

Predicting HLA-I peptide immunogenicity with deep learning and molecular dynamics

Jeffrey Weber

IBM Thomas J Watson Research Center

Diego Chowell

MSKCC

Chirag Krishna

MSKCC

Timothy Chan

Memorial Sloan Kettering Cancer Center

Ruhong Zhou (✉ rz24@columbia.edu)

Columbia University <https://orcid.org/0000-0001-8624-5591>

Article

Keywords: T Cells Self-discrimination, Human Leukocyte Antigen, Immunogenicity, Convolutional Neural Network, Cancer Neoantigens

Posted Date: November 16th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-104972/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Understanding how T cells discriminate self from non-self is a fundamental question with important implications for immunology, immunotherapy, and vaccine development. Presentation of peptides by human leukocyte antigen I (HLA-I) molecules is necessary but not sufficient for T cell recognition, and peptide features that dictate immunogenicity are obscure. Here, we develop a convolutional neural network that learns features governing peptide immunogenicity, integrating molecular dynamics and sequence representations of humans, pathogen, and tumor peptides presented by HLA-I. Our model identified structural and dynamical properties correlated with immunogenicity and yielded a highly accurate classification of peptides from pathogens versus humans. Furthermore, we applied our model to classify more challenging cancer neoantigens, and it successfully predicted immunogenic neopeptides from patients with melanomas. These data demonstrate the utility of deep learning models built on molecular dynamics and reveal underlying properties that govern HLA-I peptide immunogenicity.

Introduction

Immunotherapies such as immune checkpoint blockade and personalized neoantigen vaccines can induce anti-tumor immunity (1-4). Neoantigens arise from somatic mutations in the cancer genome, and are presented on the cell surface via binding to the polymorphic MHC-I molecules; these neopeptides are the principal targets of tumor-specific T cell responses(5, 6). Accordingly, the identification of immunogenic epitopes—i.e. those that are effectively recognized by T-cells—is a fundamental problem in immunology, with important implications for vaccine development in cancer and at large.

Currently, the identification of immunogenic epitopes is largely limited to HLA-I binding predictions(7-13). Critically, peptide binding to HLA-I is necessary but not sufficient for T cell recognition(14); moreover, the majority of HLA-bound epitopes are not recognized by the immune system(7). Therefore, peptide features (ranging from amino acid identity to more complex characteristics of HLA-I peptide presentation) may play an important role in determining epitope immunogenicity. To that end, prior studies have shown that the position of the mutated amino acid in the peptide, the contacting amino acids from the T-cell receptor, and dissimilarity to self are all determinants of HLA-I peptide immunogenicity(15-23). Concordantly, recent work has suggested that a critical component of neoantigen immunogenicity is sequence similarity to pathogenic epitopes(21). However, the structure and dynamics of the peptide-HLA complex are also likely important for T cell recognition(24-28) and remain relatively unexplored.

In this work, we present an immunogenicity prediction model that combines peptide sequence and HLA-peptide molecular dynamics simulation data within an integrated deep convolutional neural network. We show that our neural network yields highly accurate classification of pathogen, cancer, and human self peptides and successfully identifies immunogenic cancer neoantigens in two individual patients with

melanoma. Analysis of the extensive molecular dynamics ensemble reveals the biophysical underpinnings of peptide immunogenicity and illuminates corresponding peptide features that complicate the classification of non-immunogenic cancer neoantigens. In particular, we find that, when presented by HLA-A*02:01, the ensemble of non-immunogenic cancer neoantigens considered in this work look far more like immunogenic pathogen peptides than non-immunogenic self peptides.

Results

To explore determinants of HLA-I peptide immunogenicity, we first assembled a dataset of 1,038 HLA-A02-restricted immunogenic 9mer epitopes from pathogens(29), as well as 1,845 non-immunogenic self 9mer peptides identified via mass spectrometry(30) (Fig. 1a). We used these data to develop a convolutional neural network that discriminates between immunogenic and non-immunogenic peptides (Fig. 1). We first applied a one-hot sequence representation (Fig. 1b) with a residue-similarity ordering, hypothesizing that both amino acid identity and arrangement within the peptide may affect immunogenicity. This input representation allows the neural network to learn aspects of amino acid similarity from a spatial encoding of proximity. By applying a 2 x 2 sliding convolutional filter, we integrated both amino acid sequence and similarity into the network (Fig. 1c). Applying a five-fold cross validation approach to our data, our current sequence convolutional neural network achieved a mean test AUC of 0.85 on the 2,883 peptide dataset. The AUCs achieved in a recent work focused on classifying immunogenic pathogen peptides range between 0.60 and 0.70 (28), suggesting enhancements in both data curation and architecture can improve prediction results (Supplementary Table 1 and Supplementary Text).

The general molecular topology of HLA-I bound to peptide antigens is widely conserved, but HLA-I/peptide complexes each interact in a specific manner with the surfaces of their conjugate TCRs. Therefore, we hypothesized that differences in peptide solvent exposure at the residue level could account for differences in immunogenicity, since the features of the presented peptide side chains represent the nexus of the immune recognition interface. In support of this hypothesis and consistent with previous analyses(16, 17), Riley et al.(28) reported that hydrophobic solvent-accessible surface area of presented peptides is correlated with heightened immunogenicity, particularly at positions 4, 5, and 8. To interrogate the structural dynamics of the peptide-HLA complex and its effects on immunogenicity, we conducted extensive atomistic molecular dynamics (MD) simulations of a representative subset of our peptide collection (Supplementary Figure 1). Notably, it has been widely established that protein dynamics simulations in explicit solvent properly capture the dynamic hydrophobic exposure and/or burial of side chains in peptide conformational ensembles, which are essential for proper function (31). MD simulations provide the added benefit of generating time-ordered conformations, facilitating the creation of temporal features that are useful in training classifiers.

In our study, peptides were selected to ensure sequence diversity within the MD subset, particularly in reference to positions 1 and 8. In total, 110 immunogenic pathogen and 123 non-immunogenic self-nonamers were loaded onto the HLA-A*02:01 peptide-binding groove (Fig. 2a) based on a common PDB structure (PDB: 5NMH) through structural alignment. Loaded peptides were then placed in an explicit solvent box, subjected to an annealing procedure to minimize clashes, equilibrated for 100 ns at 310 K and 1 atm, and then simulated at production for 500 ns at the same temperature and pressure (32-37). This procedure was also repeated for 17 immunogenic tumor neoantigens (29). The resulting simulation ensemble (comprised of 250 peptides, and 125 μ s of aggregate MD simulation time) showed that hydrophobic exposure (defined by exposure of residues G, A, V, L, I, F, W, Y, and P) is enhanced at positions 4, 5, and 8 in immunogenic epitopes (Fig. 2b), with particular emphasis on positions 5 and 8. Another important observation derived from our MD simulations concerns non-hydrophobic exposure in non-immunogenic peptides: non-hydrophobic residues are considerably more exposed at peptide positions 4, 5, and 8, suggesting that immunogenicity is strongly anti-correlated with polar or charged residues at those positions. This finding is significant, despite the fact that enhanced hydrophobic exposure occurs in immunogenic epitopes at the same positions. Such anti-correlations with non-hydrophobic exposure are likely equally useful for immunogenicity classification.

Based on the above results, we hypothesized that an integrated representation of HLA-I binders, incorporating both sequence and dynamic structural features, would enable us to learn additional properties of HLA-I peptide immunogenicity and improve prediction of immunogenic epitopes. We designed a deep convolutional neural network that combines sequential, structural, and dynamical features for immunogenicity prediction. Our featurization approach involves computing residue-specific solvent-accessible surface areas (SASAs) as a function of time for each peptide sequence, with MD frames sampled every nanosecond (Fig. 2c). This procedure results in de facto coarse-grained images of the HLA-A*02:01 peptide presentation interface, a feature type that is well suited to convolutional architectures. The images were then split into 10 subtrajectories per peptide (representing 50 ns of MD each), with the dual purpose of providing more varied samples (2500, in this case) on which to train/test and to facilitate testing on shorter simulations, if desired. Importantly, samples from any single peptide were never divided between training and test data sets to avoid trivial classification outcomes.

