

**Universidade Federal de Minas Gerais**  
**Instituto de Ciências Biológicas**  
**Departamento de Genética, Ecologia e Evolução**  
**Programa de Pós-Graduação em Genética**

Igor Kelvyn Cavalcante Lobo

**APPLICATIONS OF MACHINE LEARNING TO REVERSE VACCINOLOGY:  
PREDICTION OF VACCINE CANDIDATES BASED ON PARASITE GENOMIC  
DATA**

Belo Horizonte  
2023

Igor Kelvyn Cavalcante Lobo

**APPLICATIONS OF MACHINE LEARNING TO REVERSE VACCINOLOGY:  
PREDICTION OF VACCINE CANDIDATES BASED ON PARASITE GENOMIC  
DATA**

Tese apresentada ao Programa de Pós-Graduação em Genética da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Doutor em Genética.

Orientador: Prof. Dr. Francisco Pereira Lobo

Belo Horizonte  
2023

043

Lobo, Igor Kelvyn Cavalcante.

Applications of machine learning to reverse vaccinology: prediction of vaccine candidates based on parasite genomic data [manuscrito] / Igor Kelvyn Cavalcante Lobo. – 2023.

109 f. : il. ; 29,5 cm.

Orientador: Francisco Pereira Lobo.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Genética.

1. Genética. 2. Genoma. 3. Antígenos. 4. Vacinologia. 5. Aprendizado de Máquina. I. Lobo, Francisco Pereira. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. III. Título.

CDU: 575



UNIVERSIDADE FEDERAL DE MINAS GERAIS  
**Instituto de Ciências Biológicas**  
Programa de Pós-Graduação em Genética

### ATA DE DEFESA DE TESE

|                             |                                                       |
|-----------------------------|-------------------------------------------------------|
| ATA DA DEFESA DE TESE       | 176/2023<br>entrada<br>1º/2018<br>CPF: 883.850.352-49 |
| Igor Kelvyn Cavalcante Lobo |                                                       |

Às oito horas e trinta minutos do dia **05 de julho de 2023**, reuniu-se a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: "**Applications of machine learning to reverse vaccinology: prediction of vaccine candidates based on parasite genomic data**", requisito para obtenção do grau de Doutor em **Genética**. Abrindo a sessão, o Presidente da Comissão, **Francisco Pereira Lobo**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

| Prof./Pesq.                  | Instituição | CPF            | Indicação |
|------------------------------|-------------|----------------|-----------|
| Francisco Pereira Lobo       | UFMG        | 012.273.736-94 | APROVADO  |
| Glória Regina Franco         | UFMG        | 623.387.496-34 | APROVADO  |
| Frederico Marianetti Soriani | UFMG        | 249.486.428-35 | APROVADO  |
| Siomar de Castro Soares      | UFTM        | 056.951.826-11 | APROVADO  |
| Tetsu Sakamoto               | UFRN        | 330.246.028-79 | APROVADO  |

Pelas indicações, o candidato foi considerado: APROVADO

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

**Belo Horizonte, 05 de julho de 2023.**

Francisco Pereira Lobo-UFMG

Glória Regina Franco-UFMG

Frederico Marianetti Soriani-UFMG

Siomar de Castro Soares- UFTM



Documento assinado eletronicamente por **Tetsu Sakamoto, Usuário Externo**, em 06/07/2023, às 09:50, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Gloria Regina Franco, Professora do Magistério Superior**, em 06/07/2023, às 15:02, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Frederico Marianetti Soriani, Professor do Magistério Superior**, em 07/07/2023, às 10:18, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Siomar de Castro Soares, Usuário Externo**, em 07/07/2023, às 11:15, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



Documento assinado eletronicamente por **Francisco Pereira Lobo, Professor do Magistério Superior**, em 10/07/2023, às 00:25, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).

---



A autenticidade deste documento pode ser conferida no site [https://sei.ufmg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **2440128** e o código CRC **C21FA150**.

---



UNIVERSIDADE FEDERAL DE MINAS GERAIS

***Instituto de Ciências Biológicas***

*Programa de Pós-Graduação em Genética*

## **FOLHA DE APROVAÇÃO**

**"Applications of machine learning to reverse vaccinology: prediction of vaccine candidates based on parasite genomic data"**

**Igor Kelvyn Cavalcante Lobo**

Tese aprovada pela banca examinadora constituída pelos Professores:

Francisco Pereira Lobo

UFMG

Glória Regina Franco

UFMG

Frederico Marianetti Soriani

UFMG

Siomar de Castro Soares

UFTM

Tetsu Sakamoto

UFRN

Belo Horizonte, 05 de julho de 2023.



Documento assinado eletronicamente por **Tetsu Sakamoto, Usuário Externo**, em 06/07/2023, às 09:50, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Gloria Regina Franco, Professora do Magistério Superior**, em 06/07/2023, às 15:02, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Frederico Marianetti Soriani, Professor do Magistério Superior**, em 07/07/2023, às 10:18, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Siomar de Castro Soares, Usuário Externo**, em 07/07/2023, às 11:16, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Francisco Pereira Lobo, Professor do Magistério Superior**, em 10/07/2023, às 00:25, conforme horário oficial de Brasília, com fundamento no art. 5º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site [https://sei.ufmg.br/sei/controlador\\_externo.php?acao=documento\\_conferir&id\\_orgao\\_acesso\\_externo=0](https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0), informando o código verificador **2440224** e o código CRC **ABE6299C**.

Dedico este trabalho aos meus avós, pais de criação, Maria Ivanilde e Ismael de Souza (in memoriam), que sonharam com isso mais do que eu próprio, que com sua simplicidade me ensinaram o amor e a dedicação. É tudo por vocês e para vocês.

## **AGRADECIMENTOS**

Ao Programa de Pós-graduação em Genética da UFMG por viabilizar minha formação, sobretudo, à Raissa Campos, por ser sempre atenciosa e prestativa com todo mundo.

Agradeço, ainda, à CAPES pela bolsa de doutorado concedida, pela prorrogação em meio ao caos que foi a pandemia e pela bolsa de doutorado-sanduíche (PrINT-CAPES), através da qual realizei o sonho de estudar fora do país.

Ao meu orientador, Prof. Dr. Francisco Lobo, por ter acreditado em mim, por ter me recebido tão bem, pela estrutura que me forneceu, pelos conhecimentos, conversas descontraídas, pelas portas que me abriu. Chico, você tem um lugar especial na minha formação profissional e na pessoa que serei daqui para a frente.

Ao Prof. Dr. Felipe Campelo, por ter me recebido na Aston University durante o doutorado-sanduíche, pelo suporte na minha adaptação à cidade de Birmingham (Inglaterra), por todo o conhecimento compartilhado comigo, pelos direcionamentos no trabalho, pelos momentos de leveza sempre ao fim do expediente, pelas caminhadas nos parques e pints. Felipe, você também ocupa um espaço especial no meu caminho.

Aos amigos do Laboratório de Algoritmos em Biologia, o LAB. Aqui, somente vou me conter quanto a citar os nomes das pessoas, porque não quero ser injusto esquecendo alguém. Cabe a vocês todos a minha mais sincera gratidão por me receberem tão carinhosamente desde o primeiro dia. Construímos uma amizade que, certamente, ultrapassa os muros da universidade. Obrigado por terem me ajudado com o trabalho mas, principalmente, com a manutenção da minha saúde mental! Obrigado pelos mais de 4 anos de reuniões discutindo ciência, mas discutindo também quem era o impostor em Among Us, pelos desenhos achincalhados no Gartic, pelas inúmeras bobagens! Obrigado pela mais incrível rede de apoio que criamos e que eu levo para a vida!

Aos colegas Mestrandos e Doutorandos do projeto 'Trem Bão é Ciência' e demais colegas da Pós-graduação. Travamos longas e duras batalhas nos últimos anos. Enfrentamos com o Trem a desinformação de um governo atroz, distribuímos conhecimento online e pessoalmente, protestamos por nós e pelos nossos, tentamos nos manter vivos e minimamente bem durante a pandemia. Foi um prazer estar nas trincheiras com vocês.

Aos amigos que já me acompanhavam antes dessa aventura: Felipe, Alessandra, Hallana, Reji e Janilson. Obrigado, meus amores, por todo o apoio do mundo em absolutamente tudo! Desde que nos separamos, eu disse que os levaria comigo sempre, pois cá estão vocês. Ninguém solta a mão de ninguém e eu volto.

Aos amigos do ensino médio, que eu também não citarei nomes, muito obrigado! Cada um de vocês deixou algo valioso por aqui, que me motivou e me fez chegar mais longe. Eu volto e quero a nossa confraternização de sempre.

À minha melhor amiga e confidente Érica Siqueira. Obrigado por tanto! Obrigado por uma longa vida de conversas que me solidificaram, pelo ombro amigo, pelos momentos de reflexão, por você estar sempre acreditando em mim, quando muitas vezes eu mesmo falhei nisso. Obrigado pelo dia do seu casamento (você sabe o porquê)! Obrigado!

À minha família. Obrigado por estarem incondicionalmente comigo! Obrigado por serem minha referência de resiliência! Obrigado por termos suportado juntos a nossa perda irreparável e por termos comemorado cada vitória do nosso jeito! Obrigado por me acolherem em cada frustração e por terem celebrado comigo cada conquista! A vida com vocês tem mais significado, mais cor e motivação. É maravilhoso ter um lugar para voltar e esse lugar ser vocês.

## RESUMO

Na era da ciência de dados, com a abundância de genomas atualmente disponíveis, um tipo de análise pós-genômica é a detecção de antígenos protetivos (AgPs). Porém, iniciativas para organizar AgPs à medida que emergem da literatura são escassas. Detectar AgPs é o principal objetivo de um campo de pesquisa chamado Vacinologia Reversa (VR). Existem várias ferramentas computacionais para VR. Entretanto, estes programas apresentam algumas limitações. Assim, aqui apresentamos PADA-ONE, um novo banco de dados para hospedar AgPs e demais proteínas imunologicamente relevantes (PIRs), bem como metadados associados. Adicionalmente, apresentamos MARVL, uma pipeline para treinar um ensemble de modelos para prever AgPs em espécies bacterianas e outros patógenos. PADA-ONE é composto por 1) um componente preditivo, apto a atribuir probabilidades a títulos de artigos quanto a seu potencial para descrever AgPs; 2) um banco de dados MySQL back-end; 3) e uma interface front-end. Nosso componente preditivo foi treinado usando títulos de artigos apresentando AgPs, que foram manualmente curados por várias iniciativas independentes. Como exemplos negativos, usamos títulos de artigos do Immune Epitope Database (IEDB). Então, aplicamos o modelo em um conjunto de, aproximadamente, 32 milhões de entradas do PubMed. Para cada entrada, recuperamos também metadados associados, quando disponíveis. PADA-ONE contém mais de 500.000 entradas e já encontramos um crescente grupo de PIRs, compreendendo 30 AgPs, 18 atenuadores, 10 não-AgPs e 2 boosters. Portanto, PADA-ONE fornece uma maneira fácil de acessar uma crescente fonte de PIRs, incluindo dados existentes de outros bancos de dados. Uma vez que as buscas com PADA-ONE resultaram em novos AgPs e não-AgPs, usamos então estas proteínas para 1) analisar dissimilaridade como indicador para distinguir vacinas de não-vacinas; 2) e validar o MARVL. Primeiramente, treinamos os modelos MARVL utilizando sequências proteicas bacterianas do Protegen como exemplos positivos. Como classe negativa, utilizamos proteínas dissimilares, adicionando novos aspectos ao processo, como splitting de sequências baseado em similaridade durante o treinamento, engenharia de atributos e classificação de uma classe. Observamos que usar exclusivamente dissimilaridade como indicador para distinguir vacinas de não-vacinas para criar uma classe negativa de treinamento não apresenta suporte por evidência. Entretanto, isto não é evidência decisiva, somente uma indicação de ser uma estratégia ineficiente. Uma

vez treinados, aplicamos nossos modelos em nosso conjunto de teste e mesmo quando testados isoladamente, alguns deles apresentaram performance robusta. Quando unidos em um ensemble, MARVL foi comparado com programas estado-da-arte e se mostrou capaz de superá-los na maioria das métricas de performance, quando validado utilizando sequências resultantes das buscas com PADA-ONE. Portanto, PADA-ONE e MARVL são novidades que contribuem com a comunidade científica em termos de organizar e sistematizar as buscas por novas PIRs e metadados associados, bem como predizer AgPs e não-AgPs melhor do que ferramentas existentes, respectivamente.

Palavras-chave: aprendizado de máquina; vacinologia reversa; AgPs.

## ABSTRACT

In the data-driven science era, with a wealth of genomes currently available, one type of post-genomic analysis is the detection of protective antigens (PAGs), but initiatives to organize PAGs as they emerge from the literature are scarce. Detecting PAGs is the main purpose of a research field named Reverse Vaccinology (RV). There are several RV computational tools. However, those programs present some limitations. Thus, here we present PADA-ONE, a new database to host PAGs and other immunologically relevant proteins (IRPs), as well as associated metadata. Additionally, we introduce MARVL, a pipeline to train an ensemble of models to predict PAGs for bacterial species and other pathogens. PADA-ONE is composed of 1) a predictive component, able to assign probabilities to paper titles regarding their potential to describe PAGs; 2) a back-end MySQL database; and 3) a front-end interface. Our predictive component was trained using titles of papers presenting PAGs that were manually curated by several independent initiatives. As negative instances, we used paper titles from the Immune Epitope Database (IEDB). Then, we deployed the model on a set of approximately 32 million PubMed entries. For each entry, we also retrieved associated metadata, when available. PADA-ONE contains more than 500,000 entries and we have already found a growing set of IRPs, comprising 30 PAGs, 18 attenuators, 10 non-PAGs, and two boosters. Therefore, PADA-ONE provides an easy way to access an expanding source of IRPs, including the already invaluable data from existing databases. Once PADA-ONE screening has resulted in new bacterial PAGs and non-PAGs, we then used those proteins to 1) perform an analysis of dissimilarity as a proxy to distinguish between vaccines and non-vaccines; 2) and validate MARVL. Firstly, we trained MARVL models using bacterial protein sequences from Protegen as positive instances. As a negative class, we used dissimilar proteins, adding new aspects to the process, such as similarity-aware data splitting in training, feature engineering, and one-class classification. We found that exclusively using dissimilarity as a proxy to distinguish between vaccines and non-vaccines to create a negative class is not supported by evidence. However, this is not decisive evidence, but only an indication of being a poor strategy. Once fitted, we deployed our models on our testing dataset and even when isolatedly tested, some of them presented robust performance. Taken together in an ensemble, MARVL was compared with state-of-the-art tools and was

able to outperform them for most performance metrics, when validated using the sequences resulting from PADA-ONE screening. Therefore, PADA-ONE and MARVL are novelties that contribute to the scientific community in terms of organizing and systematizing the search for new IRPs and metadata associated, as well as predicting PAgS and non-PAgS better than existing tools, respectively.

Keywords: machine-learning; reverse vaccinology; PAgS.

## LISTA DE ILUSTRAÇÕES

### GENERAL INTRODUCTION.....22

**Figure 1.** Development and application of a machine learning model for an image classification example. During the training step, the algorithm processes the extracted features to develop the best model to afterward predict the labels using new images as input..... 26

### Chapter 1. PADA-ONE: an online tool to explore proteins of immunological relevance from vertebrate pathogens with bibliographic, taxonomic, and molecular metadata... 33

**Figure 1.** A) Distribution of probability scores for the first component of PADA-ONE when used to survey approximately 32 million Pubmed entries for articles describing PAGs. Symbols above the bars are the following holdouts: titles of PAG articles from Protegen (positive class - '+') and from IEDB (negative class - 'o'). Black circles represent articles from Protegen describing PAGs from non-bacterial origin (third holdout). Probabilities above 75% are highlighted as light red bars. Please notice that the y-scale is compressed beyond  $1.5 \times 10^6$ , to allow the visualization of the disproportionately high peak located at predicted probability values around 0.7. B) Number of IRPs in PADA-ONE, Protegen and VirMugenDB.....40

**Figure S1.** Development of the first component of PADA-ONE. A: Database of article titles describing PAGs from Protegen (positive class) and bacterial immunogenic peptides in general from IEDB (negative class). B: Selection of the second holdout (50 PAG-related and 50 IEDB titles) as one of our validation datasets. C: Corpora of most frequent terms. D: Selection of samples for model training (75%) and testing (25%, first holdout). E: Training of the Logistic Regression model. F: Performance evaluation. G: The second holdout (validation dataset) was used to evaluate the classifier performance when considering the approximately 32 million article titles published in PubMed from 1983 to 2022. PAG: Protective antigen; IEDB: Immune Epitope Database.....44

**Figure S2.** PADA-ONE entity-relationship diagram. The database entities are "metadataPub" (publication-related information, such as article title, abstract, year, citations, and other), "classification\_1stLay" (first layer of probability), "classification\_2ndLay" (second layer of probability), "geneIDs\_PMIDs" (geneIDs associated with an article), "curationStatus\_User" (curation status of an article by the user), "geneIDs\_taxInfo\_AccNumb" (information related to taxa and proteins involved in a study), "taxPath" (organism lineage information according to the Taxonomy database) and "taxID\_taxName" (a split version of "taxPath"), "geneIDsPerPMIDs\_Counts" (geneIDs counting per PMID), "taxIDsPerPMIDs\_Counts" (taxIDs counting per PMID)...50

**Figure S3.** Classifier performance and the outcome of the analysis of the most important terms to the classification task. A) ROC Curve of the predictor in the holdout depicted in Figure S1 step E (AUC of 0.940) B) Top-10 most relevant terms for PAG-related and IEDB article titles (PAG-related in blue and IEDB in green).....51

### Chapter 2. MARVL - Machine-learning Assisted Reverse Vaccinology.....54

**Figure 1.** Estimated probability density functions of pairwise alignment similarities (as raw values and as log transformed values) of known vaccines and non-vaccines (gold-standard dataset) to an independent set of known vaccines (Protegen). Red solid line: negative class (experimentally validated non-vaccines); Green dashed line: positive class (experimentally validated vaccines). The vertical gray dotted line highlights the usual 30% similarity threshold. The two distributions cannot be distinguished at the 95% confidence level (Kolmogorov-Smirnov test,  $D = 0.014$ ,  $p = 0.487$ ). Queries that did not return any alignment were considered as having zero similarity and were not included in this figure (See Figure S4)..... 63

**Figure 2.** Correlation plot of the performance of individual models in the testing dataset.

Color intensities/circle sizes are proportional to correlation (blue: positive correlations, red; negative correlations). Models selected to build MARVL are highlighted in red. ocAE\_\*: autoencoder (one-class classifier); if\_\*: isolation forest (one-class classifier); ocSvm\_\*: support vector machine (one-class classifier); bcSvm\_\*: support vector machine (binary classifier); rf\_\*: random forest (binary classifier); \*\_Norm: normalized dataset; \*\_OutNorm: normalized dataset with truncated outliers; \*\_Stand: standardized dataset; \*\_OutStand: standardized dataset with truncated outliers..... 66

**Figure 3.** Visual comparison of performance estimates for MARVL, Vaxi-DL and Vaxign2. In red, MARVL; in green, Vaxi-DL; in blue, Vaxign2. ACC: accuracy; F1: F1 score; MCC: Matthews Correlation Coefficient; NPV: negative predictive value; PCPU: performance criteria for positive-unlabeled classification; PPV: positive predictive value; SENS: sensitivity; SPEC: specificity.....69

**Figure 4.** Top 20 most relevant features for each model that compose MARVL. In green, Biological features; in purple, Physicochemical features; and in orange, Statistical features. Models captured different attributes of protective antigens, with 66 of them not shared across predictors. BCsvm\_N: binary classifier support vector machine (SVM) with normalized dataset; BCsvm\_OS: binary classifier support vector machine (SVM) with standardized and truncated outliers dataset; OCsvm\_S: one-class support vector machine (SVM) with standardized dataset; RF\_ON: binary classifier random forest (RF) with normalized and truncated outliers dataset; and RF\_S: binary classifier random forest (RF) with standardized dataset..... 70

**Figure S1.** Development and validation of MARVL. Top: Development of a vaccine candidate predictor for bacteria, where data sources are publicly available, namely Protegen [21] and Uniprot [22]. Physicochemical and biological features were extracted from the amino acid sequences. Then, the instances were split into training and testing datasets, according to CD-HIT results. Only the positive dataset was used to fit OCCs, while both positive and unlabeled datasets were used to fit BCs. Bottom: Full ensemble models (with no dataset splitting) were then developed to benchmark our method against the most recent tools available in the literature: Vaxign2 [23] and Vaxi-DL [24]. New positive and real negative instances were retrieved from the literature to work as an additional validation dataset at this step. BC: binary classifier; OCC: one-class classifier... 88

**Figure S2.** Correlation between contiguous residues in a protein of 30 residues of size. A. Correlation between the most contiguous residues. B. Correlation between all the second-most contiguous residues. C. Correlation between all the third-most contiguous residues. The calculation of correlations between contiguous residues continues until the established value for lambda. Adapted from [25].....89

**Figure S3.** Distribution of pairwise alignment identities (as raw values and as log-transformed values) of known vaccines and non-vaccines (gold-standard) and an independent set of known vaccines (Protegen). Red solid line: negative class (experimentally validated non-vaccines); Green dashed line: positive class (experimentally validated vaccines); Sequences that did not return any alignment were considered as having zero identity and were not included in this figure (See Figure S4).90

**Figure S4.** Non-alignments in the gold-standard dataset against Protegen instances. In red, gold-standard negative instances. In blue, gold-standard positive instances. A reasonable percentage of non-alignments is observed for both classes of the gold-standard dataset: 23.05% for positives and 19.85% for negatives examples.....91

**Figure S5.** Highest similarity values of the gold-standard dataset to Protegen instances. In red, gold-standard negative instances. In blue, gold-standard positive instances. Not a single example from the gold-standard dataset presented a maximum similarity to Protegen lower than the usual threshold of 30%..... 92

**Figure S6.** ROC curves of binary classifiers (BC) in the testing dataset. A. Random Forest (RF) models. RF with normalized dataset (N) in solid dark blue line; RF with

normalized and truncated outliers dataset (ON) in solid light blue line; RF with standardized dataset (S) in dashed pink line; and RF with standardized and truncated outliers dataset (OS) in dashed red line. B. Support Vector Machine (SVM) models. SVM with normalized dataset (N) in solid dark blue line; SVM with normalized dataset and truncated outliers dataset (ON) in solid light blue line; SVM with standardized dataset (S) in dashed pink line; and SVM with standardized and truncated outliers dataset (OS) in dashed red line. Binary SVMs and RFs have delivered robust models..... 93

## LISTA DE TABELAS

|                                                                                                                                                                                                                                                                                                   |           |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>GENERAL INTRODUCTION.....</b>                                                                                                                                                                                                                                                                  | <b>22</b> |
| <b>Table 1.</b> Concepts of some terms used in ML approaches. Adapted from ERICKSON et al., 2017).....                                                                                                                                                                                            | 23        |
| <b>Chapter 1. PADA-ONE: an online tool to explore proteins of immunological relevance from vertebrate pathogens with bibliographic, taxonomic, and molecular metadata... 33</b>                                                                                                                   |           |
| <b>Supplemental Table 1.</b> Article titles used to train, test, and validate the predictive component of PADA-ONE. Full supplemental table available at Supplemental_Table_1_Chapter1.xlsx.....                                                                                                  | 51        |
| <b>Supplemental Table 2.</b> Uncovered immunologically relevant proteins (IRPs) after manual curation in PADA-ONE. Full supplemental table available at Supplemental_Table_2_Chapter1.xlsx.....                                                                                                   | 52        |
| <b>Chapter 2. MARVL - Machine-learning Assisted Reverse VaccinoLogy.....</b>                                                                                                                                                                                                                      | <b>54</b> |
| <b>Table 1.</b> Comparison of MARVL with Vaxi-DL and Vaxign2 on the PADA-ONE dataset. Reported values follow the convention “estimate ± standard error”. AUC values estimated for Vaxign2 and Vaxi-DL are unreliable due to lack of probability scores and should not be used for comparison..... | 68        |
| <b>Supplemental Table 1.</b> General aspects of the main reverse vaccinology (RV) tools available in the literature. Full supplemental table available at Supplemental_Table_1_Chapter2.xlsx.....                                                                                                 | 93        |
| <b>Supplemental Table 2.</b> Positive, “pseudo-negative”, and gold-standard datasets used to train, test, and validate MARVL. Full supplemental table available at Supplemental_Table_2_Chapter2.xlsx.....                                                                                        | 93        |
| <b>Supplemental Table 3.</b> Description of features extracted in MARVL pipeline. Full supplemental table available at Supplemental_Table_3_Chapter2.xlsx.....                                                                                                                                    | 94        |
| <b>Supplemental Table 4.</b> Summary of the R packages, functions, and parameters/arguments used to train the models. Full supplemental table available at Supplemental_Table_4_Chapter2.xlsx.....                                                                                                | 94        |
| <b>Supplemental Table 5.</b> Performance of all binary classifiers (BCs) and one-class classifiers (OCCs) on the testing dataset. Full supplemental table available at Supplemental_Table_5_Chapter2.xlsx.....                                                                                    | 94        |
| <b>Supplemental Table 6.</b> Feature counts before and after feature reduction. Full supplemental table available at Supplemental_Table_6_Chapter2.xlsx.....                                                                                                                                      | 94        |

## LISTA DE ABREVIATURAS

**ACC** - Accuracy

**AE\_\*** - Autoencoder

**AI** - Artificial Intelligence

**API** - Application Programming Interface

**AUC** - Area Under the Curve

**BC** - Binary Classifiers

**BCSVM\_\*** - Binary-class Support Vector Machine

**BLAST** - Basic Local Alignment Search Tool

**CV** - Cross-validation

**FN** - False Negatives

**FP** - False Positives

**HTTP** - Hypertext Transfer Protocol

**IEDB** - Immune Epitope Database

**IF\_\*** - Isolation Forest

**IRP** - Immunologically Relevant Protein

**LASSO** - Least Absolute Shrinkage and Selection Operator

**MARVL** - Machine-learning Assisted Reverse Vaccinology

**MCC** - Matthews Correlation Coefficient

**MHC** - Major Histocompatibility Complex

**ML** - Machine Learning

**MSE** - Mean Square Error

**NCBI** - National Center for Biotechnology Information

**NPV** - Negative Predictive Value

**OCC** - One-class classification/classifier

**OCSVM\_\*** - One-class Support Vector Machine

**PADA-ONE** - Protective Antigen Database - OnliNe Explorer

**PAg** - Protective Antigen

**PCA** - Principal Component Analysis

**PCPU** - Performance criteria for positive-unlabeled classification

**PMID** - PubMed Identifier

**PPV** - Positive Predictive Value

**PR** - Precision

**RF\_\*** - Random Forest

**ROC** - Receiver Operating Characteristic

**RV** - Reverse Vaccinology

**SENS** - Sensitivity

**SPEC** - Specificity

**SQL** - Structured Query Language

**TN** - True Negatives

**TP** - True Positives

**\*\_N** - Normalized dataset

**\*\_ON** - Normalized dataset with truncated outliers

**\*\_S** - Standardized dataset

\*\_OS - Standardized dataset with  
truncated outliers

## SUMÁRIO

|                                                                                                                                                                                 |           |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>GENERAL INTRODUCTION.....</b>                                                                                                                                                | <b>22</b> |
| The data-driven science era.....                                                                                                                                                | 22        |
| Machine learning - definitions, and perspectives.....                                                                                                                           | 23        |
| Vaccine development: the conventional approach and Reverse Vaccinology.....                                                                                                     | 27        |
| ML in reverse vaccinology.....                                                                                                                                                  | 29        |
| Challenges for new ML-based RV programs.....                                                                                                                                    | 30        |
| <b>Chapter 1. PADA-ONE: an online tool to explore proteins of immunological relevance from vertebrate pathogens with bibliographic, taxonomic, and molecular metadata... 33</b> |           |
| 1. INTRODUCTION.....                                                                                                                                                            | 35        |
| 2. IMPLEMENTATION.....                                                                                                                                                          | 37        |
| 3. RESULTS.....                                                                                                                                                                 | 38        |
| 4. CONCLUSIONS.....                                                                                                                                                             | 41        |
| 5. REFERENCES.....                                                                                                                                                              | 42        |
| 6. SUPPLEMENTAL DATA.....                                                                                                                                                       | 43        |
| 6.1. Supplemental Methods.....                                                                                                                                                  | 43        |
| 6.1.1. Machine-learning component.....                                                                                                                                          | 43        |
| 6.1.1.1. Dataset construction.....                                                                                                                                              | 44        |
| 6.1.1.2. Data pre-processing and analysis of most frequent terms.....                                                                                                           | 46        |
| 6.1.1.3. Logistic regression model.....                                                                                                                                         | 46        |
| 6.1.1.4. Evaluation metrics.....                                                                                                                                                | 47        |
| 6.1.2. Second component of PADA-ONE: article-level metadata and relational database.....                                                                                        | 47        |
| 6.1.2.1. Anchoring article, sequence, and taxonomic metadata to PADA-ONE entries.....                                                                                           | 47        |
| 6.1.3. Third component of PADA-ONE: front-end user interface and common query filters.....                                                                                      | 50        |
| 6.2. Supplemental Results.....                                                                                                                                                  | 50        |
| 6.2.1. Classifier performance.....                                                                                                                                              | 50        |
| 6.3. Supplemental Tables.....                                                                                                                                                   | 51        |
| 6.4. Supplemental References.....                                                                                                                                               | 52        |
| <b>Chapter 2. MARVL - Machine-learning Assisted Reverse VaccinoLogy.....</b>                                                                                                    | <b>54</b> |
| 1. INTRODUCTION.....                                                                                                                                                            | 56        |
| 2. METHODS.....                                                                                                                                                                 | 58        |
| 2.1. Datasets.....                                                                                                                                                              | 58        |
| 2.2. Distribution of similarity for known vaccines and non-vaccines.....                                                                                                        | 59        |
| 2.3. Feature calculation.....                                                                                                                                                   | 59        |
| 2.4. Pre-processing of features and dimensionality reduction.....                                                                                                               | 60        |
| 2.5. Building individual machine learning models.....                                                                                                                           | 60        |
| 2.6. Performance assessment, model comparison, and ensemble building.....                                                                                                       | 61        |
| 3. RESULTS.....                                                                                                                                                                 | 62        |
| 3.1. Sequence dissimilarity is a poor proxy to distinguish between vaccines and non-vaccines.....                                                                               | 62        |

|                                                                                                                            |            |
|----------------------------------------------------------------------------------------------------------------------------|------------|
| 3.2. Dataset splitting and feature preprocessing.....                                                                      | 64         |
| 3.3. Training and performance evaluation of OCCs and BCs.....                                                              | 64         |
| 3.4. Building and validation of an ensemble classifier for reverse vaccinology.....                                        | 65         |
| 3.5. Feature importance.....                                                                                               | 69         |
| 4. DISCUSSION.....                                                                                                         | 72         |
| 5. CONCLUSION.....                                                                                                         | 75         |
| 6. REFERENCES.....                                                                                                         | 77         |
| 7. SUPPLEMENTARY MATERIAL.....                                                                                             | 79         |
| 7.1. Development of MARVL.....                                                                                             | 79         |
| 7.1.1. BLAST alignment of gold-standard and Protegen entries.....                                                          | 79         |
| 7.1.2. Feature description.....                                                                                            | 80         |
| 7.1.2.1. Statistical and physicochemical features.....                                                                     | 80         |
| 7.1.2.2. Biologically functional and engineered features.....                                                              | 83         |
| 7.1.3. Data transformation.....                                                                                            | 85         |
| 7.1.4. Models description.....                                                                                             | 85         |
| 7.1.5. Model training, tuning, and evaluation.....                                                                         | 86         |
| 7.2. Supplemental Figures.....                                                                                             | 88         |
| 7.3. Supplemental Tables.....                                                                                              | 93         |
| 7.4. References.....                                                                                                       | 94         |
| <b>GENERAL DISCUSSION.....</b>                                                                                             | <b>97</b>  |
| <b>GENERAL CONCLUSION.....</b>                                                                                             | <b>102</b> |
| <b>REFERENCES.....</b>                                                                                                     | <b>103</b> |
| <b>APPENDIX A - Co-author in: Organism-specific training improves performance of linear B-cell epitope prediction.....</b> | <b>109</b> |

## GENERAL INTRODUCTION

### The data-driven science era

The advances in DNA sequencing technology make available an increasing number of genomes and transcriptomes at low costs, more quickly, and with increasing quality, ranging from viruses and prokaryotes to complex eukaryotes (YANDELL; ENCE, 2012). A considerable part of this sequence data is from pathogenic organisms due to their huge importance for medicine and agriculture (MUKHERJEE *et al.*, 2017). Additionally, several of these pathogenic species do not have life cycles established in laboratory conditions.

