OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–14
https://doi.org/10.1093/bib/bbac192
Problem Solving Protocol

$HLA_{nc}Pred$: a method for predicting promiscuous non-classical HLA binding sites

Anjali Dhall†, Sumeet Patiyal† and Gajendra P. S. Raghava

Corresponding author: Prof. Gajendra P. S. Raghava, Head and Professor, Department of Computational Biology, Indraprastha Institute of Information Technology, Delhi, Okhla Industrial Estate, Phase III, (Near Govind Puri Metro Station), 110020 Office: A-302, New Delhi, India (R&D Block), Tel.: 011-26907444;
E-mail: raghava@iiitd.ac.in
†Anjali Dhall and Sumeet Patiyal contributed equally.

Abstract

Human leukocyte antigens (HLA) regulate various innate and adaptive immune responses and play a crucial immunomodulatory role. Recent studies revealed that non-classical HLA-(HLA-E & HLA-G) based immunotherapies have many advantages over traditional HLA-based immunotherapy, particularly against cancer and COVID-19 infection. In the last two decades, several methods have been developed to predict the binders of classical HLA alleles. In contrast, limited attempts have been made to develop methods for predicting non-classical HLA binding peptides, due to the scarcity of sufficient experimental data. Of note, in order to facilitate the scientific community, we have developed an artificial intelligence–based method for predicting binders of class-Ib HLA alleles. All the models were trained and tested on experimentally validated data obtained from the recent release of IEDB. The machine learning models achieved more than 0.98 AUC for HLA-G alleles on validation dataset. Similarly, our models achieved the highest AUC of 0.96 and 0.94 on the validation dataset for HLA-E*01:01 and HLA-E*01:03, respectively. We have summarized the models developed in the past for non-classical HLA and validated the performance with the models developed in this study. Moreover, to facilitate the community, we have utilized our tool for predicting the potential non-classical HLA binding peptides in the spike protein of different variants of virus causing COVID-19, including Omicron (B.1.1.529). One of the major challenges in the field of immunotherapy is to identify the promiscuous binders or antigenic regions that can bind to a large number of HLA alleles. To predict the promiscuous binders for the non-classical HLA alleles, we developed a web server HLAncPred (https://webs.iiitd.edu.in/raghava/hlancpred) and standalone package.

Keywords: non-classical HLA, binders, COVID-19, prediction, web server, standalone

Introduction

Human leukocyte antigens (HLAs) are an essential part of our immune system and are displayed on the cell surface for antigen presentation to activate immune responses [1, 2]. In humans, the major histocompatibility complex, known as HLA, is the most polymorphic region of the human genome and located at chromosome 6 (6p21.3) [3, 4]. More than 23 000 class-I and 8600 class-II HLA alleles have been already reported across the globe in different ethnic groups according to IMGT/HLA database, 2020 version [5]. HLA class-I genes are further categorized into two major groups, i.e. classical (HLA-A, -B, -C) and non-classical (HLA-G, -E, -F) genes. The classical genes present antigenic peptide ligands on infected cells to CD8+ T cells and activate the immune response. On the other side, non-classical class-I alleles moderate the immune response by inhibiting/activating the natural killer and CD8+ T cells [6]. HLA alleles protect us from several diseases by inducing and regulating immune responses [7–9]. At the same time, adverse effects such as autoimmune disorders, cancer development, metastatic progression and poor prognosis have been shown in various ethnic groups due to the alteration in the expression of HLA alleles [10–13].

Recently, several studies report that the non-classical alleles (HLA-G and HLA-E) play immunomodulatory roles [9, 14–16] in both innate and adaptive immune system (Figure 1). Of note, HLA-G possesses four membrane-bound isoforms and three soluble isoforms; they interact with immunoglobulin-like transcript (ILT2 and ILT4), killer cell immunoglobulin-like receptor (KIR2DL4) and natural killer cell receptors (NKG2A/CD94) [17–19]. Previously, researchers believed that HLA-G alleles are only identified at the maternal–fetal interface. However, recent studies reported that the HLA-G expression is significantly higher during several disease conditions

Anjali Dhall is currently working as a PhD in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Sumeet Patiyal is currently working as a PhD in Computational Biology from Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Gajendra P. S. Raghava is currently working as Professor and Head of Department of Computational Biology, Indraprastha Institute of Information Technology, New Delhi, India.
Received: December 8, 2021. **Revised:** March 23, 2022. **Accepted:** April 27, 2022

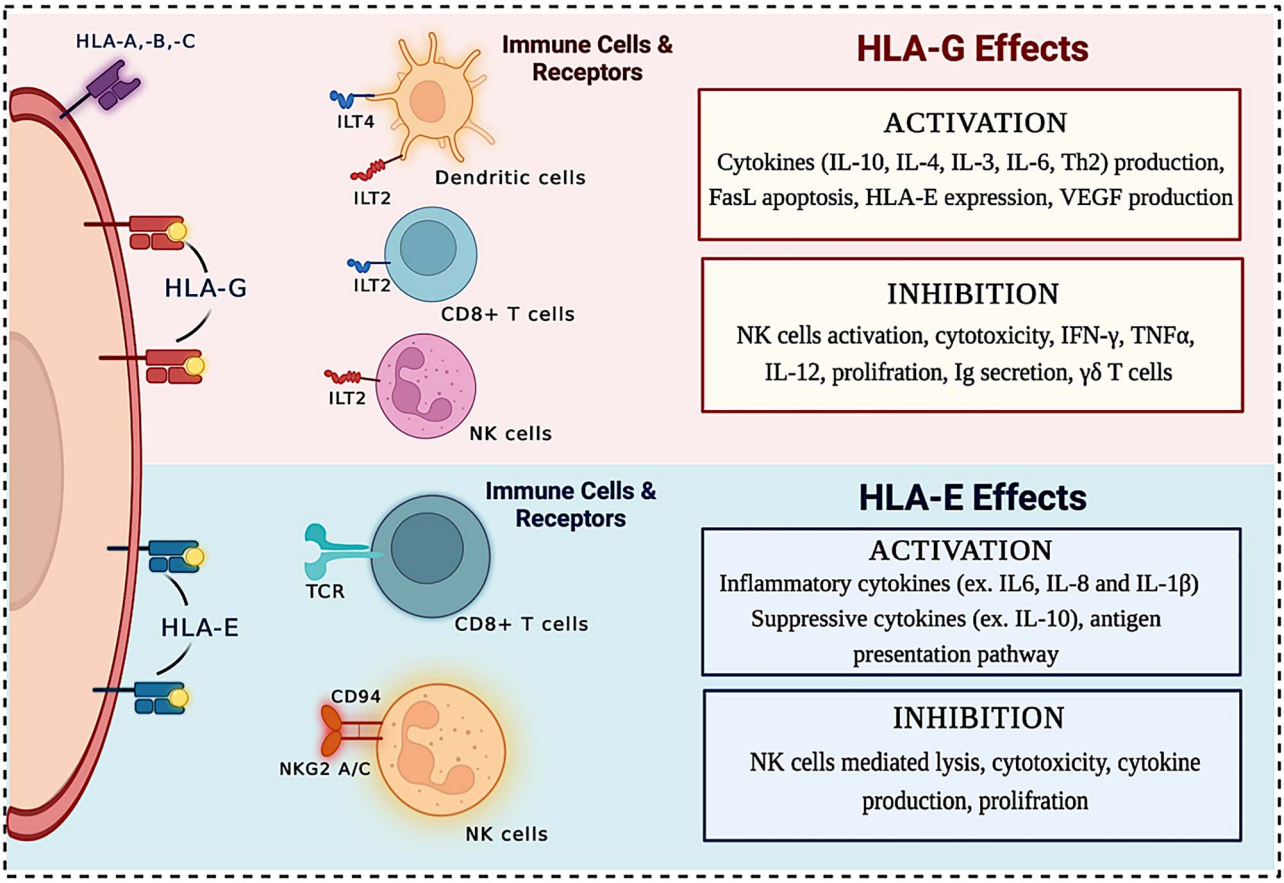

Figure 1. Schematic representation of non-classical HLAs (HLA-G and HLA-E) with immunomodulatory functions (inhibition/activation effects) when interacting with the effective immune cells.

such as cancer, COVID-19 infection, auto-immune and inflammatory diseases [20–28]. In addition, HLA-G inhibits immune cell activation, including CD8+ T, dendritic and natural killer cells during parasite and viral infections (such as influenza A virus, herpes, coronavirus) [29–31]. These viral infections lead to the over-expression of HLA-G and build an immune tolerance microenvironment. Whereas, HLA-E possesses low polymorphism and binds to the highly conserved peptides/epitopes. HLA-E regulates immune cells (natural killer and cytotoxic T cells) by interacting with inhibiting receptors (NKG2A/CD94, NKG2B/CD94) and activating receptor (NKG2C/CD94) [32].

There are two well-known mechanisms for antigen representation by HLA-E alleles, which decide the cells' fate. HLA-E binds to the peptide fragments descended from the signal sequence of other class Ia HLA alleles. This representation leads to the inhibition of NK cell functions by interacting with the NKG2A/CD94 receptors. However, some studies revealed that the peptides from the virus (such as SARS-CoV-2, Epstein-Barr virus, cytomegalovirus and hepatitis C virus) are presented by HLA-E on the cell surface and recognized by virus-specific immune cells, which further activate the immune responses [33–40]. Moreover, HLA-E-restricted CD8+ T-cell leads to the production of anti-inflammatory cytokines like transforming growth factor (TGF-β), interleukin 4 (IL4) and interleukin 10 (IL10), which is responsible for the down-regulation of pro-inflammatory cytokine production and, therefore, inhibiting the cytokine storm, which plays a crucial role in the COVID-19 pathogenesis. The inhibition of cytokine storm also lowers the degree of tissue damage [41]. Several studies reveal that HLA-E inhibits NK-mediated lysis, cytotoxicity, cytokine secretion and tumor proliferation [40, 42–44], as represented in Figure 1. These findings suggest that HLA-G and HLA-E could be essential immune checkpoint molecules for designing novel immunotherapies or subunit vaccines against many diseases.

