Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Deepitope: Prediction of HLA-independent T-cell epitopes mediated by MHC class II using a convolutional neural network

Raphael Trevizani[a,*], Fábio Lima Custódio[b]

[a] *FIOCRUZ - Fundação Oswaldo Cruz, Eusébio, CE, Brazil*
[b] *LNCC - Laboratório Nacional de Computação Científica, Petrópolis, RJ, Brazil*

ARTICLE INFO

Keywords:
Epitope prediction
CD8 +
citoxicity
MHC-I
deep learning

ABSTRACT

Computational linear T-cell epitope prediction tools allow cost and labor reduction in downstream *in vitro* testing, but the quality of currently available methods is compromised by the scarcity of experimental data and extensive HLA polymorphism. However, it is possible to improve prediction quality by forgoing HLA-dependency that allows treating all immunogenic sequences as a single group. This reduces the problem to a much simpler two-classes classification of determining whether a peptide is immunogenic or not. Here, we use a deep convolutional neural network capable of predicting linear T-cell epitope regions in primary structures trained using all peptides deposited in the IEDB website. We also investigate the possibility of using peptides derived from known human proteins as non-immunogenic counterexamples. We compared our model with a state-of-the-art tool and analyze the benefits of using larger databases. Our results corroborate the usefulness of HLA-free methods for practical applications that require the identification of immunogenic sequences. Deepitope is an open source project that can be found at https://github.com/raphaeltrevizani/deepitope.

1. Introduction

A major problem with biopharmaceuticals is their inherent potential of triggering unwanted immune responses that lead to the production of anti-drug antibodies that may cause adverse side effects and alter drug pharmacokinetics, ultimately hindering the treatment [1–7]. One pathway that triggers the immune response starts with peptides derived from proteolytically processed proteins that are presented to T-cells by membrane proteins located on the surface of the major histocompatibility complex class II (MHC-II) of antigen-presenting cells.

Due to the role T-cells play in driving humoral immunity, understanding their underlying mechanism is particularly relevant, and computational tools are valuable assets that help mapping T-cell binders [8,8–12]. These tools have been successfully employed a number of times [13–16] in practical applications such as the engineering of protein-based diagnostics, biopharmaceuticals, vaccines [17–20], and safer bio drugs [21,22]

One of the greatest challenges of predicting MHC class II molecules relates to their inherently diverse nature, which makes it impractical to quantify the interactions with each existing MHC variant. They are encoded in 3 extremely polymorphic Human Leukocyte Antigen loci (HLA-DR, DP, DQ [23]) with 9182 HLA Class II alleles as of May 2022 [https://www.ebi.ac.uk/ipd/imgt/hla/stats.html].

To overcome the problem posed by the number of HLAs, the concept of “pan” predictors was developed and relied on epitope promiscuity to generalize the properties of MHC variants [11,12]. The results reported by pan-methods such as NetMHCIIpan [11] proved the practical advantages of expanding the existing experimental data to represent the variability of most MHC interactions in humans [24].

To further increase generalization, a recent work presented by Dhanda and collaborators shows it is possible to learn some aspects of the immunogenicity encoded in the primary structure without the complications of MHC allelic variation with an HLA-independent model [25]. This HLA-independent approach takes advantage of the high degree of epitope promiscuity, compensates insufficient data on MHC molecules, and accounts for immunogenicity beyond MHC binding [25]. In addition to that, forgoing HLA-specificity has been suggested to boost prediction [25], which is made possible by the pervasiveness of epitope promiscuity [24].

In this work, we investigate the HLA-independent hypothesis by supplying experimentally determined T-cell epitopes available on the Immune Epitope Database and Analysis Resource website (IEDB, [26]) to a deep convolutional neural network and trained it to discriminate between immunogenic and non-immunogenic peptides, irrespectively of their associated HLAs.

A recent work has successfully applied a Convolutional Neural Network (CNN) to learn the preferences of MHC-I/peptide binding [27]. The

[*] Corresponding author.
E-mail address: raphael@raphaeltrevizani.com (R. Trevizani).

https://doi.org/10.1016/j.ailsci.2022.100038
Received 26 May 2022; Accepted 26 May 2022
Available online 29 May 2022

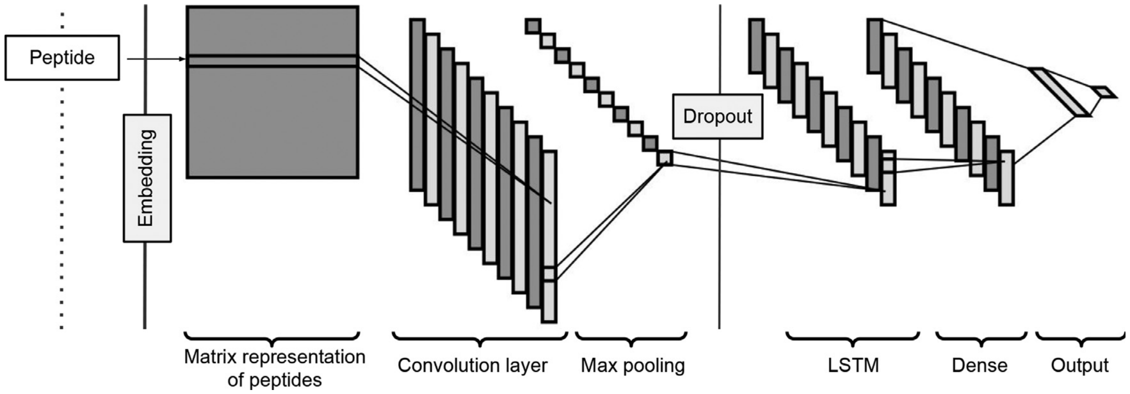

Fig. 1. Convolutional neural network model.

method proposed here relies solely on the sequence of curated epitopes, thus allowing the expansion of the database to all known immunogenic peptides. We explored different HLA-independent data sets of immunogenic peptides along with the possibility of using non-autoimmune, human-derived peptides as non-immunogenic. We compare our results to those of CD4episcore, a state-of-the-art tool that predicts CD4 immunogenicity in human populations [25].