Once genomic data is available, several post-genomic analyses can be used to go deeper into the molecular aspects of an organism. Regarding pathogenic species, these analyses can generate relevant information about characteristics of the pathogenicity phenotype, such as the high-throughput screening for druggable gene products (COWELL *et al.*, 2018) and the global analysis of host-pathogen interactions. In this latter, the pathogen's ability to respond to environmental stimuli is investigated from the point of view of gene expression (MILITELLO *et al.*, 2008). An interesting class of post-genomic study in pathogenic organisms consists of the identification of potentially immunogenic peptides and proteins from genomic data, a knowledge field named Reverse Vaccinology (RAPPUOLI, 2001), and the main focus of our work.

The scientific community is developing several recent coordinated efforts to organize and share sequencing data. As a non-exhaustive list, we mention the catalogs of normal variation (GENOMICS ENGLAND RESEARCH CONSORTIUM, 2019) and disease-causing genomic alterations (TCGA RESEARCH NETWORK) in several organisms, the endeavors to catalog the genomic properties of an increasing group of model species (THE ENCODE PROJECT CONSORTIUM, 2020) and the environmental and functional composition of microorganisms in distinct niches (HUMAN MICROBIOME PROJECT CONSORTIUM, 2012). All these platforms harbor petabytes of raw data with the potential to provide key biological insights. Since it is infeasible for human beings to analyze these large, heterogeneous, and complex databases, Artificial Intelligence (AI) constitutes a valuable resource to

automatically search these databases and extract biologically meaningful knowledge. Machine learning (ML), a sub-field of AI, has become increasingly useful among biologists for that purpose (CAMACHO *et al.*, 2018).

ML is an AI technique to develop computer programs that optimize an inference model by experience, learned from example data, to predict outcomes (LARRAÑAGA *et al.*, 2006). These algorithms are named classifiers (or predictors) and are capable of classifying or grouping samples based on previous learning experiences. The next section defines important concepts of ML that will be used in this document.

### Machine learning - definitions, and perspectives

As in any scientific field, ML has a vocabulary that is used to formally define the conceptual steps needed to develop classifiers and evaluate their efficiency through defined metrics. The terms commonly used in an ML-based application are summarized in Table 1:

**Table 1. Concepts of some terms used in ML approaches. Adapted from ERICKSON *et al.*, 2017).**

| Term              | Concept                                                                                                                                                                                 |
|-------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Class             | The label or the correct answer of each sample/instance. For instance, an image labeled as "dog" and another image labeled as "non-dog". "Dog" and "non-dog" are the labels or classes. |
| Labeled data      | Group of samples/instances, each one with the correct "answer" (or class). For instance, a set of dog images and images of other animals.                                               |
| Feature/Attribute | Descriptive property extracted from a                                                                                                                                                   |

sample/instance, i. e. shape, color, size.

Training set

Fraction of samples/instances from the labeled dataset that is used to train classifiers.

Training

The step where the ML algorithm is given a training dataset and their respective features to learn feature distribution patterns and, eventually, which ones are more useful to distinguish classes.

Testing set

Fraction of samples from the labeled dataset that is not used to train classifiers, but is used later on to evaluate classifier efficiency.

Classification

The act of a classifier assigning a class or label to a group of samples/instances, i. e. if an image corresponds to a dog or not.

Validation set

Sometimes, a third set of examples is used, after the testing step, for mimicking a real scenario of prediction. In this case, the validation set is not used to train or test classifiers.

---

Regarding data labeling, there are two major types of ML approaches: unsupervised and supervised learning algorithms. The unsupervised one attempts to determine the ideal number of labels, as well as the label for each data instance, without previously known labels. Supervised algorithms are the ones that use all concepts defined in Table 1 and are able to learn the relationships between descriptive features and labeled data from the training set, and can be used to predict the classes of new instances never evaluated before on the testing and validation sets (ERICKSON *et al.*, 2017).

An important aspect of labeled datasets, mainly when they are composed of biological data, is the natural tendency to have the number of a given sample class under-represented, or imbalanced. For instance, a disease that affects only a small portion of a population. In some cases, there is a slight imbalance (for example, 1:2) between sample classes, but sometimes a more severe imbalance (for example, 1:100) is observed as well (KRAWCZYK, 2016; XU; JACKSON, 2019). Usually, to solve this problem, there are some strategies to reach balance between classes and the most commonly used are 1) the under-sampling of the major class until reaching the balance with the minor one; and 2) over-sampling of the minor class until reaching balance with the major one (LEMAÎTRE *et al.*, 2017). This processing is performed at the step of creating the training dataset, hence before starting the learning phase of the ML-based algorithm.

When the training set is created, the learning strategy consists in presenting the training samples in order to develop an inference model (Figure 1; ERICKSON *et al.*, 2017). At this step, the classifier learns the best feature combination to classify samples according to their classes. After each training round, a subset of samples not used for training is used to evaluate the classifier efficiency (Figure 1). Some examples of supervised algorithms are decision trees, support vector machines, and neural networks (LARRAÑAGA *et al.*, 2006; XU; JACKSON, 2019).

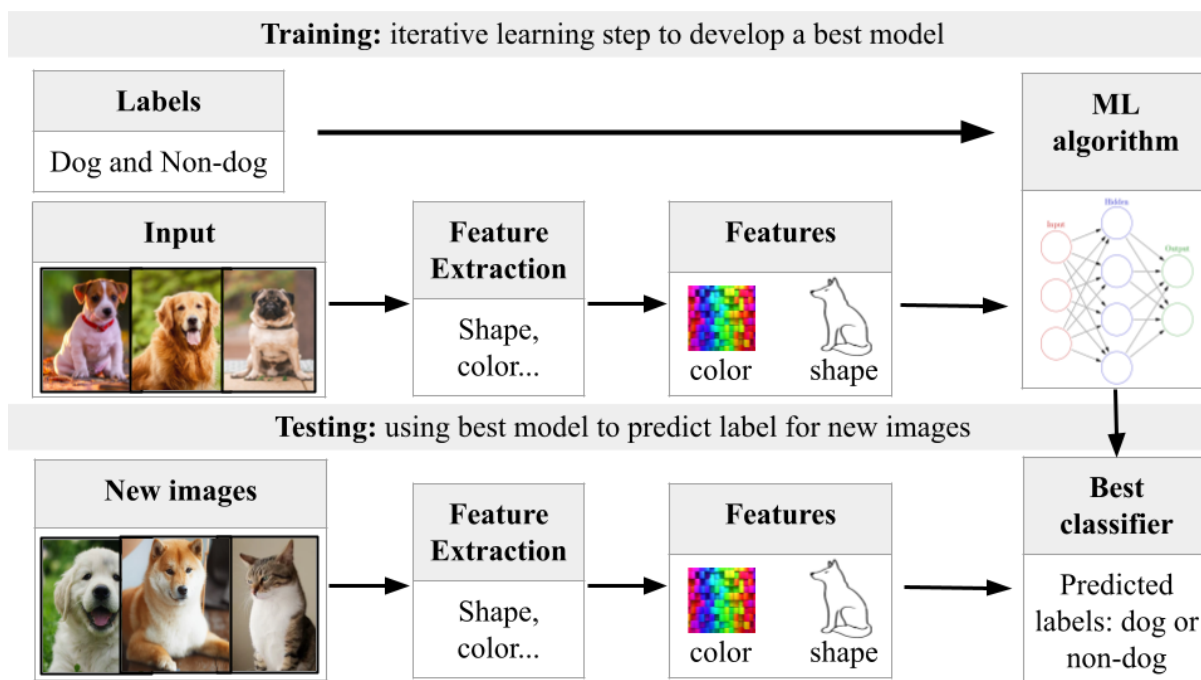


**Figure 1.** Development and application of a machine learning model for an image classification example. During the training step, the algorithm processes the extracted features to develop the best model to afterward predict the labels using new images as input.

AI is increasingly becoming a part of the biological sciences, as it already is in several other knowledge fields and industries, such as climate prediction, particle physics, telecom, electricity, gas, water, and transport (BALMER; LEVIN; SCHMIDT, 2020). ML is achieving performance equivalent to or better than trained humans in tasks as heterogeneous as image recognition (KRIZHEVSKY; SUTSKEVER; HINTON, 2012), speech recognition/translation (HINTON *et al.*, 2012), and particle detection in accelerator data (CIODARO *et al.*, 2012).

In biology, the aforementioned scenario of the wealth of sequencing data is already being surveyed using ML-based algorithms. The first classical examples of such algorithms are the tools to predict gene and protein function, RNA and protein structure, and gene and protein structural data (LARRAÑAGA *et al.*, 2006). More recently, ML algorithms have been applied in more specialized fields such as cancer diagnosis, which traditionally requires expert humans for sample labeling, with performance comparable to or better than physicians (ESTEVA *et al.*, 2017).

A new generation of ML-based algorithms is being actively developed to survey large genomic databases for activities such as variant calling (POPLIN *et al.*, 2018), prediction of regulatory regions (ZOU *et al.*, 2019), and estimation of transcript

abundance (WASHBURN *et al.*, 2019). Different from the most usual binary and supervised classifiers, another type of classification method that has been applied for medical purposes is the one-class classification (OCC). OCC methods create classification models in situations of either absence, poor sampling, or not labeling negative examples. Those models must be able to draw a classification boundary around the positive instances in the multidimensional features space to admit as many examples as possible with positive labels, while it reduces the number of admitted outliers (KHAN; MADDEN, 2014). Studies are using OCC for screening of myopathies (BURLINA *et al.*, 2019), detecting outliers in Alzheimer's and breast cancer data (GAUTAM *et al.*, 2020), and in pulmonary disease diagnostics by detecting abnormal lung sounds (LANG *et al.*, 2020). In immunology, AI is also having an increasing role to distinguish between protective and non-protective proteins in protein-coding genes from parasitic species for vaccine purposes, which is the focus of the present work.

#### Vaccine development: the conventional approach and Reverse Vaccinology

The vast majority of vaccines available or under development are the product of a traditional research pipeline, where a pathogenic organism is first grown in a laboratory under controlled conditions, followed by either the obtention of non-virulent strains (first vaccines) or the detection of potentially immunogenic proteins or peptides from such organisms, as done frequently in more recent approaches. Both strategies are followed by the evaluation of their capability to elicit protective immunity through controlled experimental designs (HEINSON; WOELK; NEWELL, 2015). Even though this traditional strategy has been highly successful for several parasitic diseases, this approach does not always find suitable vaccine candidates due to several reasons.

An important issue is that traditional immunological assays to detect vaccine candidates tend to preferentially select abundant and highly immunogenic candidates, which are frequently decoy sequences. Those sequences are a product of natural selection found in parasitic genomes that drive immune responses away from functional regions in parasite proteins, such as multi-gene surface proteins (HOU *et al.*, 2020). Furthermore, the conventional strategy usually requires the maintenance of parasites in a laboratory for their molecular characterization, leading

to failures to develop vaccines against pathogens with no known protective antigens or non-cultivable in controlled environments (RAPPUOLI, 2001).

Ideally, a good vaccine candidate should possess at least some desirable properties. These possible vaccines should be located in the host/parasite molecular interface, such as transmembrane or extracellular proteins in cellular organisms, or capsid proteins in viruses, as targeting these proteins may lead to a disruption of the molecular mechanisms of host-parasite interaction (ONG; WONG; HE, 2017). Vaccine candidates should also play essential roles during a parasite's life cycle and, consequently, the lack of their biological activities should decrease the parasite's fitness. Searching for stage-specific candidates requires the evaluation of several stages of a parasite's life cycle, for instance, *Plasmodium falciparum*, which presents different antigenic profile expressions during sporozoites, merozoites, and gametocytes phases (RAPPUOLI, 2000). Vaccine candidates are also expected to elicit a protective humoral and/or cellular immune response, therefore they should contain B- and/or T-cell epitopes. Finally, vaccine candidates should be conserved among distinct parasite strains to ensure a vaccine will be effective against distinct lineages of a parasite species found in nature (D'MELLO *et al.*, 2019; RAPPUOLI, 2000).

Hence, many efforts have been made towards improving vaccine development by establishing automated methods to overcome the known limitations of traditional vaccine development pipelines. The increasing availability of genomic information for pathogenic organisms is the starting point of a research field named reverse vaccinology (RV) (RAPPUOLI, 2001). RV has boosted the progress in vaccine development by starting the selection of good vaccine candidates through an initial computational screening of all known protein-coding genes in genomic sequences from parasite genomes (PIZZA *et al.*, 2000).

RV presents key advantages when compared to the conventional one. Specifically, by scanning all potentially immunogenic proteins and peptides from non-cultivable organisms for suitable vaccine candidates, there is no need to purify protein molecules for its initial characterization (DEL TORDELLO; RAPPUOLI; DELANY, 2017). RV also makes it possible to predict proteins located at the host-parasite molecular interface, as well as those with known epitopes, and playing essential roles in a parasitic lifestyle. Therefore, RV may save time and reduce costs when compared to conventional approaches.

## ML in reverse vaccinology

As RV protocols matured, a few initiatives to develop and share RV pipelines have been made available in the scientific literature. Generally, these RV pipelines may be classified into two types based on the algorithm behind the decision process: filtering and classifying programs (DALSASS *et al.*, 2019; HEINSON; WOELK; NEWELL, 2015).

The filtering algorithms evaluate protein sequences regarding several features (e.g. epitope presence, prediction of cellular location, conservation across parasite lineages) and remove those that fail to achieve a previously established cutoff, proceeding with a subset of promising antigens. NERVE (VIVONA; BERNANTE; FILIPPINI, 2006) is an archetypal example of a filtering RV pipeline. NERVE annotates bacterial protein sequences using the following functional attributes: subcellular localization, adhesion probability, secondary structure, sequence similarity with host proteins, and conservation in strains of the same species. The decision process is based on cutoffs arbitrarily defined by the authors for each one of the features, although the user may modify them if needed. Other examples of filtering RV pipelines are Jenner-predict (JAISWAL *et al.*, 2013), Vacsol (RIZWAN *et al.*, 2017), and PanRV (NAZ *et al.*, 2019), which uses functional attributes such as subcellular localization, transmembrane topology, match to essential genes and virulence factors, B-cell epitope prediction and other.

The classifying programs, on the other hand, are based on ML techniques to make decisions about potential immunogenic proteins and mitigate human interference. These algorithms are expected to learn feature importance in order to choose the best set of features to detect interesting vaccine candidates without relying on arbitrary, previously established cutoffs. To the best of our knowledge, Vaxijen was the first RV pipeline to use ML to select vaccine candidates (DOYTCHINOVA; FLOWER, 2007). Since ML-based approaches are capable of handling the massive amount of omics data available so far, it is not surprising to observe that these strategies are increasingly becoming the state-of-the-art method to select vaccine candidates from genomic data. Furthermore, ML-based methods are also capable of working with many more features than the filtering algorithms and return as a result a ranked list of vaccine candidates instead of providing just a list of

good vaccine candidates that have a set of interesting features above an arbitrary cutoff (DALSASS *et al.*, 2019). Over the years, new tools have been developed, aggregating ML algorithms to their structure, such as Vaxijen, the Bowman-Heinson predictor (BOWMAN *et al.*, 2011; HEINSON *et al.*, 2017), Vacceed (GOODSWEN; KENNEDY; ELLIS, 2014) and culminating in the current state-of-the-art, a new version of Vaxign-ML (ONG *et al.*, 2020), Vaxign2 (ONG *et al.*, 2021), and Vaxi-DL (RAWAL *et al.*, 2022). As recently evaluated by (DALSASS *et al.*, 2019), ML-based RV protocols outperform filtering RV pipelines, with the Bowman-Heinson arguably the most successful one at that point, recalling 75% of bacterial protective antigens, in contrast with the most used Vaxijen and Vaxign (DOYTCHINOVA; FLOWER, 2007; HE; XIANG; MOBLEY, 2010).

Vaxi-DL is the most recent ML-based RV pipeline available and was specifically developed to select bacterial, protozoan, fungal, and viral vaccine candidates from genomic data. Vaxi-DL combines several physicochemical and functional features to predict potential vaccine candidates. As examples of these features we mention 1) physicochemical properties, quasi-sequence-order distributor, pseudo-amino acid composition, and autocorrelation descriptors; and 2) biological functional attributes, such as likelihood for adhesion, prediction of transmembrane helices and instability index (RAWAL *et al.*, 2022). However, despite all of the efforts to develop more reliable and efficient RV tools, there are still opportunities for improvement, as we expose below.

### Challenges for new ML-based RV programs

Despite the successes of existing RV tools, there are still several opportunities for improvement. Current RV approaches present limitations that need to be addressed to generate reliable and robust ML-based strategies. For one, most existing programs were developed based on training datasets containing only sequences from a few groups of pathogens, with the majority having been developed and validated using only datasets of prokaryotic organisms. Vaxijen, Vacceed, Vaxign2, and Vaxi-DL are the only pipelines with alternatives to search for viral, tumoral, and eukaryotic antigens. Furthermore, programs developed based on prokaryotic data are likely to be unsatisfactory when evaluating eukaryotic genomes

(GOODSWEN; KENNEDY; ELLIS, 2013), suggesting that the development of taxon-specific tools for RV may increase detection performance.

A second issue relates to the selection of the training sets for ML models used for vaccine discovery. These models are binary classifiers (BC) expected to learn separation surfaces between PAgS and non-PAgS from training datasets ideally containing positive (i.e., known PAgS) and negative (i.e., known non-vaccines) examples. However, throughout the literature, most, if not all, models are trained on sets containing known PAgS, but with counterexamples selected from sets of unlabeled sequences (either actual proteins or synthetic sequences) based on sequence dissimilarity to the positive instances. The use of sequence dissimilarity as a proxy for the difference in terms of immunogenicity is based on the assumption that any sequence that is sufficiently dissimilar from a known vaccine candidate can be safely selected as a non-vaccine. A cutoff of 30% similarity is commonly adopted to select proteins sufficiently dissimilar from known vaccines to be used as training counterexamples (ONG *et al.*, 2021; RAWAL *et al.*, 2022).

Another point of concern for several of the currently available ML-based RV programs is the use of computationally or experimentally demanding features, including functional properties, such as gene expression data, features based on network topology, and functional domain features computed from experimental data, such as RNA-Seq analysis, availability of protein-protein interaction and a search on protein databases, respectively. Although these features may be available for well-studied pathogens, this is not true for non-model, recently sequenced organisms, such as emergent pathogens and neglected tropical diseases (ELLEGRÉN, 2014; METCALF; LESSLER, 2017), which may limit the usability of these models. Therefore, the development of predictors based on features that can be computed from primary protein sequence data alone is essential, if we aim to develop more general classifiers (MARQUES DE CASTRO *et al.*, 2022; NIGATU *et al.*, 2017).

Finally, we also list another important issue, which is searching for true positive and negative instances of protective antigens in the literature. As those proteins emerge from several studies, arranging them in a searchable and updated manner is essential to track this information over groups of pathogens, as well as to provide the scientific community with accurate data and metadata to develop new tools to predict protective antigens. However, databases hosting these data are

scarce and present some limitations (see section ‘1. Introduction’, Chapter 1). In this regard, as there are gaps to be fulfilled, we developed PADA-ONE, an online explorer for proteins of immunological relevance in vertebrate pathogens, including their associated metadata. The PADA-ONE’s predictive component was trained using titles of papers describing protective antigens as positive instances and, as the negative ones, paper titles from Immune Epitope Database (IEDB). Once deployed on PubMed, PADA-ONE supplied new examples of protective and non-protective proteins, an important sequence dataset used afterward to validate MARVL, the second product of this work (Chapter 2). MARVL is an end-to-end pipeline for training an ensemble of classifiers to predict vaccine candidates from sequence data. As a final step, using protective and non-protective proteins obtained with PADA-ONE, MARVL was compared with state-of-the-art programs.

## **CHAPTER 1**

**PADA-ONE: an online tool to explore  
proteins of immunological relevance from  
vertebrate pathogens with bibliographic,  
taxonomic, and molecular metadata**

## **PADA-ONE: an online tool to explore proteins of immunological relevance from vertebrate pathogens with bibliographic, taxonomic, and molecular metadata**

**Igor Lobo<sup>1</sup>, Maycon Douglas de Oliveira<sup>1</sup>, Leonardo Vinícius Dias da Silva<sup>2</sup>, Agnello Picorelli<sup>3</sup>, Daniella Castanheira Bartholomeu<sup>4</sup>, Felipe Campelo<sup>5</sup>, Francisco Pereira Lobo<sup>6\*</sup>**

<sup>1</sup>Graduate Program in Genetics, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil. <sup>2</sup>Department of Social Pharmacy, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil. <sup>3</sup>Laboratory of Evolutionary Genomics, Universidade Estadual de Campinas, Campinas, 13083-970, Brazil. <sup>4</sup>Department of Parasitology, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil. <sup>5</sup>Department of Computer Science, Aston University, Birmingham B4 7ET, UK. <sup>6</sup>Department of General Biology, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil. \*Corresponding author

**Keywords:** Vaccines, Database, Machine Learning, Text mining

**Abstract:** The computational screening of parasite genomes for new protective antigens (PAgs) has become a common task for immunologists. However, initiatives to provide user-friendly access to systematized data and metadata about PAgs and other classes of immunologically relevant proteins, such as boosters and attenuators of the immune response, are still scarce. We introduce PADA-ONE (Protective Antigen DAtabase - OnliNe Explorer), a web tool that hosts curated articles describing immunologically relevant proteins (IRPs), as well as non-curated articles predicted to contain information about IRPs using a robust machine learning classifier trained on article titles describing PAgs. PADA-ONE provides users with several types of filters based on classification scores, taxonomy, and publication-related metadata to search a database composed of more than 500,000 papers enriched in entries potentially describing IRPs.

**Availability and Implementation:** PADA-ONE is accessible at <https://padaone.org/>.

**Abbreviations:** PAg - Protective Antigens, non-PAgs – non-protective antigens, IRP – Immunologically Relevant Proteins.

**Contact:** franciscolobo@ufmg.br

**Supplementary information:** Supplemental data will be available at Bioinformatics online.

## 1. INTRODUCTION

A considerable fraction of the genomes available in public databases belong to pathogens of medical and veterinary interest (Yandell and Ence 2012, Mukherjee et al. 2017). Several post-genomics analyses can be performed on this data to generate a deeper understanding of the molecular *modus operandi* of pathogens infecting vertebrate hosts from an immunogenomics perspective. A major interest is the detection of protective antigens (PAgs), proteins capable of triggering specific immunological protection against a pathogenic organism from the host immune system by the stimulation of humoral and/or cellular immunity (Flower et al. 2010; Yang et al. 2011). In practice, exposure to a PAg elicits effective protection against challenges by a disease-causing agent following immunization (Doytchinova and Flower 2007), making them an essential component of most vaccines.

The computational screening of parasite genomes for PAgs has become a common task for immunologists and has been used in numerous reverse vaccinology studies (Sharma et al. 2021). These searches are based on the properties expected to be found in PAgs, which are used to develop computational tools to search for new vaccine candidates (*de novo* search). Other strategies involve similarity searches using sequences of known PAgs from phylogenetically close species (similarity-based search), or hybrid strategies based on both *de novo* and similarity-based searches. Besides vaccines, PAgs are also used for immunotherapies (Manish et al. 2020) and as molecular markers for diagnosis (Evans et al. 2018; Wang et al. 2020). These proteins are also an essential resource to design prophylactic, therapeutic, and diagnostic strategies, and initiatives to organize data about PAgs and related metadata are welcome, although still scarce.

In addition to PAgs, several other classes of proteins found in vertebrate pathogens are relevant from an immunological perspective (defined here as *Immunologically Relevant Proteins* - IRPs). Some classes of proteins, when removed or introduced, can create attenuated versions of pathogenic organisms suitable for vaccine development (referred to as *attenuators*; Vetter et al. 2018), while others can

boost immune system responses when used as a component of vaccine preparations (referred to as *boosters*; Kardani et al. 2016).

The experimental characterization of putative PAgS from a list of potential vaccine candidates also generates a byproduct of non-protective proteins, experimentally validated during the search for PAgS and found as not capable of inducing protective immunity (*non-PAgS*). The information that these negative results failed to induce protection can be valuable for research groups interested in characterizing similar proteins. Non-PAgS also represent an interesting source of information for the development and validation of predictive models that learn to distinguish PAgS (positive class) from non-PAgS (negative class) to detect vaccine candidates. Nowadays, these methods rely on arbitrary sequence alignment cutoffs to select proteins that are sufficiently dissimilar from the set of PAgS used to train them to define the negative class (Ong et al. 2021; Rawal et al. 2022). This is based on the premise that only a conserved and small fraction of the proteomes of pathogens are PAgS, even though dissimilarity has not been demonstrated to be a good proxy for the separation between PAgS and non-PAgS.

To the best of our knowledge, there are two databases hosting experimentally validated PAgS: AntigenDB (Ansari et al. 2010) and the Protegen database (Yang et al. 2011). Both are focused on hosting PAgS from bacterial and viral species, with very little information related to other important groups of human pathogens. Moreover, even though these databases provide a relevant source of PAgS, they have not been updated in years, undoubtedly due to the high costs of manually curating the scientific literature for the description of new PAgS. As for the other classes of IRPs, VirmugenDB is a curated database of virulence factors that have been demonstrated to produce attenuated strains of pathogenic organisms (Racz et al. 2013).

We present PADA-ONE (Protective Antigen DAtabase - OnliNe Explorer; <https://padaone.org/>), a searchable, user-friendly web tool that hosts curated and non-curated scientific literature predicted to describe IRPs (PAgS, non-PAgS, attenuators, and boosters), together with the embedded sequence, taxonomical and journal-related metadata, when available. PADA-ONE covers a taxonomically diverse set of pathogens, expanding the diversity of IRPs available to the scientific community in a structured database.

## 2. IMPLEMENTATION

PADA-ONE is composed of three main components:

1) A predictive component, which consists of a classification model trained and evaluated using titles of curated scientific literature describing known PAgs (positive class) and describing immunogenic peptides in general (negative class) (Figure S1, Supplemental File, section “1.1. Dataset construction”). This component assigns probabilities to articles as belonging to the positive class based on textual features extracted from its title. This component was used to survey the entirety of the scientific literature listed in PubMed for articles potentially describing IRPs.

2) A MySQL database back-end component that stores approximately 500,000 PubMed articles selected by the first component as possible sources of scientific literature describing PAgs, together with article metadata covering sequence, taxonomic and literature-based sources, when available (the entity-relationship diagram of PADA-ONE is depicted in Figure S2; see also Supplemental File 1, section “Anchoring article, sequence and taxonomic metadata to PADA-ONE entries”, for the full description of the sources of associated metadata).

3) An interface component, which is a web front-end and a custom API built in TypeScript with NextJS and running with a NodeJS runtime in a Linux server. The back-end API receives HTTP requests from the front-end and retrieves information from the database using the Prisma API. The web interface provides browsing, searching, and filtering functions for all articles stored in the database. Each individual article has a dedicated page where more in-depth information is available, for instance, bibliometric information, external links to PubMed, NCBI Taxonomy, and a downloadable list of genes associated with the article. A comprehensive description of the Home and Results pages, as well as the filters used in the search, is available on the 'Help' page.

### 3. RESULTS

The first component of PADA-ONE comprises a statistical model trained to distinguish between the titles of articles describing known bacterial vaccines from Protegen and standalone initiatives (positive class, known PAGs) and the same number of article titles randomly selected from the metadata of entries in the Immune Epitope Data Base entries (negative class) (Figure S1, Supplemental File 1, section “Dataset construction”). The choice of the negative class intended to train a classifier capable of selecting terms specifically associated with vaccine development from the arguably much larger set of conceptually related terms used to describe immunogenic proteins used, for instance, for diagnosis.

We selected three sets of articles as holdouts to evaluate our model. The first contains a random selection of 120 negative and 114 positive examples used to evaluate our model after training (Figure S1 – step B). The second comprises 50 positive and 50 negative examples, used to evaluate classifier performance in the real-world scenario when it is used to survey PubMed for articles describing PAGs (Figure S1 – step G). The third consists of 300 Protegen articles describing PAGs of non-bacterial species and was used to evaluate whether our model is able to uncover articles describing PAGs of non-bacterial origin. The model achieved a high classification performance (Figure S3A, performance evaluated using the first holdout – sensitivity: 0.92; precision: 0.85), being able to select sets of biologically meaningful terms that are useful to discriminate between the two classes (Figure S3B).

When deployed on the full set of approximately 32 million PubMed title entries from 1983 to June 2022, the model produced a skewed distribution of predicted probabilities, with a longer tail of articles with low predicted values (Figure 1A). Visual inspection of the distribution of predicted probabilities for the second holdout (positive class – crosses, negative class – empty circles) demonstrates how our tool effectively discriminates between the two classes. The third holdout (PAGs articles from non-bacterial species, Figure 1A, black circles) also receives scores clearly biased towards greater predicted values, demonstrating the ability of our tool to detect articles describing PAGs regardless of the group of pathogens.

