Bioinformatics, 36(10), 2020, 3185–3191
doi: 10.1093/bioinformatics/btaa119
Advance Access Publication Date: 25 February 2020
Original Paper

Systems biology

Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens

Edison Ong[1], Haihe Wang[2,3], Mei U Wong[3], Meenakshi Seetharaman[4], Ninotchka Valdez[4] and Yongqun He[3,5,6,*]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, USA, [2]Department of Pathogenobiology, Daqing Branch of Harbin Medical University, Daqing 163319, China, [3]Unit for Laboratory Animal Medicine, [4]College of Literature, Science, and the Arts, University of Michigan, [5]Department of Microbiology and Immunology and [6]Center of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, USA

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on September 11, 2019; revised on February 10, 2020; editorial decision on February 16, 2020; accepted on February 18, 2020

Abstract

Motivation: Reverse vaccinology (RV) is a milestone in rational vaccine design, and machine learning (ML) has been applied to enhance the accuracy of RV prediction. However, ML-based RV still faces challenges in prediction accuracy and program accessibility.

Results: This study presents Vaxign-ML, a supervised ML classification to predict bacterial protective antigens (BPAgs). To identify the best ML method with optimized conditions, five ML methods were tested with biological and physiochemical features extracted from well-defined training data. Nested 5-fold cross-validation and leave-one-pathogen-out validation were used to ensure unbiased performance assessment and the capability to predict vaccine candidates against a new emerging pathogen. The best performing model (eXtreme Gradient Boosting) was compared to three publicly available programs (Vaxign, VaxiJen, and Antigenic), one SVM-based method, and one epitope-based method using a high-quality benchmark dataset. Vaxign-ML showed superior performance in predicting BPAgs. Vaxign-ML is hosted in a publicly accessible web server and a standalone version is also available.

Availability and implementation: Vaxign-ML website at http://www.violinet.org/vaxign/vaxign-ml, Docker standalone Vaxign-ML available at https://hub.docker.com/r/e4ong1031/vaxign-ml and source code is available at https://github.com/VIOLINet/Vaxign-ML-docker.

Contact: yonqunh@med.umich.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

As the most successful intervention in modern medicine, vaccination is still facing the huge difficulty of developing safe and effective vaccines against many infectious diseases such as tuberculosis, HIV, and malaria (WHO, 2014). The advance of high-throughput sequencing technology has fostered an innovative genome-based vaccine design approach in the early 1990s, termed reverse vaccinology (RV) (Rappuoli, 2000). The first RV study identified meningococcal protein vaccine candidates using the whole genome sequences of Group B meningococcus. The authors in that study selected and verified 28 immunogenic proteins using bioinformatics approach followed by experimental validation (Pizza, 2000). The Bexsero vaccine, formulated using 5 out of the 28 protein candidates, has been licensed in Europe and the United States (Folaranmi *et al.*, 2015; Vernikos and Medini, 2014).

The great success of the first RV study has led to the creation of many RV prediction programs (Dalsass *et al.*, 2019). The currently reported open-source RV programs could be characterized based on the algorithmic approaches or input feature types. The algorithmic approaches include rule-based filtering and machine learning (ML) classification methods. NERVE, the first publicly available rule-based filtering RV program, is a standalone software published in 2006 (Vivona *et al.*, 2006). Four years later, the first web-based filtering RV program, Vaxign, was developed similar to NERVE but with additional analyses (He *et al.*, 2010) and has been applied in vaccine design studies against more than 10 pathogenic bacteria such as *Helicobacter pylori* (Navarro-Quiroz *et al.*, 2018), *Acinetobacter baumannii* (Singh *et al.*, 2016), and *Mycobacterium* spp. (Hossain *et al.*, 2017). Following NERVE and Vaxign, two other filtering-based RV programs, Jenner-predict server (Jaiswal *et al.*, 2013) and VacSol (Rizwan *et al.*, 2017) were published (Jaiswal *et al.*, 2013; Rizwan *et al.*, 2017). All these currently available rule-based filtering RV programs use only biological features as the data input.

ML classification has also been used for RV prediction. VaxiJen was the first ML classification RV program published in 2007 (Doytchinova and Flower, 2007). Bowman *et al.* (2011) and Heinson *et al.* (2017) extended the training data of VaxiJen and revised the ML algorithm. Their final training data, termed 200BPA, consisted of 200 bacterial protective antigens (BPAgs), and 200 non-protective proteins. The non-protective proteins were selected if it had no homology (BLASTp *E*-value $\leq$ 10E−3) to the BPAgs. A major difference between VaxiJen and Bowman–Heinson was that VaxiJen used physicochemical features of the input proteins while the later program used biological features. The second lineage of ML-based RV prediction program originated from the development of ANTIGENpro in 2010 (Magnan *et al.*, 2010). ANTIGENpro collected protective antigens (PAgs) from the literature in combination with the positive and negative samples tested via protein microarrays probed with sera from naive, exposed, and vaccinated individuals. Rahman *et al.* (2019) revised the algorithmic method of ANTIGENpro and developed Antigenic using the same training data. Both ANTIGENpro and Antigenic used physicochemical features as the data input. The authors of these two papers argued that proteins being able to elicit a significant antibody response could be considered as ‘protective antigens.’ However, data collected based on antibody responses did not guarantee to be protective. Such data lack the results from protection assays in at least one laboratory animal model. More importantly, antibody production does not capture cell-mediated immunity, which is often an essential protective immune mechanism. For example, *Brucella* vaccine RB51-induced protection is purely based on cell-mediated immunity, and its induced antibody response does not offer any observed protection (Jimenez de Bagues *et al.*, 1994).