2. Methods

2.1. Model

We supplied sequences of MHC class II-restricted T-cell epitopes to a CNN-LSTM (Long-Short Term Memory) model we called Deepitope and trained it to differentiate between immunogenic and non-immunogenic.

Deepitope was implemented in Python v3.8 using Tensorflow v2.0. The model was trained using a 1-D CNN-LSTM architecture that uses a Convolutional Neural Network (CNN) for feature extraction on the input sequence followed by an LSTM network for class prediction (Fig. 1). A similar model has been shown to work well for text-based cases [28]. First, the encoded sequence is passed to an embedding layer that turns positive integers into dense vectors of fixed size. Each embedded sequence is passed to the convolution layer, and the result is pooled to sample the most prominent feature extracted. A dropout layer is applied to avoid overfitting. A similar model without the LSTM layer could be used, but the LSTM layer captures long-term dependencies features of preceding and following tokens, enabling the perception of the long-term relationship between distant aminoacids[29]. The resulting forward and backward outputs are combined and passed to a fully-connected dense layer using a ReLU activation function followed by a layer with a single neuron using a sigmoid activation function.

2.2. Data set

Peptides are represented as a string of one-letter code for each amino acid and read as positive (immunogenic) or negative (non-immunogenic). A collection of peptides was obtained from various sources:

1. **T+ set**: Sequences of epitopes curated by the IEDB team that have been experimentally associated with a T cell response mediated by MHC class II receptors. These sequences were obtained by running an empty search on the main page of the IEDB website using the following criteria: Epitope: Linear; Host: Humans; Assay: T Cell Assays, Positives Assays Only; MHC Restriction: MHC Class II.
2. **T- set**: This set contains the sequences of peptides that demonstrably do not trigger a T cell response. This set was obtained by repeating the T+ set search without the Positive Assays Only box checked, meaning all peptides are exported. The peptides from the T+ set were then removed from the output, resulting in what we called the T- set.
3. **Promiscuous T+**: As of this writing, the csv file output by the IEDB search does not contain the HLAs associated with each epitope, but it is possible to use the IEDB epitope ID to search for this information. We automated a script to search and exclude the epitopes with unknown associated HLAs and those associated with fewer than two HLAs, thus ensuring a set of verified promiscuous epitopes. As this process resulted in a small number of peptides, we added the immunogenic peptides used by Dhanda [25].
4. **Human-derived peptides**: A set of peptides excised from human proteins found on the PDB website (https://www.rcsb.org/).

A total of 82 660 human proteins with an average sequence size of 243 residues were obtained from the PDB. Thus, the amount of p peptides of size n that can be extracted from each sequence of length l is $p = l - s + 1$, which results in more than 320 billion possible peptides of sizes 9 to 25. Compared to any of the positive test sets described above, such a disproportionate amount begets new complications if left unchecked. The number of examples in the positive and negative classes should be balanced so as not to bias the training. However, the result of selecting a subset of only thousands of peptides may not be representative of the larger set. To circumvent these shortcomings, we devised a method to dynamically extract a set of non-redundant peptides from the human proteins based on the configuration of the respective positive set. For each peptide of the positive set, a random peptide of the same length was extracted from the human proteins, which keeps both sets perfectly balanced and guarantees the algorithm focuses exclusively on the amino acid distribution rather than using peptide length as a discerning feature. To certify the sets generated had peptides with similar profiles and would not bias the training, we ran 1000 simulations of Deepitope training against the T+ set. For each simulation, a new human set was created and tested to determine the difference from the expected frequency of amino acids in humans [30].

2.3. Baseline comparison

We compared Deepitope with CD4episcore, a CD4 T cell immunogenicity prediction tool available at IEDB (http://tools.iedb.org/CD4episcore). Three different sets were used to compare with CD4episcore (Sets 1–3) and one was used to assess the effect of training on human-derived peptides and testing on T- (Set 4) as summarized in Table 1. Although Deepitope can handle peptides of any size, they were restricted to 15, 16 and 17-mers for both training and testing because CD4episcore requires the peptides to be at least 15 residues long, and the number of IEDB epitopes diminishes considerably after 17-mers. For all four sets tested, Deepitope was trained using k=5 cross-validation (Fig. 2B). Each test set was submitted to the CD4

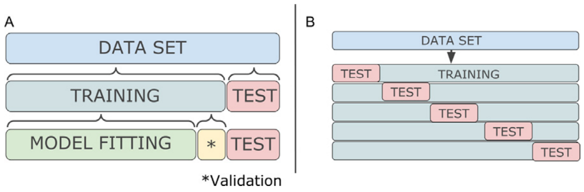

Fig. 2. A: Training, model fitting, validation and test sets. For a particular set, both positive and negative data sets are split between training and test set. During training, 90% of the data is used for model fitting and 10% of the data is used for validation, ie, an unbiased verification while tuning the model. If the loss on the validation partition is not improved for more than 5 consecutive epochs, the training is stopped and the model with the best loss is used for testing. hyperparameters. After training is complete, it is evaluated using the test set. B: Cross-validation. A 5-fold cross-validation is performed to avoid model overfitting and selection bias, where 80% of the data set is used for training and the 20% left are retained for testing. The process is repeated 5 times with each of the 20% subset used only once for testing.

immunogenicity tool using the Immunogenicity Prediction method with the default maximum immunogenicity score threshold = 50.

We also evaluated the test set using the widely used NetMHCIIpan [11]. NetMHCIIpan requires the input of a given size and specific MHC alleles, so we used 15-mers and the 7 alleles reported as optimal for general predictions [24]. All parameters were left as default. NetMHCIIpan outputs the peptides with their respective affinities and marks those that best interact with each HLA protein as Strong or Weak Binders (SB or WB, default cutoff). We parsed NetMHCIIpan predictions and considered immunogenic the peptides tagged as ’strong’ or ’weak binders’.

