

## Article

# In Silico Identification of HLA-DRB5\*01:01-Restricted Peanut Peptides

Irina Doytchinova <sup>1,2</sup> 

<sup>1</sup> Drug Design and Bioinformatics Lab, Faculty of Pharmacy, Medical University of Sofia, 1000 Sofia, Bulgaria; idoytchinova@pharmfac.mu-sofia.bg

<sup>2</sup> Centre of Excellence in Informatics and Information and Communication Technologies, 1113 Sofia, Bulgaria

## Featured Application

The computational workflow developed in this study enables the identification and prioritization of candidate HLA-DRB5\*01:01-restricted CD4+ T-cell epitopes from peanut allergens. This cost-effective approach facilitates the selection of candidate peptides for experimental validation and provides a practical framework for future studies on allergen characterization. The approach may facilitate studies of HLA-restricted antigen presentation, allergen characterization, and future immunological investigations.

## Abstract

Peanut allergy is a major food allergy associated with severe IgE-mediated hypersensitivity reactions and strong T-cell responses to peanut-derived epitopes. In this study, complementary computational approaches were developed to identify digestion-stable peanut peptides binding to HLA-DRB5\*01:01, an HLA class II molecule associated with peanut allergy. A dataset of experimentally validated binders and non-binders was used to construct one sequence logo model and three machine learning (ML) models based on Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB) algorithms. Peptide sequences were encoded using Wold's z-scale descriptors representing hydrophobic, steric, and electronic amino acid properties. All ML models showed excellent predictive performance, with ROC AUC values between 0.978 and 0.980, while the SVM model achieved the best overall accuracy (95.8%). Permutation-importance analysis identified the dominant peptide properties for HLA-DRB5 binding. The developed models were applied to major peanut allergens Ara h 1, Ara h 2, Ara h 3, and Ara h 6 following simulated gastrointestinal digestion. Multiple consensus HLA-DRB5-binding peptides were identified, including several experimentally reported allergenic epitopes. The identified peptides should be regarded as candidate HLA-DRB5\*01:01-restricted CD4+ T-cell epitopes that warrant further experimental validation using peptide–HLA binding assays and T-cell functional studies.



Academic Editors: José Antonio Canas, Blanca Cárdena and Rubén Fernández-Santamaría

Received: 2 June 2026

Revised: 3 July 2026

Accepted: 4 July 2026

Published: 10 July 2026

**Copyright:** © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

**Keywords:** peanut allergy; HLA-DRB5\*01:01; HLA binders; peptide binding prediction; machine learning; sequence logo analysis; food allergens; immunoinformatics; in silico epitope discovery

## 1. Introduction

Peanut allergy is one of the most common and severe food allergies worldwide and represents a major public health concern due to its early onset, persistence into adulthood, and risk of life-threatening anaphylaxis. Its prevalence has increased markedly over the past several decades, particularly in Western countries, affecting approximately 1–3% of

children and up to 1% of adults, with considerable geographic variability [1–3]. Unlike many other childhood food allergies, peanut allergy is typically lifelong, with spontaneous resolution occurring in only a minority of patients [4].

Peanut allergy is an immunoglobulin E (IgE)–mediated hypersensitivity reaction triggered by exposure to peanut proteins, leading to rapid activation of mast cells and basophils. Clinical manifestations range from mild cutaneous symptoms to severe systemic reactions, including bronchospasm, hypotension, and anaphylactic shock [5]. Even trace amounts of peanut can provoke reactions in sensitized individuals, making strict dietary avoidance challenging and significantly impacting quality of life for patients and caregivers [6].

At the molecular level, peanut allergy is driven by a defined set of allergenic proteins, primarily seed storage proteins such as Ara h 1, Ara h 2, Ara h 3, and Ara h 6 [7]. Among these, Ara h 2 and Ara h 6 are considered major allergens due to their high stability against heat and gastrointestinal digestion, as well as their strong IgE-binding capacity [8–10]. This structural resilience contributes to their potent allergenicity and explains the high risk of systemic reactions following oral exposure.

Management of peanut allergy has traditionally focused on strict avoidance and emergency preparedness, including the use of epinephrine autoinjectors [11]. However, emerging therapeutic approaches, such as oral, epicutaneous, and sublingual immunotherapy, aim to induce desensitization or sustained unresponsiveness, marking a paradigm shift in the treatment landscape [12–14].

Human leukocyte antigen (HLA) molecules play a central role in the development of peanut allergy by governing antigen presentation and shaping CD4<sup>+</sup> T-cell responses to peanut-derived peptides [15–17]. As peanut allergy is predominantly an IgE-mediated, T helper 2 (Th2)-driven disease, genetic variation within HLA class II genes is thought to influence individual susceptibility by determining which allergenic peptides are efficiently presented to the immune system [18].

HLA class II molecules (HLA-DR, -DQ, and -DP) are expressed on professional antigen-presenting cells (APC) and present processed peanut allergens to CD4<sup>+</sup> T cells. Several major peanut allergens, particularly Ara h 1, Ara h 2, Ara h 3 and Ara h 6, contain multiple T-cell epitopes that bind promiscuously but with variable affinity to different HLA class II alleles [19–21]. This differential binding affects T-cell activation, cytokine polarization, and downstream B-cell class switching to IgE. Experimental studies have shown that peanut-specific T-cell responses are dominated by Th2 cytokines (IL-4, IL-5, IL-13) in allergic individuals, whereas non-allergic or tolerant individuals exhibit more balanced or regulatory responses [22]. HLA polymorphisms influence this balance by shaping the HLA-peptide-T-cell receptor interaction. Mapping studies have identified immunodominant T-cell epitopes within peanut allergens that are presented in an HLA-restricted manner [23]. For example, Ara h 2-derived peptides bind efficiently to several common HLA-DR molecules and elicit strong Th2 responses in peanut-allergic individuals [19]. These findings support the concept that peanut allergy is driven by a limited set of HLA-restricted allergenic peptides rather than the entire allergen repertoire.

HLA diversity may partly explain interindividual variability in responses to peanut immunotherapy. Differences in peptide presentation can influence T-cell reprogramming, induction of regulatory T cells, and the durability of desensitization or sustained unresponsiveness. Incorporating HLA typing into clinical studies may therefore improve patient stratification and help identify biomarkers predictive of treatment success or adverse reactions [24].

Our research is focused on the association between HLA class II polymorphism and food allergy, the development of computational methods for allergenicity prediction, and the identification of HLA-restricted candidate peptides for allergy immunotherapy [25–29].