All of the ML RV programs mentioned earlier were not designed to predict eukaryotic vaccine candidates. Goodswen *et al.* (2013) developed the first, and the only ML RV targeting eukaryotic pathogens to our best knowledge. However, due to the lack of reported eukaryotic PAgs, proteins which are surface-exposed and have at least one T cell epitopes, were treated as positive samples in this study. These collected data might lack supporting experimental evidence. Furthermore, a protein with epitopes does not guarantee that this protein can elicit protective immune responses (Flower *et al.*, 2010). An independent resource, Protegen, manually collected 590 PAgs over 100 infectious diseases caused by pathogens (bacteria, viruses, and parasites) and non-infectious diseases including cancers and allergies (Yang *et al.*, 2011). Each of these collected PAgs is an antigen that can elicit a protective immune response, which has been experimentally verified by at least one laboratory animal model. A preliminary ML RV study trained on the Protegen data reported high PAgs prediction accuracy (He and Xiang, 2012). Protegen has doubled the number of annotated pathogen PAgs since its initial release in 2011.

Although a significant effort has been made on enhancing the RV prediction with ML, there are still many obstacles in ML-based RV prediction. First, all currently available programs use either biological or physicochemical properties for input protein sequence annotations. Previous studies reported that the protectiveness of BPAgs was significantly correlated to biological properties (Ong *et al.*, 2017) and physicochemical properties (Mayers *et al.*, 2003). Studies using ML algorithms trained on data with physicochemical properties annotated also showed high BPAg prediction accuracy. Therefore, the relations of BPAgs to biological and physicochemical properties deserved a more in-depth analysis and should be combined to annotate proteins in the training data for better BPAg prediction. Second, the quality of the benchmarking datasets varied in current reported studies. As aforementioned, the testing data, which was used to evaluate ANTIGENpro and Antigenic, was primarily based on the antibody responses and might not capture the cell-mediated immune responses. Therefore, the dataset of ANTIGENpro and Antigenic was excluded from this study. Last but not least, the 200BPA data from VaxiJen and Bowman–Heinson only included PAgs with supporting experimental evidence (Bowman *et al.*, 2011; Heinson *et al.*, 2017). The negative samples were randomly selected from non-homologous proteins to the PAgs. The random undersampling may not reflect the real distribution of PAgs in the proteome. Besides, VaxiJen and Bowman–Heinson model were not evaluated using an external independent dataset.

In this Vaxign-ML study, epitope information was not incorporated into the pipeline. The prediction of epitopes has been an active area of vaccine design, and the IEDB database and IEDB-AR resources (Fleri *et al.*, 2017) provides a comprehensive T cell and B cell epitope query, prediction, and analysis tools. However, the prediction of epitopes is dependent on the host information (e.g. MHC alleles and antibody). The training dataset in Vaxign-ML consisted of experimentally verified PAgs manually annotated from studies in over 10 host species. Therefore, the prediction of BPAgs in Vaxign-ML did not take host species into account, and the T cell or B cell epitope predictions were not included. Current epitope-based BPAg prediction methods such as iVAX (Moise *et al.*, 2015) often depend on the frequency or density of the epitopes located on the protein. However, such epitope measurement may not necessarily translate into protective immune responses. Despite the uncertainty of correspondence between epitope and protective immune response, we implemented an epitope-based method using IEDB epitope prediction tools. The performance of Vaxign-ML was compared to the epitope-based method in this study.

In this paper, we presented a systematic evaluation of a supervised ML classification RV program trained on the Protegen BPAgs with their biological and physicochemical features annotated. The BPAgs and the non-protective proteins were first carefully checked for homology to ensure the quality of training data. Due to imbalanced classes in the training data, three data resampling strategies were applied to the original data. Nested 5-fold cross-validation (N5CV) and leave-one-pathogen-out validation (LOPOV) were used to evaluate five supervised ML algorithms with feature selection and hyperparameter optimization. The best performing model, termed Vaxign-ML, was benchmarked using a curated external independent dataset and demonstrated superior predictive performance.

2 Materials and methods

The overall project workflow is described in Figure 1. In brief, positive and negative samples were downloaded and processed from Protegen and Uniprot (The UniProt Consortium, 2008; Yang *et al.*, 2011). The biological and physicochemical features for these protein sequences were annotated using publicly available bioinformatics software (Supplementary Table S1). Four data resampling strategies and five supervised ML classification algorithms were trained and evaluated. The performance of the best model, named as Vaxign-ML, was compared to four BPAg prediction methods and one epitope-based method using a curated external independent dataset.

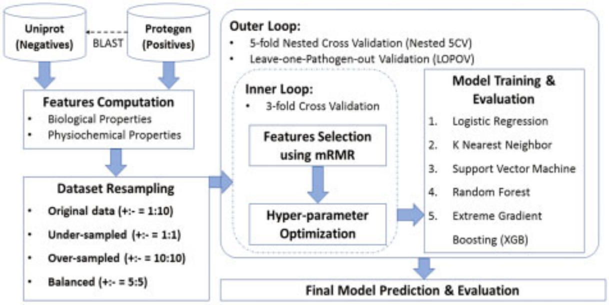
Uniprot (Negatives)
BLAST
Protegen (Positives)
Features Computation
• Biological Properties
• Physiochemical Properties
Dataset Resampling
• Original data (+:- = 1:10)
• Under-sampled (+:- = 1:1)
• Over-sampled (+:- = 10:10)
• Balanced (+:- = 5:5)
Outer Loop:
• 5-fold Nested Cross Validation (Nested 5CV)
• Leave-one-Pathogen-out Validation (LOPOV)
Inner Loop:
• 3-fold Cross Validation
Features Selection using mRMR
Hyper-parameter Optimization
Model Training & Evaluation
1. Logistic Regression
2. K Nearest Neighbor
3. Support Vector Machine
4. Random Forest
5. Extreme Gradient Boosting (XGB)
Final Model Prediction & Evaluation