2.4. Evaluation

The predictions obtained were evaluated using the following metrics: true positive (TP), true negative (TN), false positive (FP), false negative (FN). The ratio of positives among those classified as such denotes the precision of the classifier (positive predictive value, PPV=TP/(TP + FP)), and, equivalently, the negative predictive value (NPV=TN/(TN + FN)) is the proportion of negatives correctly classified as such. Specificity (true negative rate, TNR=TN/(TN + FP)) is the rate of actual negatives correctly classified. Recall measures the proportion of positives correctly classified as such (true positive rate, TPR=TP/(TP + FN)). Accuracy is the proportion of correct predictions among all predictions (TP + TN)/(P + N). The receiver operating characteristic (ROC) curve plots the TPR against the false positive rate (FP/(FP + TN)) for every possible decision rule cutoff between 0 and 1 for a model. It is common to measure the performance of a prediction method by the area under the ROC curve (AUC), where 1 represents perfect a separation of the two classes while 0.5 is a random classifier. All metrics were calculated with the aid of the Sklearn library of the Python language.

The amino acid distribution observed for each of the 1000 human sets was compared to the expected frequency of amino acids in humans [30] using the chi-squared test. A one-way ANOVA was used to reveal the statistical significance of the differences observed for each metric.

2.5. In silico experiments

The number of units in both LSTM and dense layers were systematically varied to optimize performance. Experiments were run with the number of units ranging from 10 to 90, using 10 as step [10, 20, 30..., 90], and from 100 to 2000 using 100 as step [100, 200,..., 2000]. Tests were run using 5-fold cross-validation and the average AUC was used for comparison.

All redundant sequences were removed to avoid training examples contaminating the test set so no peptide used for training was used for testing and there was only one of each peptide in the training set. However, no attempt to remove sequences based on some amino acid similarity criterion was made (eg: BLOSUM, Blast, etc). Using sequence similarity matrices to consider different amino acids as equals might group into one category residues that are similar in a different context, preventing the method from properly learning the immunogenic signature. For instance, although aspartate and glutamate are regarded as the same amino acid for numerous practical purposes due to their biophysical similarities, they have different values in many positions of TEPITOPE matrices [31], suggesting they are not recognized in the same manner by the same MHC-II molecule.

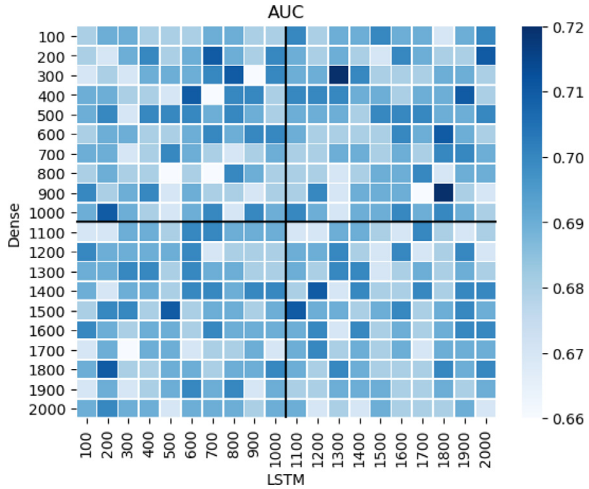

Fig. 3. AUC with different sizes of the LSTM and dense layers. LSTM: number units in the LSTM layer; Dense: number of units of the dense layer.

2.6. Comparison with other algorithms

Deep learning methods are increasingly common due to their high performance. To test if they are indeed superior to simpler methods in this problem, we compared Deepitope to four other algorithms: random forest, decision tree, gradient boost, and support vector machines using only 15-mers. In all tests, the residues of 15-mers were one-hot encoded and supplied to the methods implemented on Python v3.10.4 using Scikit v1.0.2-1.

The parameters were left as close to default as possible. The random forest classifier was trained using 100 trees and Gini impurity was used as a criterion for the optimal split of each node. No maximum depth or maximum number of leaf nodes were set. For the decision tree, at each iteration, the best split according to the Gini impurity was performed. No maximum depth was set, at least two samples were required to split an internal node, and at least one sample was required to be a leaf node. The gradient boost algorithm was applied with a learning rate of 0.1, 100 boosting stages, and optimized for the log loss function. No maximum depth was set and the quality of the split was measured by the mean squared error. The support vector machine used the radial basis function as kernel and a tolerance t = 0.001 for stopping criterion. In all cases, 90% of the data was used for training and 10% for testing.

3. Results

3.1. In silico experiments

We experimented Deepitope with different hyperparameters by varying the number of units of the LSTM and dense layers. In general, increasing the number of units of the dense layer above 1700 and decreasing the number of units of the LSTM below 400 has proven unfavorable. The best combinations were those of 900/1800 and 300/1300 for dense layers/LSTM (Fig. 3). We proceeded to use 300/1300 for the number of

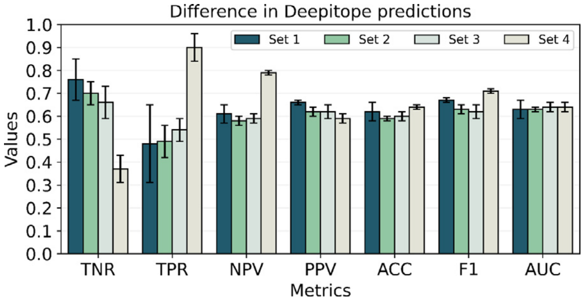

Fig. 4. Result of Deepitope predictions for different data sets (Table 1).

Table 1
Sets used to train Deepitope and compare with CD4episcore.

Set	Positive	Negative	Size*
1	T+ (15,16,17-mers)	T- (15,16,17-mers)	9385
2	T+ (15,16,17-mers)	Human	9385
3	Promiscuous T+ (15,16,17-mers)	Human	3330
4	T+ (15,16,17-mers)	Human (training) T- (testing)	9385

*Number of examples in each negative and positive sets

units of the dense/LSTM layers in all subsequent tests because it is less computationally expensive and yields a similar result.

3.2. Prediction using different data sets

Four different data sets were used in this work (Table 1). The results of the predictions using sets 1–3 show no significant difference for any of the metrics tested (Fig. 4). The high degree of similitude of TPR and PPV when sets 1 and 2 are compared to set 3 indicates the results of T+ with promiscuous T+ are statistically indistinguishable. This suggests the pattern learned by the algorithm for demonstrably promiscuous epitopes (promiscuous T+) is also found in the larger collection of epitopes (T+).