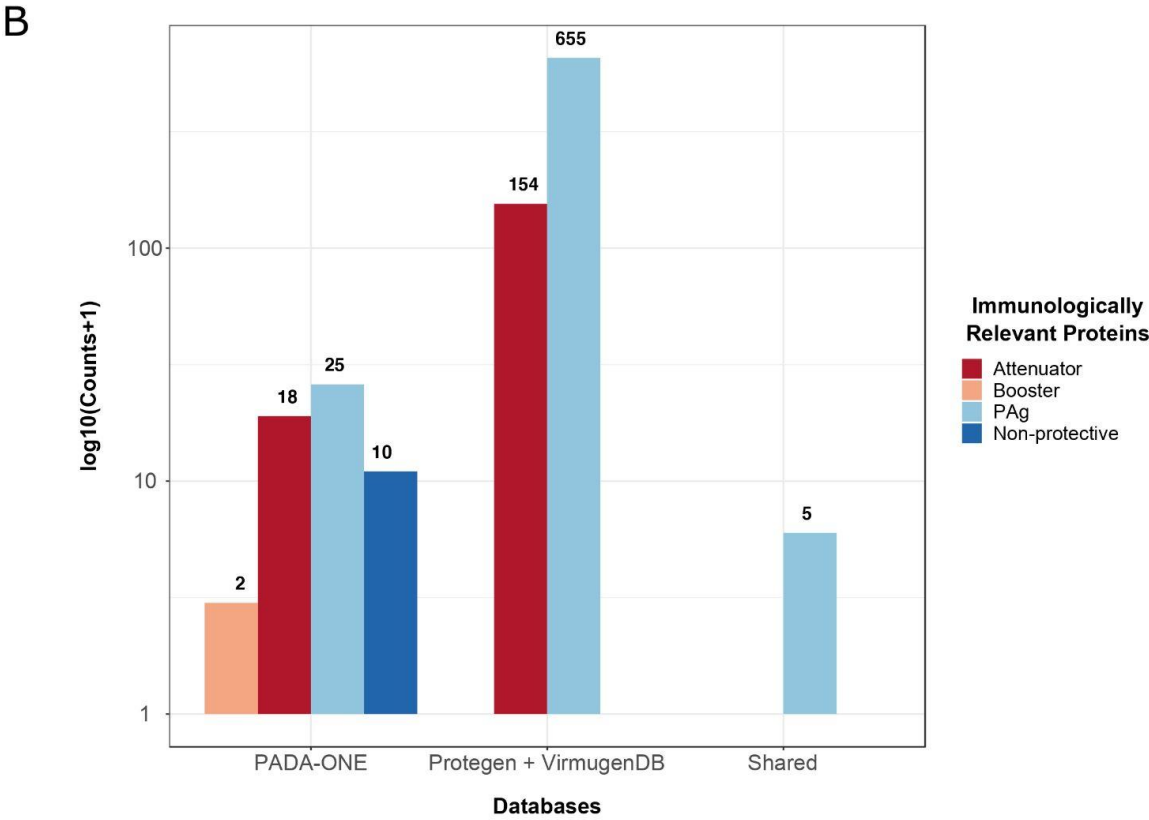
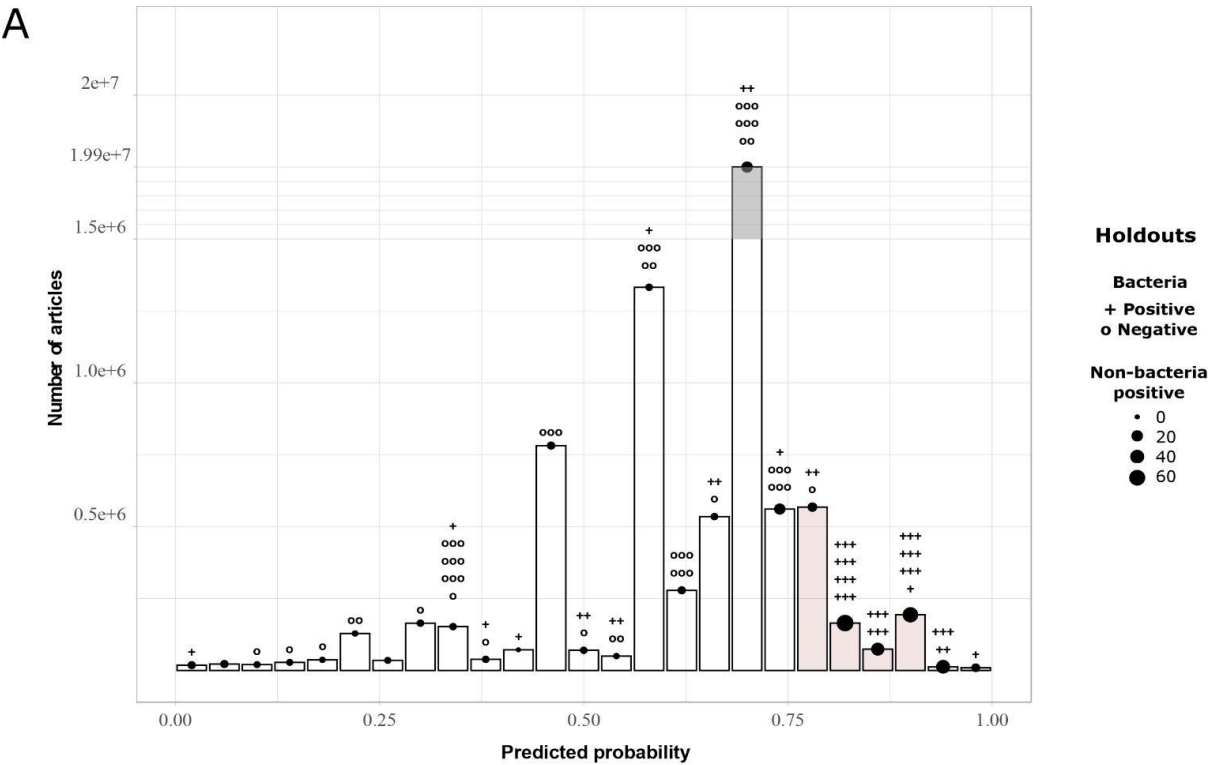


**Figure 1. A) Distribution of probability scores for the first component of PADA-ONE when used to survey approximately 32 million Pubmed entries for articles describing PAgS.** Symbols above the bars are the following holdouts: titles of PAg articles from Protegen (positive class - '+') and from IEDB (negative class - 'o'). Black circles represent articles from Protegen describing PAgS from non-bacterial origin (third holdout). Probabilities above 75% are highlighted as light red bars. Please notice that the y-scale is compressed beyond  $1.5 \times 10^6$ , to allow the visualization of the disproportionately high peak located at predicted probability values around 0.7. **B) Number of IRPs in PADA-ONE, Protegen and VirmugenDB.**

We set a cutoff of 0.75 to maximize sensitivity and specificity to build the second component of PADA-ONE, generating an initial database of 524,653 articles predicted as likely to be enriched in entries describing IRPs, together with sequence and taxonomic metadata, when available (29,212 articles, or approximately 5.6% of the total). This set of articles with sequence data available is being progressively curated by our group in an ongoing effort to provide a growing set of scientific literature describing IRPs. We also integrated the information on bacterial PAgS from Protegen into PADA-ONE by linking the sequence IDs from this database to their Pubmed IDs, when possible, further expanding the number of articles with embedded sequence metadata.

Since each sequence ID in PADA-ONE is associated with a species-level taxonomic ID through the NCBI Taxonomy Database, we developed a taxonomic search to find curated IRPs at any taxonomic level available in NCBI and present in our database (e.g., all sequences from a specific bacterial taxon). Together, the classification score and the taxonomic search comprise the two major filtering criteria provided in the PADA-ONE user interface to retrieve biomedical literature describing potential IRPs.

In the advanced search mode, users can retrieve articles based on additional filters (e.g., specific terms in the title or abstract, journal name, or number of citations). We also implemented additional filters to further reduce potential false positives and select articles associated with sequence data likely to come from pathogens of vertebrate organisms. Multiple matching results are presented in an interactive table, where articles can be sorted using different criteria. Our tool also implements community-level curation by allowing users to flag any scientific work not yet reviewed by our group as containing IRPs. This two-level curation process is expected both to produce engagement in the community using PADA-ONE and to speed up the curation process of new IRPs.

The manual curation of articles in PADA-ONE with associated bacterial gene IDs has already uncovered 60 confirmed IRPs from a diverse set of bacterial species, comprising 30 vaccines (PAGs), 18 attenuators, 10 non-PAGs, and two boosters (Supplemental Table 2, Figure 1B). Even though these numbers are relatively modest contributions to the number of bacterial PAGs and attenuators available at Protegen and VirMugenDB (660 and 154, respectively), there is only a small intersection between the sequences found in our manual curation and these databases (only five of the PAGs, and no attenuators), meaning that most entries found by our strategy are not available in these databases.

Furthermore, the latest update of the Protegen and VirMugenDB databases happened in 2019 and 2012, respectively. From the 25 new PAGs found using PADA-ONE, 20 were originally reported prior to 2019. Similarly, of the 18 new attenuators, 9 were originally reported prior to 2012. Even though these entries could have been incorporated into the last updates of Protegen and VirMugenDB, they were apparently missed by the teams working on those resources at the time of the last update. This suggests that our strategy to search for IRPs is capable of finding both examples not detected by other scientific efforts to gather this information and new examples.

#### **4. CONCLUSIONS**

We report PADA-ONE, a specialized online tool to survey a database of scientific literature enriched in entries predicted to describe PAGs and other classes of IRPs. Our tool provides different types of filters based on classification scores, taxonomy, and publication-related information, organized in a user-friendly interface. PADA-ONE also implements curation at the community level, which aims to speed up the uncovering and checking processes of new IRPs.

PADA-ONE provides an easy way to access an expanding source of IRPs, including the already invaluable data from Protegen and VirMugenDB. Taxonomic searches in PADA-ONE can be used to survey groups of interest for known IRPs that may be useful for phylogenetically close pathogens that lack efficient vaccines. As future improvements, we plan to provide curated data for IRPs of non-bacterial pathogens, as well as further develop our statistical models to explicitly predict other classes of IRPs not currently incorporated in our training database.

## 5. REFERENCES

- Ansari HR *et al.* AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res* 2010; **38**, D847-53.
- Doytchinova IA and Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007; **8**, 4.
- Evans SM *et al.* Outer Membrane Protein A Conservation among *Orientia tsutsugamushi* Isolates Suggests Its Potential as a Protective Antigen and Diagnostic Target. *Trop Med Infect Dis* 2018; **3**.
- Flower DR *et al.* Computer aided selection of candidate vaccine antigens. *Immunome Res* 2010 **6 Suppl 2**, S1.
- Kardani K *et al.* Prime-boost vaccine strategy against viral infections: Mechanisms and benefits. *Vaccine* 2016; **34**, 413–423.
- Manish M *et al.* Anthrax prevention through vaccine and post-exposure therapy. *Expert Opin Biol Ther* 2020; **20**, 1405–1425.
- Mukherjee S *et al.* Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. *Nucleic Acids Res* 2017; **45**, D446–D456.
- Ong E *et al.* Vaxign2: the second generation of the first Web-based vaccine design program using reverse vaccinology and machine learning. *Nucleic Acids Res* 2021; **49**, W671–W678.
- Pollard AJ and Bijker EM. A guide to vaccinology: from basic principles to new developments. *Nat Rev Immunol* 2021; **21**, 83–100.
- Racz R *et al.* Systematic annotation and analysis of “virmugens”-virulence factors whose mutants can be used as live attenuated vaccines. *Vaccine* 2013; **31**, 797–805.
- Rawal K *et al.* Vaxi-DL: A web-based deep learning server to identify potential vaccine candidates. *Comput Biol Med* 2022 **145**, 105401.
- Sharma A *et al.* Advanced strategies for development of vaccines against human bacterial pathogens. *World J Microbiol Biotechnol* 2021; **37**, 67.
- Vetter V *et al.* Understanding modern-day vaccines: what you need to know. *Ann Med* 2018; **50**, 110–120.
- Wang H *et al.* *Babesia microti* Protein BmSP44 Is a Novel Protective Antigen in a

Mouse Model of Babesiosis. *Front Immunol* 2020; **11**, 1437.

Yandell M and Ence D A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* 2012; **13**, 329–342.

Yang B *et al.* Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res* 2011; **39**, D1073-8.

## **6. SUPPLEMENTAL DATA**

### **6.1. Supplemental Methods**

#### **6.1.1. Machine-learning component**

Figure S1 represents the conceptual steps followed for the training and validation of our statistical model to find PAgS from article titles (the first component of PADA-ONE) and will be used as a guide through this section.

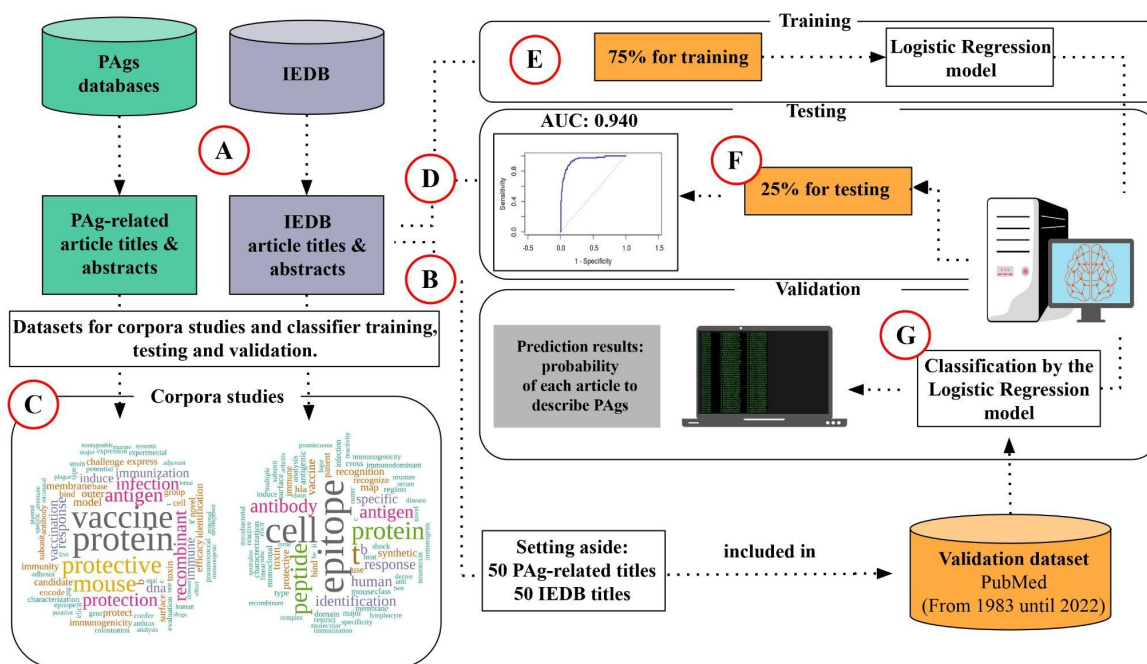


**Figure S1. Development of the first component of PADA-ONE.** **A:** Database of article titles describing PAGs from Protegen (positive class) and bacterial immunogenic peptides in general from IEDB (negative class). **B:** Selection of the second holdout (50 PAG-related and 50 IEDB titles) as one of our validation datasets. **C:** Corpora of most frequent terms. **D:** Selection of samples for model training (75%) and testing (25%, first holdout). **E:** Training of the Logistic Regression model. **F:** Performance evaluation. **G:** The second holdout (validation dataset) was used to evaluate the classifier performance when considering the approximately 32 million article titles published in PubMed from 1983 to 2022. **PAG:** Protective antigen; **IEDB:** Immune Epitope Database.

#### 6.1.1.1. Dataset construction

We performed a thorough curation of a public database plus individual initiatives describing PAGs to gather information about articles describing these sequences (Supplemental Table 1). We defined as PAGs the proteins which confer immunological protection (e.g. reduction of pathogen burden or improvement of survival rate) after immunization in an animal model followed by a subsequent challenge, with more than 50% of protection (Yang et al. 2013; Anderson et al. 2016; Park and Lee 2019). These papers were previously manually curated to be included in the Protegen database and to develop standalone programs, such as Nerve (Vivona et al. 2006), Jenner-Predict (Jaiswal et al. 2013), the Bowman-Heinson PAG predictor (Bowman et al. 2011; Heinson et al. 2017) and PanRV (Naz et al. 2019). Duplicated entries across datasets were excluded at this point.

The task of distinguishing the titles of articles describing vaccines from other articles is arguably far easier than the distinction of articles describing vaccines from articles describing immunogenic proteins and peptides in general. Nevertheless, these two fields are likely to share a substantial fraction of their vocabulary. Furthermore, the group of articles describing vaccines is likely to be a much smaller set than the group of articles describing immunogenic regions in general. For these reasons, it is desirable to train a model capable of distinguishing these two sets of articles.

We downloaded the Immune Epitope Database (IEDB, v.3, retrieved in April 2020) (Vita et al. 2015; Vita et al. 2019), a database of immunogenic peptides and associated metadata, to use as a source of articles describing immunogenic regions in bacteria in general. We used an *in-house* SQL script to select all articles PMIDs that are associated with IEDB entries (epitopes/non-epitopes from T- and B-cells) with a bacterial species described as the source and no filter for host range. As articles describing vaccines are also likely to exist in IEDB, we consider an article as a negative example (group "IEDB") when it was not observed in any of the aforementioned PAg-related datasets. We proceed by retrieving the titles and PMIDs for all negative articles.

We found a total of 818 distinct PubMed IDs describing known PAgS (positive class): 518 articles describing PAgS from bacterial origin and 300 describing PAgS from non-bacterial pathogens. At this point, we reserved all articles describing non-bacterial PAgS as a holdout to evaluate the generalization of our search strategy across phylogenetic domains (Non-bacterial positive holdout, black circles in Figure 1A). We proceed by randomly selecting 518 PubMed entries from IEDB that describe epitopes, presumed to not be associated with PAgS, from bacterial hosts (negative class), generating a total of 1,036 articles containing an equal number of entries per class.

At this point, we selected two additional holdouts from the 1,036 entries to further evaluate model performance: one holdout comprising 50 positive examples and 50 negative examples to evaluate classifier performance (Figure 1A, crosses and open circles for positive and negative examples, respectively). The second holdout comprises approximately 25% of the remaining entries to evaluate classifier performance in terms of sensitivity and specificity selected using the "rsample" R package (Silge et al. 2021) (114 and 120 positive and negative entries, respectively,

used to produce Figure S3A). The remaining 75% of the entries (354 positives and 348 negatives) were used to train our model as described below. At this point we have the data represented in Figure S1A.

#### 6.1.1.2. Data pre-processing and analysis of most frequent terms

We processed the titles of all curated articles of bacterial origin used for classifier training and evaluation for the visual inspection of word frequency and model training, as described below. We built one corpus for each group of titles and performed tokenization of article titles to a tidy data structure using the R package "tidytext" (Silge and Robinson 2016). For the transformation of titles into the tidy data structure, the following steps of pre-processing were performed: 1) stripping extra white spaces; 2) conversion of all words to lowercase; 3) removal of punctuation; 4) removal of stopwords, including pathogen names; and 5) word lemmatization using the R package "textstem" (Rinker 2018). We used the resulting data to visualize the most frequent words for each group as word clouds using the R package "wordcloud" (Fellows et al. 2018) (step C in Figure S1), as well as to train and test a logistic regression-based classifier.

#### 6.1.1.3. Logistic regression model

We used a regularized logistic regression model as our machine-learning classifier. Logistic regression is commonly used for binary classification problems and can be applied to assess the statistical significance of every predictor variable related to the probability of the outcome (Friedman et al. 2010; Shah et al. 2020). Regularization of the logistic regression model was done using LASSO (Least Absolute Shrinkage and Selection Operator), which effectively eliminates variables of low predictive power from the model (Tibshirani 1996). In other text mining works in the biomedical literature, logistic regression models have demonstrated satisfactory performance for the classification of loss- or gain-of-function mutations (Jung et al. 2015), for the classification of neuroimaging articles (Riedel et al. 2019), and, more recently, for the prediction of discharge destination (i.e. home, inpatient rehabilitation, death) of patients with COVID-19 (Fernandes et al. 2021).

The set of all positive and negative examples of bacterial origin used for model training was then converted into a sparse matrix containing the titles in the rows and the count of lemma occurrences in the columns, the latter to be used as features. The logistic regression-based classifier was trained using the R package "glmnet" (Friedman et al. 2010), which requires at least three arguments: a sparse matrix, a vector of responses in two levels, and a character string indicating the model family. Our classifier was trained using the "binomial" option, ending step E in Figure S1.

#### 6.1.1.4. Evaluation metrics

We trained the model using 10-fold cross-validation on the training dataset, followed by the evaluation step using the testing dataset (first and second holdouts), as shown on step F, Figure S1. The metrics applied for performance assessment were sensitivity and specificity, calculated as shown below:

$$\text{Sensitivity} = \text{True positives} / (\text{True Positives} + \text{False Negatives})$$

$$\text{Specificity} = \text{True Negatives} / (\text{True Negatives} + \text{False Positives})$$

We further evaluated classifier performance by considering the maximum AUC (Area Under the Curve) of the ROC (Receiver Operating Characteristic) curve (Swets 1988; Vergara et al. 2008) using R package "pROC" (Robin et al. 2011).

### 6.1.2. Second component of PADA-ONE: article-level metadata and relational database

#### 6.1.2.1. Anchoring article, sequence, and taxonomic metadata to PADA-ONE entries

After deploying the classification tool on the full set of PubMed entries as of June 2022, we selected 524,653 PubMed IDs representing articles likely to be enriched in entries describing proteins of immunological interest (prediction score > 0.75). We used the *eutils* tool (Sayers 2022) to construct queries and recover the following article-level metadata: Title, Abstract, Journal, Volume, Issue, Pages, Last author, and Number of citations. We also collected sequence and taxonomic

metadata, if available, using the following strategies: an Entrez search using *eutils* and the information available in the file “gene2pubmed” (Maglott et al. 2005). This resulted in 29,212 articles with sequence-level metadata. The entity-relationship model for our database is represented in Figure S2.

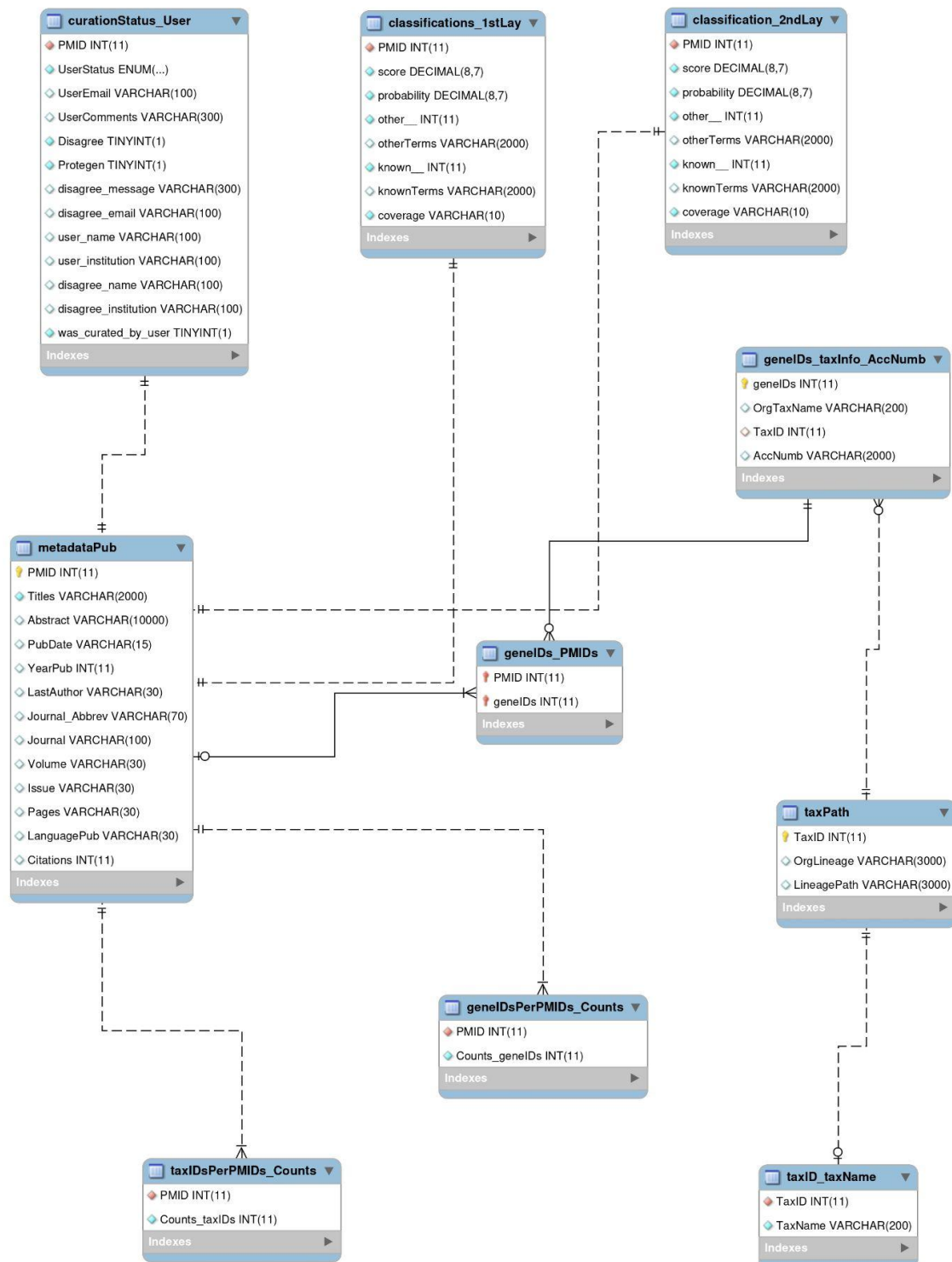


**Figure S2.** PADA-ONE entity-relationship diagram. The database entities are "metadataPub" (publication-related information, such as article title, abstract, year, citations, and other), "classification\_1stLay" (first layer of probability), "classification\_2ndLay" (second layer of probability), "geneIDs\_PMIDs" (geneIDs associated with an article), "curationStatus\_User" (curation status of an article by the user), "geneIDs\_taxInfo\_AccNumb" (information related to taxa and proteins involved in a study), "taxPath" (organism lineage information according to the Taxonomy database) and "taxID\_taxName" (a split version of "taxPath"), "geneIDsPerPMIDs\_Counts" (geneIDs counting per PMID), "taxIDsPerPMIDs\_Counts" (taxIDs counting per PMID).

### 6.1.3. Third component of PADA-ONE: front-end user interface and common query filters

We implemented filters to reduce potential false positives and select articles associated with sequence data likely to come from pathogens of vertebrate organisms as follows: 1) 'Exclude hosts' (excludes articles with sequences from the clade 'Vertebrata'); 2) 'Only papers with associated Gene IDs' (excludes articles without at least one associated gene ID); 3) 'Only true positive papers' (to include only those articles manually curated as describing IRPs); 4) and a second layer of predicted probabilities (excludes articles with a predicted score lower than a user-defined threshold).

## 6.2. Supplemental Results

### 6.2.1. Classifier performance

The classifier achieved high performance in distinguishing the lemmatized titles of positive and negative examples (Figure S3A, AUC of 0.940 on the holdout data). The most important terms used by the model for classification highlight several domain-specific words expected to distinguish between the two classes (Figure S3B). We further evaluated model performance using two holdout sets in a real-world scenario (Figure S1, step G, second holdout). The first of these holdouts comprises an additional set of 50 positive and 50 negative examples of articles describing PAgS of bacterial origin not used for classifier training (Figure S1, step E). The second holdout comprises 300 examples of PAgS of non-bacterial origin curated from Protegen. We observed greater probabilities assigned to the majority of the positive

PAGs of bacterial origin when compared with negative examples (Figure 1B, crosses and white circles, respectively). Furthermore, greater probabilities have also been assigned to most of the papers describing PAGs of non-bacterial origin (Figure 1B, black circles). Together, these results indicate that our tool is highly efficient in finding articles describing PAGs from several classes of pathogens.

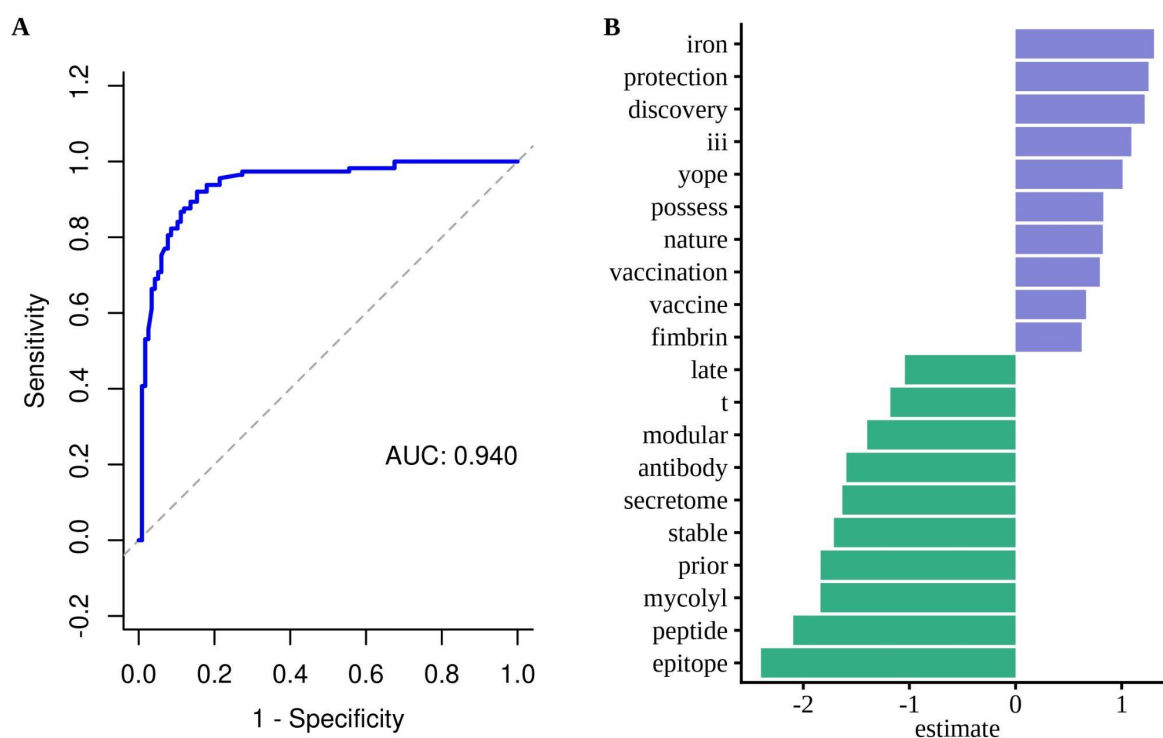


**Figure S3. Classifier performance and the outcome of the analysis of the most important terms to the classification task. A)** ROC Curve of the predictor in the holdout depicted in Figure S1 step E (AUC of 0.940) **B)** Top-10 most relevant terms for PAG-related and IEDB article titles (PAG-related in blue and IEDB in green).

### 6.3. Supplemental Tables

Supplemental Table 1. Article titles used to train, test, and validate the predictive component of PADA-ONE. Full supplemental table available at

[x Supplemental\\_Table\\_1\\_Chapter1.xlsx](#)

Supplemental Table 2. Uncovered immunologically relevant proteins (IRPs) after manual curation in PADA-ONE. Full supplemental table available at

[x Supplemental\\_Table\\_2\\_Chapter1.xlsx](#)

#### 6.4. Supplemental References

Anderson RJ *et al.* In vivo screen of genetically conserved *Streptococcus pneumoniae* proteins for protective immunogenicity. *Vaccine* 2016; **34**, 6292–6300.

Bowman BN *et al.* Improving reverse vaccinology with a machine learning approach. *Vaccine* 2011; **29**, 8156–8164.

Fellows I *et al.* Package ‘wordcloud’. *R package version* 2018; **2**, 331.

Fernandes M *et al.* Classification of the Disposition of Patients Hospitalized with COVID-19: Reading Discharge Summaries Using Natural Language Processing. *JMIR Med Inform* 2021; **9**, e25457.

Friedman J *et al.* Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* 2010; **33**, 1–22.

Heinson AI *et al.* Enhancing the biological relevance of machine learning classifiers for reverse vaccinology. *Int J Mol Sci* 2017; **18**.

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

Jung S *et al.* Identification of genomic features in the classification of loss- and gain-of-function mutation. *BMC Med Inform Decis Mak* 2015; **15 Suppl 1**, S6.

Maglott D *et al.* Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 2005 **33**, D54-8.