The present study represents a continuation of this research, extending our previous work on HLA-DRB1\*03:01 to HLA-DRB5\*01:01, another HLA class II molecule implicated in peanut allergy [19,30]. Here, we employed a bioinformatic approach to identify peanut peptides with binding affinity to HLA-DRB5\*01:01. To achieve this goal, we developed one sequence logo model and three complementary machine learning (ML) models based on Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting (GB) algorithms, which were subsequently applied to screen the amino acid sequences of the major peanut allergens Ara h 1, Ara h 2, Ara h 3, and Ara h 6. ML methods have become important tools for peptide and epitope prediction because they can identify complex sequence–activity relationships that are difficult to capture using conventional motif-based approaches. RF is an ensemble-learning method that combines multiple decision trees to improve predictive accuracy and robustness while reducing overfitting [31]. SVM constructs optimal decision boundaries in high-dimensional feature space and has strong generalization performance [32]. GB sequentially combines weak learners into a powerful predictive model by iteratively correcting previous prediction errors and has demonstrated excellent performance in numerous bioinformatics applications [33]. These algorithms have been successfully employed for prediction of peptide-MHC interactions [34], T-cell epitopes [34], antimicrobial peptides [30], and other biologically active peptide classes [35,36]. Their ability to integrate diverse physicochemical descriptors makes them particularly suitable for modeling peptide recognition by HLA molecules.

## 2. Datasets and Methods

### 2.1. Datasets Used in the Study

A set of 180 experimentally validated nonamer peptides binding to HLA-DRB5\*01:01, identified by ligand presentation mass spectrometry on cellular MHC, together with 215 experimentally confirmed non-binding peptides of varying lengths, was retrieved from the Immune Epitope Database (IEDB) [34] (<https://www.iedb.org/>, accessed on 10 February 2026, settings: assay: MHC Ligand, outcome: Positive and Negative, MHC Restriction: HLA-DRB5\*01:01). Each non-binding peptide was further represented as a set of overlapping non-binding nonamers. Duplicate sequences within and between the binding and non-binding datasets were removed. From the resulting pool, 180 non-binding nonamers exhibiting the greatest dissimilarity in amino acid composition from the 180 binding peptides were selected. Furthermore, the peptides were divided into training and test sets at a 4:1 ratio. The test set was selected to be maximally representative of the training set by including peptides with the highest similarity to the compositional centroid of the training set. The training set was used for model development, whereas the test set served for independent validation of model performance.

### 2.2. Logo Protocol

The amino acid frequencies at each of the nine positions in the sets of binders and nonbinders were mean-normalized across all positions, producing values in the range from −1 to 1. Positive values indicated residues favored at a given position, whereas negative values corresponded to disfavored residues. The normalized frequencies were arranged into a  $9 \times 20$  quantitative matrix (QM), representing nine sequence positions and the twenty natural amino acids. This QM representation, referred to as a logo model [29,37–39], was designed to capture residue preference patterns in a manner analogous to sequence logos, in which the relative abundance of amino acids at aligned positions reflects sequence conservation [40]. Separate logo models were generated from the training data for binding and non-binding peptides. The derived QMs were applied to calculate BSs and NBSs for each nonamer in the test set. Classification was performed by comparing the two scores:

peptides with BS values exceeding NBS values were assigned as binders, while peptides with BS values equal to or lower than NBS values were classified as non-binders.

### 2.3. Machine Learning Protocols

The peptide sequences were transformed into numerical representations using Wold's z-scale descriptors, which encode three major physicochemical characteristics of amino acids: hydrophobicity (z1), steric properties (z2), and electronic effects (z3) [41]. Each residue in the nonamer sequence was represented by these three parameters, yielding a 27-dimensional feature vector for every peptide (9 sequence positions  $\times$  3 descriptors).

These datasets were subsequently used for the development of supervised machine learning (ML) classification models implemented with ChatGPT-5.2 [42]. Three ML approaches were evaluated: RF [31], SVM [32], and GB [33]. Model training was performed on the training datasets, while predictive performance was assessed using the independent test sets.

The RF model was configured using 300 decision trees with Gini impurity as the splitting criterion, feature subsampling, and unrestricted tree depth to efficiently capture nonlinear relationships while reducing variance. The SVM classifier employed a radial basis function (RBF) kernel with optimized C and  $\gamma$  parameters following standard feature-scaling procedures. For the GB model, optimization included adjustment of tree depth, learning rate, number of estimators, and loss function to balance predictive accuracy against overfitting. Hyperparameter tuning followed a controlled performance-based strategy rather than exhaustive grid-search or Bayesian optimization. Considering the moderate size of the datasets and the availability of external validation sets, this approach was selected to reduce the likelihood of overfitting associated with overly aggressive parameter optimization. The complete source codes for the ML models used in this study are provided in Supplementary Files S1–S3.

To interpret model predictions, permutation-importance analysis [43] was carried out. Permutation-importance violin plots were employed to visualize the distribution of feature contributions across the test datasets. Broad distributions correspond to features exerting strong and variable effects on peptide-binding predictions, whereas distributions centered near zero indicate features with limited predictive importance.

### 2.4. Model Validation Metrics

The performance of the derived models was evaluated in terms of true positives (TP = binders predicted as binders), true negatives (TN = non-binders predicted as non-binders), false positives (FP = non-binders predicted as binders), false negatives (FN = binders predicted as non-binders), sensitivity (true binders/all binders), specificity (true non-binders/all non-binders), accuracy ((true binders and true non-binders)/all peptides), and the area under the receiver operating characteristic curve (ROC AUC).

## 3. Results

### 3.1. Logo Models for HLA-DRB5 Binding Prediction

The datasets used to derive the logo models are presented in Tables S1–S4. The amino acid frequencies at each peptide position in the training set of 144 binders were normalized towards the mean frequency over all positions and arranged into a quantitative matrix (QM) (Table S5). Amino acids with positive coefficients favor binding to HLA-DRB5\*01:01, while those with negative coefficients do not. Amino acids with coefficients around zero are neutral. Similarly, a QM was generated for the training set of 144 non-binders (Table S6). The two QMs were validated by the external test sets of 36 binders and 36 non-binders. The binding and non-binding scores for each peptide were calculated by the corresponding

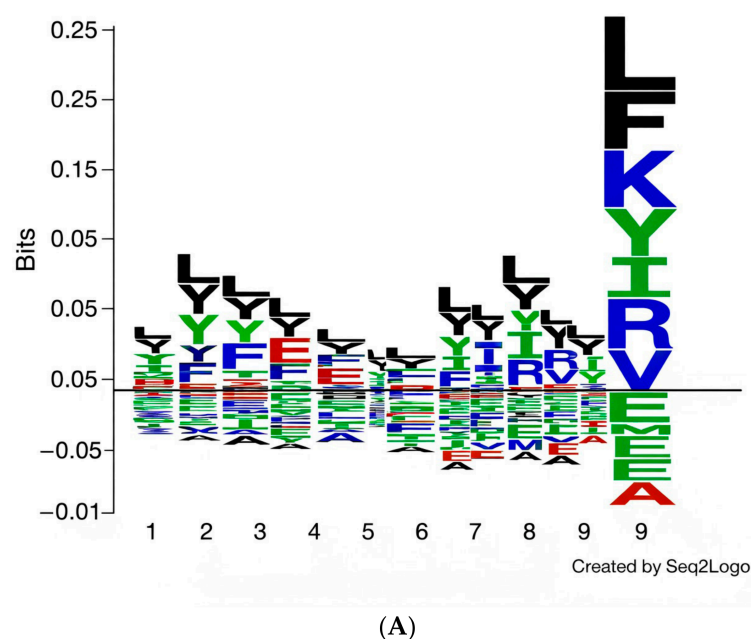
QM. If the BS is higher than the NBS, the peptide is classified as a binder, otherwise—as a non-binder. The results from the validation are summarized in Table 1.