Similarly, TPR and NPV are statistically similar when sets 2 and 3 are compared to set 1, which means the use of human-derived peptides instead of T- gives statistically similar results. The possibility of replacing T- with human-derived peptides implies it is not strictly necessary to adhere to bench-tested peptides to have reasonably useful predictions, since human-derived peptides work as a suitable negative counterpart to the positive set composed of immunogenic T+ peptides. This may prove advantageous in future works due to the abundance of peptides that may be obtained from human proteins when compared to the much smaller amount of T- available.

The similar statistical results of tests 1 and 2 do not mean human-derived peptides work as a model for the T- peptides contained in the IEDB database. The results obtained with Set 4 show that training on a set of human-derived peptides and testing on T- has a similar net result on accuracy and AUC. However, a discrepancy on TPR and TNR can be verified upon closer examination, indicating a model with low specificity. The model finds most peptides to be immunogenic, resulting in a high sensitivity due to the high number of true positives, but it wrongly classifies many T- peptides as immunogenic. This results in a high number of true positives and a low number of true negatives, suggesting the underlying signature of T- peptides is perceived as closer to T+ than to that of human-derived peptides. The tests using Sets 1 and 2 show it is equally feasible to use T- or human-derived peptides as negative examples against T+, but training on humans to test on T- means they do not precisely model the T- peptides contained in the IEDB database. Nevertheless, both sets can be separately used as a counterpart to T+ in binary classification problems.

Table 2
Average Deepitope predictions for 1000 runs of set 2*.

Metric	Avg.	St.dev	Metric	Avg.	St.dev
TNR	0.66	± 0.06	F1	0.61	± 0.03
TPR	0.51	± 0.06	ACC	0.59	± 0.01
NPV	0.58	± 0.01	AUC	0.62	± 0.02
PPV	0.60	± 0.02			

*Each individual run was performed with a different human set.

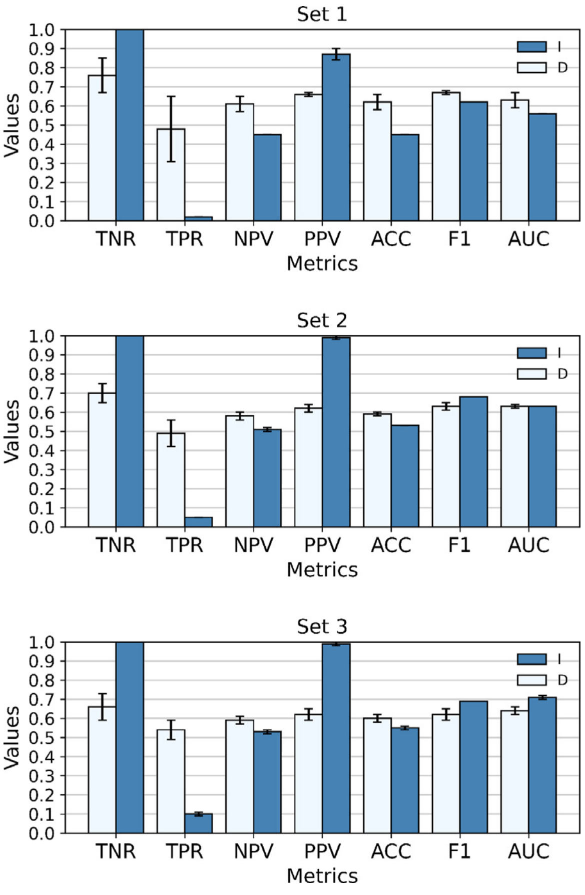

Fig. 5. D = Deepitope, I = IEDB CD4episcore; Negative examples: T+ (sets 1 and 2), promiscuous T+ (set 3), Positive examples: T- (set 1), human-derived peptides (set 3) as detailed in Table 1.

3.3. Variability of human-derived peptides

We ran 1000 simulations using the T+ set against 1000 different versions of the human set. There is virtually no difference among the results, as the highest standard deviation obtained for any of the performance metrics is 0.06. The standard deviation for AUC and accuracy were 0.02 and 0.01, respectively. Additionally, none of the human sets differ from the distribution of amino acids expected in humans, and they all have high p-values, meaning the null hypothesis cannot be rejected (average p-value = 0.9999, std: 6e-10, Table 2).

3.4. Baseline comparison

Deepitope was trained using a 5-fold cross-validation (Fig. 2). Each test set generated by the 5-fold cross-validation was submitted to the CD4episcore tool available at the IEDB website and the average results were compared to those obtained by Deepitope (Fig. 5).

Due to the different methods employed and data sets used, the results presented by Deepitope and CD4episcore diverge considerably. Over-

