use this data to evaluate model performance in predicting immunogenicity in two different settings: firstly, in a generic Pan-HLA setting in which the training and test dataset includes various class I HLA types. As each of these models were developed utilising Pan-HLA datasets, observed performance was expected to align with reported performance. Secondly, since T cell recognition of pathogens is highly HLA-restricted, we measured performance of models in an HLA-specific manner, i.e we trained and tested the models only on HLA-A02:01 assessed peptides for which there exists the highest number for a single allele. This allowed us to investigate the role of HLA restriction on predicting immunogenicity and minimise potential HLA bias. In each of these settings, we report performance of predicting immunogenic pathogen epitopes and cancer neoantigens.

Results

Curated datasets for benchmarking and modelling

As described in Methods, we identified 5 models for benchmarking: the *IEDB model, NetTepi, NetTepi-Ext, iPred* and *Repitope* (Table 1). Following this, we gathered and merged data from two publicly available datasets to produce three clean immunogenicity datasets for this study (Fig 1A).

A curated 'Standard Dataset' for the community studying immunogenicity

We first produced a clean dataset with reduced noise for benchmarking and future model development, by merging a recent version of the IEDB[20] with the data available from the *Repitope* package[13]. After data curation (see *Methods*), the resulting '*Standard dataset'* encompasses both class I and II MHC peptides of varying lengths from human hosts (Supplementary Datasheet 1). As immunogenicity predictors for class II MHC peptides are less advanced owing to additional complexities such as considerably longer peptides, we focused on the capacity of models to predict immunogenic class I presented peptides. Thus, the distribution of peptide immunogenicity is shown for the entire class I data and for the most frequent class I alleles (Fig 1B). Here, we observed that 71.81% of the peptides in this dataset are non-immunogenic. For many peptides, we do not have allele-specific information (e.g HLA-class I or HLA-A2 labels), thus during subsequent analysis these will be discarded. For a single allele, we observed a skew towards HLA-A02:01, which encompasses a large proportion of the dataset's immunogenic peptides. With the exception of HLA-B35:01 and

HLA-B08:01, all remaining single alleles comprise mostly non-immunogenic peptides. Therefore, publicly available training datasets for immunogenicity models are composed of large skews in the proportions of immunogenic peptides per allele.

Curated training and test datasets for benchmarking immunogenicity models

We next sought to curate a dataset that was both as close to the original setting of each model, and that could be used to comparatively benchmark across all models. To satisfy these criteria, we required to produce a dataset composed of peptides from multiple HLA alleles, which each model could process. However, universally benchmarking each model is difficult. *NetTepi* and the *IEDB model* are limited to a subset of alleles and the latter is only benchmarked for peptides of 9 amino acids in length (9mers). Of note, *NetTepi* is not applicable to alleles including HLA-B57:01, HLA-A33:01, HLA-A68:01 and HLA-B40:02, limiting the dataset to benchmark across all models. We have also observed variations in the distribution of HLA alleles and the proportion of immunogenic peptides across allele, between the IEDB database and the Chowell dataset[4] used to train *iPred*.

Thus, we curated the '*Pan HLA*' dataset from the previously described '*Standard Dataset*' Supplementary Datasheet 2). To minimise these cross-model variations and satisfy our criteria, we limited ourselves on 9mers from 13 HLA alleles. These 13 HLA alleles were selected as the maximal subset of alleles for which all models are applicable. The resulting *Pan-HLA* dataset comprised 4842 (64.3%) non-immunogenic (negative) and 2679 (35.6%) immunogenic (positive) peptides. As imbalanced datasets are common and may significantly affect model performance, we next explored the imbalances in this Pan-HLA dataset. A large majority of the negative peptides in the data are from *vaccinia virus,* associated with the vaccine *Dryvax* (Fig 1C). Further, many immunogenic peptides are from *polyproteins* (Fig 1D), which are involved in viral replication. Most importantly, we observed a large skew in HLA type towards HLA-A02:01, where immunogenic peptides presented by HLA-A02:01 comprise 22.44% of the entire dataset and a remarkable 63.01% of the total immunogenic peptides (Fig 1E). These large skews in publicly available datasets, in particular for HLA types, may influence model predictions through recognition of prominent sequence patterns.

We additionally observed asymmetries amongst HLA-alleles in each model's original training data. Thus, to examine and minimise effects of HLA-restriction on model performance, we curated a balanced training dataset comprised of 9mers restricted to HLA-A02:01. HLA-A02:01 was selected as it is the most observed allele in the IEDB and has a high frequency in the global population[21]. While evaluating performance using multiple allele-specific datasets would provide valuable insight, we were only able to curate a HLA-A02:01-*HLA-specific* dataset, as the next most frequent allele - HLA-A24:02 - is composed of 1042 negative peptides but only 161 positive which is not enough to reliably train a classifier. Thus, our balanced *HLA01:01-specific* dataset consists of 1168 (50.0%) immunogenic and 1168 (50.0%) non-immunogenic HLA-A02:01 peptides (Supplementary Datasheet 3). To prevent overfitting and ensure results are comparable for each model, benchmarkings were performed using these two datasets under a 10-fold cross-validation scheme (see *Methods*).
Lastly, to evaluate the utility of these models for predicting immunogenic cancer neoantigens, two independent datasets were curated. First, an in-house dataset of glioblastoma (GBM) peptides where neoantigens were functionally validated (Supplementary Datasheet 4). We selected only 9mers, and the resulting dataset is composed of 64 negative peptides (~78.0%) and 18 (~21.9%) confirmed immunogenic neoantigens (Fig 1F). This dataset is referred to as the *'GBM'* dataset. Second, we collected a publicly available dataset of 9mer cancer peptides,

tested in the context of HLA-A02:01[18]. This dataset includes 274 non-immunogenic peptides (~94.1%) and 17 confirmed immunogenic neoantigens (~5.8%) and is referred to as the *Bjerregaard* dataset[18].

Evaluation of performance of models for predicting pathogenic epitopes

Evaluation of model performance in a Pan-HLA setting

We performed 10-fold cross-validation using the Pan-HLA allele dataset. The performance of the models were diagnosed by both ROC-AUC and precision-recall area-under-the-curve (PR-AUC). ROC-AUC is commonly used in machine-learning contexts [6–8,19] while PR-AUC is less susceptible to variation from imbalanced data[22]. Here, we observed a range of ROC-AUCs between 0.65 for the *IEDB model* and 0.76 for *Repitope* (Fig 2A). PR-AUCs ranged between 0.49 for the *IEDB model* and 0.64 for *Repitope* (Fig S1A). With the exception of *iPred,* observed performances were consistent with each model's reported ROC-AUCs. *iPred* recorded the highest number of true positives (1948), but the largest number of false positives (1957). *Repitope* recorded the highest performance on all assessed metrics except Recall, for which *iPred* scored highest (Supplementary Table 1 and S3 for confusion matrices). Overall, PR-AUC scores illustrate considerable room for improvement in the Pan-HLA setting.

Evaluation of model performance in an HLA-specific setting

As previously shown, available immunogenicity datasets are substantially imbalanced by HLA type, with a skew toward HLA-A02:01. Furthermore, it is known that peptides exhibit pronounced presentation features and binders to distinct HLA alleles have different sequence patterns[5] (Fig S2A – HLA-A02:01, S2B – HLA-B35:01). Indeed, antigen presentation and TCR-recognition is tightly HLA-restricted, thus one may expect more reliable predictions when an immunogenicity classifier is trained and tested in an HLA-specific manner. Therefore, to minimise potential variation due to HLA-associated antigen-presentation features [5], we trained and tested all models in an HLA-specific manner. Thus, we performed a 10-fold cross-validation using peptides in our balanced HLA-A02:01 data (the HLA-specific dataset), which for a single allele permits the highest number of equally distributed positive and negative peptides.

Using this HLA-restricted dataset, we observed a considerable reduction in performance as measured by ROC-AUC for all models (Fig 2B), compared with the previous Pan-HLA scenario. Indeed, we observed a mean reduction of 11.9% with standard deviation of 3.84%. *iPred* experienced the largest reduction in performance of 16.2%, while *NetTepi-Ext* experienced the smallest (6.29%) (see Table 2 for all details). *Repitope* recorded the highest ROC-AUC of 0.667. Confusion matrices and detailed performance metrics can be found in S4 and Supplementary Table 2, respectively.

Given this observed reduction in performance, we first sought to exclude the possibility this could be attributed to a reduced number of samples in the HLA-specific dataset compared with the Pan-HLA. Indeed, the HLA-specific dataset has 31.06% the number of peptides that the Pan-HLA dataset comprises. After sub-sampling 30% of the Pan-HLA dataset and performing 10-fold cross-validation, we observed a mean reduction in performance of 0.0518% with 2% standard deviation. The *IEDB model*, *iPred* and *NetTepi-Ext* did not exhibit performance reductions after sub-sampling 30% of the Pan-HLA dataset and *Repitope* and *NetTepi* experienced reductions in performance of 2.54% and 2.10% respectively (Table 3 and S5A-

B). These minimal performance reductions do not seem to fully explain those observed between the Pan-HLA and HLA-specific experiments for any models, including *Repitope* and *NetTepi* where in contrast reductions of 11.7% and 14.6% were observed, respectively. Therefore, only a minute fraction of reduction in performance seems to be associated with a reduction in the number of peptides employed.

Taken together, these data show that each model's performance is substantially reduced after benchmarking with the balanced HLA-specific dataset vs. the Pan-HLA. The extent of this performance reduction cannot be primarily attributed to a reduced number of peptides in this data, which suggests differences in the composition of HLA alleles in these datasets may be at play. We will be investigating these differences and other potential contributing factors to the suboptimal performance of these models later, in section '*Exploring issues underlying model performance'*.
Our findings indicate that the models investigated here have considerable room for improvement, although *Repitope* exhibited consistent performance and superior accuracy. Therefore, *Repitope* is the preferred choice for predicting immunogenic epitopes for infectious disease settings.

<u>Evaluating performance of immunogenicity models in predicting cancer neoantigens</u>

Evaluating performance against cancer neoantigen datasets
A key application for immunogenicity classifiers is to predict immunogenic cancer neoantigens that are capable of activating CTLs for potential use as vaccine targets for personalised cancer immunotherapies[3,23]. The datasets that are used to train these models largely consist of pathogenic peptides with substantial sequence differences from the human proteome. However, mutated peptides often exhibit only a single point mutation from the corresponding wild type self-peptides[24]. This high level of sequence similarity between cancer neoantigens and self-peptides is likely subject to immune tolerance and alongside issues arising from largely pathogenic peptide datasets, makes identification of immunogenic cancer neoantigens more difficult compared to pathogenic epitopes[3]. In turn, presentation of mutated peptides alone appears to be insufficient for neoantigen immunogenicity[25] and consensus is emerging that nuanced discriminative features are required[3,24]. Lastly, the complexity of this prediction problem is compounded by differences in cancer type, between-laboratory and -patient variations and potential biases in neoantigen prediction pipelines[3]. Given the importance - albeit the complexity - of this task, it is vital to evaluate the performance and reliability in predicting immunogenic cancer neoantigens.
Here, we utilised two independent datasets. Firstly our in-house *GBM dataset* (see Fig 1F) and secondly a set of 291 9mer HLA-A02:01 peptides from multiple cancers gathered from a larger dataset by Bjerregaard et al[18], and is thus referred to as the '*Bjerregaard dataset*'. This data contains 17 functionally validated neoantigens. For full details regarding curation of these datasets, see *Methods*. For both testing scenarios, we trained all models using firstly the Pan-HLA allele dataset and secondly the HLA-specific. Both of these test data are substantially imbalanced, thus for this work PR-AUC is the more appropriate diagnostic, although the corresponding ROC curves are shown in Supplementary 6A-D.
In Fig 2C we show PR-AUCs for each model in the Pan-HLA setting when confronted with the GBM data. Performance ranged from 0.22 for *Repitope* to 0.55 for *NetTepi*. For the HLA-specific scenario against this GBM dataset, each model's performance remains poor (Fig 2D). For both training regimens, confusion matrices and other common performance metrics are shown in S7-8 and Tables S3-4 respectively, further demonstrating suboptimal performance.

We next assessed model performance against the Bjerregaard dataset. Each model was trained on the Pan-HLA and HLA-specific datasets (excluding 21 and 15 peptides observed in the test datasets respectively). In both training scenarios, all models performed poorly (Figs 2E-F). For the Pan-HLA scenario, PR-AUCs ranged from 0.05 for *NetTepi* to 0.08 for *Repitope* (Fig 2E). For the HLA-specific scenario, PR-AUCs ranged from 0.06 for *NetTepi* to 0.21 for *Repitope* (Fig 2F). *Repitope's* increased performance after training on the HLA-specific dataset largely a result of improved identification of negative peptides (S9-10).

These findings show that *NetTepi* and *Repitope* are the superior performers against the GBM and Bjerregaard datasets respectively. However, we observed suboptimal performance from all models and a great deal of inconsistency between models when applied to these two datasets.

Exploring model reliability in predicting immunogenic neoantigens

Given the large inconsistencies in model performance against these two independent datasets, we sought to determine whether any model could be deemed reliable for predicting immunogenic neoantigens. Therefore, we evaluated whether any of the models consistently demonstrated predictive capacity superior to that of a random control, by employing a bootstrapping approach. Indeed, we compared each model's benchmarked performance (shown in Figs 2C-F), with an 'estimated control distribution' of PR-AUCs, produced after randomly shuffling the true immunogenicity labels (see *Methods*). To minimise potential confounding factors, we compared the predictive capacity of the models after training with the HLA-specific data.

Density plots contrasting each models' predictive capacity relative to the 'control distribution' of PR-AUCs, are shown in Figs 3A-B against the GBM and Bjeeregaard datasets respectively. Based on this bootstrapping task, *NetTepi* was the only model where its benchmarked PR-AUC (shown by a black dashed line) was significantly higher (Z-score > 2) than the background distribution mean of PR-AUCs (Fig 3A, see Supplementary Table 5 for full Z-scores for all models). A consistent trend was observed after training with the Pan-HLA dataset (S11A). These data demonstrate that when predicting immunogenic GBM neoantigens, *NetTepi* is the only model to exhibit better performance than randomised predictions. In contrast, for the *Bjerregaard* dataset, only *Repitope* and *iPred* achieve a benchmarked PR-AUC that was distinct from the control distribution of PR-AUCs (Fig 3B, Supplementary Table 5), while after Pan-HLA training, no model achieves this (S11B). It is unclear whether these observed inconsistencies in performance result from a data- or model associated artifact or whether these models are only capable of performing in certain biological contexts.

We would expect a reliable model to exhibit consistently accurate performances - which are substantially superior to control distributions of PR-AUCs - against both neoantigen datasets. As we demonstrate, this is not the case for any model. However, we observed that while *NetTepi* performed relatively well with the *GBM* dataset, it showed rather poor performance when applied to *Bjerregaard* dataset and vice versa for *Repitope*. These inconsistent and/or suboptimal performances suggest that none of the assessed models are reliable at predicting immunogenic neoantigens. We therefore set out to further explore these inconsistencies. We observed the GBM peptides are skewed by prediction of binding affinity to HLA-A02:01 by netMHCpan v4.0 prior to functional validation. As *NetTepi* incorporates netMHC binding predictions, we sought to evaluate whether *NetTepi's* superior performance against the GBM peptides could be explained by this skew, reflecting capacity to only predict immunogenicity

in scenarios where the immunogenic peptides exhibit higher MHC binding affinities than their non-immunogenic counterparts. This is important to evaluate because presentation of mutated peptides alone is likely insufficient for immunogenicity[25].

The immunogenicity score returned by *NetTepi* is defined as a linear combination of binding affinity, stability and T cell propensity scores, which we refer to as 'combined immunogenicity score'. During the experiments to evaluate performance against pathogenic epitopes, we observed that *NetTepi's* 'combined immunogenicity score' – was highly correlated with both its binding affinity (Fig S12A) and binding stability scores (Fig S12B). In contrast, *NetTepi's* internal T cell propensity score was not correlated with its combined immunogenicity score, suggesting *NetTepi's* predictions overwhelmingly reflect antigen-presentation, rather than immunogenicity (Fig S12C). Indeed, between immunogenic and non-immunogenic GBM peptides, we observed significant differences in *NetTepi's* predicted binding affinity scores (Fig 3C) and in turn, combined immunogenicity scores (Fig 3D). In contrast, between the immunogenic and non-immunogenic Bjerregaard peptides – for which *NetTepi* performs poorly - we did not observe significant differences in *NetTepi's* predicted binding affinities (Fig 3E), nor combined immunogenicity scores (Fig 3F). This indicates *NetTepi* can only discriminate immunogenic peptides when they exhibit higher binding affinities than non-immunogenic peptides.

In summary, we show that all models perform worse for predicting immunogenic cancer neoantigens than for predicting pathogenic epitopes. Further, we supply substantial evidence that none of the investigated models – which are canonically trained primarily on pathogen epitopes – are reliable predictors of immunogenicity for cancer neoantigens. This exposes a considerable gap for the development of an accurate, publicly-available model to support neoantigen prediction.

<u>Exploring issues underlying model performance</u>

We have thus far illustrated that these immunogenicity classifiers are under-performing for predicting pathogenic epitopes, in particular if one takes into account HLA restriction. Our observations also illustrate that these models are not reliable predictors for cancer neoantigens. In the remainder of this manuscript we are going to explore the underlying issues contributing to the observed suboptimal performance.

Evaluation of HLA skewness in model predictions for pathogenic epitopes after training with Pan-HLA datasets

The demonstrated reduction in model performance in HLA-specific vs. Pan-HLA conditions for pathogenic epitopes is not solely due to a reduction in sample size. Here, we first outline our hypothesis that differences in the composition of these two datasets may explain this performance reduction.

The HLA-A02:01-restricted peptides (evidenced by MHC binding assays or binding presentation) in the Pan-HLA allele dataset comprise >60% of the positive peptides, while remaining alleles primarily comprise negative peptides (Fig 4A and see Fig 1E). Thus, the features extracted from the Pan-HLA dataset may indeed be those associated to MHC presentation - which are known to be more prominent – rather than those related to TCR recognition, resulting in predictions primarily of antigen presentation. As such features e.g conserved anchor residues are more easily discernible, these predictions may inappropriately

inflate performance of models when using an imbalanced Pan-HLA dataset. We reasoned this may lead to the reduction in performance observed when instead training with the HLA-specific dataset – where such features are rendered irrelevant due to their ubiquity – resulting in a reduced but more representative performance in predicting immunogenicity. Thus, we aimed to determine whether these models predict prominent antigen-presentation features when trained using Pan-HLA datasets.