Our integrated deep learning architecture is comprised of three modules: the sequence-based module adapted from above, and two additional sets of layers designed to encode molecular dynamics data. With respect to the MD data, each sample image corresponding to an MD trajectory was split into two channels to reflect our residue exposure observations (Fig. 3a): one for hydrophobic (G, A, V, L, I, F, W, Y, and P) residues, and a second for the complementary residues present in each peptide. These two images were then fed through respective series of three convolutional and pooling layers to generate internal tensor representations for both sets of residues. In these cases, pooling operations were only

conducted in the time dimension of each image, as the full resolution in the residue dimension needs to be retained to capture each peptide position's unique exposure statistics and relation to other residues.

The first two pooling layers in the MD convolution modules were designed to identify the most prominent features in each peptide column (a strategy often employed in convolutional neural networks called max pooling), while the third pooling layer was designed to compute an ensemble average over dynamic features representative of the underlying statistical physics of protein conformations. In standard MD data analysis, one typically employs some kind of clustering and/or ensemble averaging to understand the sampled conformational landscape. In the present case, the deep learning encoder plays the role of these traditional analytical tools in grouping and averaging conformations, using experimental immunogenicity labels as a guide.

The output of the hydrophobic and complementary convolutional modules was then concatenated and fed through a dense layer to yield a unified representation for the protein structure and dynamics. This MD-derived representation was then concatenated with the output from the sequence-based module and fed through a last dense layer, yielding an integrated sequential, structural, and dynamical representation that is used to predict a peptide's probability of immunogenicity via a final softmax layer.

Conducting a five-fold cross validation on the peptide set for which MD data was collected, we found that the structural and dynamical features improved upon the performance of the sequence-only model. ROC curves for representative test sets from this procedure are presented in Fig. 3b. The MD plus sequence model yields an average test AUC of 0.89 (Fig. 3c), as opposed to the average AUC of 0.80 derived from the sequence-only model on an identical five-fold data split of this reduced peptide set. These results suggest that our integrated model can distinguish viral immunogenic peptides from non-immunogenic self peptides at a high level of discrimination.

Considering the performance gap between the sequence-only and MD plus sequence models, one can conclude the improvement derived from MD data does not result from trivial bias in how sequences were split between training and test sets. We conducted a pair of feature analyses aimed at explaining how the deep learning model learns from molecular dynamics data: the first based on residue identity, and the second based on residue position. In the first test, we artificially restricted the hydrophobic residue list to G, A, V, L, I, F (removing W, Y, and P, which have presumed importance because of their size and abundance, but can be considered non-hydrophobic in certain environments). This procedure reduces the AUC to 0.85 (Fig. 3c), suggesting that distinguishing between hydrophobic and complementary SASAs, rather than simply using generic residual exposure as a function of position, is important for predicting

immunogenicity. Notably, the addition of MD-based features with the reduced hydrophobic residue list still significantly outperforms the sequence-only model on this same 250 peptide dataset (Fig. 3c).

In our second feature analysis, we aimed to ascertain the relative impact of residues at each peptide position (P1-P9) on immunogenicity prediction. To do so, we set all input apart from MD-based features at one specific peptide position to zero in each of nine five-fold training runs. While underlying correlations in residue dynamics present in MD trajectories cannot be removed, our ablation procedure ensures that features are only explicitly considered from one peptide position at a time.

This feature ablation procedure illustrates differences in immunogenicity prediction depending on the position considered (Fig. 3d). Perhaps unsurprisingly, the model is able to learn the least from dynamics of P2 and P9, which correspond to peptide anchor residue positions for HLA-A*02:01. This result further reinforces the point that peptide immunogenicity is a concept distinct from peptide-HLA binding, since properties of residues at P2 and P9 are known to be essential to the binding process. Outside of the anchor residues, model performance is relatively uniform across the positions considered, suggesting the model is able to distinguish a variety of peptide presentation modes in determining immunogenicity. The worst model performance at a non-anchor position occurs at P4; past work has suggested residue exposure at P4 is particularly important to peptide immunogenicity²⁶, but we cannot draw the same conclusion based on the present data. Interestingly, our model is able to perform relatively well with residue solvent exposure information from P1, a peptide position not typically associated with TCR interactions. Regardless, the result from training on the full set of MD features simultaneously (mean test AUC ~0.89 with sequence information; ~0.84 without sequence information) is still significantly higher than that of any model trained on a single residue position, implying that explicit correlations between positions do contribute positively to classification.

The above data thus indicate that a deep learning model that combines molecular dynamics data with sequence features can be used to improve peptide immunogenicity prediction with respect to distinguishing pathogen peptides from self. Importantly, the addition of high-level MD features serves to add to the performance of a sequence-only baseline computed on identical datasets, suggesting that MD-derived improvements do not result from trivial similarities across train-test divides.

We next interrogated how well our convolutional model can be extended to classifying cancer neoantigens, a task that is useful for many therapeutic applications. For example, neoantigen classification would be useful if one would like to discriminate between immunogenic and non-

immunogenic neopeptides for antigen selection in vaccine design, an approach that goes beyond just making generating HLA-I binding predictions. Though our initial model was trained on features from 17 HLA-A*02:01 immunogenic neoantigens(15), it was not exposed to any non-immunogenic neopeptides, which themselves may be distinct from standard self-peptides. Accordingly, we conducted a full MD simulation and featurization on 15 non-immunogenic neopeptides (and one additional immunogenic neopeptide, yielding 18 immunogenic peptides and 33 peptides in total) to create a neoantigen test set (hereafter called “Small Neo”), and evaluated this test set using on models trained on a range of training sets and features. For reference, the non-self/self (NS/S) MD training set listed in Supplementary Table 3 includes the 95 immunogenic pathogen and 123 self-non-immunogenic peptides fully featurized with MD, and is used as the single training set for the remainder MD-based models discussed in this work. Sequence-only models were constructed using all relevant training sets, and the 291 neoantigen dataset presented in Riley et al.(28) (hereafter known as “Full Neo”) was used as an additional test set for the purpose of comparison. Single MD plus sequence models were trained on the NS/S MD training set; for completeness, MD-only models were trained only on MD-derived features.

Not too surprisingly, the sequence-only models trained on either large peptide training (n=2883/n=3955) set could not classify either of the reduced (“Small Neo”; n=33) and full (“Full Neo”; n=291) neoantigen test data sets (Supplementary Table 2). This lack of neoantigen classification is consistent with previous results(28), where an AUC of 0.50 was reported for a sequence-only model and 0.60 for a structural model on the full neoantigen test set. Our smaller, diversity-conditioned peptide training set yields better classification on both neoantigen test sets (AUC = 0.67 – Small Neo; AUC = 0.57 – Full Neo) (Supplementary Table 3). Our sequence-only result on the full neoantigen set approximates that of previously described more expensive structural model(28) (See Supplementary Information).