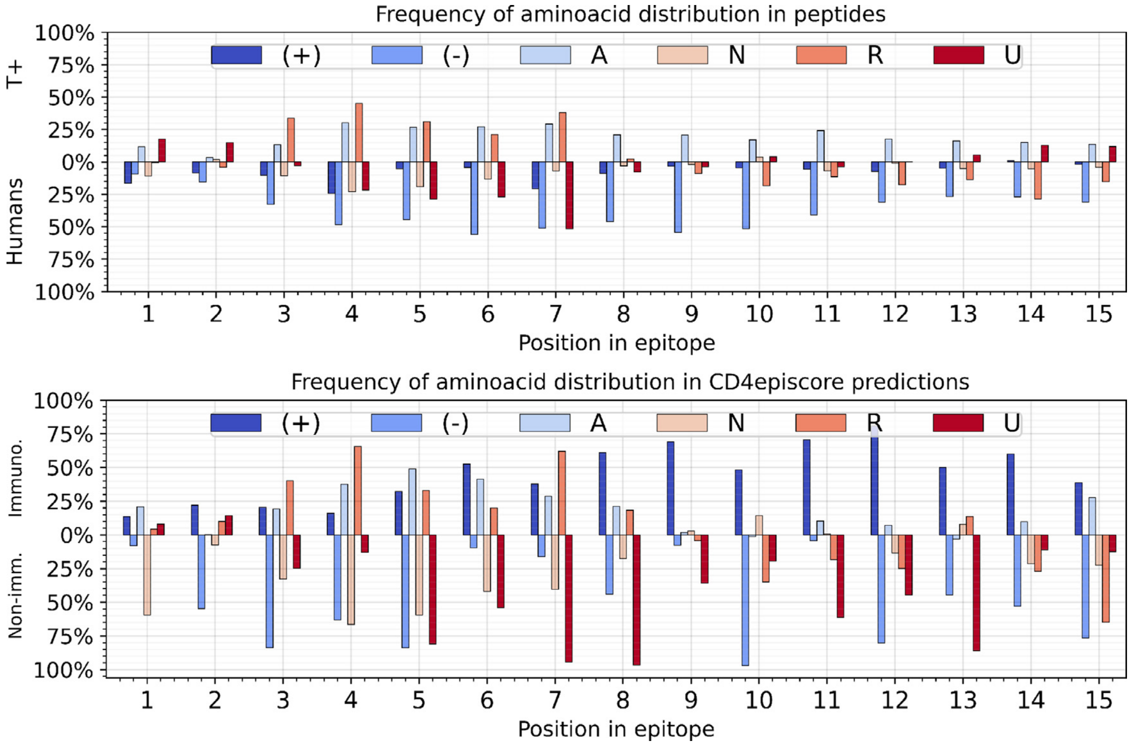

Fig. 6. Top: Relative distribution of amino acids in each position of all T+ 15-mers compared to the frequency of amino acids in human-derived peptides. Bottom: Relative distribution of amino acids in peptides predicted as immunogenic and non-immunogenic by CD4episcore. Aliphatic: Met, Ile, Leu, Ala, Val; **(+)**-charged: His, Arg, Lys; **(-)**-charged: Asp, Glu; **N**eutral: Gln, Thr, Asn, Ser, Cys; a**R**omatic: Trp, Tyr, Phe; **U**ncategorized: Pro, Gly.

all, CD4episcore selects only a few peptides as immunogenic, rendering a direct comparison somewhat incongruous. As CD4episcore classifies almost all peptides as non-immunogenic, it favors negative results (ie, non-immunogenic sequences, eg: T- and humans), which causes the rate of true negatives to be high at the cost of low true positives. This leads to a very good performance in all the metrics that do not penalize false negatives, results in very high values for metrics that reward identifying negatives as such (e.g.: TNR, PPV) and very low values when false negatives are taken into account (TPR, NPV), culminating in high AUC values at the expense of low accuracy.

One might assume that CD4episcore identifies a lower number of immunogenic peptides because it was trained using a restricted, possibly less diverse dataset. However, the analysis of the distribution of amino acids per position for CD4episcore predictions indicates it closely matches that of T+/human-derived peptides (Fig. 6). The top panel on Fig. 6 shows how frequent each amino acid category is in T+ compared to humans for each position. The bottom panel shows the same peptides as classified by CD4episcore. Firstly, it is evident that the values are more extreme for the CD4episcore predictions, which can be at least partially attributed to the reduced number of peptides predicted as immunogenic. Nevertheless, CD4episcore accurately matches the amino acid pattern for all categories in all positions, except for one: positively charged amino acids are more common in human-derived peptides and yet CD4episcore chiefly predicts peptides rich in positively-charged amino acids as immunogenic (Fig. 6 shows the (+) bars going in opposite directions).

Individualizing the frequency of each positively-charged amino acids (H, R, K) reveals the inconsistencies between experimental data (Fig. 7, top) and CD4episcore predictions (Fig. 7, bottom). Histidines in positions 2 to 5 are approximately 50% more common in human-derived peptides, something that is not detected by CD4episcore predictions. Likewise, lysine is very common in all positions of human-derived peptides, but CD4episcore frequently classifies peptides rich in lysine as immunogenic.

Table 3
Comparison with NetMHCpanII.

Metric	NetMHCIIpan	Deepitope
ACC	0.59	0.60
AUC	0.59	0.63
TNR	0.73	0.72
TPR	0.45	0.48

The results of the comparison with NetMHCIIpan are shown in Table 3. The results of NetMHCIIpan show an AUC of 0.59, which is on par with those of Deepitope. It is also noteworthy the best result of NetMHCIIpan is the true negative rate (0.73), which shows that NetMHCIIpan is able to predict human-derived peptides despite being trained using non-binding peptides. This reinforces the fact that human-derived peptides can be used to model negative binders. NetMHCIIpan does not directly model T-cell activation, but the affinity between the peptide and specific MHCs. While many factors are needed for a peptide to be an epitope, binding to the MHC is a key factor, which is why it is common to infer high-binding sequences as epitopes to specific MHCs. Therefore, we acknowledge this is technically a different context from NetMHCIIpan's direct output. Nevertheless, given the pervasiveness of NetMHCIIpan and how frequently it is to infer immunogenicity from MHC binding, we think this is comparison is useful as it provides insight into the differences of modelling immunogeniciy by direct TCR activation compared to MHC binding affinity.

3.5. *Comparison with other algorithms*

We tested Deepitope against some well-known, traditional machine learning algorithms using 15-mers as a test case. Deepitope was retrained to work only with 15-mers and evaluated only on 15-mers, so its performance was not as good as the one obtained for all peptide sizes. Table 4 shows the comparison of the AUC values.