We first sought to evaluate whether prediction scores are skewed toward dominant HLA motifs. Therefore, we gathered the model predictions produced after the Pan-HLA pathogenic cross-validation to determine whether these scores differed between peptides that are labelled as restricted to HLA-A02:01 or not (the latter is referred to as HLA-A02:01-). Consistent with our hypothesis, for all models, we observed that the immunogenicity scores were higher for peptides in the HLA-A02:01+ group, than for remaining alleles (Fig 4B). Next, we selected only the negative peptides where we also observed higher immunogenicity scores for peptides in the HLA-A02:01+ group (Fig 4C). With the exception of *NetTepi-Ext*, these observations stood after we selected only the positive peptides (Fig 4D). Taken together, these data indicate that with the exception of *NetTepi-Ext*, model immunogenicity prediction scores are significantly higher for peptides comprising HLA-A02:01 binding motifs, regardless of true immunogenicity status.

We next sought to establish the extent these models predict HLA-associated antigen-presentation features by utilising the prediction scores generated during the pathogenic Pan-HLA immunogenicity analysis, to instead predict whether a peptide engages with HLA-A02:01 (Fig 4E). Here, we observed remarkable consistency with the ROC-AUCs of the aforementioned experiment (see Fig 2A). This shows that when trained using data where the composition of immunogenic peptides is skewed amongst HLA alleles, models can predict enriched HLA motifs as accurately as they predict immunogenicity. Thus, it is unclear whether the models identify immunogenic peptides or dominant antigen presentation features.

We next minimised the effect of such skewed features with the Pan-HLA dataset, to confirm the extent that models can predict HLA-A02:01 restricted peptides when both immunogenicity and HLA-associated features are more balanced. Thus, we sampled a balanced dataset of 1982 peptides labelled as restricted to HLA-A02:01 (HLA-A02:01^{+} peptides: 50% immunogenic and 50% non-immunogenic) and 1982 HLA-A02:01^{-} peptides (50% immunogenic and 50% non-immunogenic). We then performed 10-fold cross-validation, training models on the aforementioned HLA groups to assess performance in predicting HLA-A02:01+ peptides. Remarkably, for *iPred* and *Repitope*, we observed increased performance here, compared to their performance in predicting immunogenicity using the Pan-HLA dataset (Fig 4F). For *NetTepi*, *IEDB model* and *NetTepi-Ext*, performance was similar to the experiment using Pan-HLA data to predict immunogenicity for pathogenic epitopes presented in Fig 2A. This capacity to predict HLA associated features - together with HLA biases in Pan-HLA datasets - suggests that these models primarily predict antigen-presentation features.

We have shown that when trained on skewed Pan-HLA datasets these models predict peptides associated with alleles prominent in the training data e.g HLA-A02:01 as well as they can predict immunogenicity. Interestingly, both models which do not require an HLA allele for a peptide-immunogenicity prediction (*Repitope* and *iPred*), can more accurately predict peptides which are associated with HLA-A02:01 than immunogenic peptides. Furthermore, we had previously observed that when these enriched HLA-associated antigen-presentation features are attenuated with the HLA-specific dataset, all model performance in predicting

immunogenicity is reduced. These results indicate that when trained on imbalanced Pan-HLA data, model predictions are inappropriately skewed by prominent HLA motifs e.g conserved anchor residues, resulting in predictions that appear to primarily represent antigen-presentation rather than immunogenicity.

In turn, these findings support the hypothesis that model performance may be inappropriately inflated when using Pan-HLA datasets through recognition of more easily discernable skewed HLA features. This is consistent with the reasoning that the reduction in performance observed with the HLA-specific dataset may stem from attenuation of these skewed features, leading to perfomances with this dataset that – while reduced - are more representative of predicting immunogenicity.

Further data-associated complexities

While *Repitope* remains the superior predictor of choice for pathogenic epitopes, we have highlighted concerns with each model in predicting both pathogen epitopes and immunogenic neoantigens. In what follows, we aim to shed light on complexities that may arise from two model's original training data and remaining imbalances in our Pan-HLA dataset.

Challenges associated with iPred's training data

As published, *iPred* uses the Chowell dataset[4] for training, for which we observed considerable HLA imbalance in the data. For HLA-A engaging peptides, we observed that 70.5% are positive, while conversely for HLA-B ~72.4% of the peptides are negative (Fig S13A). For peptides from HLA-B alleles e.g HLA-B57:01, the scores generated by *iPred* during training are distributed towards lower, less immunogenic probabilities whilst those for HLA-A02:01 are distributed toward higher probabilities (Fig S13B). This is unsurprising given the proportion of immunogenic peptides amongst HLA genes and is consistent with the prior finding that *iPred* is susceptible to HLA skewness in the training data.

To evaluate the extent the distributions of scores predicted by *iPred* are appropriate, we utilised an unseen test dataset, equally distributed amongst immunogenic and non-immunogenic HLA-B peptides and peptides presented by HLA-A02:01 (labelled HLA-A). *iPred* was trained canonically and prediction profiles were generated. For immunogenic HLA-B peptides, prediction scores are appropriately distributed around higher immunogenicity probability scores, however scores for HLA-A02:01 peptides are distributed amongst higher immunogenicity probabilities regardless of whether the peptide is immunogenic (S13C). This indicates that large HLA skews in the training data propagates through to inappropriate predictions by *iPred*, where HLA-A02:01 peptides are more likely to be predicted as immunogenic even when this is not the casae. Therefore, when *iPred* is employed, a more balanced training dataset should be used, which may reduce inaccuracy.

Challenges associated with IEDB Model's training data

For the published *IEDB Model* (and consequently for the T cell propensity component of *NetTepi*), the publicly available training data (prior to the redundancy filtering step), includes only one non-immunogenic peptide from humans (Fig 5A). Indeed, to maximise the number of true negatives, their criteria excluded humans as a host for negative sequences. Additionally, this dataset contains a substantial number of peptides presented in mice for murine and human alleles, with the third most frequent allele being H-2-Kd (Fig 5B). Lastly, the proportion of

positive and negative peptides per allele is enormously skewed toward positive sequences, and overall is skewed toward HLA-A02:01 (Fig 5C). Therefore, numerous sources of variation in training data is evident, for which it is unclear how this affects model predictions.

Remaining imbalances in the Pan-HLA dataset

The Pan-HLA dataset was curated to be representative of publicly-available immunogenicity datasets and we have presented key imbalances in this data in Figs 1C, 1D and 1E for the origin pathogenic organism, origin antigen and HLA alleles, respectively. Given that we have found model predictions can be skewed toward dominating HLA groups, we here aim to illustrate further imbalances in this data and how they may skew predictions. After grouping predictions by HLA supertype, minor differences in the distributions of immunogenicity scores by *iPred* and *Repitope* are observed (Fig 5C). This is unsurprising as we have determined that both models are able to accurately extract antigen-presentation features. Next, we grouped peptides by whether their antigenic organism contained the phrase '*virus*' or not. For all models, the distribution of immunogenicity scores for the 'viral' group were significantly lower than the 'non-viral' group (Fig 5D). This corresponds with an imbalance in the dataset shown in Fig 1C, where many non-immunogenic peptides were associated with *vaccinia virus*. Indeed, for each model if scores are grouped by whether the antigenic organism contained the phrase '*vaccinia virus'*, a similar pattern was observed (Fig 5E). Lastly, we grouped scores by whether the source antigen is '*polyprotein*' and observed a preference for higher scores for peptides where this is the case, although, it is difficult to draw a conclusion here as the number of peptides in the 'polyprotein' group is small (Fig S13D). Together, these data provide further evidence that large imbalances in immunogenicity training datasets have considerable potential to skew model predictions.

Discussion

Identification of T cell epitopes has been intensively studied due to its important applications in cancer immunotherapy and infectious disease. Recently, the requirement for accurate prediction of T cell immunogenicity has exploded. Indeed, efficient development of vaccines against pandemic diseases such as COVID-19 is paramount. Accurate identification of vaccine targets from viral genomes would fine-tune this process to help mitigate and/or prevent future pandemics. Multiple studies have employed predictive algorithms to identify potential SARS-CoV-2 targets, ranging from publicly available binding predictors to novel systematic workflows[26–32] . For predictions of immunogenicity it is unclear which publicly-available models are the best-performing, and in particular for cancer peptides, evidence is emerging that available predictors of immunogenicity are suboptimal [3,19,24].

In this study, we evaluated performance for several publicly available models for predicting immunogenicity of pathogenic epitopes and cancer neoantigens. Despite common use, many of these models are unreliable in both settings, with the exception of *Repitope* for predicting immunogenic pathogen epitopes. This model showed consistent performance, recording a ROC-AUC of 0.667 against a balanced, HLA-specific dataset. A notable distinction for *Repitope* is that it is the only model to incorporate TCR information. It additionally permits an intense feature selection to obtain a minimum subset of features (out of nearly 6000) that are associated with immunogenicity. This large number of features may increase susceptibility to overfitting, although the authors claim their choice of algorithm permits 'robustness' against this. Given *Repitope's* superior performance, we recently employed this model for identification of potential crossreactive SARS-CoV-2 epitopes [32].

To assist in the development and benchmarking of immunogenicity predictors, we have presented the *Standard, Pan-HLA* and *HLA-specific* datasets. In accordance with the challenges highlighted by this paper and to provide further clarity, we have supplied a portrait of the imbalances that remain within our Pan-HLA dataset. The current study is limited by these imbalances; while we have aimed to analyse one large source of variation e.g HLA type, others remain. However, these cleaned data are a reflection of the current availability of immunogenicity datasets, thus we have aimed to minimise variation where possible.

At respective times of publication, the *IEDB model* and *NetTepi* were important advances in the field, but studies are beginning to emerge which show these models are suboptimal[19]. For *NetTepi*, we supply evidence this is because its predictions are dominated by antigen-presentation components, suggesting that *NetTepi* is more suited to scenarios where many presented peptides are immunogenic[33].

Furthermore, we found model predictions are skewed in favour of dominant HLA types in imbalanced Pan-HLA training data and it appears that in this scenario most models primarily predict antigen-presentation features. Our analysis that showed these models predict HLA-A02:01 restrictions is based upon the labelled MHC restrictions in the dataset which provides evidence (such as MHC binding assays or predictions) that despite immunogenicity status, a peptide is presented by the corresponding allele. Such evidence is common in immunogenicity datasets, although a potential limitation of this approach is that we have found small numbers of nonimmunogenic peptides with experimentally evidenced MHC binding, which in contrast are not predicted to bind the corresponding allele by netMHCpan v4.1. Further work will be required to determine the cause of this discrepancy.

Together, our findings strongly suggest that using imbalanced Pan-HLA data, models primarily recognise prominent HLA motifs. This finding is supported by Bassani-Sternberg et al [5], in which they accurately determined HLA-I motifs from peptidomics datasets through clustering peptides on their sequence similarities, without using known information about HLA allele. This confirms information extracted from peptide sequences can be deconvoluted to determine HLA motifs. Our findings imply that antigen-presentation features such as HLA binding motifs, as well as TCR recognition components should of course be captured in the next generation of predictors, however they should be deconvoluted to appropriately reflect both sets of components. Indeed, we strongly believe that a *bona fide* immunogenicity predictor for pathogen epitopes should perform well using an HLA-specific mode such as HLA-A02:01 epitope data, of which only *Repitope* is close to achieving.
For predicting immunogenicity of cancer neoantigens, none of the assessed models exhibit reliable performance, exposing an important research avenue to enhance personalised cancer immunotherapies. By investigating performance against two independent datasets, we observed large inconsistencies amongst the same models for predicting immunogenic neoantigens. These datasets comprise very small numbers of immunogenic neoantigens, which may affect predictive capacity of these models. Contributing to these inconsistencies, is likely a combination of differences in cancer type, between-laboratory and patient variations, biases in neoantigen prediction pipelines and scant training data. Indeed, the Bjerregaard dataset[18] spans multiple cancers, while the GBM dataset contains only one cancer type. Additionally, the immunogenic neoantigens amongst the GBM dataset were all confirmed by pMHC tetramer staining in healthy donors, while the Bjerregaard dataset was originally curated from 13 studies, comprising an array of techniques – and patients – used to assess immunogenicity. These types of inconsistencies among cancer neoantigen predictions have been investigated

and reported by Wells et al[3]. It is likely that a tapestry of these effects contribute to the varied performances observed in the current work.
In fact alongside vital contributions, recent studies have also reported challenges in predicting immunogenic neoantigens[3,19,24]. Baker et al[19] presented a structure-based approach, reporting increased performance against other models including the *IEDB model* and *NetTepi*. Despite this success, the authors suggest improvements to their model are necessary before wide adoption of their methodology. Capietto et al[24] recently supplied a framework for how mutation position contributes to neoantigen immunogenicity and propose that poor success rate of neoantigen predictions stems from a lack of methods which capture a variety of features. Indeed, to our knowledge no publicly-available model captures the breadth of features they propose. However, very recently, Wells et al[3] have characterised immunogenic tumor epitopes through an array of features such as half-life, strong MHC binding and foreignness to self etc, culminating in an integrated model of tumor epitope immunogenicity. To our knowledge this model is not publicly available. Despite these contributions, the authors suggest that neoantigen identification approaches are confounded by biases in individual pipelines and that parameters associated with immunogenicity may require calibration for individual use cases. Alongside vital advances, these studies depict key challenges, which taken together with our work, demonstrate that development of a publicly available, integrative, multimodal approach is a vital avenue of research.
A recent study by Croft et al[33], found in vaccinia-virus infected C57BL/6 mice the majority (> 80%) of presented peptides by H-2D^b or H-2D^k are immunogenic. This finding implies that algorithms originally developed for predicting antigen presentation can be used to predict immunogenic vaccinia-virus epitopes in a mouse model. This was evaluated in a recent study by Paul et al[8] where the authors benchmarked 17 models that are used primarily to predict antigen presentation. They showed each of these models can be used to predict immunogenic vaccina-virus epitopes at relatively high level of confidence, albeit with some variation in their predictive power. They observed that netMHCpan 4.0 was the best performing model. Here, while we did observe a preference for higher predicted binding affinity (BA) amongst immunogenic peptides from various pathogens, we did not observe high ROC-AUCs when confronting netMHCpan 4.0 with our pathogenic datasets (see S14 A-D). This was not even true for vaccina-virus peptides in our dataset for which, netMHCpan 4.0 was not able to accurately predict immunogenicity although we did again observe a preference for higher predicted BA amongst these immunogenic peptides (S14 E-F).
Therefore, our study suggests the same principle doesn't seem to be true in other settings and as a result dedicated immunogenicity predictors with capacity to discriminate between peptides' presentation and T cell activation features will benefit the community. Multiple distinctions between humans and mouse models may explain the differences observed in our *in silico* observations where we used human peptides, compared to those functionally validated peptides employed in the work of Paul et al. Firstly, Paul et al evaluated ligands eluted from two murine MHC molecules, whereas there is high variation in human populations, including substantially more diverse HLAs. Indeed, as observed with our GBM dataset (see Supplementary Datasheet 4) and by Strønen et al[34], immunogenicity of peptides can vary depending on the donor. Further, we cannot exclude the possibility that the negative peptides used in our study may not be objectively negative. Therefore, further targeted experimental approaches are required to shed further light on the extent that peptides displayed by humans' class I MHC are immunogenic.

Immunogenicity prediction is conceptualised as predicting likelihood of an immune response to a peptide. However, the magnitude, duration and shape of an immune response is dictated by many factors. Full CTL activation, accumulation and survival is influenced by a wide range

of factors, including co-stimulatory molecules such as CD80/86, CD137, CD70, and OX40, as well as inflammatory cytokines and cytokine mediated proliferation[35–38]. Furthermore, the local chemokine and cytokine environment in secondary lymphoid organs, immunosuppressive cells and cytokines which reduce T cell responses, as well as T cell promiscuity, cross-reactivity, and differences between murine and human immunology all add further complexity to such predictions[9,39–44]. Therefore, a peptide sequence and their physicochemical properties are not the only contributors to effective immunogenicity and these factors are not yet included in existing models. Predicting peptides which can elicit a T cell response is vital for more efficient medicine, however it remains a portion of the requirements for an effective CTL response, convoluting the ability to predict likelihood of an immune response from a peptide sequence.

Here, we sought to benchmark under four scenarios, two comprising data from external pathogenic epitopes and two cancer peptide datasets, to place the study in the context of two vital immunological scenarios. As illustrated by the era of pandemic, accurate prediction of pathogenic peptides which elicit T cell responses would augment future therapeutic development. Further, prediction of patient-specific epitopes and neoantigen immunogenicity could assist in revolutionising cancer outcomes. We have shown that although existing models perform better for predicting pathogen epitopes compared to immunogenic cancer neoantigens, for both applications there is room for improvement. In particular, improving accuracy of neoantigen prediction remains a challenging goal which if accomplished would be highly appreciated by the medical community. Furthermore, we have highlighted systemic concerns and sources of variation in the data used to train immunogenicity predictors and provide evidence these contribute to suboptimal model predictions.

We believe that our work will provide a useful and informative resource in the community that will help design of future experiments for fighting cancers and infectious/pathogenic diseases. Our work highlights an urgent need for approaches capable of modelling various confounding factors of pMHC recognition by TCRs for more accurate prediction of immunogenic epitopes, in particular for cancer neoantigens. We envision that addressing these concerns would create vast potential for highly accurate immunogenicity predictors, which could augment the efficiency of medical research and the response to pandemic disease.

Methods

Models selected for analysis

As a first step, we gathered all publicly available models and excluded those which we were unable to perform comparative assessment (Supplementary Text). After exclusion, we selected 4 models: the *IEDB model, NetTepi iPred,* and *Repitope*. Given that NetTepi was developed in 2014, one NetTepi author recommended an improvement to the algorithm. We constructed an initial implementation for this which we refer to as *NetTepi-Ext.* Therefore, the final list contained 5 models and each is described below and further details are provided in Table 1.

The IEDB model[16] is an immunogenicity predictor available on the IEDB-AR website, which uses sequence-based enrichment of amino acid prevalence in immunogenic versus non-immunogenic peptides to predict immunogenicity. The IEDB model reports a receiver-operating-characteristic area-under-the-curve (ROC-AUC) of 0.66. To improve the accuracy of T cell epitope discovery, *NetTepi* - as a natural extension to the IEDB model – integrates pMHC binding affinity and stability with the *IEDB model* (so-called T cell propensity component) into a weighted sum approach to predict immunogenicity [15]. As recommended to update *NetTepi*, we constructed *NetTepi-Ext*, by replacing stability and affinity scores with netMHCpan 4.0 results. Pogorelyy et al[14] interrogated a database of immunogenic and non-immunogenic peptides for physicochemical properties of amino acids to stratify immunogenicity. Their work resulted in development

of a predictor called *iPred* (named after the github repository). *iPred* reports a ROC-AUC of 0.80. Finally, *Repitope* is a framework which leverages physicochemical properties to predict immunogenicity by mimicking thermodynamic interactions between pMHC and public TCRs [45]. *Repitope* reports a ROC-AUC of 0.76.