Thus, there is a need to develop methods for predicting non-classical HLA binders. In the past, numerous computational methods have been developed for predicting HLA binder but majorly focused on classical HLA [45–53]. Only a few tools incorporate models for predicting binders for non-classical HLA alleles. Best of our knowledge, no computational tool has been explicitly developed for predicting non-classical HLA binders. This study is dedicated for non-classical HLA, where a systematic attempt has been made to develop models for predicting non-classical HLA binders. We obtained and examined all experimentally validated non-classical HLA binders from the immune epitope database (IEDB). Based on the availability of sufficient data in IEDB, we

developed models for predicting binders for the following non-classical alleles HLA-G*01:01, HLA-G*01:03, HLA-G*01:04, HLA-E*01:01 and HLA-E*01:03. We have used various machine learning models to better predict non-classical HLA binders.

Material and methods

Dataset collection and pre-processing

One of the significant challenges in bioinformatics is obtaining a sufficient amount of experimentally validated data. In the current study, we have collected the non-classical class-I HLA binding peptides from the IEDB, accessed on 26 October 2021. We obtained a total of 1135 HLA-E and 5151 HLA-G binding peptides. Then, we removed identical peptides from each dataset to prepare non-redundant datasets. Further, we filtered-out those peptides having length greater than 15 or less than 8 residues from each dataset. Finally, we obtained 142 and 723 unique peptides for HLA-E*01:01 and -E*01:03 alleles, respectively. Similarly, we get 2633, 751 and 812 unique binding peptides for HLA-G*01:01, -G*01:03 and -G*01:04 alleles, respectively. The HLA-G alleles-associated binding peptides used in this study were derived from the mass spectrometry data.

On the other side, HLA-E alleles-associated binders were majorly derived from fluorescence-based (biophysical techniques) and mass spectrometry techniques. In case of HLA-E*01:03, most of the data were derived from mass spectrometry (i.e. 632 unique positive binders with 8–15 residues range), and 87 peptides were derived using fluorescence-based techniques and 4 peptides from X-ray crystallography. In order to maintain the uniformity in the datasets, we considered only mass spectrometry–derived peptides for HLA-G*01:01, -G*01:03 and -G*01:04 -E*01:03. However, HLA-E*01:01 has very limited number of experimentally validated binders derived from mass spectrometry; hence, we have considered the complete dataset, i.e. 142 binding peptides (114 derived from fluorescence based and 28 peptides derived from mass spectrometry).

Due to the limited number of negative peptides in IEDB, we randomly generated the HLA-G and HLA-E non-binding peptides having length 8–15 residues from the Swiss-Prot [54] database (March 2021 release). Here, we have created two separate datasets; one is a balanced dataset that incorporates an equal number of negative peptides as positive peptides for each allele. The other is the imbalanced/realistic dataset that comprises ten times negative data in comparison to the positive dataset. The complete distribution of positive and negative datasets is shown in Table 1.

Amino acid composition

Amino acid composition (AAC) of the positive and negative dataset for each allele is computed to understand the compositional similarity in different peptide sequences. The following equation is used to calculate the AAC for HLA-G and HLA-E allele binder/non-binder peptides.

$$AAC_i = \frac{AAR_i}{\text{Total number of residues}} \times 100$$

where AAC_i and AAR_i are the percentage composition and number of residues of type i in a peptide, respectively.

Sequence logo

We have generated sequence logos with the help of WebLogo software (http://weblogo.threeplusone.com) [54] for each HLA allele. WebLogo provides the graphical representation with residue positions on x-axis, and y-axis represents the bit score signifying the conservation of residues at a particular position. Each position exhibits the stack of amino acids that are conserved at that position, where the height of each residue signifies the relative frequency. The binding nonameric peptides restricted to each HLA allele is used to generate the sequence logos.

Feature generation

In this study, we used Pfeature [55] to calculate the binary profile for each peptide belongs to the positive and negative datasets of non-classical HLA-alleles. To calculate the binary profiles, the length of the variable should be fixed, but the length of the peptides was varying from 8 to 15. Thus, to generate the fixed length vector, we extracted the same number of residues from the N- and C-terminal. Since the minimum length of the peptides was 8, we chose to select the eight residues from each end and designed different patterns to calculate features. First, eight residues were selected from N-terminal and designated as N_8 patterns; similarly, eight residues were chosen from C-terminal and referred to as C_8 residues. Other patterns of length 16 were generated by joining the eight residues from N- and C-terminal and referred as N_8C_8. For the aforementioned patterns, each amino acid is represented by the vector of length 20, where each element represents the presence/absence of that residue, where presence was designated as '1' and absence was presented by '0'. For instance, residue 'A' was represented by vector 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0; hence, total length of vector generated for patterns N_8 and C_8 was 160 (20*8), and for N_8C_8, it was 320 (20*16).

Likewise, patterns of peptides with maximum length (i.e. 15 amino acids) were generated and termed as AA_{15} and binary profile was calculated. In this case, residues 'X' were added in the peptides having length less than 15. For example, 7 'X' were added in the peptide having length equal to 8, in order to make its length 15, such as peptide 'VYIKHPVS' became 'VYIKHPVSXXXXXXX'. In this scenario, each amino acid is represented by the vector size of 21, where the last element exhibited presence/absence of 'X'. Residue 'V' is represented by vector 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0 and 'X' is presented as 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1.

Table 1. Distribution of positive and negative peptides for HLA-G and HLA-E alleles obtained from IEDB and Swiss-Prot database

HLA	HLA allele	Binding peptides (positive dataset)	Non-binding peptides (negative dataset)	Total peptides
Balanced dataset				
HLA-G	HLA-G*01:01	2633	2633	5266
	HLA-G*01:03	751	751	1502
	HLA-G*01:03	812	812	1624
HLA-E	HLA-E*01:01	142	142	284
	HLA-E*01:03	632	632	1282
Realistic dataset (imbalanced)				
HLA-G	HLA-G*01:01	2633	26 330	28 963
	HLA-G*01:03	751	7510	8261
	HLA-G*01:03	812	8120	8932
HLA-E	HLA-E*01:01	142	1420	1562
	HLA-E*01:03	632	6320	6952

Machine learning techniques

Several machine learning methods have been employed to classify the positive and negative non-classical HLA binding peptides. Here, we have used numerous machine learning classifiers such as decision tree (DT), support vector classifier (SVC), random forest (RF), XGBoost (XGB), logistic regression (LR), ExtraTree classifier (ET), Gaussian Naive Bayes (GNB) and k-nearest neighbors (KNN) using scikit-learn python-based package [56].

Five-fold cross-validation approach

We have applied a 5-fold cross-validation technique to evade the curse of biasness and overfitting in the generated models. It is one of the most crucial steps to evaluate the prediction model. In this approach, the entire dataset is partitioned into five parts, out of which four are used in training, and the resulted model is tested on the left one. The exact process is reproduced five times so that each part gets the opportunity to act as the testing dataset. Ultimately, the final performance is represented as the average of the performance of the five models that resulted from the five iterations.

Evaluation parameters

Several parameters can be employed to assess the prediction models, which can be broadly classified into two categories termed threshold-dependent and -independent parameters. In this study, we have determined sensitivity, specificity, accuracy, F1-score and Matthews correlation coefficient (MCC) as the threshold-dependent parameters. In contrast, area under receiver operating characteristics (AUC) curve is calculated as the threshold-independent parameter. Sensitivity (equation 1) measures ability of the model to correctly predict the binders, whereas specificity (equation 2) accounts for the percentage of correctly predicted non-binders. Accuracy (equation 3) exhibits the percentage of the correctly predicted binders and non-binders, F1-score (equation 4) captures the balance between precision and recall, and MCC (equation 5) explains the relation between the predicted and observed values. AUC is a plot between the sensitivity and 1-specificity, which captures the ability of the model to distinguish between the classes.

$$\text{Sensitivity} = \frac{T_P}{T_P + F_N} \tag{1}$$

$$\text{Specificity} = \frac{T_N}{T_N + F_P} \tag{2}$$

$$\text{Accuracy} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \tag{3}$$

$$\text{F1} - \text{Score} = \frac{2T_P}{2T_P + F_P + F_N} \tag{4}$$

$$\text{MCC} = \frac{(T_P * T_N) - (F_P * F_N)}{\sqrt{(T_P + F_P)\,(T_P + F_N)\,(T_N + F_P)\,(T_N + F_N)}} \tag{5}$$

where T_P, T_N, F_P and F_N stand for true positive, true negative, false positive and false negative, respectively.

Results

Composition analysis

The average AAC of HLA-G and HLA-E binding and non-binding peptides is represented in Figure 2. As shown in the graphs, the compositional difference in the positive and negative datasets is clearly visible. HLA-G*01:01, -G*01:03, -G*01:04 binders (i.e. positive peptides) have a higher composition of residues such as isoleucine (I), lysine (K), leucine (L) and proline (P) as compared to non-binding peptides as depicted in Figure 2 A–C). However, the average composition of alanine (A), leucine (L), methionine (M), proline (P) and valine (V) residues is higher in HLA-E*01:01, -E*01:03 binding peptides in contrast to the negative dataset as shown in Figure 2 D and E).