Fig. 1. Overall study workflow. This flowchart depicted the entire process to train and evaluate ML-based RV models. See main text for details

2.1 Data preparation

2.1.1 BPAgs and non-protective protein sequences

BPAgs with supporting experimental evidence were downloaded from the Protegen database (Yang *et al.*, 2011). As of July 31, 2019, Protegen included 584 BPAgs from 50 Gram+ and Gram− pathogenic bacteria (Supplementary Table S2). BPAgs with sequence similarity over 30% were considered homologous proteins, which is a commonly accepted threshold for homologous proteins (Pearson, 2013), and removed from the study to avoid potential bias (Supplementary Fig. S1). The final positive samples in the original data consisted of 397 BPAgs. A set of 'gold-standard' non-protective proteins does not exist. Therefore, the negative samples in the training dataset were selected based on its sequence dissimilarity to the BPAgs (Supplementary Fig. S1), as described in previous ML-based BPAg prediction studies (Bowman *et al.*, 2011; Doytchinova and Flower, 2007; Heinson *et al.*, 2017). The whole pathogen proteomes of the 50 pathogenic bacteria were downloaded from the Uniprot database (The UniProt Consortium, 2008). Any pathogen proteins with sequence similarity <30% to the BPAgs were kept, and homologous proteins were also removed.

2.1.2 Protein sequence annotations

Vaxign-ML used two categories of features for each of the protein sequences: biological and physicochemical features. The biological features, including the Gram(±) stain, subcellular localization (Yu *et al.*, 2010), adhesin probability (Sachdeva *et al.*, 2005), transmembrane helix (Krogh *et al.*, 2001), signal peptide (Petersen *et al.*, 2011) and immunogenicity (Fleri *et al.*, 2017), were computed using publicly available bioinformatics software programs. The analyzed physicochemical features included the compositions, transitions, and distributions (Dubchak *et al.*, 1995), quasi-sequence-order (Chou, 2000), Moreau–Broto autocorrelation (Feng and Zhang, 2000; Lin and Pan, 2001) and Geary autocorrelation (Sokal and Thomson, 2006) of various physicochemical properties such as charge, hydrophobicity, polarity and solvent accessibility (Ong *et al.*, 2007). A total of 509 biological and physicochemical features were annotated for each of the protein sequences in the original data (Supplementary Table S1).

2.1.3 Balance data with resampling

The original data were imbalanced and had a dimension of 4367 samples (positive-to-negative classes ratio = 1:10) and 509 features. To study the effect of class imbalance, three data resampling strategies were implemented. First, the negative samples in the original data were randomly sampled without replacement (positive-to-negative classes ratio = 1:1). Second, the Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al.*, 2002) was applied to the original data to increase the number of positive samples (positive-to-negative classes ratio = 10:10). Finally, the balanced resampling strategy was applied by combining both under- and oversampling strategies. The negative samples in the original training data were randomly sampled without replacement to have five times the size of positive samples. Then, the positive samples were oversampled using SMOTE (positive-to-negative ratio = 5:5).

2.2 Supervised ML classification

Five supervised ML classification algorithms were used in this study including logistic regression (LR), support vector machine (SVM), k-nearest neighbor (KNN), random forest (RF) (Pedregosa *et al.*, 2012) and extreme gradient boosting (XGB) (Chen and Guestrin, 2016). The best performing model was trained and named 'Vaxign-ML.' The output of Vaxign-ML is the percentile rank score from the final ML classification model, termed 'protegenicity' in Equation (1) where c_l is the count of all scores in Vaxign-ML prediction which are less than the score of interest, f_i is the frequency of the score of interest and N is the number of sample in the original data.

$$\text{Protegenicity score} = \frac{c_l + 0.5 f_i}{N} \times 100\%. \qquad (1)$$

2.2.1 Nested 5-fold cross-validation

An N5CV was applied to evaluate all supervised ML classification models. The training data, including original data, undersampled, oversampled, and balanced, were randomly split into five parts while preserving the percentage of positive and negative samples. One important note is that the data resampling was only performed after the N5CV splitting to avoid duplicated positive samples being included in both training and testing data. Among the five parts, four parts were used for training. Feature selection with mRMR (Ding and Peng, 2003) and hyperparameter optimization was applied before training all classification models. The remaining part was used as the testing set for model evaluation.

To determine whether the discriminative power of the prediction models depended on the immunogenic potential in the Protegen dataset rather than sequence dissimilarity, the negative dataset was randomly split into two sets with sequence identity <30%. The same N5CV was applied to confirm that the discriminative performance depended on the PAg potential in the Protegen database rather than sequence dissimilarity.

2.2.2 Leave-one-pathogen-out validation

To have a more unbiased estimation of the classification performance and to mimic the situation where vaccine candidates would be needed for a newly emerging pathogen, a LOPOV was implemented. Ten tested pathogens included four Gram+ pathogens (*Mycobacterium tuberculosis*, *Staphylococcus aureus*, *Streptococcus pneumoniae,* and *Streptococcus pyogenes*) and six Gram− pathogens (*Helicobacter pylori*, *Neisseria meningitidis*, *Brucella abortus*, *Escherichia coli*, *Yersinia pestis*, and *Haemophilus parasuis*). The positive and negative samples from these 10 pathogens were held out as testing sets. The remaining samples were used for training. Data sampling, feature selection, and hyperparameter optimization were applied before training all classification models similar to the N5CV.

