PLOS One

RESEARCH ARTICLE

Machine learning application to predict binding affinity between peptide containing non-canonical amino acids and HLA-A0201

Shan Jiang[1☯], **Zhaoqian Su**[1☯], **Nathaniel Bloodworth**[2], **Yunchao Liu**[1], **Cristina E. Martina**[1], **David G. Harrison**[2], **Jens Meiler**[1,3,4,5*]

1 Department of Chemistry and Center for Structural Biology, Vanderbilt University, Nashville, Tennessee, United States of America, **2** Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tennessee, United States of America, **3** Institute for Drug Discovery, Institute for Computer Science, Wilhelm Ostwald Institute for Physical and Theoretical Chemistry, University Leipzig, Leipzig, Germany, **4** Center for Scalable Data Analytics and Artificial Intelligence ScaDS.AI and School of Embedded Composite Artificial Intelligence SECAI, Dresden/Leipzig, Germany, **5** Department of Pharmacology, Institute of Chemical Biology, Center for Applied Artificial Intelligence in Protein Dynamics, Vanderbilt University, Nashville, Tennessee, United States of America

☯ These authors contributed equally to this work.
* jens@meilerlab.org

Citation: Jiang S, Su Z, Bloodworth N, Liu Y, Martina CE, Harrison DG, et al. (2025) Machine learning application to predict binding affinity between peptide containing non-canonical amino acids and HLA-A0201. PLoS One 20(6): e0314833. https://doi.org/10.1371/journal.pone.0314833

Editor: Yuval Garini, Technion Israel Institute of Technology, ISRAEL

Received: November 18, 2024

Accepted: June 1, 2025

Published: June 27, 2025

Data availability statement: All relevant data for this study are publicly available from the GitHub repository (https://github.com/meilerlab/ML_PLS_MHC_peptide_binding_pred).

Abstract

Class I major histocompatibility complexes (MHC-I), encoded by the highly polymorphic HLA-A, HLA-B, and HLA-C genes in humans, are expressed on all nucleated cells. Both self and foreign proteins are processed to peptides of 8–10 amino acids, loaded into MHC-I, within the endoplasmic reticulum and then presented on the cell surface. Foreign peptides presented in this fashion activate CD8 + T cells and their immunogenicity correlates with their affinity for the MHC-I binding groove. Thus, predicting antigen binding affinity for MHC-I is a valuable tool for identifying potentially immunogenic antigens. While quite a few predictors for MHC-I binding exist, there are no currently available tools that can predict antigen/MHC-I binding affinity for antigens with explicitly labeled post-translational modifications or unusual/non-canonical amino acids (NCAAs). However, such modifications are increasingly recognized as critical mediators of peptide immunogenicity. In this work, we propose a machine learning application that quantifies the binding affinity of epitopes containing NCAAs to MHC-I and compares its performance with other commonly used regressors. Our model demonstrates robust performance, with 5-fold cross-validation yielding an R^2 value of 0.477 and a root-mean-square error (RMSE) of 0.735, indicating strong predictive capability for peptides with NCAAs. This work provides a valuable tool for the computational design and optimization of peptides incorporating NCAAs, potentially accelerating the development of novel peptide-based therapeutics with enhanced properties and efficacy.

Funding: J.M. is supported by a Humboldt Professorship of the Alexander von Humboldt Foundation. J.M. acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG) through SFB1423 (421152132), SFB 1664 (514901783), TRR (514664767), and SPP 2363 (460865652). J.M. is supported by the Federal Ministry of Education and Research (BMBF) through the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), through the German Network for Bioinformatics Infrastructure (de.NBI), and through the German Academic Exchange Service (DAAD) via the School of Embedded Composite AI (SECAI 15766814). Work in the Meiler laboratory is further supported through the National Institute of Health (NIH) through R01 HL122010, R01 DA046138, R01 AG068623, U01 AI150739, R01 CA227833, R01 LM013434, S10 OD016216, S10 OD020154, S10 OD032234, 5T32HL144446-05. D.G.H. is supported by NIH grants IH R35HL140016 and NIH AG076785. N.B. is supported by NIH grant 5T32HL144446-05.

Competing interests: The authors have declared that no competing interests exist.

Introduction

The class I major histocompatibility complex (MHC-I) enables the adaptive immune response by presenting antigens to patrolling cytotoxic T cells [1,2]. Peptides presented by MHC-I originate in the cytoplasm and are usually length limited, having only 8–10 amino acids. This system has evolved principally to enable rapid identification and elimination of viral infected or malignant cells while minimizing the risk of aberrant recognition of self-peptides and consequential autoimmunity [2]. The MHC-I protein products are themselves encoded by the Human Leukocyte Antigen (HLA) genes in humans; both the co-dominantly expressed subtypes (A, B, and C) and the high degree of polymorphism observed in the peptide-binding domain of these genes enable MHC-I to complex with a large repertoire of peptides [2,3]. Post-translational modification of proteins and peptides (resulting in the incorporation of NCAAs) can further broaden the immunogenic landscape of peptides presented by MHC-I. Peptides containing various NCAAs are implicated as immunogens in a variety of diseases [4] including rheumatoid arthritis [5], hypertension and cardiometabolic inflammation [6], and cancer [7].

Recent advances in immunotherapy targeting cancer and autoimmune diseases, coupled with advances in data science have incentivized the creation of computational tools that predict peptides likely to bind to MHC-I and induce immune responses [8,9]. These tools encompass a wide range of methodologies to analyze peptide-MHC interactions. Among these are sequence-based approaches like NetMHCPan [10–12] and MHCflurry [13] that utilize amino acid sequences to forecast binding affinities. Additionally, structure-based approaches such as Rosetta FlexPepDock [14–16] employ three-dimensional structural data to provide a detailed understanding of peptide-MHC binding dynamics and conformational stability. The most advanced and effective of these tools leverage machine learning techniques to construct predictive models. These models are trained on extensive datasets comprising antigen-MHC-I pairs and their corresponding binding affinity data. A significant portion of these data are derived from the Immune Epitope Database (IEDB) [17], which provides a comprehensive repository of experimentally validated immune epitopes. These methods are thoroughly benchmarked and reviewed by Zhao et al [8].