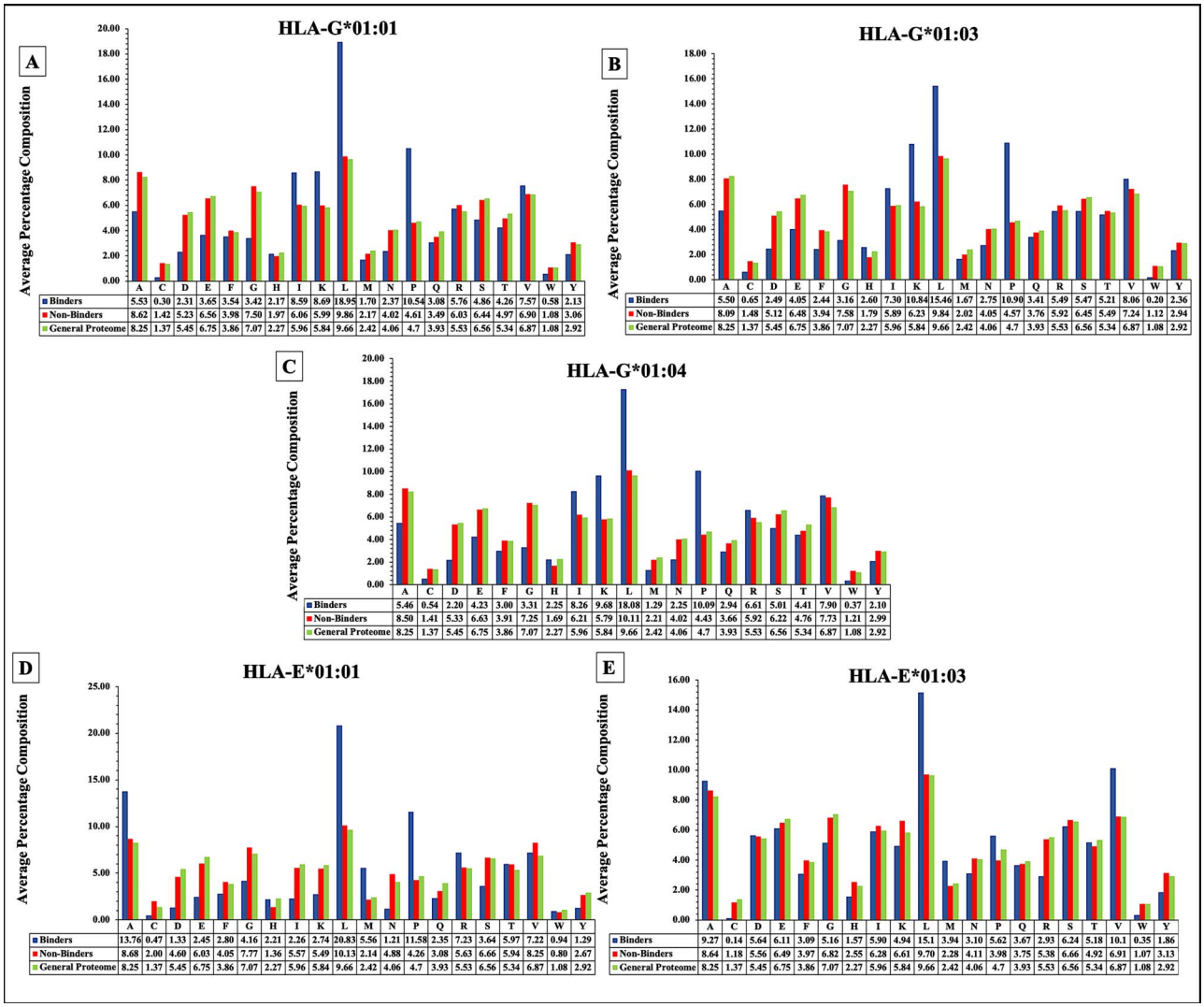

Figure 2. The average amino acid composition of HLA-G*01:01, -G*01:03, -G*01:04, -E*01:01 and -E*01:03 binding/non-binding peptides and general proteome.

Position conservation analysis

Here, we represent the sequence logo for each non-classical HLA-allele using WebLogo for nonameric binders only. The conserved residue and its specific position in the nonameric sequences are determined using each logo. In the case of HLA-G, K/R anchor residues located at first anchor position (P1), while P at third position (P3), and L at ninth position (P9) exhibits very strong abundance (see Figure 3 A–C). In case of HLA-E, conserved binding residues are mostly located at P2 and P9 with predominant hydrophobic residues. In case of HLA-E, M/L is the anchor residues at second anchor position (P2); at P9, L is most dominating residue in case of HLA-E01:01, and V/L for HLA-E01:03 (see Figure 3D and E).

Moreover, the conserved motif residues and frequencies of the specific amino acids from position 1 to 9 in the 9mer binding peptides corresponding to HLA-G and HLA-E alleles are provided in Table 2. Here, we have utilized Jalview software [57] to identify the conservation and percentage of occurrence frequency. We first extract the 9mer binders restricted to each allele, and then we used the Jalview software to understand the conservation at particular position. From the position conservation analysis, we observed that most of the residue motifs are positively charged and hydrophobic in nature. HLA-G*01:01/G*01:03/G*01:04 restricted peptides anchored residues predominantly located at P1, P3 and P9, whereas the % frequency of K/R at P1 is greater than 57%, P3 is rich with P (>54%) and P9 is highly conserved with amino acid L (>89%).

As illustrated in Table 2, in case of HLA-E*01:01/E*01:03, the anchored motif residues for P2 are M/L (>62%). Moreover, L (92.20%) is strongly preferred at P9 for HLA-E*01:01 restricted 9mers, whereas in case of HLA-E*01:03, V/L (72.20%) is rich at P9. The comprehensive information corresponding to each allele is provided in Supplementary Table S1.

Machine learning–based prediction

In this study, we have implemented several classifiers such as GNB, XGB, RF, DT, SVC, ET, KNN and LR to develop

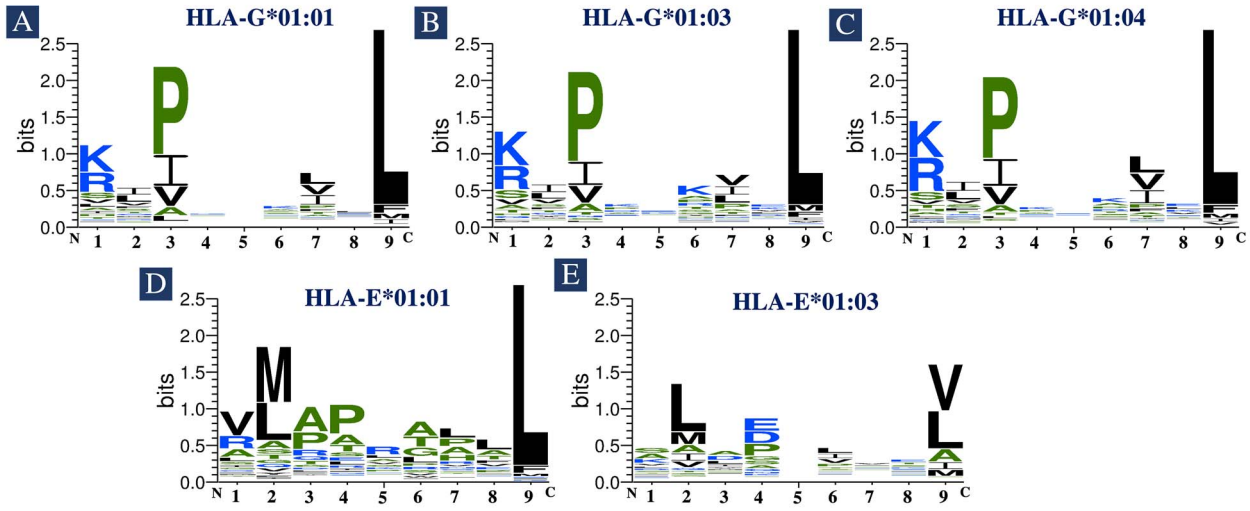

Figure 3. WebLogo representation of nonameric sequence binders for non-classical HLA alleles.

Table 2. Frequency of specific amino acid residue at positions 1 to 9 in the nonameric peptides for non-classical HLA alleles

HLA-allele / Position		P1	P2	P3	P4	P5	P6	P7	P8	P9
HLA-G*01:01	Residue	K/R	I/L/V	P	-	-	-	L/V/I	-	L
	% Frequency	57.72%	47.31%	54.96%	-	-	-	60.18%	-	89.46%
	Chemical Property	Positive	Hydrophobic	Proline	-	-	-	Hydrophobic	-	Hydrophobic
HLA-G*01:03	Residue	K/R	I/L/V	P	-	-	-	V/I/L	-	L
	% Frequency	61.35%	46.5%	57.80%	-	-	-	55.30%	-	89.20%
	Chemical Property	Positive	Hydrophobic	Proline	-	-	-	Hydrophobic	-	Hydrophobic
HLA-G*01:04	Residue	K/R	I/L/V	P	-	-	-	L/V/I	-	L
	% Frequency	67.10%	49.7%	54.40%	-	-	-	67.60%	-	89.30%
	Chemical Property	Positive	Hydrophobic	Proline	-	-	-	Hydrophobic	-	Hydrophobic
HLA-E*01:01	Residue	V	M/L	-	-	-	-	-	-	L
	% Frequency	35.20%	69.50%	-	-	-	-	-	-	92.20%
	Chemical Property	Hydrophobic	Hydrophobic	-	-	-	-	-	-	Hydrophobic
HLA-E*01:03	Residue	-	L/M	-	-	-	L/I/V	-	-	V/L
	% Frequency	-	62.40%	-	-	-	47.10%	-	-	72.20%
	Chemical Property	-	Hydrophobic	-	-	-	Hydrophobic	-	-	Hydrophobic

prediction models. We calculate binary profile–based features of positive and negative datasets (i.e. HLA-G*01:01, -G*01:03, -G*01:04, -E*01:01 and -E*01:03 binding and non-binding peptides). Initially, we generate four feature sets (i.e. N_8, C_8, N_8C_8 and AA_{15} binary profiles) using Pfeature standalone package. Then, we developed several machine learning models on each feature set for HLA-G and HLA-E alleles.