**Table 1.** Model validation metrics on the external test set of 36 binders and 36 non-binders to HLA-DRB5\*01:01. The training set contains 144 binders and 144 non-binders.

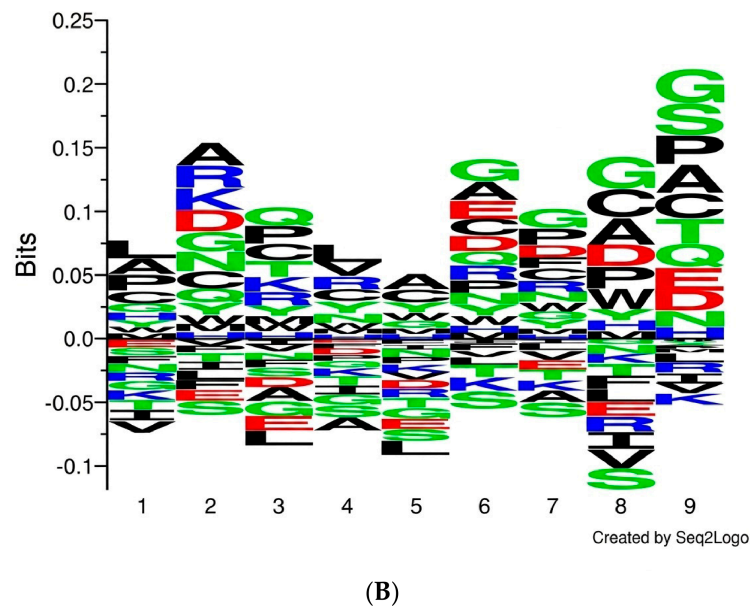
| Metrics           | Logo Model            | Support Vector Machine (SVM) Model | Random Forest (RF) Model | Gradient Boosting (GB) Model |
|-------------------|-----------------------|------------------------------------|--------------------------|------------------------------|
| Optimal threshold | BS > NBS <sup>1</sup> | 0.5                                | 0.5                      | 0.5                          |
| True positives    | 36                    | 33                                 | 31                       | 32                           |
| True negatives    | 36                    | 36                                 | 36                       | 35                           |
| False positives   | 0                     | 0                                  | 0                        | 1                            |
| False negatives   | 0                     | 3                                  | 5                        | 4                            |
| Sensitivity       | 1.000                 | 0.917                              | 0.861                    | 0.889                        |
| Specificity       | 1.000                 | 1.000                              | 1.000                    | 0.972                        |
| Accuracy          | 1.000                 | 0.958                              | 0.931                    | 0.931                        |
| F1 score          | 1.000                 | 0.957                              | 0.925                    | 0.928                        |
| MCC               | 1.000                 | 0.920                              | 0.870                    | 0.864                        |
| ROC AUC           | 1.000                 | 0.978                              | 0.980                    | 0.979                        |

<sup>1</sup> If the BS is higher than the NBS, the peptide is classified as a binder, and otherwise, it is classified as a non-binder.

The sequence logo for peptides binding to HLA-DRB5\*01:01 is given in Figure 1A. Both logo model and sequence logo reveal that p9 is the dominant anchor position, displaying strong enrichment of bulky hydrophobic, aromatic and positively charged residues. Among the remaining anchor positions, only p6 showed a noticeable preference for hydrophobic amino acids, whereas p1 and p4 exhibited minimal sequence conservation. This suggests that peptide binding is primarily driven by interactions within the p9 pocket, with p6 providing an additional hydrophobic stabilization contribution. However, the observed motif may also be influenced by the selective analysis of exact nonamers extracted from a larger set of variable-length binders. Because MHC class II molecules accommodate multiple binding registers and peptide flanking regions, restricting the dataset to fixed-length cores may reduce detectable preferences at other anchor positions and accentuate the apparent dominance of the p9 anchor. The identified preferences at p6 and p9 are consistent with known the known binding motif for HLA-DRB5 [44,45]. The observed motif does not, by itself, validate peptide presentation, T-cell recognition, or allergenic relevance.



**Figure 1.** Cont.



**Figure 1.** The derived sequence logo for peptides binding (A) and not binding (B) to HLA-DRB5\*01:01. The peptide positions are shown along the X-axis. The sequence logos were generated by Seq2Logo (<https://services.healthtech.dtu.dk>, accessed on 7 May 2026).

Unlike the logo model for binders, which is defined by strong, position-specific anchor residues, the logo model (Table S6) and sequence logo (Figure 1B) for non-binders are dominated by broadly tolerated, low-information amino acids and lack the positional specificity required for stable HLA-DRB5 engagement.