Data Analysis
All analyses were conducted in R 3.6.3. All visualisations (excluding ROC and PR curves) were produced using either the *ggpubr* or *ggplot2* packages. ROC curves were produced using the *pROC* package and PR curves were produced using the *yardstick* package. Confusion matrices were produced using the *caret* package.

Original model training data acquisition
We obtained the IEDB model training data from the supporting information of Calis et al 2013. This version of the data is prior to their redundancy filtering step and to our knowledge the final training data is not publicly available. The 'Chowell' dataset used to train iPred was obtained from the github repository hosting the classifier: https://github.com/antigenomics/ipred/tree/master/classifier

The *MHCI_Human* training data for the human class I Repitope model was obtained from the R package, hosted at https://github.com/masato-ogishi/Repitope

Standard Dataset
We created the standard dataset by merging a recent version of the IEDB database (accessed 22-05-20) with the data available from the *Repitope* package. Peptides where the 'host-organism' did not contain either the phrases '*human*' or '*homo sapien*' were removed. We systematically scanned '*MHCRestriction*' information and made annotation consistent across different peptides. Of note, the HLA annotation was processed to have a consistent format e.g 'HLA-A*02:01' for peptides where allelic information was available. At this stage, peptides which do not have a specific allele (for example these are labelled as 'HLA class I' or 'HLA-A2') were retained, however models that require a specific HLA allele to make a prediction would not be able to process these peptides. We removed redundant peptides presented on identical MHC alleles. Furthermore, to reduce noise inherent in the dataset, we labelled peptidea having contradictory immunogenicity annotations as 'Positive' i.e immunogenic. Lastly, all peptides were capitalised and peptides containing non-amino acid characters were removed.

Pan-HLA Allele Dataset
From the standard dataset, we selected only peptides of nine amino acids in length and excluded Class II MHC peptides. Collapsed 'MHCRestrictions' columns were separated into individual rows, resulting in one row per allele. Commas were removed from fields for 'MHCRestriction', 'AntigenName', 'AntigenOrganism' and 'HostOrganism' columns. We next sought to retain peptides where specific allele information was available and for alleles where each models were applicable. Indeed, we applied a filter to keep peptides from the following thirteen class I HLA alleles: HLA-A01:01, HLA-A02:01,HLA-A03:01,HLA-A11:01,HLA-A24:02,HLA-A26:01,HLA-B07:02,HLA-B15:01,HLA-B27:05,HLA-B35:01, HLA-B39:01, HLA-B40:01, HLA-B58:01. Next, the '*MHCRestriction*' column was renamed as '*HLA_Allele*'. To ease processing by models, the '*' was removed from rows in the *HLA_Allele* column. Finally, to avoid a peptide appearing in both training and test folds, the dataset was shuffled at random and sampled to select only one observation per peptide-immunogenicity combination. This was required to prevent model overfitting.

HLA-specific (HLA-A02:01) Dataset
We filtered only HLA-A02:01 assessed peptides from the Pan-HLA allele dataset. Lastly, we sampled the highest number of peptides possible for a balance between positive and negative peptides.

GBM Dataset
These peptides were predicted using an in-house neoantigen identification pipeline and subsequently functionally validated. Prior to functional validation, these peptides were filtered to be of netMHCpan priority score > 50. In-vitro priming of healthy donors peripheral blood mononuclear cells (PBMC) and peptide stimulation of patient's autologous PBMC was performed to find peptide specific T cells. Positive peptides were those where peptide specific T cells were identified by pMHC(HLA-A02:01)-tetramer

staining. A peptide titration curve was then performed. To minimise variation in the current study, we additionally filtered for length 9mer.

Bjerregaard Dataset
This pool of peptides were gathered from a large peptide dataset composed of multiple cancers primarily from skin-cutaneous melanomas (SKCM) and non-small-cell lung cancer (NSCLC) patients. These data were curated by Bjerregaard et al [18] from 13 publications, where immunogenicity was evaluated using both healthy donors and autologous cells from cancer patients. The full set of ~1400 potential neoantigens was downloaded from the Supplementary Material from Bjerregaard et al and filtered for HLA-A02:01 presented 9mers, resulting in 291 peptides, 17 of which are reported as immunogenic. These 17 immunogenic neoantigens were identified by an array of methods and the dataset has also been used for testing model performance by Riley et al [19]. We refer to this dataset as the '*Bjerregaard dataset*'.

Retraining the IEDB Model
As input, the IEDB model takes 'log enrichment scores' for amino acids in immunogenic vs non-immunogenic peptides and weights based on positional importance. As presented in Calis et al, per amino acid enrichment is calculated 'as the ratio between the fraction of that amino acid in the immunogenic versus non-immunogenic datasets'. Here, this was performed in R, by calculating the appearance frequency of amino acids in immunogenic and non-immunogenic datasets and dividing the percentage frequency in the immunogenic set by the percentage frequency of the non-immunogenic dataset for each amino acid. The natural logarithm was taken, resulting in the log enrichment scores for each amino acid. To illustrate this, we use the example published in the original paper, which our algorithm reproduces: if Tyr occurs with a frequency of 2.5% in immunogenic and 1.5% in immunogenic peptides, the enrichment in immunogenic peptides is 2.5/1.5 = ~1.7. The natural logarithm of this enrichment is ~0.54. After training occurs to produce these values for each amino acid, these scores were used as input into the original prediction model (in place of their pre-defined log enrichment values) for subsequent testing. Specifically, a python dictionary file was created based on the newly trained log enrichment values and the original model code was altered to read this dictionary into the code. The remaining original code remained the same and is called from within R to generate predictions. For further explanation of the IEDB immunogenicity model, the reader is directed to Calis et al.

Retraining NetTepi
NetTepi integrates 3 metrics: binding affinity using NetMHCcons1.1, stability predictions using NetMHCstab1.0, and the T cell propensity using the IEDB model. To retrain NetTepi for our analysis, we follow the above for retraining the IEDB model. This was used as input to the original NetTepi code, which calculates the integrated predictions for a test dataset. For consistency, instead of using NetTepi's rank or 'Epitopes' binary classification, we compute a threshold using the youden index on the combined score 'Comb' to classify immunogenicity predictions.

Retraining iPred
iPred provides an R function to compute kidera factors based on a set of peptides. A matrix of these ten features for each peptide is used to train a multivariate gaussian mixture model, using expectation maximisation via the '*EMCluster'* R package. Retraining iPred is performed by simply passing a new training dataset to the original model training script, where kidera factors are computed and an EM-based clustering multivariate gaussian model is trained. To predict immunogenicity for one or more test peptides, a function is provided which calculates the kidera features, which are passed to the model to predict probabilities.

Retraining Repitope
All model re-training and testing was performed using the R package '*Repitope'*. To train a model in Repitope, a TCR fragment library is created first from the package's public TCR clonotype set using the CPP_FragmentLibrary function. Subsequently, a data-frame of features is computed for each dataset. The 'Immunogenicity_TrainModels' function then takes as input the features data-frame (for k-fold cv this is subset based on the training data data partition dictated by the fold), the metadata data-frame (which comprises the association of each peptide with a true immunogenicity classification). This function performs pre-processing and trains extremely randomised tree models, and computes immunogenicity scores on the provided training dataset. The Immunogenicity_Predict function predicts immunogenicity scores for an

external dataset – i.e. the test dataset. Five seeds were used for all feature computation and for training each model.

NetTepi-Ext Initial Implementation
Firstly, netMHCpan4.0 was used to generate binding predictions for each model, where the 1-logk50 score is used. The *Tcell* score was computed and the final immunogenicity score *I* was calculated as per below;

$$I = (t \times Tcell + (1 - t) \times netMHCpanScore)$$

where *t* falls in the range of 0-1, and where *Tcell* is the T cell propensity component (extracted from *NetTepi*) and netMHCpanScore is the prediction value from netMHCpan 4.0 with a range of 0-1. Weights were optimised to maximise Receiver-Operating-Characteristic Area Under the Curve (ROC-AUC) for each experiment.

Performance Evaluation
Model performance was evaluated through a combination of ROC-AUC, PR-AUC and balanced accuracy. Other metrics such as F1 score, precision and recall were also calculated. ROC-AUC curves show the performance of a model by perturbing thresholding and visualising the true positive rate (fraction of true positives / all true positives) against the false positive rate (fraction of false positives / all true negatives). Curve information is summarised using the area under the curve. Given a balanced dataset for binary classification (50% each classification), a random, unskilled model will have a ROC-AUC of 0.5, reflecting only the balance in the dataset. In contrast, a perfect model would have a ROC-AUC of 1.0. In a similar fashion, PR curves is a visualisation of model precision and recall (equations are described below) after perturbing thresholds. A perfect model would have a PR-AUC of 1.0, and a no skill classifier would reflect the balance in the data, i.e. using the above example, PR-AUC would be 0.5.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

where 'tp' stands for 'true positives' and fp stands for 'false negatives'.

Pathogenic epitopes: 10-fold Cross Validation Analysis
A set of training/test partitions were created as each data were split into 10 groups using the createFolds function in the R package *caret.* For each fold, each model was trained (as above) on 90% of the dataset and predictions were made for the remaining test dataset. These predictions were compiled, generating a full dataset where each peptide has been used in training, and each peptide has one prediction score generated per model. This is appropriate as we are benchmarking performance and not performing parameter estimation. For each model ROC curves were built using the *pROC* package and PR curves built using *yardstick*. From ROC curves, optimal threshold values for binary classification (Positive or Negative) were generated using the youden index. The youden index uses ROC curves to compute a threshold value which maximises the equation (1-sensitivity+specificity). For each model, the individual computed threshold value was used to classify prediction scores into 'Positive' or 'Negative' sequences and compiled in an additional '*ImmunogenicityPrediction'* column. For each model, confusion matrices were generated using the *confusionMatrix* function in the *caret* R package.

Pathogenic epitopes: Downsampling analysis
For each sample size (N = 10% , 20% , 30%…90%), the Pan-HLA allele dataset was randomly sampled. Following, for each sample size, training and test partitions were created using the *createFolds* function of *caret* as above. 10-fold cross-validation was performed for each model, under each sample size. For example, for the sample size 10% - corresponding to 10% the size of the entire Pan-HLA dataset - the training data partition for fold 1 (out of ten) contained 689 peptides, whereas the corresponding test set contained 61.

Cancer neoantigens: GBM Analysis

Each model was trained separately on the Pan-HLA and HLA-specific datasets and tested separately against the GBM dataset. PR curves and ROC curves were generated using *yardstick* and *pROC* packages respectively.

Cancer neoantigens: Bjerregaard Analysis
Each model was trained on the Pan-HLA and HLA-specific dataset and tested against the 291 candidate neoantigens. PR curves and ROC curves were generated using *yardstick* and *pROC* packages respectively.

Cancer neoantigens: generating 'shuffled PR-AUCs' to represent unskilled models
For each training / test scenario (Pan-HLA / HLA-specific against GBM / Bjerregaard data), we separately gathered the generated prediction scores. These scores were not altered, however we then randomly shuffled the immunogenicity label for each peptide. The distribution of immunogenicity labels was preserved. We then calculated PR-AUCs with the original model prediction scores against the newly shuffled immunogenicity labels. This was repeated 1000 times, resulting in a distribution of 'shuffled PR-AUCs', reflecting performance of a distribution of random, unskilled models. We then compared the 'benchmarked PR-AUCs' against this distribution of 'shuffled PR-AUCs', to provide a representation of how superior the model performance observed during benchmarking, is to a distribution of unskilled models. Z-scores were calculated by subtracting the mean of the 'shuffled PR-AUCs' from the benchmarked PR-AUCs, and then dividing by the standard deviation of the 'shuffled PR-AUCs'.

HLA imbalance in model predictions for pathogenic epitopes: Exploratory Analysis
We gathered the prediction scores generated by the Pan-HLA pathogenic epitopes experiment to evaluate performance in predicting immunogenicity. If a peptide was observed to bind HLA-A02:01 in this data, it was labelled as 'HLA-A02:01 Positive', while the remaining peptides were labelled 'HLA-A02:01 negative'.

HLA imbalance in model predictions for pathogenic epitopes: Predicting HLA-A02:01+ peptides from Pan-HLA Immunogenicity Prediction Scores
Prediction scores generated by each model after the Pan-HLA pathogenic epitopes experiment to evaluate performance in predicting immunogenicity were gathered. Instead of producing ROC curves against the 'Immunogenicity' column however, ROC curves were produced against the binary classification of whether the peptide engages with HLA-A02:01 or not (i.e. HLA-A02:01+/-).

HLA imbalance in model predictions for pathogenic epitopes: Predicting HLA-A02:01+ peptides after training on HLA-A02:01+/- groupings
The data was grouped by whether the peptide is presented by HLA-A02:01 or not and by immunogenicity status. We then performed balanced sampling, resulting in 1982 HLA-A02:01+ and 1982 HLA-A02:01- peptides, with 50% immunogenic peptides per group. Employing 10-fold cross-validation, data were partitioned into training and test sets, and training occurred on the HLA-A02:01 groupings, and models were tested against the HLA-A02:01 groupings. ROC curves were visualised using the *pROC* package.

Seqlogos for HLA-A02:01 and HLA-B35:01
Positive binders were gathered for HLA-A02:01 and HLA-B35:01 and the R package *PepTools* in combination with *ggseqlogo* were used to generate the seqlogo visualisations.

Exploring the composition of NetTepi's combined immunogenicity predictions
NetTepi's predictions were gathered after the pathogenic 10-fold-cross-validation experiment using the HLA-specific dataset. Scatterplots and pearson R are calculated using the *ggpubr* package.

Exploring whether netMHCpan 4.0 can predict immunogenicity for human tested peptides
NetMHCpan 4.0 BA mode was used with default parameters and prediction scores were generated for the Pan-HLA allele dataset and the HLA-specific. The resulting 1-log50k scores were used to construct ROC-Curves and visualise violin plots. To explore whether netMHCpan 4.0 could discriminate immunogenicity for vaccinia-virus peptides, all peptides with 'AntigenOrganism'annotations which contained the phrase '*Vaccinia'* or '*vaccinia'* were selected from the standard-dataset, and only peptides where the assay to evaluate T cell response was 'IFNg release' were kept (for consistency with Croft et al).

Acknowledgements

H.K and P.B, I.W were funded by MRC HIU core grant MC_UU_00008. A.S was funded by a Wellcome Investigator Award (219523/Z/19/Z), the UK Medical Research Council, NIHR, awards from Bristol-Myers Squibb and UCB. A.S is an NIHR Senior Investigator and acknowledges support from the Oxford NIHR Biomedical Research Centre. C.H.L was funded by UK National Institute of Health Research (NIHR). A.A was supported by Bristol Myers Squibb. D.S.S and M.S are supported by Ministry of Science and Higher Education of the Russian Federation grant No 075-15-2019-1789. MR was funded by CRUK Programme Grant awarded to Prof Cerundolo-C399/A2291 MRC HIU Core Grant MC_UU_00008/1. RM was funded by CRUK clinical research training fellowship.

Conflicts of Interest

No conflict of interest

Ethics Statement

References

1. Zhang, N. & Bevan, M. J. CD8+ T Cells: Foot Soldiers of the Immune System. *Immunity* **35**, 161–168 (2011).
2. Pennock, N. D. *et al.* T cell responses: naive to memory and everything in between. *AJP Adv. Physiol. Educ.* **37**, 273–283 (2013).
3. Wells, D. K. *et al.* Key Parameters of Tumor Epitope Immunogenicity Revealed Through a Consortium Approach Improve Neoantigen Prediction. *Cell* **0**, (2020).
4. Chowell, D. *et al.* TCR contact residue hydrophobicity is a hallmark of immunogenic CD8+ T cell epitopes. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E1754–E1762 (2015).
5. Bassani-Sternberg, M. & Gfeller, D. Unsupervised HLA Peptidome Deconvolution Improves Ligand Prediction Accuracy and Predicts Cooperative Effects in Peptide–HLA Interactions. *J. Immunol.* **197**, 2492–2499 (2016).
6. Reynisson, B., Alvarez, B., Paul, S., Peters, B. & Nielsen, M. NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* **48**, W449–W454 (2020).
7. Mei, S. *et al.* A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief. Bioinform.* **21**, 1119–1135 (2020).
8. Paul, S. *et al.* Benchmarking predictions of MHC class I restricted T cell epitopes in a comprehensively studied model system. *PLoS Comput. Biol.* **16**, e1007757 (2020).
9. Lee, C. H. *et al.* Predicting Cross-Reactivity and Antigen Specificity of T Cell Receptors. *Front. Immunol.* **11**, 565096 (2020).
10. Paludan, S. R., Pradeu, T., Masters, S. L. & Mogensen, T. H. Constitutive immune mechanisms: mediators of host defence and immune regulation. *Nature Reviews Immunology* 1–14 (2020) doi:10.1038/s41577-020-0391-5.
11. Gfeller, D. & Bassani-Sternberg, M. Predicting antigen presentation-What could we learn from a million peptides? *Frontiers in Immunology* vol. 9 1 (2018).
12. Joglekar, A. V. & Li, G. T cell antigen discovery. *Nature Methods* 1–8 (2020) doi:10.1038/s41592-020-0867-z.
13. Ogishi, M. & Yotsuyanagi, H. Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* **10**, 827 (2019).
14. Pogorelyy, M. V. *et al.* Exploring the pre-immune landscape of antigen-specific T cells. *Genome Med.* **10**, 1–14 (2018).
15. Trolle, T. & Nielsen, M. NetTepi: An integrated method for the prediction of T cell epitopes. *Immunogenetics* **66**, 449–456 (2014).
16. Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, (2013).
17. Luksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
18. Bjerregaard, A. M. *et al.* An analysis of natural T cell responses to predicted tumor neoepitopes. *Front. Immunol.* **8**, 1566 (2017).