Despite notable advances in both sequence- and structure-based epitope binding predictors, there are currently no tools capable of rapidly predicting antigen/ MHC-I binding affinity for antigens with post-translational modifications or NCAAs. These modifications are increasingly recognized as critical mediators of peptide immunogenicity. The main scope of this research is to develop a new model that would be able to predict the binding affinity of epitopes containing NCAAs to MHC-I.

Machine learning models have demonstrated superior performance in predicting binding affinity due to their ability to capture complex patterns and interactions within the data. The development and refinement of these models involve rigorous processes including feature generation, model training, and validation. Several popular algorithms are widely used for property prediction in the fields of chemistry and biology, including support vector machines (SVM), artificial neural networks (ANNs), principal component analysis (PCA), and partial least squares (PLS) regression [18,19].

In this work, we develop a simple encoder capable of creating feature vectors from peptides based on chemical structure. We then systematically benchmark several different supervised machine learning models on a filtered, publicly available dataset containing peptides with NCAAs and experimentally determined binding affinities.

Methods and results

This section details the study results from three perspectives, data preparation, feature generation and model testing and validation. The data preparation subsection will explain the source and structure of the data used, focusing on data exploration and filtration. The feature generation subsection, the key part of this section, will introduce how peptides with NCAAs are encoded. The model testing and validation subsection will evaluate and compare performance metrics such as R^2 and RMSE across different datasets with five-fold cross-validation using various algorithms.

Data preparation

The initial dataset, a table with 100,141 rows and 29 columns, was exported from the publicly available Immune Epitope Database (IEDB). Among the 29 columns, five are of particular interest for this study: "Name", "Qualitative Measurement", "Quantitative Measurement", "Response Measured", and "HLA". Table 1 lists possible or example values for these five columns. The "Name" column shows the peptide sequence within the binding complex of interest; in the given "Name" column example in the table, GILGFVFTV + OTH(V9), the text between the "+" sign and parentheses indicates the modification method applied to the peptide, and the text within the parentheses lists the amino acids modified by this method. The "HLA" column shows the HLA gene responsible for encoding the MHC binding to the peptide. The "Qualitative Measurement" column has values ranging from strong to weak, representing binding strength. The "Quantitative Measurement" column provides a numerical value obtained from experiments, with the type of measurement explained in the "Response Measured" column.

Fig 1 demonstrates the data preparation process. Starting with the original dataset of 100,141 rows, it was confirmed that each peptide contains at least one NCAA. Since the objective of our research is to predict quantitative binding affinity, each row needed a non-NA value in the "Quantitative Measurement" column. Additionally, to ensure consistency of "HLA" and "Response Measured" across the training and test datasets, the most populated "HLA" and "Response Measured" values, which were HLA-A*02:01 and IC50 with a unit of nanomolar(nM) were selected. Finally, a dataset of 166 rows was prepared for further analysis.

Feature generation

With the sequences of peptides and their quantitative binding values prepared, the next step was to determine how to encode them for machine learning model building. Protein/amino acid encoding involves representing a protein or amino acid with an n-dimensional numerical vector. According to published studies, there are multiple encoding methods, which can be either whole sequence-based or amino acid-based [20]. In the latter approach, each amino acid is first encoded individually, and then the combination of feature vectors from all amino acids in the protein sequence constitutes the encoding of the entire peptide sequence.

Table 1. Table listing the five columns of most interests, example values, and the number of unique values.

Column Name	Example Values	Number of Unique Values
Name	GILGFVFTV + OTH(V9)	61813
Qualitative Measurement	positive, negative	5
Quantitative Measurement	0.1–65,000	223
HLA	HLA-A*02:01	122
Response Measured	half maximal inhibitory concentration (IC50)	8

https://doi.org/10.1371/journal.pone.0314833.t001

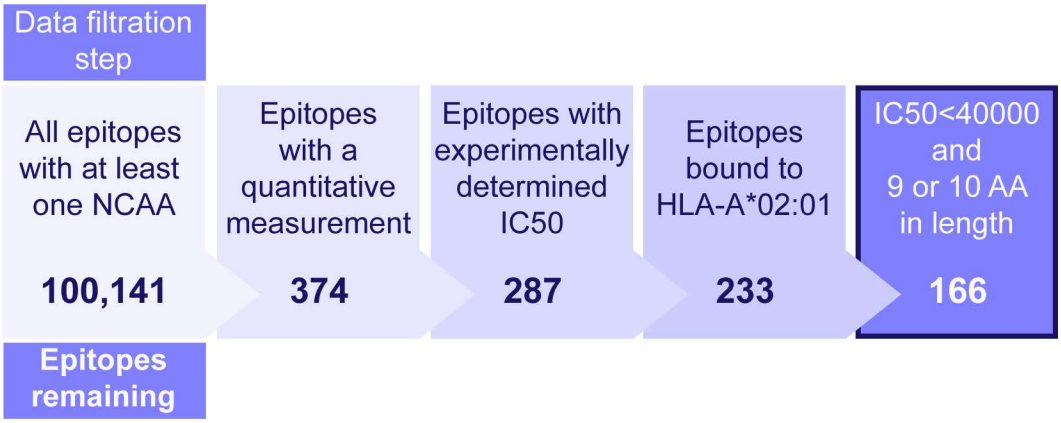


Fig 1. Dataset generation for model training and validation. Epitopes with experimentally determined IC50(nM) values were extracted from the IEDB and filtered as shown to generate the dataset used to generate the model.

https://doi.org/10.1371/journal.pone.0314833.g001

Since all HLA species across the dataset used for this study are the same, only the peptides of the binding complex need to be considered for generating the input vector for the next step's model building. This simplifies the process and makes it more time efficient. Given that the target peptides in this study contain at least one NCAA, which implies potential chemical modifications at the same amino acid position, it is intuitive to use chemistry or structural encoding rather than sequence encoding to retain residue-specific information.