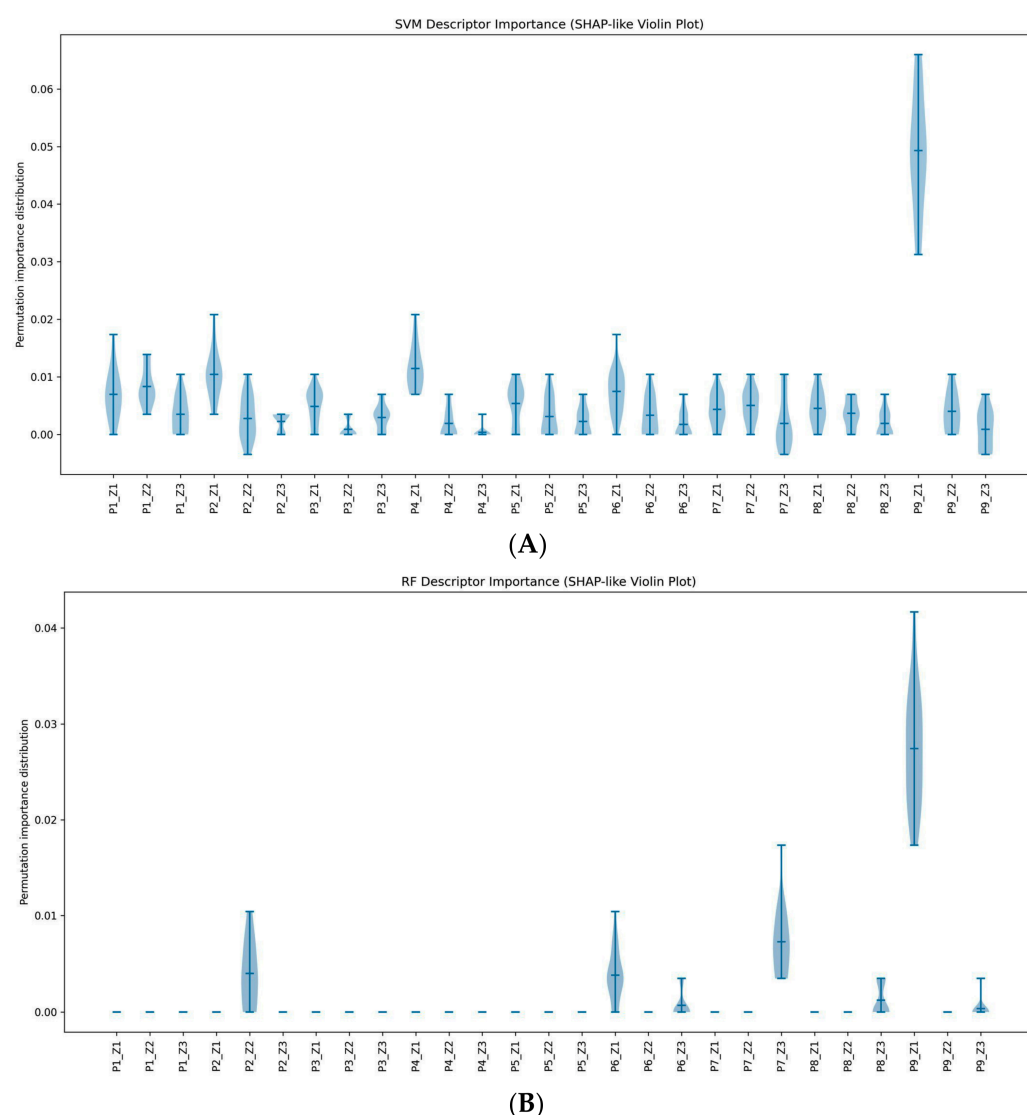
### 3.2. Machine Learning Models for HLA-DRB5 Binding Prediction

The datasets used to derive the ML models are presented in Tables S7–S10. The training set of 144 binders and 144 non-binders was used to derive the ML models, while the test set of 36 binders and 36 non-binders was used to validate them. The RF model was constructed with 300 decision trees ( $n\_estimators = 300$ ), while the SVM model employed a radial basis function (RBF) kernel with probability estimation enabled. The GB model was generated using the default GradientBoostingClassifier settings. For all models, the `random_state` parameter was set to 42 to ensure reproducibility. Binder probabilities  $\geq 0.5$  were classified as binders. The validation metrics are given in Table 1.

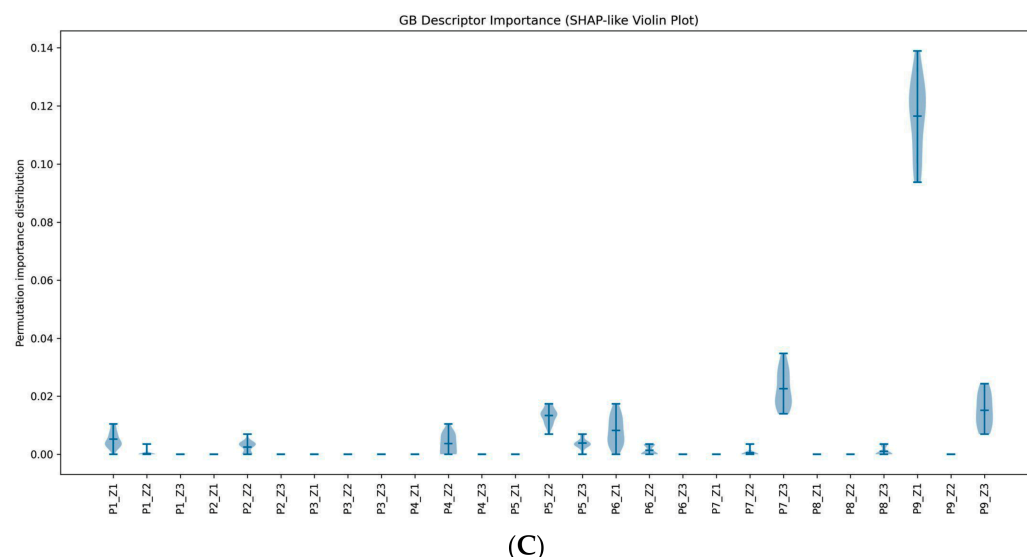
All three ML models showed excellent predictive performance, indicating that the peptide encoding with z-scales effectively captured the physicochemical determinants of DRB5 peptide binding. The SVM model showed the best overall performance, achieving the highest accuracy (95.8%), sensitivity (91.7%), F1 score (0.957), and MCC (0.92), while maintaining perfect specificity (100%) with no false positives. The RF and GB models also performed very well, both reaching an accuracy of 93.1% and MCC values around 0.87, indicating strong predictive stability and balanced classification. The very high AUC values for all models (0.978–0.980) demonstrate excellent discrimination between binders and non-binders. All models produced extremely low numbers of false positives, suggesting reliable identification of true binders, which is particularly important for binder screening applications. Overall, the results indicate that the developed ML models are robust, highly predictive, and suitable for virtual screening of DRB5-binding peptides.

The permutation-importance violin plots revealed that the three ML models relied predominantly on a limited subset of Wold's z-scale descriptors for peptide classification. In all models, the descriptor `p9_z1` exhibited the highest importance, indicating that hydrophobicity-related properties at peptide p9 play a dominant role in DRB5 bind-

ing prediction. In the SVM model (Figure 2A), p9\_z1 showed both the highest mean importance and the broadest distribution, while additional descriptors, including p2\_z1, p4\_z1, p6\_z1, p7\_z2, and p8\_z3, also contributed moderately, suggesting a more distributed classification strategy involving multiple physicochemical variables. In contrast, the RF model (Figure 2B) displayed a sparser importance profile dominated mainly by p9\_z1, with secondary contributions from p7\_z3, p6\_z1, and p2\_z2, whereas most other descriptors exhibited near-zero importance and very narrow violin distributions, indicating stable importance estimates across permutation repeats. Similarly, the GB model (Figure 2C) showed an exceptionally strong dominance of p9\_z1 together with smaller contributions from p7\_z3, p6\_z1, and p8\_z2, while the majority of descriptors contributed minimally to classification. The analyses indicate that hydrophobicity at peptide p9 represents the most influential physicochemical determinant across all three ML models, although the SVM classifier distributes predictive importance more broadly than the tree-based RF and GB approaches.



