ImmunoInformatics 20 (2025) 100061

Contents lists available at ScienceDirect

ImmunoInformatics

journal homepage: www.journals.elsevier.com/immunoinformatics

Research Article

IApred: A versatile open-source tool for predicting protein antigenicity across diverse pathogens

Sebastian Miles *, Gonzalo Menafra, Andrés Iriarte, Jose Alejandro Chabalgoity

Unidad Académica de Desarrollo Biotecnológico, Instituto de Higiene, Facultad de Medicina, Universidad de la República, Montevideo, Uruguay

ARTICLE INFO

Keywords:
Antigenicity
Immunology
Bioinformatics
Immunoinformatics
Predictor
Vaccines

ABSTRACT

Accurate prediction of protein antigenicity is crucial for vaccine development, diagnostic test design, and therapeutic protein engineering. However, existing tools face limitations in accessibility, computational efficiency, and pathogen diversity. Here, we present IApred, an open-source intrinsic antigenicity predictor that addresses these challenges. IApred employs a Support Vector Machine (SVM) model trained on a comprehensive dataset of 918 high-antigenicity proteins from diverse pathogens, including Gram-positive and Gram-negative bacteria, viruses, fungi, protozoa, and helminths. The model incorporates features derived from physicochemical properties, *E*-descriptors, amino acid dimers and small linear motifs (SLiMs) to predict the probability of a protein eliciting a humoral immune response. In external validation, IApred demonstrated superior balanced performance (ROC AUC = 0.761, sensitivity = 0.702, specificity = 0.706) compared to existing tools (VaxiJen 2.0, VaxiJen 3.0 and ANTIGENpro), while maintaining high computational efficiency (approximately 1000 sequences per minute). IApred's host-and-pathogen-agnostic nature and integration capability into bioinformatic pipelines makes it versatile for diverse applications. A web-based version of the software is available at https://smilesinformatics.com/iapred, while the software and training code are freely available on GitHub (https://github.com/sebamiles/IAPred) and Zenodo (https://doi.org/10.5281/zenodo.14578279)

Introduction

The challenge of antigenicity prediction

Immunogenicity, as a concept, refers to the ability to induce a humoral and/or cell-mediated immune response, while antigenicity is the ability to specifically combine with the final products of the immune response (ie., secreted antibodies and/or surface T-cell receptors) [1]. In humoral responses, antigens are typically recognized by both T-cells and B-cells. Peptides derived from the antigen are presented on MHC II molecules to T-cells, which then collaborate with B-cells that recognize specific epitopes of the antigen through the B-cell receptor (BCR). B-cell epitopes exhibit greater flexibility, as they can be composed of not only proteins but also polysaccharides or lipids. T-independent antibody responses can also occur for antigens with repetitive epitopes (e.g., polysaccharides); however, these responses are less studied and do not develop immunological memory, limiting their utility in biotherapies [2].

Understanding antigenicity is crucial for numerous biotechnological applications. For instance, identifying antigenic proteins is essential for vaccine and diagnostic test development, while non-antigenic proteins are desirable for therapeutic applications. Additionally, decoding the antigenicity of proteins can provide insights into host-pathogen interactions and inform therapeutic strategies. However, predicting antigenicity is a complex challenge. It depends not only on the physicochemical properties of the antigen [3] but also on its three-dimensional structure [4], post-translational modifications [5], the accessibility of epitopes, and the nature of antigen-antibody interactions [6]. Furthermore, the diversity of immune responses across individuals and species adds another layer of complexity, with differences in MHC haplotypes being a key factor influencing the success of T-dependent antibody responses [7].

Limitations of existing tools

At the time of writing, there are four reliable and available antigenicity predictors: VaxiJen 2.0 [8] and 3.0 [9] ANTIGENpro [3], and APRANK [10]. VaxiJen is the most widely used antigenicity predictor,

* Corresponding author at: Unidad Académica de Desarrollo Biotecnológico, Instituto de Higiene, 1st floor., Av. Dr. Alfredo Navarro 3051, 16000, Montevideo, Uruguay.
E-mail address: smiles@higiene.edu.uy (S. Miles).

https://doi.org/10.1016/j.immuno.2025.100061
Received 30 December 2024; Received in revised form 30 September 2025; Accepted 7 October 2025
Available online 9 October 2025

and its latest version (VaxiJen 3.0) uses machine learning to predict the antigenicity of a protein using previously described *E*-descriptors [11], which are five numerical values derived from principal component analysis (PCA) of 237 physicochemical properties. Each protein is transformed into a uniform vector, which are then used to train three machine learning algorithms (SVM, XGBoost and RSM-*knn*). A consensus of these models classifies proteins as immunogen or non-immunogen.

ANTIGENpro, on the other hand, uses a two-stage architecture for antigenicity prediction. In the first stage, a Naive Bayes algorithm incorporates forty primary classifiers, while the second stage employs an SVM classifier based on probability estimates derived from protein microarray data and literature-curated antigens [3].

APRANK, though not a direct antigenicity predictor, ranks top antigen candidates within a proteome. Its prediction model is more complex, incorporating T- and B-cell epitope predictions, three-dimensional structural analysis, protein-specific site analysis, and sequence similarity between pathogen and host proteins [10]. While APRANK offers higher prediction accuracy, its complexity comes with increased computational costs, and the need of programming knowledge to run it, making it not universally available. Notably, no comprehensive comparative study has been conducted to evaluate the performance between these three tools.

Several other antigenicity predictors have also been developed, including NERVE [12], Vaxign [13], Jenner-predict [14], Vacceed [15], iVAX [16], Protectome analysis [17] and VacSol [18]. These tools typically use a series of filters, such as subcellular localization, adhesion probability, topology, homology with human proteins, and presence of T and B cell epitopes, to identify antigenic proteins. However, most of these tools are no longer available or rarely used.

It is worth noting that while a newer version of Vaxign (Vaxign2) has been recently described [19], it is reported to have unresolved bugs in the algorithm. Similarly, Vaxi-DL, a web-based deep learning server for identifying potential vaccine candidates, has been recently presented [20], but we were unable to obtain results from the web server. Vaxign-DL, a deep learning-based method for vaccine design, has also been proposed in a preprint [21], but the software is not currently accessible.

The need for IApred

While existing antigenicity predictors are accurate, they have significant limitations. VaxiJen and ANTIGENpro are not open-source and lack APIs, restricting their large-scale use and integration into automated bioinformatics pipelines. Additionally, ANTIGENpro only allows predictions for one sequence at a time, with results sent via email, making it impractical for analysing large datasets. APRANK, although open-source, demands substantial computational resources and technical expertise. It also relies on third-party software (e.g., BepiPred 1.0 [22] and NetMHCIIPan 2.0 [23]), which may require licenses and further complicate its use. Moreover, APRANK is designed to rank top antigenic proteins within a proteome rather than predict the antigenicity of individual proteins. Its reliance on T-cell epitope predictions, which are specific to certain MHC molecules, also limits its applicability to hosts with those MHC haplotypes.

Another critical limitation of VaxiJen and ANTIGENpro is their reliance on training datasets dominated by bacterial antigens, which may not generalize well to other pathogen types, such as viruses, fungi, protozoa, or helminths.

Development of IApred

To address these limitations, we aimed to develop a simple yet accurate, open-source antigenicity predictor with low computational requirements. IApred is designed to predict the intrinsic antigenicity of proteins derived from a wide range of infectious disease pathogens. It is pathogen-and-host agnostic, trained on a diverse dataset of manually curated antigens from Gram-positive and Gram-negative bacteria, viruses, fungi, protozoa, and helminths.

IApred employs a Support Vector Machine (SVM) model trained on a comprehensive set of features, including physicochemical parameters, *E*-descriptors, amino acids dimers and small linear amino acid motifs (SLiMs). The model assigns a numeric value representing the intrinsic antigenicity of a protein sequence, defined as the probability of the protein inducing a specific humoral immune response in the context of natural infection or vaccination. This definition focuses solely on the protein's amino acid sequence, excluding factors such as post-translational modifications, physical availability, pathogen origin, or host immune variability.

Materials and methods

High-antigenicity proteins

To train the machine learning model, we used a set of high- and low-antigenicity proteins datasets. For the high-antigenicity set, we primarily relied on Serological Proteome Analysis (SERPA) reports, and vaccine candidates. In SERPA, a pathogen protein fraction is separated in a two-dimensional electrophoresis gel (2DEG), followed by Western blot analysis using sera from infected or vaccinated mammalian hosts. Each developed spot is then identified by mass spectrometry, considering those identified proteins as antigens. For our set of high-antigenicity proteins, we included SERPA antigens identified using sera from infected or vaccinated humans, pigs, dogs, goats, rabbits, guinea pigs, cattle, ewe and sheep.