The feature generation process includes four main steps, as illustrated in Fig 2. First, each peptide sequence is tokenized into amino acid tokens. According to summary after step 1, for all 166 rows of data, totally 20 canonical and 28 non-canonical tokens were generated. Fig 3 shows the count of unique tokens across the entire dataset, while Fig 4 illustrates the distribution of tokens at each amino acid position. Tokens with names longer than one character indicate NCAAs. There are two types of NCAA tokens: those containing an underscore ("_") were defined by the authors of this paper during tokenization step, while the others were derived by their resources and named uniquely, for example, "Phg" refers to a racemic mixture of DL-phenylglycine, as described in the referenced literature [21]. The structure of each amino acid token, particularly the NCAAs, was verified using referenced literature searched from IEDB by "Epitope IRI", and chemical structures were converted to SMILES [22] strings. Third, feature vectors for each token were generated using RDKit [23] from the SMILES strings obtained in the previous step. According to RDKit, these vectors describe various physicochemical properties such as molecular weight, partial charge, and the number of specific functional groups, resulting in a total of 208 features. Given the size of the prepared dataset, the feature vector dimension is large, and many features are highly correlated, so principal component analysis (PCA) was applied to reduce the dimensionality from 208 to 10. The choice of 10 components was made because they cover 99.75% of the variance of the original feature vector.

At this point, a map was created with an amino acid token as the key and its corresponding feature vector of size 10 as the value. The final step is to combine the features of all tokens obtained in the first step to generate the feature vector for the entire peptide sequence. With each token's vector size being 10 and the peptide length being nine or ten, the resulting feature vector dimensions for each peptide sequence would be 90 or 100. To ensure consistency of input data for building a machine learning model, an additional 10 zeros were appended to the feature vectors of peptides with a length of nine.

Model testing and validation

To predict binding affinity values of HLA-A0201 with peptide based on both canonical and non-canonical amino acid composition, a machine learning model was established. Fig 5 demonstrates the framework of the model. The model follows

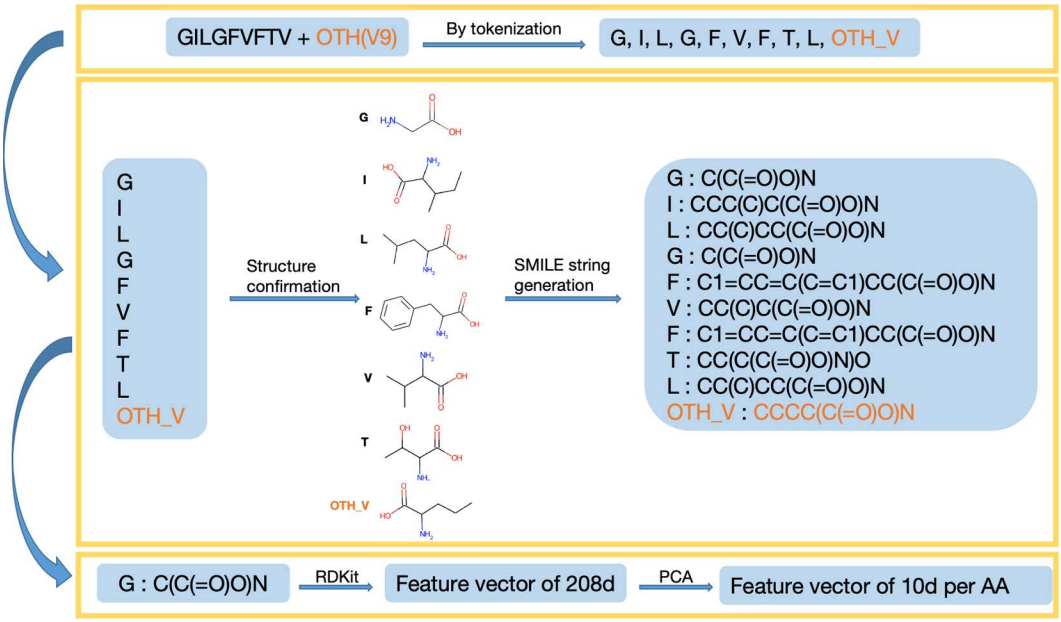


Fig 2. Process for encoding naturally occurring and non-canonical amino acids. Peptides were tokenized by individual amino acid, structures of NCAAs manually confirmed, and SMILE strings for each structural representation generated. These SMILE strings were vectorized using RDKit followed by feature reduction with PCA.

https://doi.org/10.1371/journal.pone.0314833.g002

a structural feature-based supervised learning architecture where each input is a feature vector with 100 dimensions representing the structural and physicochemical properties of the peptide, and the output is the logarithm-transformed IC50 (nM) binding affinity.

During model building, a five-fold cross-validation was applied to the dataset. Root mean square error (RMSE) and R-square (R^2) were used as evaluation metrics. To compare the training results of Partial Least Squares (PLS) with other commonly used algorithms, an open-source tool named Lazy Predict was applied to the same dataset.

Three components were selected for building the Partial Least Squares (PLS) model because, among the range of 2–10 components tested, using 3 components yielded the best performance in terms of cross-validated R-squared (R^2) and root mean square error (RMSE) using five-folds. The detailed results of this comparison are listed in Table 2. Fig 6 illustrates the correlation between the original binding affinity and the predicted binding affinity for both the training set and the test set, using PLS from each individual cycle of five-fold cross-validation. The scatter plots reveal a clear correlation between the actual and predicted values, demonstrating the model's effectiveness despite the relatively small dataset. This strong correlation in both training and test datasets indicates that the model generalizes well and is not overfitted.