Adding MD-derived features to our reduced NS/S training set and small neoantigen test set improved the AUC from 0.67 to 0.71 (Supplementary Table 3). This result falls far short of the 0.89 reported from cross-validation, suggesting that neoantigens are much more challenging to differentiate in terms of immunogenicity than peptides from pathogens. This observation could have origins in the fact that neoantigens are derived from the human genome, and thus might be more similar to one another than or have structural and dynamic features distinct from pathogen and human self peptides. However, an AUC of 0.71 still represents a reasonable level of classification, and the MD plus sequence, MD-only, and sequence-only results suggest that those feature sets can cooperate in determining immunogenicity. Fig. 4a shows box-and-whisker plots of predicted probabilities for the immunogenic and non-immunogenic peptides in the small neoantigen test set derived from the best-performing MD plus sequence model. The model does differentiate the two classes of peptides with reasonable separation: the average predicted probability for immunogenic samples is 0.69, while the average value for non-immunogenic samples is 0.42. By contrast, the corresponding values are approximately 0.90 and 0.05 in models trained and tested

on pathogen/self peptides. These differences may suggest that non-immunogenic peptides share features correlated with immunogenicity and that immunogenic neoepitopes diverge less from nonimmunogenic neoepitopes than pathogen peptides diverge from self peptides.

To further assess whether our model could identify immunogenic cancer neoepitopes among HLA-A*02:01 binders from individual tumor genomes, we generated neoantigen test sets from peptides derived from two individual patients with melanoma (n=8 peptides for "Patient 1" (38), and n=15 peptides for "Patient 2"(23)). The patient-specific test results are shown in Fig 4b and Supplementary Table 3. Intriguingly, the results for each patient diverge with respect to dominant feature type. However, the results for both patients were favorable overall (Fig. 3b); with the MD plus sequence model, both of Patient 1's top ranked peptides and three of the top four-ranked Patient 2 peptides were identified as immunogenic. It is important to note that these patients' data were obtained under specific conditions and limited by screening assay sensitivity and the presence of potential immune evasion. It is nonetheless encouraging that our model's predictions feature strong enrichment of immunogenic peptides, providing an example of how neopeptides might be prioritized in personalized vaccine design.

Fig. 3c provides two anecdotal examples in which MD-derived features appear to improve classification outcomes. In the case of AVGSYVYSV, the MD simulation reveals conformational dynamics that expose Y5, a residue for which such exposure is likely correlated with immunogenicity (and indeed seems to advance the immunogenicity probability prediction from 6% in the sequence-only model to 99% in the MD plus sequence model). By contrast, the non-immunogenic LMASISSFL features large, hydrophobic residues at positions 5 and 8, which triggers a strong immunogenicity prediction in the sequence-only model (97%); however, the MD simulation has the effect of burying F8, reducing the predicted immunogenicity probability to 31% in the MD-only model. The presence of I5/F8 seems to be so strongly correlated with immunogenicity that even the MD plus sequence model produces a high immunogenicity probability (94%). While neural networks learn from a complicated array of internal representations that are often difficult to interpret, these two peptides provide examples of how MD-derived features can improve immunogenicity prediction in a deep learning framework.

To expand our exploration of non-immunogenic neoantigens, we featurized the full (n=274) non-immunogenic neoantigen set by collecting a single 50 ns sample for each peptide. All peptides were equilibrated for 25 ns prior to production. Combining these non-immunogenic neopeptide data with those previously generated for 17 immunogenic neoantigens and feeding all samples into the MD plus sequence model, we see a modest improvement in AUC relative to the sequence only model (0.61 vs. 0.57). These results are consistent with those reported by Riley et al.(28) for structural featurization (0.60). It is possible that a full (10 sample) featurization could improve our model's performance on these

data, though the improvement derived from MD is consistent with that observed for the neoantigens that were fully featurized (0.71 vs. 0.67).

These results on the full non-immunogenic neoantigen set raises the question of why neopeptides are apparently more difficult to discriminate than immunogenic pathogen and self non-immunogenic peptides. A part of the answer may lie in the characteristics of the non-immunogenic neopeptides available for study. Our MD featurization of the 274 neoantigens allows for a direct comparison of A02-presented peptide interfaces with those from other peptide classes. Non-immunogenic neopeptide SASA differences relative to the same self-peptide set featured in Fig. 2b are shown in Fig 4d. Surprisingly, the presented interfaces of these non-immunogenic neopeptides are quite similar to those of immunogenic pathogen peptides, with respect to both our chosen hydrophobic and complementary residue exposure metrics.

Discussion

Our simulation data on non-immunogenic neoantigens thus suggest that cancer neopeptide immunogenicity exists on a spectrum, and that neoantigens labeled as “negative” by common assays appear more immunogenic (in terms of HLA-I presentation) than standard human/self controls. Accordingly, our deep neural network must pick up on much more subtle differences between presented neoantigens to identify nonimmunogenic peptides than those that are typically reliable for distinguishing pathogen peptides from self. In short, the non-immunogenic neoantigens considered here are presented in strikingly similar ways to immunogenic peptides, complicating the classification problem. Interestingly, this difficulty perhaps mirrors the physiological characteristics of neoantigen recognition: while robust and durable T cell responses are normally observed in the presence of pathogen peptides, the same often cannot be said for cancer neoantigens.

Despite these complications, our current model has proven capable of ranking true-positive immunogenic peptides at or near the top of patient-specific neoantigen sets, an important ability for vaccine design. It remains possible that features orthogonal to current representations of sequence and solvent exposure may play a significant role in determining neoantigen immunogenicity. More feature-agnostic deep learning architectures like graph convolutions could perhaps also help identify such differentiating factors.

We believe our results add to knowledge of specific positional features of the HLA-I bound peptides that are associated with immunogenicity. We observe the most significant hydrophobic exposure occurs at positions 5 and 8, whereas Riley et al.(28) report their most certain enhancement at position 4. Since

both the previous Riley *et al.*(28) study and the present work identify the importance of nuanced hydrophobic exposure with respect to observed immunogenicity, a more rigorous featurization approach, like the current full MD simulation in explicit solvent, is helpful to improve immunogenicity prediction. It's also important to note that molecular dynamics simulations produce more physically rigorous conformational ensembles that better approximate dynamic functional context.

Single-sample (50 ns) MD trajectories of a peptide-HLA-I complex take approximately one node-day to collect on our present hardware, which makes featurizing datasets of the size presented here expensive but tractable. To further accelerate featurization going forward, advances in generative deep learning (e.g., dynamical autoencoders) are worth exploring in the context of generating MD-derived features with a neural network rather than molecular dynamics code(39, 40). Other recent work has shown that HLA allele-dependent factors can have a considerable (and potentially negative) impact on neoantigen presentation(41), suggesting that the impact of various properties of the cancer peptidome, and HLA diversity(42), need to be assessed in future immunogenicity prediction. Pressing needs related to vaccine development for emerging viruses suggest additional use cases for our models in antigen selection, should such advanced strategies be required. The present work suggests that combining molecular dynamics with artificial intelligence can play a foundational role in immunogenicity prediction in difficult biomedical applications, which are certainly not limited to personalized cancer vaccine design.

Methods

Construction of Datasets

The peptides used to train and test models in this study were all of length nine. For analyses involving pathogen peptides, all peptides restricted to the HLA-I supertype A02 were derived from the Immune Epitope Database(29) (IEDB) under T cell assays, last accessed July 2017. Peptides were considered immunogenic if they had an annotated T cell response (column "Qualitative Measure") as positive (N = 1038). HLA-A02 self non-immunogenic binders (N = 1845) were derived from Pearson *et al.*(30), in which mass-spectrometry was used to elute human self peptides bound to HLA-A*02:01 from the cell surface. Neoantigen data were obtained from four main sources. Immunogenic and non-immunogenic neoantigens restricted to HLA A*02:01 were obtained from Bjerregaard *et al.*(15) (n=18 immunogenic; 15 non-immunogenic). and Riley *et al.* (n=291) (28). For the analyses ranking Patient 1 and Patient 2 immunogenic neoepitopes, data were obtained from Robbins *et al.*(38) (n=8) and Stronen *et al.*(23) (n=15). respectively. The reduced dataset on which full MD featurization was selected to maximize sequence diversity at peptide positions 1 and 8, so as to ensure a range of amino acid compositions at both peptide termini. As such, peptides sets were selected from the larger self and non-self databases with roughly equal proportions of available amino acids at the aforementioned positions.