**Figure 2.** *Cont.*



**Figure 2.** Permutation-importance violin plots of the 27 Wold’s z-scale descriptors used in the SVM (A) RF (B) and GB (C) models for HLA-DRB5\*01:01 peptide-binding prediction. In all three models, p9\_z1 displayed the highest importance, indicating a dominant contribution of hydrophobicity-related properties at peptide p9, while the SVM model showed a broader distribution of contributing descriptors compared with the sparser RF and GB importance profiles.

### 3.3. Identification of Peanut Peptides Binding to HLA-DRB5

The amino acid sequences of peanut allergens were retrieved from UniProt [46] and subjected to in silico gastrointestinal (GI) digestion using PeptideCutter [47], applying cleavage rules for pepsin (pH 1.3 and pH > 2), trypsin, and chymotrypsin (high and low specificity). Digestion products shorter than nine amino acid residues were excluded from further analysis. The resulting digestion-stable peptides longer than nine amino acids were further fragmented into overlapping nonamer sequences. Binding to HLA-DRB5\*01:01 was predicted using the developed computational models, and peptides identified as binders by the logo model and all three ML models were classified as binders. Table 2 summarizes the digestion-stable peptides and the consensus HLA-DRB5\*01:01-binding nonamers derived from them. Where available, experimental evidence supporting the predicted binders is also included in Table 2.

Several of the consensus HLA-DRB5\*01:01 binders identified in the present study overlap with experimentally reported peanut T-cell epitopes, providing independent support for the biological relevance of the developed prediction models. In Ara h 1, the digestion-stable fragment IVQIEAKPNTL (214–224) contains the predicted binder QIEAKPNTL, which overlaps a previously reported T-cell epitope identified by tetramer-guided epitope mapping [19]. Similarly, the Ara h 3 fragment RRNALRRPFYSNAPLE (77–92) yielded the predicted binder RPFYSNAPL, while the fragment EQGAIVTVKGGLRILSPDEED (296–316) generated the predicted binders AIVTVKGGL and TVKGGLRIL; both regions have been reported as immunologically active peanut T-cell epitopes [30]. These findings demonstrate that the computational workflow successfully recovered peptide regions previously identified by experimental studies. Importantly, the current approach further prioritizes digestion-stable peptide fragments that are more likely to survive gastrointestinal processing and become available for HLA presentation in vivo. The agreement between computational predictions and experimentally validated epitopes supports the utility of the developed models for identifying candidate peptides for peanut allergy immunotherapy.

**Table 2.** Digestion-stable peanut allergen fragments and their binders to HLA-DRB5\*01:01 predicted by logo and ML models.

| Peanut Allergens        | Positions | Digestion-Stable Peptide Sequence | HLA-DRB5*01:01 Binders |
|-------------------------|-----------|-----------------------------------|------------------------|
| Ara h 1<br>E5G076 *     | 18–26     | ASVSATHAK                         | ASVSATHAK              |
|                         | 214–224   | IVQIEAKPNTL **                    | QIEAKPNTL              |
|                         | 237–253   | VIQQGQATVTVANGNNR                 | QQGQATVTVA             |
|                         | 288–298   | ISMPVNTPGQF                       | no binder              |
|                         | 357–367   | SSENEGIVVK                        | ENNENEGVIVK            |
|                         | 387–400   | GSEEEGDITNPINL                    | EEEGDITNP, GDITNPINL   |
|                         | 452–460   | AMVIVVVNK                         | AMVIVVVNK              |
|                         | 481–495   | EEEEDEDEEEGSNR **                 | no binder              |
|                         | 511–526   | IMPAAHPVAINASSEL                  | no binder              |
|                         | 548–556   | DNVIDQIEK                         | no binder              |
|                         | 562–572   | AFPGSGEQVEK                       | AFPGSGEQV, PGSGEQVEK   |
|                         | 605–615   | EDQEEENQGGK                       | QEEENQGGK              |
| Ara h 2<br>A5Z1Q9 *     | 33–42     | CQSQLERANL                        | QSQLERANL              |
|                         | 43–51     | PCEQHLMQK                         | PCEQHLMQK              |
|                         | 109–117   | CMCEALQQI                         | CMCEALQQI              |
|                         | 129–139   | RQQEQQFKREL                       | QEQQFKREL              |
|                         | 150–164   | APQRCDLEVESGGRDR                  | no binder              |
| Ara h 3<br>E5G077 *     | 46–58     | RIESEGGYIETWN                     | SEGGYIETW              |
|                         | 59–72     | PNNQEFQCAGVALS                    | EFQCAGVAL              |
|                         | 77–92     | RRNALRRPFYSNAPLE ***              | RPFYSNAPL              |
|                         | 96–116    | IYVQQGSGYFGLIFPGCPSTY             | no binder              |
|                         | 140–163   | FQVGQDDPSQQQQDSHQKVHRFDE          | DSHQKVHRF              |
|                         | 184–203   | DETDVVTVTLSDTSSIHNQ               | DTDVVTVTL              |
|                         | 252–270   | GDEQENEGSNIFSGFAQEF               | no binder              |
|                         | 273–293   | LQHAFQVDRQTVENLRGENERE            | AFQVDRQTV, VDRQTVENL   |
|                         | 296–316   | EQGAIVTVKGGLRILSPDEED ***         | AIVTVKGGL, TVKGGLRIL   |
|                         | 317–340   | ESSRSPNRRREEFDEDRSRPQQRG          | SSRSPNRR, SPPNRRREEF   |
|                         | 361–379   | ASVKKNLGRSSNPDIYNPQ               | ASVKKNLGR, LGRSSNPDI   |
|                         | 381–403   | AGSLRSVNELDLPILGWLGLSAQ           | DLPLGLWL, LPILGWLGL    |
|                         | 446–466   | YDEELQEGHVLVVPQNFAVAA             | GHVLVVPQN              |
|                         | 487–507   | ANQAGENSIIDNLPEEVVANS             | no binder              |
| Ara h 6<br>A0A5Q0V3Y4 * | 33–45     | SSCERQVDRVNLK                     | SSCERQVDR, RQVDRVNLK   |
|                         | 46–54     | PCEQHIMQR                         | PCEQHIMQR              |
|                         | 109–121   | LQDRQMVMQQFKRE                    | RQMVMQQFKR             |

\* UniProt/GenBank Protein number; \*\* [19]; \*\*\* [30].