To provide a comprehensive comparison, the same 5-fold cross validation was performed using various regressors employed by the Lazy Predict(version 0.2.11) [24]. A total of 36 different regressors were included. Fig 7 displays the test set R-squared from the first validation cycle for all the regressors, with algorithm names labeled for reference. Additional figures displaying RMSE and R^2 across all regressors for each validation cycle are available in the GitHub repository for further reference, the link to which is provided in the code and data section.

The test set R^2 and RMSE results for the top three performing algorithms among these 36, along with those from PLS, are summarized in Table 3. Performance varied by data split; the best-performing algorithm differed between validation cycles. However, certain algorithms appeared more frequently in the top ranks. Among the 15 entries in Table 3, the most frequent high-performing models were ExtraTrees(3 appearance), GradientBoosting (2 appearances), and

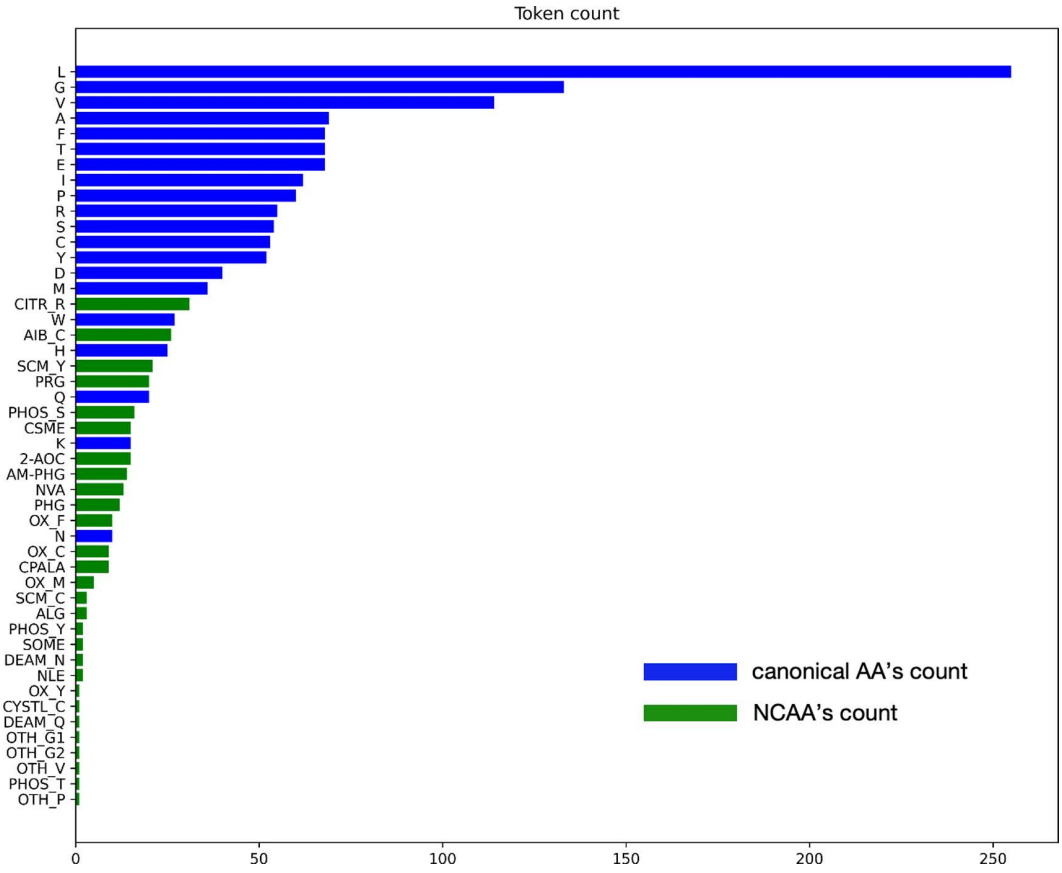


Fig 3. Distribution of canonical and NCAA tokens for every epitope in the dataset.

https://doi.org/10.1371/journal.pone.0314833.g003

Tweedie(2 appearances). Detailed R^2 and RMSE values of test set for these three algorithms along with PLS across all five validations are summarized in Table 4.

For all algorithms, model performance varied by fold – suggesting sensitivity to the specific data splits. This variability may be caused by the relatively small size of the test sets, which can result in low variance in the target variable (y-values). Such low variance can, in turn, lead to cases where certain validation cycles exhibit lower R^2 values despite achieving lower RMSE. Additionally, with smaller datasets, R^2 is more susceptible to the influence of outliers, further complicating the interpretation of model performance.

Discussion

Compared with other sequence-based prediction tools such as NetMHCPan [10,11], the most important improvement our model achieves is its ability to significantly expand the coverage of amino acid species in the involved peptides. Not only does it include the 20 canonical amino acids, but it also takes NCAAs into account without compromising structural accuracy. As long as the structure of an NCAA is known, applying this protocol to predict affinity is straightforward. Additionally, to make the model even more user-friendly, we have eliminated the need for MHC involvement in the model-building process. This means that, when compared with structure- or model-docking based approaches such as Rosetta FlexPepDock [15], our model provides results much faster with minimal human intervention. This is because our method does not require the provision and fine-tuning of large and complex protein structures, thereby accelerating the prediction process and reducing the potential for user error.