2.3 Benchmarking with an independent dataset

A curated external independent dataset was created to benchmark the best performing model (Vaxign-ML). Dalsass *et al.* (2019) collected a list of 100 BPAgs termed 100BPA. However, 100BPA only includes positive samples. Another dataset consisted of 200 positive and 200 negative samples (200BPA) was initially created for the development of VaxiJen program (Doytchinova and Flower, 2007) and later extended by Bowman *et al.* (2011) and Heinson *et al.* (2017). Both 100BPA and 200BPA were combined and used for benchmarking. To ensure the quality of this external independent dataset, all positive samples in 100BPA and 200BPA were checked against the BPAgs in Protegen. Any duplicated positive samples were then removed (Supplementary Table S3). Meanwhile, the negative samples in 200BPA were also evaluated to ensure no supporting experimental evidence from the literature to ensure the 'true negative' dataset for ML-based BPAgs prediction. The final curated external independent dataset consisted of 131 positive and 118 negative samples, named 'iBPA.'

Vaxign-ML was compared to four BPAg candidate prediction programs: Vaxign (He *et al.*, 2010); VaxiJen (Doytchinova and Flower, 2007); Heinson–Bowman (Bowman *et al.*, 2011; Heinson *et al.*, 2017); Antigenic (Rahman *et al.*, 2019); and one epitope-based prediction method using IEDB-AR epitope prediction tools (Dhanda *et al.*, 2019). For Vaxign prediction, we used two suggested criteria: surface-exposed proteins (subcellular localization in the cell wall, outer membrane or extracellular space) and adhesin probability >0.51. The recommended cutoff (0.5) was used for BPA prediction by VaxiJen. For Heinson–Bowman method, a nested cross-validated SVM prediction model was tested with the iBPA dataset annotated by the top 10 significant biological properties (Heinson *et al.*, 2017). Vaxign-ML had major differences compared to the Heinson–Bowman method including the quality of training

Table 1. N5CV evaluation metrics of five ML algorithms trained using four data resampling methods

Original data				Undersampled				Oversampled				Balanced			
AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC	AUROC	AUPRC	WF1	MCC
LR															
0.95	0.77	0.93	0.60	0.94	0.75	0.93	0.60	0.95	0.78	0.93	0.63	0.95	0.8	0.91	0.58
SVM															
0.95	0.84	0.96	0.76	0.95	0.77	0.92	0.58	0.95	0.8	0.94	0.67	0.95	0.87	0.87	0.03
KNN															
0.91	0.67	0.93	0.59	0.89	0.6	0.83	0.43	0.84	0.58	0.83	0.41	0.76	0.6	0.87	0.31
RF															
0.96	0.87	0.96	0.76	0.94	0.75	0.93	0.58	0.94	0.79	0.95	0.69	0.94	0.91	0.87	0.06
XGB															
0.96	0.87	0.96	0.79	0.95	0.84	0.95	0.71	0.95	0.83	0.95	0.72	0.96	0.93	0.95	0.72

Note: The underlined denotes the highest values for each metric.

AUROC, area under receiver operating characteristic curve; AUPRC, area under precision recall curve; WF1, weighted F1 score; MCC, Matthew's correlation coefficient.

data, selection of ML algorithms, data resampling methods, and annotated features. For Antigenic, the default settings and cutoff values were used to call BPAgs from the iBPA dataset. An epitope-based prediction method was implemented by thresholding the percentile ranking of the epitope frequency in the iBPA dataset, compared to 10 000 randomly selected background proteins. The epitope frequency of a protein was calculated by summing the top 1% predicted MHC-I restricted epitopes and top 10% predicted MHC-II restricted epitopes across the reference set alleles (Greenbaum *et al.*, 2011; Weiskopf *et al.*, 2013) using the IEDB-AR epitope prediction tools (Dhanda *et al.*, 2019). A percentile ranking threshold of 58% was used after optimization for the true and false positive rate. Proteins in the iBPA dataset with epitope frequency above 58% percentile rank compared to the random background were considered to have significant immunogenic potential.

2.4 Performance evaluation

The receiver operating characteristics (ROC) curve, precision–recall (PR) curve, weighted F1 score (WF1), and Matthew's correlation coefficient (MCC) were computed for both N5CV and LOPOV. An additional evaluation was performed for the LOPOV. For the benchmarking of the Vaxign-ML with iBPA, the PR, WF1, and MCC metrics were calculated. Finally, the protegenicity scores of 20 proteins from five *Mycobacterium tuberculosis* (MTB) vaccines undergoing clinical trials and the licensed DTP (*Corynebacterium diphtheriae*, *Clostridium tetani*, and *Bordetella pertussis*) combination vaccine were calculated.

3 Results

3.1 Effect of data resampling strategies on the classification

All ML classification algorithms performed worse when trained on under- and oversampled data compared to original or balanced data (Table 1). When evaluating the performance of different data resampling strategies based on area under ROC (AUROC), almost all ML classification algorithms had high values of AUROC (ranging from 0.89 to 0.96) except KNN (AUROC = 0.76). Since the data resampling step was only performed on the training data during the N5CV but not the testing data, the area under precision recall curve (AUPRC), WF1, and MCC metrics were less prone to the imbalanced classes in the data than AUROC. All ML algorithms trained on under- and oversampled data consistently had lower AUPRC, WF1, and MCC. The balanced data did not significantly improve the performance of the ML algorithms used in this study. Furthermore, the MCC values of the SVM and RF trained on balanced data were dramatically reduced, which indicated high degrees of overfitting in these two ML models. KNN algorithm was more sensitive to the changes in sample class ratio because all four metrics of the models trained on under-, oversampled, and balanced data were lower than the original data. Although the LR and XGB trained on balanced data had slightly higher AUPRC, these two ML models had lower WF1 and MCC when trained on original data. Therefore, balancing the positive and negative samples did not significantly improve the BPAg prediction.