Naz K *et al.* PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. *BMC Bioinformatics* 2019; **20**, 123.

Park BS and Lee N. A bivalent fusion vaccine composed of recombinant Apx proteins shows strong protection against *Actinobacillus pleuroneumoniae* serovar 1 and 2 in a mouse model. *Pathog Dis* 2019; **77**.

Riedel MC *et al.* Automated, efficient, and accelerated knowledge modeling of the cognitive neuroimaging literature using the ATHENA toolkit. *Front Neurosci*

- 2019; **13**, 494.
- Rinker TW. textstem: Tools for stemming and lemmatizing text. 2018.
- Robin X *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; **12**, 77.
- Sayers E. A General Introduction to the E-utilities. 2022.
- Shah K *et al.* A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augment Hum Res* 2020; **5**, 12.
- Silge J and Robinson D. tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *JOSS* 2016; **1**.
- Silge J *et al.* CRAN - Package rsample. 2021.
- Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988; **240**, 1285–1293.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 1996; **58**, 267–288.
- Vergara IA *et al.* StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics* 2008; **9**, 265.
- Vita R *et al.* The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res* 2019; **47**, D339–D343.
- Vita R *et al.* The immune epitope database (IEDB) 3.0. *Nucleic Acids Res* 2015; **43**, D405–12.
- Vivona S *et al.* NERVE: new enhanced reverse vaccinology environment. *BMC Biotechnol* 2006; **6**, 35.
- Yang Y *et al.* Identification of an outer membrane protein of *Salmonella enterica* serovar Typhimurium as a potential vaccine candidate for Salmonellosis in mice. *Microbes Infect* 2013; **15**, 388–398.

## **CHAPTER 2**

### **MARVL**

### **Machine-learning Assisted Reverse VaccinoLogy**

## MARVL - Machine-learning Assisted Reverse VaccinoLogY

Igor Lobo<sup>1</sup>, Felipe Campelo<sup>2\*</sup>, Francisco Lobo<sup>3\*</sup>

<sup>1</sup>Graduate Program in Genetics, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil, <sup>2</sup>Department of Computer Science, Aston University, Birmingham B4 7ET, UK, <sup>3</sup>Department of General Biology, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil. \*Shared corresponding authorship

### Abstract

Reverse Vaccinology (RV) approaches select vaccine candidates using computational screening of protein-coding genes from parasite genomes. Despite the relative success of RV programs in supporting the development of prophylactic approaches against infectious diseases, current approaches based on Machine Learning (ML) still present important methodological issues, such as using sequence dissimilarity to known protective antigens (PAgs) as a proxy to define the set of counterexamples used for training ML models. In this work, we report results that indicate that sequence dissimilarity with known PAgs is not an efficient proxy to indicate a lack of ability to elicit protective immunogenicity. To address this issue, we created a gold-standard dataset of PAgs and experimentally validated non-protective antigens retrieved from the published literature, which can be used to develop, and in particular to evaluate, models aimed at discovering new PAgs from general protein data. We also present MARVL (Machine-learning Assisted Reverse VaccinoLogY), an end-to-end RV pipeline for training general ensemble classifiers for predicting vaccine candidates from sequence data alone. We used MARVL to develop an ensemble classifier composed of five distinct models, including a one-class classifier trained on positive data only, as well as different binary classifiers. We show that this ensemble outperforms the state-of-the-art tools when tested on the gold-standard dataset containing true negative examples instead of dissimilarity-derived ones. MARVL is an open-source pipeline that can be trained for other pathogens, and is freely available at <https://github.com/oigorlobo/MARVL>.

**Keywords:** machine-learning, reverse vaccinology

**Acronyms:** RV – Reverse Vaccinology, ML – Machine-Learning, OCC – One-class classifier, BC – Binary Classifiers, IF – Isolation Forest, OCSVM – One-class Support Vector Machine, RF – Random Forest, BCSVM – Binary-class Support Vector Machine, PCPU - Performance criteria for positive-unlabeled classification.

## 1. INTRODUCTION

Many efforts have been made to improve vaccine development by establishing automated methods to overcome the known limitations of traditional vaccine development pipelines. Out of those constraints, we mention 5-15 years to get a final product, only cultivable pathogens can be approached, and only antigens expressed *in vitro* can be identified [1,2]. A major initiative in this direction leveraged by the increasing availability of genomic information is a research field named Reverse Vaccinology (RV). RV has boosted progress in vaccine development by starting the selection of vaccine candidates through an initial computational screening of all known protein-coding genes in genomic sequences from parasite genomes [1,3].

As RV matured, a few initiatives to develop and share RV pipelines have been made available in the scientific literature (Supplemental Table 1). The first developed RV tools were based on filtering strategies: the decision process was based on somewhat arbitrary feature cutoffs defined by the developers or users [4,5]. Later, in trying to mitigate this human interference factor, machine-learning (ML) techniques were incorporated into RV pipelines and, since then, have been successfully used to handle the massive amount of biological data available in several complex tasks [6].

To the best of our knowledge, Vaxijen was the first RV pipeline to use ML to select vaccine candidates [7]. Other methods followed, including the Bowman-Heinson predictor [8,9], Vacceed [10], culminating in the current state-of-the-art, a new version of Vaxign-ML [11], Vaxign2 [12], and Vaxi-DL [13]. As assessed by (Dalsass et. al 2019), ML-based RV protocols outperform filtering RV pipelines, with the Bowman-Heinson predictor arguably being the most successful one at that point, recalling 75% of bacterial protective antigens.

Despite the successes of existing RV tools, there are still several opportunities for improvement, once they present limitations. Some of these limitations are related to the composition of training datasets. A major one is related to the selection of the

training sets for ML models used for vaccine discovery. These models are usually binary classifiers (BC) expected to learn separation surfaces between protective antigens (PAGs) and non-PAGs from training sets ideally containing positive (i.e., known PAGs) and negative (i.e., known non-vaccines) examples. However, throughout the literature, most, if not all, models are trained on sets containing known PAGs, but with counterexamples selected from sets of unlabeled sequences based on sequence dissimilarity to the positive instances. The use of sequence dissimilarity as a proxy for the difference in immunogenicity is based on the assumption that any sequence sufficiently dissimilar from a known vaccine candidate can be safely selected as a non-vaccine. A cutoff of 30% similarity is commonly adopted to select proteins sufficiently dissimilar from known vaccines to be used as training negative examples [12,13].

Although this scarcity of experimentally validated negative instances in the context of RV cannot be easily remedied, there are possible ways to minimize this problem. One-class classifiers (OCCs) are ML models that are trained for binary classification in the absence of counterexamples. OCCs have achieved satisfactory results in some biological knowledge fields, such as in determining metabolite dynamics on breast cancer [14], and have outperformed regular BCs in the identification of genes associated with certain diseases [15,16]. Even though OCCs may be an interesting approach to detect vaccine candidates without relying on the absence of a true negative dataset, we are not aware of any initiative in this direction.

In this work, we present MARVL (Machine-learning Assisted Reverse VaccinoLogY), an end-to-end RV pipeline for training general ensemble ML classifiers to predict vaccine candidates. MARVL aims to address issues identified in existing ML-based RV tools by using several complementary strategies to develop a robust and general ML approach. Here, we use bacteria as a case study to enable comparability with existing approaches, providing curated sets of experimentally validated positive and negative examples extracted from the literature to be used as holdouts. These are expected to provide a gold-standard to evaluate not only the current work but also future RV initiatives. Our tool computes several statistical, physicochemical, and functional attributes that can be extracted from protein sequence data alone, together with engineered features to improve classifier performance. We evaluated several classes of ML models, including OCCs, for their relative performance on our gold-standard dataset. Then, we selected the best ones

to compose an ensemble model that was compared against the state-of-the-art tools. MARVL is freely available at <https://github.com/oigorlobo/MARVL>.

## 2. METHODS

Figure S1 represents the conceptual steps for the initial development and testing of candidate models for MARVL, including data sources and similarity-aware data splitting for model training. To explore how distinct ML models work, we trained two types of models: OCCs and several classes of BCs. Next, we describe procedures used for the creation of the positive, unlabeled, and negative datasets for model training and performance assessment.

### 2.1. Datasets

We selected known bacterial vaccines from the Protegen Database, which hosts experimentally validated proteins able to induce immunological protection [17]. As of November 2022, there were 660 known PAGs from 14 Gram-positive and 36 Gram-negative species. We downloaded a fasta file consisting of all the bacterial amino acid sequences and performed data cleaning steps, namely the removal of sequences containing non-canonical amino acids and of sequences shorter than 10-amino acid residues, and the consolidation of highly similar sequences ( $\geq 97\%$  of identity and  $\geq 98\%$  of similarity) by retaining only one sequence from each high similarity group and removing the others. The resulting set of positive examples comprises 588 sequences (TableS2, “Positive\_Protegen” sheet).

We created a set of “pseudo-negative” examples using the same sequence similarity-based strategy used in existing RV methods, as detailed in the introduction, for use in the training (but, critically, not the evaluation) of BCs. To this end, we downloaded the available proteomes from species present in Protegen from Uniprot [18] (Supplemental Table 2, “Genome\_IDs” sheet). Following the standard used in the literature ([9]; [11]), sequences with similarity to those in the positive set lower than 30% were used as potential counterexamples in the training of BCs. We

randomly selected 588 of these sequences as the “pseudo-negative” set (Supplemental Table 2, “Negative\_similarity” sheet).

We also produced a manually curated gold-standard of 25 positive and 10 negative examples using PADA-ONE, a web tool recently developed by our group that hosts curated proteins of immunological relevance [19] (Supplemental Table 2, “Gold\_standard” sheet). Positive examples are new vaccinal proteins not found in Protegen, while negative examples are curated entries from the scientific literature comprising proteins that failed to confer protective immunity after experimental validation. This negative dataset was not used at any point for model development but was kept as a holdout to benchmark our method against the most recent available tools (Figure S1).

## 2.2. Distribution of similarity for known vaccines and non-vaccines

We used the gold-standard dataset described above to evaluate the similarity distribution between the new experimentally validated positive and negative examples and the Protegen entries. For that purpose, we used BLASTP to compute pairwise alignments between all positive and negative examples from the gold-standard dataset and the Protegen proteins, obtaining one similarity value for each unique pair of sequences from the gold-standard and Protegen (Supplementary material, section ‘1.2. BLAST alignment gold-standard and Protegen entries’).

## 2.3. Feature calculation

We developed an *in-house* python script to extract distinct groups of attributes from a variety of computational sources. We included features expected to capture statistical, physicochemical, and biologically meaningful attributes commonly used to develop ML-based RV tools. We also include engineered features (see Supplementary material, section ‘1.3. Feature description’. See also Supplemental Table 3 for full feature description). This script is distributed with MARVL.

## 2.4. Pre-processing of features and dimensionality reduction

As a first pre-processing step, a skewness analysis was performed and highly skewed features were log-transformed, both with and without outlier truncation. Then, we further compared two types of data transformation, namely standardization and normalization of data (See Supplementary material, section ‘1.4. Data transformation’).

Our dimensionality reduction strategy was guided by two principles: not relying on class labels, since only positive examples were considered reliably labeled in our training datasets; and retaining direct explainability of features (i.e., not using data transformations such as PCA). To this end, we first removed features with near-zero variance, which commonly do not contribute to the predictive ability of models. To this end, we used the function `nearZeroVar()` from the R package “caret” [20]. Subsequently, we performed correlation-based filtering of redundant features using the function `findCorrelation()` from the same R package, with a cutoff value of 0.75 to remove highly correlated attributes.

## 2.5. Building individual machine learning models

As mentioned earlier, we tested both OCCs and BCs for building vaccine predictors. The OCCs tested are Isolation Forest (IF), One-class Support Vector Machine (OCSVM), and Autoencoder (AE). The binary ones are Random Forest (RF) and Binary-class Support Vector Machine (BCSVM) (See Supplementary material, section ‘1.5. Models description’).

We used 70% of our dataset of positive and unlabeled data for model training, with the remaining observations reserved for testing (Figure S1). The splitting step was performed taking into account sequence coverage and identity to ensure that sequences sufficiently similar belong to the same data split to minimize data leakage through sequence similarity from the training and testing datasets. We used CD-HIT v.4.8.1 [21] to split sequences while considering pairwise identity to ensure that all sequence pairs with more than 75% of identity were kept in the same split. OCC models were trained using only positive instances from the training dataset, whereas both positive and unlabeled instances (taken as a “pseudo-negative” set) were used

to train BCs.

For both OCCs and BCs, positive and unlabeled testing datasets were used to evaluate individual classifier performance. Supplemental Table 4 summarizes the R packages, functions, and parameters/arguments used to train the models. RF, OC, and BCSVM were trained using 10-fold cross-validation (CV). Full models were trained with no dataset splitting, using all the instances from positive and unlabeled datasets (only the positive dataset in the OCCs case).

## 2.6. Performance assessment, model comparison, and ensemble building

Multiple quality measures were calculated in order to assess different aspects of performance of the fitted models: the counts of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN); accuracy (Acc); sensitivity (Sens) and specificity (Spec); precision (Pr; also known as positive predictive value, PPV) and negative predictive value (NPV); the Matthews Correlation Coefficient (MCC); area under the ROC curve (AUC); and a performance criterion for positive-unlabeled classification (PCPU) [22]. This last one, PCPU, is a performance measure applied to classification tasks where there is a positive and an unlabeled dataset, a typical scenario of one-class classification problems. The calculation of PCPU was performed as follows:

$$PCPU = \frac{Sens^2}{\hat{p}(y+)}$$

where  $\hat{p}(y+)$  is the estimated probability that an instance is labeled as positive, which is calculated from the output of a model. Importantly, this performance measure can be used in the testing step to compare OCCs and BCs [22].

We detected highly correlated models by using the vectors of predictions of the testing dataset from each model. Then, we retained only one model from each high pairwise correlation (correlation  $\geq 0.85$ ) and removed the other. From the remaining models, we selected the 4 best BCs based on AUC values, together with the OCC with the most balanced correct predictions between TPs and unlabeled instances (“TN”, Supplemental Table 5) to compose a final ensemble. Our ensemble

model uses a simple majority rule voting schema where each of the five models classifies a sample as either belonging to the positive or negative class.

We used the gold-standard (25 positive and 10 real negative examples retrieved from PADA-ONE) to compare our ensemble model against Vaxign2 [12] and Vaxi-DL [13]. As the gold-standard is not present in the training dataset of our classifiers nor in the training datasets of Vaxign2 and Vaxi-DL, the performance measure values obtained here work as an indicator of the generalization capacity of all models.

### 3. RESULTS

#### 3.1. Sequence dissimilarity is a poor proxy to distinguish between vaccines and non-vaccines

A common premise in the field of ML-based RV is that any sequence that is sufficiently dissimilar (usually similarity  $\leq 30\%$ ) from known vaccines can be labeled as a non-vaccine. However, this premise has never been evaluated as far as we are aware. We objectively evaluated this premise by computing pairwise alignments of all positive and negative entries from our gold-standard (25 and 10 entries, respectively) against the filtered Protegen dataset (removal of 72 sequences, resulting in 588 entries), which produced 14,700 and 5,880 data points, respectively. If the assumptions for the construction of entries belonging to the negative class are true, one would expect most negative examples from our gold-standard to have considerably smaller similarity scores with known vaccines when compared with our gold-standard positives.

We found the distributions of the alignment similarities from both positive and negative gold-standards to be highly similar, with both heavily biased towards values smaller than 30% (Figure 1). The alignment identities have a similar distribution profile (Figure S3). In fact, only 347 and 162 positive and negative alignments, respectively, have similarity greater than 30% (2.4% and 2.76% of the data points for each class, respectively). We observed a reasonable number of non-alignments (which we considered as meaning a similarity to Protegen entries equivalent to zero) in both the positive (3387/14699 alignments, 23.05%) and negative (1167/5879

alignments, 19.85%) examples of the gold-standard dataset. These were removed from the data used to generate the plots in Figure 1.

In terms of maximum similarity to any Protegen entry (which would be used to select counterexamples in the similarity-based approaches) we again report a high overlap in terms of the values observed for both the positive and negative examples in our gold-standard dataset (Figure S5). Curiously, not a single example from the gold-standard dataset presented a maximum similarity to Protegen lower than the usual threshold of 30% - meaning that none of the actual negatives in our dataset would have been selected as a counterexample.

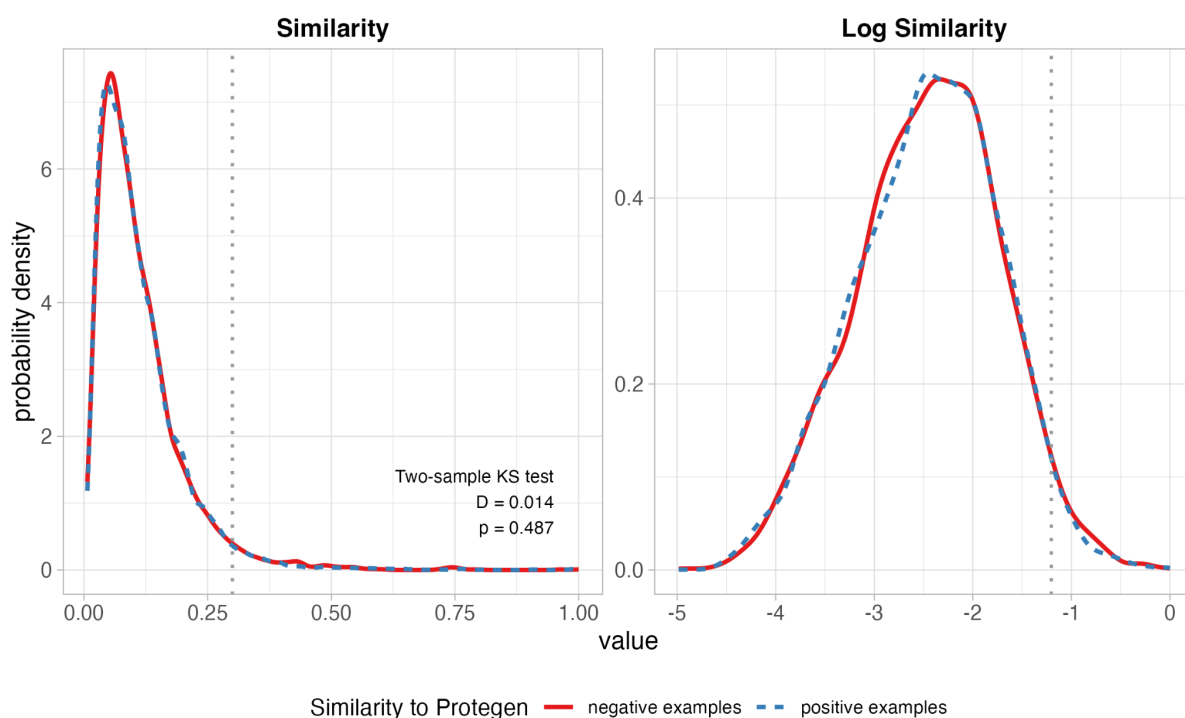


**Figure 1. Estimated probability density functions of pairwise alignment similarities (as raw values and as log transformed values) of known vaccines and non-vaccines (gold-standard dataset) to an independent set of known vaccines (Protegen).** Red solid line: negative class (experimentally validated non-vaccines); Green dashed line: positive class (experimentally validated vaccines). The vertical gray dotted line highlights the usual 30% similarity threshold. The two distributions cannot be distinguished at the 95% confidence level (Kolmogorov-Smirnov test,  $D = 0.014$ ,  $p = 0.487$ ). Queries that did not return any alignment were considered as having zero similarity and were not included in this figure (See Figure S4).

### 3.2. Dataset splitting and feature preprocessing

We downloaded from Uniprot all the available bacterial proteomes that originated the 588 positive examples from Protegen, resulting in 119,288 amino acid sequences. Those instances were aligned against the positive entries using BLAST (e-value cutoff  $< 1e-5$ ), resulting in the removal of 10,553 protein sequences. We randomly selected 588 examples from these sequences and labeled them as putative non-protective antigens (“pseudo-negative”) in order to produce a balanced training dataset.

We proceed by using a similarity-aware splitting where highly similar sequences are all allocated to the same split to avoid potential similarity-induced data leakage between the splits [23]. Our final training dataset has 409 positive (Protegen) and 414 “pseudo-negative” (unlabeled) instances (~70%), while 179 positive and 174 “pseudo-negative” examples were kept for testing (~30%) (Supplemental Table 2). Our pipeline computes a total of 9,799 features, including physicochemical, biologically functional, and engineered attributes (Supplemental Table 3). After the removal of features with small variance and filtering based on those high correlations, the final number of features used to train models ranged from 703 to 3613 (Supplemental Table 6 contains feature counts before and after feature reduction).

### 3.3. Training and performance evaluation of OCCs and BCs

We trained OCCs using only positive instances from the training dataset, while BCs were fitted using all positive and “pseudo-negative” instances from the training split. Two out of the three OCC models used in our analysis also required a cutoff to define relevant/anomalous predictions: autoencoders and IFs require the setting of Reconstruction MSE and Anomaly Score thresholds, respectively. PCPU is commonly used as a performance criterion when it is necessary to define any important parameter, as in the original paper describing this metric [22]. The authors applied it as a metric to set a regularization parameter or to find the best combination of parameters to tune an ML algorithm [24]. To define these parameters, we considered the minimum and maximum values for Reconstruction MSE and Anomaly

Score during the training step, as described in Supplementary material, section ‘1.6. Model training, tuning, and evaluation’.

[22] also demonstrated how PCPU can be used to compare models in a dataset of positive and unlabeled instances, as is our case when considering the dataset composed of positive examples (true vaccines) and “pseudo-negative” examples built using dissimilarity as a criterion. In general, BCs have a better performance when compared with OCCs (Supplemental Table 5, ‘PCPU’ column). When considering OCC performance, we found autoencoders and isolation forests to be overly conservative, with high false negative rates (Supplemental Table 5, ‘FN’ column). OCSVMs have more success classifying negative instances, in particular when analyzing the standardized dataset (OCsvm\_S model, Supplemental Table 5). When considering BC performance, both SVMs and RFs have delivered robust models (Figure S6, Supplemental Table 5).

### 3.4. Building and validation of an ensemble classifier for reverse vaccinology

Building model ensembles is a strategy adopted by major reverse vaccinology tools available, such as Vacceed and Vaxi-DL [10,13]. We build MARVL (Machine-learning Assisted Reverse VaccinoLogY), an ensemble model aimed at offering a robust performance in a wide range of scenarios, by considering several aspects of our individual models, including model class diversity, normalization procedures, performance, and the correlation of models in the output vector of the testing dataset (Methods, section ‘2.6. Performance assessment, model comparison, and ensemble building’). We proceed by selecting five models (four BC and one OCC) that include a diverse set: BCsvm\_N, BCsvm\_OS, RF\_ON, RF\_S, and OCsvm\_S (in bold in Supplemental Table 5). Figure 2 contains the correlation for all models in the testing dataset.

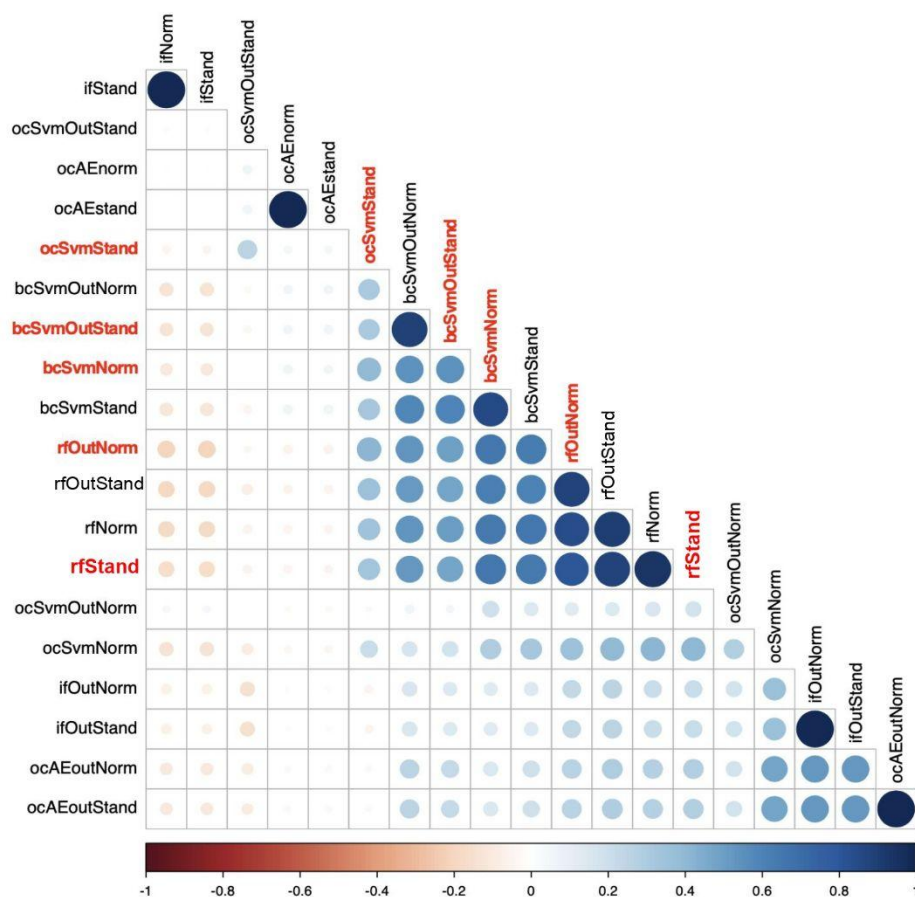


**Figure 2. Correlation plot of the performance of individual models in the testing dataset.** Color intensities/circle sizes are proportional to correlation (blue: positive correlations, red; negative correlations). Models selected to build MARVL are highlighted in red. **ocAE\_\***: autoencoder (one-class classifier); **if\_\***: isolation forest (one-class classifier); **ocSvm\_\***: support vector machine (one-class classifier); **bcSvm\_\***: support vector machine (binary classifier); **rf\_\***: random forest (binary classifier); **\*\_Norm**: normalized dataset; **\*\_OutNorm**: normalized dataset with truncated outliers; **\*\_Stand**: standardized dataset; **\*\_OutStand**: standardized dataset with truncated outliers.

We evaluated the performance of Vaxi-DL and Vaxign2 when compared with MARVL using the holdout dataset of 25 positive and 10 negative examples manually curated from the PADA-ONE database. We stress that these 10 negative examples are proteins experimentally confirmed not to elicit protective immunity after challenge, and represent, to the best of our knowledge, the first set of true negatives to objectively evaluate reverse vaccinology tools.

To compare MARVL against the state-of-the-art RV tools using our gold-standard dataset of experimentally-validated positive and negative examples

(PADA-ONE dataset), we focus primarily on using the Matthews Correlation Coefficient (MCC), a data imbalance-resistant quality measure that assesses the overall performance of the models [25]. Other performance metrics are also reported in Table 1, together with their bootstrap standard errors. It is important to highlight that although we report AUC for completeness, the AUC values for the competing models are very likely to be negatively biased since predicted probabilities, which are essential to a proper calculation of AUC, are not returned by those models. As such, comparisons against the benchmark models using AUC values should not be considered as valid.

The MCC scores of current RV tools suggest a performance no better than random classification (-0.04 for Vaxi-DL, -0.12 for Vaxign2), mostly due to their very low specificity values. In contrast, MARVL achieved a reasonably solid performance of 0.326, indicating a balanced predictive performance across both the positive and negative examples. This is further evidenced by the good value of sensitivity and excellent specificity of MARVL in comparison with the existing RV tools.

Vaxi-DL and Vaxign2 were able to correctly predict most of the positive examples (19 and 17 for Vaxi-DL and Vaxign2, respectively. See Table 1), with MARVL also finding the majority of positives (14), although it was more conservative than the available tools. This more conservative behavior resulted in better PPV, meaning that the proportion of false positives was particularly low. The available tools failed to correctly label the majority of gold-standard negative observations (2 out of 10 correct classifications for both), whereas MARVL was able to correctly predict 8 out of 10 negative examples from the holdout set. This classification performance indicates MARVL as having a more balanced classification profile, as evidenced by the results shown in Table 1 and further illustrated in Figure 3. Overall, the observed performance of our tool suggests that it is indeed capable of distinguishing class labels better than current state-of-the-art tools when evaluated in a dataset more closely mirroring real-world scenarios.

**Table 1. Comparison of MARVL with Vaxi-DL and Vaxign2 on the PADA-ONE dataset. Reported values follow the convention “estimate  $\pm$  standard error”. AUC values estimated for Vaxign2 and Vaxi-DL are unreliable due to lack of probability scores and should not be used for comparison.**

|                   | MARVL            | Vaxi-DL           | Vaxign2          |
|-------------------|------------------|-------------------|------------------|
| TP                | 14               | 19                | 17               |
| FP                | 2                | 8                 | 8                |
| TN                | 8                | 2                 | 2                |
| FN                | 11               | 6                 | 8                |
| Accuracy          | 0.63 $\pm$ 0.083 | 0.6 $\pm$ 0.081   | 0.54 $\pm$ 0.086 |
| Sensitivity       | 0.56 $\pm$ 0.1   | 0.76 $\pm$ 0.085  | 0.68 $\pm$ 0.097 |
| Specificity       | 0.8 $\pm$ 0.14   | 0.2 $\pm$ 0.13    | 0.2 $\pm$ 0.13   |
| PPV               | 0.88 $\pm$ 0.086 | 0.7 $\pm$ 0.087   | 0.68 $\pm$ 0.095 |
| NPV               | 0.42 $\pm$ 0.12  | 0.25 $\pm$ 0.16   | 0.2 $\pm$ 0.14   |
| F1                | 0.68 $\pm$ 0.086 | 0.73 $\pm$ 0.069  | 0.68 $\pm$ 0.079 |
| MCC               | 0.33 $\pm$ 0.15  | -0.043 $\pm$ 0.16 | -0.12 $\pm$ 0.16 |
| Balanced Accuracy | 0.68 $\pm$ 0.086 | 0.48 $\pm$ 0.076  | 0.44 $\pm$ 0.082 |
| PCPU              | 0.69 $\pm$ 0.16  | 0.75 $\pm$ 0.11   | 0.65 $\pm$ 0.13  |
| AUC               | 0.65             | 0.48*             | 0.44*            |

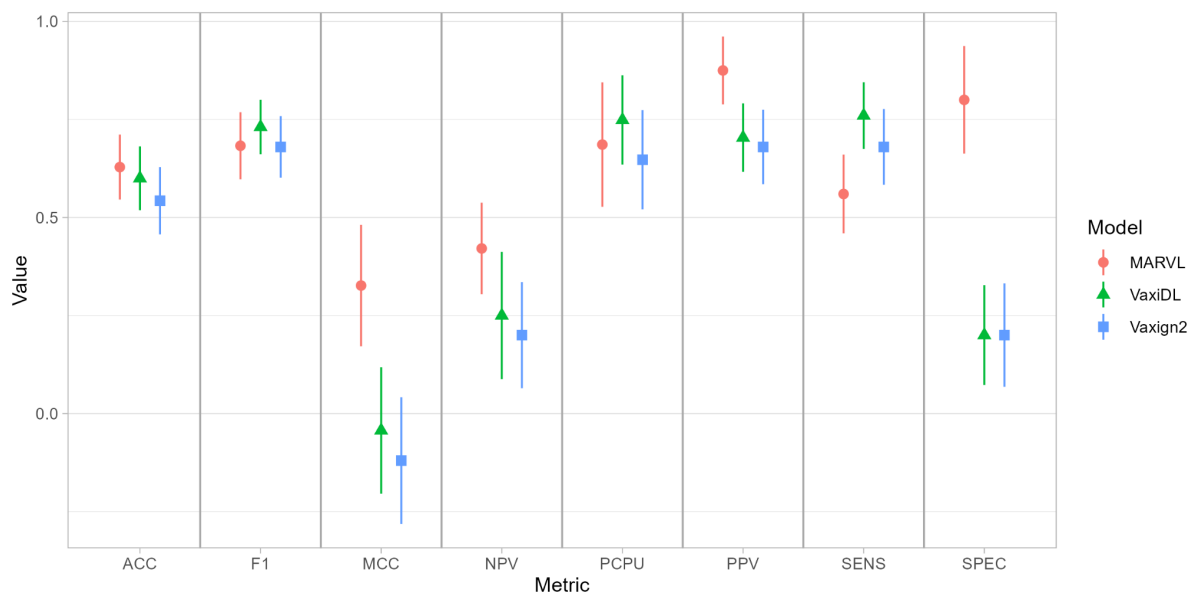


**Figure 3. Visual comparison of performance estimates for MARVL, Vaxi-DL and Vaxign2.** In red, MARVL; in green, Vaxi-DL; in blue, Vaxign2. ACC: accuracy; F1: F1 score; MCC: Matthews Correlation Coefficient; NPV: negative predictive value; PCPU: performance criteria for positive-unlabeled classification; PPV: positive predictive value; SENS: sensitivity; SPEC: specificity.