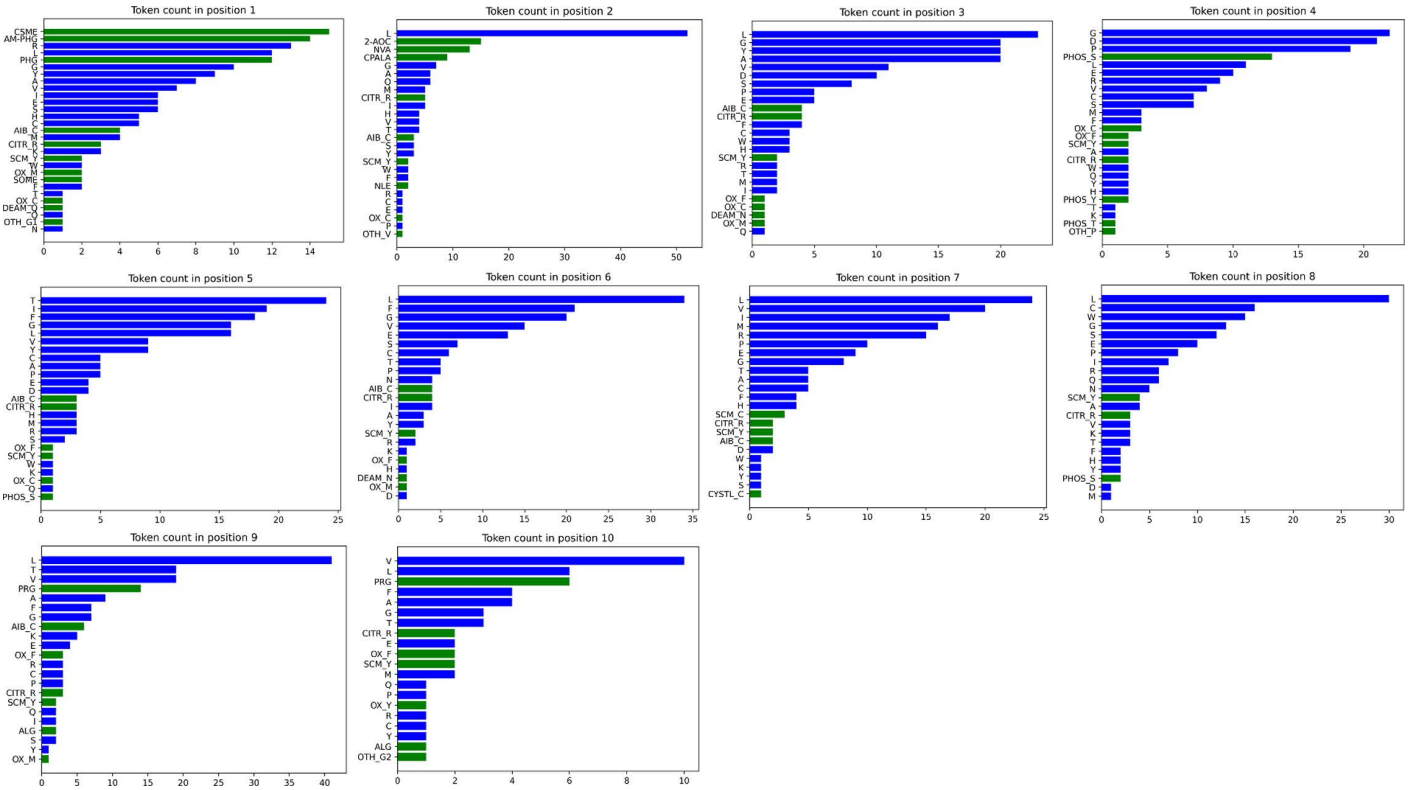


Fig 4. Distribution of canonical and NCAA tokens at each residue position (labeled N to C terminus).

https://doi.org/10.1371/journal.pone.0314833.g004

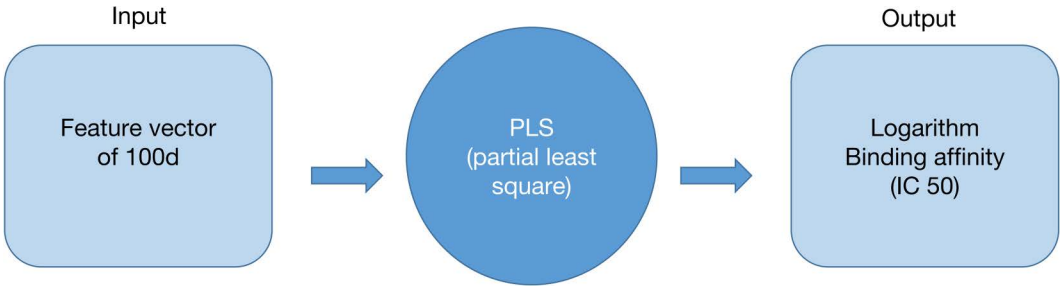


Fig 5. Overview of the predictive model framework.