19. Riley, T. P. *et al.* Structure based prediction of neoantigen immunogenicity. *Front. Immunol.* **10**, (2019).
20. Dhanda, S. K. *et al.* IEDB-AR: immune epitope database - analysis resource in 2019. *Nucleic Acids Res.* **47**, W502–W506 (2019).
21. Gonzalez-Galarza, F. F. *et al.* Allele frequency net database (AFND) 2020 update: Gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Res.* **48**, D783–D788 (2020).
22. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, (2015).
23. Roudko, V., Greenbaum, B. & Bhardwaj, N. Computational Prediction and Validation of Tumor-Associated Neoantigens. *Frontiers in Immunology* vol. 11 27 (2020).
24. Capietto, A.-H. *et al.* Mutation position is an important determinant for predicting cancer neoantigens. (2020) doi:10.1084/jem.20190179.
25. Yadav, M. *et al.* Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, (2014).
26. La Porta, C. A. M. & Zapperi, S. Estimating the Binding of Sars-CoV-2 Peptides to HLA Class I in Human Subpopulations Using Artificial Neural Networks. *Cell Syst.* 1–6 (2020) doi:10.1016/j.cels.2020.08.011.
27. Gao, A. *et al.* Predicting the Immunogenicity of T cell epitopes: From HIV to SARS-CoV-2. *bioRxiv Prepr. Serv. Biol.* (2020) doi:10.1101/2020.05.14.095885.
28. Grifoni, A. *et al.* A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. *Cell Host Microbe* **27**, 671-680.e2 (2020).
29. Crooke, S. N., Ovsyannikova, I. G., Kennedy, R. B. & Poland, G. A. Immunoinformatic identification of B cell and T cell epitopes in the SARS-CoV-2 proteome. *Sci. Rep.* **10**, 14179 (2020).
30. Kiyotani, K., Toyoshima, Y., Nemoto, K. & Nakamura, Y. Bioinformatic prediction of potential T cell epitopes for SARS-Cov-2. *J. Hum. Genet.* **65**, 569–575 (2020).
31. Lee, C. H. *et al.* In silico identification of vaccine targets for 2019-nCoV [version 2; peer review: 3 approved]. *F1000Research* **9**, 1–10 (2020).
32. Lee, C. H. *et al.* Potential CD8+ T Cell Cross-Reactivity Against SARS-CoV-2 Conferred by Other Coronavirus Strains. *Front. Immunol.* **11**, 2878 (2020).
33. Croft, N. P. *et al.* Most viral peptides displayed by class I MHC on infected cells are immunogenic. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 3112–3117 (2019).
34. Strønen, E. *et al.* Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science (80-.).* **352**, 1337–1341 (2016).
35. Kim, M. T. & Harty, J. T. Impact of inflammatory cytokines on effector and memory CD8+ T cells. *Front. Immunol.* **5**, (2014).
36. Pardigon, N. *et al. Role of co-stimulation in CD8+ T cell activation*. *International Immunology* vol. 10 (1998).
37. Bansal-Pakala, P., Halteman, B. S., Cheng, M. H.-Y. & Croft, M. Costimulation of CD8 T Cell Responses by OX40. *J. Immunol.* **172**, 4821–4825 (2004).
38. Croft, M., So, T., Duan, W. & Soroosh, P. The significance of OX40 and OX40L to T-cell biology and immune disease. *Immunol. Rev.* **229**, 173–191 (2009).
39. Lian, J. & Luster, A. D. Chemokine-guided cell positioning in the lymph node orchestrates the generation of adaptive immune responses. *Curr. Opin. Cell Biol.* **36**, 1–6 (2015).
40. Gajdasik, D. W. *et al.* Th1 responses in vivo require cell-specific provision of OX40L dictated by environmental cues. *Nat. Commun.* **11**, 1–15 (2020).
41. Monu, N. R. & Frey, A. B. Myeloid-derived suppressor cells and anti-tumor T cells: A complex relationship. *Immunol. Invest.* **41**, 595–613 (2012).
42. Law, A. M. K., Valdes-Mora, F. & Gallego-Ortega, D. Myeloid-Derived Suppressor Cells as a Therapeutic Target for Cancer. *Cells* **9**, 561 (2020).
43. Mason, D. A very high level of crossreactivity is an essential feature of the T- cell receptor. *Immunol. Today* **19**, 395–404 (1998).
44. Mestas, J. & Hughes, C. C. W. Of mice and not men: differences between mouse and human immunology. *J. Immunol.* **172**, 2731–2738 (2004).
45. Ogishi, M. & Yotsuyanagi, H. Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* **10**, (2019).

Figures

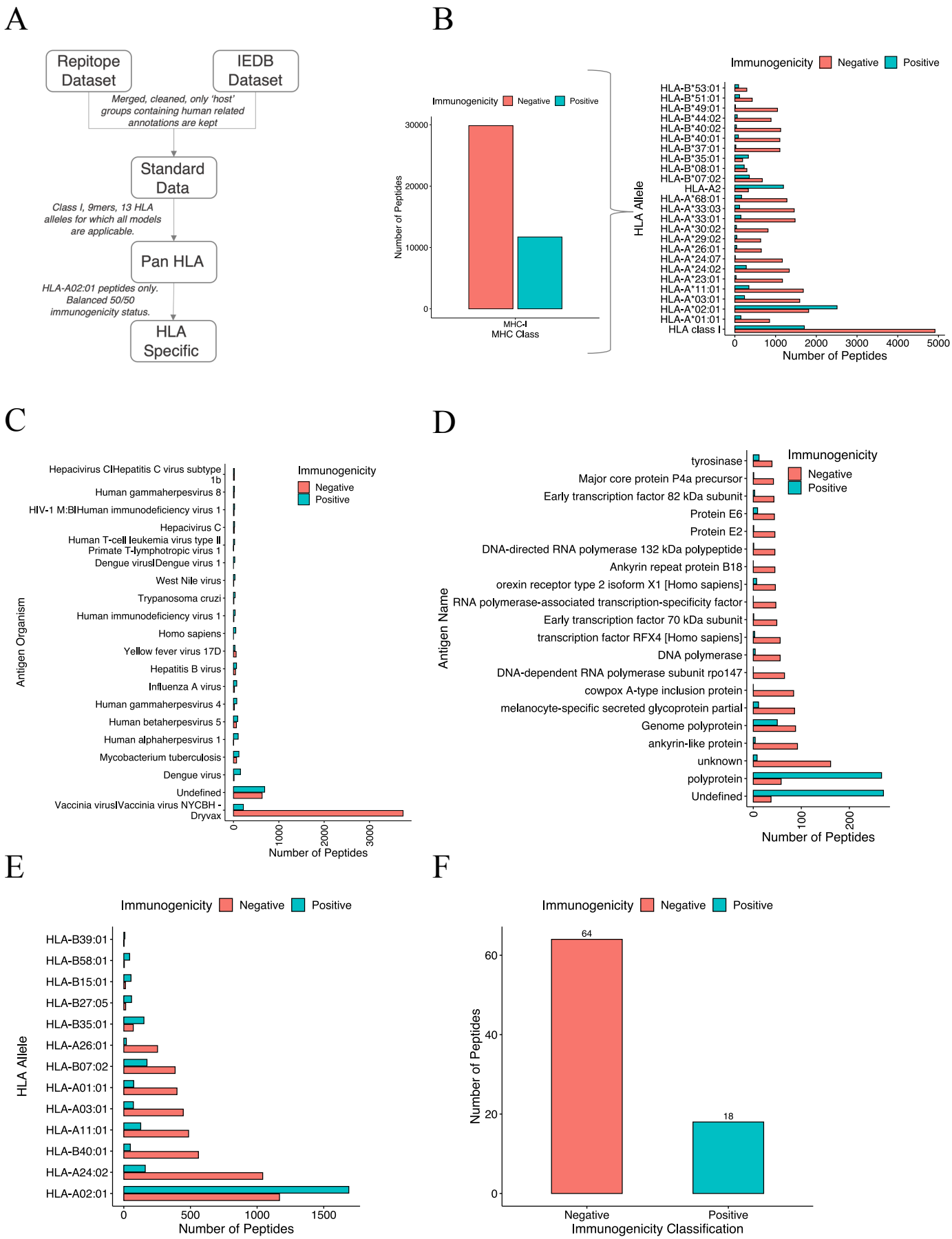


Figure 1. Characteristics of the Gold Standard, Pan-HLA allele and GBM Datasets. A) Flowchart showing the data curation process. B-F) Bar charts showing B) Distribution of immunogenicity status for class I peptides in the Standard Data and the corresponding distribution of immunogenicity status for the peptides amongst the most frequent class I alleles. C) Distribution of immunogenic and non-immunogenic peptides per antigen-organisms in the Pan-HLA dataset. D) Distribution of immunogenic and non-immunogenic peptides per source antigen for the Pan-HLA dataset. E) Distribution of immunogenic and non-immunogenic peptides amongst HLA alleles in the 'Pan HLA-Allele' dataset. F) Distribution of immunogenic and non-immunogenic peptides amongst the GBM dataset.

Figures

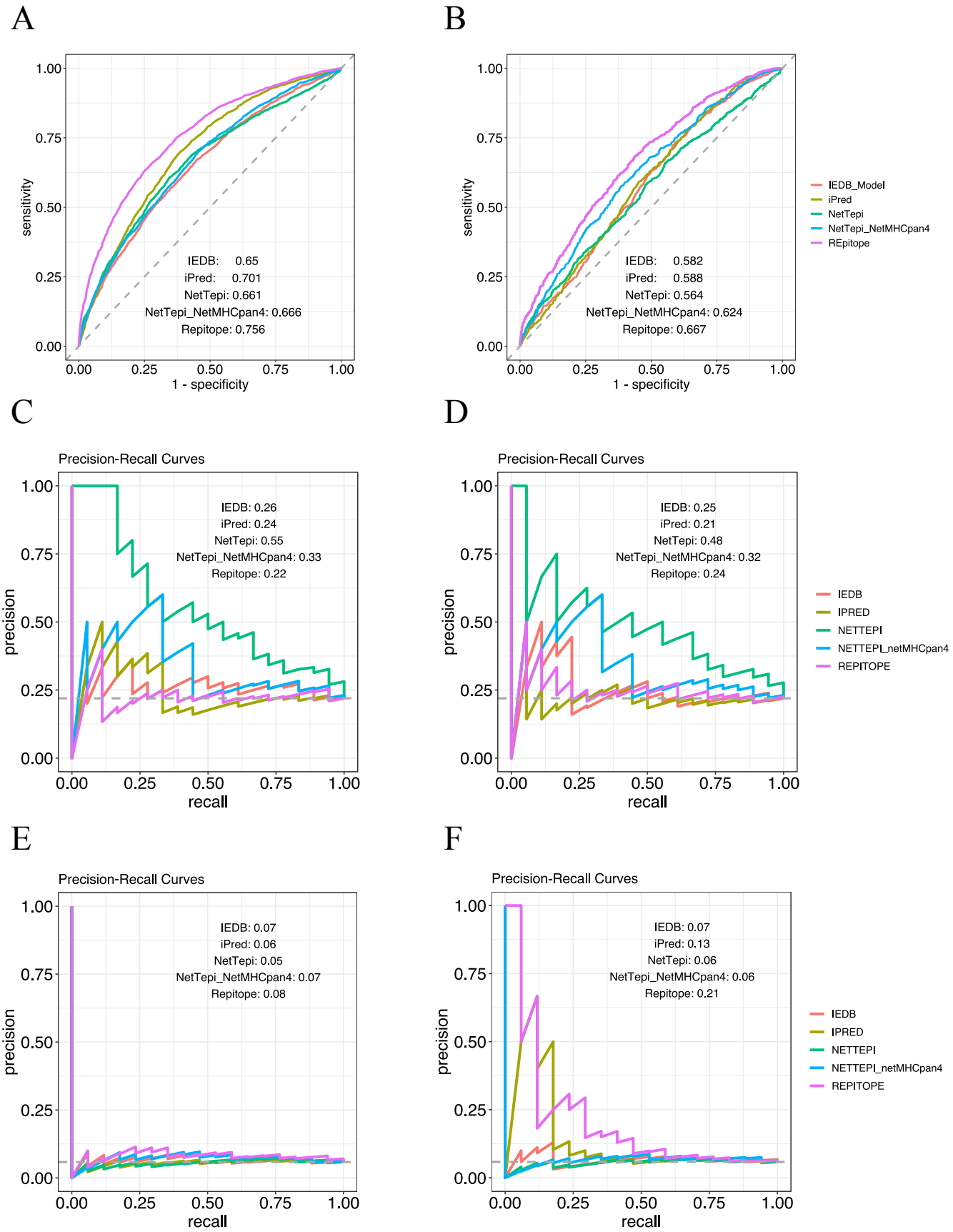


Figure 2: Evaluating model performance against pathogenic epitopes and cancer neoantigens: A) ROC curve showing the performance of the benchmarked models in the pathogenic setting after a 10-fold cross-validation experiment using the 'Pan HLA Allele dataset'. B) ROC curve showing the reduction in performance of the benchmarked models in the pathogenic setting after a 10-fold cross-validation experiment using the HLA-specific dataset. C-E) Precision-recall AUC plots showing model performance in the cancer neoantigen setting C) against the GBM dataset after training each model on the Pan-HLA-allele dataset. D) against the GBM dataset after training each model on the HLA-Specific dataset. E) against the Bjerregaard dataset after training models on the Pan-HLA dataset. F) against the Bjerregaard dataset after training models on the HLA-specific dataset.

Figures

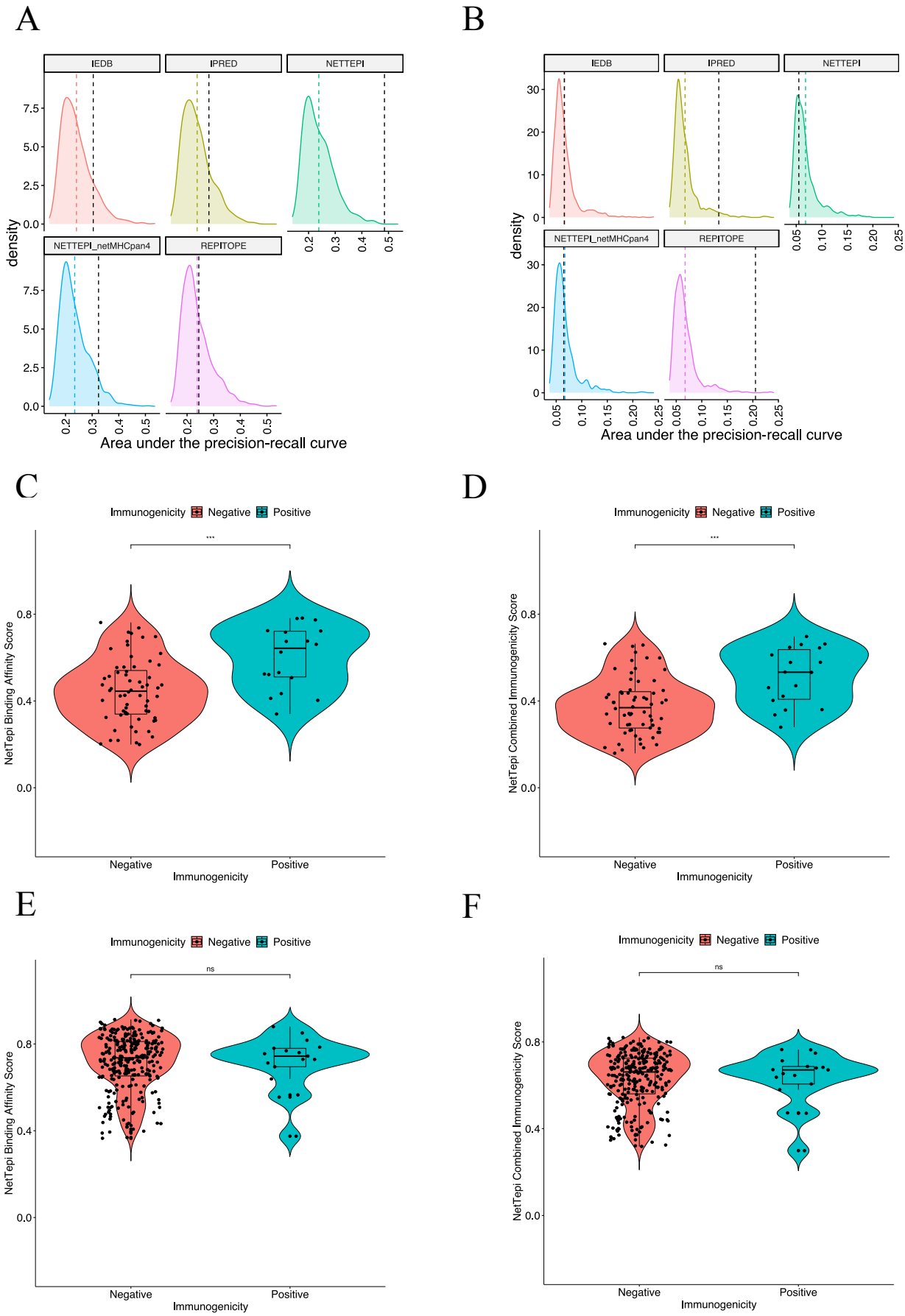


Figure 3: Each model is unreliable for predicting immunogenic neoantigens: A-B) Density plots contrasting the benchmarked PR-AUC scores (blacked dashed line) with the estimated control distribution of PR-AUCs generated after randomly shuffling immunogenicity labels when testing against the A) GBM dataset after training on the HLA-specific dataset. B) Bjeeregaard dataset after training on the HLA-specific dataset. The coloured dashed lines show the mean of the 'estimated control distribution' of PR-AUCs. C) Violin plot of NetTepi's binding affinity scores for the GBM dataset. D) Violin plot of NetTepi's combined immunogenicity scores for GBM data. E) Violin plot of NetTepi's binding affinity scores for the Bjerregaard neoantigen dataset. F) Violin plot of NetTepi's combined immunogenicity scores for the Bjerregaard neoantigen dataset.