3.2 The best performance by XGB trained on original data in N5CV and LOPOV

In N5CV, the XGB model consistently had the highest performance compared to the other four ML algorithms when trained on four different data resampling methods (Table 1). Three models, including XGB trained on original data (XGB-original) and balanced data (XGB-balance), and RF trained on original data had the highest AUROC curve. XGB-original had the highest WF1 and MCC while XGB-balance had the highest AUPRC. The N5CV results of five ML prediction models to discriminate two sets of randomly selected dissimilar non-BPAgs were approximately equivalent to random prediction (Supplementary Fig. S6 and Table S5). The discriminative power of the current BPAg prediction pipeline was indeed dependent on the immunogenic potential in the Protegen dataset rather than sequence dissimilarity.

Both XGB-original and XGB-balance were evaluated with the LOPOV and had similar AUROCs for the 10 pathogens held out in LOPOV and the average of these pathogens (Fig. 2a and b). However, the XGB-original had higher average AUPRC (Fig. 2c and d), WF1, and MCC (Supplementary Table S4). The ROC and PRC curves of all ML models were plotted (Supplementary Figs S2–S5). The best performing XGB-original was selected as the final BPAg prediction model and termed Vaxign-ML for benchmarking. The online version is available at http://www.violinet.org/vaxign/vaxign-ml. A Docker standalone Vaxign-ML version is available at https://hub.docker.com/r/e4ong1031/vaxign-ml and source code is available at https://github.com/VIOLINet/Vaxign-ML-docker. It is recommended that the users search in the Protegen database to identify known BPAgs before predicting novel BPAgs with Vaxign-ML.

3.3 Biological and physicochemical features in Vaxign-ML

The mRMR feature selection and hyperparameter optimization steps in Vaxign-ML suggested an optimal set of 180 features (Supplementary Table S6). The biological features, including subcellular localization, adhesin probability, transmembrane helix, and immunogenicity score, are frequently used in filtering-based vaccine prediction programs (e.g. NERVE and Vaxign). However, in Vaxign-ML, these features only accounted for 11.4% of the importance in the final XGB model (Fig. 3).

The pathogen's Gram(±) stain was excluded from the Vaxign-ML due to its lack of contribution to the outcome. Although the physicochemical properties are often difficult to be interpreted

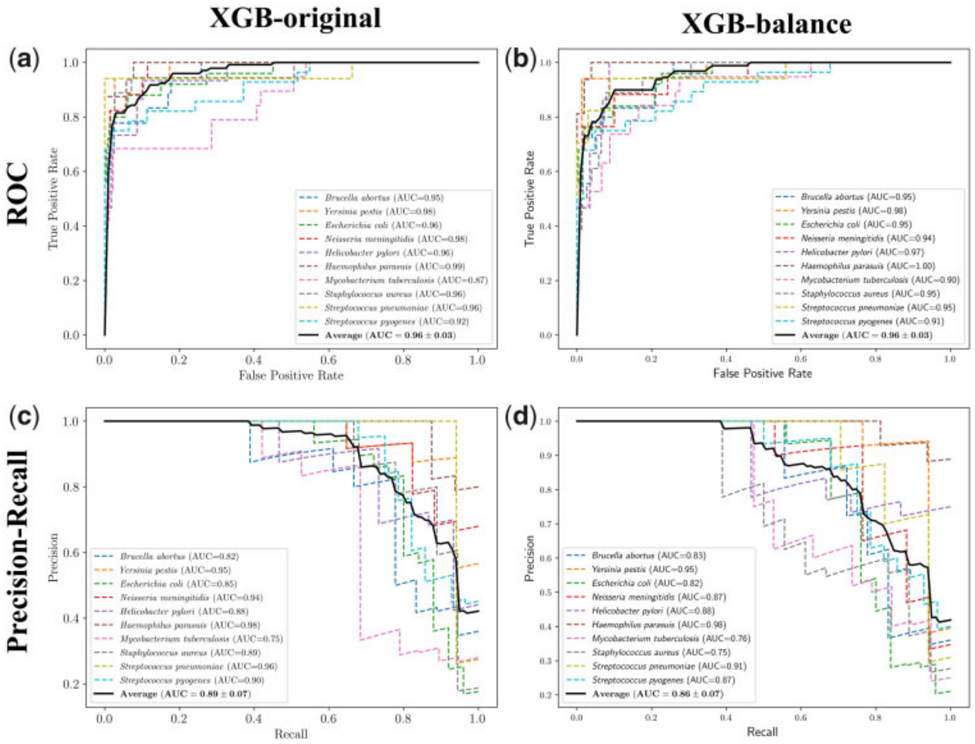

Fig. 2. LOPOV of the two best performing models, XGB-original and XGB-balance. The top and bottom row plotted the ROC and PR curves, respectively. Ten pathogens were tested in the LOPOV (color dash lines), as shown in the legend. The average of the ROC and PR curves were also plotted (black line). The average AUROC curve of both XGB-original (**a**) and XGB-balance (**b**) performed equally well. However, XGB-original had a higher average AUPRC than the XGB-balance. In addition, XGB-balance had lower performance when tested on the *M.tuberculosis*, *S.aureus* while XGB-original had lower performance only on the former one. (Color version of this figure is available at *Bioinformatics* online.)