The proteins in the dataset were classified by their organisms of origin and pathogen group, corresponding to virus, Gram-negative and Gram-positive bacteria, fungi, protozoans and helminths. To reduce overfitting, we removed sequences with >90 % of identity inter- and intra-specie, retaining the longest sequence. This process resulted in a SERPA dataset comprising 553 proteins from 25 different organisms.

In addition to the SERPA set, we included the datasets used to train other antigenicity prediction tools, such as the one reported by Dimitrov et al. [9], composed of 315 bacteria antigens used to train VaxiJen 3.0. For viral proteins, we generated a non-redundant dataset from VaxiJen ViralDB comprising 95 proteins corresponding to 22 viruses. The final training set combine these sources, totalling 918 proteins (Table 1).

For external evaluation, we utilize the Protegen database [24], a curated repository of protective antigens against infectious and non-infectious diseases, including 1371 antigens derived from 216 pathogens or host organisms (https://violinet.org/protegen/; last accessed November 2024). Given the diversity of organisms in this database, we created a reduced version of this dataset by randomly selecting 5 antigens from each infectious-disease related organism with at least 5 reported protective antigens and removed those with a 90 % similitude against any antigen in the training dataset. An exception was made for *Coccidioides immitis*, where only 4 proteins were available out of a total of 9 fungal proteins, and *Treponema pallidum*, where only 4 non-redundant proteins were obtained. The final reduced evaluation set consists of 218 proteins, including 110 Gram-negative, 40 Gram-positive, 9 fungal, 54 protozoan and 5 helminth proteins.

Low-antigenicity proteins

Given that we were unable to obtain an experimental dataset of low-antigenicity proteins comparable in size to the high-antigenicity dataset, we developed a workaround based on the assumption that most proteins have low-antigenicity. For each protein in the high-antigenicity SERPA set, we randomly selected a protein of comparable size (amino acid length ± 10 %) from the same organism, ensuring that the selected protein had <70 % sequence similarity with the antigen. This approach aimed to create a balance dataset of low-antigenicity proteins

However, this method was not feasible for most viruses due to their

Table 1
Description of the proteins used as antigens for the SVM training. The antigens used were classified into pathogens Classes. The column Source describes briefly how the antigens were identified.

Class	Organism	# Antigens	Source	Reference
Gram(-) *N* = 83	Actinobacillus pleuropneumoniae	7	SERPA using Pig sera	[33]
	Bordetella pertussis	50	SERPA using Human sera	[34]
	Chlamydia pneumoniae	26	SERPA using Human sera	[35]
Gram(+) *N* = 189	*Bacillus anthracis*	55	SERPA using Rabbit sera and immune guinea pig sera	[36]
	Corynebacterium pseudotuberculosis	16	SERPA using Goat sera	[37]
	Mycobacterium bovis	13	SERPA using Cattle sera	[38]
	Staphylococcus aureus	68	SERPA using Ewe sera	[39]
	Staphylococcus pseudintermedius	14	SERPA using Dog sera	[40]
	Tropheryma whipplei	23	SERPA using Human sera	[41]
Fungi *N* = 101	Candida albicans	51	SERPA using Human sera and reported antigenic proteins	[42]
	Aspergillus fumigatus	15	Reported antigens	[43]
	Cryptococcus deuterogattii	13	Reported antigens	[43]
	Cryptococcus neoformans	3	Reported antigens	[43]
	Coccidioides posadasii	6	Reported antigens	[43]
	Paracoccidioides brasiliensis	8	Reported antigens	[43]
	Talaromyces marneffei	5	Reported antigens	[43]
	Ajellomyces capsulatus	25	Immunoprecipitation followed by mass spectrometry	[44]
Protozoan *N* = 91	Plasmodium vivax	13	Antigens proteins discovered or validated using protein microarrays	[45]
	Trypanosoma cruzi	65	Previously reported antigens used to create a protein microarray	[46]
	Toxoplasma gondii	13	Reported Vaccine Candidates	[47]
Helminth *N* = 64	Ascaris suum	13	SERPA using Pig sera	[48]
	Echinococcus granulosus	17	SERPA using Mice sera	[49]
	Fasciola hepatica	10	SERPA using Sheep sera	[50]
	Schistosoma mansoni	11	Reported Vaccine Candidates	[51]
	Trichinella spiralis	13	SERPA using Human sera	[52]
Virus	Virus	95	Manually curated non-redundant virus vaccine candidates	[2]
VaxiJen	VaxiJen Antigens	315	Manually curated antigens used to train VaxiJen 3.0	[2]

small proteomes. To address this, we supplemented the dataset by randomly selecting 95 low-antigenicity proteins from all pathogens included in the SERPA set, matching number of viral antigens. Additionally, we incorporated the dataset of non-antigenic proteins used to train VaxiJen 3.0 [9].

Features used for training

To train the machine learning algorithm, we generated a Python (version 3.12.4) script, were we analysed 838 distinct features derived from protein sequences. A complete list of these features can be found in Table 2 and Supplementary Table 1. Using the ProteinAnalysis module from the Biopython's (version 1.83) ProtParam library [25], we calculated basic physical and chemical properties of each protein. These included molecular weight, isoelectric point, proportions of secondary structural elements (helices, sheets, and coils), aliphatic index, sequence entropy, charge distribution, hydrophobicity and hydrophobic moment. We also examined the ratio between polar and non-polar amino acids, as well as proline and cysteine content, and the proportion of amino acids by size, classifying them into tiny, small, and large residues.

E-descriptors, originally reported by Venkatarajan and Braun [11] and used in the VaxiJen 3 prediction algorithm, were included as features in the analysis to provide a mathematical representation of amino acid properties. For this, we represented each amino acid as a five-dimensional vector and analysed the generated vectors using PyTorch (version 2.7.1) [26]. From these vectors, we extracted both basic features (averages and sums) and advanced geometric properties. The advanced properties included sphericity, planarity, and linearity of the combined vectors. We also analysed changes along the protein sequence through vector analysis and examined local patterns by studying protein segments of 5, 10, and 15 amino acids. Position-weighted averages were also calculated to account for the possibility that certain protein regions contribute differentially to the antigenicity. Finally, we measure the complexity and diversity of *E*-descriptors using entropy-based calculations. At the end, 49 different features were obtained derived from *E*-descriptors

In addition to these features, we also included specific small linear amino acid patterns (SLiMs) that are known to have biological importance. We obtained a list of 353 well-studied pattern classes from the Eukaryotic Linear Motif database (http://elm.eu.org/; last access November 2024) [27]. Then, we calculated the frequency of each SLiM and every possible pair of amino acids in each protein sequence, to find indications of other important protein motifs.

All these features together provide a comprehensive mathematical description of each protein sequence, allowing our machine learning algorithm to learn patterns that might be related with the intrinsic antigenicity of a protein.

Table 2
Features initially taken into consideration for the machine learning training.

Feature type	# of features	Description
Physicochemical properties	7	IP, GRAVY, Aromaticity, Instability Index, Flexibility, Charge and Hydrophobicity
Secondary structure	3	Predicted fraction of a-helix, b-sheet and random coil
Amino acid size	3	Fraction of tiny, small and large residues
Aliphatic indez	1	Relative volumen occupied by aliphatic side chains
Sequence Enttropy	1	Measure of sequence complexity based on amino acid composition
Hydrophobic moment	1	Using Eisenberg scale and assuming a-helix with 100° angle
Charge Distribution	1	Standar deviation of charges along the sequence
Polar/Non-Polar Ratio	1	Ratio of polar to non-polar residues
Proline Content	1	Percentage of proline residues
Cysteine Content	1	Percentage of cysteine residues
Bigram transition	16	Frequency of transition between polar, non-polar, acidic and basic residues
E-descriptors	49	Features derived from vectors generated using E1-E5 descriptors
amino acid dimers	400	Frequency of each possible combination of two amino acids
SLiMs	353	Frequency of specific Short Linear Motifs

Machine learning algorithm

Scikit-learn (version 1.5.2) [28] python module was used to train the Support Vector Machine (SVM) model. Data Preprocessing, feature selection and hyperparameter tuning was performed before obtaining the final SVM model.

Data preparation and preprocessing

The protein sequences of our datasets (corresponding to the SERPA, non-redundant Virus and VaxiJen proteins) were first encoded into numerical vectors using the previously discussed features. To address potential scale differences between features, we normalized the feature values using StandardScaler. Constant features, which provide no discriminative power, were removed using VarianceThreshold to reduce dimensionality and improve model efficiency. Synthetic Minority Oversampling Technique (SMOTE) [29]), from imbalanced-learn (version 20.12.3) was employed to mitigate any potential effect of class imbalance in the dataset.