### 3.5. Feature importance

After evaluating MARVL performance, we selected the 20 features deemed as the most important by each of the five models that compose our ensemble to gain a better understanding of the underlying behavior of MARVL. As a general profile, we found the most important features to be model-specific (66 out of the 100 features are specific to individual models, see Figure 4). This suggests our strategy to eliminate highly associated models may have successfully selected models that captured distinct properties of vaccines, which may partially explain the better performance of MARVL in comparison with other tools.

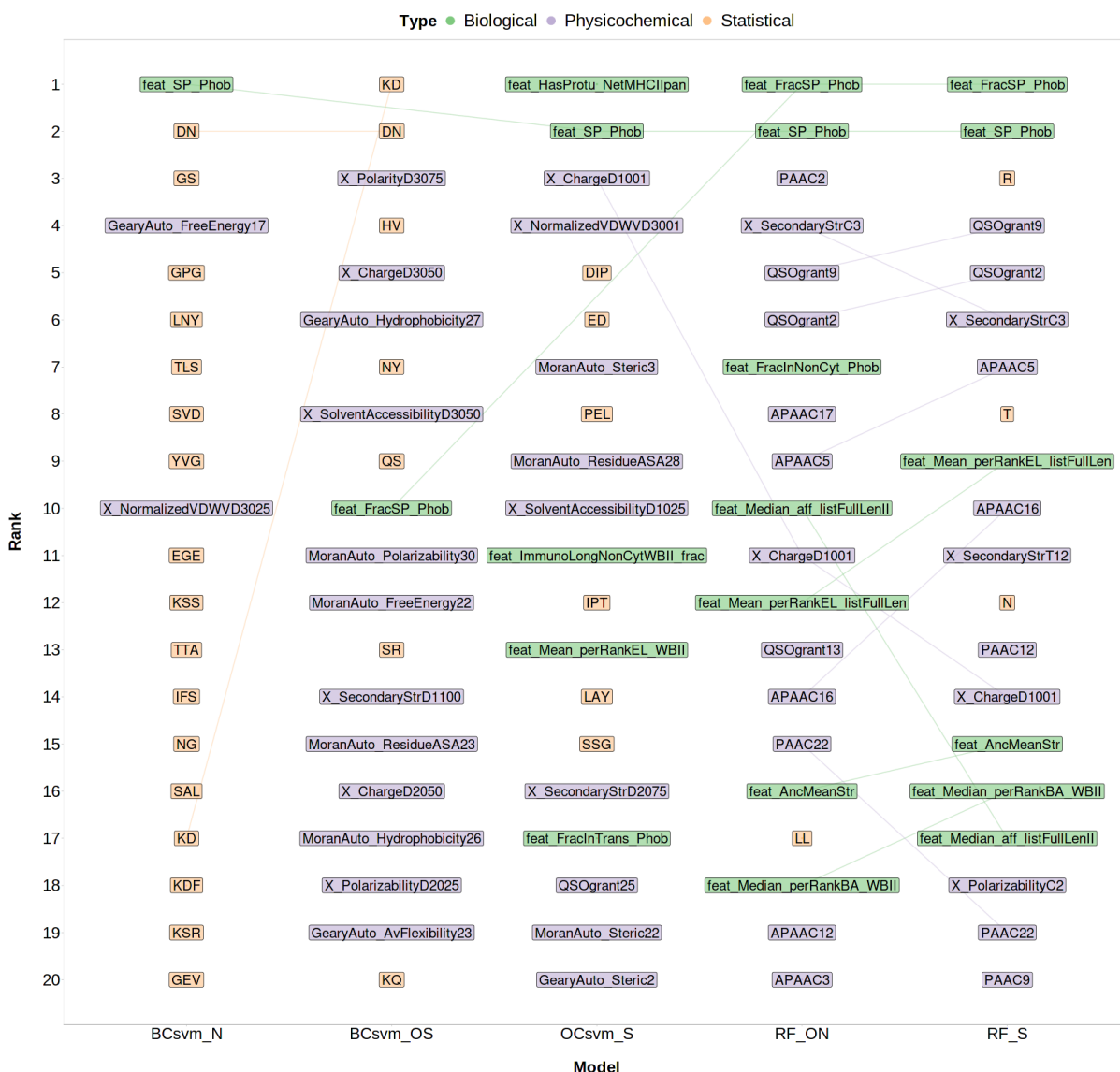


**Figure 4. Top 20 most relevant features for each model that compose MARVL.** In green, Biological features; in purple, Physicochemical features; and in orange, Statistical features. Models captured different attributes of protective antigens, with 66 of them not shared across predictors. **BCsvm\_N**: binary classifier support vector machine (SVM) with normalized dataset; **BCsvm\_OS**: binary classifier support vector machine (SVM) with standardized and truncated outliers dataset; **OCsvm\_S**: one-class support vector machine (SVM) with standardized dataset; **RF\_ON**: binary classifier random forest (RF) with normalized and truncated outliers dataset; and **RF\_S**: binary classifier random forest (RF) with standardized dataset.

We also observe that most models appear to select mostly physicochemical or biological attributes, with the exception of SVM trained using normalized data (BCsvm\_N), which was clearly enriched on statistical attributes. Random Forest

models are more influenced by physicochemical features, followed by the biologically functional ones. On the other hand, BCsvm\_OS and OCsvm\_S were more affected by physicochemical and statistical features, such as autocorrelation attributes and di- and tripeptide composition. The BCsvm\_N model was mostly influenced by statistical features, mainly by tripeptide composition attributes.

The analysis of features shared across models highlights the importance of specific biological and physicochemical properties successfully to identify vaccines. We mention the relevance of signal peptides, an attribute affecting all of the models, either through the feature “feat\_SP\_Phob” (presence/absence of signal peptide, selected amongst the two most important features by four out of five models) or through the feature “feat\_FracSP\_Phob” (fraction of the protein occupied by the signal peptide region, selected amongst the ten most important features by three out of five models).

Signal peptides are amino acid stretches identified by other cytosolic proteins to address proteins containing this sequence to the membrane, where they can be inserted or secreted to the extracellular medium [26]. Features representing signal peptide have played, along with subcellular localization, a relevant role to train ML-based RV tools [11].

Most of the other shared attributes are observed in the two Random Forest models, potentially as a result of their shared model induction algorithm. Those include features related to scores of binding affinity of MHC ligands, such as the feature “feat\_Median\_perRankBA\_WBII” - the median value of the binding affinity scores for weak ligands considering the MHCII alleles used in our study.

Other interesting functional features can affect the classification task of some models, such as “feat\_FracInNonCyt\_Phob” for the model RF\_ON and “feat\_ImmunoLongNonCytWBII\_frac” for the model OCsvm\_S. The attribute “feat\_FracInNonCyt\_Phob” means the portion of the protein that is outside the cytoplasm and “feat\_ImmunoLongNonCytWBII\_frac” corresponds to the relative length of the protein that corresponds to MHCII weak ligand regions, present in a long domain outside the cytoplasm. This latter is one of our new engineered features.

## 4. DISCUSSION

The first ML-based RV tool designed to predict protective antigens in bacterial species was Vaxijen [7], followed by several other initiatives, including the Bowman & Heinson predictor [8,9], Vaxign-ML [11], which was afterward upgraded to Vaxign2 [12], and, more recently, Vaxi-DL [13]. All these tools implement binary classifiers which, in addition to positive examples (known vaccines), require examples of non-vaccines for model training and assessment.

We found that virtually all available tools rely on somewhat arbitrary similarity cutoffs to define negative examples. For instance, both Vaxign-ML and Vaxi-DL have considered a threshold of 30% of similarity to define negative classes. This strategy is based on the assumptions that sequence dissimilarity acts as a good proxy for distinct immunogenic properties and that most proteins from any given pathogen are not expected to share functional similarities with known vaccines and, therefore, are unlikely to work as vaccines. However, despite high sequence similarity being indeed a frequent good proxy for functional similarity, we are aware of no good evidence or theoretical argument suggesting the reverse to be true, that is, that PAg should exhibit a greater average within-group similarity than the average PAg-non-PAg one. This means that true vaccine candidates that are sufficiently dissimilar from known vaccines may be wrongly included in the set of training negative examples. Indeed, blind confidence in this assumption could be interpreted as a formal logical fallacy, namely *denying the antecedent* [27]. More importantly in the context of RV, it could lead not only to impacts on the generalization ability of ML models trained using "pseudo-negative" sets derived using exclusively this strategy (since the models may end up learning separation surfaces that are little more than sequence similarity filters), but also to an impossibility of properly assessing the expected error rates of these models when deployed to classify new sequences as PAg vs non-PAg.

This sequence dissimilarity-based strategy for composing both the training and validation sets is widely adopted in the field, mainly due to the scarcity of published negative evidence of protective immunogenicity. Vaxijen, Vaxign-ML, and Vaxi-DL all selected negative examples based on sequence dissimilarity, while Vacceed went even further and developed the full model using putative vaccine candidates (instead of experimentally validated examples) as positive instances.

Despite the use of sophisticated modeling approaches in existing RV methods, the definition of datasets for model development and validation has received less attention. This makes a critical evaluation and comparison of distinct ML-based RV algorithms particularly difficult.

Thus, in this study, we used PADA-ONE [19] to extract a set of experimentally validated non-PAGs, together with a list of new vaccines that have not been used to train any of the currently available tools. This required a considerable curation effort of articles describing the experimental assessment of vaccine candidates to find true negative examples, defined as vaccine candidates that fail to provide protection. Nevertheless, we believe using these entities as negative holdout examples when evaluating RV tools was necessary due to the fact that distinguishing between true positives and true negative examples in the context of vaccine discovery is expected to be a far more challenging task than discriminating between proteins that are intentionally selected to be very dissimilar.

The true negative examples we found are mostly antigens that were reported as a byproduct of screenings for vaccine candidates. Therefore, those sequences are expected to share several interesting properties from an immunological perspective that led to scientists being willing to undergo the efforts of further characterizing them through costly and time-consuming experiments [28]. As a consequence, the true performance of RV tools is likely to be overestimated, as these tools were evaluated in a potentially much easier classification task than the true challenge of distinguishing between the experimentally validated positive and negative examples represented in our gold-standard dataset.

The distribution of sequence similarities of both positive and negative examples of the gold-standard dataset with previously known vaccines is remarkably similar (Figure 1), demonstrating how sequence similarity fails to be a good proxy to distinguish between classes when building datasets for model training and evaluation. This observation may also explain the poor performance of state-of-the-art models when evaluated with the gold-standard dataset, as these models have been previously evaluated using putative negative classes built using similarity as the criterion for class pertinence.

We built our ensemble classifier (MARVL) through the systematic evaluation of a diverse set of features, data preprocessing protocols, and classes of machine learning models. These include, to the best of our knowledge, the first report on the

performance of OCCs in the field of RV. Even though the OCC models did not outperform the BC approaches, some OCCs presented interesting properties regarding their ability to correctly identify positive entries without being overly optimistic (e.g. the OCsvm\_S model).

We observed several criteria to select the models to build MARVL. In particular, we highlight the selection of diverse models in terms of model type, data preprocessing, as well as reasonable performance and low correlation of prediction results (Figure 2). By selecting models with satisfactory performance in the testing dataset while preserving model class diversity and actively selecting models with low prediction correlations, we aimed at building a robust ensemble expected to present good performance in a broader range of scenarios. Importantly, as some of our models were trained using “pseudo-negative” examples analogous to those used by existing tools, the strategy to build our ensemble from a diverse set of individual models and to include OCC as part of the ensemble could help explain the better-observed performance of MARVL in classifying the examples from the gold-standard dataset.

Existing ML-based RV tools have been shown to be overly optimistic when evaluated with our gold-standard, presenting higher TP rates but also considerably higher FP rates (and, consequently, proportionally lower TN rates) when compared with MARVL. The classification pattern of the existing tools generates a performance close to the one of a random classifier (see e.g. MCC values in Table 1). Even though MARVL is still levels of performance seen in other predictive problems in bioinformatics, our tool outperforms available pipelines by a considerable margin in aggregate metrics (such as MCC and AUC) and, importantly, a substantially better performance in terms of Precision, indicating that the positive predictions returned by MARVL are generally more likely to correspond to actual vaccine candidates. This indicates that our tool is likely to be more effective in the real-world scenario comprising the much harder task of distinguishing between experimentally validated examples of vaccines and non-vaccines.

Another point of concern for several of the currently available ML-based RV programs is the use of features mostly available for well-studied pathogens, such as gene expression data, features based on network topology, and functional domain attributes (Supplemental Table 1). This may limit the usability of these models and prevent their usage for groups of recently sequenced parasites, such as the

emergent ones and neglected tropical diseases [29,30]. Therefore, the development of predictors based on features that can be computed from primary protein sequence data alone is essential, if we aim to develop more general classifiers [31,32].

Thus, MARVL computes features that can be essentially extracted from amino acid sequences. In this regard, most of the features deemed as of high importance for model performance in our ensemble are exclusive to one model, suggesting our approach to select models with small classification correlation in the testing dataset successfully selected models that captured distinct properties of the training dataset. Two attributes representing the presence of signal peptide were the only ones deemed as important in more than two models, highlighting the importance of proteins located at the host-parasite molecular interface for the development of RV tools [11]. Interestingly, the addition of signal sequences to antigens for immunogenicity purposes has recently presented satisfactory results [33,34], further demonstrating the crucial role of signal peptides in the immunogenicity of vaccines.

Among the features deemed as relevant for individual models, we mention two of them that are biologically expected to be of importance in the immunogenicity of PAgS. The attribute “feat\_FracInNonCyt\_Phob” corresponds to the fraction of a protein that is outside the cytoplasm at the molecular interface of the host-pathogen interaction. The second one, “feat\_ImmunoLongNonCytWBII\_frac”, represents weak MHCII ligands in long non-cytoplasmic domains. This attribute adds to the known relevance of T-cell activation for the immunological role of PAgS [35,36]. It is worth highlighting that this feature is also an engineered one and one of the novelties of our work.

## 5. CONCLUSION

The tools currently available for RV, and the databases used to validate these tools, lack statistical rigor in their development and performance evaluation. The first RV programs, such as Nerve, Jenner-predict, Vacsol, and even more recently Pan-RV, were developed as filtering-based softwares. Thus, there is an essential human interference in establishing which features are the most important ones to label an antigen as a PAg or not. The emergence of ML-based approaches based on two-class classifiers provided a solution for this issue; however, these procedures may have introduced other issues in the prediction of PAgS, such as the need for the

definition of a secondary class comprising negative examples in order to train binary classifiers [12,13]. All of the ML-based tools available create a set of counterexamples based on sequence dissimilarity. By producing a list of curated non-vaccinal proteins and comparing it with known vaccine candidates we demonstrated that this premise, although reasonable, does not appear to be a good proxy for negative examples. Furthermore, the lack of a similarity threshold to define the training and testing datasets in model training is likely to produce data leakage and overestimate classifier performance [23].

Furthermore, as our set of negative examples has been experimentally evaluated for vaccinal protection through costly and time-consuming experiments, they are expected to share enough shared properties with successful vaccines to be labeled as potential vaccines. Due to this fact, distinguishing between our set of true negative and true positive examples of vaccines is expected to be a more challenging task.

We conclude that exclusively using similarity as a proxy to distinguish between vaccines and non-vaccines in order to build a negative class is not supported by evidence. This is not, however, decisive evidence that it is a poor strategy, only an indication of the lack of evidence of it being a sound one. We posit that methods trained using datasets derived using sequence dissimilarity as a proxy for different class may yield reasonably good performance in tests not because this represents a good strategy, but rather due to the expected low base rate of PAgS in any pathogen's genome, which makes most sequences, regardless of their similarity to known PAgS, likely to be negative. We argue, however, that this remains a dangerous strategy, since ML models may learn to distinguish differences in similarity rather than in the ability to elicit protective immunity, leading to potentially much higher misclassification rates when deployed.

MARVL is an ensemble machine-learning strategy designed to account for these biases using a machine-learning strategy that involves a similarity-aware data splitting in training, together with feature engineering to capture biological knowledge expected to be relevant for vaccine prediction. We also investigated the performance of a large variety of machine-learning models, including one-class classifiers that have not been previously evaluated in the context of vaccine candidate prediction. We built our ensemble classifier to include robust individual models while removing highly correlated models in the testing dataset. When using our curated dataset of

true positive and negative examples, we found MARVL to outperform current state-of-the-art ML-based RV tools for most performance metrics. Our tool is freely available at <https://github.com/oigorlobo/MARVL>.

## 6. REFERENCES

1. Rappuoli R. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 2001; **19**:2688–2691
2. Del Tordello E, Rappuoli R, Delany I. Reverse Vaccinology. *Human Vaccines* 2017; 65–86
3. Pizza M, Scarlato V, Masignani V, *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 2000; **287**:1816–1820
4. Heinson AI, Woelk CH, Newell M-L. The promise of reverse vaccinology. *Int. Health* 2015; **7**:85–89
5. Dalsass M, Brozzi A, Medini D, *et al.* Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery. *Front. Immunol.* 2019; **10**:113
6. Greener JG, Kandathil SM, Moffat L, *et al.* A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* 2022; **23**:40–55
7. Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007; **8**:4
8. Bowman BN, McAdam PR, Vivona S, *et al.* Improving reverse vaccinology with a machine learning approach. *Vaccine* 2011; **29**:8156–8164
9. Heinson AI, Gunawardana Y, Moesker B, *et al.* Enhancing the biological relevance of machine learning classifiers for reverse vaccinology. *Int. J. Mol. Sci.* 2017; **18**
10. 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**:2381–2383
11. Ong E, Wang H, Wong MU, *et al.* Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. *Bioinformatics* 2020; **36**:3185–3191
12. Ong E, Cooke MF, Huffman A, *et al.* Vaxign2: the second generation of the first Web-based vaccine design program using reverse vaccinology and machine

- learning. *Nucleic Acids Res.* 2021; **49**:W671–W678
13. Rawal K, Sinha R, Nath SK, *et al.* Vaxi-DL: A web-based deep learning server to identify potential vaccine candidates. *Comput. Biol. Med.* 2022; **145**:105401
  14. Migdadi L, Telfah A, Hergenröder R, *et al.* Novelty detection for metabolic dynamics established on breast cancer tissue using 2D NMR TOCSY spectra. *Comput. Struct. Biotechnol. J.* 2022; **20**:2965–2977
  15. Yousef A, Charkari NM. A novel method based on physicochemical properties of amino acids and one class classification algorithm for disease gene identification. *J. Biomed. Inform.* 2015; **56**:300–306
  16. Vasighizaker A, Sharma A, Dehzangi A. A novel one-class classification approach to accurately predict disease-gene association in acute myeloid leukemia cancer. *PLoS ONE* 2019; **14**:e0226115
  17. Yang B, Sayers S, Xiang Z, *et al.* Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res.* 2011; **39**:D1073-8
  18. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.* 2008; **36**:D190-5
  19. PADA-ONE: an online tool to explore proteins of immunological relevance in vertebrate pathogens with bibliographic, taxonomic, and molecular metadata. <https://padaone.org/> (14 July 2023, date last accessed).
  20. Kuhn M. Building Predictive Models in R Using thecaret Package. *Journal of statistical software* 2008; **28**:1–26
  21. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012; **28**:3150–3152
  22. Lee WS, Liu B. Learning with positive and unlabeled examples using weighted logistic regression. *ICML* 2003; **448**
  23. Ashford J, Reis-Cunha J, Lobo I, *et al.* Organism-specific training improves performance of linear B-cell epitope prediction. *Bioinformatics* 2021; **37**:4826–4834
  24. Mack B, Roscher R, Waske B. Can I Trust My One-Class Classification? *Remote Sens (Basel)* 2014; **6**:8779–8802
  25. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020; **21**:6
  26. Kaushik S, He H, Dalbey RE. Bacterial Signal Peptides- Navigating the Journey of Proteins. *Front. Physiol.* 2022; **13**:933153

27. Smith P. An introduction to formal logic. 2020;
28. Musser JM. The next chapter in reverse vaccinology. *Nat. Biotechnol.* 2006; **24**:157–158
29. Ellegren H. Genome sequencing and population genomics in non-model organisms. *Trends Ecol. Evol.* 2014; **29**:51–63
30. Metcalf CJE, Lessler J. Opportunities and challenges in modeling emerging infectious diseases. *Science* 2017; **357**:149–152
31. Nigatu D, Sobetzko P, Yousef M, *et al.* Sequence-based information-theoretic features for gene essentiality prediction. *BMC Bioinformatics* 2017; **18**:473
32. Marques de Castro G, Hastenreiter Z, Silva Monteiro TA, *et al.* Cross-species prediction of essential genes in insects. *Bioinformatics* 2022; **38**:1504–1513
33. Maruggi G, Chiarot E, Giovani C, *et al.* Immunogenicity and protective efficacy induced by self-amplifying mRNA vaccines encoding bacterial antigens. *Vaccine* 2017; **35**:361–368
34. Sheerin D, Dold C, Silva-Reyes L, *et al.* Inclusion of a dual signal sequence enhances the immunogenicity of a novel viral vectored vaccine against the capsular group B meningococcus. *Cell Biosci.* 2022; **12**:86
35. Blum JS, Wearsch PA, Cresswell P. Pathways of antigen processing. *Annu. Rev. Immunol.* 2013; **31**:443–473
36. Saylor K, Gillam F, Lohneis T, *et al.* Designs of Antigen Structure and Composition for Improved Protein-Based Vaccine Efficacy. *Front. Immunol.* 2020; **11**:283

## 7. SUPPLEMENTARY MATERIAL

### 7.1. Development of MARVL

#### 7.1.1. BLAST alignment of gold-standard and Protegen entries

We used BLASTP to compute pairwise alignments for the 25 positive and 10 negative entries of our curated gold-standard of proteins not found in Protegen and the 588 vaccine examples from this database. We turned off the seg filter (*-seg no*) to ensure alignments would be returned even for low-complexity regions, and removed redundant hits with smaller quality (*-culling\_limit 1*). To compute overall identity and

similarity, we considered the sums of similar and identical amino acids from pairwise alignments divided by the length of the shortest sequence. Pairwise comparisons returning no significant alignment were considered as having identity and similarity of zero.

### 7.1.2. Feature description

We designed our pipeline to use only features that can be quickly computed from sequence data alone to allow the evaluation of uncharacterized proteins from non-model organisms that lack additional sequence metadata, such as gene expression or protein-protein interaction. For this purpose, we developed an *in-house* python script to compute and extract distinct groups of attributes (Supplemental Table 3). These features are described below.

#### 7.1.2.1. Statistical and physicochemical features

Our *in-house* script invokes methods from Propy 1.0 [1] to extract the following statistical and physicochemical features:

- Amino acid composition: amino acid, dipeptide, and tripeptide composition, totaling 8420 features.
- Autocorrelation: MARVL uses three distinct types of autocorrelation descriptors: 1) Normalized Moreau-Broto autocorrelation; 2) Moran autocorrelation; and 3) Geary autocorrelation. In terms of amino acid sequences, autocorrelation descriptors measure the level of correlation between two residues considering eight amino acid properties: hydrophobicity, average flexibility, polarizability, free energy in water, residue accessible surface area, residue volumes, steric parameters, and mutability.

Moreau-Broto autocorrelation is defined as follows:

$$AC(lag) = \sum_{i=1}^{N-lag} P_i P_{i+lag},$$

where  $lag$  is the distance between contiguous amino acids to be considered in the calculation and  $P_i P_{i+lag}$  represents one of the eight aforementioned properties of the amino acids in the positions  $i$  and  $i+lag$ . Those descriptors are obtained by summing up the products of properties values of contiguous residues at  $lag$  distance [2]. The other two autocorrelation descriptors differ from the Moreau-Broto in the usage of the property values for calculation: Moran autocorrelation uses deviations from the average values and Geary autocorrelation applies square-difference of property values, as described in [3]. Here, we used a  $lag$  value of 30, totaling 720 features.

- Composition, transition, and distribution: this set of features consists of descriptors that refer to composition, transition, and distribution of seven physicochemical or structural properties along the amino acid sequence: polarizability, solvent accessibility, secondary structures, charge, polarity, normalized Van der Waals volume, hydrophobicity [3]. For each property type, amino acids are included in three different groups based on their characteristics. As an example, for polarity, the defined categories are: hydrophobic residues (CVLIMFW), the neutral (GASTPHY), and the polar (RKEDQN) ones. Consider the following amino acid sequence:

K L L K L L L L K K K L L L K L K L L L L K K

There are 10 group K residues and 16 group L residues. The composition for this sequence is calculated as follows:  $100n_K/(n_K + n_L) = 38.46\%$  for K and  $100n_L/(n_K + n_L) = 61.53\%$  for L. The transition consists of the frequency in percentage with which K is followed by L and vice-versa. Therefore, there are 10 events like that, which is  $100(10/25) = 40\%$ . The rationale behind the distribution concept consists of five chain lengths in which the first, 25%, 50%, 75%, and 100% of residues with a given property are contained. Using again the example sequence, the first group K residue is right at the beginning of the

sequence, so the first value for the distribution descriptor is 0. Twenty-five percent of group K residues are in the first four residues of the chain, then:  $100(4/26) = 15.38\%$ . The second value of this descriptor is then 15.38%. The same is applied to the third (50%), fourth (75%), and fifth (100%) descriptor values [4]. This set of calculations results in 147 features.

- Quasi-sequence order: these descriptors are based on physicochemical distances between a pair of each one of the 20 amino acids, taking into account the following properties: hydrophilicity, hydrophobicity, side-chain, and polarity. Those distances are derived from the Schneider-Wrede physicochemical distance matrix [5] and the Grantham chemical distance matrix [6] and are mathematically represented as follows:

$$\tau_{lag} = \sum_{i=1}^{N-lag} (lag_{i,i+lag})^2,$$

where  $\tau_{lag}$  is the  $lag$ -th rank sequence order coupling number that represents the coupling between contiguous residues at positions  $i$  and  $i+lag$ . Considering each distance matrix, this step results in 60 features, for  $lag = 30$ . When the  $lag$  value is incorporated along with the weighting factor into the calculation, it results in 100 additional features, corresponding to the effect of sequence order [3,7].

- Pseudo-amino acid composition: analogous to quasi-sequence order, the order effect is directly related to the sequence size and thus a correlation between contiguous amino acids is measured. The difference between pseudo-amino acid composition and quasi-sequence order is the replacement of the sequence order coupling number  $\tau_{lag}$  by the sequence order correlation factor  $\theta_{\lambda}$  [3], as described:

$$\theta_{\lambda} = \frac{1}{N-\lambda} \sum_{i=1}^{L-\lambda} \theta(R_i, R_{i+\lambda}),$$

where,  $\Theta(R_i, R_{i+\lambda})$  is the correlation factor and  $\theta_\lambda$  is the first sequence order correlation factor when between contiguous residues, when  $\lambda = 1$ . The parameter  $\lambda$  is the maximal distance between residues in the sequence and the correlation between them is based on the difference between values of hydrophobicity, hydrophilicity, and side chain mass of the contiguous residues. Therefore, for instance, considering a protein of 30 residues in length, the maximal distance between residues is 29, then  $\lambda = 29$ . Thus, there will be 29 rounds of calculation of difference between values of hydrophobicity, hydrophilicity, and side chain mass of contiguous residues (Figure S2). In this case, besides the 20 features related to the composition of each residue in a protein, there will be  $\lambda$  additional features related to the order effect of them. We used two types of pseudo amino acid calculation. The second one does not take into account the side chain mass property and distinctly finds the correlation factor, as described in [1]. We used here  $\lambda = 10$ , therefore resulting in 30 features for Type I and 40 features for Type II pseudo-amino acid composition.

#### 7.1.2.2. Biologically functional and engineered features

Our *in-house* script computes a set of biologically meaningful features expected to aid in the prediction of vaccine candidates. These attributes require third-party programs for feature computation and extraction (Supplemental Table 3). Our script also combines these features to engineer new attributes that, to the best of our knowledge, have not been evaluated for their suitability for RV. The features and software used are as follows:

- Signal peptide prediction: presence of a signal peptide in the protein sequence. Moreover, the number of transmembrane helices, cytoplasmic/non-cytoplasmic domains, and short/long non-cytoplasmic domains are also extracted, as well as coordinates and additional pieces of information (Supplemental Table 3). We used Phobius v.1.01 [8] for feature computation, resulting in 24 attributes.

- Identification of disordered protein regions and disordered binding regions: We compute the mean and median of disordered and binding regions scores for the entire protein, as well as the counts of disordered/structured regions, and their coordinates. The software used to compute these properties are IUPRED v.2 and ANCHOR v.2 [9]. This step results in 26 features.
- Prediction of peptide binding to MHC I and II molecules: We computed the mean and median of eluted ligands and binding affinity scores for the entire protein while considering strong (SB) and weak (WB) binders. We also computed the numbers of SB and WB, their coordinates, and the fraction of these regions in the protein sequence, as well as information about protrusions and their start positions. We used NetMHCpan v.4.1 for the prediction of MHCI molecules and NetMHCIIpan v.4.0 for the prediction of MHCII molecules [10]. The total number of features obtained in this step depends on the number of alleles submitted to these tools. We extracted  $38 + (4 \times \text{number of alleles})$  features for each MHC molecule. For MHCI molecules, we considered the following alleles: HLA-A01:01, HLA-A02:01, HLA-A03:01, HLA-A24:02, HLA-A26:01, HLA-B07:02, HLA-B08:01, HLA-B27:05, HLA-B39:01, HLA-B40:01, HLA-B58:01, and HLA-B15:01. Those HLA class I alleles are representatives of each one of the 12 HLA supertype clusters, which cover a broad bulk of known HLA class I ligation specificities [11–13]. For MHCII molecules, we used the following alleles: HLA-DQA10101-DQB10201, HLA-DQA10101-DQB10301, HLA-DQA10101-DQB10602, DRB1\_0701, DRB1\_1501, DRB1\_0301. Those HLA class II alleles were chosen based on the alleles' population frequencies, according to [14].
- Engineered biological features: All the biologically functional features were combined and engineered into new attributes. Using the coordinates of strong and weak MHCI and II ligands, this module is able to capture the number of ligands, and the fraction occupied by them, in domains such as transmembrane helices, cytoplasmic domains, short and long

non-cytoplasmic domains, disordered and structured regions. This engineering of features step is performed by our in-house python script and results in 84 features.