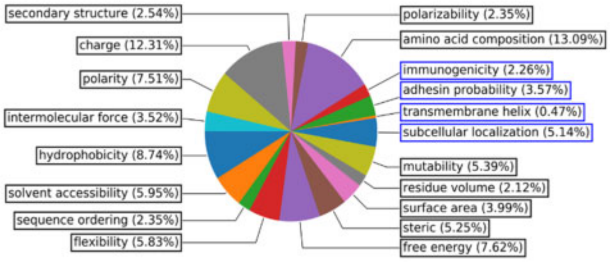

Fig. 3. Categories of features and its weight in the Vaxign-ML model. Features in blue and black boxes are biological and physicochemical features, respectively. (Color version of this figure is available at *Bioinformatics* online.)

during vaccine design, these features accounted for 88.6% of the importance. In particular, amino acid composition (13.1%), charge (12.3%), hydrophobicity (8.7%), free energy (7.6%), and polarity (7.5%) were the top five important categories of the physicochemical properties in the Vaxign-ML model (Fig. 3).

3.4 Benchmarking Vaxign-ML

Table 2. Benchmarking of Vaxign-ML compared to three publicly available protective antigen prediction programs (Vaxign, VaxiJen and Antigenic), the Heinson–Bowman SVM-based method and the epitope-based method

	Recall	Precision	WF1	MCC
Vaxign-ML	0.81	0.75	0.76	0.51
Heinson–Bowman	0.72	0.69	0.68	0.37
VaxiJen	0.69	0.68	0.66	0.32
Vaxign	0.32	0.79	0.56	0.27
Antigenic	0.50	0.52	0.49	−0.02
Epitope-based	0.63	0.65	0.62	0.24

WF1, weighted F1 score; MCC, Matthew's correlation coefficient.

The final constructed model, Vaxign-ML, was benchmarked and compared to currently publicly available BPAg candidate prediction programs (Vaxign, VaxiJen, and Antigenic), Heinson–Bowman and epitope-based method. The performance of these programs was evaluated based on the iBPA dataset described in Section 2.3. Vaxign-ML had the highest performance in three out of four metrics (recall = 0.81, WF1 = 0.76 and MCC = 0.51) among all methods (Table 2). The ROC and PRC curves of Vaxign-ML tested on the iBPA dataset were also plotted (Supplementary Fig. S7). The rule-based Vaxign program had a higher precision value (0.79) than the Vaxign-ML (0.75), likely due to the more restrictive rules in the Vaxign program. Vaxign-ML also out-performed the ML-based Heinson–Bowman method, suggesting that the enhancement of Vaxign-ML in terms of training data quality, annotated features, and selected ML algorithm improves the BPAg prediction. Finally, the epitope-based prediction method had lower performance than Vaxign-ML, Heinson–Bowman, and VaxiJen across all four metrics in the context of BPAg prediction.

3.5 Vaxign-ML predicting current clinical trial or licensed vaccines

As a final validation, Vaxign-ML was used to calculate and rank the corresponding protegenicity scores of five clinical trial MTB vaccines and one licensed DTP vaccine (*C.diphtheriae*, *C.tetani*, and *B.pertussis*) (Table 3). The protegenicity score was the percentile rank score generated by Vaxign-ML trained on the entire original data. A total of 20 proteins were included in these six vaccines, and all of them had a predicted protegenicity score of over 90%. In other words, these 20 proteins were ranked in the top 10% best BPAg candidates by the Vaxign-ML.

Table 3. Vaxign-ML prediction of five MTB vaccines currently in clinical trial and one licensed DPT vaccine

Vaccine	Protein	Protegenicity score (%)
M.tuberculosis		
H1, H4, H56	Ag85B	95.21
H1, H56	ESAT-6	94.89
H4	EsxH	90.91
H56	Rv2660	91.23
M72	PPE18	92.05
	PepA	94.28
ID93	EsxW	90.95
	PPE42	91.89
	EsxV	91.53
	Rv1813	91.09
B.pertussis		
Pertussis vaccine	Pertussis toxin subunit 1	94.07
	Pertussis toxin subunit 2	91.53
	Pertussis toxin subunit 3	90.91
	Pertussis toxin subunit 4	91.37
	Pertussis toxin subunit 5	91.62
	Filamentous hemagglutinin	98.9
	Pertactin autotransporter	95.35
	Fimbrial protein	95.99
C.diphtheriae		
Diphtheria vaccine	Diphtheria toxin	97.73
C.tetani		
Tetanus vaccine	Tetanus toxin	99.79

4 Discussion

Overall, Vaxign-ML showed superior performance in BPAg prediction compared to all other BPAg prediction methods. Our study also demonstrated the significance of both biological and physicochemical properties in ML-based RV prediction. Finally, the results of Vaxign-ML highlighted the critical role of physicochemical properties and might have an implication in structural vaccinology.

Our study showed that Vaxign-ML (XGB trained on original data with mRMR feature section and hyperparameter optimization) was the best performing supervised ML classification model with an unbiased N5CV and LOPOV validations. The LOPOV validation also assessed how well the model could predict BPAgs when encountering a new emerging pathogen. The benchmarking of Vaxign-ML using a curated external independent dataset suggested the superior performance of Vaxign-ML to its predecessor with the highest recall, WF1 score, and Matthew's correlation coefficient. Notably, the iBPA dataset was derived and curated from the VaxiJen program. Vaxign-ML was trained on the Protegen dataset and did not encounter any samples in the iBPA, and yet Vaxign-ML had better predictive power than VaxiJen. Although the preceding rule-based Vaxign program missed a lot of the potential candidates (recall = 0.32, Table 2), the rule-based RV method had better potential in filtering out non-protective proteins, as demonstrated by the highest precision value among all four programs being studied. A combination of Vaxign-ML followed by a filtering step similar to Vaxign might be a future direction to enhance the predictive performance.

