



VirusHound-I: prediction of viral proteins involved in the evasion of host adaptive immune response using the random forest algorithm and generative adversarial network for data augmentation

Jorge F. Beltrán , Lisandra Herrera Belén, Jorge G. Farias, Mauricio Zamorano , Nicolás Lefin, Javiera Miranda and

Fernanda Parraguez-Contreras

Corresponding author: Jorge F. Beltrán, Department of Chemical Engineering, Faculty of Engineering and Science, Universidad de La Frontera, Ave. Francisco Salazar 01145, Temuco, Chile. Tel.: +56 999 281342; E-mail: beltran.lissabet.jf@gmail.com

Abstract

Throughout evolution, pathogenic viruses have developed different strategies to evade the response of the adaptive immune system. To carry out successful replication, some pathogenic viruses encode different proteins that manipulate the molecular mechanisms of host cells. Currently, there are different bioinformatics tools for virus research; however, none of them focus on predicting viral proteins that evade the adaptive system. In this work, we have developed a novel tool based on machine and deep learning for predicting this type of viral protein named VirusHound-I. This tool is based on a model developed with the multilayer perceptron algorithm using the dipeptide composition molecular descriptor. In this study, we have also demonstrated the robustness of our strategy for data augmentation of the positive dataset based on generative antagonistic networks. During the 10-fold cross-validation step in the training dataset, the predictive model showed 0.947 accuracy, 0.994 precision, 0.943 F1 score, 0.995 specificity, 0.896 sensitivity, 0.894 kappa, 0.898 Matthew's correlation coefficient and 0.989 AUC. On the other hand, during the testing step, the model showed 0.964 accuracy, 1.0 precision, 0.967 F1 score, 1.0 specificity, 0.936 sensitivity, 0.929 kappa, 0.931 Matthew's correlation coefficient and 1.0 AUC. Taking this model into account, we have developed a tool called VirusHound-I that makes it possible to predict viral proteins that evade the host's adaptive immune system. We believe that VirusHound-I can be very useful in accelerating studies on the molecular mechanisms of evasion of pathogenic viruses, as well as in the discovery of therapeutic targets.

Keywords: virus; pathogen; machine learning; neural network; deep learning; protein

Jorge F. Beltrán is an expert in Bioinformatics and Data Science, focusing on Machine Learning and Deep Learning applied to biology and chemistry. He holds a PhD in Applied Cellular and Molecular Biology. Additionally, he has specialized training in Artificial Intelligence, excelling in the analysis and interpretation of large biological and chemical data sets. His primary fields of study are bioinformatics, artificial intelligence, and molecular and cellular biology, tackling scientific challenges with advanced data analysis techniques.

Lisandra Herrera Belén is an academic and researcher at the University of Santo Tomás in Temuco, Chile, specializing in bioinformatics, genetic engineering, and pharmaceutical chemistry. Her work focuses on the design of a chimeric (recombinant) enzyme using advanced bioinformatics tools, a project that merges aspects of biochemistry and pharmacology. Additionally, she has conducted research in the genetic modification of asparaginase, an enzyme from *Escherichia coli*, marking a significant advancement in the field of pharmaceutical chemistry and enzyme therapy.

Jorge Farias Avendaño, newly elected as the dean of the Faculty of Engineering and Sciences at the University of La Frontera, holds a distinguished academic and scientific profile. He is a biochemist from the Pontifical Catholic University of Valparaíso and holds a PhD in Biotechnology from the Federico Santa María Technical University and Pontifical Catholic University of Valparaíso. Dr. Farias is an active member of several national and international scientific societies, including the Chilean Society of Reproduction and Development, the Pharmacology Society of Chile, The American Society For Cell Biology, The American Physiology Society, the American Society of Andrology, and the American Society for Biochemistry and Molecular Biology.

Mauricio Zamorano is an Assistant Professor in the Department of Chemical Engineering at the University of La Frontera in Temuco, Chile. His primary research interest focuses on stem cell processing in three-dimensional (3D) environments, the design of medical devices, and clinical-scale bioreactors. Zamorano has a special interest in the production of bone, cartilaginous, and dental tissues on biodegradable scaffolds. In his research, he employs advanced techniques such as media perfusion, biophysical signaling, and the use of physiological levels of cytokines to optimize cell growth and differentiation in these environments. His work significantly contributes to the field of tissue engineering and regenerative medicine.