#### 7.1.3. Data transformation

Due to the distinct types of classifiers evaluated in this study, two different data transformations were performed, generating four types of transformed data to be used to train the models. The transformations comprise the combinations of scaling methods (normalization or standardization) and quantile truncation (with and without). When applied, truncation used the 25th and 75th quantiles as threshold values.

#### 7.1.4. Models description

IFs and RFs share the same overall design concept: an ensemble of trees whose each outcome is aggregated into a final result [15,16]. The main differences between both approaches are that 1) IFs detect anomalies (observations that deviate markedly from those used for training), while RFs profile normal (non-anomalous) points; and 2) the definition of nodes and splits: while IFs establish nodes and split values randomly [15,16], RFs define nodes and split values based on the attribute of the highest information gain [15]. Conceptually, anomalies are expected to present attribute values considerably different from the normal points, which makes anomalies more susceptible to isolation by space-partitioning decision trees (such as the ones used in IF) than non-anomalous points. In the context of this work, we trained IFs using positive examples and treated (potential) negative examples as “anomalies” to be detected. IFs generate a score (Anomaly Score) to label an instance as an anomaly or not [15,16].

OCSVM and BCSVM also share similar concepts in their overall designs, as both starts by representing the instances in a hyperdimensional feature space. In the case of instances from two defined classes, BCSVM establishes a hyperplane of separation between the group of points from each class that maximizes the margin between each group and the hyperplane. On the other hand, an OCSVM aims at creating the minimum hypersphere that encompasses all examples in the training

data (consisting of a single class) and putatively defines a boundary of separation between those instances and other possible examples. OCSVM labels new instances as either normal or as outliers [17,18].

Autoencoders are neural networks that can learn to reconstruct input data by compressing it into a lower-dimensional representation and then decoding it back to its original form. In unsupervised learning, autoencoders learn to capture the underlying structure and patterns of the data without explicit labels. The Mean Squared Error of Reconstruction (Reconstruction MSE), defined as the average of the squared differences between the input data and the corresponding reconstructed output, is commonly used in autoencoders to measure the quality of the reconstructed output compared to the original input and, consequently, as a cutoff to determine whether models are capturing the important features of the input and producing accurate reconstructions [19,20].

#### 7.1.5. Model training, tuning, and evaluation

As described in the main text's section '2.5. Building individual machine learning models', IFs and Autoencoders were trained using only positive instances. When deployed to classify a dataset, these models generate values of Anomaly Score and Reconstruction MSE, respectively, for every instance. Those values are used to label an instance as an anomaly or not: if the output is higher than a threshold, the instance is predicted as an anomaly (in our case, as a non-vaccine, since the models are trained with vaccine examples). Our thresholds for Anomaly Score and Reconstruction MSE were set as follows:

- 1) IFs and Autoencoders were trained using commands presented in Supplemental Table 4;
- 2) We used the models to generate predictions for the testing dataset (main text, section '2.5. Building individual machine learning models') and obtained the Anomaly Scores/Reconstruction MSE for the testing instances.
- 3) We varied the threshold of Anomaly Score/Reconstruction MSE between the minimum and the maximum values observed in either the

training or testing data (step sizes: 0.025 for Anomaly Score, 0.015 for Reconstruction MSE), and at each point we calculated the PCPU metric for the testing data (main text, section '2.6. Performance assessment, model comparison, and ensemble building'). The threshold resulting in the highest PCPU value for the IF (using Anomaly Score) or the Autoencoder (using Reconstruction MSE) was selected as the optimal one and incorporated as part of the modeling pipeline (Supplemental Table 5).

## 7.2. Supplemental Figures

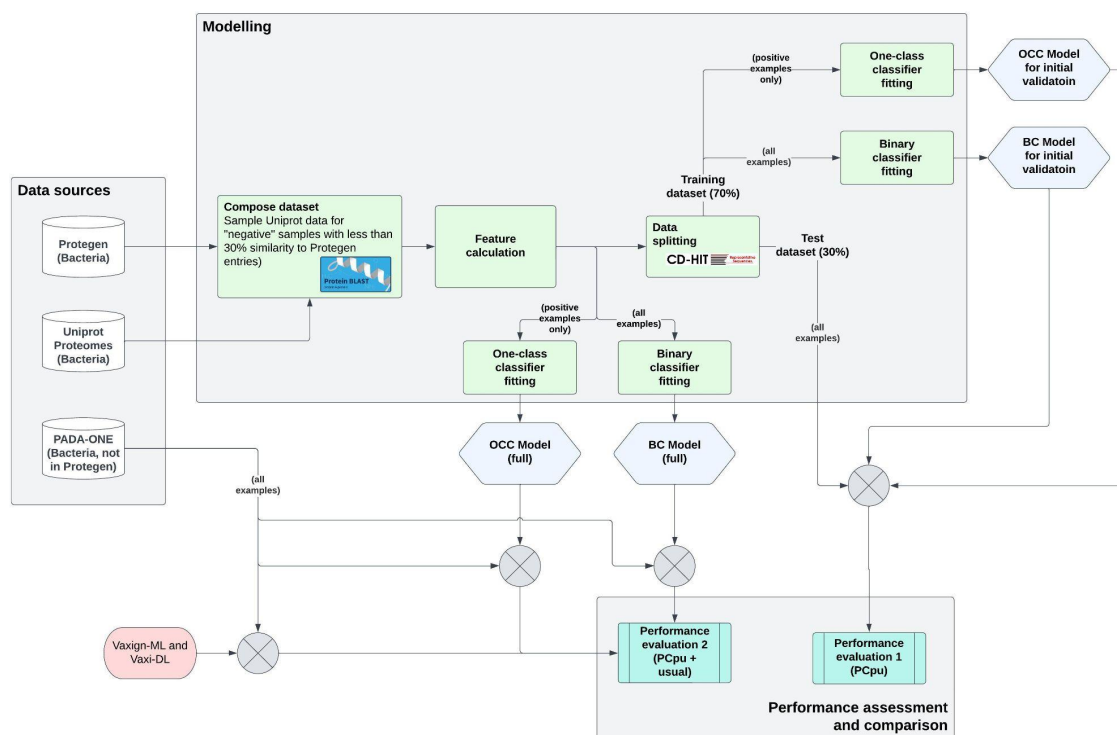


**Figure S1. Development and validation of MARVL.** *Top:* Development of a vaccine candidate predictor for bacteria, where data sources are publicly available, namely Proteogen [21] and Uniprot [22]. Physicochemical and biological features were extracted from the amino acid sequences. Then, the instances were split into training and testing datasets, according to CD-HIT results. Only the positive dataset was used to fit OCCs, while both positive and unlabeled datasets were used to fit BCs. *Bottom:* Full ensemble models (with no dataset splitting) were then developed to benchmark our method against the most recent tools available in the literature: Vaxign2 [23] and Vaxi-DL [24]. New positive and real negative instances were retrieved from the literature to work as an additional validation dataset at this step. BC: binary classifier; OCC: one-class classifier.

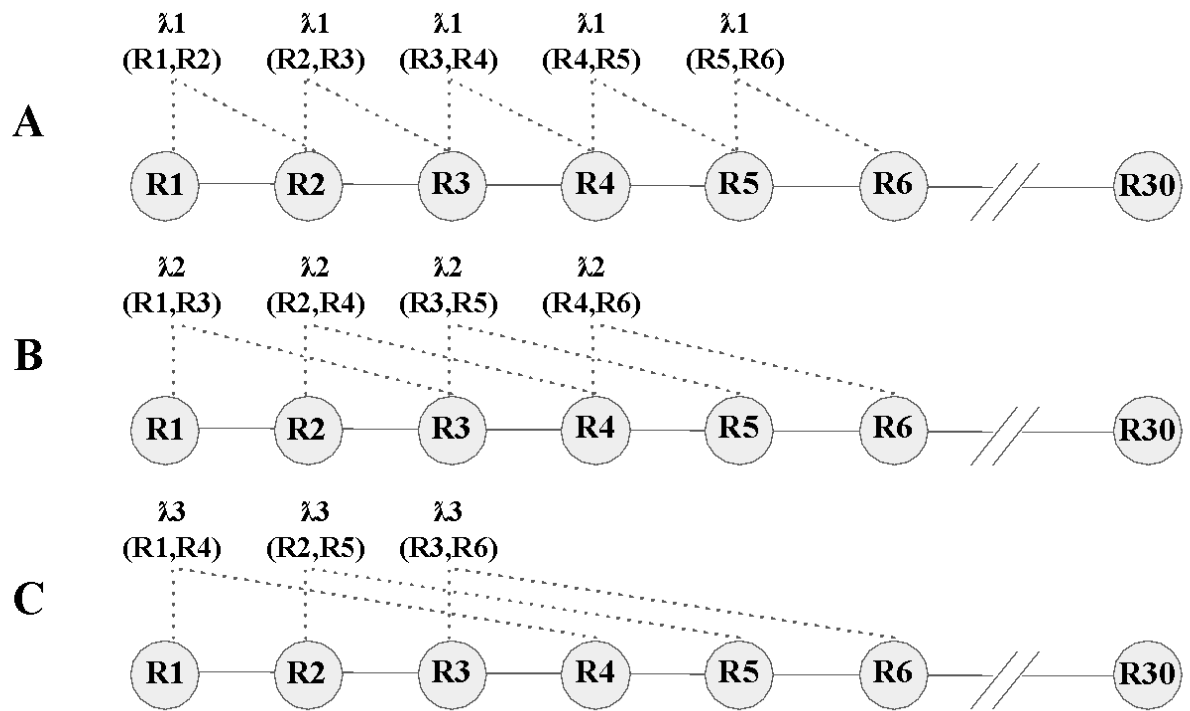


**Figure S2. Correlation between contiguous residues in a protein of 30 residues of size.** **A.** Correlation between the most contiguous residues. **B.** Correlation between all the second-most contiguous residues. **C.** Correlation between all the third-most contiguous residues. The calculation of correlations between contiguous residues continues until the established value for *lambda*. Adapted from [25].

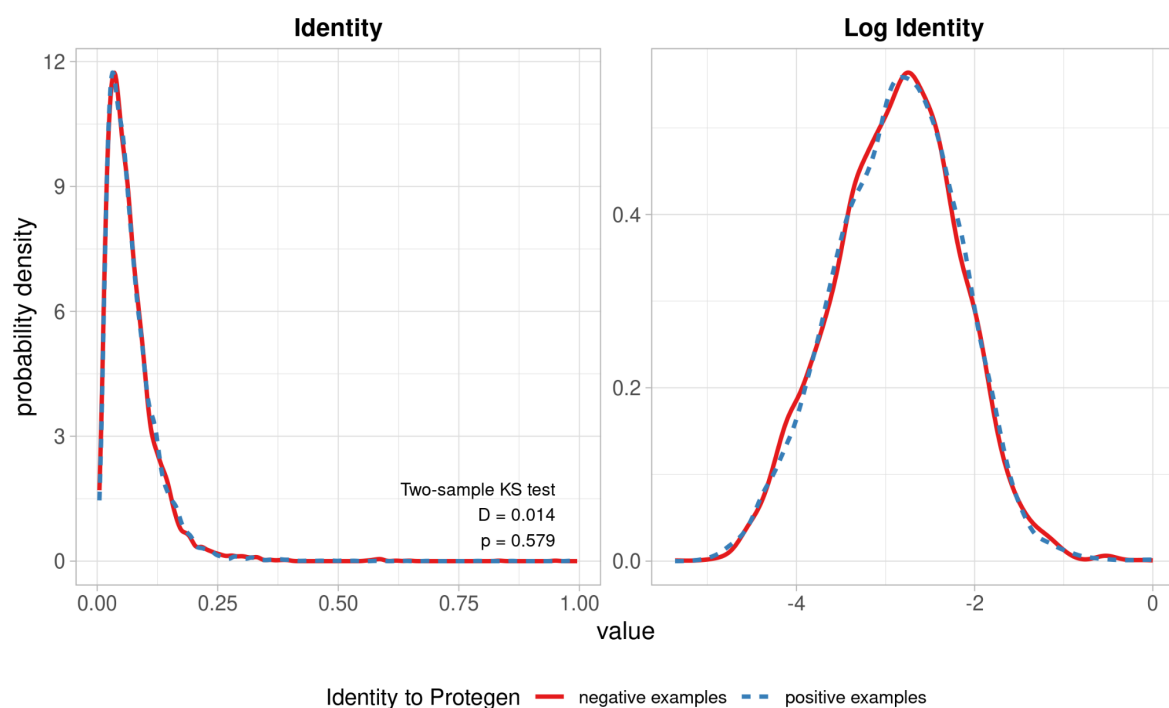


**Figure S3. Distribution of pairwise alignment identities (as raw values and as log-transformed values) of known vaccines and non-vaccines (gold-standard) and an independent set of known vaccines (Protegen).** Red solid line: negative class (experimentally validated non-vaccines); Green dashed line: positive class (experimentally validated vaccines); Sequences that did not return any alignment were considered as having zero identity and were not included in this figure (See Figure S4).

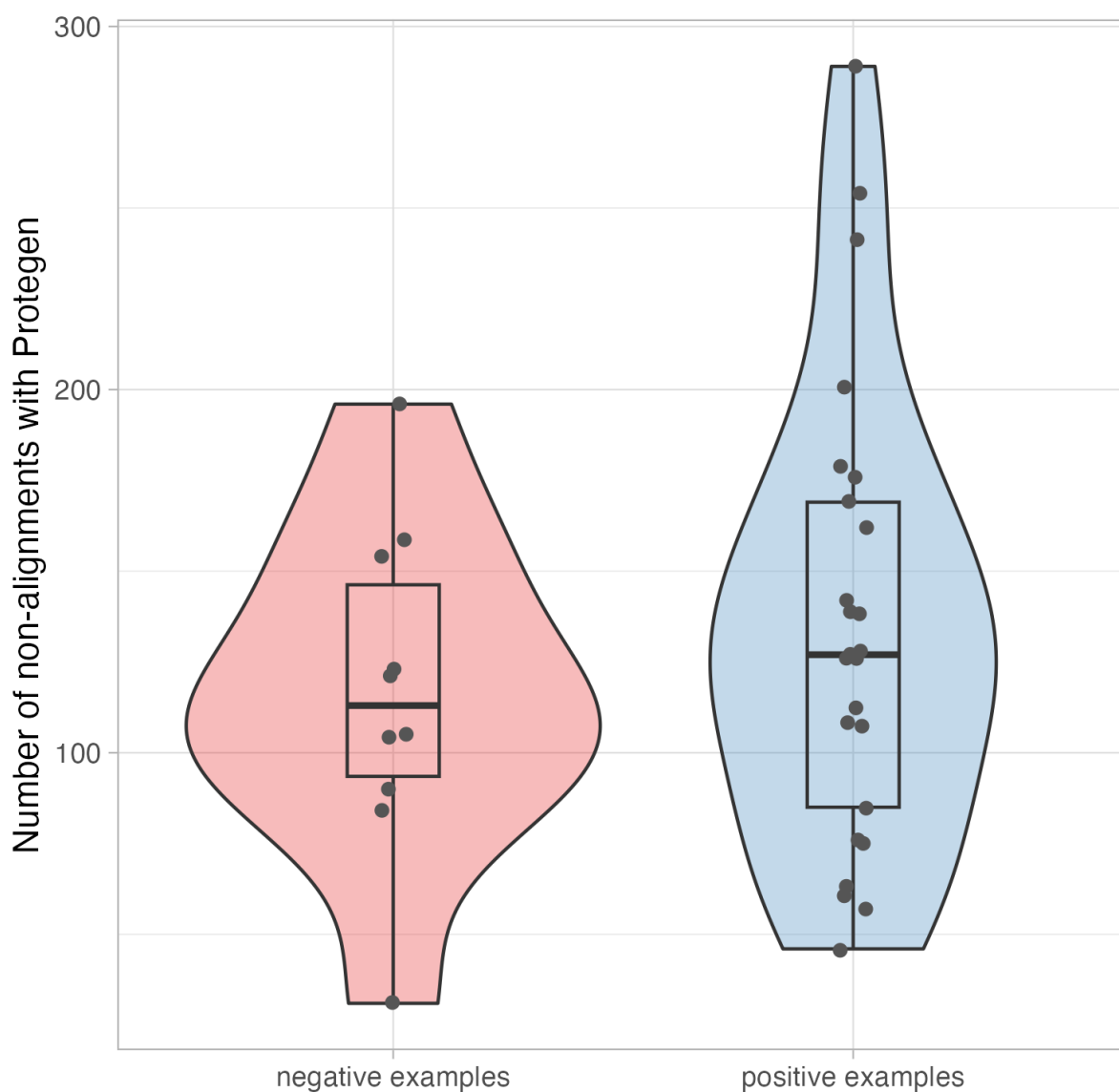


**Figure S4. Non-alignments in the gold-standard dataset against Protegen instances.** In red, gold-standard negative instances. In blue, gold-standard positive instances. A reasonable percentage of non-alignments is observed for both classes of the gold-standard dataset: 23.05% for positives and 19.85% for negatives examples.

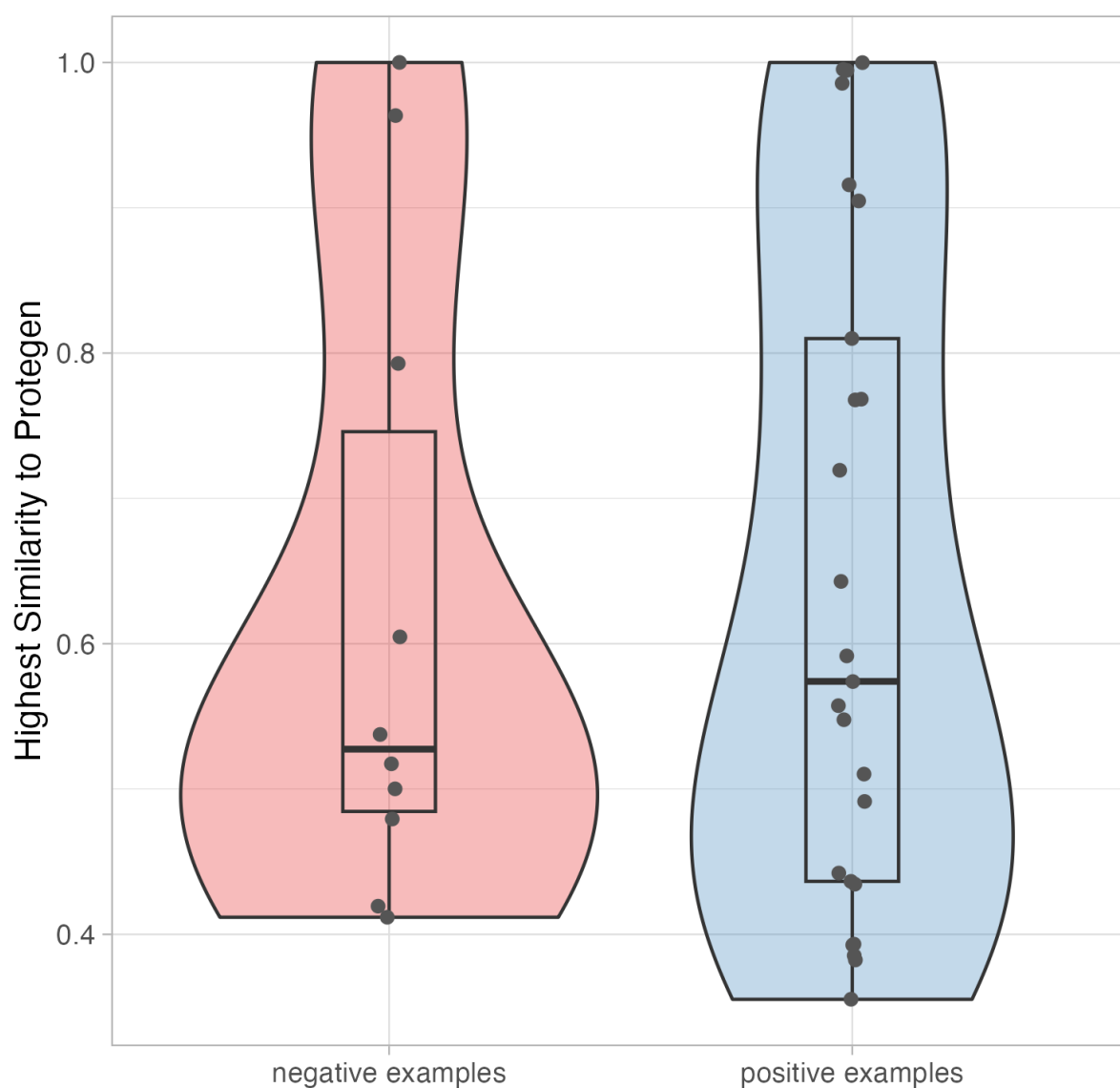


**Figure S5. Highest similarity values of the gold-standard dataset to Protegen instances.** In red, gold-standard negative instances. In blue, gold-standard positive instances. Not a single example from the gold-standard dataset presented a maximum similarity to Protegen lower than the usual threshold of 30%.

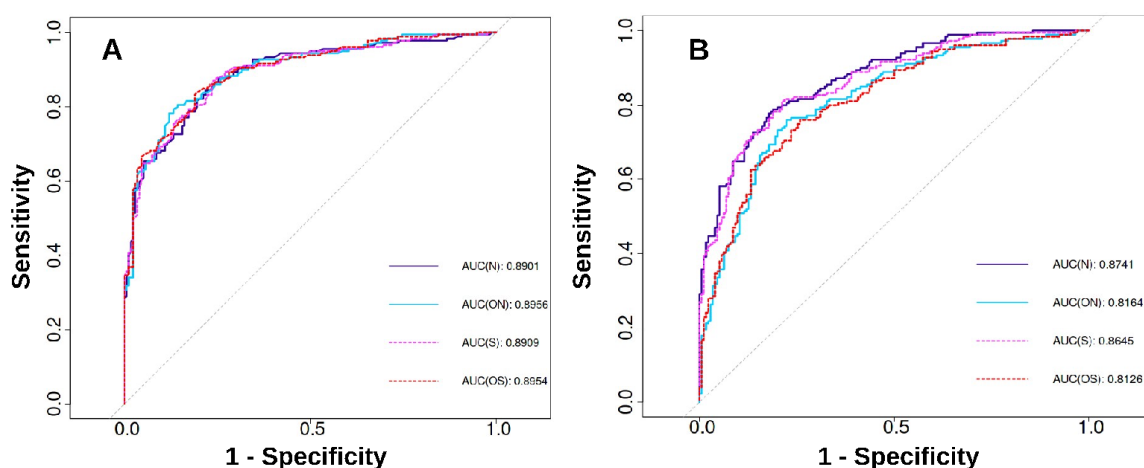


**Figure S6. ROC curves of binary classifiers (BC) in the testing dataset.** **A.** Random Forest (RF) models. RF with normalized dataset (N) in solid dark blue line; RF with normalized and truncated outliers dataset (ON) in solid light blue line; RF with standardized dataset (S) in dashed pink line; and RF with standardized and truncated outliers dataset (OS) in dashed red line. **B.** Support Vector Machine (SVM) models. SVM with normalized dataset (N) in solid dark blue line; SVM with normalized dataset and truncated outliers dataset (ON) in solid light blue line; SVM with standardized dataset (S) in dashed pink line; and SVM with standardized and truncated outliers dataset (OS) in dashed red line. Binary SVMs and RFs have delivered robust models.

### 7.3. Supplemental Tables

Supplemental Table 1. General aspects of the main reverse vaccinology (RV) tools available in the literature. Full supplemental table available at

[x Supplemental\\_Table\\_1\\_Chapter2.xlsx](#)

Supplemental Table 2. Positive, “pseudo-negative”, and gold-standard datasets used to train, test, and validate MARVL. Full supplemental table available at

[x Supplemental\\_Table\\_2\\_Chapter2.xlsx](#)

Supplemental Table 3. Description of features extracted in MARVL pipeline. Full supplemental table available at [x Supplemental\\_Table\\_3\\_Chapter2.xlsx](#)

Supplemental Table 4. Summary of the R packages, functions, and parameters/arguments used to train the models. Full supplemental table available at [x Supplemental\\_Table\\_4\\_Chapter2.xlsx](#)

Supplemental Table 5. Performance of all binary classifiers (BCs) and one-class classifiers (OCCs) on the testing dataset. Full supplemental table available at [x Supplemental\\_Table\\_5\\_Chapter2.xlsx](#)

Supplemental Table 6. Feature counts before and after feature reduction. Full supplemental table available at [x Supplemental\\_Table\\_6\\_Chapter2.xlsx](#)

#### 7.4. References

1. Cao D-S, Xu Q-S, Liang Y-Z. propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 2013; **29**:960–962
2. Caballero J, Fernández L, Abreu JI, *et al.* Amino Acid Sequence Autocorrelation vectors and ensembles of Bayesian-Regularized Genetic Neural Networks for prediction of conformational stability of human lysozyme mutants. *J. Chem. Inf. Model.* 2006; **46**:1255–1268
3. Ong SAK, Lin HH, Chen YZ, *et al.* Efficacy of different protein descriptors in predicting protein functional families. *BMC Bioinformatics* 2007; **8**:300
4. Dubchak I, Muchnik I, Holbrook SR, *et al.* Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci USA* 1995; **92**:8700–8704
5. Schneider G, Wrede P. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys. J.* 1994; **66**:335–344
6. Grantham R. Amino acid difference formula to help explain protein evolution. *Science* 1974; **185**:862–864
7. Chen Y, Li W, Cai N. Predicting Protein Structural Class Based on Ensemble Binary Classification. *Proc IEEE Inst Electr Electron Eng* 2009; 167–170

8. Käll L, Krogh A, Sonnhammer ELL. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.* 2004; **338**:1027–1036
9. Mészáros B, Erdos G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* 2018; **46**:W329–W337
10. Reynisson B, Alvarez B, Paul S, *et al.* NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Res.* 2020; **48**:W449–W454
11. Sette A, Sidney J. Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* 1999; **50**:201–212
12. Lund O, Nielsen M, Kesmir C, *et al.* Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* 2004; **55**:797–810
13. Nielsen M, Lundegaard C, Blicher T, *et al.* NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PLoS ONE* 2007; **2**:e796
14. Saylor K, Donnan B, Zhang C. Computational mining of MHC class II epitopes for the development of universal immunogenic proteins. *PLoS ONE* 2022; **17**:e0265644
15. Breiman L. Random Forests. *Machine Learning* 2001; **45**:5–32
16. Liu FT, Ting KM, Zhou Z-H. Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining* 2008; 413–422
17. Schölkopf B, Platt JC, Shawe-Taylor J, *et al.* Estimating the support of a high-dimensional distribution. *Neural Comput.* 2001; **13**:1443–1471
18. Shin HJ, Eom D-H, Kim S-S. One-class support vector machines—an application in machine fault detection and classification. *Computers & Industrial Engineering* 2005; **48**:395–408
19. Zhou C, Paffenroth RC. Anomaly Detection with Robust Deep Autoencoders. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17* 2017; 665–674
20. Salekshahrezaee Z, Leevy JL, Khoshgoftaar TM. A reconstruction error-based framework for label noise detection. *J. Big Data* 2021; **8**:57
21. Yang B, Sayers S, Xiang Z, *et al.* Protegen: a web-based protective antigen database and analysis system. *Nucleic Acids Res.* 2011; **39**:D1073-8
22. UniProt Consortium. The universal protein resource (UniProt). *Nucleic Acids Res.*

2008; **36**:D190-5

23. Ong E, Cooke MF, Huffman A, *et al.* Vaxign2: the second generation of the first Web-based vaccine design program using reverse vaccinology and machine learning. *Nucleic Acids Res.* 2021; **49**:W671–W678

24. Rawal K, Sinha R, Nath SK, *et al.* Vaxi-DL: A web-based deep learning server to identify potential vaccine candidates. *Comput. Biol. Med.* 2022; **145**:105401

25. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001; **43**:246–255

## GENERAL DISCUSSION

Most of the classical infectious diseases that affect humans were initially parasites of domestic and peridomestic animals that jumped through these species by their close contact with humans for many generations (DIAMOND, 1997). The successful vaccine development for these pathogens was achieved mostly through the classic development protocols. As climate change and environmental degradation expand our contact with wild animals, the number of emerging and re-emerging infectious diseases in humans is expected to be a major health issue in the near future (OLIVAL *et al.*, 2017; WU *et al.*, 2016). In this scenario, the development of computational tools capable of detecting immunologically relevant peptides and proteins may be a key factor in providing fast and reliable responses to these threats, as already happened during the Sars-Cov-2 pandemic (GRIFONI *et al.*, 2020). However, some major challenges remain to develop such computational tools: 1) the lack of freely available public datasets to develop and objectively evaluate these prediction tools; 2) the development of either a general, robust strategy for all pathogens or customized solutions for major classes of parasites; 3) user-friendly interfaces to access these resources.

Our work aims at providing new solutions for RV tasks while taking into account the issues mentioned above. Regarding the datasets for model training and validation, most of the available software does not provide direct access to them, therefore preventing the objective comparison across them to determine the best strategies. Furthermore, for RV programs, there are currently no defined sets of curated, experimentally validated vaccinal proteins and non-vaccine ones.

Even though there are a few isolated initiatives to provide curated examples of known antigens for a few taxa in databases, like the Protegen (YANG *et al.*, 2011) database, there is a lack of examples of vaccine proteins for other major taxa. As for negative examples, to the best of our knowledge, they are retrieved through similarity analysis via sequence alignment followed by the removal of protein sequences that are arbitrarily defined as sufficiently similar to known vaccines, as is the case of the most recent RV tools Vaxijen (DOYTCHINOVA; FLOWER, 2007), Vaxign2 (ONG *et al.*, 2021) and Vaxi-DL (RAWAL *et al.*, 2022). Vacceed (GOODSWEN; KENNEDY; ELLIS, 2014), an RV software developed for eukaryotic antigen prediction, was

trained on proteins that are only likely to be vaccine candidates based on criteria such as protein localization and T-cell epitope occurrence. Therefore, there is currently no RV program validated using experimentally validated positive and negative sequence datasets.

Thus, considering the lack of vaccine proteins for many groups of major pathogens, as well as the absence of experimentally validated non-protective proteins (negative examples) to objectively evaluate vaccine candidate predictors, we firstly developed PADA-ONE. This web tool hosts scientific papers predicted to describe PAgS, together with their sequence, taxonomy, and literature metadata, if available. The searches on our database can be performed using filters related to the metadata associated with each article. A major novelty is the association of a taxon ID to a publication: using this information as a filter, it is possible to search papers with associated proteins belonging to any major group of pathogens, such as apicomplexans, fungi, and helminths. By using these filters to search parasites lacking vaccines, it is possible to find potential PAgS for downstream characterization in species not covered by Protegen and AntigenDB. Those taxa are important human pathogens, composing, for instance, the set of neglected tropical diseases (NTD; (ORGANIZATION, 2020). Taking into account that the number of people treated for NTD decreased by 34% during the COVID-19 pandemic (ORGANIZATION, 2023), it is essential to provide the scientific community with curated information regarding those species, besides those bacterial and viral ones already covered by Protegen and AntigenDB.