Vaxign-ML is the first RV method that incorporates both biological and physicochemical properties. Historically, the biological and physicochemical features had been treated as two isolated silos in the field of BPAgs prediction. Several ML RV studies predicted BPAgs based on the physicochemical properties of the input proteins (Doytchinova and Flower, 2007; Magnan *et al.*, 2010; Rahman *et al.*, 2019). In this paper, all the individual physicochemical features were grouped into 15 categories for a better interpretation (Fig. 3). Mayers *et al.* (2003) reported that known protein vaccine antigens had distinct characteristics in amino acid composition, hydrophobicity, flexibility, and mutability, which accounted for 13.1%, 8.7%, 5.8%, and 5.4% of the Vaxign-ML feature importance, respectively. Polarity (7.5%) and charge (12.3%) had an important implication in vaccine design. Studies showed that antibody–antigen interfaces are likely polar (Hebditch and Warwicker, 2019). However, highly negatively charged vaccines often possess limited cell uptake ability, whereas highly positively charged vaccines exert significant cytotoxicity (Zhang *et al.*, 2018). Positively charged nanoparticles induce a more robust and systemic antibody response in a recent nano-based vaccine delivery study (Fromen *et al.*, 2015). Finally, free energy is an essential factor in the structural design of chimeric subunit vaccine (Nazarian *et al.*, 2012) as well as describing the binding between epitope and major histocompatibility complex (Patronov and Doytchinova, 2013).

The significance of biological property profiles in BPAgs (Ong *et al.*, 2017) had been utilized by both rule-based RV programs (He *et al.*, 2010; Jaiswal *et al.*, 2013; Rizwan *et al.*, 2017; Vivona *et al.*, 2006) and supervised ML BPAg classifications (Bowman *et al.*, 2011; Heinson *et al.*, 2017). Vaxign-ML took a substantial consideration into the biological properties, including subcellular localization, adhesin probability, and immunogenicity. However, some biological features (e.g. Gram stain and transmembrane helix) might not be significantly associated with protectiveness and were considered for practical reasons (pathogen characterization and efficacy in recombinant protein isolation) (He *et al.*, 2010). The biological features of the protein sequences in the training data are predictions and are dependent on the performance of the corresponding bioinformatics tools. Although some of these bioinformatics tools already included physicochemical properties in the prediction pipelines, these properties were utilized to address specific scientific questions (e.g. subcellular location prediction, signal peptide). In Vaxign-ML, these biological features of attributed to 11.4% of the importance in BPAg prediction and could be the key factor leading to better prediction performance by Vaxign-ML compared to VaxiJen and Antigenic.

Currently, Vaxign-ML does not consider the epitopes and structure in the prediction model. The comparison of Vaxign-ML and the epitope-based method, which was benchmarked using the iBPA dataset (Table 2), showed that Vaxign-ML had better BPAg prediction than the epitope-based method. Epitope prediction does not necessarily correlate with the immune protection due to the host diversity, amino acid properties, location of epitope and the coevolution between pathogen and host immune system (Halling-Brown *et al.*, 2008, 2009). Undoubtedly, epitopes still play a role in antibody and cell-mediated immunity, and the integration of BPAg prediction with epitope identification and antigen structural analysis will be investigated in the future.

Acknowledgements

We would like to express our deepest appreciation to Professors Denise Kirschner, Kayvan Najarian, Yang Zhang, Zhenhua Yang for their contribution in study design and discussion.

Funding

This work was supported by the National Institutes of Health (NIH) - National Institute of Allergy and Infectious Diseases (NIAID) [1R01AI081062].

Conflict of Interest: none declared.

References

Bowman,B.N. *et al.* (2011) Improving reverse vaccinology with a machine learning approach. *Vaccine*, **29**, 8156–8164.

Chawla,N.V. *et al.* (2002) SMOTE: synthetic minority over-sampling technique Nitesh. *J. Artif. Intell. Res.*, **16**, 321–357.

Chen,T. and Guestrin,C (2016) XGBoost: a scalable tree boosting system. In: *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, August 13–17, pp. 785–794.

Chou,K.-C. (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.*, **278**, 477–483.

Dalsass,M. *et al.* (2019) Comparison of open-source reverse vaccinology programs for bacterial vaccine antigen discovery. *Front. Immunol.*, **10**, 1–12.

Dhanda,S.K. *et al.* (2019) IEDB-AR: immune epitope database—analysis resource in 2019. *Nucleic Acids Res.*, **47**, W502–506.

Ding,C. and Peng,H. (2003) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 185–205.

Doytchinova,I.A. and Flower,D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics*, **8**, 4.

Dubchak,I. *et al.* (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **92**, 8700–8704.

Feng,Z.P. and Zhang,C.T. (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.*, **19**, 269–275.

Fleri,W. *et al.* (2017) The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Front. Immunol.*, **8**, 1–16.

Flower,D.R. *et al.* (2010) Computer aided selection of candidate vaccine antigens. *Immunome Res.*, **6**, S1–16.

Folaranmi,T. *et al.* (2015) Use of serogroup B meningococcal vaccines in persons aged >/=10 years at increased risk for serogroup B meningococcal disease: recommendations of the advisory committee on immunization practices, 2015. *Morb. Mortal. Wkly. Rep.*, **64**, 608–612.

Fromen,C.A. *et al.* (2015) Controlled analysis of nanoparticle charge on mucosal and systemic antibody responses following pulmonary immunization. *Proc. Natl. Acad. Sci.*, **112**, 488–493.