#### 4. Discussion

This study demonstrates that HLA-DRB5\*01:01 peptide binding can be predicted with high accuracy using complementary computational approaches based on sequence logos and machine learning. By combining residue-frequency-based quantitative matrices with physicochemical descriptor-driven ML models, we identified a set of digestion-stable peanut-derived peptides with strong predicted affinity for HLA-DRB5\*01:01. Several of the predicted binders correspond to experimentally reported allergenic epitopes, supporting the biological relevance of the developed models and confirming the utility of in silico approaches for epitope discovery in food allergy research.

The study confirmed the dominant contribution of the p9 anchor position in the HLA-DRB5\*01:01 binding motif [44,45]. Both the sequence logo and the quantitative matrix revealed a strong enrichment of bulky hydrophobic, aromatic, and positively charged residues at p9. This pattern is consistent with the structural organization of HLA class II peptide-binding grooves, where specific pockets contribute differently to peptide stabi-

lization and T-cell receptor recognition. The observed preference at p9 suggests that the corresponding pocket in HLA-DRB5\*01:01 imposes a major selectivity filter for peptide binding. In contrast, the relatively weak conservation at p1 and p4 may reflect greater permissiveness of these pockets or the existence of multiple compatible binding registers. Because MHC class II molecules bind peptides in an open-ended groove, peptides longer than the canonical nine-residue core may shift within the groove and adopt alternative registers. Consequently, restriction of the analysis to exact nonamers may emphasize the strongest anchor positions while underestimating weaker but still biologically relevant preferences at other sites. The differences between the binder and non-binder logos are also informative. While the binder dataset exhibited clear positional enrichment patterns, the non-binder sequences were characterized by diffuse amino acid distributions lacking strong positional specificity.

The three ML approaches evaluated in this work—RF, SVM, and GB—all showed excellent predictive performance, with ROC AUC values close to 1.0. Among them, the SVM model achieved the best balance between sensitivity and specificity. These results indicate that Wold's z-scale descriptors successfully captured the physicochemical determinants underlying HLA-DRB5 peptide recognition. Hydrophobicity, steric effects, and electronic properties are known to play major roles in peptide–MHC interactions, particularly in determining compatibility with anchor pockets and stabilizing intermolecular contacts. The strong performance of the models therefore supports the relevance of these descriptors for representing HLA-binding peptides in compact numerical form.

An important advantage of the current strategy is the use of complementary modeling approaches rather than reliance on a single predictor. The logo model provides mechanistic interpretability by directly visualizing amino acid preferences at each peptide position, whereas ML models capture nonlinear relationships and interactions between residues that may not be evident from sequence frequencies alone. The consensus strategy requiring agreement between the logo and ML models likely reduced false-positive predictions and increased confidence in the identified binders. This is particularly relevant for allergen epitope discovery, where high specificity is essential to avoid unnecessary experimental validation of irrelevant candidates.

Application of the models to the major peanut allergens Ara h 1, Ara h 2, Ara h 3, and Ara h 6 identified numerous digestion-stable peptides predicted to bind HLA-DRB5\*01:01. Several of these peptides are derived from regions previously implicated in peanut allergenicity and T-cell activation [19,30]. T-cell epitopes from peanut allergens have been experimentally identified using tetramer-guided epitope mapping (TGEM), in which the complete allergen sequence was synthesized as overlapping 20-mer peptides with a 12-amino-acid overlap [19,30]. Peptide pools were loaded onto recombinant HLA-DR molecules to generate class II tetramers, and PBMCs from peanut-allergic subjects were stimulated with these pools and screened for tetramer-positive T cells. This methodology, however, does not account for gastrointestinal digestion of the proteins, and many of the predicted T-cell epitopes are unlikely to survive transit through the gastrointestinal tract. Consequently, a substantial proportion of the identified epitopes may never be generated *in vivo* or become available for presentation by HLA class II molecules, potentially limiting their physiological relevance.

The focus on digestion-stable peptides is biologically important. Proteins susceptible to rapid gastrointestinal degradation are less likely to reach the intestinal immune system in an immunologically intact form. In contrast, digestion-resistant fragments may persist long enough to be taken up by antigen-presenting cells and processed for HLA presentation. The combination of simulated gastrointestinal digestion with HLA-binding prediction therefore represents a rational strategy for prioritizing physiologically relevant allergenic

peptides. This approach may also have applications beyond peanut allergy, including prediction of immunogenic food peptides, assessment of novel food proteins, and design of peptide-based immunotherapies. Although simulated gastrointestinal digestion provides a practical approximation of proteolysis, it does not account for interindividual differences in digestive physiology or food-matrix effects, and the resulting digestion-stable peptides should therefore be considered computational candidates pending validation by standardized in vitro digestion assays.

Despite the strong predictive performance of the models, several limitations should be considered. First, the datasets used for model development were relatively moderate in size, reflecting the limited number of experimentally validated HLA-DRB5 ligands currently available in public databases. Although the use of external test sets reduced the risk of overfitting, larger and more diverse datasets would likely improve model generalizability. Second, the current analysis focused exclusively on HLA-DRB5\*01:01, whereas peanut allergy is influenced by multiple HLA class II alleles. Because allergen presentation often involves promiscuous peptide binding across several HLA molecules, extension of the approach to HLA-DQ and HLA-DP alleles associated with peanut allergy would provide a more comprehensive view of population-level susceptibility. Third, peptide-HLA binding alone does not fully determine allergenicity. T-cell receptor recognition, antigen processing efficiency, cytokine polarization, and immune regulatory mechanisms also contribute critically to the development of allergic responses. Consequently, the peptides identified in this study should be regarded as candidate HLA-DRB501:01-restricted CD4<sup>+</sup> T-cell epitopes rather than confirmed allergenic or immunotherapeutic peptides. Experimental validation using HLA-typed PBMCs from peanut-allergic donors, peptide-induced CD4<sup>+</sup> T-cell activation assays, Th2/Treg cytokine profiling, serum IgE-binding assays, and basophil activation studies will be required to establish their biological significance.

## 5. Conclusions

This study demonstrates that computational modeling can effectively identify HLA-DRB5\*01:01-binding peptides from major peanut allergens and provides a practical framework for systematic in silico epitope discovery in food allergy research. The combination of interpretable sequence-based models with ML approaches offers both mechanistic insight and strong predictive power, supporting their future application in allergen characterization. The identified peptides represent candidate HLA-DRB5\*01:01-restricted CD4<sup>+</sup> T-cell epitopes that provide a basis for future experimental studies aimed at validating their immunological relevance and potential utility in allergy research.