Feature selection

The best number of features to use was obtained by performing a feature selection step using the SelectKBest method with the f_classif scoring function. This function first ranks all features and then analyses the ROC-AUC value for an SVM model trained with each k value of the top ranked features. For this, we performed an 80/20 stratified split of the dataset, fitted the training pipeline on the 80 % subset, and computed ROC-AUC on the 20 % hold-out validation set for each k

Model training and hyperparameter tuning

Radial basis function (RBF) was selected as the best performing kernel. To optimize the model's performance, we conducted a 5-fold cross-validation grid search over a range of values for the regularization parameter C (from 0.001 to 1000) and the kernel coefficient gamma (from 1 to 1 × 10–7). The 5-fold cross-validation was performed over the training subset only (after the 80/20 split).

Model evaluation

The final model (kernel=RBF, $k = 529$, $C = 1$, gamma=0.001) was trained and evaluated, obtaining the Learning and Precision-Recall curves, and confusion matrix, using the 80/20 training and validating sets. The learning curve was obtained from 5-fold CV on the training subset at increasing training sizes, while the Precision-Recal curve and the confusion matrix are computed on the validation set. In addition, the correlation between features was analysed and the importance of each feature was calculated using the permutation importance method. We then performed a stratified 10-fold cross validation and two modifications of LOO—CV (Leave-One-Out Cross-Validation) defined as LOCO—CV (Leave-One-Class-Out Cross-Validation) and LOPO—CV (Leave-One-Pathogen-Out Cross-Validation). LOCO—CV consisted of training the model while leaving out one pathogen class (Gram+ or Gram- bacteria, fungi, protozoan, or helminth) and calculating the ROC-AUC score. LOPO—CV (not applied to viruses) leaves out one organism per run.

Internal and external performance evaluation

To assess the predictive performance of the model, we conducted both internal and external evaluations. For the internal evaluation, we calculated the intrinsic antigenicity scores for all proteins in the training dataset, including both antigens and non-antigens, as well as the entire proteome of each corresponding organism. We then compared the score distributions among groups. For the external evaluation, we utilized the reduced version of the Protegen dataset. The performance of IApred was benchmarked against three widely used antigenicity predictors—VaxiJen 2.0, VaxiJen 3.0, and ANTIGENpro—using this dataset. Its worth noting that, as VaxiJen 3.0 only has a model for bacteria preduction, we included two versions of this for the external evaluation, one with only the bacteria results, and one with all the pathogens. Finally, we compared the performance metrics of all tools to highlight the strengths and limitations of IApred in relation to existing methods.

Predictor and data availability

The prediction model was compiled into a python script, available to be downloaded from GitHub (https://github.com/sebamiles/IAPred), Zenodo (https://doi.org/10.5281/zenodo.14578279) and a google colab notebook was created to run single-sequence or multiple-sequences FASTA files prediction as an online predictor (https://colab.research.google.com/github/sebamiles/IAPred/blob/main/IApred-colab.ipynb). In addition, an easy-to-use web server was made available at https://smilesinformatics.com/iapred, with the limitation of only be able to process up to 1000 sequences at a time. The training, evaluating and prediction codes were also uploaded to GitHub, as well as the manually curated high-antigenicity SERPA set, the non-redundant virus set and the VaxiJen Sets.

Results and discussion

An antigen is a molecule capable of binding to a specific antibody or a major histocompatibility complex (MHC) [30]. In this context, we define the intrinsic antigenicity of a protein as the probability of an amino acid sequence being targeted by the immune system, regardless of other factors. This concept is highly relevant for a wide range of biotechnological applications, such as vaccine development or biotherapeutics.

Here, we present IApred, an open-source pathogen-and-host agnostic intrinsic-antigenicity predictor from an amino acid sequence. IApred can be used as a standalone tool or in a combination of other bioinformatic programs. Unlike other available predictors that classify proteins as either antigenic or non-antigenic, IApred provides a continuous IAscore, which reflects the reality that virtually any protein can be targeted by the immune system, albeit with variable effectiveness. We believe this approach is particularly valuable for diverse biotechnological applications. For example, in the development of biotherapeutics, low-antigenicity molecules are often desired to minimized unwanted immune responses; however, for diagnosis or identification of vaccine candidates, high-antigenicity proteins are preferred.

High- and low-antigenicity datasets

We assembled a dataset of 918 high-antigenicity proteins, compiled from previously reported antigenic proteins [9], as well as literature curated proteins (details provided in Table 1 and Supplementary Table 1). Creating a corresponding dataset of low-antigenicity proteins presented significant challenges, as we found no straightforward experimental method to identify these proteins. To address this limitation, we developed an alternative approach based on the premise that most proteins within an organism exhibit low-antigenicity. Thus, for each high-antigenicity protein in our dataset, we selected a random protein from the same organism's proteome with similar size and sequence identity lower than 70 %. While this approach enabled us to construct a balanced dataset, we are well aware that this approach might negatively affect the performance of the predictor, since some high-antigenicity proteins may inadvertently be included in the low-antigenicity dataset.

Model training

Once obtained the High- and Low-antigenicity datasets, we proceeded by selecting the features used to train the model. For this, we obtained a combination of physicochemical and amino acids compositional properties, as well as protein sequence specific features, and features derived from *E*-descriptors. Table 2 shows a description of all the evaluated features, with more detailed information found in Supplementary Table 2. Many of these features have been previously reported as determinants of antigenicity and specificity in immune responses, including protein sequence properties of composition, isoelectric point, change distribution, secondary structure and solvent accessibility [31]. Other features have been used in different antigenicity predictors, such as the Molecular Weight, GRAVY, aliphatic index [3] and *E*-descriptors [9]. The inclusion of SLiMs as an antigenic correlation is novel to IApred, supported by the idea that even though most SLiMs correspond to eukaryotic proteins, many pathogens mimic these motifs as an evolutionary advantage [32], and thus, some motifs could be over-or-underrepresented in antigenic proteins. Amino acid dimers frequency or sequence repetitiveness could correlate with evolutionary conserveness between pathogens and potentially creating stable structural motifs for antibody recognition. Abundance of cysteine or presence of cysteine-rich regions could derive into formation of disulfide bonds, masking potential epitopes. In the current SVM model, many of the analysed features describe low-antigenicity protein characteristics, helping with the discrimination process.

The SVM model was trained using the high- and low-antigenicity datasets, evaluating all 838 features in each sequence, using RFB as the optimal kernel (data not shown). Firstly, 47 features that showed constant or very-low variance were removed, which were primarily SLiMs. Then, we performed 5-fold stratified cross-validation on the training set to select the best parameters to use and evaluate its performance. In this sense, although Fig. 1A shows a plateau around 15–20 top features under fixed hyperparameters ($C = 1$, gamma='scale'), the fully tuned pipeline selected $k = 529$ as it consistently provided higher stability and predictive performance, with mid-ranked features contributing complementary information that was lost at smaller k values. Thus, 529 features were selected as the best performing k number as shown by the ROC AUC in Fig. 1A, where C and gamma values for the RFB kernel were optimized, selecting $C = 1$ and gamma=0.001 (Fig. 1B). Using these parameters the final model was trained, and the performance was evaluated by means of the Learning Curve (Fig. 1C) and the Precision-Recall Curve (Fig. 1D). Fig. 1E shows the confusion matrix composed of the true and predicted antigens and non-antigens, using the validations sets after the 80/20 split, considering a threshold score of 0 to discriminate between antigens and non-antigens. The learning curve shows only a slight improvement on performance with increasingly bigger training sets. Thus, since the ROC-AUC score (0.8017) and the precision-recall curve (AP= 0.79) show that the model exhibits a quite high accuracy, we determined that expanding the datasets was unnecessary as the model's accuracy was sufficient for its intended application as an antigenicity predictor. It is worth noting that alternative models, such as Random Forest and XGBoost, were also evaluated; however, they demonstrated lower discriminating power compared to the SVM model (data not shown).