Performance of HLA-G allele models

We compute the performance of each allele dataset using various machine learning classifiers with four feature sets as depicted in Supplementary Table S2. Further, we observed that AA_{15} binary profile–based models outperform other feature sets with balanced sensitivity and specificity. As shown in Table 3, HLA-G*01:01 dataset achieved maximum AUC of 0.99 (Figure 4) and accuracy

Table 3. The performance of machine leaning models developed using AA_{15} binary profile-based features of HLA-G alleles on training and validation datasets

Classifier	Training dataset					Validation dataset				
	Sensitivity	Specificity	Accuracy	AUC	MCC	Sensitivity	Specificity	Accuracy	AUC	MCC
HLA-G*01:01										
DT	88.65	88.03	88.34	0.93	0.77	89.37	89.75	89.56	0.94	0.79
RF	95.25	95.11	95.18	0.99	0.90	93.93	95.83	94.88	0.98	0.90
LR	94.26	94.40	94.33	0.98	0.89	92.79	95.07	93.93	0.98	0.88
XGB	95.44	95.39	95.42	0.99	0.91	94.12	92.98	93.55	0.98	0.87
KNN	92.93	92.83	92.88	0.97	0.86	91.27	94.12	92.69	0.97	0.85
GNB	93.31	79.77	86.54	0.87	0.74	91.08	86.34	88.71	0.90	0.78
ET	95.39	95.39	95.39	0.99	0.91	93.93	96.02	94.97	0.99	0.90
SVC	95.30	95.58	95.44	0.99	0.91	94.50	95.83	95.16	0.99	0.90
HLA-G*01:03										
DT	82.36	84.67	83.51	0.89	0.67	80.67	82.78	81.73	0.88	0.64
RF	92.18	92.33	92.26	0.98	0.85	87.33	94.04	90.70	0.97	0.82
LR	91.51	91.67	91.59	0.97	0.83	88.00	94.70	91.36	0.97	0.83
XGB	92.18	92.33	92.26	0.98	0.85	90.00	93.38	91.69	0.98	0.83
KNN	89.35	88.83	89.09	0.95	0.78	85.33	95.36	90.37	0.94	0.81
GNB	93.84	57.83	75.85	0.76	0.55	91.33	65.56	78.41	0.78	0.59
ET	92.85	92.67	92.76	0.98	0.86	89.33	94.04	91.69	0.97	0.84
SVC	92.35	92.50	92.42	0.98	0.85	88.00	95.36	91.69	0.97	0.84
HLA-G*01:04										
DT	79.54	80.74	80.14	0.86	0.60	86.42	76.69	81.54	0.87	0.63
RF	93.08	93.07	93.07	0.98	0.86	96.30	93.87	95.08	0.98	0.90
LR	92.31	92.30	92.30	0.98	0.85	96.30	92.64	94.46	0.98	0.89
XGB	92.62	92.76	92.69	0.98	0.85	95.06	94.48	94.77	0.98	0.90
KNN	89.85	89.99	89.92	0.96	0.80	95.06	90.80	92.92	0.97	0.86
GNB	90.77	67.49	79.14	0.79	0.60	93.21	67.49	80.31	0.80	0.63
ET	93.39	93.22	93.30	0.98	0.87	96.30	93.87	95.08	0.98	0.90
SVC	93.08	93.07	93.07	0.98	0.86	96.91	93.87	95.39	0.98	0.91

DT; decision tree, RF; random forest, LR; logistic regression, XGB; XGBoost, KNN; k-nearest neighbors, GNB; Gaussian Naive Bayes, ET; ExtraTree, SVC; support vector classifier; AUC; area under receiver operating curve, MCC; Matthews correlation coefficient.

>95% on both training and validation dataset using SVC classifier. ET-based models also show similar results on training and validation datasets with an AUC of 0.99 and accuracy >95% (Table 3). XGB classifier shows comparable performance on HLA-G*01:03 dataset, with maximum AUC 0.98; accuracy 91.69% on validation dataset, whereas RF, ET and SVC classifiers outperform the other models and perform equivalent on HLA-G*01:04 dataset. The maximum AUC of 0.98 (Figure 4) and accuracy 93.07 and 95.39% was achieved on training and validation dataset as shown in Table 3. On the other side, DT-, LR-, KNN- and GNB-based models performed poorly on each dataset. The complete results for each feature set are provided in Supplementary Table S2.

Performance of HLA-E allele models

In this, we have used positive and negative datasets of HLA-E*01:01 and -E*01:03 alleles and developed various prediction models. For this dataset, AA_{15} binary profile–based features prevail over other classifiers as indicated in the previous section results. As demonstrated in Table 4, the performance of ET-based models for HLA-E*01:01 allele outperforms the other classifiers with an accuracy of 87.67 and 89.47%; AUC of 0.96 (Figure 4) on training and validation dataset, respectively. Besides, RF- and XGB-based models perform pretty well with balanced sensitivity and specificity (Table 4). However, on HLA-E*01:03 dataset, SVC performed quite well, with AUC of 0.93 and 0.94 (Figure 4); accuracy of 84.08 and 84.98% on training and validation dataset, respectively, as shown in Table 4. Similarly, models based on ET also perform equivalent with very less difference in sensitivity and specificity (Table 4). The complete results for all the other feature sets are provided in Supplementary Table S3.

In addition, we also checked the performance of models by separating the datasets based on different sources from which the HLA-E binders were obtained. As provided in Supplementary Table S4, we developed models for HLA-E*01:01 and HLA-E*01:03 using mass spectrometry- and florescence-derived datasets. We observed that both techniques performed quite well in case of both the alleles. However, due to the limitation of experimental data for HLA-E*01:01 binding peptides, we reported the results on the combined dataset in this study.

Moreover, the robustness of the developed models was checked by shuffling the data 10 times, and the performance was calculated. The mean ±standard deviation of the performance of best-performing classifiers for HLA-G and HLA-E alleles is reported in Table 5. The comprehensive results of 10 times reshuffling data for each allele were integrated into Table 5. We observe that the

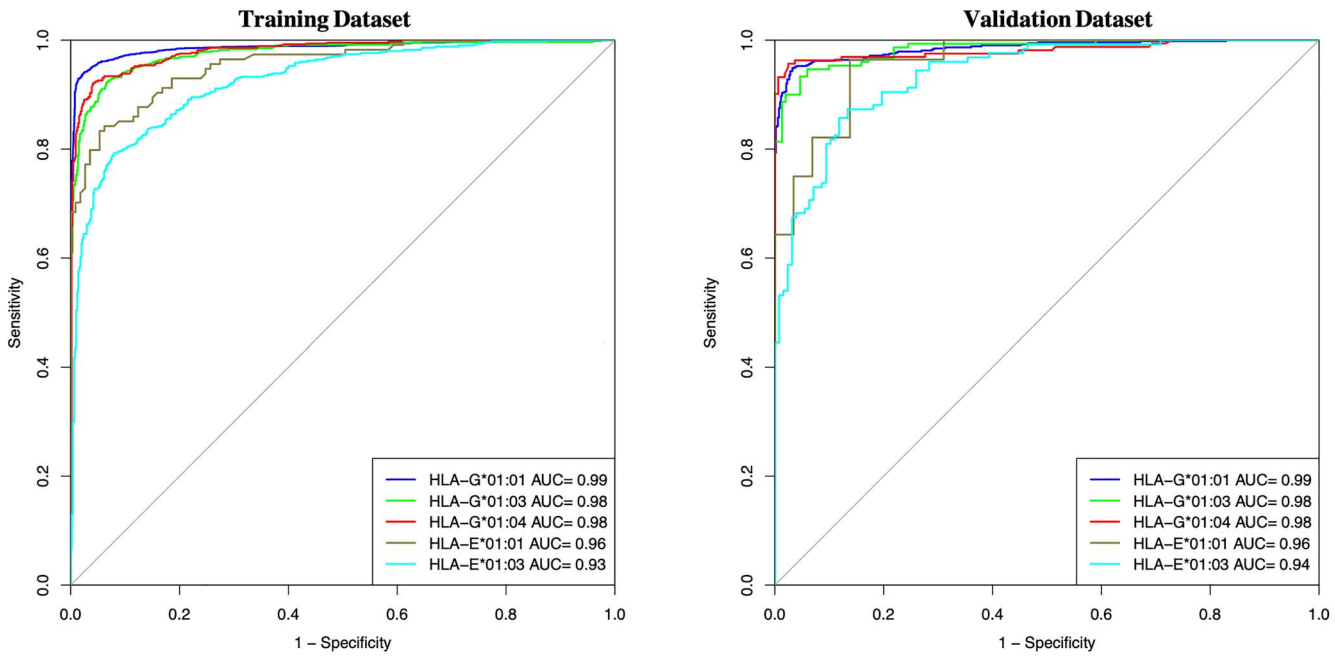

Figure 4. AUC curves represent the performance of best models on training and validation datasets for each allele.