Nicolás Lefin is a graduate student at the University of La Frontera, Chile. In 2023, he graduated with honors in Chemical Civil Engineering from the same university, where he received the 'Faculty Award' in recognition of his academic excellence and outstanding contributions. Currently, Nicolás is continuing his academic training as a master's student in Biotechnology Engineering at the University of La Frontera, in collaboration with University of São Paulo, Brasil. Specializing in the fields of bioprocessing, molecular biology, and protein engineering.

Javiera Miranda is a graduate student at the University of La Frontera, Chile. In 2023, he graduated with honors in Biotechnology Engineering from the same university, where he received the 'Faculty Award' in recognition of his academic excellence and outstanding contributions. Currently, Javiera is continuing his academic training as a master's student in Biotechnology Engineering at the University of La Frontera, in collaboration with University of São Paulo, Brasil. Specializing in the fields of bioprocessing, molecular biology, and protein engineering.

Fernanda Parraguez is a master's student at the University of La Frontera, recently graduated with academic excellence from the Civil Engineering in Biotechnology program. She is currently continuing her studies in the Master of Science in Engineering program, specializing in Biotechnology at the University of La Frontera, focusing on water treatment, nanotechnology, and photocatalysis.

Received: May 25, 2023. **Revised:** October 18, 2023. **Accepted:** November 5, 2023

© The Author(s) 2023. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

INTRODUCTION

There have been significant pandemics in the last century, such as the 1918 H1N1 influenza and HIV, which caused millions of deaths. In the twenty-first century, there have been zoonotic outbreaks, including SARS, MERS, Ebola, Hendra, Nipah and COVID-19, which have become a major crisis with global consequences [1]. Pathogenic viruses have evolved various strategies to evade the immune system and successfully infect host cells [2, 3]. The adaptive immune response is a key component of the host's immune system against viral pathogens, and many of these pathogens have evolved to evade it [4]. Among the most common mechanisms used by pathogenic viruses to evade this type of response are inhibition of host major histocompatibility complex class I (MHC-I) molecule presentation [5–7], MHC class II (MHC-II) molecule presentation, proteasome antigen processing, transporter associated with antigenic processing (TAP), and tapasin [7]. For example, the human papillomavirus (HPV) encodes a protein called E5 (HPV E5) that facilitates successful infection by inducing loss of surface MHC-I expression in infected basal cells, thereby preventing viral antigen presentation to effector T cells. HPV16 E5 can bind to the transmembrane domain of MHC-I, retaining it inside the Golgi apparatus, thus preventing its trafficking to the cell surface [8, 9]. The US2 protein of human cytomegalovirus inhibits the MHC-II antigen presentation pathway by degrading human leukocyte antigen (HLA)-DR- α and -DM- α , thereby preventing recognition by CD4⁺ T cells [10]. Finally, the Epstein–Barr virus encoded nuclear antigen 1 protein interrupts proteasome substrate processing [11], among many other examples. The study of virus proteins that evade adaptive immune responses is crucial in the development of vaccines and therapeutic drugs that help combat these pathogens more efficiently [12]. However, despite advances in understanding the molecular mechanisms by which viruses evade the immune system, demonstrating that a virus protein confers the ability to evade the immune system remains a difficult and resource-intensive task. In this direction, the development of alternative tools that allow for an accelerated process is an area of research that must be considered to combat these unpredictable and dangerous pathogens.

In recent years, different immunoinformatic tools based on machine learning algorithms have been developed to address various problems in the field of immunology, which have been trained with experimental datasets of immunogenic proteins and peptides. Most of these tools focus on predicting the binding affinity of peptides to MHC-I and MHC-II, as well as predicting B-cell epitopes, based solely on the primary sequences of proteins [13, 14], which facilitates the large-scale study of potentially immunogenic peptides. On the other hand, there are other immunoinformatic tools for the study of viruses based on machine learning such as VirVACPRED and Vaxijen, which allow for specific predictions of viral antigens [15, 16]. Other alternatives for the study of viruses are tools for predicting the subcellular localization of these pathogens such as MSLVP [17], Virus-mPLoc [18] and pLoc_Deep-mVirus [19], among many other tools compiled in an excellent review by Kumar and colleagues [20]. To date, there is no tool for predicting virus proteins that evade the host adaptive immune response. In this direction, the present work aims to develop a novel tool to address this important problem. This tool could be of great use in the study of the molecular mechanisms by which these pathogens camouflage themselves against this type of response, as well as facilitating the discovery of new therapeutic targets.