Another valuable feature of PADA-ONE is the distinction of four classes of IRPs: attenuators, boosters, PAgS, and non-protective proteins can be found and obtained through searches on our database. We highlight the inclusion of curated non-protective proteins produced through manual curation from studies where the authors characterized several vaccine candidates and found some to fail to induce protection. This information is of great relevance to develop new vaccine candidate predictors, as those models can be now evaluated using PAgS and true non-protective proteins, which solves the problem of the absence of experimentally validated non-protective proteins for that purpose. Additionally, PADA-ONE presents a user-friendly interface that, besides enabling searches combining different types of filters, also allows users to provide feedback regarding papers still not curated by our team, which is expected to increase the speed of the curation process.

The initial screening performed by PADA-ONE in the PubMed literature assigning probabilities to each paper drastically reduced the number of articles of interest to be curated. Those filtered results can then be analyzed according to the study's purpose. In our case, we used PADA-ONE to mine new examples of PAgS exclusively found in the PADA-ONE database (true positives), together with examples of non-protective proteins (true negatives). We used those sequences to benchmark MARVL, our vaccine candidate predictor, against state-of-the-art softwares (Chapter 2). To the best of our knowledge, all of the existing tools to predict vaccine candidates using a machine learning strategy were not evaluated using real non-protective antigens: instead, the training and testing steps of those softwares were performed using putative negative instances based on sequence dissimilarity, as previously described. As examples of the true negative examples recovered from PADA-ONE, we mention PcpA of *Streptococcus pneumoniae*, HspX of *Mycobacterium tuberculosis H37Rv*, and a putative S-layer protein of *Bacillus anthracis str. Ames*. Those antigens presented, respectively, 0%, 20%, and 30% of survival rates in animal model experiments (KUMAR *et al.*, 2020; VERHOEVEN; PERRY; PICHICHERO, 2014; WIECZOREK *et al.*, 2014).

MARVL presents three novelties aimed at increasing its performance: 1) splitting the training dataset based on CD-HIT results to prevent data leakage in the training and testing steps; 2) newly engineered features; 3) using a diverse set of machine learning strategies, together with the selection of models with high performance and low intersection of predictions, to compose our final ensemble.

Our group has previously discussed the process of splitting the training dataset based on a similarity threshold (ASHFORD *et al.*, 2021). In summary, we have developed a strategy that incorporates similarity awareness to effectively assign similar sequences to the same data partition during model training. This approach aims to mitigate the risk of assigning similar sequences to different groups, such as training and validation sets, as it can lead to an overestimation of model performance. By preventing the model from relying on similarity to distinguish between classes, our strategy promotes a more accurate assessment of the model's true capabilities.

This is also the first study in the RV knowledge field containing newly engineered features. The first ML-based RV tool, Vaxijen (DOYTCHINOVA; FLOWER, 2007), was introduced to the scientific community using only

physicochemical features to find vaccine candidates. Later, Vacceed (GOODSWEN; KENNEDY; ELLIS, 2014), like the previous filtering RV tools, was developed to extract only biologically functional features from amino acid sequences. The first predictor to combine physicochemical and functional attributes was the Bowman-Heinson predictor (BOWMAN *et al.*, 2011; HEINSON *et al.*, 2017). This combination of features has become a tendency for the next tools until the last one, Vaxi-DL (RAWAL *et al.*, 2022). Here, we introduced some more informative attributes in our method by integrating the output of NetMHCpan v.4.1 and NetMHCIIpan v.4.0 with the output of Phobius v.1.01, Iupred v.2 and ANCHOR v.2 to produce new attributes. One of those new attributes, the presence of weak MHCII ligands in long non-cytoplasmic domains, has never been discussed by any of the existing tools in the literature and was found as a relevant feature in MARVL, as discussed in section 4, Chapter 2.

We also surveyed a wide range of machine learning strategies, including one-class classifiers that, to the best of our knowledge, have never been evaluated for RV before, to build MARVL as an ensemble RV machine learning approach. Furthermore, by actively selecting models from distinct machine learning classes with high classification performance and relatively low superposition in their prediction vectors, we aimed at selecting models that captured distinct patterns in the training data.

We used the gold-standard of true positives and true negatives curated from PADA-ONE to objectively evaluate MARVL and the other ML-based RV tools. We found that the state-of-the-art tools predicted most of the TPs, but failed to correctly label most of the TNs. Our ensemble model (MARVL), in contrast, performed considerably better for most metrics.

As future improvements of our tools, we intend to use PADA-ONE to obtain more true examples of non-protective bacterial proteins. These would then be included in the training stage of MARVL, and not only in the benchmarking step, which can considerably increase its performance. Moreover, PADA-ONE can be used to mine PAgS and non-protective antigens specific to other groups of pathogens, like protozoans, fungi, and helminths, which enables the development of robust predictors for those taxa as well. We also plan to train a new version of PADA-ONE to specifically learn how to find other classes of IRPs, such as boosters, attenuators, and non-PAgS.

Our second tool (MARVL) can also be improved in many directions. We notice that the OCCs fitted here were trained with no great refinements. Moreover, although we have employed criteria to choose the models to compose our ensemble, there are improvements to be made also in this aspect. Currently, the decision-make process is performed based on the majority of the votes of the ensemble's models. However, extensive surveys have been published on ensemble learning describing other output fusion techniques, such as weighting and meta-learning methods, as well as the algorithms executing those methods (SAGI; ROKACH, 2018).

## GENERAL CONCLUSION

Considering the still existing gaps to be fulfilled in the RV tools development, we presented in this work two solutions that can now be used by the scientific community and improved in future studies. We developed PADA-ONE, a database with a user-friendly interface to search scientific articles predicted to describe vaccines and other classes of immunologically relevant proteins. Later, we also trained, tested, and validated MARVL, an ensemble ML-based RV tool to predict vaccine candidates using bacterial sequence amino acids. Both of the products of this thesis deal with issues still not explored so far by any available initiative.

Besides PAGs, PADA-ONE also sheds light on other 3 different types of IRPs, namely attenuators, boosters, and non-protective proteins. This latter class was used to solve an ongoing issue: providing a true set of negative examples to evaluate the true performance of predictors of vaccine candidates. PADA-ONE also contains data from major groups of pathogens underrepresented in initiatives like Protegen and AntigenDB, such as protozoans, fungi, and helminths. Our database combines several types of filters and organizes results in a user-friendly interface and is available at <https://padaone.org/>.

Our second product is MARVL, an ensemble predictor of vaccine candidates that outperforms currently available tools when validated with an independent, curated true set of positive and negative examples of vaccines extracted from PADA-ONE. Together, these tools provide a rich source of data and a statistical scaffold to create new solutions for reverse vaccinology and are expected to provide the scientific community with powerful instruments to deal with known and emerging infectious diseases of humans and other animals.

## REFERENCES

- ASHFORD, Jodie *et al.* Organism-specific training improves performance of linear B-cell epitope prediction. **Bioinformatics**, v. 37, n. 24, p. 4826–4834, 11 Dec. 2021.
- BALMER, Roberto E.; LEVIN, Stanford L.; SCHMIDT, Stephen. Artificial Intelligence Applications in Telecommunications and other network industries. **Telecommunications policy**, v. 44, n. 6, p. 101977, Jul. 2020.
- BOWMAN, Brett N *et al.* Improving reverse vaccinology with a machine learning approach. **Vaccine**, v. 29, n. 45, p. 8156–8164, 19 Oct. 2011.
- BURLINA, Philippe *et al.* Deep embeddings for novelty detection in myopathy. **Computers in biology and medicine**, v. 105, p. 46–53, Feb. 2019.
- CAMACHO, Diogo M *et al.* Next-Generation Machine Learning for Biological Networks. **Cell**, v. 173, n. 7, p. 1581–1592, 14 Jun. 2018.
- CIODARO, T *et al.* Online particle detection with Neural Networks based on topological calorimetry information. **Journal of Physics: Conference Series**, v. 368, p. 012030, 21 Jun. 2012.
- COWELL, Annie N *et al.* Mapping the malaria parasite druggable genome by using in vitro evolution and chemogenomics. **Science**, v. 359, n. 6372, p. 191–199, 12 Jan. 2018.
- D'MELLO, Adonis *et al.* ReVac: a reverse vaccinology computational pipeline for prioritization of prokaryotic protein vaccine candidates. **BMC Genomics**, v. 20, n. 1, p. 981, 16 Dec. 2019.
- DALSASS, Mattia *et al.* Comparison of Open-Source Reverse Vaccinology Programs for Bacterial Vaccine Antigen Discovery. **Frontiers in Immunology**, v. 10, p. 113, 14 Feb. 2019.
- DEL TORDELLO, E; RAPPUOLI, R; DELANY, I. Reverse Vaccinology. **Human Vaccines**. [S.l.]: Elsevier, 2017. p. 65–86.
- DIAMOND, J. Guns. **Germs and Steel: The Fates of Human Societies**, Vintage, v. 186, 1997.
- DOYTCHINOVA, Irini A; FLOWER, Darren R. VaxiJen: a server for prediction of

protective antigens, tumour antigens and subunit vaccines. **BMC Bioinformatics**, v. 8, p. 4, 5 Jan. 2007.

ELLEGREN, Hans. Genome sequencing and population genomics in non-model organisms. **Trends in Ecology & Evolution**, v. 29, n. 1, p. 51–63, Jan. 2014.

ERICKSON, Bradley J *et al.* Machine learning for medical imaging. **Radiographics**, v. 37, n. 2, p. 505–515, 17 Feb. 2017.

ESTEVA, Andre *et al.* Dermatologist-level classification of skin cancer with deep neural networks. **Nature**, v. 542, n. 7639, p. 115–118, 2 Feb. 2017.

GAUTAM, Chandan *et al.* Minimum variance-embedded deep kernel regularized least squares method for one-class classification and its applications to biomedical data. **Neural Networks**, v. 123, p. 191–216, Mar. 2020.

GENOMICS ENGLAND RESEARCH CONSORTIUM. **The 100,000 Genomes Project, c2019**. Disponível em: <https://www.genomicsengland.co.uk/>. Acesso em: 11 Jul. 2023.

GOODSWEN, Stephen J; KENNEDY, Paul J; ELLIS, John T. A guide to in silico vaccine discovery for eukaryotic pathogens. **Briefings in Bioinformatics**, v. 14, n. 6, p. 753–774, Nov. 2013.

GOODSWEN, Stephen J; KENNEDY, Paul J; ELLIS, John T. Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology. **Bioinformatics**, v. 30, n. 16, p. 2381–2383, 15 Aug. 2014.

GRIFONI, Alba *et al.* A Sequence Homology and Bioinformatic Approach Can Predict Candidate Targets for Immune Responses to SARS-CoV-2. **Cell Host & Microbe**, v. 27, n. 4, p. 671–680.e2, 8 Apr. 2020.

HEINSON, Ashley I; WOELK, Christopher H; NEWELL, Marie-Louise. The promise of reverse vaccinology. **International health**, v. 7, n. 2, p. 85–89, Mar. 2015.

HEINSON, Ashley I *et al.* Enhancing the biological relevance of machine learning classifiers for reverse vaccinology. **International Journal of Molecular Sciences**, v. 18, n. 2, 1 Feb. 2017.

HE, Yongqun; XIANG, Zuoshuang; MOBLEY, Harry L T. Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine

development. **Journal of Biomedicine & Biotechnology**, v. 2010, p. 297505, 4 Jul. 2010.

HINTON, Geoffrey *et al.* Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. **IEEE signal processing magazine**, v. 29, n. 6, p. 82–97, Nov. 2012.

HOU, Nan *et al.* Low-Complexity Repetitive Epitopes of Plasmodium falciparum Are Decoys for Humoural Immune Responses. **Frontiers in Immunology**, v. 11, p. 610, 15 Apr. 2020.

HUMAN MICROBIOME PROJECT CONSORTIUM. A framework for human microbiome research. **Nature**, v. 486, n. 7402, p. 215–221, 14 Jun. 2012.

JAISWAL, Varun *et al.* Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions. **BMC Bioinformatics**, v. 14, p. 211, 1 Jul. 2013.

KHAN, Shehroz S.; MADDEN, Michael G. One-class classification: taxonomy of study and review of techniques. **The Knowledge engineering review**, v. 29, n. 3, p. 345–374, Jun. 2014.

KRAWCZYK, Bartosz. Learning from imbalanced data: open challenges and future directions. **Progress in Artificial Intelligence**, v. 5, n. 4, p. 221–232, Nov. 2016.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, n. 6, p. 84–90, 24 May 2012.

KUMAR, M *et al.* BA3338, a surface layer homology domain possessing protein augments immune response and protection efficacy of protective antigen against Bacillus anthracis in mouse model. **Journal of Applied Microbiology**, v. 129, n. 2, p. 443–452, Aug. 2020.

LANG, Rongling *et al.* Graph-based semi-supervised one class support vector machine for detecting abnormal lung sounds. **Applied mathematics and computation**, v. 364, p. 124487, Jan. 2020.

LARRAÑAGA, Pedro *et al.* Machine learning in bioinformatics. **Briefings in Bioinformatics**, v. 7, n. 1, p. 86–112, Mar. 2006.

MARQUES DE CASTRO, Giovanni *et al.* Cross-species prediction of essential genes

in insects. **Bioinformatics**, v. 38, n. 6, p. 1504–1513, 4 Mar. 2022.

METCALF, C Jessica E; LESSLER, Justin. Opportunities and challenges in modeling emerging infectious diseases. **Science**, v. 357, n. 6347, p. 149–152, 14 Jul. 2017.

MILITELLO, Kevin T *et al.* Antisense RNA and RNAi in protozoan parasites: working hard or hardly working? **Molecular and Biochemical Parasitology**, v. 157, n. 2, p. 117–126, Feb. 2008.

MUKHERJEE, Supratim *et al.* Genomes OnLine Database (GOLD) v.6: data updates and feature enhancements. **Nucleic Acids Research**, v. 45, n. D1, p. D446–D456, 4 Jan. 2017.

NAZ, Kanwal *et al.* PanRV: Pangenome-reverse vaccinology approach for identifications of potential vaccine candidates in microbial pangenome. **BMC Bioinformatics**, v. 20, n. 1, p. 123, 12 Mar. 2019.

NIGATU, Dawit *et al.* Sequence-based information-theoretic features for gene essentiality prediction. **BMC Bioinformatics**, v. 18, n. 1, p. 473, 9 Nov. 2017.

OLIVAL, Kevin J *et al.* Host and viral traits predict zoonotic spillover from mammals. **Nature**, v. 546, n. 7660, p. 646–650, 29 Jun. 2017.

ONG, Edison; WONG, Mei U; HE, Yongqun. Identification of New Features from Known Bacterial Protective Vaccine Antigens Enhances Rational Vaccine Design. **Frontiers in Immunology**, v. 8, p. 1382, 26 Oct. 2017.

ONG, Edison *et al.* Vaxign2: the second generation of the first Web-based vaccine design program using reverse vaccinology and machine learning. **Nucleic Acids Research**, v. 49, n. W1, p. W671–W678, 2 Jul. 2021.

ONG, Edison *et al.* Vaxign-ML: supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. **Bioinformatics**, v. 36, n. 10, p. 3185–3191, 1 May 2020.

ORGANIZATION, World Health. **Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021–2030**. 2020.

ORGANIZATION, World Health. **Global report on neglected tropical diseases 2023: executive summary**. 2023.

PIZZA, M *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. **Science**, v. 287, n. 5459, p. 1816–1820, 10 Mar. 2000.

POPLIN, Ryan *et al.* A universal SNP and small-indel variant caller using deep neural networks. **Nature Biotechnology**, v. 36, n. 10, p. 983–987, Nov. 2018.

RAPPUOLI, R. Reverse vaccinology, a genome-based approach to vaccine development. **Vaccine**, v. 19, n. 17–19, p. 2688–2691, 21 Mar. 2001.

RAPPUOLI, R. Reverse vaccinology. **Current Opinion in Microbiology**, v. 3, n. 5, p. 445–450, Oct. 2000.

RAWAL, Kamal *et al.* Vaxi-DL: A web-based deep learning server to identify potential vaccine candidates. **Computers in biology and medicine**, v. 145, p. 105401, Jun. 2022.

RIZWAN, Muhammad *et al.* VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology. **BMC Bioinformatics**, v. 18, n. 1, p. 106, 13 Feb. 2017.

SAGI, Omer; ROKACH, Lior. Ensemble learning: A survey. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 8, n. 4, Jul. 2018.

TCGA RESEARCH NETWORK. **The Cancer Genome Atlas Research Network**. Disponível em: <https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga>. Acesso em: 11 Jul. 2023.

THE ENCODE PROJECT CONSORTIUM. **ENCODE: Encyclopedia of DNA Elements, c2020**. Disponível em: <https://www.encodeproject.org/>. Acesso em: 11 Jul. 2022.

VERHOEVEN, David; PERRY, Sheldon; PICHICHERO, Michael E. Contributions to protection from *Streptococcus pneumoniae* infection using the monovalent recombinant protein vaccine candidates PcpA, PhtD, and PlyD1 in an infant murine model during challenge. **Clinical and Vaccine Immunology**, v. 21, n. 8, p. 1037–1045, Aug. 2014.

VIVONA, Sandro; BERNANTE, Filippo; FILIPPINI, Francesco. NERVE: new enhanced reverse vaccinology environment. **BMC Biotechnology**, v. 6, p. 35, 18 Jul.

2006.

WASHBURN, Jacob D *et al.* Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. **Proceedings of the National Academy of Sciences of the United States of America**, v. 116, n. 12, p. 5542–5549, 19 Mar. 2019.

WIECZOREK, Agatha E *et al.* HspX vaccination and role in virulence in the guinea pig model of tuberculosis. **Pathogens and disease**, v. 71, n. 3, p. 315–325, Aug. 2014.

WU, Xiaoxu *et al.* Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. **Environment International**, v. 86, p. 14–23, Jan. 2016.

XU, Chunming; JACKSON, Scott A. Machine learning and complex biological data. **Genome Biology**, v. 20, n. 1, p. 76, 16 Apr. 2019.

YANDELL, Mark; ENCE, Daniel. A beginner's guide to eukaryotic genome annotation. **Nature Reviews. Genetics**, v. 13, n. 5, p. 329–342, May 2012.

YANG, Brian *et al.* Protegen: a web-based protective antigen database and analysis system. **Nucleic Acids Research**, v. 39, n. Database issue, p. D1073-8, Jan. 2011.

ZOU, James *et al.* A primer on deep learning in genomics. **Nature Genetics**, v. 51, n. 1, p. 12–18, Jan. 2019.

**APPENDIX A - Co-author in:  
Organism-specific training improves  
performance of linear B-cell epitope  
prediction**

Data and text mining

# Organism-specific training improves performance of linear B-cell epitope prediction

Jodie Ashford<sup>1</sup>, João Reis-Cunha<sup>2</sup>, Igor Lobo<sup>3</sup>, Francisco Lobo<sup>4</sup> and Felipe Campelo <sup>1,\*</sup>

<sup>1</sup>Department of Computer Science, College of Engineering and Physical Sciences, Aston University, Birmingham B4 7ET, UK, <sup>2</sup>Department of Preventive Veterinary Medicine, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil, <sup>3</sup>Graduate Program in Genetics, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil and <sup>4</sup>Department of General Biology, Universidade Federal de Minas Gerais, Belo Horizonte 31270-901, Brazil

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on May 8, 2021; revised on July 1, 2021; editorial decision on July 15, 2021; accepted on July 19, 2021

## Abstract

**Motivation:** *In silico* identification of linear B-cell epitopes represents an important step in the development of diagnostic tests and vaccine candidates, by providing potential high-probability targets for experimental investigation. Current predictive tools were developed under a generalist approach, training models with heterogeneous datasets to develop predictors that can be deployed for a wide variety of pathogens. However, continuous advances in processing power and the increasing amount of epitope data for a broad range of pathogens indicate that training organism or taxon-specific models may become a feasible alternative, with unexplored potential gains in predictive performance.

**Results:** This article shows how organism-specific training of epitope prediction models can yield substantial performance gains across several quality metrics when compared to models trained with heterogeneous and hybrid data, and with a variety of widely used predictors from the literature. These results suggest a promising alternative for the development of custom-tailored predictive models with high predictive power, which can be easily implemented and deployed for the investigation of specific pathogens.

**Availability and implementation:** The data underlying this article, as well as the full reproducibility scripts, are available at <https://github.com/fcampelo/OrgSpec-paper>. The R package that implements the organism-specific pipeline functions is available at <https://github.com/fcampelo/epitopes>.

**Contact:** [f.campelo@aston.ac.uk](mailto:f.campelo@aston.ac.uk)

**Supplementary information:** [Supplementary materials](#) are available at *Bioinformatics* online.

## 1 Introduction

In humoral immunity, activated B-lymphocytes (B cells) produce antibodies that bind with specific antigens, and are a key component in vertebrate immune responses (Getzoff *et al.*, 1988; Lodish *et al.*, 2000). The exact portion of an antigen that an antibody binds to is known as an *epitope* or *antigenic determinant* (Paul, 2012). Identifying B-cell epitopes is a crucial process for a number of medical and immunological processes including: vaccine development, therapeutic antibody production, disease prevention and diagnosis (Dudek *et al.*, 2010; Leinikki *et al.*, 1993; Potocnakova *et al.*, 2016).

B-cell epitopes are broadly classified into two groups: linear (or continuous) epitopes, which represent contiguous stretches of amino acid (AA) residues in an antigenic sequence; and conformational (or discontinuous) epitopes, where the AA residues that constitute these antigens are separated in the sequence and brought together by

folding (Kindt *et al.*, 2007, Chap. 3). The methods used to predict B-cell epitopes differ depending on the type of epitope being predicted. Although the majority of B-cell epitopes are conformational (Van Regenmortel, 1996; Lo *et al.*, 2013), most epitope prediction methods are designed to predict linear epitopes (Alix, 1999; Blythe and Flower, 2005; EL-Manzalawy *et al.*, 2008; Kolaskar and Tongaonkar, 1990; Larsen *et al.*, 2006; Saha and Raghava, 2004, 2006; Singh *et al.*, 2013; Yao *et al.*, 2013). This is mainly due to a relative scarcity of available data on antigen 3D structures, as well as the high computational cost associated with predicting these structures (Yang and Yu, 2009). On the other hand, linear B-cell epitopes can be predicted from protein primary structure alone (AA sequence data), which is more readily available. Linear epitopes are also stable in a wide range of conditions, an interesting property for the transportation and storage of potential peptide vaccines. On the other hand, discontinuous epitopes can be disrupted by alterations

in protein secondary/tertiary structure caused by a wide range of factors, such as variations in pH, salinity and temperature, by protein-protein interactions and post-translational modifications, among many others. In fact, linear epitopes were consistently more recognized than conformational epitopes in the sera of rabbits immunized with recombinant proteins and peptides (Forsström *et al.*, 2015). The impact of mutations is also likely to be more easily estimated for linear epitopes, where most of the relevant changes are observed in the antigenic region. Conformational epitopes, on the other hand, can be affected by AA changes in other regions of the protein that result in conformational changes, which are harder to predict and model (Pandurangan and Blundell, 2020).

Several experimental methods have been traditionally used for B-cell epitope identification, including X-ray crystallography, peptide microarrays, Western Blotting and enzyme-linked immunosorbent assay (ELISA) (Arnold *et al.*, 2018; Jespersen *et al.*, 2019). These methods are both time consuming and resource intensive, which led to the development of computational methods for epitope prediction that are commonly used as pre-screening tools for prioritizing targets for experimental investigation. Early computational methods for predicting linear epitopes were based on direct prediction of different physiochemical properties of individual AA residues found to be more represented in known epitopes, such as hydrophobicity, flexibility, surface accessibility, charge and AA residue frequency (EL-Manzalawy *et al.*, 2008; Haste Andersen *et al.*, 2006; Hopp and Woods, 1981; Parker *et al.*, 1986; Pellequer *et al.*, 1991, 1993; Yang and Yu, 2009). Numerical propensity scales are often created to represent physiochemical properties like these, and commonly used as prediction methods (Alix, 1999; Pellequer *et al.*, 1991; Pellequer and Westhof, 1993) or as input features for machine learning (ML) predictors.

Though propensity scales are still used for B-cell epitope prediction, multiple works have shown that their use alone can result in poor prediction performance (Blythe and Flower, 2005; Giacobbe *et al.*, 2012; Kulkarni-Kale *et al.*, 2005; Pellequer *et al.*, 1991; Ponomarenko and Bourne, 2007). This perceived limitation in predictive power, coupled with increases both in computational resources and available protein sequence data, have led to the adoption of ML models as the main methods for epitope prediction in recent years (Supplementary File S1, Table S1-1). Several ML approaches currently exist for epitope prediction: some are trained using 3D structures, some using a combination of features from propensity scales and many more. ML methods for epitope prediction tend to outperform methods based solely on simple AA propensity scale calculations (Sanchez-Trincado *et al.*, 2017), although this is not always the case (Greenbaum *et al.*, 2007; Sanchez-Trincado *et al.*, 2017).

To our knowledge all existing epitope prediction tools are based on datasets containing labelled peptide sequences coming from a wide variety of organisms (Supplementary File S1, Table S1-1). The use of heterogeneous datasets is associated with a common goal of developing general-purpose predictors that can be pre-trained and used out-of-the-box, without requiring users to inform the source organism of the peptides submitted for classification. In fact, as recently as 2020, Collatz *et al.* (2020) suggest that having ‘a large variety of known epitopes from evolutionarily distinct organisms in the training set’ would be essential to achieve bias-free classification. This is a reasonable assumption if one is aiming at developing generalist, one-size-fits-all models; however, it may be unnecessary or even counterproductive, if the new observations for which the model is expected to generalize correspond only to a specific subset of all possible observations.

Continuous advances in processing power and the increasing amount of data for distinct pathogens suggest that organism- or taxon-specific models may become a feasible alternative. Generating predictors specifically trained for individual pathogens, rather than having a single generalist model, would result in smaller but potentially higher quality training sets, resulting in better predictive performance of new epitopes for the target organism, and potentially for its phylogenetically close relatives. Under this alternative approach of training bespoke models for distinct (groups of)

pathogens, the objective is to obtain predictors that generalize well only to the target organism(s), rather than to the whole variety of pathogens that may interact with a given host.

This work investigates the effects of using such organism-specific datasets to train ML models for linear B-cell epitope prediction. Proof-of-concept predictors are trained using organism-specific, heterogeneous and hybrid data, using data-rich pathogens representing two of the major classes of parasitic organisms: nematodes and viruses. The effects of these training sets on the generalization performance of the models is quantified to test whether organism-specific training can result in better predictors. The results obtained for three test cases not only support this idea, but also show that even relatively simple models trained on organism-specific data can generally outperform current state-of-the-art predictors in terms of several performance measures.

## 2 Materials and methods

Figure 1 summarizes the general organism-specific epitope prediction pipeline. This section details the main methodological aspects of the proposed tool.

### 2.1 Datasets

Specific datasets for each pathogen were generated based on the full XML export of the Immune Epitope Data Base (IEDB) (Vita *et al.*, 2019) retrieved on 10 October 2020, and filtered according to the following criteria:

- Only peptides marked as linear B-cell epitopes/non-epitopes of length between 8 and 25 were selected. The filtering criteria used to isolate peptides identified as linear B-cell entries were (i) those with one or more Assays containing a ‘BCell’ field name (in the Assay fields of the XML document); and (ii) those containing the field ‘FragmentOfANaturalSequenceMolecule—LinearSequence’ in the *EpitopeStructure* field of the XML document. Peptides marked as either ‘Exact Epitope’ and ‘Epitope-containing region’ in the *EpitopeStructureDefines* field were included. The upper length limit was imposed to prevent overly long sequences labelled as ‘Epitope-containing region’ from adding too much noise to the training data, whereas the lower limit was set to prevent excessive redundancy due to short windows (see below, Section 2.1.1).
- Labels ‘Positive’, ‘Positive-High’, ‘Positive-Intermediate’ and ‘Positive-Low’ were grouped under the single label ‘Positive’.
- Observations with missing or inconsistent information related to the protein information (protein ID or peptide position on the protein) were removed.
- Protein information was retrieved from NCBI (NCBI Resource Coordinators, 2015) and UniProt (UniProt Consortium, 2020) based on the protein IDs available in the epitope data. Observations with invalid protein IDs were removed.
- When different assays provided conflicting evidence for the class (Positive vs. Negative) of a given peptide the class was determined by simple majority. Ties were removed from the training sets and de-classed (reference class information set as unknown) in the hold-out sets, so as not to influence the performance calculation.

For each pathogen a number of distinct datasets were generated as follows:

1. First, all examples related to the specific pathogen were extracted based on the taxonomy ID information from the IEDB data. This includes all taxonomically dependent IDs (related, e.g. to subspecies or strains) as part of the data. Prior to any data

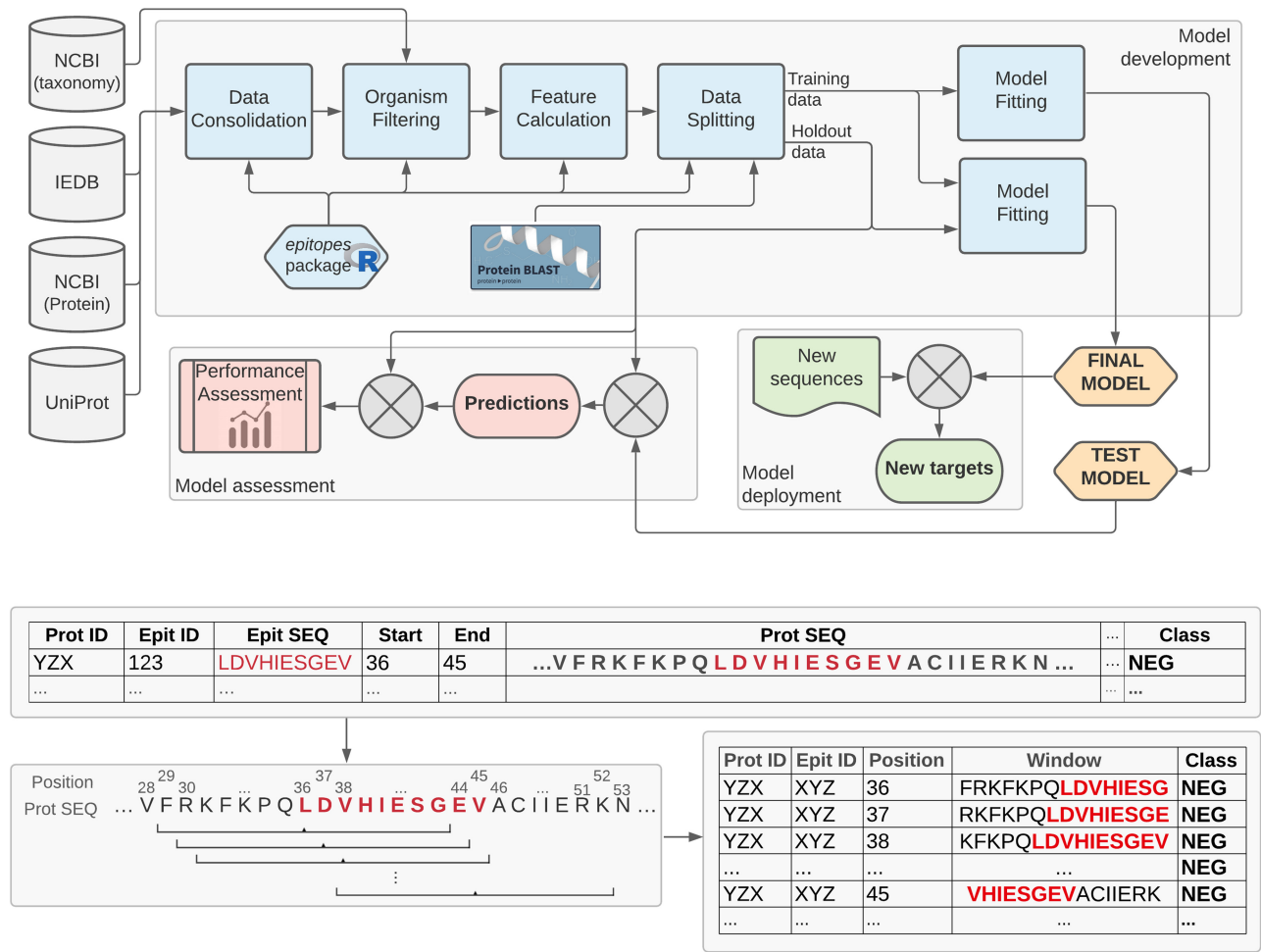


Fig. 1. Top: Organism-specific epitope prediction pipeline. Publicly available data is retrieved from IEDB (Vita et al., 2019), NCBI (NCBI Resource Coordinators, 2015) and UniProt (UniProt Consortium, 2020) to compose an organism-specific dataset. 845 simple features are calculated for each AA, based on the local neighbourhood of every position extracted using a 15-AA sliding window representation with a step size of one (bottom). The data is then split at the protein level, based on protein ID and similarity, into a training set (used for model development) and a hold-out set (used to estimate the generalization performance of the models). The *epitopes* R package, which implements the main elements of this pipeline, is available at <https://fcampelo.github.io/epitopes>

exploration or modelling, a subset of the organism-specific data consisting of approximately 25% of the available observations were set aside as a validation (*Hold-out*) set, which was not seen at any point during model construction. To minimize the chances of data leakage (Kaufman et al., 2011) the splitting of the datasets was done at the protein level, based on protein ID as well as sequence coverage and similarity. Proteins with similarity and/or coverage greater than 80% were always placed within in the same split.