Deep convolutional neural network based on peptide sequence only

Once peptide sequences have been encoded as described in the main text, the one-hot sequence representations are run through a 32-filter convolutional layer using the filter 2 x 2 configuration described above (Fig. 1c). The resulting mixed features are then max-pooled across the same filter dimensions to extract the dominant amplitudes from the convolutional procedure, as one often does in image- or graph-based convolutional neural networks. The pooled features are next passed through an L2-regularized dense layer, which is then subjected to a 50% dropout procedure during training. Finally, the surviving output from the dense layer is passed through a softmax layer to yield peptide immunogenicity probabilities. Sequence-based deep learning models were implemented in TensorFlow 1.2.1.

Deep convolutional neural network combining sequential, structural, and dynamical features

MD trajectories were collected as described in the main text using NAMD 2.9.1 using the CHARMM27 force field, Langevin thermostat (damping at 1/ps), a Langevin Piston Barostat, and 2 fs time step. All systems were solvated in explicit TIP3P water and ionized to 150 mM Na⁺ and Cl⁻. Prior to full equilibration, all systems were annealed with three iterations of 50000 steepest descent minimization steps and 5 ns molecular dynamics at 310 K and 1 atm. After collection of production simulations, residue-specific solvent-accessible surface areas (SASAs) were computed every 1 ns using the measure tool in VMD 1.9.2, split into hydrophobic/nohydrophobic channels, normalized by the maximum residue-specific SASA observed per channel across all peptides used in training/testing, and split into 10 (50 ns) samples per peptide (with the exception of the full non-immunogenic neoantigen set, which contained only one 50 ns sample per peptide). Molecular dynamics data samples were then each fed into a series of three convolutional/pooling units, leveraging 2 x 2 convolutional filters and pooling along the time axis across 33, 33, and 35 frames to yield a 2 x 9 x 128 internal representation for each channel image. In the third convolutional unit, max pooling was replaced with mean pooling to mimic the ensemble averaging that is known to characterize the underlying statistical physics. The hydrophobic and complementary representations are then flattened, concatenated, and fed through a dense neural network layer. The output of this dense layer is concatenated with the sequence-only dense layer output; this vector is finally fed through a softmax layer to generate immunogenicity probability predictions. Integrated deep learning models were implemented in TensorFlow 1.2.1.

Deep convolutional neural network training procedure

For applications not involving neoantigens, a five-fold cross validation procedure was applied in an attempt reproduce the method used by Riley et al. for immunogenicity classification. As described by Riley et al., datasets were randomly split into five equal parts, and models were trained on all possible permutations of three-part training sets, one-part validation sets, and one-part test sets. In practice,

hyperparameters were not changed as a result of training runs conducted on each validation set: all L2 regularization weights were set at standard value of 0.01 for, and all sequence and MD plus sequence models were trained for 200000 and 500000 optimization steps, respectively, with a batch size of 50 samples. Stochastic gradient descent was governed by the default settings of the ADAM optimizer in TensorFlow. Standard ReLU activation functions were used in convolutional and dense layers. Cross validation AUCs were reported as averages over withheld test sets. For applications testing neoantigen classification, models were trained with all available pathogen/self peptide samples using the same procedures described above. Out of an abundance of caution, pathogen peptides with the ambiguous label “occurrence of autoimmune disease” were excluded from training data for models used to make neoantigen predictions. In models involving molecular dynamics data, no samples drawn from the same peptide were divided between any train/test split. In other words, in cases in which 10 samples per peptide were available, all ten samples were either used as training data or test data, but never a mixture of both.

Statistical analyses

Statistical analyses were performed with available functions in TensorFlow 1.2.1 and standard tools packaged with Python 2.7.12 and MATLAB_R2015b.

Data availability

All peptides used in the present study are provided as a supplementary file.

Declarations

Author Contributions

R.Z. and T.A.C. conceived and planned the study. J.K.W. designed and constructed the deep learning architecture, and performed molecular dynamics simulations. D.C. and C.K. collected and prepared all peptide sequences and patient data. D.C. and C.K. helped develop the deep learning architecture for sequence only model. J.K.W., D.C. and C.K. analyzed data. J.K.W., D.C., C.K., T.A.C. and R.Z. interpreted results and co-wrote the manuscript. All authors contributed to the general discussion of the project and manuscript.

Acknowledgements

The authors wish to acknowledge helpful discussions with David Bell, Joseph Morrone, Leili Zhang, and Binqun Luan. This work was partially supported by W. M. Keck Foundation Grant (R.Z.), Pershing Square

Sohn Cancer Research Alliance (T.A.C), STARR Cancer Consortium (T.A.C), Stand Up To Cancer (T.A.C), and NCI R35 CA232097 (T.A.C.), and MSKCC through NIH/NCI Cancer Center Support Grant (P30 CA008748). R.Z. also gratefully acknowledges the financial support from the IBM Bluegene Science Program (W125859, W1464125 and W1464164).

Competing interests

T.A.C. is a co-founder of Gritstone Oncology and holds equity. T.A.C. holds equity in An2H. T.A.C. acknowledges grant funding from Bristol-Myers Squibb, AstraZeneca, Illumina, Pfizer, An2H and Eisai. T.A.C. has served as an advisor for Bristol-Myers, MedImmune, Squibb, Illumina, Eisai, AstraZeneca and An2H. N.A.R. is a consultant/advisory board member for AstraZeneca, BMS, Roche, Merck, Novartis, Lilly and Pfizer. T.A.C. and D.C. hold ownership of intellectual property on using TMB to predict immunotherapy response, with pending patent, which has been licensed to PGDx.

References

1. N. Hilf, S. Kuttruff-Coqui, K. Frenzel, V. Bukur, S. Stevanovic, C. Gouttefangeas, M. Platten, G. Tabatabai, V. Dutoit, S. H. van der Burg, P. Thor Straten, F. Martinez-Ricarte, B. Ponsati, H. Okada, U. Lassen, A. Admon, C. H. Ottensmeier, A. Ulges, S. Kreiter, A. von Deimling, M. Skardelly, D. Migliorini, J. R. Kroep, M. Idorn, J. Rodon, J. Piro, H. S. Poulsen, B. Shraibman, K. McCann, R. Mendrzyk, M. Lower, M. Stieglbauer, C. M. Britten, D. Capper, M. J. P. Welters, J. Sahuquillo, K. Kiesel, E. Derhovanessian, E. Rusch, L. Bunse, C. Song, S. Heesch, C. Wagner, A. Kemmer-Bruck, J. Ludwig, J. C. Castle, O. Schoor, A. D. Tadmor, E. Green, J. Fritsche, M. Meyer, N. Pawlowski, S. Dorner, F. Hoffgaard, B. Rossler, D. Maurer, T. Weinschenk, C. Reinhardt, C. Huber, H. G. Rammensee, H. Singh-Jasuja, U. Sahin, P. Y. Dietrich, W. Wick, Actively personalized vaccination trial for newly diagnosed glioblastoma. *Nature* **565**, 240-245 (2019).
2. P. A. Ott, Z. Hu, D. B. Keskin, S. A. Shukla, J. Sun, D. J. Bozym, W. Zhang, A. Luoma, A. Giobbie-Hurder, L. Peter, C. Chen, O. Olive, T. A. Carter, S. Li, D. J. Lieb, T. Eisenhaure, E. Gjini, J. Stevens, W. J. Lane, I. Javeri, K. Nellaiappan, A. M. Salazar, H. Daley, M. Seaman, E. I. Buchbinder, C. H. Yoon, M. Harden, N. Lennon, S. Gabriel, S. J. Rodig, D. H. Barouch, J. C. Aster, G. Getz, K. Wucherpfennig, D. Neuberg, J. Ritz, E. S. Lander, E. F. Fritsch, N. Hacohen, C. J. Wu, An immunogenic personal neoantigen vaccine for patients with melanoma. *Nature* **547**, 217-221 (2017).
3. U. Sahin, E. Derhovanessian, M. Miller, B. P. Kloke, P. Simon, M. Lower, V. Bukur, A. D. Tadmor, U. Luxemburger, B. Schrors, T. Omokoko, M. Vormehr, C. Albrecht, A. Paruzynski, A. N. Kuhn, J. Buck, S. Heesch, K. H. Schreeb, F. Muller, I. Ortseifer, I. Vogler, E. Godehardt, S. Attig, R. Rae, A. Breitkreuz, C. Tolliver, M. Suchan, G. Martic, A. Hohberger, P. Sorn, J. Diekmann, J. Ciesla, O. Waksman, A. K. Bruck, M. Witt, M. Zillgen, A. Rothermel, B. Kasemann, D. Langer, S. Bolte, M. Diken, S. Kreiter, R. Nemecek, C. Gebhardt, S. Grabbe, C. Holler, J. Utikal, C. Huber, C. Loquai, O. Tureci, Personalized RNA mutanome vaccines mobilize poly-specific therapeutic immunity against cancer. *Nature* **547**, 222-226 (2017).