MATERIAL AND METHODS

Datasets

In this study, we identified 98 virus proteins involved in evading the host adaptive immune response (VPEs) from the scientific literature. The primary reference sequences of these proteins were identified and downloaded from the UniProt database [21] to construct a positive dataset (positive dataset). On the other hand, to create the negative dataset, we randomly selected 285 virus proteins (negative dataset) lacking this biological functionality (non-VPEs) in the UniProt database. These were also compared with the scientific literature to ensure that they are not involved in invading the host adaptive immune response.

Feature computation

From both datasets, four different types of molecular descriptors were calculated: amino acid composition (AAC, 20 features), amphiphilic pseudo amino acid composition (APAAC, 50 features), dipeptide composition (DPC, 400 features) and pseudo amino acid composition (PAAAC, 50 features). All calculations of the molecular descriptors were performed using the propy3 package (<https://propy3.readthedocs.io/>).

Data augmentation using generative adversarial network

GAN architecture

Considering the disproportion between the positive and negative datasets in terms of the number of virus protein sequences, we proceeded to balance this imbalance by generating synthetic data from the positive dataset. For this purpose, a Generative Adversarial Network (GAN) was used, and it was separately fed with each of the molecular descriptors calculated from the positive dataset. The implementation of our strategy for generating synthetic data consists of three models: the generator, the discriminator and the GAN. The generator was provided with a noise input, and synthetic data resembling the real data were generated from this input. Subsequently, the architecture was configured so that the discriminator received real or synthetic data and attempted to classify them as true or false. The GAN developed in this work combines the generator and the discriminator in a neural network to train them together. The generator seeks to improve its ability to generate increasingly realistic synthetic data, while the discriminator aims to improve its ability to distinguish between real and synthetic data. The GAN was trained in a zero-sum game process where the generator tries to deceive the discriminator, and the discriminator tries to correctly identify the fake data. This feedback process was iteratively repeated until the GAN was capable of generating synthetic data that is indistinguishable from the real data.

Generator configuration

A neural network was configured consisting of three central layers. First, an input layer with a dimensionality of 128 was used. Subsequently, hidden layers were implemented, each composed of two dense layers with 128 units. These layers are followed by LeakyReLU activations with a slope factor of 0.01 and dropout layers with a retention rate of 50%. Finally, the output layer was configured as a dense layer designed to generate synthetic data, maintaining the same dimensionality as the input dataset.

Discriminator configuration

An input layer with a dimensionality equal to the number of entities in the dataset was used. The hidden layers consist of

two dense layers, each comprising 128 units. These layers were followed by LeakyReLU activations with a slope factor of 0.01 and dropout layers with a retention factor of 0.5. The output layer was configured as a dense layer with a sigmoid activation for binary classification of data as real or synthetic. The discriminator was trained to distinguish between real and synthetic data using the binary cross-entropy loss function and was optimized with the Adam algorithm, configured with a learning rate of 0.0002 and a beta factor of 0.5. Simultaneously, the GAN was trained over 1000 epochs, with a batch size of 30 examples per epoch, aiming to deceive the discriminator by generating synthetic data resembling real data. Losses for both the discriminator and the generator were recorded in each epoch, and these losses were summed as the evaluation metric.

The entire strategy carried out for generating synthetic data was written using the TensorFlow 2.0 framework (<https://www.tensorflow.org/>). The code written for this task was deposited along with all the code used in this work and the datasets in the GitHub repository: <https://github.com/jfbldevs/virushound-1>.