Model evaluation

To further evaluate the performance of the obtained model, we performed a 10-fold cross validation, dividing all the training data into ten segments, training the model with 9 of those segments and evaluating the performance over the remaining segment, repeating the process for each leave-out segment. Fig. 2A shows the ROC curve for each iteration, as well as the mean ROC and standard deviation between runs. In addition, we performed a similar cross-validation, described as Leave-One-Class-Out (LOPO—CV), and Leave-One-Pathogen-Out (LOPO—CV) [21], where all proteins from one class or pathogen (referring to only SERPA antigens, always including the VaxiJen-antigen set and the

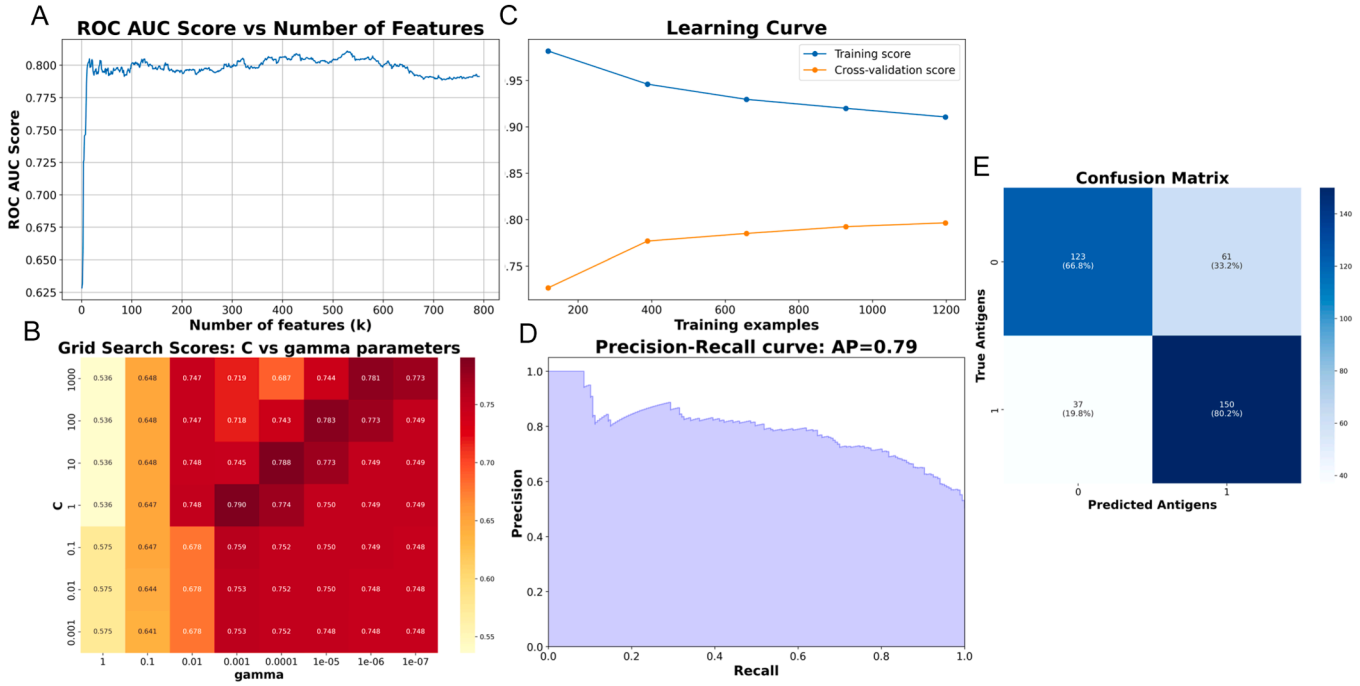


Fig. 1. Machine Learning Training and Optimization. An SVM machine learning algorithm was trained using the antigens and non-antigens datasets. (A) ROC-AUC score of the model using different numbers of features (k) under fixed hyperparameters ($C = 1$, gamma='scale'). (B) Optimization of C and gamma for $k = 529$, with values corresponding to ROC-AUC. (C) Learning Curve of the model using optimized parameters ($k = 529$, $C = 1$ and gamma=0.001); the training score shows the ROC-AUC for the training set, while the cross-validation score corresponds to the 5-fold cross-validation ROC-AUC, both at increasing dataset size. (D) Precision-Recall curve computed on the validation set. (E) Confusion matrix displaying the model's classification results based on the validation sets, presented in both absolute values and percentage.

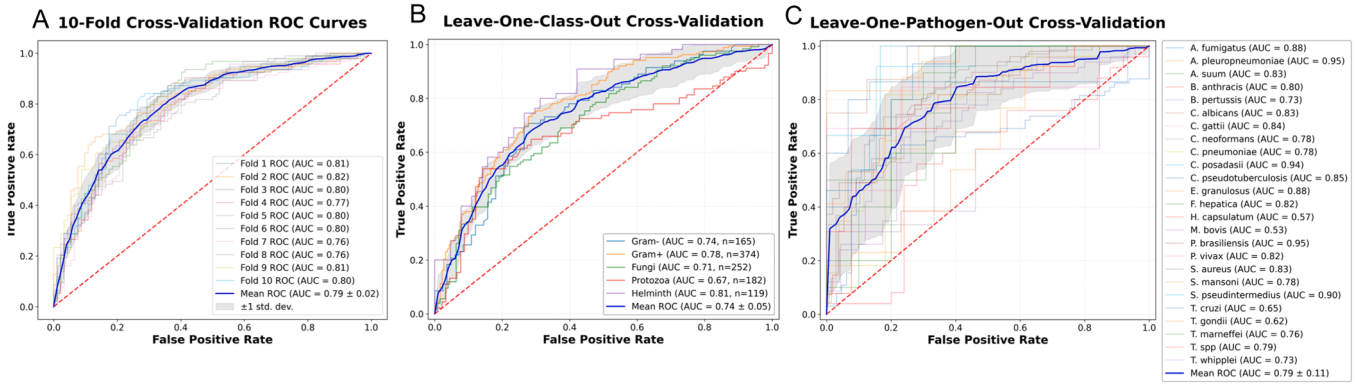


Fig. 2. IApred Cross-Validation. The performance of the IApred algorithm was evaluated using three cross-validation strategies, using $k = 529$, $C = 1$ and gamma=0.001. (A) 10-fold Cross-Validation, (B) Leave-One-Class-Out Cross-Validation (LOCO—CV), and (C) Leave-One-Pathogen-Out Cross-Validation (LOPO—CV).

non-redundant viral set due to low-antigenicity proteins limitations) were removed from the training, and then, the model was evaluated over the left-out set of antigen, resulting in the curves shown in Fig. 2B and 2C. These last results show that our model is capable of predicting the intrinsic antigenicity of proteins independently of the class, with a modest difference but positive prediction power depending on the pathogen.

Over the final model, feature correlation and feature contribution analysis were performed. Supplementary Figure 1 shows a correlation heatmap and a contribution histogram for all the selected features, as well as detailed information about each of the 529 selected feature. Most features are slightly positively correlated (correlation > 0), with only a small number of features exhibiting a high degree of correlation, either positive or negative (correlation $> |0.5|$). Of the selected features, most of them (410) have a positive impact on the classification process of protein as high-antigenicity. The frequency of large and small residues, the aliphatic index, sequence repetitiveness, isoelectric point and KN dimer are the top positive features, contributing each with at least 0.8 % to the prediction model. On the other hand, the motifs ELME000057, ELME000058, ELME000096, ELME000235, ELME000340, ELME000407, ELME000426, ELME000553 and ELME000561 are the top negative features, with a contribution of at least 0.3 % to the prediction model. It's worth noting that almost all features that negatively impact the IAscore (except the proportion of coils) are related to the frequency of SLiMs or amino acid dimers, which can reflect similarities to the host proteins, and a lack of response due to tolerance processes, although further investigation are needed to corroborate this hypothesis.

Internal evaluation

To evaluate the performance of the model to correctly classify proteins according to their intrinsic antigenicity, we performed an internal and external evaluation. As an internal evaluation, we firstly predicted the intrinsic antigenicity of every protein in the SERPA set (both antigens and non-antigens), as well as in the whole proteome of the corresponding organisms, and compared the antigenicity distribution. From the results exhibited in Table 3, we can observe that, in most cases, the manually curated antigens have a statistically higher IAscore compared to the corresponding proteome (Mann-Whitney-Wilcoxon $p < 0.05$), and that the low-antigenicity proteins have a similar distribution to the proteome. The observed antigenicity of the manually curated antigens and non-antigens datasets is expected, as they were used to train the model, but the relatively low-antigenicity of the rest of the proteome partially corroborate the accuracy of our model, as it has been reported that most protein have low-antigenicity.