Goodswen,S.J. *et al.* (2013) A novel strategy for classifying the output from an *in silico* vaccine discovery pipeline for eukaryotic pathogens using machine learning algorithms. *BMC Bioinformatics*, **14**, 315.

Greenbaum,J. *et al.* (2011) Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes. *Immunogenetics*, **63**, 325–335.

Halling-Brown,M. *et al.* (2008) Are bacterial vaccine antigens T-cell epitope depleted? *Trends Immunol.*, **29**, 374–379.

Halling-Brown,M. *et al.* (2009) Proteins accessible to immune surveillance show significant T-cell epitope depletion: implications for vaccine design. *Mol. Immunol.*, **46**, 2699–2705.

He,Y. and Xiang,Z. (2012) Bioinformatics analysis of bacterial protective antigens in manually curated Protegen database. *Procedia Vaccinol.*, **6**, 3–9.

He,Y. *et al.* (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. *J. Biomed. Biotechnol.*, **2010**, 1–15.

Hebditch,M., and Warwicker,J. (2019) Web-based display of protein surface and pH-dependent properties for assessing the developability of biotherapeutics. *Sci. Rep*, **9**, 1.

Heinson,A.I. *et al.* (2017) Enhancing the biological relevance of machine learning classifiers for reverse vaccinology. *Int. J. Mol. Sci.*, **18**, 312.

Hossain,M.S. *et al.* (2017) Computational identification and characterization of a promiscuous T-cell epitope on the extracellular protein 85B of *Mycobacterium* spp. for peptide-based subunit vaccine design. *Biomed. Res. Int.*, **2017**, 1–14.

Jaiswal,V. *et al.* (2013) Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. *BMC Bioinformatics*, **14**, 211.

Jimenez de Bagues,M.P. *et al.* (1994) Vaccination with *Brucella abortus* rough mutant RB51 protects BALB/c mice against virulent strains of *Brucella abortus*, *Brucella melitensis*, and *Brucella ovis*. *Infect. Immun.*, **62**, 4990–4996.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Lin,Z. and Pan,X.M. (2001) Accurate prediction of protein secondary structural content. *Protein J.*, **20**, 217–220.

Magnan,C.N. *et al.* (2010) High-throughput prediction of protein antigenicity using protein microarray data. *Bioinformatics*, **26**, 2936–2943.

Mayers,C. *et al.* (2003) Analysis of known bacterial protein vaccine antigens reveals biased physical properties and amino acid composition. *Comp. Funct. Genomics*, **4**, 468–478.

Moise,L. *et al.* (2015) iVAX: an integrated toolkit for the selection and optimization of antigens and the design of epitope-driven vaccines. *Hum. Vaccin. Immunother.*, **11**, 2312–2321.

Navarro-Quiroz,E. *et al.* (2018) Prediction of Epitopes in the Proteome of *Helicobacter pylori*. *Glob. J. Health Sci.*, **10**, 148.

Nazarian,S. *et al.* (2012) An in silico chimeric multi subunit vaccine targeting virulence factors of enterotoxigenic Escherichia coli (ETEC) with its bacterial inbuilt adjuvant. *J Microbiol Methods*, **90**, 36–45.

Ong,E. *et al.* (2017) Identification of new features from known bacterial protective vaccine antigens enhances rational vaccine design. *Front. Immunol.*, **8**, 1–11.

Ong,S.A.K. *et al.* (2007) Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics*, **8**, 300–314.

Patronov,A. and Doytchinova,I. (2013) T-cell epitope vaccine design by immunoinformatics. *Open Biol*, **3**.

Pearson,W.R. (2013) An introduction to sequence similarity ('homology') searching. *Curr. Protoc. Bioinform,* 42, 3.1.1–3.1.8.

Pedregosa,F. *et al.* (2012) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Petersen,T.N. *et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*, 8, 785–786.

Pizza,M. (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science (80-)*, **287**, 1816–1820.

Rahman,M.S. *et al.* (2019) Antigenic: an improved prediction model of protective antigens. *Artif. Intell. Med.*, **94**, 28–41.

Rappuoli,R. (2000) Reverse vaccinology. *Curr. Opin. Microbiol.*, **3**, 445–450.

Rizwan,M. *et al.* (2017) VacSol: a high throughput *in silico* pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. *BMC Bioinformatics*, **18**, 1–7.

Sachdeva,G. *et al.* (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics*, **21**, 483–491.

Singh,R. *et al.* (2016) Immunoprotective efficacy of *Acinetobacter baumannii* outer membrane protein, FilF, predicted *in silico* as a potential vaccine candidate. *Front. Microbiol.*, **7**, 158.

Sokal,R.R. and Thomson,B.A. (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am. J. Phys. Anthropol.*, **129**, 121–131.

The UniProt Consortium. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D193–197.

Vernikos,G. and Medini,D. (2014) Bexsero H chronicle. *Pathog. Glob. Health*, **108**, 305–311.

Vivona,S. *et al.* (2006) NERVE: new enhanced reverse vaccinology environment. *BMC Biotechnol.*, **6**, 35.

Weiskopf,D. *et al.* (2013) Comprehensive analysis of dengue virus-specific responses supports an HLA-linked protective role for CD8+ T cells. *Proc. Natl. Acad. Sci. USA*, **110**, E2046–2053.

WHO. (2014) *MDG 6: Combat HIV/AIDS, Malaria and Other Diseases*. World Health Organization.

Yang,B. *et al.* (2011) Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res.*, **39**, 1073–1078.

Yu,N.Y. *et al.* (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.

Zhang,R. *et al.* (2018) Peptide amphiphile micelle vaccine size and charge influence the host antibody response. *ACS Biomater. Sci. Eng.,* **4**, 2463–2472.