https://doi.org/10.1371/journal.pone.0314833.g005

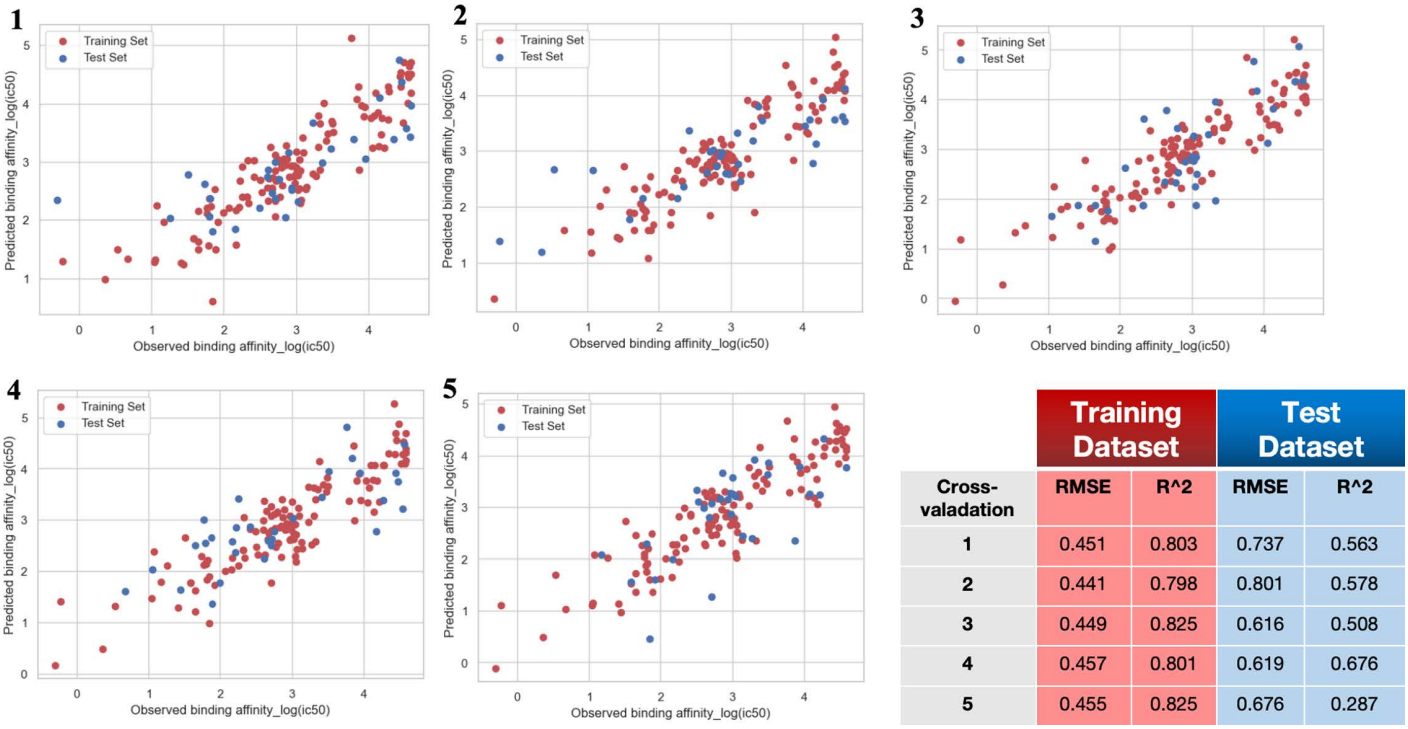
Despite these promising results, it is important to acknowledge that the current size of the training and test datasets is relatively small, which may limit the model's performance. Although the Immune Epitope Database (IEDB) contains a substantial amount of data regarding peptide-MHC binding affinities, only a small percentage of these data includes quantitative binding values, and an even smaller portion pertains to peptides containing NCAAs. Collecting more data would enhance the model's ability to capture a broader range of patterns and interactions, thereby improving its robustness and reliability.

Another future effort involves expanding the scope of the model to include MHCs from other species. For this study, we used data related solely to HLA-A0201 to ensure consistency, but extending the protocol to incorporate other MHC types would significantly widen the prediction coverage and improve the model's reliability. By encompassing a larger variety

Table 2. Performance of PLS with different components – cross-validated R^2 and RMSE.

Components	Cross-validated R^2	Cross-validated RMSE
2	0.444	0.759
3	***0.477***	***0.735***
4	0.463	0.743
5	0.451	0.750
6	0.425	0.766
7	0.395	0.786
8	0.355	0.810
9	0.325	0.827
10	0.283	0.852

https://doi.org/10.1371/journal.pone.0314833.t002



	Training Dataset		Test Dataset	
Cross-valadation	RMSE	R^2	RMSE	R^2
1	0.451	0.803	0.737	0.563
2	0.441	0.798	0.801	0.578
3	0.449	0.825	0.616	0.508
4	0.457	0.801	0.619	0.676
5	0.455	0.825	0.676	0.287

Fig 6. PLS model performance (5-fold cross validation) shown as actual vs. predicted $\log_{10}$(IC50). After splitting the model into 5 equal sized training and testing data sets, the correlation between predicted and experimentally determined IC50 values was calculated. Training dataset shown in red, testing in blue.

https://doi.org/10.1371/journal.pone.0314833.g006

of MHC alleles, we can better understand the nuances of peptide-MHC interactions across different biological contexts, making the model more universally applicable.

An additional consideration is the potential advantage of using ensemble regressors instead of relying on a single algorithm. Given the observed variability in performance across different models and validation folds, an ensemble approach that combines predictions from multiple algorithms may yield more reliable results. By aggregating output – such as through voting or averaging – ensemble models can help mitigate the impact of data-specific fluctuations, potentially improving overall prediction accuracy.