Table 4. The performance of machine leaning models developed using AA_{15} binary profile–based features of HLA-E alleles on training and validation datasets

Classifier	Training dataset					Validation dataset				
	Sensitivity	Specificity	Accuracy	AUC	MCC	Sensitivity	Specificity	Accuracy	AUC	MCC
HLA-E*01:01										
DT	78.07	76.99	77.53	0.82	0.55	75.00	75.86	75.44	0.81	0.51
RF	86.84	86.73	86.78	0.95	0.74	89.29	86.21	87.72	0.96	0.76
LR	88.60	89.38	88.99	0.94	0.78	85.71	89.66	87.72	0.97	0.76
XGB	86.84	86.73	86.78	0.95	0.74	89.29	86.21	87.72	0.96	0.76
KNN	83.33	83.19	83.26	0.91	0.67	82.14	82.76	82.46	0.93	0.65
GNB	74.56	76.99	75.77	0.76	0.52	89.29	79.31	84.21	0.84	0.69
ET	87.72	87.61	87.67	0.96	0.75	92.86	86.21	89.47	0.96	0.79
SVC	86.84	86.73	86.78	0.94	0.74	85.71	86.21	85.97	0.96	0.72
HLA-E*01:03										
DT	71.74	70.50	71.12	0.78	0.42	71.43	66.93	69.17	0.76	0.38
RF	84.19	83.37	83.78	0.92	0.68	92.06	77.95	84.98	0.93	0.71
LR	83.60	83.76	83.68	0.91	0.67	88.89	77.17	83.00	0.90	0.67
XGB	82.81	82.77	82.79	0.91	0.66	82.54	77.95	80.24	0.90	0.61
KNN	80.44	79.80	80.12	0.88	0.60	88.10	72.44	80.24	0.90	0.61
GNB	93.08	44.95	69.04	0.69	0.43	90.48	46.46	68.38	0.69	0.41
ET	84.19	83.96	84.08	0.93	0.68	93.65	77.95	85.77	0.93	0.73
SVC	83.99	84.16	84.08	0.93	0.68	90.48	79.53	84.98	0.94	0.70

DT; decision tree, RF; random forest, LR; logistic regression, XGB; XGBoost, KNN; k-nearest neighbors, GNB; Gaussian Naive Bayes, ET; ExtraTree, SVC; support vector classifier; AUC; area under receiver operating curve, MCC; Matthews correlation coefficient.

performance of models on each dataset is maintained even after the reshuffling, which supports the robustness of the developed models.

Performance on realistic dataset

In the realistic scenario, the negative dataset is much bigger than the positive dataset. But, due to the low availability of experimentally validated non-binders, we have generated the random negative peptide datasets from the Swiss-Prot database corresponding to each allele, which is ten times bigger than the positive dataset. Further, we developed prediction models using AA_{15} binary features on the imbalanced datasets, and the complete results on each dataset are provided in Supplementary Table S5. It was observed that even after increasing the datasets ten times, the performance of the models is comparable for HLA-G and HLA-E alleles.

Table 5. The performance of best machine learning models developed after 10 times reshuffling the datasets

Allele	Classifier	Training dataset					Testing dataset				
		Sensitivity	Specificity	Accuracy	AUROC	MCC	Sensitivity	Specificity	Accuracy	AUROC	MCC
HLA-G*01:01	SVC	95.22 ± 0.22	95.25 ± 0.28	95.24 ± 0.15	0.99 ± 0.00	0.90 ± 0.00	95.18 ± 0.75	95.56 ± 0.79	95.37 ± 0.58	0.99 ± 0.00	0.91 ± 0.01
HLA-G*01:03	XGB	90.47 ± 0.34	92.71 ± 0.76	91.59 ± 0.40	0.97 ± 0.00	0.83 ± 0.01	91.31 ± 2.59	92.02 ± 2.02	91.66 ± 1.34	0.97 ± 0.01	0.83 ± 0.03
HLA-G*01:04	ET	93.61 ± 0.61	92.71 ± 0.50	93.16 ± 0.45	0.98 ± 0.00	0.86 ± 0.01	93.98 ± 2.18	93.10 ± 2.46	93.54 ± 1.37	0.98 ± 0.00	0.87 ± 0.03
HLA-E*01:01	ET	87.20 ± 2.92	87.07 ± 1.81	87.14 ± 1.51	0.95 ± 0.01	0.74 ± 0.03	86.75 ± 7.96	87.99 ± 6.26	87.37 ± 3.77	0.95 ± 0.02	0.75 ± 0.08
HLA-E*01:03	**SVC**	83.12 ± 0.97	84.08 ± 0.87	83.60 ± 0.75	0.92 ± 0.00	0.67 ± 0.02	83.89 ± 3.23	84.18 ± 3.3	84.03 ± 2.12	0.93 ± 0.01	0.68 ± 0.04

Comparison with existing methods

It is very important to compare this new method with the existing methods to understand the advantages/disadvantages. To validate our method, we compare the performance of our models with the existing methods (MHCflurry 2.0 and NetMHCpan 4.1). We have trained our models on the previous used dataset, provided by MHCflurry 2.0 and NetMHCpan 4.1 for HLA-G and HLA-E alleles. Further, we validate the performance of all the methods on the updated IEDB data i.e. binders of HLA-G*01:01, -G*01:03, -G*01:04, -E*01:01 and -E*01:03 alleles. As shown in the results (Table 6), HLA_{nc}Pred outperforms other methods for predicting the HLA-G binding peptides with balanced sensitivity, specificity and maximum accuracy (Table 6). Although MHCflurry 2.0 performs quite well on the prediction of binder/non-binding peptides of HLA-G*01:01 dataset, it performs poorly on other datasets. Similarly, we observed NetMHCpan 4.1 performs quite well on HLA-E*01:01 validation dataset; however, it performs inadequately on the validation dataset of HLA-G alleles as shown in Table 6.

Web server and standalone implementation

In the current study, to assist the scientific community in predicting and scanning of non-classical HLA binder and non-binder peptides, we have developed a web-based service 'HLA_{nc}Pred' (https://webs.iiitd.edu.in/raghava/hlancpred/). In this web server, we have used our best models to better predict non-classical HLA binders. A detailed description of HLA_{nc}Pred modules is given below.

(i) PREDICT: The prediction module allows the users to identify the most promiscuous HLA-G (-G*01:01, -G*01:02, -G*01:03) and HLA-E (-E*01:01, -E*01:03) binders and non-binder peptides. Users can submit multiple peptides in the standard FASTA format in the provided box or upload the input files and can select either single or more than one allele for binding prediction. Server provides tabular format results, with the input sequence, score and prediction (binder/non-binder).

(ii) SCAN: This module allows the facility to recognize the protein regions that may bind to the non-classical alleles, such as HLA-G*01:01, -G*01:02, -G*01:03, -E*01:01, -E*01:03, using their binary profiles. It also facilitates the users to choose any length of sequence to predict binders. The other way round, users can also scan a protein sequence to find out the novel peptides with the binding ability to HLA-G and HLA-E alleles. It will generate the fragments of a length selected by the users and predict their activity. Users can submit one or multiple protein sequences in FASTA format and choose the allele(s) for the prediction. In addition, the user can also choose the result mode, i.e. graphical or tabular, as represented in Figure 5.

Table 6. Comparison of performance of $HLA_{nc}Pred$ and other methods on the updated IEDB dataset

HLA allele	$HLA_{nc}Pred$				MHCflurry 2.0				NetMHCpan 4.1			
	Sensitivity	Specificity	Accuracy	MCC	Sensitivity	Specificity	Accuracy	MCC	Sensitivity	Specificity	Accuracy	MCC
HLA-G*01:01	92.6	94.3	93.4	0.87	88.7	93.3	91.0	0.82	47.2	98.7	72.9	0.53
HLA-G*01:03	72.2	61.1	66.7	0.33	27.8	94.4	61.1	0.29	8.30	97.2	52.8	0.12
HLA-G*01:04	73.5	70.6	72.1	0.44	32.4	97.1	64.7	0.39	11.8	100	55.9	0.25
HLA-E*01:01	92.1	88.1	90.1	0.80	85.7	84.9	85.4	0.71	82.5	92.1	87.3	0.75
HLA-E*01:03	71.3	83.8	77.6	0.56	61.2	91.0	76.5	0.55	50.5	95.3	72.9	0.51