**Supplementary Materials:** The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/app16146926/s1>: Table S1: Positive training set for development of logo model for binders to HLA-DRB5; Table S2: Negative training set for development of logo model for non-binders to HLA-DRB5; Table S3: Positive test set of binders to HLA-DRB5 for validation of the logo models; Table S4: Negative test set of non-binders to HLA-DRB5 for validation of the logo models; Table S5: Logo model for binders to HLA-DRB5; Table S6: Logo model for non-binders to HLA-DRB5; Table S7: Positive training set for development of ML models for peptide binding to HLA-DRB5; Table S8: Negative training set for development of ML models for peptide binding to HLA-DRB5; Table S9: Positive test set of binders to HLA-DRB5 for validation of the ML models; Table S10: Negative test set of non-binders to HLA-DRB5 for validation of the ML models; File S1: Source code of the RF model; File S2: Source code of the SVM model; File S3: Source code of the GB model.

**Funding:** This research was funded by GIANT LEAPS, HORIZON-CL6-2021-FARM2FORK-01, European Research Executive Agency 2022–2026.

**Data Availability Statement:** All data used in this study are provided in the Supplementary Materials.

**Acknowledgments:** This research was performed in the Centre of Excellence in Informatics and ICT under Grant No. BG16RFPR002-1.014-0018-C01, funded by the Research, Innovation and Digitalization for Smart Transformation Programme, EU 2025–2029. AI-assisted tools were used for English language editing and to improve the quality of the figures and tables.

**Conflicts of Interest:** The author declares no conflicts of interest.

## References

1. Sicherer, S.H.; Sampson, H.A. Food allergy: Epidemiology, pathogenesis, diagnosis, and treatment. *J. Allergy Clin. Immunol.* **2014**, *133*, 291–307. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Gupta, R.S.; Springston, E.E.; Warrier, M.R.; Smith, B.; Kumar, R.; Pongratic, J.; Holl, J.L. The prevalence, severity, and distribution of childhood food allergy in the United States. *Pediatrics* **2011**, *128*, e9–e17. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Nwaru, B.I.; Hickstein, L.; Panesar, S.S.; Muraro, A.; Werfel, T.; Cardona, V.; Dubois, A.E.J.; Halken, S.; Hoffmann-Sommergruber, K.; Poulsen, L.K.; et al. Prevalence of common food allergies in Europe: A systematic review and meta-analysis. *Allergy* **2014**, *69*, 992–1007. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Savage, J.H.; Matsui, E.C.; Skripak, J.M.; Wood, R.A. The natural history of peanut allergy: Extending our knowledge beyond childhood. *J. Allergy Clin. Immunol.* **2007**, *120*, 717–719. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Muraro, A.; Roberts, G.; Worm, M.; Bilò, M.B.; Brockow, K.; Fernández Rivas, M.; Santos, A.F.; Zolkipli, Z.Q.; Bellou, A.; Beyer, K.; et al. Anaphylaxis: Guidelines from the European Academy of Allergy and Clinical Immunology. *Allergy* **2014**, *69*, 1026–1045. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Cummings, A.J.; Knibb, R.C.; King, R.M.; Lucas, J.S. The psychosocial impact of food allergy and food hypersensitivity in children, adolescents and their families. *Allergy* **2010**, *65*, 933–945. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Palladino, C.; Breiteneder, H. Peanut allergens. *Mol. Immunol.* **2018**, *100*, 58–70. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Koppelman, S.J.; Vlooswijk, R.A.; Knippels, L.M.; Hessing, M.; Knol, E.F.; van Reijssen, F.C.; Bruijnzeel-Koomen, C.A. Quantification of major peanut allergens Ara h 1 and Ara h 2 in the peanut varieties. *Allergy* **2001**, *56*, 132–137. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Beyer, K.; Morrow, E.; Li, X.-M.; Bardina, L.; Bannon, G.A.; Burks, A.W.; Sampson, H.A. Effects of cooking methods on peanut allergenicity. *J. Allergy Clin. Immunol.* **2001**, *107*, 1077–1081. [\[CrossRef\]](#) [\[PubMed\]](#)
10. Hemmings, O.; Du Toit, G.; Radulovic, S.; Lack, G.; Santos, A.F. Ara h 2 is the dominant peanut allergen despite similarities with Ara h 6. *J. Allergy Clin. Immunol.* **2020**, *146*, 621–630. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Sampson, H.A.; Aceves, S.; Bock, S.A.; James, J.; Jones, S.; Lang, D.; Nadeau, K.; Nowak-Wegrzyn, A.; Oppenheimer, J.; Perry, T.T.; et al. Food allergy: A practice parameter update—2014. *J. Allergy Clin. Immunol.* **2014**, *134*, 1016–1025. [\[CrossRef\]](#) [\[PubMed\]](#)
12. PALISADE Group of Clinical Investigators. AR101 oral immunotherapy for peanut allergy. *N. Engl. J. Med.* **2018**, *379*, 1991–2001. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Fleischer, D.M.; Greenhawt, M.; Sussman, G.; Bégin, P.; Nowak-Wegrzyn, A.; Petroni, D.; Beyer, K.; Brown-Whitehorn, T.; Hebert, J.; Hourihane, J.O.; et al. Effect of Epicutaneous Immunotherapy vs Placebo on Reaction to Peanut Protein Ingestion Among Children With Peanut Allergy: The PEPITES Randomized Clinical Trial. *JAMA* **2019**, *321*, 946–955. [\[CrossRef\]](#)
14. Vickery, B.P.; Scurlock, A.M.; Kulis, M.; Steele, P.H.; Kamilaris, J.; Berglund, J.P.; Burk, C.; Hiegel, A.; Carlisle, S.; Christie, L.; et al. Sustained unresponsiveness to peanut in subjects who have completed peanut oral immunotherapy. *J. Allergy Clin. Immunol.* **2014**, *133*, 468–475. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Howell, W.M.; Holgate, S.T. HLA genetics and allergic disease. *Thorax* **1995**, *50*, 815–818. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Howell, W.M.; Turner, S.J.; Hourihane, J.O.; Dean, T.P.; Warner, J.O. HLA class II DRB1, DQB1 and DPB1 genotypic associations with peanut allergy: Evidence from a family-based and case-control study. *Clin. Exp. Allergy* **1998**, *28*, 156–162. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Hong, X.; Hao, K.; Ladd-Acosta, C.; Hansen, K.D.; Tsai, H.-J.; Liu, X.; Xu, X.; Thornton, T.A.; Caruso, D.; Keet, C.A.; et al. Genome-wide association study identifies peanut allergy-specific loci and evidence of epigenetic mediation in US children. *Nat. Commun.* **2015**, *6*, 6304. [\[CrossRef\]](#) [\[PubMed\]](#)
18. Hemler, J.A.; Phillips, E.J.; Mallal, S.A.; Kendall, P.L. The evolving story of human leukocyte antigen and the immunogenetics of peanut allergy. *Ann. Allergy Asthma Immunol.* **2015**, *115*, 471–476. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Prickett, S.R.; Voskamp, A.L.; Dacumos-Hill, A.; Symons, K.; Rolland, J.M.; O’HEhir, R.E. Ara h 2 peptides containing dominant CD4+ T-cell epitopes: Candidates for a peanut allergy therapeutic. *J. Allergy Clin. Immunol.* **2011**, *127*, 608–615.e5. [\[CrossRef\]](#) [\[PubMed\]](#)