External evaluation

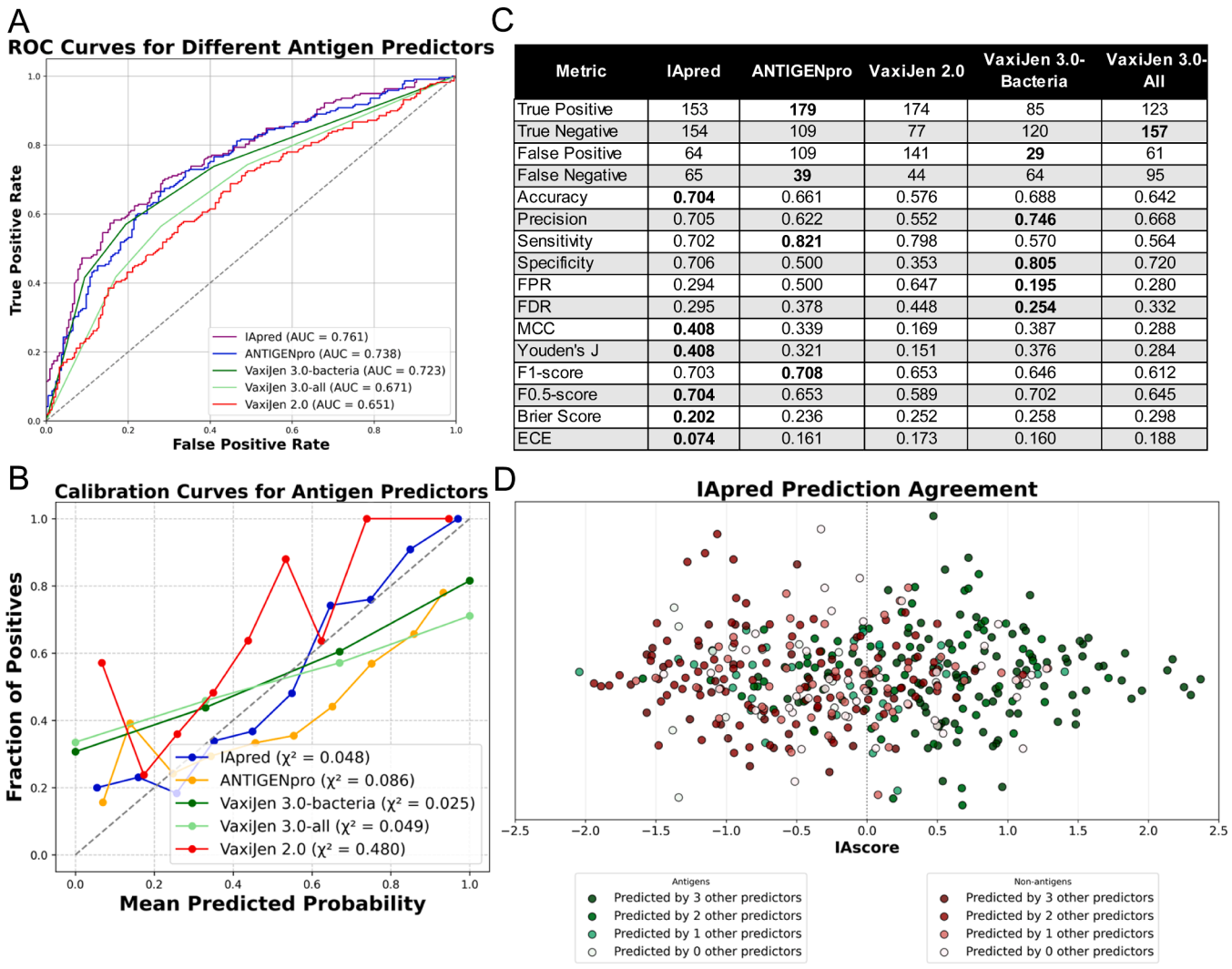
For the external evaluation of IApred, we utilized the reduced version of the Protegen dataset, along with a corresponding non-antigen dataset generated using the same methodology as described earlier. The antigenicity of both datasets was predicted using IApred, as well as three widely used tools: ANTIGENpro, VaxiJen 2.0, and VaxiJen 3.0. For Vaxijen 3.0, due to limitations of the software to evaluate different pathogen classes, we display the results as two different predictions, one using only the bacteria-derived proteins (named as Vaxijen 3.0-bacteria), and one including all pathogens proteins (named Vaxijen 3.0-all). We did not use APRANK in the external evaluation, as it is designed to rank top antigenic proteins within a proteome rather than predict the antigenicity of individual proteins; and its reliance on T-cell epitope predictions -which are specific to certain MHC molecules- also limits its applicability to hosts with those MHC haplotypes. To classify proteins as antigens or non-antigens, we applied the following thresholds: for IApred, proteins with positive scores were considered antigens; for ANTIGENpro, a threshold of 0.5 was used across all pathogen classes [3]; for VaxiJen 2.0, thresholds of 0.4 for bacteria and 0.5 for other pathogens were applied, running each prediction with the appropriate model based on the analysed pathogen [8]; and for VaxiJen 3.0, a numeric value (0–1) was assigned based on the consensus of three machine learning models, with each predictor contributing 0.33. A protein was classified as an antigen if at least two models agreed ($\geq$0.66).

The performance of the four predictors was compared using ROC curves (Fig. 3A), calibration plots (Fig. 3B), and various evaluation metrics (Fig. 3C). ANTIGENpro exhibited the highest sensitivity (0.821) but low specificity (0.507), indicating a tendency to overpredict antigens. Conversely, VaxiJen 3.0-bacteria, (when only the bacteria proteins are analysed), achieved the highest specificity (0.805) but low sensitivity (0.570), reflecting a cautious approach that misses many true antigens. As expected, if all pathogens' proteins are predicted using the VaxiJen 3.0 bacterial mode, the specificity (0.720) and sensitivity (0.564) show an overall worst performance. VaxiJen 2.0 showed high sensitivity (0.798) but very low specificity (0.353) meaning that it correctly predicts most antigens at the cost of a high percentage of false positives. In contrast IApred emerged as the most balanced predictor, with sensitivity (0.702) and specificity (0.706) closely matched, resulting in the highest accuracy (0.704). It also led in Matthews Correlation Coefficient (MCC = 0.408), a single summary score that combines true and false positives and negatives into a value between –1 and 1 to reflect overall prediction quality, and Youden's J ($J = 0.408$), which subtracts the false positive rate from the true positive rate to identify the optimal threshold for distinguishing antigens. Additionally, IApred achieved the lowest Brier score (0.202), which indicate more reliable probability between predicted probabilities and actual outcomes, and the smallest

Table 3
Distribution of predicted antigens across training datasets and proteomes.