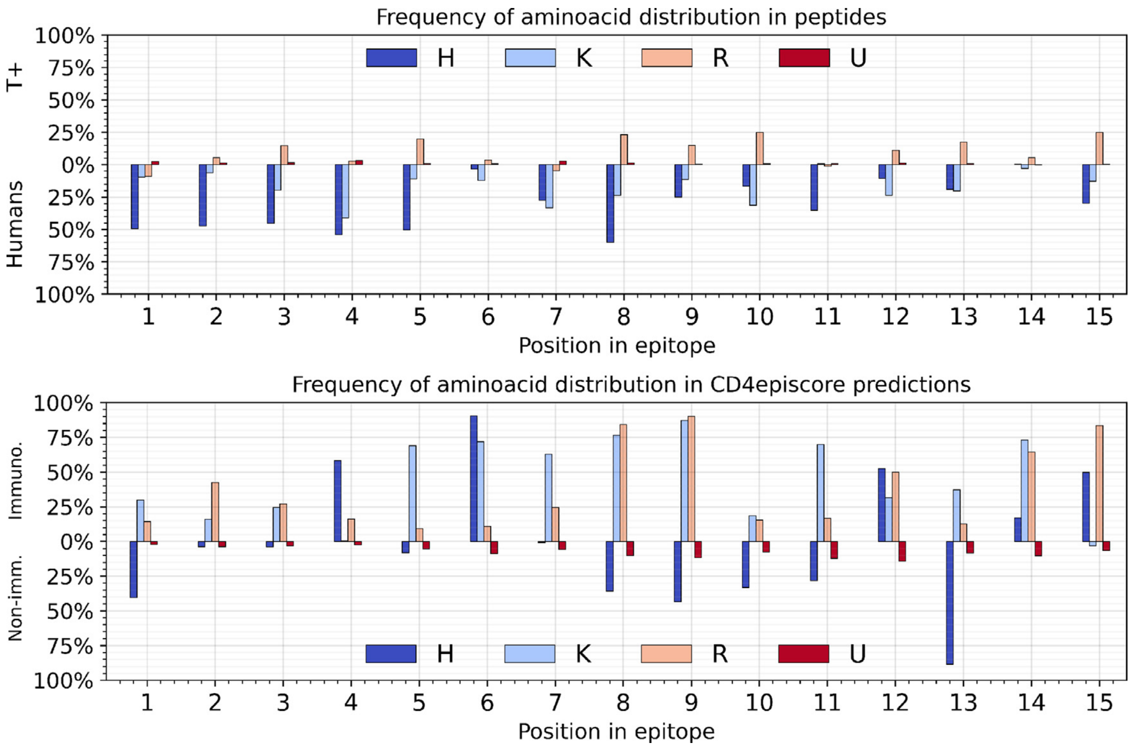

Fig. 7. Top: Relative distribution of positively-charged amino acids in each position of all T+ 15-mers compared to the frequency of amino acids in human-derived peptides. Bottom: Relative distribution of amino acids in peptides predicted as immunogenic and non-immunogenic by CD4episcore. **H**: His, **R**: Arg, **K**: Lys, **U**ncategorized: all other amino acids .

Table 4
Comparison of Deepitope and other machine learning for 15-mers.

Algorithm	AUC
RF	0.50
DT	0.50
SVM	0.50
GB	0.50
Dp	0.62

RF = random forest, DT = decision tree, SVM = support vector machines, GB = gradient boosting, Dp = Deepitope.

Because the results are no better than random guessing, we did not test them any further or attempted to examine how they learn the positional preferences of each residue in human-derived peptides or T+. Each algorithm could arguably be tweaked for better performance, but given the poor initial results, we think this was not worth investigating. The CNN-LSTM employed by Deepitope does not contain any complexities that deviate from a basic CNN-LSTM and achieved considerably superior results. This suggests the need for more complex algorithms that are able to capture higher-order interactions and, for each position, model the interplay of forces encoded in neighboring residues.

4. Discussion

This work aimed to show HLA-independent epitope mapping can aid the isolation of immunogenic sequences from an ensemble of peptides by taking advantage of features shared across MHC-II epitopes regardless of their corresponding HLAs. We proposed a deep learning-based method to differentiate immunogenic from non-immunogenic peptides because other machine learning methods have been successful in the past in this field and newer approaches based on deep learning are promising in protein science [32].

Ignoring HLA-specificity allows all known epitopes to be treated as a single, homogeneous group that is large enough to be useful in training. Thence, instead of predicting binding affinities, the problem is ultimately reduced to a straightforward two-classes classification problem where non-HLA-specific immunogenic signatures are identified. Therefore, a robust HLA-independent method can compensate for some aspects of the inadequately small data sets that currently supply the training of HLA-specific methods.

Several types of peptides could be used as immunogenic. The IEDB database contains peptides that have been experimentally shown to bind to the MHC-II clefts [26] and many methods use MHC data relying on the well-established fact that peptides that bind more strongly to the MHC cleft remain there longer and therefore have greater chances of being identified by T-cell receptors [33–36]. Using MHC data is also more common because there is substantially more data available on peptides/MHC complex than epitopes that cause T-cell activation. While this gives precise information about the interaction with the epitope, it requires costly, laborious bench tests resulting in an insufficient amount of data. Nonetheless, triggering the immune response also requires the peptide to be recognized by a T-cell, so we used epitopes identified as "T-cell assay" on the IEDB website. As we wanted to model T-cell activation, filtering epitopes that have been experimentally shown to bind to the T-cell receptor seemed a fitting choice, even though it resulted in a smaller data set.

While lack of information of the related HLAs grants only a crude classification of each peptide as immunogenic or not, HLA-independent models can be useful to detect some underlying pattern of what makes an epitope beyond the specificity of each HLA. This may help to better establish a set of fundamental signatures that ideally spans the human population as broadly as possible. What must be inquired is whether it is feasible to generalize patterns from a set of epitopes for which the associated HLA remains to be determined. Because not all epitopes in the IEDB have known associated HLAs, we ran a separate test using T+ epitopes that bind to more than one HLA. The results obtained using the different immunogenic sets are very similar and suggest T+ and promiscuous T+ work as virtually the same (Fig. 4). As promiscuous T+ shows no substantial difference from the full T+ set, it is possible to hypothesize the attainment of the same features of promiscuous epitopes using a set that is not demonstrably promiscuous.

There are also different possibilities for non-immunogenic peptides. As the model proposed here intends to recognize epitopes at the population level, a reasonable approach was to gather peptides from human proteins. The human peptides were extracted from the PDB website, as it allows all human FASTA files to be exported with ease. Interestingly, the intersection between IEDB T-cell epitopes and PDB human peptides coincides with the IEDB auto-immune epitopes (data not shown) and was discarded to simplify the model. The theoretical advantages of using the human-derived peptides are natural abundance, diversity, and null cost, but can they be shown to work as a replacement for bench-verified, non-immunogenic peptides in practice? Virtually no difference can be detected between the results of the positive metrics of set 1 to sets 2 and 3, suggesting the human peptides (sets 2 and 3) can adequately replace non-immunogenic peptides like those of set 1. Over a thousand human sets were generated in the course of this study and all proved consistent with the amino acid distribution in humans, while the standard deviation for any of the evaluation metrics did not vary more than 6% (Table 2). We have also shown that the different subsets of human-derived peptides always obeyed the amino acid distribution expected for humans (Table 2) and granted predictions by Deepitope that bore a remarkable similitude with those of the non-immunogenic data set (Fig. 4), thus hinting at the realistic use of human-derived peptides in lieu of experimentally verified non-immunogenic peptides.