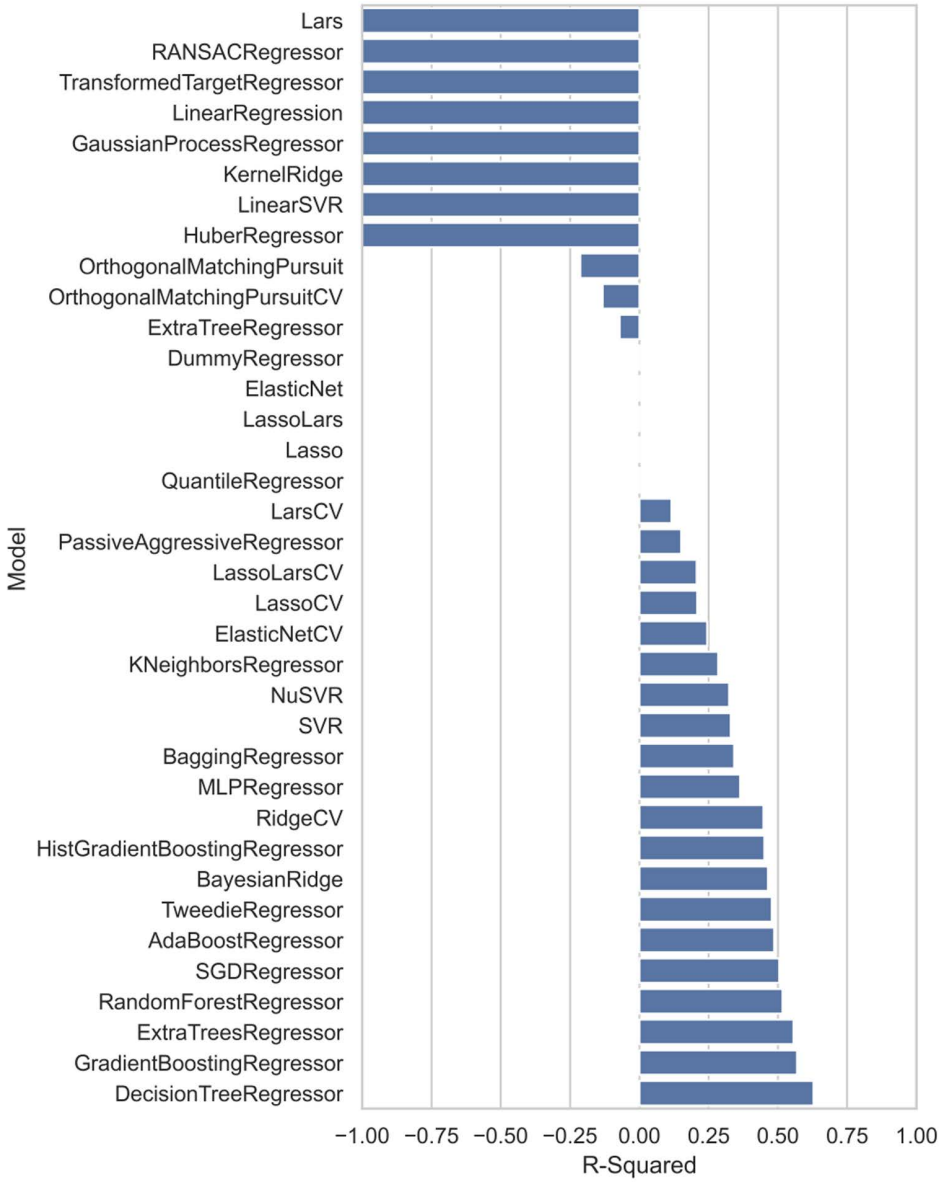


Fig 7. Test set R-squared of different regressors from Lazy Precit in the first cross-validation cycle.

https://doi.org/10.1371/journal.pone.0314833.g007

In conclusion, our model presents a notable advancement in peptide-MHC binding affinity predictions by expanding amino acid coverage and simplifying the prediction process. Future enhancements through increased dataset size and broader MHC coverage will further solidify its utility and accuracy, making it a powerful tool for computational immunology and related fields.

Author contributions

Conceptualization: Shan Jiang, Zhaoqian Su, Nathaniel Bloodworth, Yunchao Liu, Cristina E. Martina, David G. Harrison, Jens Meiler.

Table 3. Comparison of test set R^2 and RMSE for top three performing models and PLS regression for each validation.

Cross-validation 1			Cross-validation 2			Cross-validation 3		
Regressor	**R^2**	**RMSE**	**Regressor**	**R^2**	**RMSE**	**Regressor**	**R^2**	**RMSE**
DecisonTree	0.6280	0.6819	GradientBoosting	0.6947	0.6812	AdaBoost	0.6658	0.5084
GradientBoosting	0.5691	0.7339	Bagging	0.6292	0.7507	ExtraTrees	0.6063	0.5518
ExtraTrees	0.5559	0.7451	BayesianRidge	0.6248	0.7552	RandomForest	0.6028	0.5542
PLS	0.5639	0.7385	PLS	0.5780	0.8010	PLS	0.5089	0.6163
Cross-validation 4			**Cross-validation 5**					
Regressor	**R^2**	**RMSE**	**Regressor**	**R^2**	**RMSE**			
ExtraTrees	0.6304	0.6615	NuSVR	0.5528	0.5358			
AdaBoost	0.5931	0.8666	SVR	0.5418	0.5424			
Tweedie	0.5840	0.7018	Tweedie	0.4836	0.5758			
PLS	0.6761	0.6093	PLS	0.2871	0.6766			

https://doi.org/10.1371/journal.pone.0314833.t003

Table 4. Comparison of test set R^2 and RMSE for PLS and the top three frequently high-performing regressors across all validation cycles.

	ExtraTrees		GradientBoosting		Tweedie		PLS	
	R^2	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**	**R^2**	**RMSE**
Cross-validation 1	0.5559	0.7451	0.5691	0.7340	0.4780	0.8079	0.5639	0.7385
Cross-validation 2	0.5920	0.7875	0.6915	0.6847	0.4891	0.8812	0.5780	0.8010
Cross-validation 3	0.6063	0.5518	0.5092	0.6160	0.5890	0.5638	0.5089	0.6163
Cross-validation 4	0.6304	0.6615	0.4925	0.7752	0.5841	0.7018	0.6761	0.6093
Cross-validation 5	0.3784	0.6317	0.4106	0.6152	0.4836	0.5759	0.2871	0.6766

https://doi.org/10.1371/journal.pone.0314833.t004

Data curation: Shan Jiang, Nathaniel Bloodworth.

Formal analysis: Shan Jiang.

Funding acquisition: David G. Harrison, Jens Meiler.

Methodology: Shan Jiang, Zhaoqian Su, Yunchao Liu.

Software: Shan Jiang, Zhaoqian Su.

Supervision: David G. Harrison, Jens Meiler.

Validation: Shan Jiang.

Visualization: Shan Jiang.

Writing – original draft: Shan Jiang.

Writing – review & editing: Shan Jiang, Zhaoqian Su, Nathaniel Bloodworth, Cristina E. Martina, David G. Harrison, Jens Meiler.