Class	Organism	Antigens dataset					Non-Antigens dataset					Proteome dataset					Statistics
		#Prot	#Ag	#N-Ag	%Ag	IAscore (Mean ± std)	#Prot	#Ag	#N-Ag	%Ag	IAscore (Mean ± std)	#Prot	#Ag	#N-Ag	%Ag	IAscore (Mean ± std)	
Gram(-)	Actinobacillus pleuropneumoniae	7	7	0	100.0 %	1.20 ± 0.32	6	2	4	33.3 %	−0.28 ± 0.64	2155	489	1666	22.7 %	−0.48 ± 0.71	@#
	Bordetella pertussis	50	36	14	72.0 %	0.35 ± 0.64	50	5	45	10.0 %	−0.70 ± 0.64	3258	673	2585	20.7 %	−0.50 ± 0.64	@#$
	Chlamydia pneumoniae	26	16	10	61.5 %	0.30 ± 0.75	26	2	24	7.7 %	−0.84 ± 0.61	1113	109	1004	9.8 %	−0.79 ± 0.63	@#
Gram(+)	*B. anthracis*	55	45	10	81.8 %	0.79 ± 0.88	55	12	43	21.8 %	−0.59 ± 0.76	5493	1054	4439	19.2 %	−0.56 ± 0.71	@#
	Corynebacterium pseudotuberculosis	15	15	0	100.0 %	1.09 ± 0.39	15	7	8	46.7 %	−0.09 ± 0.72	2001	574	1427	28.7 %	−0.33 ± 0.66	@#
	Mycobacterium bovis	13	7	6	53.8 %	0.02 ± 0.65	13	4	9	30.8 %	−0.40 ± 0.62	3891	1016	2875	26.1 %	−0.35 ± 0.63	@#
	Staphylococcus aureus	68	55	13	80.9 %	0.64 ± 0.76	68	13	55	19.1 %	−0.70 ± 0.67	2889	648	2241	22.4 %	−0.52 ± 0.75	@#$
	Staphylococcus pseudintermedius	13	11	2	84.6 %	0.64 ± 0.62	13	0	13	0.0 %	−0.97 ± 0.53	2449	490	1959	20.0 %	−0.59 ± 0.71	@#
	Tropheryma whipplei	23	7	16	30.4 %	−0.18 ± 0.51	23	2	21	8.7 %	−0.83 ± 0.43	805	106	699	13.2 %	−0.71 ± 0.61	@#
Fungi	Candida albicans	51	44	7	86.3 %	0.73 ± 0.51	51	10	41	19.6 %	−0.52 ± 0.53	6036	1832	4204	30.4 %	−0.29 ± 0.62	@#
	Aspergillus fumigatus	15	13	2	86.7 %	0.60 ± 0.54	15	1	14	6.7 %	−0.61 ± 0.54	9577	2709	6868	28.3 %	−0.35 ± 0.66	@#$
	Cryptococcus deuterogattii	13	10	3	76.9 %	0.48 ± 0.60	13	2	11	15.4 %	−0.60 ± 0.53	2945	981	1964	33.3 %	−0.26 ± 0.62	@#$
	Cryptococcus neoformans	3	2	1	66.7 %	0.66 ± 0.59	3	1	2	33.3 %	−0.09 ± 0.75	7429	2248	5181	30.3 %	−0.30 ± 0.61	#
	Coccidioides posadasii	6	5	1	83.3 %	0.74 ± 0.45	6	1	5	16.7 %	−0.50 ± 0.51	27,565	6170	21,395	22.4 %	−0.44 ± 0.60	@#
	Paracoccidioides brasiliensis	8	8	0	100.0 %	0.75 ± 0.36	8	1	7	12.5 %	−0.71 ± 0.64	8399	1987	6412	23.7 %	−0.40 ± 0.58	@#$
	Talaromyces marneffei	5	4	1	80.0 %	0.21 ± 0.68	5	1	4	20.0 %	−0.52 ± 0.41	10,448	2961	7487	28.3 %	−0.34 ± 0.66	@#$
	Ajellomyces capsulatus	25	10	15	40.0 %	0.01 ± 0.59	25	3	22	12.0 %	−0.58 ± 0.54	9214	2099	7115	22.8 %	−0.42 ± 0.58	
Protozoa	Plasmodium vivax	13	13	0	100.0 %	0.81 ± 0.35	13	2	11	15.4 %	−0.59 ± 0.59	5389	1729	3660	32.1 %	−0.31 ± 0.64	@#
	Trypanosoma cruzi	65	56	9	86.2 %	0.62 ± 0.54	65	12	53	18.5 %	−0.59 ± 0.71	19,242	4685	14,557	24.3 %	−0.45 ± 0.74	@#$
	Toxoplasma gondii	13	8	5	61.5 %	0.22 ± 0.67	13	2	11	15.4 %	−0.67 ± 0.53	8315	2183	6132	26.3 %	−0.39 ± 0.60	
Helminth	Ascaris suum	13	11	2	84.6 %	0.60 ± 0.48	4	0	4	0.0 %	−0.71 ± 0.59	10,331	1743	8588	16.9 %	−0.60 ± 0.63	@#
	Echinococcus granulosus	17	10	7	58.8 %	0.26 ± 0.49	17	0	17	0.0 %	−0.91 ± 0.43	10,233	1442	8791	14.1 %	−0.62 ± 0.57	@#$
	Fasciola hepatica	10	6	4	60.0 %	0.14 ± 0.70	10	0	10	0.0 %	−0.95 ± 0.36	11,190	1810	9380	16.2 %	−0.58 ± 0.57	@#$
	Schistosoma mansoni	11	7	4	63.6 %	−0.01 ± 0.69	11	0	11	0.0 %	−0.93 ± 0.27	10,770	1946	8824	18.1 %	−0.56 ± 0.58	@#$
	Trichinella spiralis	13	10	3	76.9 %	0.35 ± 0.54	13	1	12	7.7 %	−0.77 ± 0.58	18,572	2333	16,239	12.6 %	−0.67 ± 0.56	@#
	mean	23.65	18.35	5.30	71.0 %	0.44	23.65	3.65	20.00	17.0 %	−0.59	8818.35	2091.35	6726.00	23.0 %	−0.45	

Number of total proteins (#Prot), predicted antigens (#Ag), predicted non-antigens (#N-Ag), percentage of predicted antigens (%Ag), and mean IAscore (±SD) are shown for each dataset (Antigens, Non-Antigens, and Proteome) per organism.
@ Significant difference between antigens and non-antigens ($p < 0.05$).
# Significant difference between antigens and proteome ($p < 0.05$).
$ Significant difference between non-antigens and proteome ($p < 0.05$).
Statistical significance determined using Mann-Whitney-Wilcoxon test.



Metric	IApred	ANTIGENpro	VaxiJen 2.0	VaxiJen 3.0-Bacteria	VaxiJen 3.0-All
True Positive	153	**179**	174	85	123
True Negative	154	109	77	120	**157**
False Positive	64	109	141	**29**	61
False Negative	65	**39**	44	64	95
Accuracy	**0.704**	0.661	0.576	0.688	0.642
Precision	0.705	0.622	0.552	**0.746**	0.668
Sensitivity	0.702	**0.821**	0.798	0.570	0.564
Specificity	0.706	0.500	0.353	**0.805**	0.720
FPR	0.294	0.500	0.647	**0.195**	0.280
FDR	0.295	0.378	0.448	**0.254**	0.332
MCC	**0.408**	0.339	0.169	0.387	0.288
Youden's J	**0.408**	0.321	0.151	0.376	0.284
F1-score	0.703	**0.708**	0.653	0.646	0.612
F0.5-score	**0.704**	0.653	0.589	0.702	0.645
Brier Score	**0.202**	0.236	0.252	0.258	0.298
ECE	**0.074**	0.161	0.173	0.160	0.188

Fig. 3. Comparison between antigenicity predictors: The performance of ANTIGENpro, VaxiJen 2.0, VaxiJen 3.0 and IApred was evaluated using a dataset derived from the Protegen database. (A) ROC-AUC curves are comparing the predictive accuracy of each tool. (B) Calibration plot displaying the mean chi-square values for the four predictors. (C) Table summarizing key performance metrics for each predictor. (D) Graphical representation of predictors agreement based on IAscore, with darker colors indicating higher consensus among predictors regarding a protein antigen classification.

Expected Calibration Error (ECE = 0.074), representing the average gap between predicted confidence levels and observed frequencies. Its mean chi-squared calibration statistic of 0.047 further confirms that predicted probabilities closely match real-world results. (Fig. 3B). Taking all together, IApred showed to be the most reliable tool for general antigenicity prediction, particularly in scenarios were minimizing both false positives and false negatives is critical. Fig. 3D illustrates the agreement between IApred and the other predictors (excluding Vaxijen 3.0-bacteria as it uses a smaller dataset), with antigens and non-antigens arranged according to their IAscore values. Antigens are depicted in green, with darker shades indicating higher consensus among predictors, while non-antigens are shown in red, with darker shades representing stronger agreement in their classification. It is worth noting that the external evaluation dataset is heavily composed of bacterial proteins (150 out of 218), which may influence the comparison between predictors. However, we were unable to generate a sufficiently large and pathogen-class-balanced dataset. This skew should be considered when using the predictors or referring to this comparison in the future.

Interestingly, antigen scores showed a consistent length dependence for all predictors except VaxiJen 2.0: Pearson correlations were positive and modest ($r \approx 0.29$–0.35) with $p < 0.0001$ for IApred, ANTIGENpro, Vaxijen 3.0-all, and Vaxijen 3.0-bacteria. In contrast, non-antigens showed no convincing association between score and length; although IApred exhibited a weak correlation ($r \approx 0.17$, $p = 0.014$), its magnitude is small, and it might not survive multiple-testing correction. When combining antigens and non-antigens, correlations were attenuated ($r \approx$ 0.12–0.18; VaxiJen 2.0 remained nonsignificant), consistent with the effect being mostly specific to antigens. For IApred, this correlation is expected because the model includes features directly or indirectly related to protein length (e.g., molecular weight). In any case, the antigenicity prediction interpretation should account for protein length.

Conclusion

IApred establishes itself as a robust and versatile solution for intrinsic protein antigenicity prediction, outperforming widely used tools by achieving a balanced sensitivity (0.702) and specificity (0.706) across diverse pathogen classes. Its open-source, host-and-pathogen-agnostic design, combined with the ability to process approximately 1000 sequences per minute using the standard google colab configuration (and up to 10,000 sequences per minute in the web server, although limited to 1000 sequences per run), enables seamless integration into bioinformatics pipelines and large-scale analyses relevant to vaccine development, diagnostic design, and therapeutic protein engineering.