20. Pascal, M.; Masilamani, M.; Lieberman, J.; Sampson, H.A. Identification of Ara h 2 T-cell epitopes for use in a T cell-peptide vaccine. *J. Allergy Clin. Immunol.* **2011**, *127*, AB31. [\[CrossRef\]](#)
21. Renand, A.; Farrington, M.; Whalen, E.; Wambre, E.; Bajzik, V.; Chinthrajah, S.; Nadeau, K.C.; Kwok, W.W. Heterogeneity of Ara h component-specific CD4 T cell responses in peanut-allergic subjects. *Front. Immunol.* **2018**, *9*, 1408. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Berin, M.C. Targeting type 2 immunity and the future of food allergy treatment. *J. Exp. Med.* **2023**, *220*, e20221104. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Møiniche, M.; Corneliussen, J.; Johansen, K.H.; Ruiz-Carrasco, A.; Paulsen, C.; Li, Y.; Barra, C.; Bangaru, S.; Fernández-Quintero, M.L.; Bartko, E.; et al. Mapping allergen B- and T-cell epitopes: Technological advances and their role in precision allergy therapy. *Allergy* **2026**, 1–18. [\[CrossRef\]](#)
24. Kanchan, K.; Shankar, G.; Huffaker, M.F.; Bahnson, H.T.; Chinthrajah, R.S.; Sanda, S.; Manohar, M.; Ling, H.; Paschall, J.E.; Du Toit, G.; et al. HLA-associated outcomes in peanut oral immunotherapy trials identify mechanistic and clinical determinants of therapeutic success. *Front. Immunol.* **2022**, *13*, 941839. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Dimitrov, I.; Bangov, I.; Flower, D.R.; Doytchinova, I. AllerTOP v.2—A server for in silico prediction of allergens. *J. Mol. Model.* **2014**, *20*, 2278. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Dimitrov, I.; Naneva, L.; Bangov, I.; Doytchinova, I. AllergenFP: Allergenicity prediction by descriptor fingerprints. *Bioinformatics* **2014**, *30*, 846–851. [\[PubMed\]](#)
27. Dimitrov, I.; Naneva, L.; Bangov, I.; Doytchinova, I. Allergenicity prediction by artificial neural networks. *J. Chemom.* **2014**, *28*, 282–286. [\[CrossRef\]](#)
28. Dimitrov, I.; Doytchinova, I. Associations between main food allergens and HLA-DR/DQ polymorphism. *Int. Arch. Allergy Immunol.* **2016**, *169*, 33–39. [\[CrossRef\]](#) [\[PubMed\]](#)
29. Doytchinova, I.; Atanasova, M.; Sotirov, S.; Dimitrov, I. In silico identification of peanut peptides suitable for allergy immunotherapy in HLA-DRB1\*03:01-restricted patients. *Pharmaceuticals* **2024**, *17*, 1097. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Gomez, R.A.; Hou, J.; Gersuk, V.H.; Chow, I.; Farrington, M.L.; Robinson, D.; Kwok, W.W. Ara h 2105–124-specific TH2A cells drive peanut allergy in DRB1\*15:01 individuals: A detailed epitope analysis. *Clin. Exp. Allergy* **2025**, *55*, 820–833. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)
32. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [\[CrossRef\]](#)
33. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [\[CrossRef\]](#)
34. Vita, R.; Blazeska, N.; Marrama, D.; Members, I.C.T.; Shackelford, D.; Zalman, L.; Foos, G.; Zarebski, L.; Chan, K.; Reardon, B.; et al. The Immune Epitope Database (IEDB): 2024 update. *Nucleic Acids Res.* **2025**, *53*, D436–D443. [\[PubMed\]](#)
35. Manavalan, B.; Basith, S.; Lee, G. Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2. *Brief. Bioinform.* **2022**, *23*, bbab412. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Basith, S.; Manavalan, B.; Hwan Shin, T.; Lee, G. Machine intelligence in peptide therapeutics: A next-generation tool for rapid disease screening. *Med. Res. Rev.* **2020**, *40*, 1276–1314. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Doytchinova, I.; Atanasova, M.; Fernandez, A.; Moreno, F.J.; Koning, F.; Dimitrov, I. Modeling peptide–protein interactions by a logo-based method: Application in peptide–HLA binding predictions. *Molecules* **2024**, *29*, 284. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Doytchinova, I.; Sotirov, S.; Dimitrov, I. Molecular insights into tumor immunogenicity. *Curr. Issues Mol. Biol.* **2025**, *47*, 641. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Doytchinova, I.; Dimitrov, I.; Atanasova, M.; Mihaylova, N.M.; Tchorbanov, A. AI-driven identification of candidate peptides for immunotherapy in non-obese diabetic mice: An in silico study. *AI* **2026**, *7*, 140. [\[CrossRef\]](#)
40. Schneider, T.D.; Stephens, R.M. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **1990**, *18*, 6097–6100. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. Peptide quantitative structure–activity relationships, a multivariate approach. *J. Med. Chem.* **1987**, *30*, 1126–1135. [\[CrossRef\]](#) [\[PubMed\]](#)
42. OpenAI. *ChatGPT, Version 5.2; Large Language Model*; OpenAI: San Francisco, CA, USA, 2026.
43. Altmann, A.; Toloşi, L.; Sander, O.; Lengauer, T. Permutation importance: A corrected feature importance measure. *Bioinformatics* **2010**, *26*, 1340–1347. [\[CrossRef\]](#) [\[PubMed\]](#)
44. Scholz, E.M.; Marcilla, M.; Daura, X.; Arribas-Layton, D.; James, E.A.; Alvarez, I. Human leukocyte antigen (HLA)-DRB1\*15:01 and HLA-DRB5\*01:01 present complementary peptide repertoires. *Front. Immunol.* **2017**, *8*, 984. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Kaabinejadian, S.; Barra, C.; Alvarez, B.; Yari, H.; Hildebrand, W.H.; Nielsen, M. Accurate MHC motif deconvolution of immunopeptidomics data reveals a significant contribution of DRB3, DRB4 and DRB5 to the total DR immunopeptidome. *Front. Immunol.* **2022**, *13*, 835454. [\[CrossRef\]](#) [\[PubMed\]](#)

46. UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53*, D609–D617. [[CrossRef](#)] [[PubMed](#)]
47. Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.E.; Wilkins, M.R.; Appel, R.D.; Bairoch, A. Protein identification and analysis tools on the ExPASy server. In *The Proteomics Protocols Handbook*; Walker, J.M., Ed.; Humana Press: Totowa, NJ, USA, 2005; pp. 571–607.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.