References

1. Janeway CA, Travers P, Walport M, Shlomchik MJ. The major histocompatibility complex and its functions. Immunobiology: The Immune System in Health and Disease. 5th edition. Garland Science; 2001. Available from: https://www.ncbi.nlm.nih.gov/books/NBK27156/
2. Matsumura M, Fremont DH, Peterson PA, Wilson IA. Emerging principles for the recognition of peptide antigens by MHC class I molecules. Science. 1992;257(5072):927–34. https://doi.org/10.1126/science.1323878 PMID: 1323878
3. Archbold JK, Macdonald WA, Gras S, Ely LK, Miles JJ, Bell MJ, et al. Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition. J Exp Med. 2009;206(1):209–19. https://doi.org/10.1084/jem.20082136 PMID: 19139173

4. Dendrou CA, Petersen J, Rossjohn J, Fugger L. HLA variation and disease. Nat Rev Immunol. 2018;18(5):325–39. https://doi.org/10.1038/nri.2017.143 PMID: 29292391
5. James EA, Rieck M, Pieper J, Gebe JA, Yue BB, Tatum M, et al. Citrulline-specific Th1 cells are increased in rheumatoid arthritis and their frequency is influenced by disease duration and therapy. Arthritis Rheumatol. 2014;66(7):1712–22. https://doi.org/10.1002/art.38637 PMID: 24665079
6. Kirabo A, Fontana V, de Faria APC, Loperena R, Galindo CL, Wu J, et al. DC isoketal-modified proteins activate T cells and promote hypertension. J Clin Invest. 2014;124(10):4642–56. https://doi.org/10.1172/JCI74084 PMID: 25244096
7. Kacen A, Javitt A, Kramer MP, Morgenstern D, Tsaban T, Shmueli MD, et al. Post-translational modifications reshape the antigenic landscape of the MHC I immunopeptidome in tumors. Nat Biotechnol. 2023;41(2):239–51. https://doi.org/10.1038/s41587-022-01464-2 PMID: 36203013
8. Zhao W, Sher X. Systematically benchmarking peptide-MHC binding predictors: From synthetic to naturally processed epitopes. PLoS Comput Biol. 2018;14(11):e1006457. https://doi.org/10.1371/journal.pcbi.1006457 PMID: 30408041
9. Mei S, Li F, Leier A, Marquez-Lago TT, Giam K, Croft NP, et al. A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. Brief Bioinform. 2020;21(4):1119–35. https://doi.org/10.1093/bib/bbz051 PMID: 31204427
10. Hoof I, Peters B, Sidney J, Pedersen LE, Sette A, Lund O, et al. NetMHCpan, a method for MHC class I binding prediction beyond humans. Immunogenetics. 2009;61(1):1–13. https://doi.org/10.1007/s00251-008-0341-z PMID: 19002680
11. Nielsen M, Andreatta M. NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. 2016;8(1):33. https://doi.org/10.1186/s13073-016-0288-x PMID: 27029192
12. 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. 2020;48(W1):W449–54. https://doi.org/10.1093/nar/gkaa379 PMID: 32406916
13. O'Donnell TJ, Rubinsteyn A, Bonsack M, Riemer AB, Laserson U, Hammerbacher J. MHCflurry: Open-Source Class I MHC Binding Affinity Prediction. Cell Syst. 2018;7(1):129–132.e4. https://doi.org/10.1016/j.cels.2018.05.014 PMID: 29960884
14. Bloodworth N, Barbaro NR, Moretti R, Harrison DG, Meiler J. Rosetta FlexPepDock to predict peptide-MHC binding: An approach for non-canonical amino acids. PLoS One. 2022;17(12):e0275759. https://doi.org/10.1371/journal.pone.0275759 PMID: 36512534
15. Raveh B, London N, Zimmerman L, Schueler-Furman O. Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS One. 2011;6(4):e18934. https://doi.org/10.1371/journal.pone.0018934 PMID: 21572516
16. Liu T, Pan X, Chao L, Tan W, Qu S, Yang L, et al. Subangstrom accuracy in pHLA-I modeling by Rosetta FlexPepDock refinement protocol. J Chem Inf Model. 2014;54(8):2233–42. https://doi.org/10.1021/ci500393h PMID: 25050981
17. Vita R, Blazeska N, Marrama D, IEDB Curation Team Members, Duesing S, Bennett J, et al. The Immune Epitope Database (IEDB): 2024 update. Nucleic Acids Res. 2025;53(D1):D436–43. https://doi.org/10.1093/nar/gkae1092 PMID: 39558162
18. Ferreira LLG, Andricopulo AD. ADMET modeling approaches in drug discovery. Drug Discov Today. 2019;24(5):1157–65. https://doi.org/10.1016/j.drudis.2019.03.015 PMID: 30890362
19. Jiang L, Yu H, Li J, Tang J, Guo Y, Guo F. Predicting MHC class I binder: existing approaches and a novel recurrent neural network solution. Brief Bioinform. 2021;22(6):bbab216. https://doi.org/10.1093/bib/bbab216 PMID: 34131696
20. Jing X, Dong Q, Hong D, Lu R. Amino Acid Encoding Methods for Protein Sequences: A Comprehensive Review and Assessment. IEEE/ACM Trans Comput Biol Bioinform. 2020;17(6):1918–31. https://doi.org/10.1109/TCBB.2019.2911677 PMID: 30998480
21. Hoppes R, Oostvogels R, Luimstra JJ, Wals K, Toebes M, Bies L, et al. Altered peptide ligands revisited: vaccine design through chemically modified HLA-A2-restricted T cell epitopes. J Immunol. 2014;193(10):4803–13. https://doi.org/10.4049/jimmunol.1400800 PMID: 25311806
22. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6. https://doi.org/10.1021/ci00057a005
23. Landrum G, Tosco P, Kelley B, Rodriguez R, Cosgrove D, Vianello R, et al. rdkit/rdkit: 2025_03_1 (Q1 2025) Release. Zenodo. 2025. Available from: https://doi.org/10.5281/zenodo.15115844
24. lazypredict. Available from: https://pypi.org/project/lazypredict/.