Figures

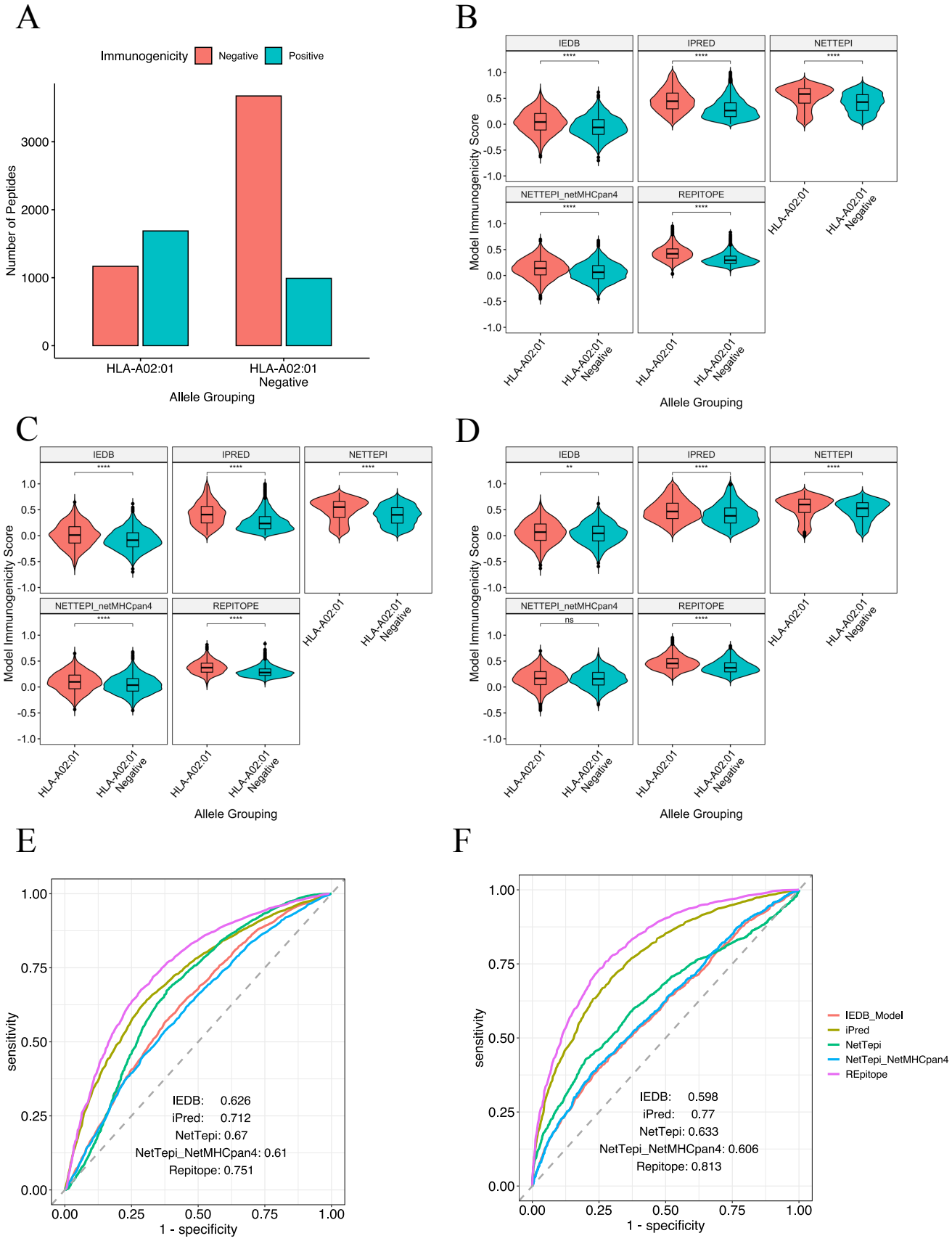


Figure 4: Evaluating effects of HLA skewness on model predictions for pathogenic epitopes: A) Barchart showing the distribution of immunogenic and non-immunogenic peptides for peptides that are restricted to HLA-A02:01 and those which are not (HLA-A02:01 Negative). B-D) Violin plots showing the distribution of each model's predicted immunogenicity score, grouped by whether the peptide is restricted to HLA-A02:01 or not after selecting B) all peptides in the Pan-HLA dataset, C) only the negative peptides in the Pan-HLA dataset, D) only the positive peptides in the Pan-HLA dataset. E-F) ROC-curves of each model after: E) training all models on the immunogenicity groupings but instead used the cross-validation scores to predict HLA-A02:01+ peptides. F) assessing performance of models to predict HLA-A02:01+ peptides after training on a balanced dataset of HLA-A02:01+/-, immunogenic and non-immunogenic peptides

Figures

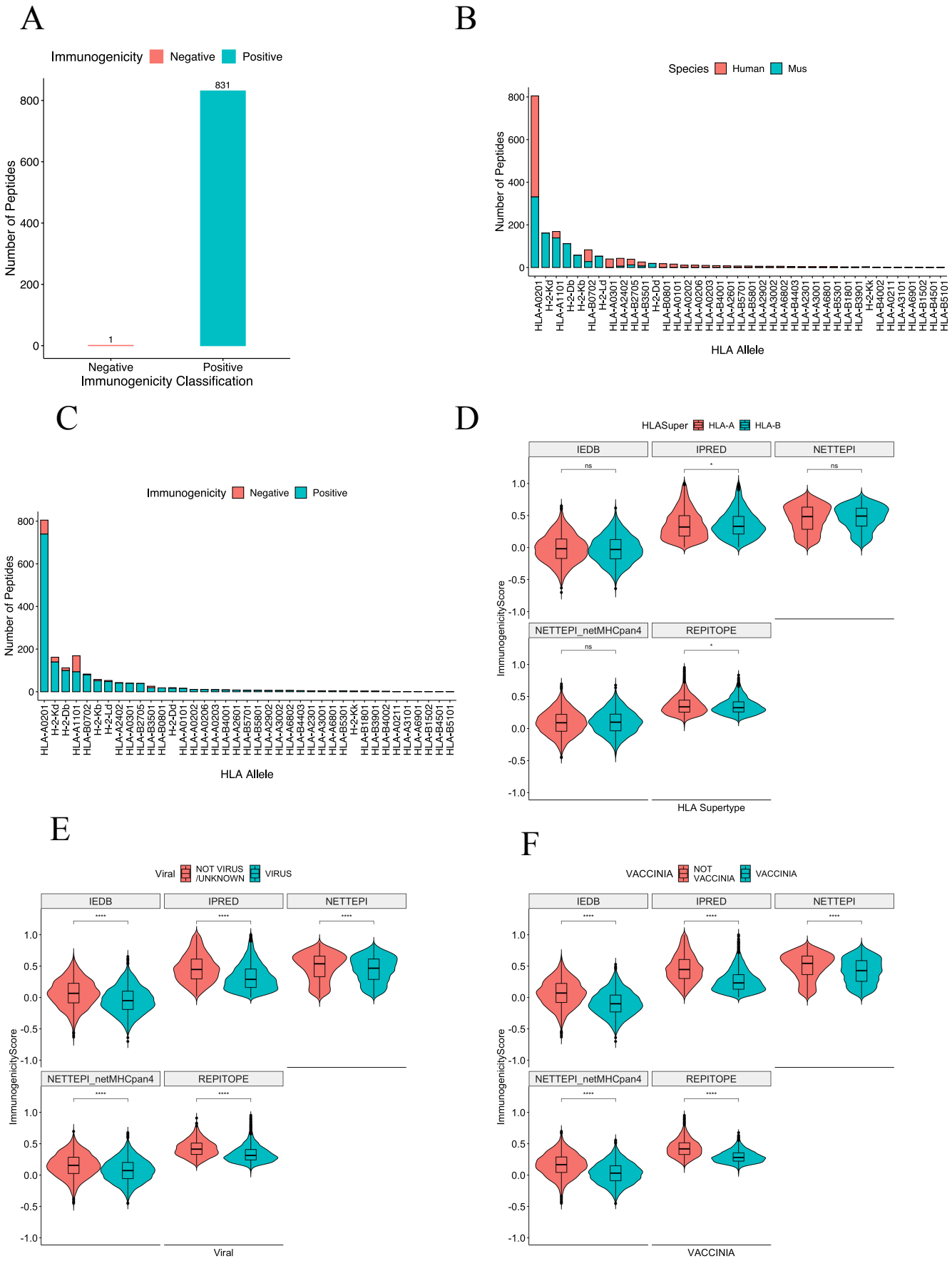


Figure 5: Further data-associated complexities in predicting immunogenicity: A) Bar-plot showing the skewed distribution of human host immunogenic and non-immunogenic peptides amongst the IEDB model training data. B) Bar-plot of the distribution of IEDB model training data, showing large amounts of murine peptides and HLA A02:01 as the dominant allele. C) Bar-plot showing the distribution of immunogenicity classification per allele. D-F) Violin plots showing distributions of scores generated by each model after the Pan-HLA pathogenic epitopes cross-validation experiment: D) labelled by HLA supertype, (E) labelled by whether the peptide's source organism is a virus, (F) labelled by whether the peptide source organism is vaccinia virus.

Model	Features and Calculations	Training Data	Language	Input Data	HLA Restriction	Reported Performance
IEDB Model Calis, J. J. A. *et al.* Properties of MHC Class I Presented Peptides That Enhance Immunogenicity. *PLoS Comput. Biol.* **9**, (2013).	The IEDB model captures sequence-based frequencies from specific amino acids describing those which are more prevalent in immunogenic peptides than non-immunogenic peptides. These 'enrichment scores' are combined with weights on amino acid positional importance for immunogenicity. The user is provided with a score between -1 and 1, to suggest immunogenicity probability.	Data is compiled from the IEDB and three murine studies. Strict data curation process excludes humans as a host for non-immunogenic peptides. Nine-mers are then selected and a redundancy filtering process is performed to avoid oversampling. Dataset after redundancy filtering is not publicly available.	Python	Peptide(s) and HLA Allele	Yes.	ROC-AUC after 3-fold cross-validation: 0.65
NetTepi Trolle, T. & Nielsen, M. NetTepi: An integrated method for the prediction of T cell epitopes. *Immunogenetics* **66**, 449–456 (2014)	NetTepi is defined as the linear combination of binding affinity scores which are obtained by netMHCpan algorithm, binding stability scores which are obtained by netMHCstab (v1, citation) and T cell propensity scores which are obtained similarly to the IEDB model: $M = tTCellPropensity + sNetMHCstab + (1 - t$ Developers optimize weights *t* and *s*, and defaults are given. *netMHCstab* and *netMHCcons* are called directly and canonically from NetTepi and individual scores for these algorithms are produced. These are integrated as above with the T cell propensity predictions. To our knowledge, the only difference between the T cell propensity component of NetTepi and the IEDB model, is that only default masking (the ones used for HLA-A0201) to encode for positional importance are employed by NetTepi.	T cell propensity component uses training data as above per IEDB.	Python	Peptides, HLA Allele, peptide length (8-14)	Yes, 13 HLA-A and HLA-B alleles	Average AUC0.1 values range between 0.9305 and 0.9652
NetTepi & NetMHCpan 4.0 (NetTepi-Ext)	The model was constructed by replacing NetTepi's stability and affinity predictions with netMHCpan4.0 results. Naturally, netMHCpan scores replace the affinity and stability terms in the equation for *NetTepi*. This model produces an immunogenicity score *I,* and aims to mirror the *NetTepi* equation; $I = (t \times Tcell + (1 - t) \times netMHCpanScore)$	N/A	Python and R	Peptides, HLA allele, peptide length (8-14)	Yes, 13 HLA-A and HLA-B alleles	N/A

Table 1

	where *t* falls in the range of 0-1, and where Tcell is the T cell propensity component and netMHCpanScore is the prediction value from netMHCpan 4.0 with a range of 0-1.					
iPred Pogorelyy, M. V. *et al.* Exploring the pre-immune landscape of antigen-specific T cells. *Genome Med.* **10**, 1–14 (2018	iPred is a Multinomial Gaussian Process classifier, utilising expectation-maximisation and as input *iPred* takes peptide sequence(s) and computes 10 Kidera Features [17] averaged over all residues. The output of the model is a probability score reflecting the likelihood of T cell immunogenicity.	Chowell dataset is used to compute kidera factor vector sums for model training.	R	Peptide sequence	No	ROC-AUC: ~0.80
Repitope Ogishi, M. & Yotsuyanagi, H. Quantitative prediction of the landscape of T cell epitope immunogenicity in sequence space. *Front. Immunol.* **10**, (2019)	Repitope is a framework which probes public TCRs to discriminate immunogenicity. Repitope computes physiochemical properties based on mimicking the thermodynamics between pMHC and public TCR interactions. This framework termed 'TCR-peptide Contact Potential Profiling' (CPP) focuses on the TCRβ CDR3 sequences, which are primarily involved in the interactions with peptides presented onto MHC. After a highly dimensional feature (~6000) computation occurs, feature selection can be performed and then a probabilistic estimate of immunogenicity is computed for each epitope. Models can be extrapolated to sequences for which it was not trained to make predictions. Repitope can compute probabilistic estimates of the immunogenicity for an original epitope and for all single amino acid variants, as well as the largest difference of immunogenicity between any variants to explore escape potential. Repitope also provides utility to predict CD4+ T cell immunogenicity for class II MHC.	Calis 2013 dataset, Chowell 2015 dataset, EPIMHC dataset, LANL HCV/HIV dataset, POPISK dataset, MHCBN dataset, TANTIGEN dataset. Please see following link for more details: https://github.com/masato-ogishi/Repitope	R or Python.	Peptide sequences	No	ROC-AUC 0.76

Table 1

Table 1: Overview of Models identified for benchmarking. The table shows from left to right, an overview of the model, information regarding the data used to train the model as published, the programming language the model is implemented in, the information required for the model to make a prediction, whether the model incorporates HLA restriction and – if applicable – performance reported in original publications.

Table 2

Model	Pan-HLA ROC-AUC	HLA-specific ROC-AUC	Percentage Change
IEDB Model	0.65	0.594	-10.5
iPred	0.701	0.588	-16.2
NetTepi	0.661	0.566	-14.6
NetTepi & NetMHCpan 4.0	0.666	0.630	-6.29
Repitope	0.756	0.667	-11.7

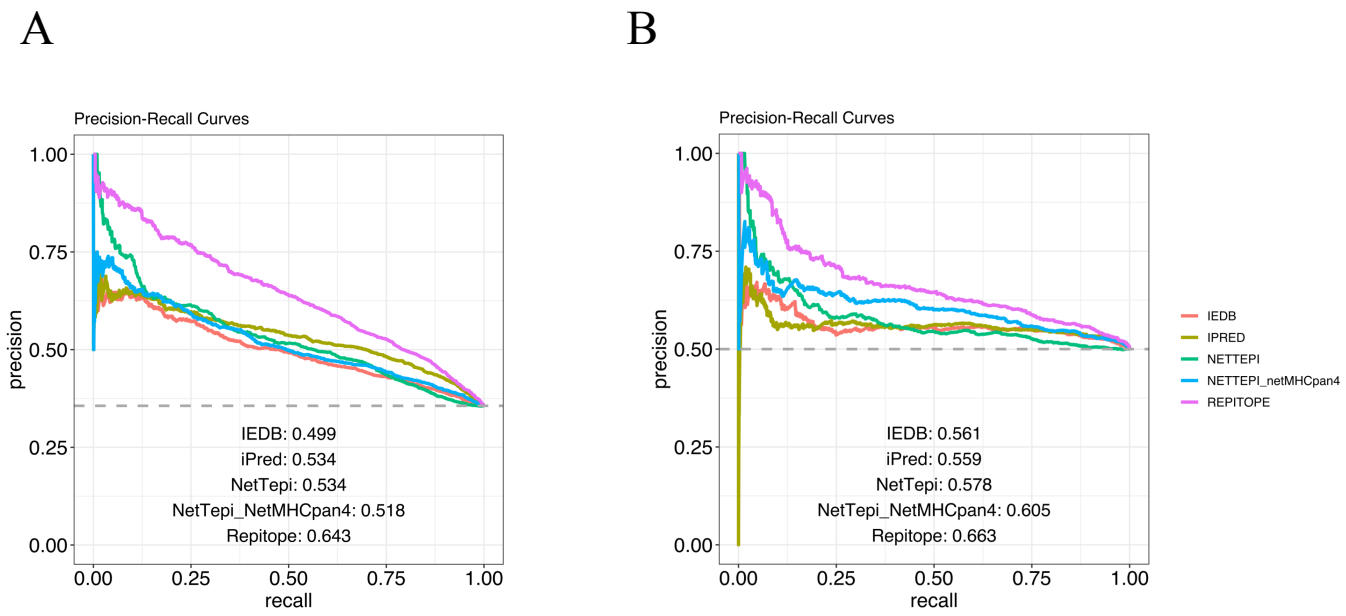
Table 2: Model performance reductions observed from Pan-HLA training to HLA-specific training in the pathogenic setting.

Table 3

Model	Pan-HLA ROC-AUC	30% Down Sampled ROC-AUC	Percentage Change
IEDB Model	0.65	0.667	2.63
iPred	0.701	0.702	0.0898
NetTepi	0.661	0.647	-2.10
NetTepi & NetMHCpan 4.0	0.666	0.673	1.07
Repitope	0.756	0.737	-2.54

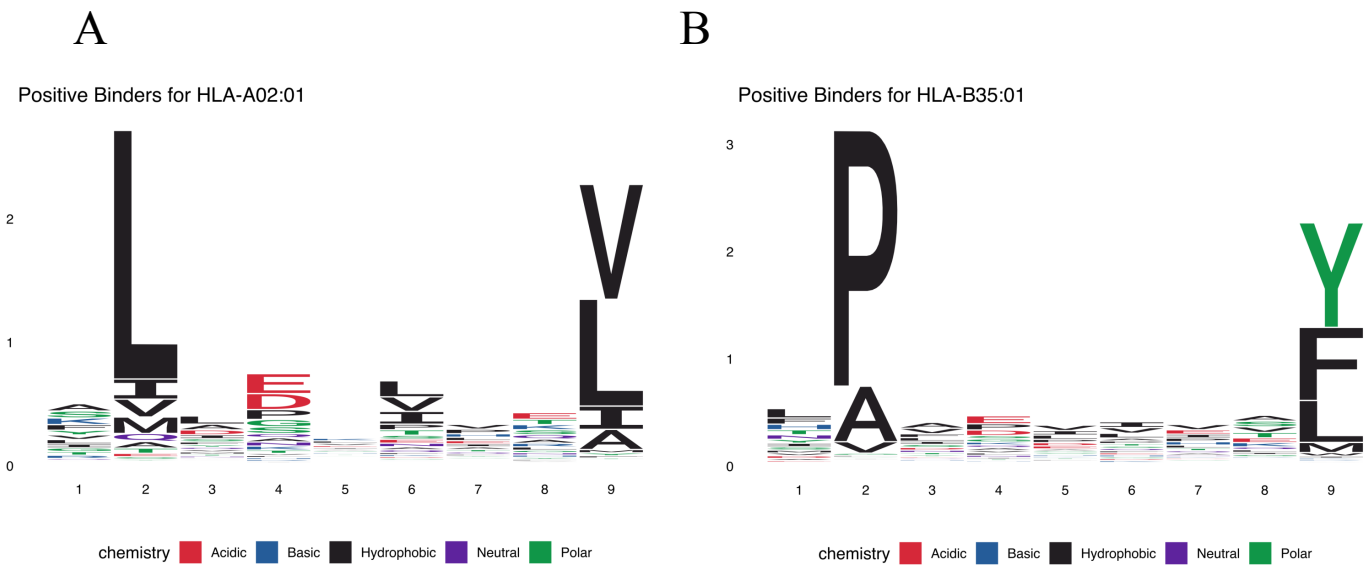
Table 3: Model Performance observed after sampling 30% the size of the Pan-HLA allele dataset compared with the Pan-HLA training performance in the pathogenic setting.