By addressing key limitations of existing predictors—such as limited accessibility, lack of transparency, and restricted applicability to non-bacterial antigens—IApred provides the scientific community with a transparent and adaptable framework. The comprehensive training dataset, which encompasses diverse pathogen classes, ensures reliable performance across a wide range of organisms, making IApred particularly valuable for studying emerging pathogens and non-bacterial antigens. The availability of the training code and datasets further empowers users to retrain or extend the model for specialized applications.

While current results demonstrate strong and reliable performance, there is significant potential for future enhancements. Expanding the training dataset, especially for underrepresented non-bacterial pathogens, could further improve prediction accuracy. Incorporating immunological and structural features, such as T- and B-cell epitope predictions, could further enhance prediction accuracy and scope. An advanced version of IApred, currently under development, aims to address these aspects for specialized use cases, while the present version remains highly effective for general applications.

In summary, IApred offers a balanced, accurate, and accessible alternative to existing antigenicity predictors. Its open-source foundation and comprehensive design not only address current challenges but also lay the groundwork for future advances in antigenicity prediction and immunoinformatic research

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Universidad de la República, Agencia Nacional de Investigación e Innovación (ANII - Uruguay) and PEDECIBA-Química (Uruguay) are acknowledged for their general financial support.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used Claude.ai in order to improve the English writing. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the published article.

CRediT authorship contribution statement

Sebastian Miles: Writing – review & editing, Writing – original draft, Software, Formal analysis, Data curation, Conceptualization. **Gonzalo Menafra:** Formal analysis, Data curation. **Andrés Iriarte:** Writing – review & editing, Supervision, Conceptualization. **Jose Alejandro Chabalgoity:** Writing – review & editing, Supervision, Conceptualization.

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.

Supplementary materials

Supplementary Figure 1: Features correlation and importance.

The heatmap shows the correlations between the used features, while the histogram shows the importance of each feature in the final model. Positive importances correlates with

Supplementary Figure 2: Correlation between antigenicity and protein length.

The Pearson Correlation between the antigenicity score generated by each of the four analysed predictors and the length of the protein. Each figure shows the Pearson Correlation value for the Antigen, Non-antigen and combined datasets (box), as well as a linear regression and Locally Estimated Scatterpoint Smoothing (LOESS) Curves. (A) IApred; (B) ANTIGENpro; (C) Vaxijen 2.0; (D) Vaxijen 3.0-bacteria; (F) Vaxijen 3.0-all.

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.immuno.2025.100061.

Data availability

All the code and data is open-source.

References

[1] Owen JA, Punt J, Stranford SA, Jones PP. Kuby immunology. 7th edn. Susan Winslow; 2013.
[2] Weller S, Sterlin D, Fadeev T, Coignard E, Verge de Los Aires A, Goetz C, Fritzen R, Bahuaud M, Batteux F, Gorochov G, Weill JC, Reynaud CA. T-independent responses to polysaccharides in humans mobilize marginal zone B cells prediversified against gut bacterial antigens. Sci Immunol 2023;8(79):eade1413. https://doi.org/10.1126/sciimmunol.ade1413. Epub 2023 Jan 27. PMID: 36706172; PMCID: PMC7614366.
[3] Magnan CN, Zeller M, Kayala MA, Vigil A, Randall A, Felgner PL, Baldi P. High-throughput prediction of protein antigenicity using protein microarray data. Bioinformatics 2010;26(23):2936–43. https://doi.org/10.1093/bioinformatics/btq551.
[4] Qiu J, Qiu T, Yang Y, et al. Incorporating structure context of HA protein to improve antigenicity calculation for influenza virus A/H3N2. Sci Rep 2016;6:31156. https://doi.org/10.1038/srep31156.
[5] Doyle HA, Mamula MJ. Post-translational protein modifications in antigen recognition and autoimmunity. Trends Immunol 2001;22(8):443–9. https://doi.org/10.1016/s1471-4906(01)01976-7. PMID: 11473834.
[6] Lollier V, Denery-Papini S, Larré C, Tessier D. A generic approach to evaluate how B-cell epitopes are surface-exposed on protein structures. Mol Immunol 2011;48(4):577–85. https://doi.org/10.1016/j.molimm.2010.10.011. Epub 2010 Dec 15. PMID: 21111484; PMCID: PMC7112657.
[7] Jawa V, Terry F, Gokemeijer J, Mitra-Kaushik S, Roberts BJ, Tourdot S, De Groot AS. T-cell dependent immunogenicity of protein therapeutics pre-clinical assessment and mitigation-updated consensus and review 2020. Front Immunol 2020;11:1301. https://doi.org/10.3389/fimmu.2020.01301. PMID: 32695107; PMCID: PMC7338774.
[8] Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. BMC Bioinform 2007;8:4. https://doi.org/10.1186/1471-2105-8-4.
[9] Dimitrov I, Zaharieva N, Doytchinova I. Bacterial immunogenicity prediction by machine Learning methods. Vaccines 2020;8(4):709. https://doi.org/10.3390/vaccines8040709.
[10] Ricci AD, Brunner M, Ramoa D, Carmona SJ, Nielsen M, Agüero F. APRANK: computational prioritization of antigenic proteins and peptides from complete pathogen proteomes. Front Immunol 2021;12:702552. https://doi.org/10.3389/fimmu.2021.702552.
[11] Venkatarajan M, Braun W. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. J Mol Model 2001;7:445–53. https://doi.org/10.1007/s00894-001-0058-5.
[12] Vivona S, Bernante F, Filippini F. NERVE: new enhanced reverse vaccinology environment. BMC Biotechnol 2006;6:35. https://doi.org/10.1186/1472-6750-6-35.
[13] He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010;2010:297505. https://doi.org/10.1155/2010/297505.
[14] Jaiswal V, Chanumolu SK, Gupta A, Chauhan RS, Rout C. Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. BMC Bioinform 2013;14:211. https://doi.org/10.1186/1471-2105-14-211.
[15] Goodswen SJ, Kennedy PJ, Ellis JT. Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. Bioinformatics 2014;30(16):2381–3. https://doi.org/10.1093/bioinformatics/btu300.
[16] Moise L, Gutierrez A, Kibria F, Martin R, Tassone R, Liu R, Terry F, Martin B, De Groot AS. iVAX: an integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. Hum Vaccin Immunother 2015;11(9):2312–21. https://doi.org/10.1080/21645515.2015.1061159.
[17] Altindis E, Cozzi R, Di Palo B, Necchi F, Mishra RP, Fontana MR, Soriani M, Bagnoli F, Maione D, Grandi G, Liberatori S. Protectome analysis: a new selective bioinformatics tool for bacterial vaccine candidate discovery. Mol Cell Proteom 2015;14(2):418–29. https://doi.org/10.1074/mcp.M114.039362.
[18] Rizwan M, Naz A, Ahmad J, Naz K, Obaid A, Parveen T, Ahsan M, Ali A. VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. BMC Bioinform 2017;18(1):106. https://doi.org/10.1186/s12859-017-1540-0.
[19] Ong E, Cooke MF, Huffman A, Xiang Z, Wong MU, Wang H, Seetharaman M, Valdez N, He Y. Vaxign2: the second generation of the first web-based vaccine