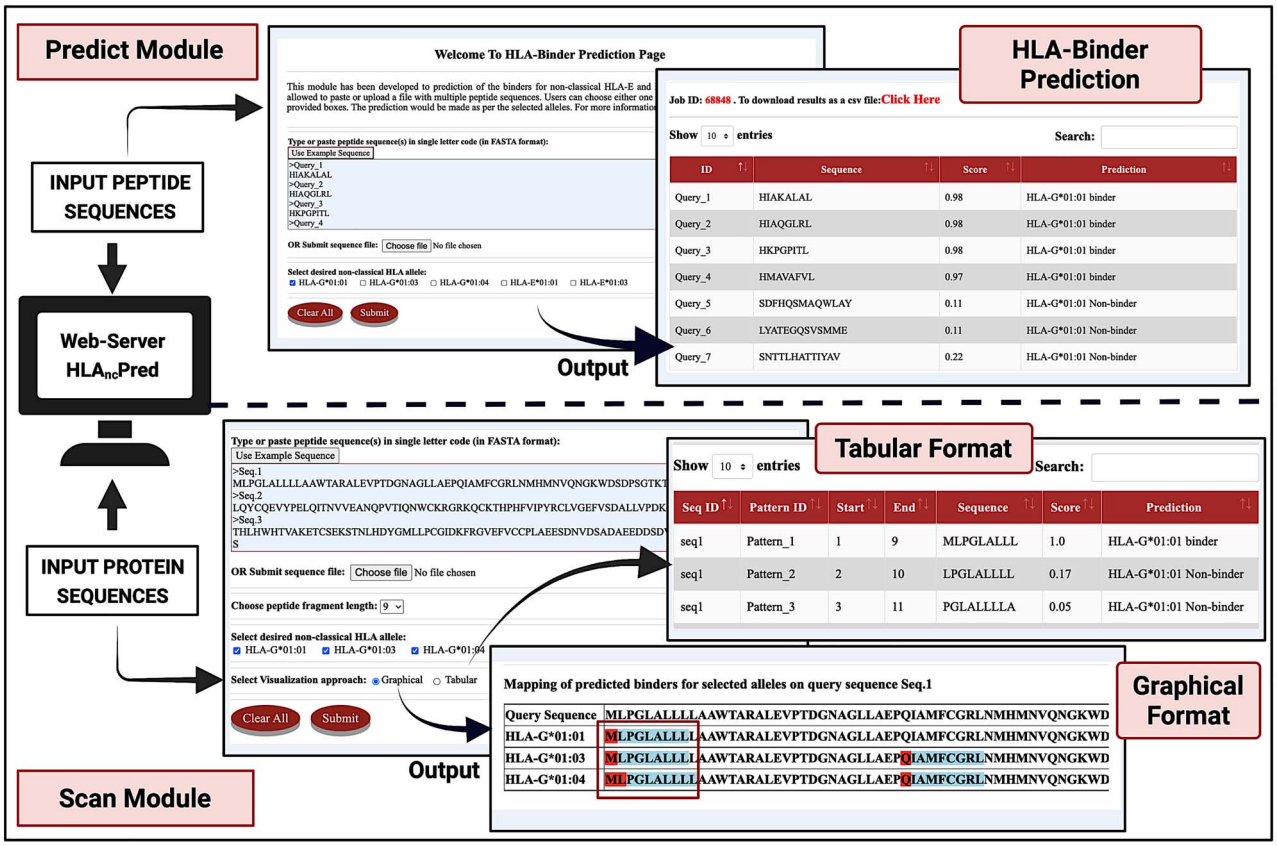

Figure 5. Usage of predict and scan module of HLAncPred.

(iii) PACKAGE: To serve the scientific community, we also provide a python- and Perl-based command-line package (https://webs.iiitd.edu.in/raghava/hlancpred/stand.html) for prediction of non-classical binders at a large scale and in the absence of the internet. In addition, we have also developed a docker-based standalone package of $HLA_{nc}Pred$ and integrated into the 'GPSRdocker' package https://webs.iiitd.edu.in/gpsrdocker/ [58].

(iv) DOWNLOAD: Users can download the datasets used in this study using this module.

The '$HLA_{nc}Pred$' web server is compatible with a number of devices (iPhone, iPad, laptops, android mobile phones) and was built using HTML, PHP and JAVA scripts.

Case study: non-classical HLA binders in COVID-19 variants

Several studies have shown that HLA allele binding sites significantly influence the severity of COVID-19 disease [8, 27, 41, 59]. Recently, World Health Organization (WHO) reported that the spike protein of coronavirus has shown more than 30 mutations in the new SARS-CoV-2 variant B.1.1.529 (Omicron). To investigate the effect of mutations on the HLA binding regions, we have used the reference spike protein of SARS-CoV-2 from NCBI. Further, we identified the substituted mutations named as A67V, Δ69–70, T95I, G142D, Δ143–145, Δ211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493K, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K and L981F,

Table 7. Alterations in the binding sites of non-classical HLA alleles by mutations in Spike protein of SARS-CoV-2 variant B.1.1.529 (Omicron)

Alleles	Mutation	Reference peptide	Mutated peptide	Prediction (binder/non-binder)	
				(Reference)	(Mutated)
HLA-G*01:01	**Δ 211, L212I**	DLPQGFSAL	DPIGEPESA	Binder	Non-binder
	S371L	VLYNSASFS	VLYNSASFL	Non-binder	Binder
	P681H	PSVASQSII	HSVASQSII	Binder	Non-binder
HLA-G*01:03	**Δ 211, L212I**	DLPQGFSAL	DPIGEPESA	Binder	Non-binder
	S371L	VLYNSASFS	VLYNSASFL	Non-binder	Binder
	D614G	DEVPVAIHA	GEVPVAIHA	Binder	Non-binder
	Q954H	QNAQQLNTL	QNAQHLNTL	Non-binder	Binder
	N969K	NSSVLNDIL	KSSVLNDIL	Non-binder	Binder
HLA-G*01:04	**Δ 211, L212I**	DLPQGFSAL	DPIGEPESA	Binder	Non-binder
	G339D	LCPFGEVFG	LCPFGEVFD	Non-binder	Binder
	T547K	TLTESNKKF	KLTESNKKF	Non-binder	Binder
	N679K	NSPRNARSV	NSPRKAHSV	Non-binder	Binder
	Q954H	QNAQQLNTL	QNAQHLNTL	Non-binder	Binder
	N969K	NSSVLNDIL	KSSVLNDIL	Non-binder	Binder
HLA-E*01:01	**Δ 211, L212I**	DLPQGFSAL	DPIGEPESA	Binder	Non-binder
	S371L	VLYNSASFS	VLYNSASFL	Non-binder	Binder
	T547K	NGLTGTGTL	NGLTGTGKL	Non-binder	Binder
HLA-E*01:03	**A67V, Del 69**	NVTWFHAIH	NVTWFHVIV	Non-binder	Binder
	Δ 211, L212I	DLPQGFSAL	DPIGEPESA	Binder	Non-binder
	G339D	LCPFGEVFG	LCPFGEVFD	Non-binder	Binder
	S371L	VLYNSASFS	VLYNSASFL	Non-binder	Binder
	S371L, S373P, S375F	FSTSKSYGV	FLTPKFYGV	Non-binder	Binder
	T547K	GLTGTGTLT	GLTGTGKLT	Binder	Non-binder
	N679K	NSPRNARSV	NSPRKAHSV	Non-binder	Binder
	D796Y	GGDNFSQIL	GGYNFSQIL	Non-binder	Binder

associated with the spike protein of B.1.1.529 variant from Centers for Disease Control and Prevention (CDC portal). In addition, we have also considered mutations in other variants such as in alpha variant (B.1.1.7) where seven mutations named as N501Y, A570D, D614G, P681H, T716I, S982A and D1118H were reported; similarly, in beta variant (B.1.351), nine mutations such as D80A, D215G, K417N, E484K, N501Y, D614G, A701V, L18F and R246I and in delta (B.1.617.2) strain nine mutations namely T19R, T95I, G142D, R158G, L452R, T478K, D614G, P681R and D950N were reported in spike protein according to CDC portal and Indian SARS-CoV-2 Genomics Consortium.

We created the mutated spike protein sequence by substituting the new mutations in the reference protein sequence. Then, we utilized the 'SCAN' module (with peptide length = 9) of HLA_{nc}Pred server, to identify HLA binding regions in the reference and altered spike proteins. After that, we mapped the disparities in the reference and variant spike protein sequences. We found that in alpha strain, a single substitution T716I in peptide 'NSIAIPINF' results in the gain of binding ability for HLA-G*01:01; similarly, E484K substitution in 'KGFNCYFPL' peptide made it a binder for HLA-G*01:03, HLA-G*01:04 and HLA-E*01:01. However, mutation L452R in peptide 'KVGGNYNYR' results in the loss of binding ability for HLA-G*01:03 and HLA-G*01:04 in delta strain. In the case of recent strain Omicron (B.1.1.529), mutations have resulted in the changes in the binding ability of various regions of spike proteins as shown in Table 7. The complete results for the binding prediction for an alpha, beta, delta and Omicron strains are provided in Supplementary Tables S6, S7, S8 and S9, respectively. We hope that these findings can be used by the scientific community for further investigation and designing potential immunotherapies/vaccines against the new or future COVID-19 strains.

Discussion and conclusion

The non-classical HLA, such as HLA-G, acts as an immunomodulatory molecule and natural guard while providing protection during fetus development [60, 61]. HLA-E triggers immune responses via activating inflammatory cytokines during viral infections [9, 33, 62, 63]. Of note, the over-expression of HLA-G may induce the immune-suppressive microenvironment, which may help in evading tumor cells from our innate and adaptive immune system. Studies have also shown that the abundant and aberrant expression of HLA-G leads to immune-mediated disorders such as multiple sclerosis and systemic lupus erythematosus [28, 60, 64]. Due to the low polymorphism of non-classical HLA, the HLA-E-restricted T-cell immunotherapy may be given to a heterogenic population, which may possess many benefits over classical HLA-based therapies. Moreover, the activation of anti-inflammatory immune response by HLA-E interaction with CD8+ T cell leads to the inhibition of cytokine storm and hence reduces the collateral tissue damage, which could be beneficial in treating COVID-19 patients [41]. On the other hand, HLA-G-based immunotherapies have been shown to achieve encouraging results in solid-cancer treatments. Researchers have