Supplementary Figures



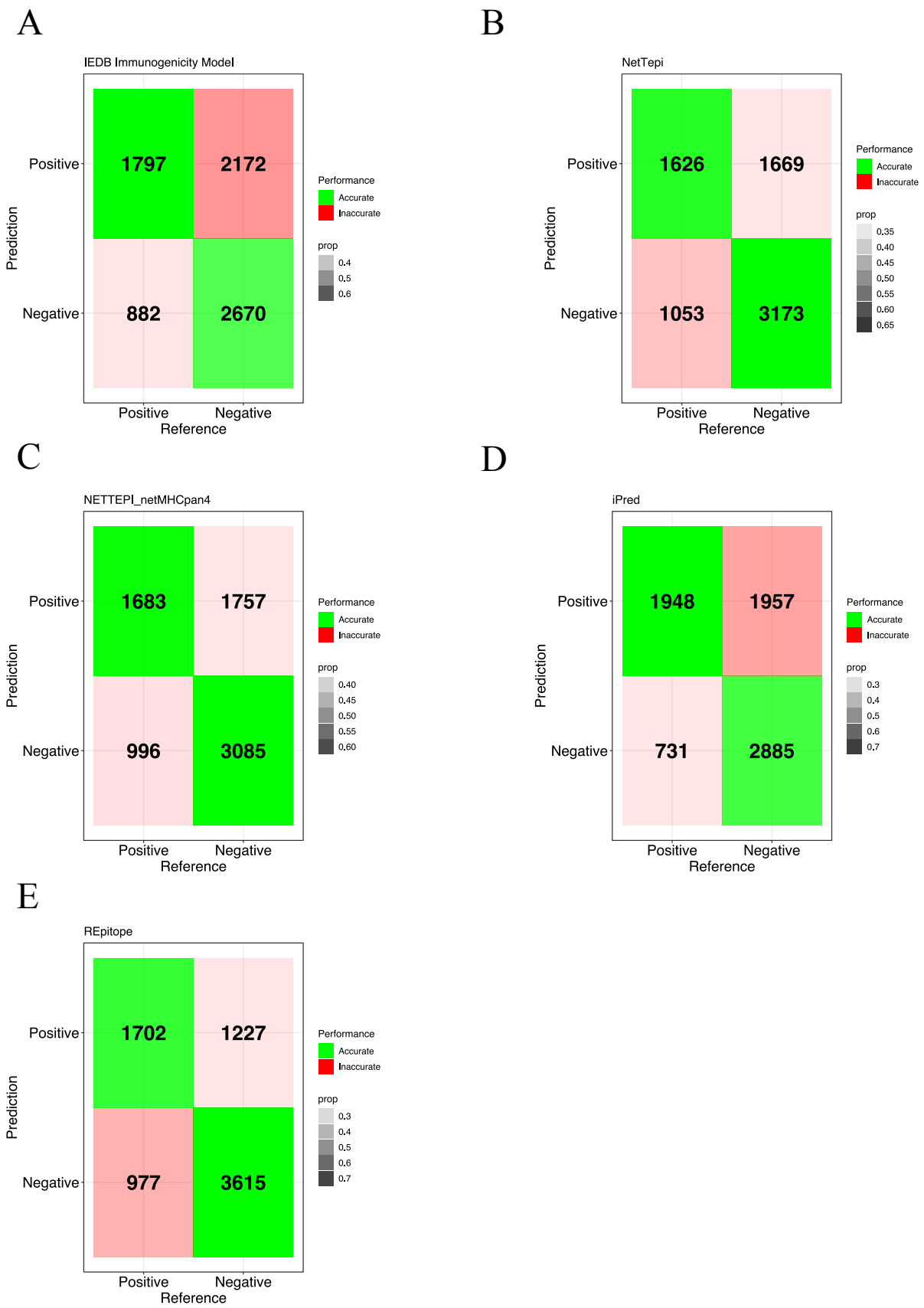
Supplementary Figure 1. A) PR-AUC curves showing performance against the Pan-HLA dataset. B) PR-AUC curves showing performance against the HLA-specific dataset. For PR-AUC curves, a dashed line represents the frequency of the 'Positive' class in the data.

Supplementary Figures



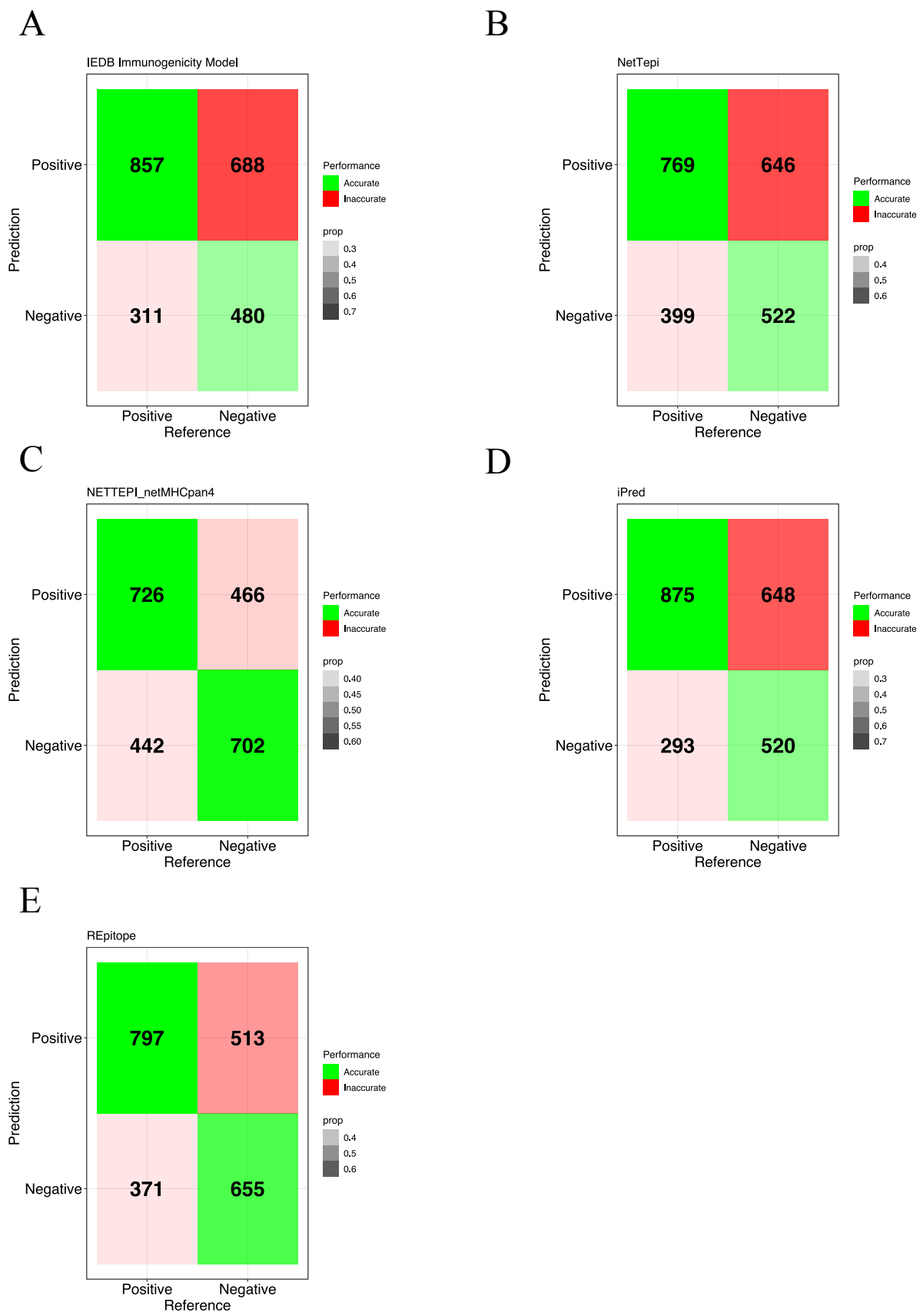
Supplementary Figure 2. A) 'Seq logo' plot showing the distribution of amino acids for peptides that bind to HLA-A02:01. B) 'Seq logo' plot showing the distribution of amino acids for peptides that bind to HLA-B35:01.

Supplementary Figures



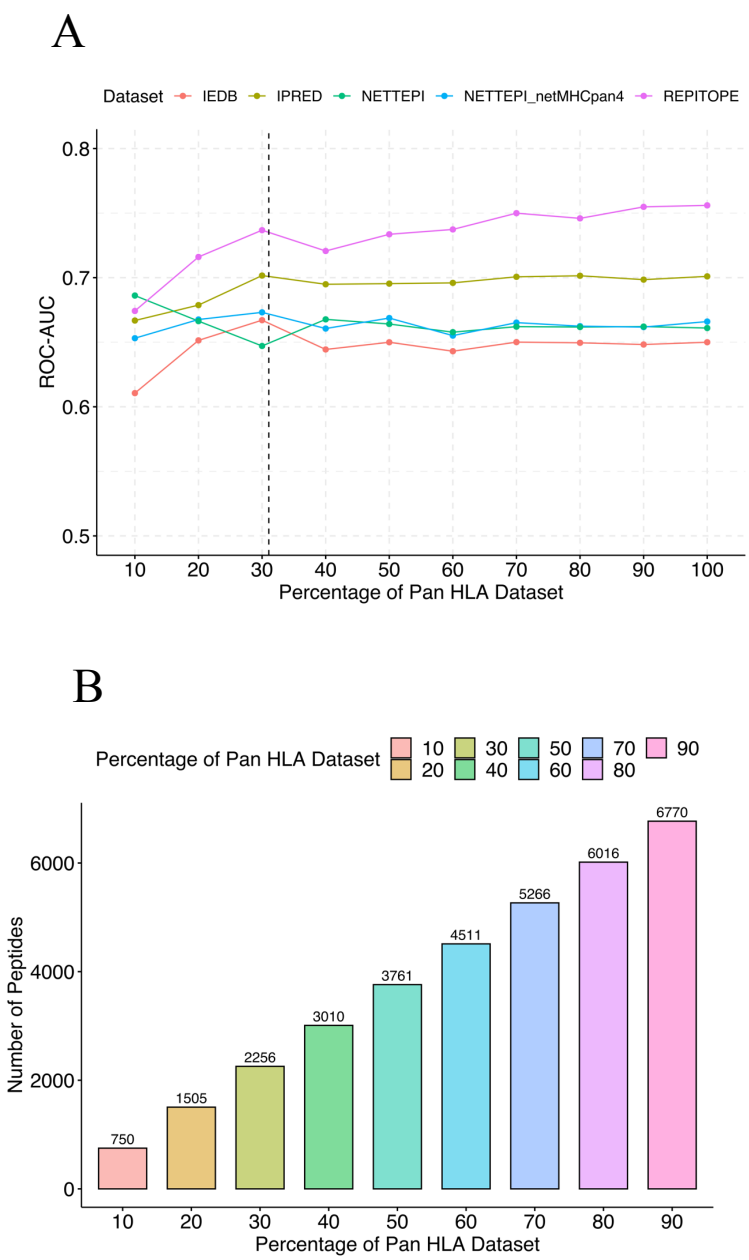
Supplementary Figure 3. Confusion matrices for the pathogenic Pan HLA Allele 10-fold cross-validation experiment for A) IEDB Model B) NetTepi C) NetTepi Extended D) iPred E) Repitope.

Supplementary Figures



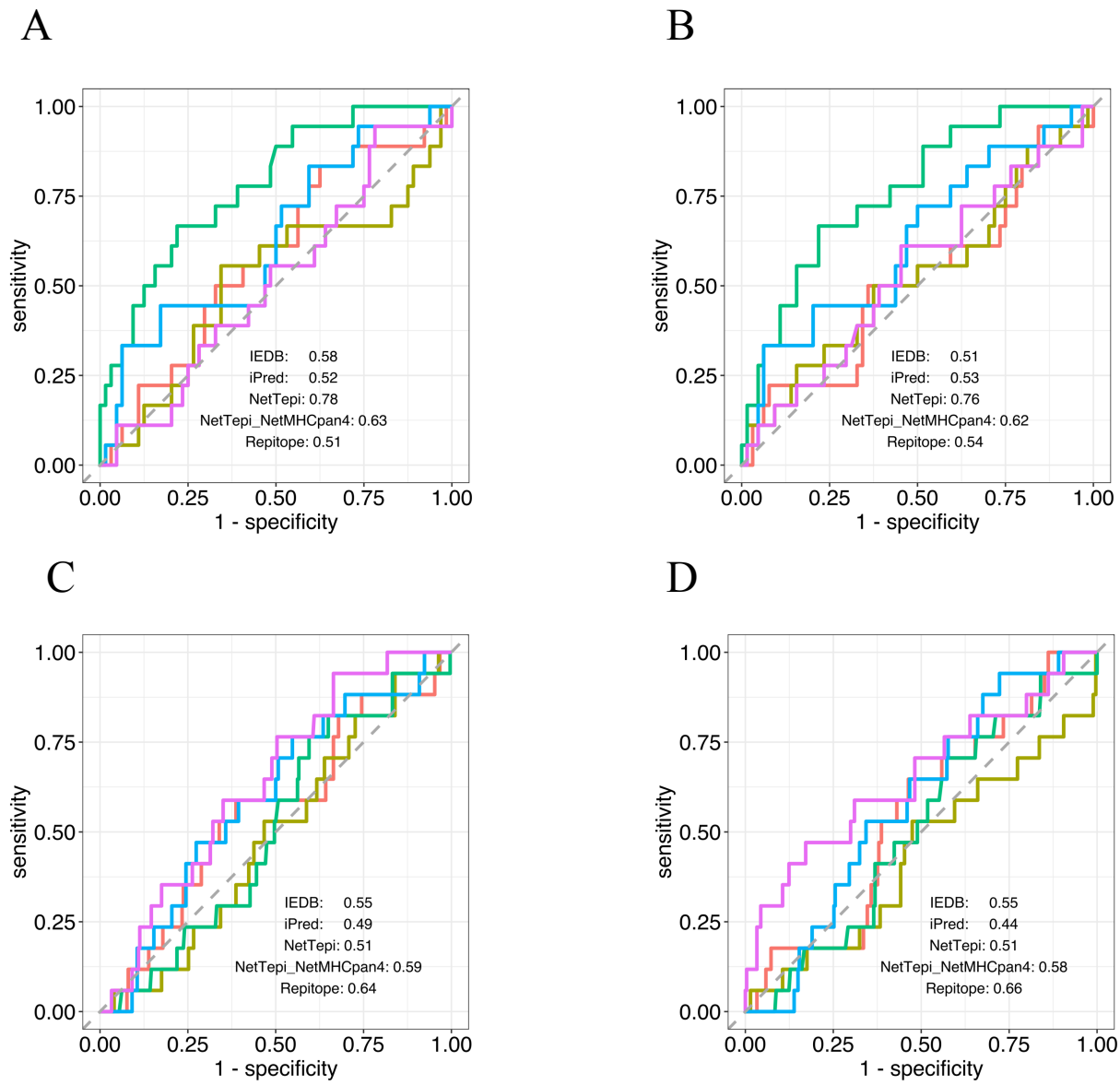
Supplementary Figure 4. Confusion matrices for the pathogenic HLA-specific 10-fold cross-validation experiment for A) IEDB Model B) NetTepi C) NetTepi Extended D) iPred E) Repitope.

Supplementary Figures



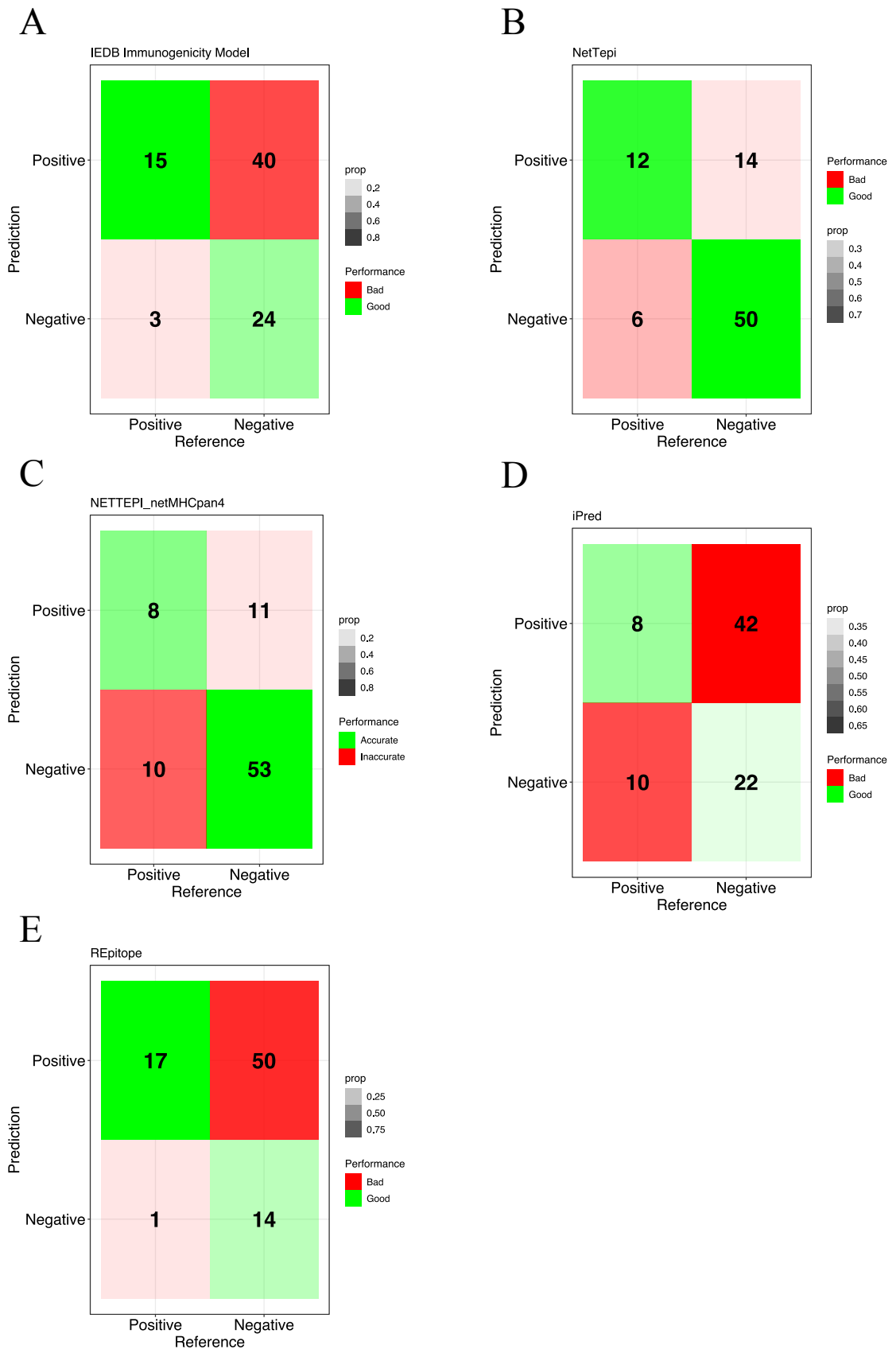
Supplementary Figure 5. A) Line graph showing the ROC-AUCs of models after reducing the sample size of the Pan HLA allele dataset by 10% increments. B) Bar chart showing the number of peptides for each down-sampling scenario.