4. B. M. Carreno, V. Magrini, M. Becker-Hapak, S. Kaabinejadian, J. Hundal, A. A. Petti, A. Ly, W. R. Lie, W. H. Hildebrand, E. R. Mardis, G. P. Linette, Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. *Science* **348**, 803-808 (2015).
5. E. Tran, M. Ahmadzadeh, Y. C. Lu, A. Gros, S. Turcotte, P. F. Robbins, J. J. Gartner, Z. Zheng, Y. F. Li, S. Ray, J. R. Wunderlich, R. P. Somerville, S. A. Rosenberg, Immunogenicity of somatic mutations in human gastrointestinal cancers. *Science* **350**, 1387-1390 (2015).
6. N. van Rooij, M. M. van Buuren, D. Philips, A. Velds, M. Toebes, B. Heemskerk, L. J. van Dijk, S. Behjati, H. Hilkmann, D. El Atmioui, M. Nieuwland, M. R. Stratton, R. M. Kerkhoven, C. Kesmir, J. B. Haanen, P. Kvistborg, T. N. Schumacher, Tumor exome analysis reveals neoantigen-specific T-cell reactivity in an ipilimumab-responsive melanoma. *J Clin Oncol* **31**, e439-442 (2013).
7. The problem with neoantigen prediction. *Nat Biotechnol* **35**, 97 (2017).
8. J. G. Abelin, D. B. Keskin, S. Sarkizova, C. R. Hartigan, W. Zhang, J. Sidney, J. Stevens, W. Lane, G. L. Zhang, T. M. Eisenhaure, K. R. Clauser, N. Hacohen, M. S. Rooney, S. A. Carr, C. J. Wu, Mass Spectrometry Profiling of HLA-Associated Peptidomes in Mono-allelic Cells Enables More Accurate Epitope Prediction. *Immunity* **46**, 315-326 (2017).
9. J. Racle, J. Michaux, G. A. Rockinger, M. Arnaud, S. Bobisse, C. Chong, P. Guillaume, G. Coukos, A. Harari, C. Jandus, M. Bassani-Sternberg, D. Gfeller, Robust prediction of HLA class II epitopes by deep motif deconvolution of immunopeptidomes. *Nat Biotechnol*, (2019).
10. M. Yadav, S. Jhunjunwala, Q. T. Phung, P. Lupardus, J. Tanguay, S. Bumbaca, C. Franci, T. K. Cheung, J. Fritsche, T. Weinschenk, Z. Modrusan, I. Mellman, J. R. Lill, L. Delamarre, Predicting immunogenic tumour mutations by combining mass spectrometry and exome sequencing. *Nature* **515**, 572+ (2014).
11. B. Bulik-Sullivan, J. Busby, C. D. Palmer, M. J. Davis, T. Murphy, A. Clark, M. Busby, F. Duke, A. Yang, L. Young, N. C. Ojo, K. Caldwell, J. Abhyankar, T. Boucher, M. G. Hart, V. Makarov, V. T. De Montpreville, O. Mercier, T. A. Chan, G. Scagliotti, P. Bironzo, S. Novello, N. Karachaliou, R. Rosell, I. Anderson, N. Gabrail, J. Hrom, C. Limvarapuss, K. Choquette, A. Spira, R. Rousseau, C. Voong, N. A. Rizvi, E. Fadel, M. Frattini, K. Jooss, M. Skoberne, J. Francis, R. Yelensky, Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification. *Nature Biotechnology* **37**, 55+ (2019).
12. V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, NetMHCpan-4.0: Improved Peptide-MHC Class I Interaction Predictions Integrating Eluted Ligand and Peptide Binding Affinity Data. *J Immunol* **199**, 3360-3368 (2017).
13. M. Bassani-Sternberg, E. Braunlein, R. Klar, T. Engleitner, P. Sinitcyn, S. Audehm, M. Straub, J. Weber, J. Slotta-Huspenina, K. Specht, M. E. Martignoni, A. Werner, R. Hein, H. B. D, C. Peschel, R. Rad, J. Cox, M. Mann, A. M. Krackhardt, Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry. *Nat Commun* **7**, 13404 (2016).
14. M. C. W. Feltkamp, M. P. M. Vierboom, W. M. Kast, C. J. M. Melief, Efficient Mhc Class I-Peptide Binding Is Required but Does Not Ensure Mhc Class I-Restricted Immunogenicity. *Molecular*