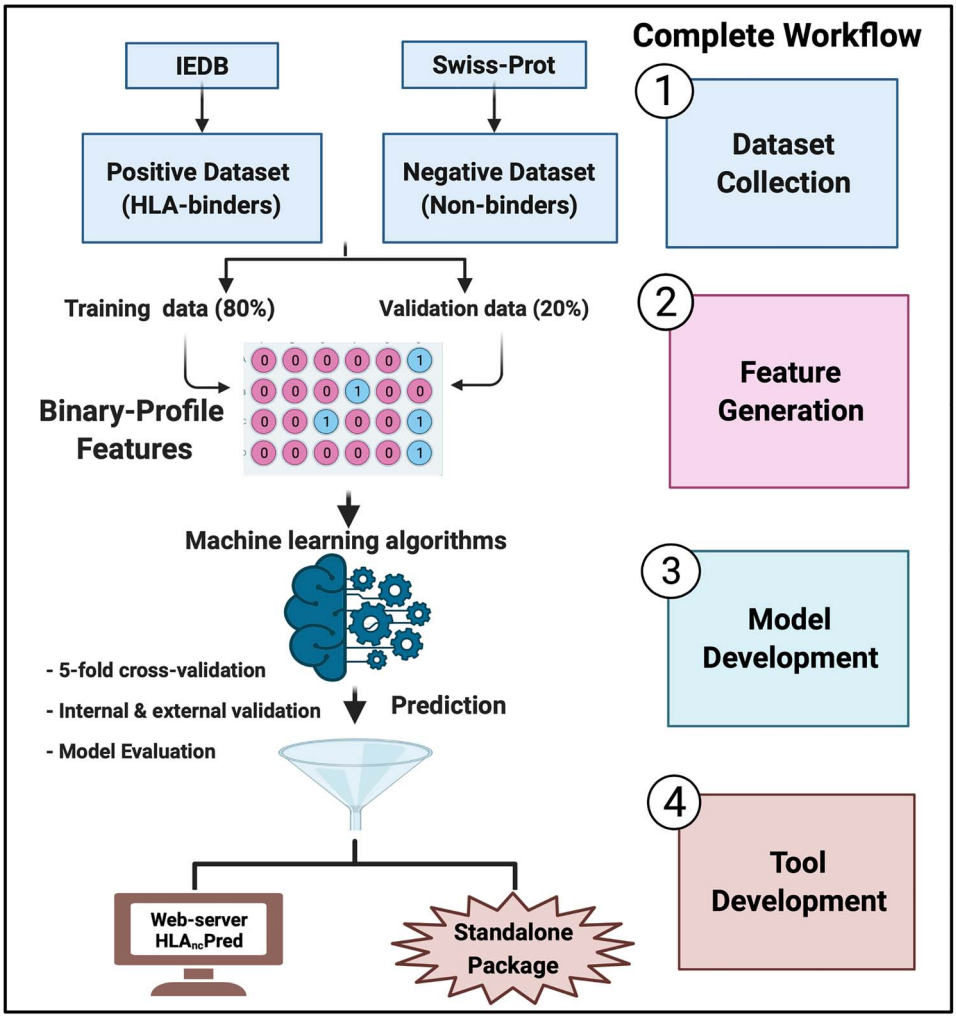

Figure 6. Complete architecture of HLAncPred including the dataset collection, feature generation, models and tool development.

designed an anti-HLA-G CAR-T cells immunotherapy for the treatment of acute lymphoblastic leukemia and B-cell malignancies [65].

Therefore, it is the utmost need to develop an accurate prediction method for the identification of non-classical HLA-binder peptides. In last few decades, a number of HLA binding peptide prediction methods have been developed, hitherto very limited prediction methods have been deployed in the non-classical binder prediction. To strength the previous works and to facilitate the researchers working in this area, we developed a highly accurate and efficient tool dedicated to the binder prediction of non-classical HLA alleles. The dataset takes an important role in machine learning model development; thus, we have constructed our major dataset from IEDB. For the training and validation dataset, we have used experimentally validated positive peptides i.e. HLA-G and HLA-E binders and negative data are randomly generated using Swiss-Prot. The composition and position conservation analysis indicates that non-classical HLA binding peptides are enriched in proline and leucine amino acids. The anchor motif residues K, L, P and V predominantly occurred at P1, P3 and P9 in HLA-G alleles, whereas M, L, V are highly conserved P2 and P9 in HLA-E alleles. Furthermore, various models were developed using N_8, C_8, N_8C_8 and AA_{15} binary profiles for each allele datasets. Our results indicate that models developed on AA_{15} binary profile–based features achieve highest performance (i.e. AUC 0.99 and accuracy >95%) on training and validation dataset for HLA-G*01:01 allele using SVC classifier. We have used the best models for each non-classical HLA binders and developed a web server named 'HLA_{nc}Pred' to predict and scan non-classical HLA binding peptides. We also provide the standalone package for the bulk scale non-classical HLA binder prediction.

Additionally, we utilized the SCAN module of our server for the prediction of non-classical HLA binding/non-binding peptides in the spike protein of new strain of SARS-CoV-2 i.e. B.1.1.529 (Omicron). We identified 431 promiscuous binders, where 80, 155, 56, 57 and 83 binders were predicted for HLA-E*01:01, - E*01:03, -G*01:01, -G*01:03, -G*01:04 allele, respectively. From the prediction, we observed that due to the new mutations occurred in B.1.1.529 (Omicron) variant, several peptide regions have shown the opposite binding trait in comparison with the reference spike protein (Table 7).

For instance, mutations Δ 211 and L212I changed the binding behavior of 'DLPQGFSAL' peptide for all non-classical HLA alleles. These observations can benefit the scientific community for designing vaccine against the deadly virus. In addition, researches can use this tool to predict the binding peptides of non-classical HLA alleles against different pathogenic, autoimmune and viral infections. We anticipate that this method will benefit the community working in the area of vaccine and HLA-based immunotherapy designing. The complete architecture of the HLA_{nc}Pred is shown in Figure 6.

Key Points

- Non-classical HLAs play immunomodulatory roles in the immune system.
- HLA-E-restricted T-cell therapy may reduce COVID-19-associated cytokine storm.
- In silico models were developed for the prediction of HLA-G and HLA-E binding peptides.
- Non-classical HLA-binding peptides for several strains of COVID-19 were predicted.
- A web-server and standalone package is available for the prediction of non-classical HLA alleles binding peptides.

Authors' contributions

AD, SP and GPSR collected and processed the datasets. AD, SP and GPSR implemented the algorithms and developed the prediction models. AD, SP and GPSR analyzed the results. SP and AD created the back-end of the web server the front-end user interface. AD, SP and GPSR penned the manuscript. GPSR conceived and coordinated the project. All authors have read and approved the final manuscript.

Data Availability

All the datasets generated in this study are available at 'HLA_{nc}Pred' web server, https://webs.iiitd.edu.in/raghava/hlancpred/down.php.

Biorxiv DOI: https://doi.org/10.1101/2021.12.04.471207

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors are thankful to the Department of Bio-Technology (DBT) and Department of Science and Technology (DST-INSPIRE) for fellowships and the financial support and Department of Computational Biology, IIITD, New Delhi, for infrastructure and facilities.

Funding Source

The current work has not received any specific grant from any funding agencies.

References

1. Marshall JS, Warrington R, Watson W, *et al.* An introduction to immunology and immunopathology. *Allergy Asthma Clin Immunol* 2018;**14**:49.
2. Chaplin DD. Overview of the immune response. *J Allergy Clin Immunol* 2010;**125**:S3–23.
3. Choo SY. The HLA system: genetics, immunology, clinical testing, and clinical implications. *Yonsei Med J* 2007;**48**:11–23.
4. Beck S, Trowsdale J. The human major histocompatability complex: lessons from the DNA sequence. *Annu Rev Genomics Hum Genet* 2000;**1**:117–37.
5. Robinson J, Barker DJ, Georgiou X, *et al.* IPD-IMGT/HLA database. *Nucleic Acids Res* 2020;**48**:D948–55.
6. Uzhachenko RV, Shanker A. CD8(+) T lymphocyte and NK cell network: circuitry in the cytotoxic domain of immunity. *Front Immunol* 2019;**10**:1906.
7. Blackwell JM, Jamieson SE, Burgner D. HLA and infectious diseases. *Clin Microbiol Rev* 2009;**22**:370–85 Table of Contents.
8. Tavasolian F, Rashidi M, Hatam GR, *et al.* HLA, immune response, and susceptibility to COVID-19. *Front Immunol* 2020;**11**:601886.
9. Crux NB, Elahi S. Human leukocyte antigen (HLA) and immune regulation: how do classical and non-classical HLA alleles modulate immune response to human immunodeficiency virus and hepatitis C virus infections? *Front Immunol* 2017;**8**:832.
10. Sabapathy K, Nam SY. Defective MHC class I antigen surface expression promotes cellular survival through elevated ER stress and modulation of p53 function. *Cell Death Differ* 2008;**15**:1364–74.
11. Aptsiauri N, Cabrera T, Mendez R, *et al.* Role of altered expression of HLA class I molecules in cancer progression. *Adv Exp Med Biol* 2007;**601**:123–31.
12. Mendez R, Aptsiauri N, Del Campo A, *et al.* HLA and melanoma: multiple alterations in HLA class I and II expression in human melanoma cell lines from ESTDAB cell bank. *Cancer Immunol Immunother* 2009;**58**:1507–15.
13. Johansen LL, Lock-Andersen J, Hviid TV. The pathophysiological impact of HLA class Ia and HLA-G expression and regulatory T cells in malignant melanoma: a review. *J Immunol Res* 2016;**2016**:6829283.
14. Amiot L, Vu N, Samson M. Immunomodulatory properties of HLA-G in infectious diseases. *J Immunol Res* 2014;**2014**:298569.
15. Murdaca G, Contini P, Negrini S, *et al.* Immunoregulatory role of HLA-G in allergic diseases. *J Immunol Res* 2016;**2016**:6865758.
16. Rouas-Freiss N, Khalil-Daher I, Riteau B, *et al.* The immunotolerance role of HLA-G. *Semin Cancer Biol* 1999;**9**:3–12.
17. Rizzo R, Trentini A, Bortolotti D, *et al.* Matrix metalloproteinase-2 (MMP-2) generates soluble HLA-G1 by cell surface proteolytic shedding. *Mol Cell Biochem* 2013;**381**:243–55.
18. Tronik-Le Roux D, Renard J, Verine J, *et al.* Novel landscape of HLA-G isoforms expressed in clear cell renal cell carcinoma patients. *Mol Oncol* 2017;**11**:1561–78.
19. Ho GT, Celik AA, Huyton T, *et al.* NKG2A/CD94 is a new immune receptor for HLA-G and distinguishes amino acid differences in the HLA-G heavy chain. *Int J Mol Sci* 2020;**21**(12):4362.
20. Carosella ED, Gregori S, Rouas-Freiss N, *et al.* The role of HLA-G in immunity and hematopoiesis. *Cell Mol Life Sci* 2011;**68**:353–68.
21. Kovats S, Main EK, Librach C, *et al.* A class I antigen, HLA-G, expressed in human trophoblasts. *Science* 1990;**248**:220–3.
22. Schmidt CM, Orr HT. Maternal/fetal interactions: the role of the MHC class I molecule HLA-G. *Crit Rev Immunol* 1993;**13**:207–24.
23. Shih IM. Application of human leukocyte antigen-G expression in the diagnosis of human cancer. *Hum Immunol* 2007;**68**:272–6.