Supplementary Figures



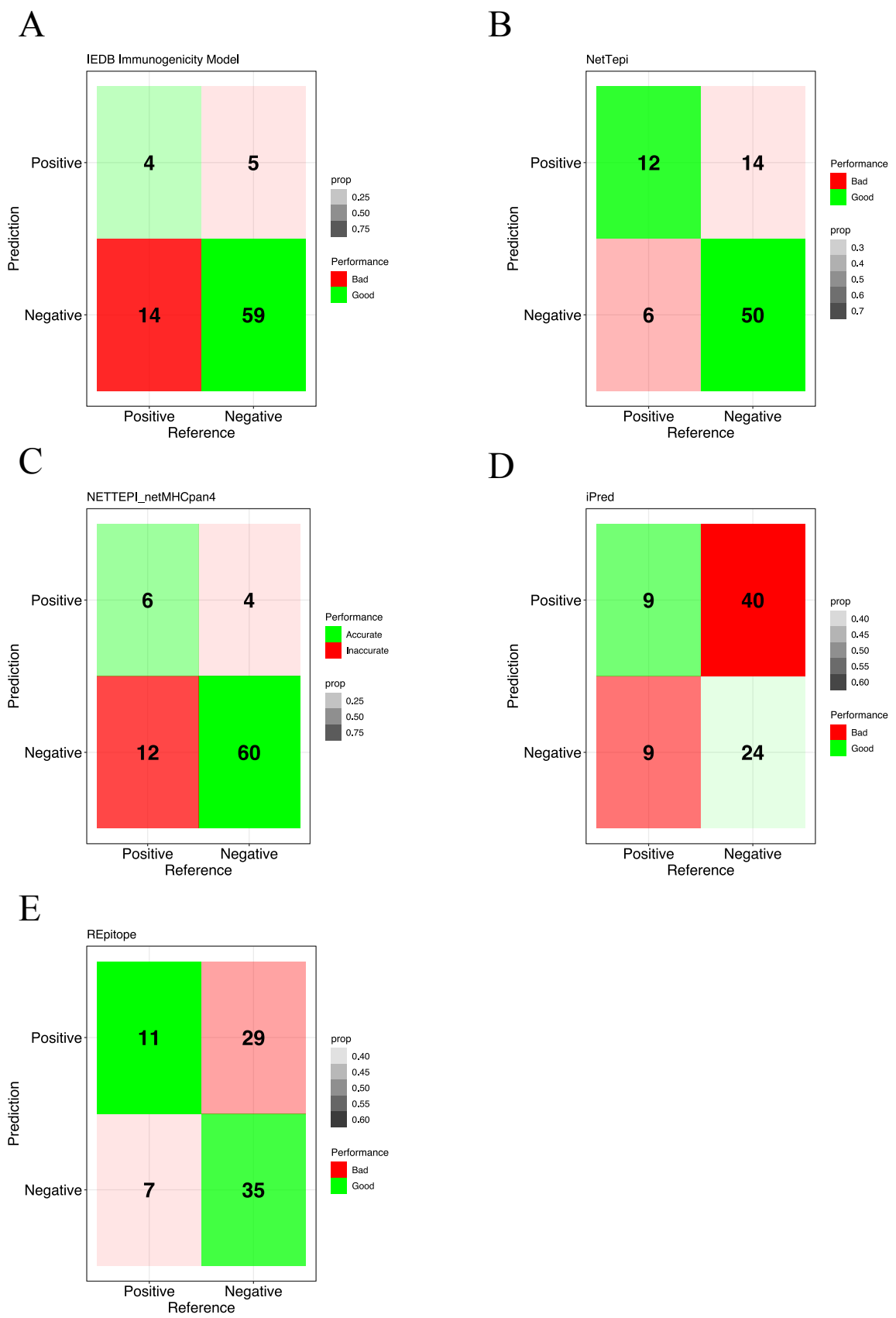
Supplementary Figure 6. A) ROC-AUC plots showing the performance of each model: A) After training on the Pan-HLA dataset and testing against the GBM data. B) After training on the HLA-specific dataset and testing against the GBM data. C) After training on the Pan-HLA dataset and testing against Bjerregaard cancer peptides. D) After training on the HLA-specific dataset and testing against Bjerregaard cancer peptides.

Supplementary Figures



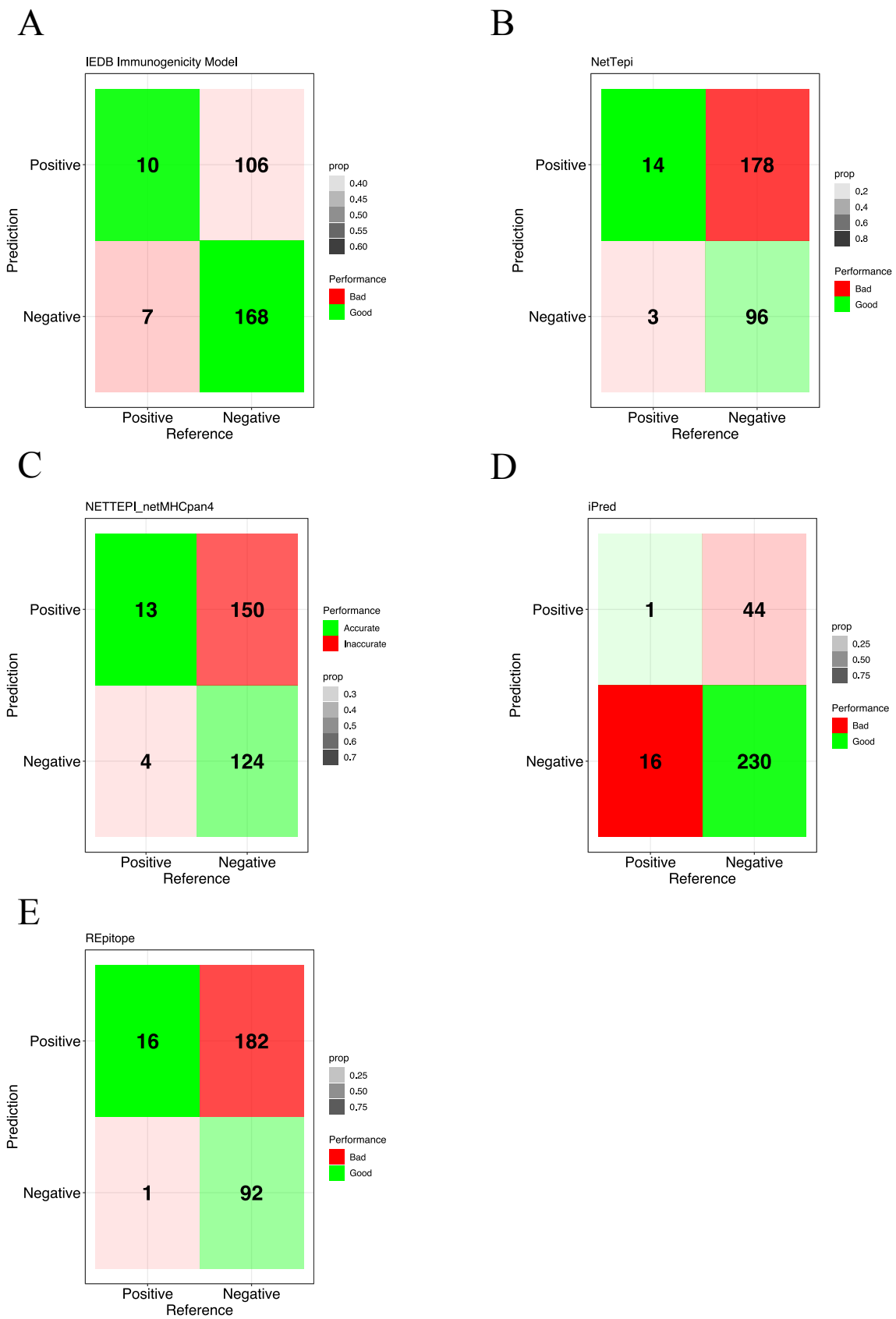
Supplementary Figure 7. Confusion matrices for the cancer neoantigen experiment where models were trained on the Pan HLA dataset and tested against the GBM data. A) IEDB Model B) NetTepi C) NetTepi Extended D) iPred E) Repitope.

Supplementary Figures



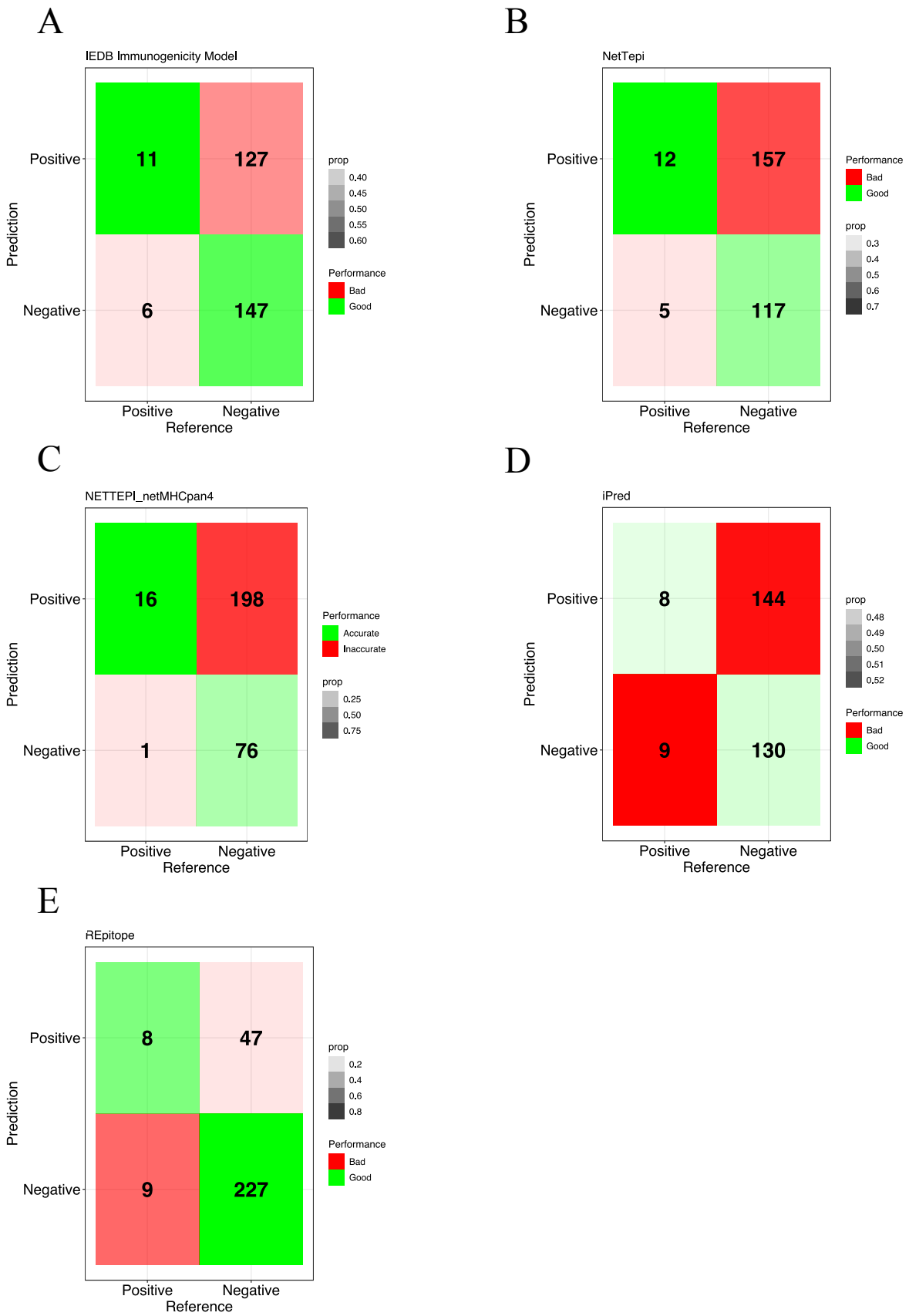
Supplementary Figure 8. Confusion matrices for the cancer neoantigen experiment where models were trained on the HLA-specific dataset and tested against the GBM data. A) IEDB Model B) NetTepi C) NetTepi Extended D) iPred E) Repitope.

Supplementary Figures



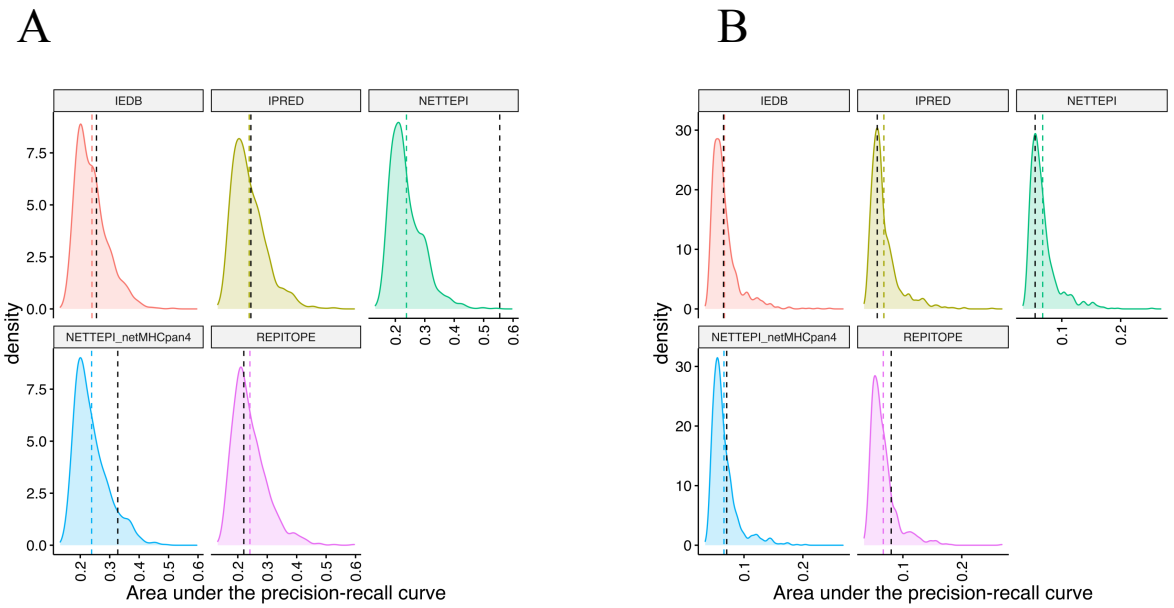
Supplementary Figure 9. Confusion matrices for the cancer neoantigen experiment where models were trained on the Pan-HLA dataset and tested against the Bjerregaard data. A) IEDB Model B) NetTepi C) NetTepi Extended D) iPred E) Repitope.

Supplementary Figures



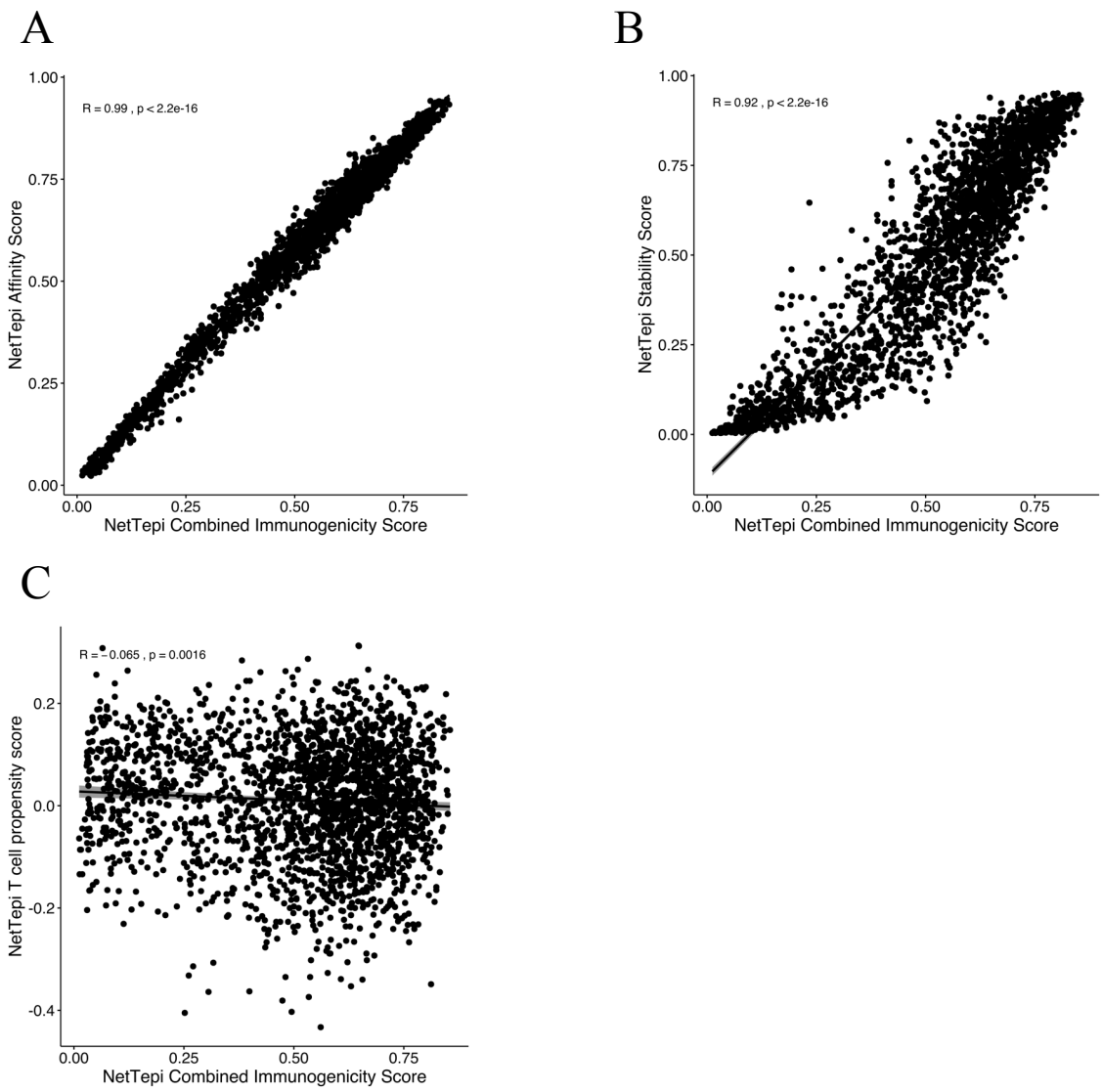
Supplementary Figure 10. Confusion matrices for the cancer neoantigen experiment where models were trained on the HLA-specific dataset and tested against the Bjerregaard data. A) IEDB Model B) NetTepi C) NetTepi Extended D) iPred E) Repitope.