The use of human-derived peptides as non-immunogenic is a leap that should be taken with care. Considering the immune system is trained by means of negative selection and the negative dataset is composed of curated epitopes that show T-cell activation in lab tests, they should in theory work as opposing data sets with different patterns. The question is whether they can be verifiably used for the aforesaid purpose. Here, we demonstrated the usefulness of peptides excised from human proteins as a positive set to oppose immunogenic peptides, which seemed adequate since the task at hand is to isolate peptides that are immunogenic to humans.

The comparison with CD4episcore predictions shows CD4episcore only classifies a few peptides of the data set as epitopes. Nevertheless, it shows a good recall (TPR) and AUC, and effectively captures the characteristics of most T+ epitopes. However, it classifies too many peptides rich in positively-charged amino acids as immunogenic. Taken together, these results illustrate the effects of smaller, unrepresentative input data that may effectively bias the predictions and reinforces the need for expanding and diversifying the database. The use of broader databases in combination with deep learning methods may open the way for new insights into several elemental aspects of immunogenic peptides.

5. Conclusion

The efficiency of epitope mapping tools depends on experimental data that are scarce due to the high polymorphism of HLA genes.

In this work, we trained a CNN-LSTM model to appraise peptides for their alikeness to immunogenic/non-immunogenic categories using MHC-II mediated T-cell epitopes. We propose HLA-independent linear T-cell epitope mapping tools are advantageous to find immunogenic peptides due to the existence of underlying non-HLA-specific immunogenic patterns. Therefore, HLA-specificity was dismissed and the model was trained solely to predict whether a query peptide is best classified as immunogenic or non-immunogenic.

We show peptides excised from human proteins are an adequate non-immunogenic negative set for the immunogenic T+, even if they do not model the negative T- set. It remains to be determined if the T- set is still too restricted to span all non-immunogenic sequences due to the incompleteness of the IEDB dataset.

Finally, it is shown that supplying T-cell epitopes allows learning a pattern similar to that presented by a general set of T-cell epitopes that have been demonstrated to be truly promiscuous. Forgoing HLA-specificity limits the scope of applications as it gives less information, but it improves the quality of predictions and may be useful to explore the fundamentals of immunogenic peptides.

Deepitope is an open source project that can be found at https://github.com/raphaeltrevizani/deepitope.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the Santos Dumont Supercomputer and Fiocruz for the use of their computational facilities and NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] Eckardt K-U, Casadevall N. Pure red-cell aplasia due to anti-erythropoietin antibodies. Nephrol Dialy Transplant 2003;18(5):865–9. doi:10.1093/ndt/gfg182.
[2] Casadevall N, Nataf J, Viron B, Kolta A, Kiladjian J-J, Martin-Dupont P, et al. Pure red-cell aplasia and antierythropoietin antibodies in patients treated with recombinant erythropoietin. N top N Engl J Med 2002;346(7):469–75. doi:10.1056/nejmoa011931.
[3] Tatarewicz SM, Wei X, Gupta S, Masterman D, Swanson SJ, Moxness MS. Development of a maturing t-cell-mediated immune response in patients with idiopathic parkinson's disease receiving r-methugdnf via continuous intraputaminal infusion. J Clin Immunol 2007;27(6):620–7. doi:10.1007/s10875-007-9117-8.
[4] Jawa V, Hokom M, Hu Z, El-Abaadi N, Zhuang Y, Berger D, et al. Assessment of immunogenicity of romiplostim in clinical studies with ITP subjects. Ann Hematol 2010;89(S1):75–85. doi:10.1007/s00277-010-0908-2.
[5] Shankar G, Pendley C, Stein KE. A risk-based bioanalytical strategy for the assessment of antibody immune responses against biological drugs. Nat Biotechnol 2007;25(5):555–61. doi:10.1038/nbt1303.
[6] Li J. Thrombocytopenia caused by the development of antibodies to thrombopoietin. Blood 2001;98(12):3241–8. doi:10.1182/blood.v98.12.3241.
[7] Baert F, Noman M, Vermeire S, Assche GV, Haens GD, Carbonez A, et al. Influence of immunogenicity on the long-term efficacy of infliximab in crohn's disease. N top N Engl J Med 2003;348(7):601–8. doi:10.1056/nejmoa020888.
[8] Nielsen M, Lund O. Nn-align. an artificial neural network-based alignment algorithm for mhc class ii peptide binding prediction. BMC Bioinformatics 2009;10:296. doi:10.1186/1471-2105-10-296.
[9] Karosiene E, Rasmussen M, Blicher T, Lund O, Buus S, Nielsen M. Netmhciipan-3.0, a common pan-specific mhc class ii prediction method including all three human mhc class ii isotypes, hla-dr, hla-dp and hla-dq. Immunogenetics 2013;65(10):711724. doi:10.1007/s00251-013-0720-y.
[10] Reynisson B, Alvarez B, Paul S, Peters B, Nielsen M. Netmhcpan-4.1 and netmhciipan-4.0: improved predictions of mhc antigen presentation by concurrent motif deconvolution and integration of ms mhc eluted ligand data. Nucleic Acids Res 2020;48(W1):W449–54. doi:10.1093/nar/gkaa379.
[11] Andreatta M, Karosiene E, Rasmussen M, Stryhn A, Buus S, Nielsen M. Accurate pan-specific prediction of peptide-mhc class ii binding affinity with improved binding core identification. Immunogenetics 2015;67:641–50. doi:10.1007/s00251-015-0873-y.
[12] Zhang L, Chen Y, Wong H-S, Zhou S, Mamitsuka H, Zhu S. Tepitopepan: extending tepitope for peptide binding prediction covering over 700 hla-dr molecules. PLoS ONE 2012;7:e30483. doi:10.1371/journal.pone.0030483.
[13] De Groot AS, Knopp PM, Martin W. De-immunization of therapeutic proteins by t-cell epitope modification. Dev Biol (Basel) 2005;122:171–94.
[14] Salvat RS, Verma D, Parker AS, Kirsch JR, Brooks SA, Bailey-Kellogg C, et al. Computationally optimized deimmunization libraries yield highly mutated enzymes with low immunogenicity and enhanced activity. Proc Natl Acad Sci USA 2017;114:E5085–93. doi:10.1073/pnas.1621233114.
[15] Parker AS, Choi Y, Griswold KE, Bailey-Kellogg C. Structure-guided deimmunization of therapeutic proteins. J Comput Biol 2013;20:152–65. doi:10.1089/cmb.2012.0251.
[16] Osipovitch DC, Parker AS, Makokha CD, Desrosiers J, Kett WC, Moise L, et al. Design and analysis of immune-evading enzymes for adept therapy. Prot Eng Des Select 2012;25:613–23. doi:10.1093/protein/gzs044.
[17] Groot ASD, Bosma A, Chinai N, Frost J, Jesdale BM, Gonzalez MA, Martin W, Saint-Aubin C. From genome to vaccine: in silico predictions, ex vivo verification. Vaccine 2001;19(31):4385–95. doi:10.1016/s0264-410x(01)00145-1.
[18] Ahlers JD, Belyakov IM, Thomas EK, Berzofsky JA. High-affinity t helper epitope induces complementary helper and apc polarization, increased ctl, and protection against viral infection. J Clin Invest 2001;108:1677–85. doi:10.1172/JCI13463.