24. Sheu J, Shih IM. HLA-G and immune evasion in cancer cells. *J Formos Med Assoc* 2010;**109**:248–57.
25. Amiot L, Ferrone S, Grosse-Wilde H, *et al.* Biology of HLA-G in cancer: a candidate molecule for therapeutic intervention? *Cell Mol Life Sci* 2011;**68**:417–31.
26. Rizzo R, Bortolotti D, Bolzani S, *et al.* HLA-G molecules in autoimmune diseases and infections. *Front Immunol* 2014;**5**:592.
27. Zidi I. Puzzling out the COVID-19: therapy targeting HLA-G and HLA-E. *Hum Immunol* 2020;**81**:697–701.
28. Contini P, Murdaca G, Puppo F, *et al.* HLA-G expressing immune cells in immune mediated diseases. *Front Immunol* 2020;**11**:1613.
29. Sabbagh A, Sonon P, Sadissou I, *et al.* The role of HLA-G in parasitic diseases. *HLA* 2018;**91**:255–70.
30. Catamo E, Zupin L, Crovella S, *et al.* Non-classical MHC-I human leukocyte antigen (HLA-G) in hepatotropic viral infections and in hepatocellular carcinoma. *Hum Immunol* 2014;**75**:1225–31.
31. Dias FC, Castelli EC, Collares CV, *et al.* The role of HLA-G molecule and HLA-G gene polymorphisms in Tumors. *Viral Hepatitis, and Parasitic Diseases, Front Immunol* 2015;**6**:9.
32. Kraemer T, Blasczyk R, Bade-Doeding C. HLA-E: a novel player for histocompatibility. *J Immunol Res* 2014;**2014**:352160.
33. Joosten SA, Sullivan LC, Ottenhoff TH. Characteristics of HLA-E restricted T-cell responses and their role in infectious diseases. *J Immunol Res* 2016;**2016**:2695396.
34. Romagnani C, Pietra G, Falco M, *et al.* Identification of HLA-E-specific alloreactive T lymphocytes: a cell subset that undergoes preferential expansion in mixed lymphocyte culture and displays a broad cytolytic activity against allogeneic cells. *Proc Natl Acad Sci U S A* 2002;**99**:11328–33.
35. Garcia P, Llano M, de Heredia AB, *et al.* Human T cell receptor-mediated recognition of HLA-E. *Eur J Immunol* 2002;**32**:936–44.
36. Jorgensen PB, Livbjerg AH, Hansen HJ, *et al.* Epstein-Barr virus peptide presented by HLA-E is predominantly recognized by CD8(bright) cells in multiple sclerosis patients. *PLoS One* 2012;**7**:e46120.
37. Pietra G, Romagnani C, Mazzarino P, *et al.* HLA-E-restricted recognition of cytomegalovirus-derived peptides by human CD8+ cytolytic T lymphocytes. *Proc Natl Acad Sci U S A* 2003;**100**:10896–901.
38. Mazzarino P, Pietra G, Vacca P, *et al.* Identification of effector-memory CMV-specific T lymphocytes that kill CMV-infected target cells in an HLA-E-restricted fashion. *Eur J Immunol* 2005;**35**:3240–7.
39. Romagnani C, Pietra G, Falco M, *et al.* HLA-E-restricted recognition of human cytomegalovirus by a subset of cytolytic T lymphocytes. *Hum Immunol* 2004;**65**:437–45.
40. Crew MD, Cannon MJ, Phanavanh B, *et al.* An HLA-E single chain trimer inhibits human NK cell reactivity towards porcine cells. *Mol Immunol* 2005;**42**:1205–14.
41. Caccamo N, Sullivan LC, Brooks AG, *et al.* Harnessing HLA-E-restricted CD8 T lymphocytes for adoptive cell therapy of patients with severe COVID-19. *Br J Haematol* 2020;**190**:e185–7.
42. Lee N, Llano M, Carretero M, *et al.* HLA-E is a major ligand for the natural killer inhibitory receptor CD94/NKG2A. *Proc Natl Acad Sci U S A* 1998;**95**:5199–204.
43. Yang Y, Liu Z, Wang H, *et al.* HLA-E binding peptide as a potential therapeutic candidate for high-risk multiple myeloma. *Front Oncol* 2021;**11**:670673.
44. Zhen Z, Yang K, Ye L, *et al.* HLA-E inhibitor enhances the killing of neuroblastoma stem cells by co-cultured dendritic cells and cytokine-induced killer cells loaded with membrane-based microparticles. *Am J Cancer Res* 2017;**7**:334–45.
45. Singh H, Raghava GP. ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 2001;**17**:1236–7.
46. Singh H, Raghava GP. ProPred1: prediction of promiscuous MHC class-I binding sites. *Bioinformatics* 2003;**19**:1009–14.
47. Chen B, Khodadoust MS, Olsson N, *et al.* Predicting HLA class II antigen presentation through integrated deep learning. *Nat Biotechnol* 2019;**37**:1332–43.
48. Jurtz V, Paul S, Andreatta M, *et al.* NetMHCpan-4.0: improved peptide-MHC class I interaction predictions integrating eluted ligand and peptide binding affinity data. *J Immunol* 2017;**199**:3360–8.
49. O'Donnell TJ, Rubinsteyn A, Laserson U. MHCflurry 2.0: improved pan-allele prediction of MHC class I-presented peptides by incorporating antigen processing. *Cell Syst* 2020;**11**:42–48 e47.
50. Ye Y, Wang J, Xu Y, *et al.* MATHLA: a robust framework for HLA-peptide binding prediction integrating bidirectional LSTM and multiple head attention mechanism. *BMC Bioinformatics* 2021;**22**:7.
51. Bhasin M, Raghava GP. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci* 2007;**32**:31–42.
52. Mei S, Li F, Xiang D, *et al.* Anthem: a user customised tool for fast and accurate prediction of binding between peptides and HLA class I molecules. *Brief Bioinform* 2021;**22**(5):bbaa415.
53. Mei S, Li F, Leier A, *et al.* A comprehensive review and performance evaluation of bioinformatics tools for HLA class I peptide-binding prediction. *Brief Bioinform* 2020;**21**:1119–35.
54. Crooks GE, Hon G, Chandonia JM, *et al.* WebLogo: a sequence logo generator. *Genome Res* 2004;**14**:1188–90.
55. Pande A, Patiyal S, Lathwal A, *et al.* Computing wide range of protein/peptide features from their sequence and structure. *BioRxiv*. 2019;**1**:599126.
56. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: machine learning in python. *Journal of Machine Learning Research* 2012;**12**:2825–30.
57. Waterhouse AM, Procter JB, Martin DM, *et al.* Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 2009;**25**:1189–91.
58. Agrawal P, Kumar R, Usmani SS, *et al.* GPSRdocker: a Docker-based resource for genomics, proteomics and systems biology. *BioRxiv*. 2019;**1**:827766.
59. Bouayad A. Features of HLA class I expression and its clinical relevance in SARS-CoV-2: what do we know so far? *Rev Med Virol* 2021;**31**:e2236.
60. Amodio G, Gregori S. HLA-G genotype/expression/disease association studies: success. *Hurdles, and Perspectives, Front Immunol* 2020;**11**:1178.
61. Xu X, Zhou Y, Wei H. Roles of HLA-G in the maternal-Fetal immune microenvironment. *Front Immunol* 2020;**11**:592010.
62. Kanevskiy L, Erokhina S, Kobyzeva P, *et al.* Dimorphism of HLA-E and its disease association. *Int J Mol Sci* 2019;**20**(21):5496.
63. Sharpe HR, Bowyer G, Brackenridge S, *et al.* HLA-E: exploiting pathogen-host interactions for vaccine development. *Clin Exp Immunol* 2019;**196**:167–77.
64. Morandi F, Rizzo R, Fainardi E, *et al.* Recent advances in our understanding of HLA-G biology: lessons from a wide Spectrum of human diseases. *J Immunol Res* 2016;**2016**:4326495.
65. Anna F, Bole-Richard E, LeMaoult J, *et al.* First immunotherapeutic CAR-T cells against the immune checkpoint protein HLA-G. *J Immunother Cancer* 2021;**9**(3).