Supplementary Figures



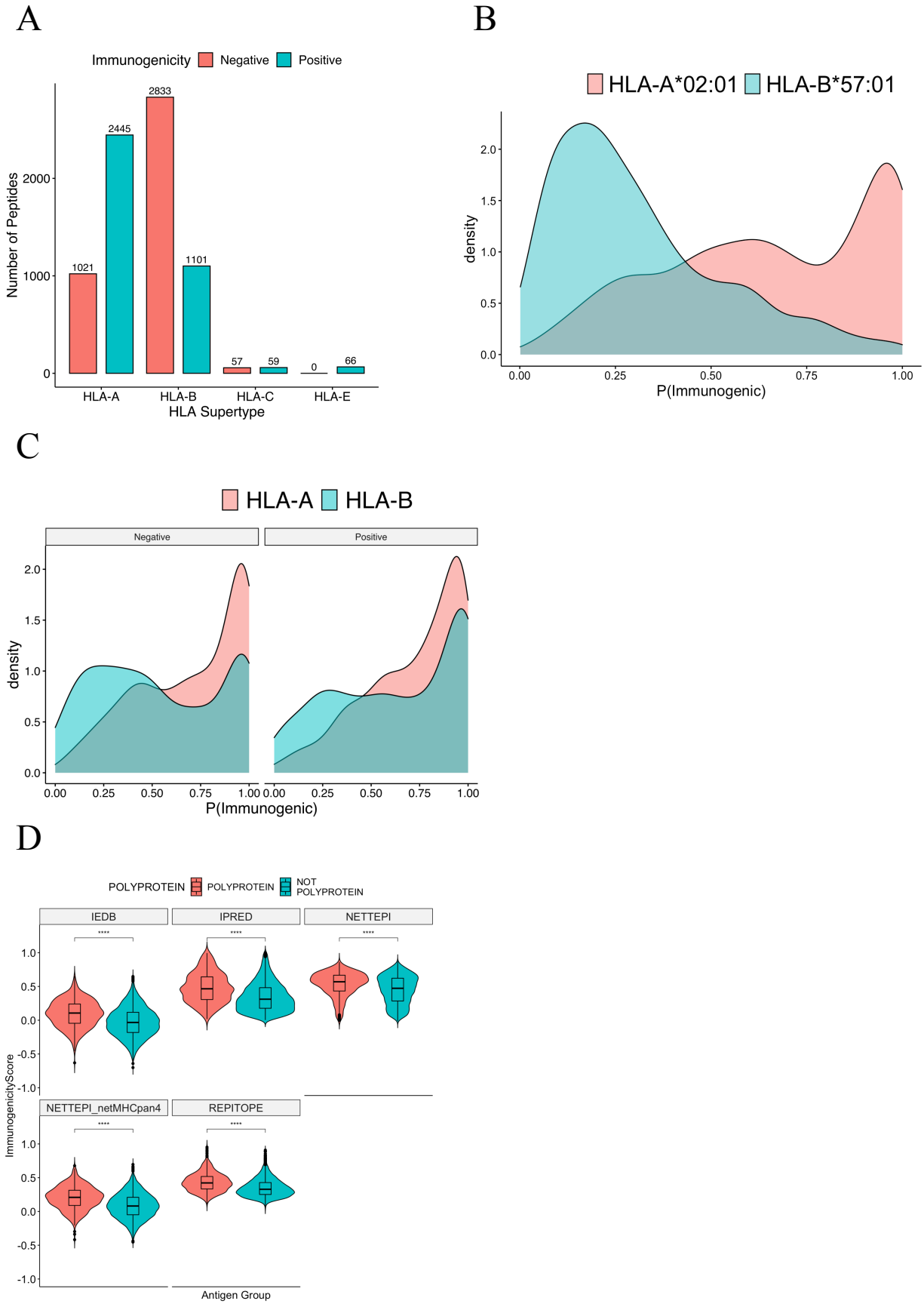
Supplementary Figure 11. Density plots contrasting a control distribution of randomly generated PR-AUCs (in colour) with the PR-AUC achieved after model benchmarking (dashed black line): A) After training on the Pan-HLA dataset and testing against the GBM data. B) After training on the Pan-HLA dataset and testing against the Bjerregard data.

Supplementary Figures



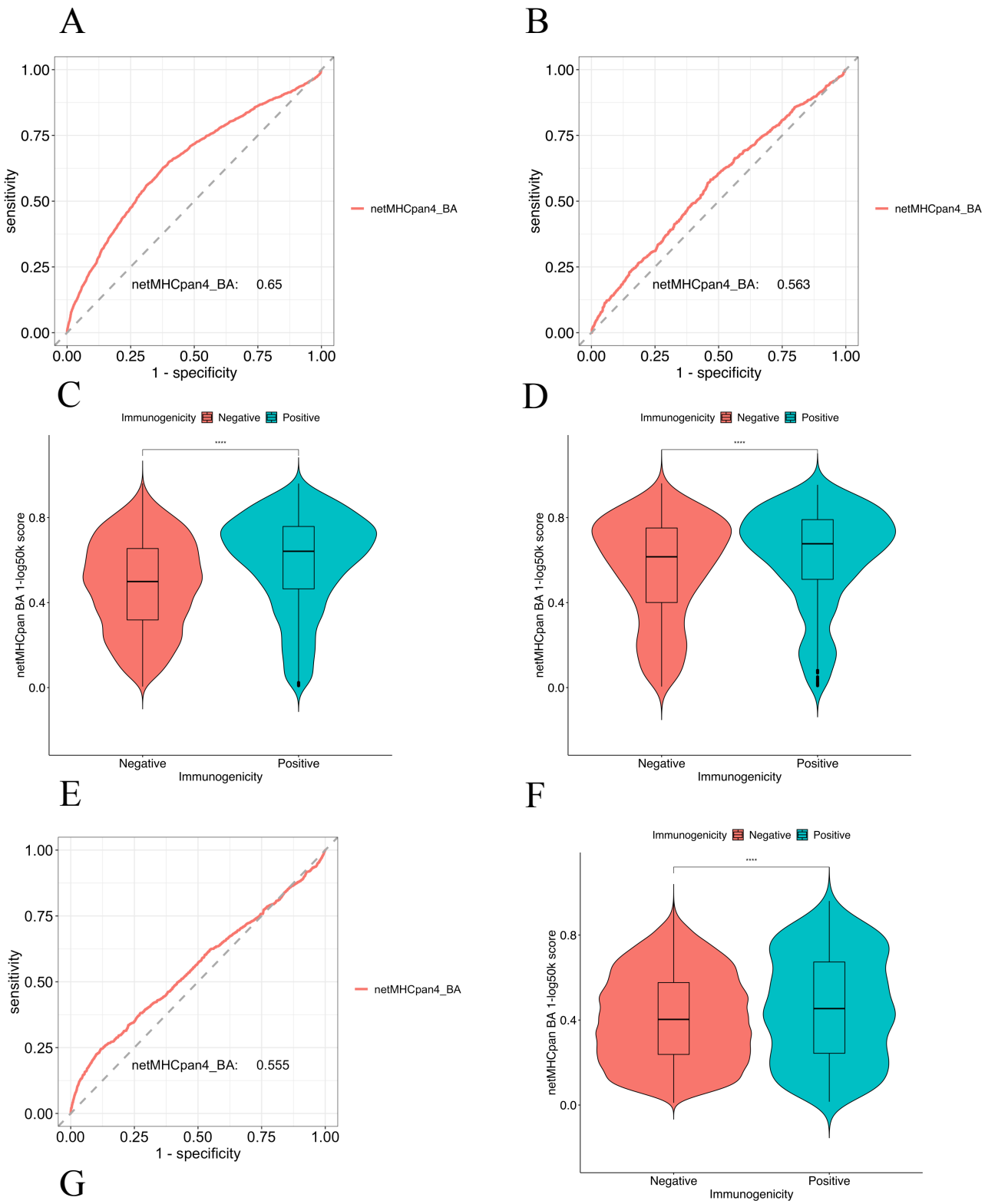
Supplementary Figure 12: A) Scatter plot showing a strong correlation between predicted immunogenicity score and binding affinity score. B) Scatter plot showing a strong correlation between predicted immunogenicity score and stability affinity score. C) Scatter plot showing no correlation between T cell propensity score and immunogenicity score.

Supplementary Figures



Supplementary Figure 13. A) Bar-plot showing the distribution of HLA supertypes per immunogenicity classification for the Chowell dataset. B) Density plot of the distribution of internal training scores for two HLA types generated by iPred when trained as published on the Chowell data. C) Density plot showing the distribution of scores for a test dataset composed of HLA-A02:01 sequences (HLA-A) and sequences from HLA-B alleles (HLA-B). D) Violin plots showing distributions of scores generated by each model after the pan-HLA pathogenic epitopes cross-validation experiment labelled by whether the peptide's source antigen is polyprotein.

Supplementary Figures



Supplementary Figure 14. Binding affinity is not predictive of immunogenicity using diverse sets of human peptides although preference for higher binding affinity is observed amongst immunogenic peptides:
A-B) ROC-AUC of netMHCpan 4.0 binding affinity output when trained using default parameters and tested against the A) Pan-HLA dataset and B) HLA-specific dataset. C-D) Violin plots of the netMHCpan 4.0 binding affinity output 1-log50k scores for both nonimmunogenic and immunogenic peptides for the C) HLA-specific dataset and D) Pan-HLA dataset. F) Violin plot of the netMHCpan 4.0 binding affinity output for immunogenic and non-immunogenic vaccinia virus peptides tested in humans, from our entire standard dataset. G) Barplot showing the percentage of immunogenic and non-immunogenic vaccinia virus peptides from the standard dataset, per allele and for lengths 9 and 10.

Supplementary Tables

Model	Accuracy	F1	Precision	Recall	F0.5	Balanced Accuracy
IEDB Model	0.59	0.54	0.45	0.67	0.48	0.61
iPred	0.64	0.59	0.50	0.73	0.53	0.66
NetTepi	0.61	0.54	0.49	0.61	0.51	0.63
NetTepi & NetMHCpan	0.63	0.55	0.49	0.63	0.51	0.63
Repitope	0.71	0.61	0.58	0.64	0.59	0.69

Supplementary Table 1. Performance metrics for each model after the pathogenic epitopes cross-validation experiment in the Pan-HLA setting.

Model	Accuracy	F1	Precision	Recall	F0.5	Balanced Accuracy
IEDB	0.57	0.63	0.55	0.73	0.58	0.57
iPred	0.60	0.65	0.57	0.75	0.60	0.60
NetTepi	0.55	0.60	0.54	0.66	0.56	0.55
NetTepi & NetMHCpan	0.61	0.62	0.61	0.62	0.61	0.61
Repitope	0.62	0.64	0.61	0.68	0.62	0.62

Supplementary Table 2. Performance metrics for each model after the pathogenic epitopes cross-validation experiment in the HLA-specific setting.

Model	Accuracy	F1	Precision	Recall	F0.5	Balanced Accuracy
IEDB	0.48	0.41	0.27	0.83	0.32	0.60
iPred	0.37	0.24	0.16	0.44	0.18	0.39
NetTepi	0.76	0.55	0.46	0.67	0.49	0.72
NetTepi & NetMHCpan	0.74	0.43	0.42	0.44	0.43	0.64
Repitope	0.38	0.40	0.25	0.94	0.30	0.58

Supplementary Table 3. Performance metrics for each model after the cancer neoantigen cross-validation experiment in the Pan-HLA setting.

Model	Accuracy	F1	Precision	Recall	F0.5	Balanced Accuracy
IEDB	0.77	0.3	0.44	0.22	0.37	0.57
iPred	0.40	0.27	0.18	0.50	0.21	0.44
NetTepi	0.76	0.55	0.46	0.67	0.49	0.72
NetTepi & NetMHCpan	0.80	0.43	0.60	0.33	0.52	0.64
Repitope	0.56	0.38	0.28	0.61	0.31	0.58

Supplementary Table 4. Performance metrics for each model after the cancer neoantigen cross-validation experiment in the HLA-specific setting.

Supplementary Tables

Model	**Pan-HLA GBM Dataset Z-score**	**HLA-specific GBM Dataset Z-score**	**Pan-HLA Bjerregaard Dataset Z-score**	**HLA-specific Bjerregaard Dataset Z-score**
IEDB	0.28	1.1	-0.07	-0.04
iPred	0.1	0.80	-0.45	2.67
NetTepi	5.67	4.41	-0.53	-0.53
NetTepi & NetMHCpan	1.54	1.66	0.21	-0.1
Repitope	-0.35	0.11	0.57	5.63

Supplementary Table 5. Z-scores computed in each scenario during the exploration of reliability for each model against cancer neoantigen datasets.

Supplementary Datasheets

External files:

Supplementary Datasheet 1: The Standard Dataset.
Supplementary Datasheet 2: The Pan-HLA Dataset.
Supplementary Datasheet 3: The HLA-specific Dataset.
Supplementary Datasheet 4: The Functionally Validated GBM Dataset

Supplementary Text

Exclusion of Models

In this study, we limited ourselves on a direct comparison of four publicly available models, and one first implementation serving as a potential update to *NetTepi*. There exist several other potentially relevant features for which their relevance to immunogenicity have begun to be explored recently, culminating in models which we were unable to include in this study for reasons such as inaccessibility of the source code, inability to produce their input features, etc. For the sake of completeness, in the following paragraphs we aim to highlight functionality and exclusion of such models.

Structure-based peptide immunogenicity predictors are becoming increasingly available, and a recently reported neural-network model trained on structural information within the context of TCRs and peptide binding, outperformed other models including the *IEDB* and *NetTepi* models [1]. The reported ROC-AUC is in the region of *Repitope's* reported performance, illustrating the potential that additional features including but not limited to rich structural information can provide to circumventing this complex problem. This model records a reasonable performance against a cancer peptide dataset, although their model performance reduces against neoantigens compared to a wider epitope dataset, an observation consistent with our study. This particular model was not benchmarked due to our inability to generate the structural data required for training the model for benchmarking.

The neoantigen fitness model from Luksza et al[2] comprises a component which computes 'R', a neoantigen immunogenicity metric. R represents the probability that a tumour neoantigen is recognised by the TCR repertoire. Peptides for testing are aligned to known viral epitopes from the IEDB, and sequence similarities mimic binding energies for a thermodynamic model to compute R, the probability that a given sequence *S* is bound to a TCR specific to some epitope from the pool. We implemented this model and re-produced 'R' calculations from their work, although we observed that the 'R' component alone of this model produced poor performance against all of our datasets. We determined that 'R' component's intended use in-conjunction with remaining model components for neoantigen fitness makes a direct comparison to models here inappropriate. A recent preprint presents a Cytotoxic T lymphocyte (CTL) immunogenicity predictor inspired by this model from Luksza, which reports a ROC-AUC of 0.76. This predictor was utilised to suggest that a level of pre-existing $CD8^+$ T cell immunity against COVID-19 is possessed by some individuals prior to infection [3].

Another model that we could not include in our direct comparison is NetTCR, which predicts binding probability between the CDR3 sequence of a TCR and an HLA-A02:01-peptide complex [4]. NetTCR is a convolution neural network model, detecting patterns in input data with convolution filters and reports a ROC-AUC of 0.727. NetTCR's prediction to specific TCR sequences means we could not fairly assess its performance in the current study. POPISK was selected as an appropriate model for benchmarking however we were unable to access the webserver or locate a copy of the model [5]. POPISK was the first immunogenicity model to construct and utilise a large-scale dataset. This model employs a support vector machine and reports ability to predict immunogenicity for 68% of HLA-A2 binding peptides.

iNeoEpp[6] is another very recent immunogenicity predictor inspired by the enrichment-score approach utilised in the IEDB model by Calis et al. Their model extracts sequence physicochemical information, eluted ligand rank, peptide entropy and 'immunogenic frequency score' which reflects amino acid distribution frequency differences between immunogenic and non-immunogenic peptides at TCR contact sites. Consistent with Chowell et al's analysis, the *iNeoEpp* authors observe that TCRs tend to identify hydrophobic amino

acids. *iNeoEpp* comprises a neoantigen prediction model distinct from its 'antigen' epitope predictor, utilising mutational information and sequence-based amino acid characteristics. For the 'antigen-epitopes' model, they report a ROC-AUC against a validation set of 0.77, and for the 'neoantigen-epitope' model they report a ROC-AUC of 0.785. The source-code for this model was not available to enable a direct comparison. We did test this model as published against the GBM dataset and unfortunately while a reasonable ROC-AUC was achieved, zero precision was recorded as every peptide was predicted non-immunogenic.

Thresholding

Throughout the process of benchmarking, we observed that the threshold for determining a positive or negative prediction, was a key factor for each predictor's accuracy. Some publications provide a mechanism for thresholding, with variation in the strictness, e.g the *IEDB* model as published does not classify epitopes as immunogenic or non-immunogenic, while *NetTepi* does. For all models, we observed that optimising the classification threshold per experiment improves performance. Therefore, consideration to calibrating thresholds for individual use cases may be taken into consideration for future model development.

NetTepi Extended

We have presented a first implementation of *NetTepi Extended,* although major improvements in performance are not observed compared to *NetTepi* for the Pan-HLA allele dataset although this model is the second-best performing model using only the HLA-A02:01 dataset, outperforming *NetTepi* and closely following *Repitope*. Despite a preliminary analysis to maximise ROC-AUC through optimising model weights, further work is required to calibrate an implementation of *NetTepi Extended* and a potential avenue for improvement may be to optimise weights for both PR-AUC and ROC-AUC, as well as inclusion of other netMHCpan metrics.

Supplementary References

1. Sgourakis, N. G. *et al.* Structure Based Prediction of Neoantigen Immunogenicity. *Front. Immunol. | www.frontiersin.org* **1**, (2019).
2. Luksza, M. *et al.* A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. *Nature* **551**, 517–520 (2017).
3. Gao, A. *et al.* Predicting the Immunogenicity of T cell epitopes: From HIV to SARS-CoV-2. doi:10.1101/2020.05.14.095885.
4. Isabell Jurtz, V. *et al.* NetTCR: sequence-based prediction of TCR binding to peptide-MHC complexes using convolutional neural networks. doi:10.1101/433706.
5. Tung, C. W., Ziehm, M., Kämper, A., Kohlbacher, O. & Ho, S. Y. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* **12**, (2011).
6. Wang, G. *et al.* INeo-Epp: T-cell HLA class I immunogenic or neoantigenic epitope prediction via random forest algorithm based on sequence related amino acid features. (2019) doi:10.1101/697011.