2. The other sub-set, containing 75% of the labelled peptides belonging to the specific pathogen, was used as the organism-specific (*OrgSpec*) training set.
3. A second training dataset (*Heterogeneous*) was assembled by random sampling of observations (grouped by taxonomy ID) from the full IEDB, *excluding any observations related to the specific pathogen*. The sampling routine included as many organisms as required to assemble a class-balanced *heterogeneous* training set containing between 2000 and 3000 labelled peptides of each class (epitope/non-epitope).

4. Finally, a third training set (*Hybrid*) was assembled by combining the *OrgSpec* and *Heterogeneous* sets.

In all cases the *Hybrid* dataset was the largest one, followed by *Heterogeneous* and then *OrgSpec*. This was set up in order to allow us to investigate the hypothesis that trading sample size (which is larger if one incorporates heterogeneous observations) by sample relevance (represented by data that belongs to the organism of interest, for which the models are being developed) would result in improved performance.

The datasets that we assembled for each pathogen allowed us to:

- (i) investigate the *generalization* performance of our models to the prediction of new epitopes in proteins belonging to the specific organisms for which they were trained, by examining the predictive performance for distinct proteins that were set aside as the organism-specific *Hold-out*;
- (ii) investigate the effect of using only organism-specific data on predictive performance, by contrasting models developed using the *OrgSpec*, *Hybrid* and *Heterogeneous* datasets (since all pre-processing, feature development and classification models were the same for all cases, any systematic differences in performance can be attributable to the pre-selection of training data);
- (iii) compare the performance of organism-specific models

against usual approaches in the literature. This last point was the main motivating factor for using a hold-out approach rather than cross-validation for model assessment, as it allowed us to estimate the generalization performance of all predictors on the same data rather than using reported performance values from the literature, which were obtained on distinct datasets or using different testing protocols.

### 2.1.1 Data representation

Each dataset was set up as a fixed-width windowed representation. A sliding window of length 15 with a step size of one was run over each peptide. The choice of length 15 was based on the smallest peptide length of interest, namely 8. The rationale was to use the longest possible window such that strictly more than half the AAs covered would belong to a labelled peptide, which translates as a window of length  $\ell_{min}/2 - 1$ , with  $\ell_{min}$  representing the shortest labelled peptide in the training sets. Based on this windowed representation the following features were calculated:

- Percent composition of the sequence in terms of each individual AA type (20 features), each dipeptide combination (400 features), each conjoint triad (Shen et al., 2007; Wang et al., 2017) (343 features) and each of nine AA types: Tiny, Small, Aliphatic, Aromatic, Non-Polar, Polar, Charged, Basic and Acidic (9 features).
- AA descriptors, averaged over the window: Cruciani properties, Kidera factors, Z scales, FASGAI indices, T scales, VHSE scales, ProtFP descriptors, ST Scales, BLOSUM indices and MS-WHIM scores (Osorio et al., 2015) (66 features).
- Total molecular weight of the window (1 feature).
- Total number of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms in the sequence (5 features).
- Entropy of the distribution of AA residues in the sequence (1 feature).

### 2.1.2 Target pathogens used

The following organisms were used to investigate the efficiency of organism-specific training:

- *Onchocerca volvulus* (taxonomy ID: 6282), a roundworm (Nematoda) which is the causative agent of Onchocerciasis, a leading cause of blindness worldwide (World Health Organization, 2019) with over 37 million people estimated to be infected, mostly in Africa and Latin America (Basáñez et al., 2006; Osei-Atweneboana et al., 2007).
- Epstein-Barr Virus (taxonomy ID: 10376), a double-stranded DNA virus of the *Herpesviridae* family that is the causative agent of infectious mononucleosis and a pathogen linked to many human neoplastic diseases (Rezki et al., 2018).
- Hepatitis C Virus (taxonomy ID: 11102), a positive-sense single-stranded RNA virus of the family *Flaviviridae* which causes Hepatitis C and is associated with the development of certain cancers (Ferri, 2015).

The main criterion used to select these pathogens was the availability of a large volume of validated positive and negative observations in the IEDB, to allow the use of the strict validation strategy outlined above (based on the use of a 25% hold-out set) while keeping enough data for model development. To that end we extracted the ten organism IDs with the greatest number of valid entries in the IEDB (after the filtering described in Section 2.1), and selected those that (i) had a reasonable balance between the positive and negative examples (this removed entries with heavily imbalanced class distributions, IDs 353153, 1314, 5833, 1392); and (ii) represented a pathogen of interest (this removed entries related to allergens or

potential self-epitopes, IDs 9606, 9913 and 3818). The test pathogens selected based on these criteria were a multi cellular parasite, an RNA virus and a DNA virus, which allowed us to evaluate how our pathogen-specific tools perform when evaluating distinct classes of parasitic organisms. Supplementary File S1, Table S1-2 documents the dataset sizes extracted for each organism.

The results obtained for these pathogens were used to assess the performance gains of models trained with organism-specific data, and to compare their performance against existing predictors. The specific results obtained for *O.volvulus* are explored in greater detail in Section 3.1, to illustrate some particular aspects of the proposed approach. A fourth pathogen, the bacterium *Streptococcus Pyogenes*, was also investigated for the sake of completeness, although it did not fulfil the selection criteria outlined above. The specific data pre-processing, modelling and results obtained for this bacterium are described in Supplementary File S2.

### 2.2 Modelling

All training sets were used to develop Random Forest (RF) predictors (Breiman, 2001). This work used the RF implementation from R package *ranger* (R Core Team, 2020; Wright and Ziegler, 2017) version 0.12.1, under standard hyper-parameter values. Experiments with hyper-parameter tuning and feature selection did not result in relevant improvements in performance. All exploratory modelling experiments are documented in Supplementary File S5.

The final output of our predictive pipeline consists of a predicted probability for each position on each protein queried, which is converted to a binary prediction by thresholding at the level 0.5 (no tuning was performed for the threshold value). From these AA-wise predictions, arbitrary-length predicted epitopes are extracted. To reduce prediction noise, positive regions shorter than 8-AAs long were filtered out from the output of the random forest.

### 2.3 Performance assessment and comparison

Several performance indicators were calculated to provide comparability with different references in the literature, and to explore distinct aspects of the predictive behaviour of the models. More specifically, we assessed and compared model performance using the *Positive Predictive Value* (PPV), *Negative Predictive Value* (NPV), *Sensitivity* (SENS), *Accuracy* (ACC), *Area Under the ROC Curve* (AUC) (Tan et al., 2005) and *Matthews Correlation Coefficient* (MCC) (Chicco and Jurman, 2020). The detailed mathematical definition and interpretation of each of these measures is provided in Supplementary File S5.

All performance values reported in the Results section refer to out-of-sample prediction, i.e. observed performance on the *Hold-out* set extracted for each individual pathogen. Since this data is not used at any point in model development, the performance values reported are considered as representing a good estimation of the generalization performance of the proposed models for these organisms.

Performance was calculated based on peptide-wise correct classifications. Following standard practice, a classification was considered as correct whenever a model predicted the right class for strictly more than half the residues in a labelled peptide. Bootstrap (Davison and Hinkley, 2013) was used to calculate standard errors of estimation for each performance measure, as well as to derive *P*-values for the comparison of mean performance between our reference implementation (trained with *OrgSpec*) and all other comparison methods (999 bootstrap resamples were used in all cases). The resulting *P*-values were corrected for multiple hypothesis testing (MHT) using the Holm correction (Holm, 1979), which provides strict control of the Family-wise error rate (FWER) for each family of hypotheses. All comparisons were done at the joint  $\alpha^* = 0.05$  significance level.

Five well-known B-cell epitope predictors providing easy-to-use online interfaces were used to obtain a comparison baseline: BepiPred 2.0 (Jespersen et al., 2017), SVMTriP (Yao et al., 2012), LBtope (Singh et al., 2013), ABCpred (Saha and Raghava, 2006) and iBCE-EL (Manavalan et al., 2018). These models were used to

predict epitopes in the same *Hold-out* sets as our models, based on the default configurations of their respective online tools.

### 3 Results and discussion

#### Organism-specific training improves performance of linear B-cell epitope prediction

As detailed in Section 2.3, the performance of the organism-specific Random Forest models (*RF-OrgSpec*) was compared with (i) the same Random Forest model trained using heterogeneous and hybrid data, to investigate the effect on performance of the data selection strategy; and (ii) a number of well-known predictors, to provide a comparison against currently used approaches. In all cases the performance was calculated based on the hold-out set that was isolated for each pathogen, which was not used at any point in model development. Random Forests are ensemble learning methods that consist of the aggregation of several weaker decision tree (DT) models, with an output based on the combined output of the underlying DTs. Random forests present a good balance between computational cost and performance, and are robust and flexible to work with different data types and scales, which justifies their use in a variety of application domains including several epitope prediction methods (Jespersen *et al.*, 2017; Saravanan and Gautham, 2015). Preliminary comparative testing suggested Random Forests and Gradient Boosting models as having better performance than multi-layer perceptron neural networks and kNN classifiers, and RF was chosen for this work due to presenting lower computational costs in relation to Gradient Boosting.

Figure 2 summarizes the results obtained for the organisms described in Section 2.1.2. The strong positive effect of training models with organism-specific data was observed in all datasets. A clear performance ordering *RF-OrgSpec* > *RF-Hybrid* > *RF-Heter* can be observed across all pathogens, on all performance indices used. The corrected *P*-values indicate that the observed differences are in most cases statistically significant at the joint 0.05 significance level. This pattern corroborates the initial hypothesis that training models on organism-specific data yields improved predictive performance, even when compared with models that contain the same organism-specific data combined with examples from other organisms.

Additional analyses also confirm that performance gains of organism-specific prediction are observed for the pathogen for which the model is originally trained, but not when trying to predict epitopes for other organisms. Figure S3-1 in Supplementary File S3 contrasts the observed performance of *OrgSpec* models on the hold-out set of their specific organism with that obtained when predicting epitopes for other pathogens. These results clearly illustrate that the excellent gains in organism-specific performance (Fig. 2) come at the cost of a reduced ability to detect patterns in proteins coming from other pathogens, which further corroborates our underlying hypothesis that organism-specific training allows models to learn patterns that may be idiosyncratic to the target pathogen.

#### Organism-specific models exhibit better performance than existing generalist models

Contrasting the observed performance values of the *RF-OrgSpec* models with the selected predictors in Figure 2, it is again clear that even the standard Random Forest model used in this work (without hyper-parameter tuning or threshold adjustment) was able to significantly outperform all baseline models on most performance measures. The only predictor that presents performance values comparable to *RF-OrgSpec* is *LBtope* in the case of the hepatitis C virus. This can, however, be partly explained by the fact that part of the hold-out examples used to assess the performance of the models is present in the training data of *LBtope* (9.59% of the Hep C hold-out sequences are present in the *LBtope* training dataset). There is also significant presence of our hold-out Hep C examples in the training data *BepiPred-2.0* (16.3%) and *iBCE-EL* (8.6%). Other predictors are not substantially affected, and this is not observed in the case of the other pathogens tested. Supplementary File S1, Table S1-3 provides the performance values of all predictors on all test organisms, including performance values calculated using only the

unseen sequences (not part of the training set) for the case of the hepatitis C virus.

#### 3.1 Example: *O.volvulus* results

The predictions obtained for the *O.volvulus* data were selected to illustrate the organism-specific results in more detail. Figure S3-2 in Supplementary File S3 (right panel) shows the ROC curves obtained for all predictors on the *O.volvulus* hold-out data, clearly indicating that the organism-specific model does indeed result in substantial performance gains. The *RF-OrgSpec* model presented very good robustness to different threshold values (*AUC* = 0.83). *RF-Hybrid*, which also included organism-specific data as part of its training set, displayed reasonably good performance as well (*AUC* = 0.75).

Figures S3-3 and S3-4 (Supplementary File S3) illustrate the target regions predicted by the organism-specific pipeline for the 22 hold-out proteins of the *O.volvulus* data, using the default threshold value of 0.5 (Fig. S3-5 to S3-11 in Supplementary File S3 illustrate the corresponding results for the other pathogens tested). This illustrates not only the excellent concordance of the *RF-OrgSpec* predictions with the known labels on the hold-out proteins, but also a number of newly identified potential epitopes that may exist in those proteins. The peptides output by the *O.volvulus* model with an average probability of over 0.75 are listed in Supplementary File S1, Table S1-5.

These results show how the higher overall performance of organism-specific models, when compared with state-of-the-art predictors, can be invaluable to advance the detection and selection of diagnostic targets and vaccine candidates for infectious diseases. In particular, the higher PPV values (see Fig. 2) indicate that predicted targets have a good chance to be indeed antigenic, improving the efficiency of epitope discovery processes based on the proposed organism-specific models. This could be a consequence of idiosyncratic patterns of epitopes in different species that would be neglected by generalist predictors. For this reason, organism-specific models may be especially relevant for types of pathogens that are usually under-represented in generic epitope training data bases.

#### 3.2 Discussion

The results described in this section indicate a clear improvement in performance resulting from the use of organism-specific models, when compared to generalist predictors trained on heterogeneous, or even hybrid, data. While an in-depth exploration of the underlying causes of these differences in performance is outside the scope of this work, there are some potential, non-mutually exclusive hypotheses that could be raised.

An examination of the relevance of distinct features and feature groups is provided in Supplementary File S6. There are interesting general insights that can be derived from that exploration in terms of which feature groups contribute the most to the predictive ability of both *OrgSpec* and Heterogeneous models (Supplementary File S6, Fig. S6-1 to S6-3), such as the disproportionately large prevalence of *AA descriptors*-type features among the most relevant, of the apparent irrelevance of dipeptide frequencies or Conjoint Triads for the linear B-cell epitope problem as modelled here. However, it is potentially more valuable in the context of this particular work to focus on features that appear more consistently as relevant for *OrgSpec* models than Heterogeneous ones. As suggested in Supplementary File S6 (Supplementary Fig. S6-4–6), feature *BLOSUM1* (Georgiev, 2009) is clearly one that stands out in terms of being very relevant in general, and particularly so for the organism-specific models. This feature is very strongly correlated with hydrophobicity, with  $r^2 = 0.94$  according to (Georgiev, 2009). For the windowed data representation used in this work, it measures the average hydrophobicity of the 15-AA neighbourhood of a given position on the protein. Hydrophobicity/hydrophilicity are directly related to epitope accessibility in the protein structure. Hydrophilic polar regions are usually observed in the protein surface, been constantly exposed to antibodies, whereas hydrophobic regions often interact either with each other in the protein core or with other cellular components, and are not readily accessible to the serologic immune response



(Hopp and Woods, 1981). Other features that seem to appear consistently amongst the most relevant ones (although not as prominently as BLOSUM1), e.g. ProtFP1, Z1, VHSE8 and F5 (see [Supplementary File S6, Fig. S6-6](#)) are composite scales based on algebraic transformations of underlying physicochemical properties, and lack the same direct interpretability as BLOSUM1, which prevents the derivation of biochemical hypotheses.

Another aspect that may suggest an explanation for the increased performance of organism-specific models is a possible difference in the spatial distribution of epitopes in the feature space, conditional on the pathogen. To explore the neighbourhood structure of the data, we have employed t-SNE projections (Van der Maaten and Hinton, 2008) to investigate whether data coming from distinct pathogens present different clustering

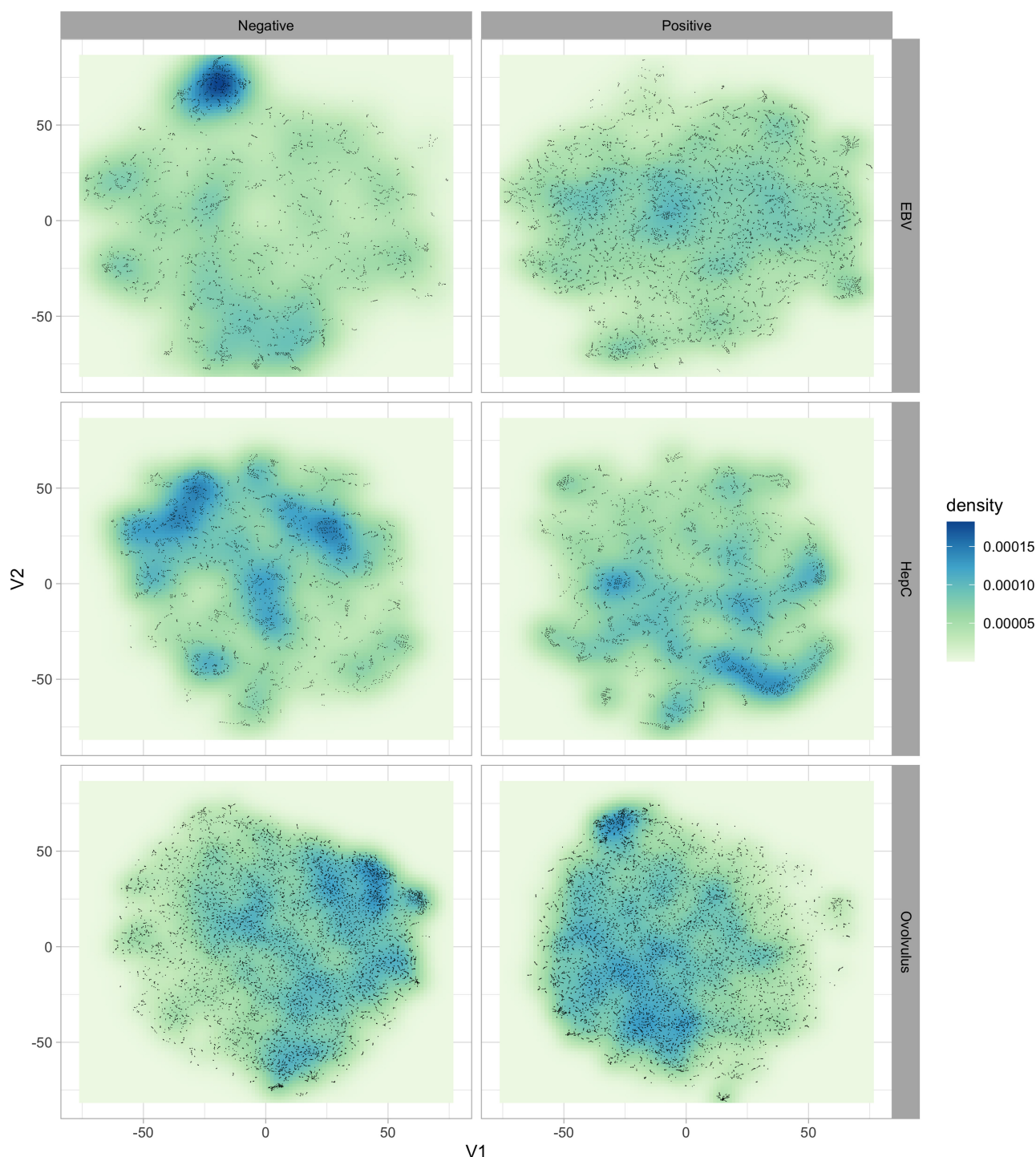


Fig. 3. Estimated probability density of epitope and non-epitope observations in the t-SNE projection. Notice the clear distinct regions of high density of positive/negative observations, which occupy different portions of the feature space. This figure clearly illustrates how epitopes (*positive* observations) of different pathogens tend to occur in very distinct regions of the space of features. More importantly, regions that present a high density of positive examples for one pathogen can simultaneously have high numbers of negative observations for another—see, e.g. how the top-left portion of the negative examples of EBV and HepC coincide with a corresponding high-density regions of positive *O.volvulus* points. Models trained on combined (heterogeneous) data would not be able to explore these patterns, and would likely fail to detect promising regions, which may explain the increased performance of the organism-specific models when compared against generalist ones trained on heterogeneous data

or neighbourhood structures in terms of positive/negative observations.

Figure 3 illustrates the estimated density of observations on the 2D t-SNE projection of the data, stratified by pathogen and class. The V1-V2 coordinates are consistent across the different panels, and the figure clearly shows how the density of positive and negative examples not only varies depending on the pathogen, but also how regions with a high density of positive examples for one organism can simultaneously contain high densities of negative examples for others. This type of pattern can help explain the success of organism-specific training from a data mining perspective (albeit not necessarily from a biological one): generalist models trained on heterogeneous data would not be able to pick up these organism-specific patterns, as they would appear as having a more mixed combination of positive and negative examples if the data from multiple pathogens were combined into a single training set. This could in effect prevent those models from detecting regions of the feature space that were potentially rich in epitopes of a specific pathogen, resulting in decreased predictive performance.

## 4 Conclusions

In this article, we investigated the use of organism-specific data for improving the performance of linear B-cell epitope prediction. Organism-specific Random Forest models developed for three distinct pathogens (Epstein-Barr virus, Hepatitis C virus and the roundworm *O.volvulus*) yielded significant performance gains when compared with similar models trained using heterogeneous and hybrid datasets, across several relevant performance indicators. These results suggest that pre-selecting the most relevant data and training bespoke models for specific pathogens is preferable to the common strategy of increasing and diversifying the training set.

Performance comparisons also indicate that this organism-specific modelling strategy is able to provide results that are at least as good as, and in several cases better than, several common predictors from the literature, despite the fact that (i) the predictors trained in this study were relatively simple proof-of-concept models, without specific refinements; and (ii) only basic features, calculated from the AA sequence alone, were employed, without any sophisticated feature engineering performed. We expect that further refinements to organism-specific predictors, such as model improvements or the use of more informative features, may result in even higher predictive performance. While these results do not obviate the utility of generalist predictors—which are still very relevant in the investigation of pathogens for which little or no specific data is available—they certainly suggest a powerful and easily generalizable new approach for researchers working with relatively data-abundant organisms.

## Acknowledgement

This work was originally inspired by discussions with Prof. Daniella Bartholomeu from the Laboratory of Parasite Genomics, UFMG, Brazil.

## Funding

This work was supported by the Engineering and Physical Sciences Research Council (J.A.: Ph.D. fees and stipend); by the Coordination for the Improvement of Higher Education Personnel, CAPES (J.R.C.: salary; I.L.: Ph.D. stipend).

*Conflict of Interest:* none declared.

## References

Alix,A.J. (1999) Predictive estimation of protein linear epitopes by using the program people. *Vaccine*, 18, 311–314.  
 Arnold,B.F. et al. (2018) Integrated serologic surveillance of population immunity and disease transmission. *Emerging Infect. Dis.*, 24, 1188–1194.

Basáñez,M.-G. et al. (2006) River blindness: a success story under threat? *PLoS Med.*, 3, e371.  
 Blythe,M.J. and Flower,D.R. (2005) Benchmarking b cell epitope prediction: underperformance of existing methods. *Protein Sci.*, 14, 246–248.  
 Chicco,D. and Jurman,G. (2020) The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 6.  
 Collatz,M. et al. (2021) EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics*, 37, 448–455.  
 Davison,A.C. and Hinkley,D.V. (2013) *Bootstrap Methods and Their Application*. Cambridge University Press, USA.  
 EL-Manzalawy,Y. et al. (2008) Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit. Interdiscipl. J.*, 21, 243–255.  
 Ferri,C. et al. (2015) HCV syndrome: a constellation of organ- and non-organ specific autoimmune disorders, B-cell non-Hodgkin's lymphoma, and cancer. *World J. Hepatol.*, 7, 327–343.  
 Forsström,B. et al. (2015) Dissecting antibodies with regards to linear and conformational epitopes. *PLoS One*, 10, e0121673.  
 Georgiev,A.G. (2009) Interpretable numerical descriptors of amino acid space. *J. Comput. Biol.*, 16, 703–723.  
 Getzoff,E.D. et al. (1988) The chemistry and mechanism of antibody binding to protein antigens. In: Dixon F.J.(ed.), *Advances in Immunology*, Vol. 43, pp. 1–98, Academic Press, Cambridge, MA, USA.  
 Giacò,L. et al. (2012) B-pred, a structure based B-cell epitopes prediction server. *Adv. Appl. Bioinf. Chem.*, 5, 11–21.  
 Greenbaum,J.A. et al. (2007) Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J. Mol. Recognit. Interdiscipl. J.*, 20, 75–82.  
 Haste Andersen,P. et al. (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.*, 15, 2558–2567.  
 Holm,S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Stat.*, 6, 65–70.  
 Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA*, 78, 3824–3828.  
 Jespersen,M.C. et al. (2017) Bepipred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.*, 45, W24–W29.  
 Jespersen,M.C. et al. (2019) Antibody specific B-cell epitope predictions: leveraging information from antibody-antigen protein complexes. *Front. Immunol.*, 10, 298.  
 Kaufman,S. et al. (2011) Leakage in data mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD'11*. ACM Press, San Diego, CA, USA.  
 Kindt,T.J. et al. (2007) *Kuby Immunology*. Macmillan Learning. New York, NY, USA.  
 Kolaskar,A. and Tongaonkar,P.C. (1990) A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Lett.*, 276, 172–174.  
 Kulkarni-Kale,U. et al. (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res.*, 33, W168–W171.  
 Dudek,N. et al. (2010) Epitope discovery and their use in peptide based vaccines. *Curr. Pharm. Des.*, 16, 3149–3157.  
 Larsen,J.E.P. et al. (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res.*, 2, 2.  
 Leinikki,P. et al. (1993) Synthetic peptides as diagnostic tools in virology. In: K. Maramorosh (eds.) et al., *Advances in Virus Research*, Vol. 42, pp. 149–186, Academic Press, Cambridge, MA, USA.  
 Lo,Y.-T. et al. (2013) Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics. *BMC Bioinformatics*, 14, S3.  
 Lodish,H. et al. (2000) *Molecular Cell Biology*, 4th edn. W.H.Freeman & Co Ltd. New York, NY, USA.  
 Manavalan,B. et al. (2018) iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front. Immunol.*, 9, 1695.  
 NCBI Resource Coordinators. (2015) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, 44, D7–D19.  
 Osei-Atweneboana,M.Y. et al. (2007) Prevalence and intensity of *Onchocerca volvulus* infection and efficacy of ivermectin in endemic communities in Ghana: a two-phase epidemiological study. *Lancet*, 369, 2021–2029.  
 Osorio,D. et al. (2015) Peptides: a package for data mining of antimicrobial peptides. *R. J.*, 7, 4–14.  
 Pandurangan,A.P. and Blundell,T.L. (2020) Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and MCSM, using machine learning. *Protein Sci.*, 29, 247–257.  
 Parker,J. et al. (1986) New hydrophobicity scale derived from high-performance liquid chromatography peptide retention data:

- correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, **25**, 5425–5432.
- Paul, W. (2012) *Fundamental Immunology*, 7th edn. Lippincott Williams & Wilkins, London.
- Pellequer, J. and Westhof, E. (1993) Preditop: a program for antigenicity prediction. *J. Mol. Graph.*, **11**, 204–210.
- Pellequer, J. *et al.* (1991) Predicting location of continuous epitopes in proteins from their primary structures. In: Langone J.J. (ed.), *Methods in Enzymology*, Vol. **203**. pp. 176–201, Elsevier, Amsterdam.
- Pellequer, J.-L. *et al.* (1993) Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol. Lett.*, **36**, 83–99.
- Ponomarenko, J.V. and Bourne, P.E. (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. *BMC Struct. Biol.*, **7**, 64.
- Potocnakova, L. *et al.* (2016) An introduction to B-cell epitope mapping and in silico epitope prediction. *J. Immunol. Res.*, **2016**, 6760830.
- R Core Team. (2020) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rezk, S.A. *et al.* (2018) Epstein–Barr virus (EBV)-associated lymphoid proliferations, a 2018 update. *Hum. Pathol.*, **79**, 18–41.
- Saha, S. and Raghava, G.P.S. (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: *Third International Conference on Artificial Immune Systems, Sicily, Italy*, pp. 197–204, Springer.
- Saha, S. and Raghava, G.P.S. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins Struct. Funct. Bioinf.*, **65**, 40–48.
- Sanchez-Trincado, J.L. *et al.* (2017) Fundamentals and methods for T- and B-cell epitope prediction. *J. Immunol. Res.*, **2017**, 2680160.
- Saravanan, V. and Gautham, N. (2015) Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor. *Omics J. Integr. Biol.*, **19**, 648–658.
- Shen, J. *et al.* (2007) Predicting protein–protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA*, **104**, 4337–4341.
- Singh, H. *et al.* (2013) Improved method for linear B-cell epitope prediction using antigen's primary sequence. *PLoS One*, **8**, e62216.
- Tan, P.-N. *et al.* (2005) *Introduction to Data Mining*. Addison Wesley, Boston, MA, USA.
- UniProt Consortium. (2020) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
- Van der Maaten, L. and Hinton, G. (2008) Visualizing data using t-SNE. *J. Mach. Learn. Res.*, **9**, 2579–2605.
- Van Regenmortel, M.H. (1996) Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods*, **9**, 465–472.
- Vita, R. *et al.* (2019) The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res.*, **47**, D339–D343.
- Wang, J. *et al.* (2017) Protein–protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. *Int. J. Mol. Sci.*, **18**, 2373.
- World Health Organization. (2019) *Onchocerciasis Fact Sheet*. <https://www.who.int/news-room/fact-sheets/detail/onchocerciasis> (22 July 2020, date last accessed).
- Wright, M.N. and Ziegler, A. (2017) ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J. Stat. Softw.*, **77**, 1–17.
- Yang, X. and Yu, X. (2009) An introduction to epitope prediction methods and software. *Rev. Med. Virol.*, **19**, 77–96.
- Yao, B. *et al.* (2012) SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PLoS One*, **7**, e45152.
- Yao, B. *et al.* (2013) Conformational B-cell epitope prediction on antigen protein structures: a review of current algorithms and comparison with common binding site prediction methods. *PLoS One*, **8**, e62249.