design program using reverse vaccinology and machine learning. Nucleic Acids Res 2021;49(W1):W671–8. https://doi.org/10.1093/nar/gkab279.
[20] Rawal K, Sinha R, Nath SK, Preeti P, Kumari P, Gupta S, Sharma T, Strych U, Hotez P, Bottazzi ME. Vaxi-DL: a web-based deep learning server to identify potential vaccine candidates. Comput Biol Med 2022;145:105401. https://doi.org/10.1016/j.compbiomed.2022.105401.
[21] Zhang Y., Huffman A., Johnson J., He Y. Vaxign-DL: a deep learning-based method for vaccine design and its evaluation. bioRxiv [Preprint]. 2023 Dec 1: 2023.11.29.569096. https://doi.org/10.1101/2023.11.29.569096.
[22] Larsen JE, Lund O, Nielsen M. Improved method for predicting linear B-cell epitopes. Immunome Res 2006;2:2. https://doi.org/10.1186/1745-7580-2-2.
[23] Nielsen M, Justesen S, Lund O, Lundegaard C, Buus S. NetMHCIIpan-2.0 - improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure. Immunome Res 2010;6:9. https://doi.org/10.1186/1745-7580-6-9.
[24] Yang B, Sayers S, Xiang Z, He Y. Protegen: a web-based protective antigen database and analysis system. Nucleic Acids Res 2011;39(Database issue). https://doi.org/10.1093/nar/gkq944. D1073-8.
[25] (ed). In: Walker John M, editor. The proteomics protocols handbook. Humana Press; 2005. p. 571–607.
[26] Paszke A., Gross S., Massa F., Lerer A., Bradbury J., Chanan G., et al. PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems. 2019.
[27] Kumar M, Michael S, Alvarado-Valverde J, Zeke A, Lazar T, Glavina J, Nagy-Kanta E, Donagh JM, Kalman ZE, Pascarelli S, Palopoli N, Dobson L, Suarez CF, Van Roey K, Krystkowiak I, Griffin JE, Nagpal A, Bhardwaj R, Diella F, Mészáros B, Dean K, Davey NE, Pancsa R, Chemes LB, Gibson TJ. ELM-the eukaryotic Linear Motif resource-2024 update. Nucleic Acids Res 2024;52(D1):D442–55. https://doi.org/10.1093/nar/gkad1058.
[28] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011.
[29] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002.
[30] Zhang J, Antigenicity Tao A. Immunogenicity, allergenicity. Allergy Bioinform 2015;8:175–86. https://doi.org/10.1007/978-94-017-7444-4_11. PMCID: PMC7123983.
[31] Wang Y, Wu W, Negre NN, White KP, Li C, Shah PK. Determinants of antigenicity and specificity in immune response for protein sequences. BMC Bioinform 2011;12:251. https://doi.org/10.1186/1471-2105-12-251. PMID: 21693021; PMCID: PMC3133554.
[32] Garg A, Singhal N, Kumar M. Investigating the eukaryotic host-like SLiMs in microbial mimitopes and their potential as novel drug targets for treating autoimmune diseases. Front Microbiol 2022;13:1039188. https://doi.org/10.3389/fmicb.2022.1039188. PMID: 36406429; PMCID: PMC9672370.
[33] Antenucci F, Magnowska Z, Nimtz M, Roesch C, Jänsch L, Bojesen AM. Immunoproteomic characterization of outer membrane vesicles from hyper-vesiculating Actinobacillus pleuropneumoniae. Vet Microbiol 2019;235:188–94. https://doi.org/10.1016/j.vetmic.2019.07.001.
[34] Zhu YZ, Cai CS, Zhang W, Guo HX, Zhang JP, Ji YY, Ma GY, Wu JL, Li QT, Lu CP, Guo XK. Immunoproteomic analysis of human serological antibody responses to vaccination with whole-cell pertussis vaccine (WCV). PLoS One 2010;5(11):e13915. https://doi.org/10.1371/journal.pone.0013915.
[35] Bunk S, Susnea I, Rupp J, Summersgill JT, Maass M, Stegmann W, Schrattenholz A, Wendel A, Przybylski M, Hermann C. Immunoproteomic identification and serological responses to novel Chlamydia pneumoniae antigens that are associated with persistent C. pneumoniae infections. J Immunol 2008;180(8):5490–8. https://doi.org/10.4049/jimmunol.180.8.5490.
[36] Chitlaru T, Gat O, Grosfeld H, Inbar I, Gozlan Y, Shafferman A. Identification of in vivo-expressed immunogenic proteins by serological proteome analysis of the *Bacillus anthracis* secretome. Infect Immun 2007;75(6):2841–52. https://doi.org/10.1128/IAI.02029-06.
[37] Seyffert N, Silva RF, Jardin J, Silva WM, Castro TL, Tartaglia NR, Santana KT, Portela RW, Silva A, Miyoshi A, Le Loir Y, Azevedo V. Serological proteome analysis of Corynebacterium pseudotuberculosis isolated from different hosts reveals novel candidates for prophylactics to control caseous lymphadenitis. Vet Microbiol 2014;174(1–2):255–60. https://doi.org/10.1016/j.vetmic.2014.08.024.
[38] Jeon HS, Shin AR, Son YJ, Kim JM, Jang Y, Kim S, Lee KI, Choi CH, Park JK, Kim HJ. Seroreactive mycobacterial proteins and lipid in cattle with bovine tuberculosis. J Bacteriol Virol 2015.
[39] Le Maréchal C, Jardin J, Jan G, Even S, Pulido C, Guibert JM, Hernandez D, François P, Schrenzel J, Demon D, Meyer E, Berkova N, Thiéry R, Vautor E, Le Loir Y. Staphylococcus aureus seroproteomes discriminate ruminant isolates causing mild or severe mastitis. Vet Res 2011;42(1):35. https://doi.org/10.1186/1297-9716-42-35.
[40] Couto N, Martins J, Lourenço AM, Pomba C, Varela Coelho A. Identification of vaccine candidate antigens of Staphylococcus pseudintermedius by whole proteome characterization and serological proteomic analyses. J Proteom 2016;133:113–24. https://doi.org/10.1016/j.jprot.2015.12.017.
[41] Kowalczewska M, Fenollar F, Lafitte D, Raoult D. Identification of candidate antigen in Whipple's disease using a serological proteomic approach. Proteomics 2006;6(11):3294–305. https://doi.org/10.1002/pmic.200500171.
[42] Vaz C, Pitarch A, Gómez-Molero E, Amador-García A, Weig M, Bader O, Monteoliva L, Gil C. Mass spectrometry-based proteomic and immunoproteomic analyses of the Candida albicans hyphal secretome reveal diagnostic biomarker candidates for invasive candidiasis. J Fungi 2021;7(7):501. https://doi.org/10.3390/jof7070501.
[43] Wangsanut T, Pongpom M. Human-fungal pathogen interactions from the perspective of immunoproteomics analyses. Int J Mol Sci 2024;25(6):3531. https://doi.org/10.3390/ijms25063531.
[44] Almeida MA, Almeida-Paes R, Guimarães AJ, Valente RH, Soares CMA, Zancopé-Oliveira RM. Immunoproteomics reveals pathogen's antigens involved in homo sapiens-histoplasma capsulatum interaction and specific linear B-cell epitopes in histoplasmosis. Front Cell Infect Microbiol 2020;10:591121. https://doi.org/10.3389/fcimb.2020.591121.
[45] Kassegne K, Abe EM, Chen JH, Zhou XN. Immunomic approaches for antigen discovery of human parasites. Expert Rev Proteom 2016;13(12):1091–101. https://doi.org/10.1080/14789450.2016.1252675.
[46] Carmona SJ, Nielsen M, Schafer-Nielsen C, Mucci J, Altcheh J, Balouz V, Tekiel V, Frasch AC, Campetella O, Buscaglia CA, Agüero F. Towards high-throughput immunomics for infectious diseases: use of next-generation peptide microarrays for rapid discovery and mapping of antigenic determinants. Mol Cell Proteom 2015;14(7):1871–84. https://doi.org/10.1074/mcp.M114.045906.
[47] Mamaghani AJ, Fathollahi A, Arab-Mazar Z, Kohansal K, Fathollahi M, Spotin A, Bashiri H, Bozorgomid A. Toxoplasma gondii vaccine candidates: a concise review. Ir J Med Sci 2023;192(1):231–61. https://doi.org/10.1007/s11845-022-02998-9.
[48] González-Miguel J, Morchón R, Gussoni S, Bossetti E, Hormaeche M, Kramer LH, Simón F. Immunoproteomic approach for identification of Ascaris suum proteins recognized by pigs with porcine ascariasis. Vet Parasitol 2014;203(3–4):343–8. https://doi.org/10.1016/j.vetpar.2014.03.031.
[49] Miles S, Magnone J, Cyrklaff M, Arbildi P, Frischknecht F, Dematteis S, Mourglia-Ettlin G. Linking murine resistance to secondary cystic echinococcosis with antibody responses targeting echinococcus granulosus tegumental antigens. Immunobiology 2020;225(3):151916. https://doi.org/10.1016/j.imbio.2020.151916.
[50] Becerro-Recio D, González-Miguel J, Ucero A, Sotillo J, Martínez-Moreno Á, Pérez-Arévalo J, Cwiklinski K, Dalton JP, Siles-Lucas M. Recognition pattern of the fasciola hepatica excretome/secretome during the course of an experimental infection in sheep by 2D immunoproteomics. Pathogens 2021;10(6):725. https://doi.org/10.3390/pathogens10060725.
[51] Al-Naseri A, Al-Absi S, El Ridi R, Mahana N. A comprehensive and critical overview of schistosomiasis vaccine candidates. J Parasit Dis 2021;45(2):557–80. https://doi.org/10.1007/s12639-021-01387-w.
[52] Grzelak S, Stachyra A, Stefaniak J, Mrówka K, Moskwa B, Bień-Kalinowska J. Immunoproteomic analysis of trichinella spiralis and trichinella britovi excretory-secretory muscle larvae proteins recognized by sera from humans infected with Trichinella. PLoS One 2020;15(11):e0241918. https://doi.org/10.1371/journal.pone.0241918.