[19] De Groot AS. Immunome-derived vaccines. Expert Opin Biol Ther 2004;4:767–72. doi:10.1517/14712598.4.6.767.

[20] De Groot AS, Ardito M, McClaine EM, Moise L, Martin WD. Immunoinformatic comparison of t-cell epitopes contained in novel swine-origin influenza a (h1n1) virus with epitopes in 2008–2009 conventional influenza vaccine. Vaccine 2009;27:5740–7. doi:10.1016/j.vaccine.2009.07.040.

[21] Inaba H, Martin W, De Groot AS, Qin S, De Groot LJ. Thyrotropin receptor epitopes and their relation to histocompatibility leukocyte antigen-dr molecules in graves' disease. J Clin Endocrinol Metab 2006;91:2286–94. doi:10.1210/jc.2005-2537.

[22] Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V. Evaluation of mhc-ii peptide binding prediction servers: applications for vaccine research. BMC Bioinformatics 2008;9 Suppl 12:S22. doi:10.1186/1471-2105-9-S12-S22.

[23] Traherne JA. Human mhc architecture and evolution: implications for disease association studies. Int J Immunogenet 2008;35:179–92. doi:10.1111/j.1744-313X.2008.00765.x.

[24] Paul S, Lindestam Arlehamn CS, Scriba TJ, Dillon MBC, Oseroff C, Hinz D, et al. Development and validation of a broad scheme for prediction of hla class ii restricted t cell epitopes. J Immunol Methods 2015;422:28–34. doi:10.1016/j.jim.2015.03.022.

[25] Dhanda S, Karosiene E, Edwards L, Grifoni A, Paul S, Andreatta M, et al. Predicting hla cd4 immunogenicity in human populations. Front Immunol 2018;9:1369.

[26] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. The immune epitope database 2.0. Nucleic Acids Res 2010;38:D854–62. doi:10.1093/nar/gkp1004.

[27] Hu Y, Wang Z, Hu H, Wan F, Chen L, Xiong Y, et al. ACME: Pan-specific peptide–MHC class i binding prediction through attention-based deep neural networks. Bioinformatics 2019. doi:10.1093/bioinformatics/btz427.

[28] Shen Q, Wang Z, Sun Y. Sentiment analysis of movie reviews based on cnn-blstm. In: Shi Z, Goertzel B, Feng J, editors. IFIP TC12 ICIS. IFIP Advances in Information and Communication Technology, vol. 510. Springer; 2017. p. 164–71. ISBN 978-3-319-68121-4. http://dblp.uni-trier.de/db/conf/ifip12/icis2017.html#ShenWS17

[29] Bepler T, Berger B. Learning the protein language: Evolution, structure, and function. Cell Systems 2021.

[30] Darby NJ, Creighton TE. Protein structure / N.J. Darby and T.E. Creighton. IRL Press at Oxford University Press Oxford ; New York; 1993. 019963310

[31] Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, et al. Generation of tissue-specific and promiscuous hla ligand databases using dna microarrays and virtual hla class ii matrices. Nat Biotechnol 1999;17:555–61. doi:10.1038/9858.

[32] Torrisi M, Pollastri G, Le Q. Deep learning methods in protein structure prediction. Comput Struct Biotechnol J 2020;18:1301–10. doi:10.1016/j.csbj.2019.12.011.

[33] Iwai LK, Yoshida M, Sidney J, Shikanai-Yasuda MA, Goldberg AC, Juliano MA, et al. In silico prediction of peptides binding to multiple hla-dr molecules accurately identifies immunodominant epitopes from gp43 of paracoccidioides brasiliensis frequently recognized in primary peripheral blood mononuclear cell responses from sensitized individuals. Mol Med (Cambridge, Mass) 2003;9:209–19.

[34] Mustafa AS, Shaban FA. Propred analysis and experimental evaluation of promiscuous t-cell epitopes of three major secreted antigens of mycobacterium tuberculosis. Tuberculosis (Edinb) 2006;86:115–24. doi:10.1016/j.tube.2005.05.001.

[35] Al-Attiyah R, Mustafa AS. Computer-assisted prediction of hla-dr binding and experimental analysis for human promiscuous th1-cell peptides in the 24 kda secreted lipoprotein (lppx) of mycobacterium tuberculosis. Scand J Immunol 2004;59:16–24.

[36] Lazarski CA, Chaves FA, Jenks SA, Wu S, Richards KA, Weaver JM, Sant AJ. The kinetic stability of mhc class ii:peptide complexes is a key parameter that dictates immunodominance. Immunity 2005;23:29–40. doi:10.1016/j.immuni.2005.05.009.