- immunology* **31**, 1391-1401 (1994).
15. A. M. Bjerregaard, M. Nielsen, V. Jurtz, C. M. Barra, S. R. Hadrup, Z. Szallasi, A. C. Eklund, An Analysis of Natural T Cell Responses to Predicted Tumor Neoepitopes. *Front Immunol* **8**, 1566 (2017).
 16. D. Chowell, S. Krishna, P. D. Becker, C. Cocita, J. Shu, X. F. Tan, P. D. Greenberg, L. S. Klavinskis, J. N. Blattman, K. S. Anderson, TCR contact residue hydrophobicity is a hallmark of immunogenic CD8(+) T cell epitopes. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E1754-E1762 (2015).
 17. J. J. Calis, M. Maybeno, J. A. Greenbaum, D. Weiskopf, A. D. De Silva, A. Sette, C. Kesmir, B. Peters, Properties of MHC class I presented peptides that enhance immunogenicity. *PLoS Comput Biol* **9**, e1003266 (2013).
 18. F. Duan, J. Duitama, S. Al Seesi, C. M. Ayres, S. A. Corcelli, A. P. Pawashe, T. Blanchard, D. McMahon, J. Sidney, A. Sette, B. M. Baker, Mandoiu, II, P. K. Srivastava, Genomic and bioinformatic profiling of mutational neoepitopes reveals new rules to predict anticancer immunogenicity. *J Exp Med* **211**, 2231-2248 (2014).
 19. E. F. Fritsch, M. Rajasagi, P. A. Ott, V. Brusic, N. Hacohen, C. J. Wu, HLA-binding properties of tumor neoepitopes in humans. *Cancer immunology research* **2**, 522-529 (2014).
 20. L. P. Richman, R. H. Vonderheide, A. J. Rech, Neoantigen Dissimilarity to the Self-Proteome Predicts Immunogenicity and Response to Immune Checkpoint Blockade. *Cell Syst* **9**, 375-382 e374 (2019).
 21. S. Kim, H. S. Kim, E. Kim, M. G. Lee, E. Shin, S. Paik, S. Kim, Neopepsee: accurate genome-level prediction of neoantigens by harnessing sequence and amino acid immunogenicity information. *Ann Oncol*, (2018).
 22. J. S. Lee, E. Ruppin, Multiomics Prediction of Response Rates to Therapies to Inhibit Programmed Cell Death 1 and Programmed Cell Death 1 Ligand 1. *Jama Oncol*, (2019).
 23. E. Stronen, M. Toebes, S. Kelderman, M. M. van Buuren, W. Yang, N. van Rooij, M. Donia, M. L. Boschen, F. Lund-Johansen, J. Olweus, T. N. Schumacher, Targeting of cancer neoantigens with donor-derived T cell receptor repertoires. *Science* **352**, 1337-1341 (2016).
 24. M. M. Davis, P. J. Bjorkman, T-Cell Antigen Receptor Genes and T-Cell Recognition. *Nature* **334**, 395-402 (1988).
 25. M. Krogsaard, M. M. Davis, How T cells 'see' antigen. *Nature Immunology* **6**, 239-245 (2005).
 26. M. G. Rudolph, R. L. Stanfield, I. A. Wilson, How TCRs bind MHCs, peptides, and coreceptors. *Annual Review of Immunology* **24**, 419-466 (2006).
 27. A. V. Joglekar, Z. Liu, J. K. Weber, Y. Ouyang, J. D. Jeppson, W. J. Noh, P. A. Lamothe-Molina, H. Chen, S. G. Kang, M. T. Bethune, R. Zhou, B. D. Walker, D. Baltimore, T cell receptors for the HIV KK10 epitope from patients with differential immunologic control are functionally indistinguishable. *Proc Natl Acad Sci U S A* **115**, 1877-1882 (2018).
 28. T. P. Riley, G. L. J. Keller, A. R. Smith, L. M. Davancaze, A. G. Arbuiso, J. R. Devlin, B. M. Baker, Structure Based Prediction of Neoantigen Immunogenicity. *Front Immunol* **10**, 2047 (2019).

29. R. Vita, J. A. Overton, J. A. Greenbaum, J. Ponomarenko, J. D. Clark, J. R. Cantrell, D. K. Wheeler, J. L. Gabbard, D. Hix, A. Sette, B. Peters, The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* **43**, D405-412 (2015).
30. H. Pearson, T. Daouda, D. P. Granados, C. Durette, E. Bonneil, M. Courcelles, A. Rodenbrock, J. P. Laverdure, C. Cote, S. Mader, S. Lemieux, P. Thibault, C. Perreault, MHC class I-associated peptides derive from selective regions of the human genome. *Journal of Clinical Investigation* **126**, 4690-4701 (2016).
31. B. J. Berne, J. D. Weeks, R. H. Zhou, Dewetting and Hydrophobic Interaction in Physical and Biological Systems. *Annual Review of Physical Chemistry* **60**, 85-103 (2009).
32. B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, M. Karplus, Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J Comput Chem* **4**, 187-217 (1983).
33. G. Bussi, M. Parrinello, Stochastic thermostats: comparison of local and global schemes. *Comput Phys Commun* **179**, 26-29 (2008).
34. W. Humphrey, A. Dalke, K. Schulten, VMD: Visual molecular dynamics. *J Mol Graph Model* **14**, 33-38 (1996).
35. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **79**, 926-935 (1983).
36. A. D. Mackerell, M. Feig, C. L. Brooks, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *J Comput Chem* **25**, 1400-1415 (2004).
37. K. Maenaka, T. Juji, D. I. Stuart, E. Y. Jones, Crystal structure of the human p58 killer cell inhibitory receptor (KIR2DL3) specific for HLA-Cw3-related MHC class I. *Structure* **7**, 391-398 (1999).
38. P. F. Robbins, Y. C. Lu, M. El-Gamil, Y. F. Li, C. Gross, J. Gartner, J. C. Lin, J. K. Teer, P. Cliften, E. Tycksen, Y. Samuels, S. A. Rosenberg, Mining exomic sequencing data to identify mutated antigens recognized by adoptively transferred tumor-reactive T cells. *Nat Med* **19**, 747-752 (2013).
39. C. X. Hernandez, H. K. Wayment-Steele, M. M. Sultan, B. E. Husic, V. S. Pande, Variational encoding of complex dynamics. *Phys Rev E* **97**, (2018).
40. A. Mardt, L. Pasquali, H. Wu, F. Noe, VAMPnets for deep learning of molecular kinetics (vol 9, 5, 2018). *Nature Communications* **9**, (2018).
41. D. Chowell, L. G. T. Morris, C. M. Grigg, J. K. Weber, R. M. Samstein, V. Makarov, F. Kuo, S. M. Kendall, D. Requena, N. Riaz, B. Greenbaum, J. Carroll, E. Garon, D. M. Hyman, A. Zehir, D. Solit, M. Berger, R. Zhou, N. A. Rizvi, T. A. Chan, Patient HLA class I genotype influences cancer response to checkpoint blockade immunotherapy. *Science* **359**, 582-587 (2018).
42. D. Chowell, C. Krishna, F. Pierini, V. Makarov, N. A. Rizvi, F. Kuo, L. G. T. Morris, N. Riaz, T. L. Lenz, T. A. Chan, Evolutionary divergence of HLA class I genotype impacts efficacy of cancer immunotherapy. *Nat Med* **25**, 1715-1720 (2019).

Figures

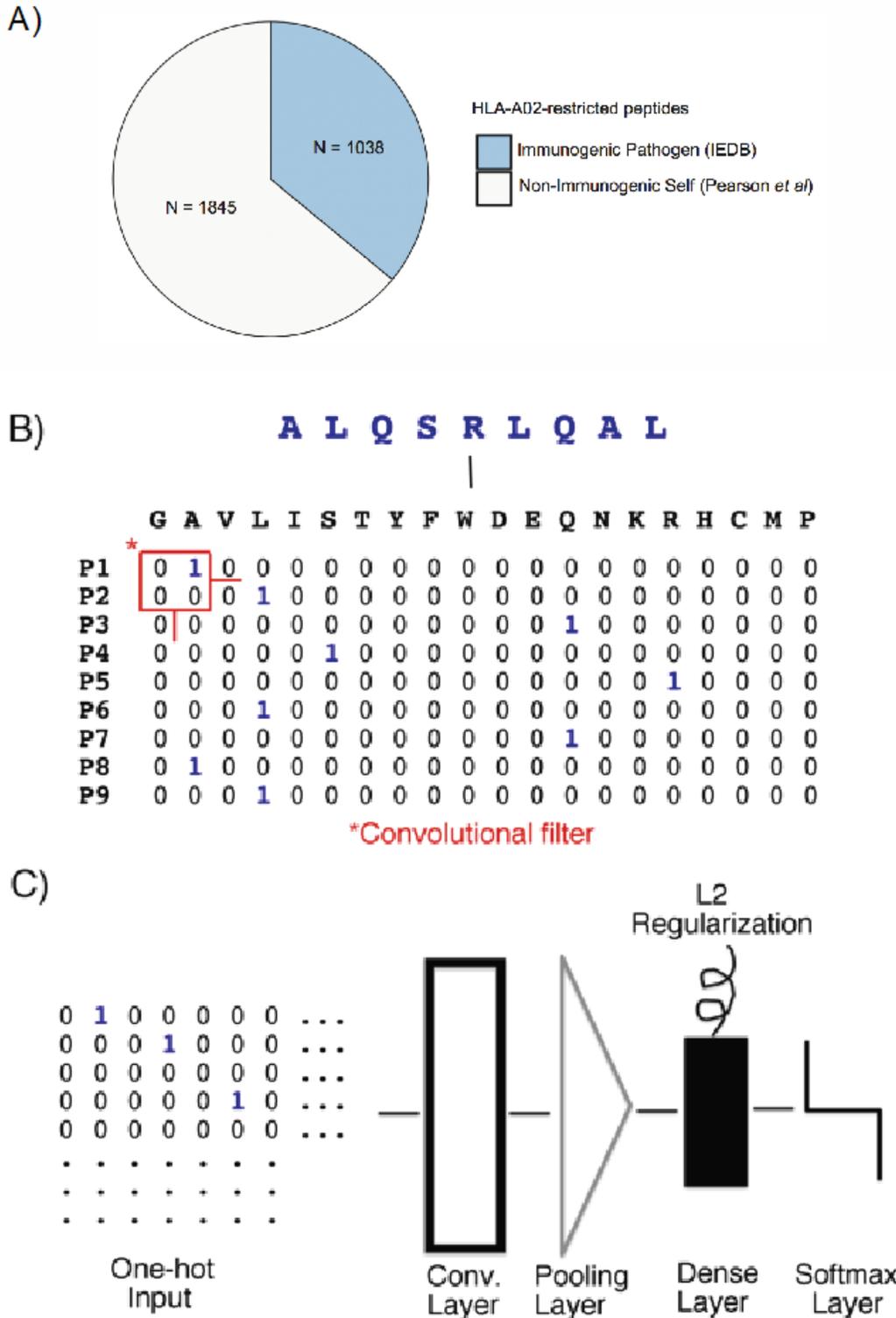


Figure 1

Featurization and architecture for sequence-based deep convolutional neural network. a, Peptide datasets used for neural network training. b, Sequence featurization schematic for input into neural network. c, Neural network architecture for sequence-based immunogenicity prediction.

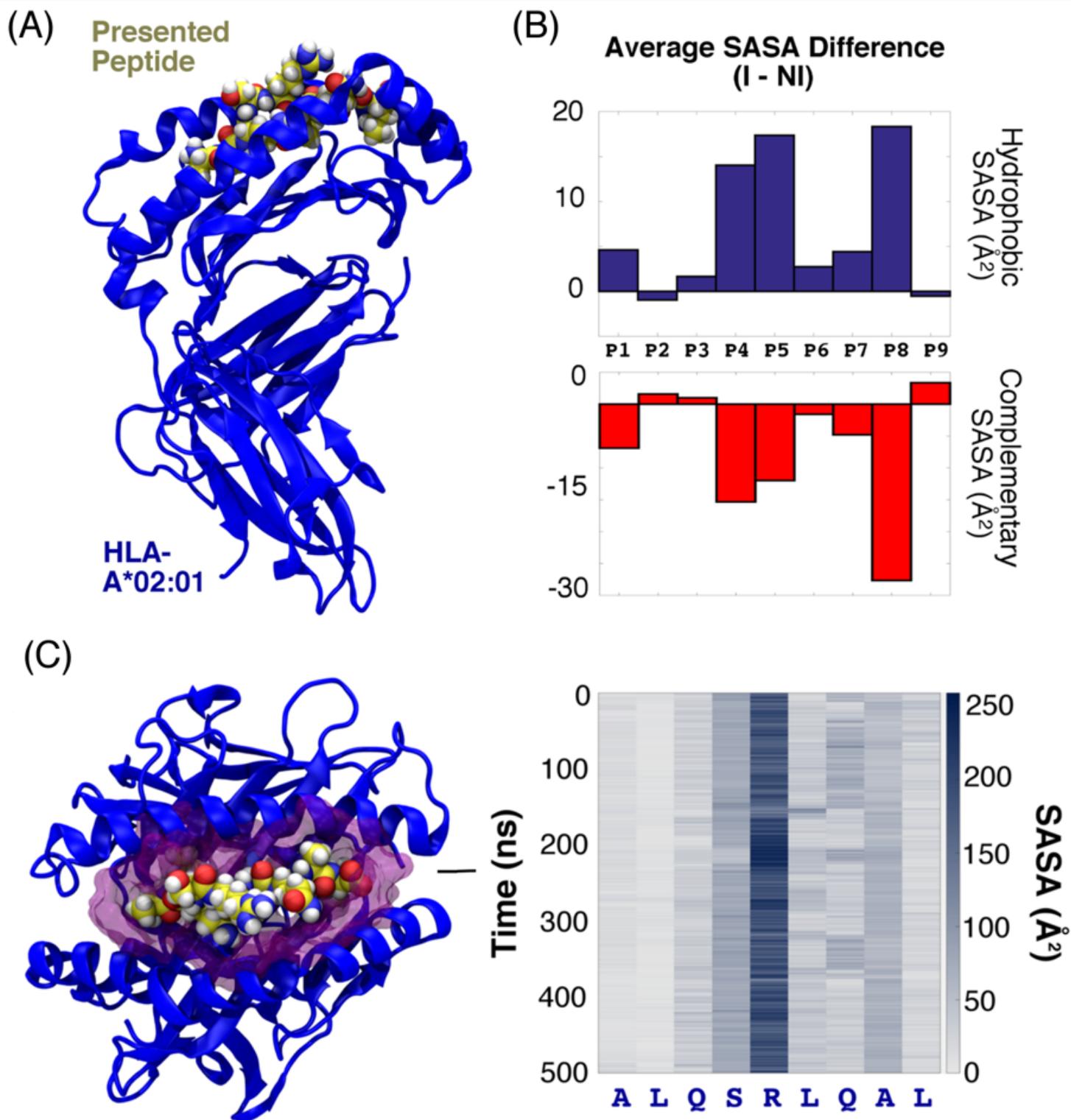


Figure 2

Molecular dynamics trajectory featurization a, Illustration of simulation system. b, Position-specific SASA differences for immunogenic pathogen and self non-immunogenic peptides for hydrophobic (G,A,V,L,I,F,W,Y,P) and complementary residue types. c, Featurization schematic for MD trajectories of HLA-peptide complexes.

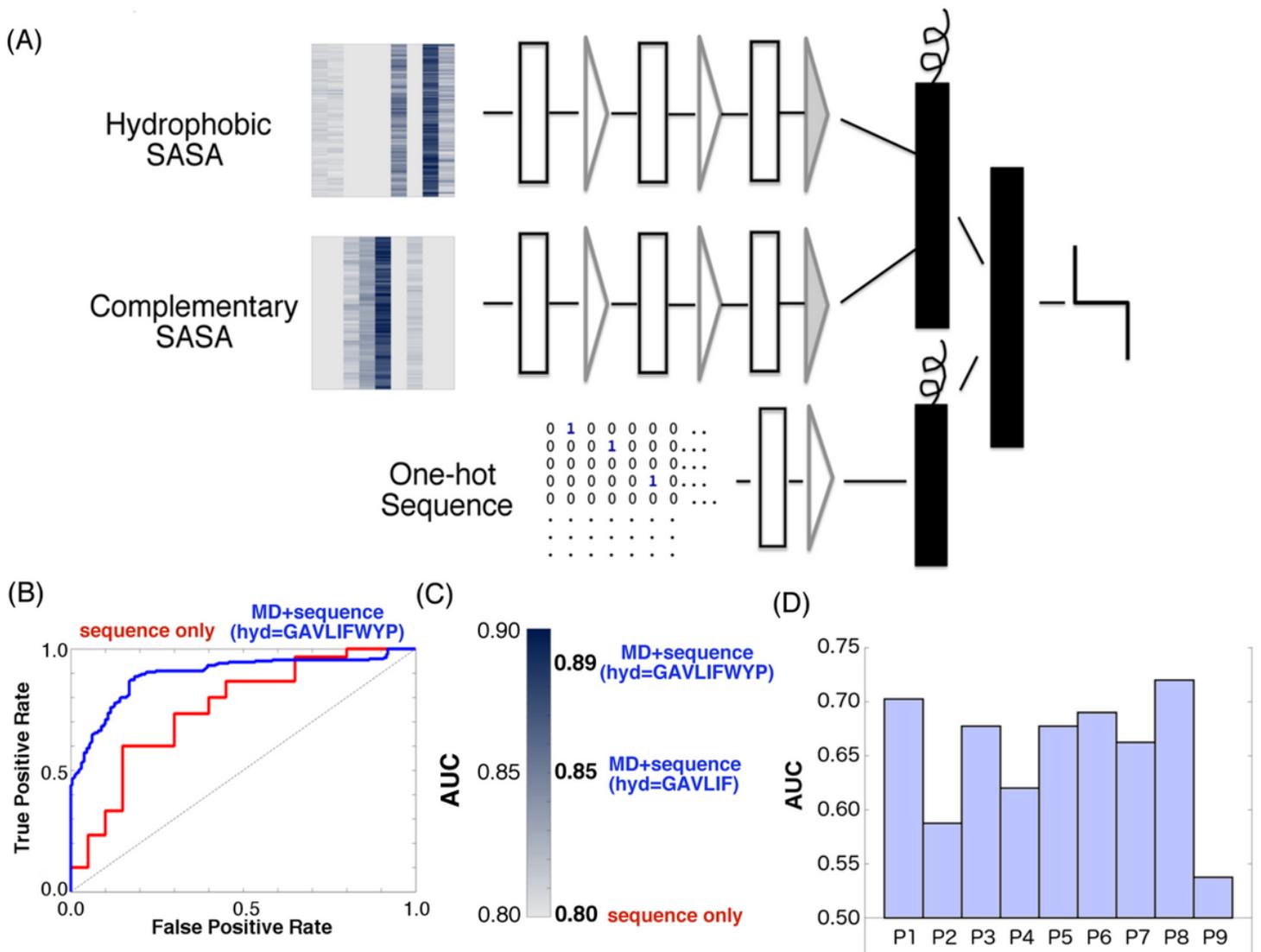


Figure 3

MD-derived immunogenicity prediction architecture and results. a, Convolutional deep learning architecture for immunogenicity prediction. Layer shapes follow the same key presented in Fig. 1; shaded pooling layers correspond to mean pooling. b, Representative ROC curves for classification of an immunogenic pathogen/non-immunogenic self peptide test set for sequence-only and MD plus sequence (hydrophobic = G,A,V,L,I,F,W,Y,P) models. c, Mean test AUC results for five-fold cross validation procedure across sequence-only, MD plus sequence (hydrophobic = G,A,V,L,I,F,W,Y,P), and MD plus sequence (hydrophobic = G,A,V,L,I,F) models. d, Position-based feature ablation study of molecular dynamics data. Each data point corresponds to an independent five-fold cross validation run in which all features but the SASAs corresponding to the indicated position were set to zero.

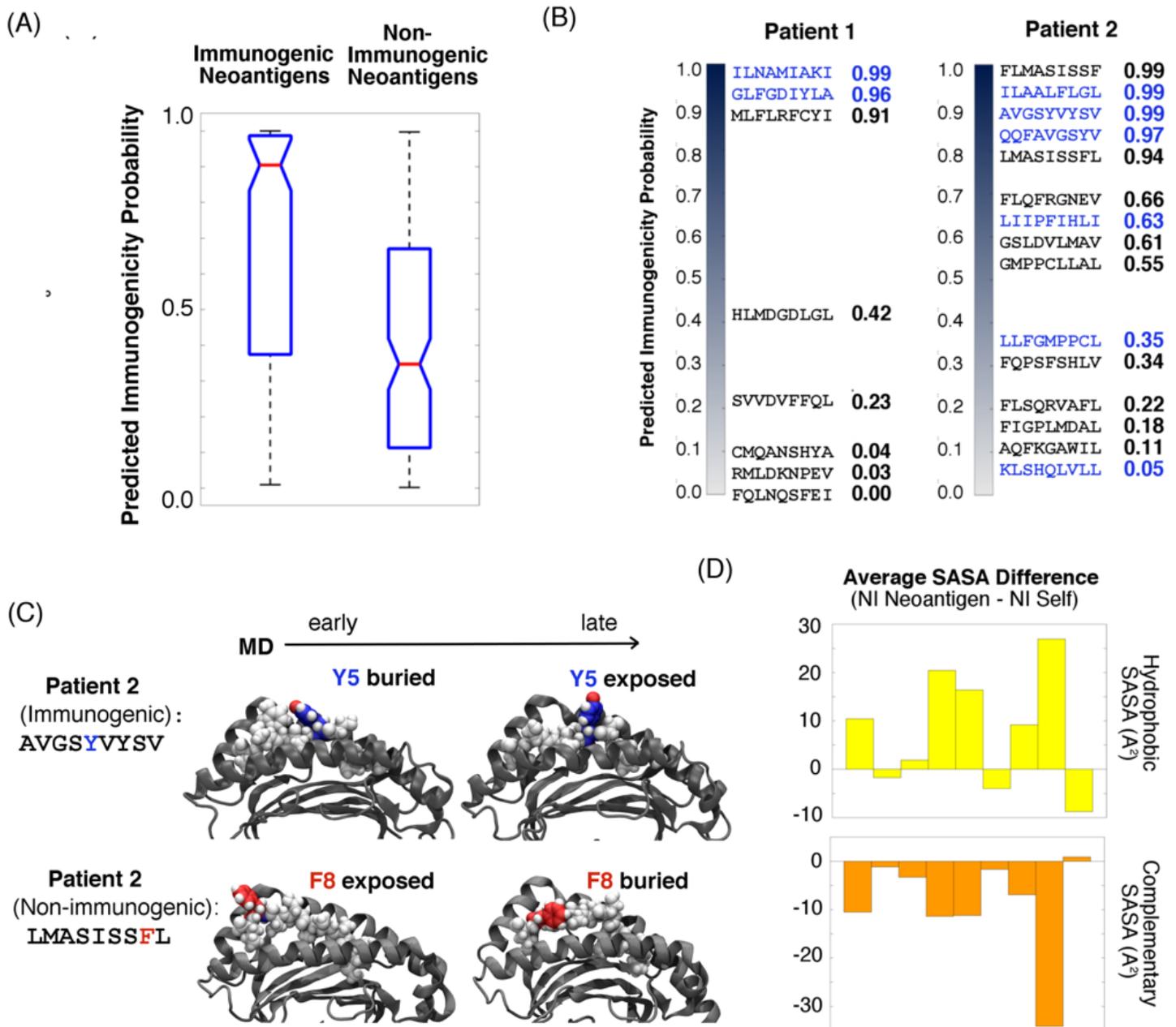


Figure 4

Cancer neoantigen immunogenicity prediction. a, Box-and-whisker plots for predicted immunogenicity probabilities corresponding to 33 immunogenic and non-immunogenic cancer neoantigens. Red lines represent median values. b, De novo classification results for neoantigens derived from two specific patients with cancer as derived from the MD plus sequence model. Peptides marked in blue are true positives, c, Examples of MD-described conformational change that impact immunogenicity predictions. d, Position-specific SASA differences for non-immunogenic neoantigens and self non-immunogenic peptides for hydrophobic (G,A,V,L,I,F,W,Y,P) and complementary residue types.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [S1Immunogenicity.docx](#)