Classification and assessments

The random forest classification algorithm (RF) was used to develop predictive models for viral proteins that evade the adaptive immune system (abbreviated as VPEs). Before developing models using each molecular descriptor tested, robust hyperparameter optimizations were performed on the training dataset, made up of 80% of the data, and subjected to a 10-fold cross-validation step. Cross-validation is a statistical technique used to assess the performance and generalizability of a predictive model. It involves dividing a dataset into multiple subsets, typically a training set and a validation set, multiple times to ensure robust model evaluation. This helps to mitigate the risk of overfitting and provides a more reliable estimate of how well the model will perform on new, unseen data [22, 23]. Subsequently, the optimized models were tested on the remaining 20% of the data (unseen data). The hyperparameters and their associated values were as follows, *n_estimators*: the number of trees in the forest. The possible values range from 100 to 1000, with increments of 100. *criterion*: The function to measure the quality of a split. The possible values are 'gini' (Gini impurity) and 'entropy' (information gain). *max_depth*: The maximum depth of the tree. The possible values are 5, 10, 15, 20, 25, 30 and None (unlimited depth). *min_samples_split*: The minimum number of samples required to split an internal node. The possible values are 2, 5 and 10. *min_samples_leaf*: The minimum number of samples required to be at a leaf node. The possible values are 1, 2 and 4. *max_features*: The number of features to consider when looking for the best split. The possible values are 'sqrt' (square root of the total number of features), 'log2' (log2 of the total number of features) and None (all features). *bootstrap*: Whether bootstrap samples are used when building trees. The possible values are True and False. *class_weight*: Weights associated with classes. The possible values are None (all classes have equal weight), 'balanced' (weights are inversely proportional to class frequencies) and 'balanced_subsample' (similar to 'balanced' but computed based on the bootstrap sample for every tree). *min_weight_fraction_leaf*: The minimum weighted fraction of the sum total of weights required to be at a leaf node. The possible values are 0.0, 0.1 and 0.2. *max_leaf_nodes*: The maximum number of leaf nodes in the tree. The possible values are None (unlimited leaf nodes), 5, 10, 20 and 50. *ccp_alpha*: Complexity parameter used for Minimal Cost-Complexity Pruning. The possible values are 0.0, 0.1 and 0.2. The scikit-learn machine learning library

(<https://scikit-learn.org/>) was used for the entire workflow during the development of all predictive models. The following performance measures were evaluated for this binary classification problem:

$$\text{Sensitivity (TPR)} = \text{TP} / (\text{TP} + \text{FN}) \quad (1)$$

$$\text{Accuracy (ACC)} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN}) \quad (2)$$

$$\text{Precision (PPV)} = \text{TP} / (\text{TP} + \text{FP}) \quad (3)$$

$$\text{F1 score (F1)} = 2\text{TP} / (2\text{TP} + \text{FP} + \text{FN}) \quad (4)$$

$$\text{Specificity (TNR)} = \text{TN} / (\text{FP} + \text{TN}) \quad (5)$$

$$k (\text{kappa}) = 2 * (\text{TP} * \text{TN} - \text{FN} * \text{FP}) / ((\text{TP} + \text{FP}) * (\text{FP} + \text{TN}) + (\text{TP} + \text{FN}) * (\text{FN} + \text{TN})) \quad (6)$$

$$\text{MCC} = (\text{TP}(\text{TN}) - (\text{FP})(\text{FN})) / \sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})} \quad (7)$$

with Matthew's correlation coefficient (MCC), true positive (TP), false positive (FP), true negative (TN) and false negative (FN). Along with these performance measures, the receiver operating characteristic (ROC) curve was also evaluated at all stages of predictive model assessment. The ROC curve compares two operating characteristics (TPR and FPR), where TPR is the sensitivity mentioned earlier and FPR is the false positive rate defined as

$$\text{FPR} = \text{FP} / (\text{FP} + \text{TN}) \quad (8)$$

On the other hand, a web application named VirusHound-I was developed using the Python 3.11 programming language (<https://www.python.org/>). This web application scores the outputs with a probability ranging from 0 to 1. Figure 1 shows all the workflow implemented in this study (Figure 1).

RESULTS

In this work, using a GAN-based strategy and starting from the four molecular descriptors evaluated, synthetic data were generated to augment the positive dataset. The plots of the molecular descriptor values for the synthetic data, when compared with the real data, showed that both exhibit a high similarity in all cases (Figure 2). Taking these results into account, datasets were created for the subsequent training and testing phases. The synthetic data was used to augment the positive datasets for VPE classification, by using RF (Figure 1). The 10-fold cross-validation on the training dataset showed that, in general, all evaluated descriptors allow obtaining predictive models with good performance according to the assessed metrics (Table 1 and Figure 3).

On the other hand, good performance measures were also observed during the testing step, demonstrating the efficiency of the models in generalizing over new data (Table 1 and Figure 3). While all models presented excellent performance measures during the mentioned stages, we highlight the predictive model based on the DPC molecular descriptor because it showed the best performance measures on the test dataset, indicating better generalization over new data (Table 1 and Figure 3H). In this regard, this model was selected and incorporated into our web application VirusHound-I for VPE predictions. Considering the limited computational resources available to us to date, we limited the analysis to only 100 virus sequences per query. However, this

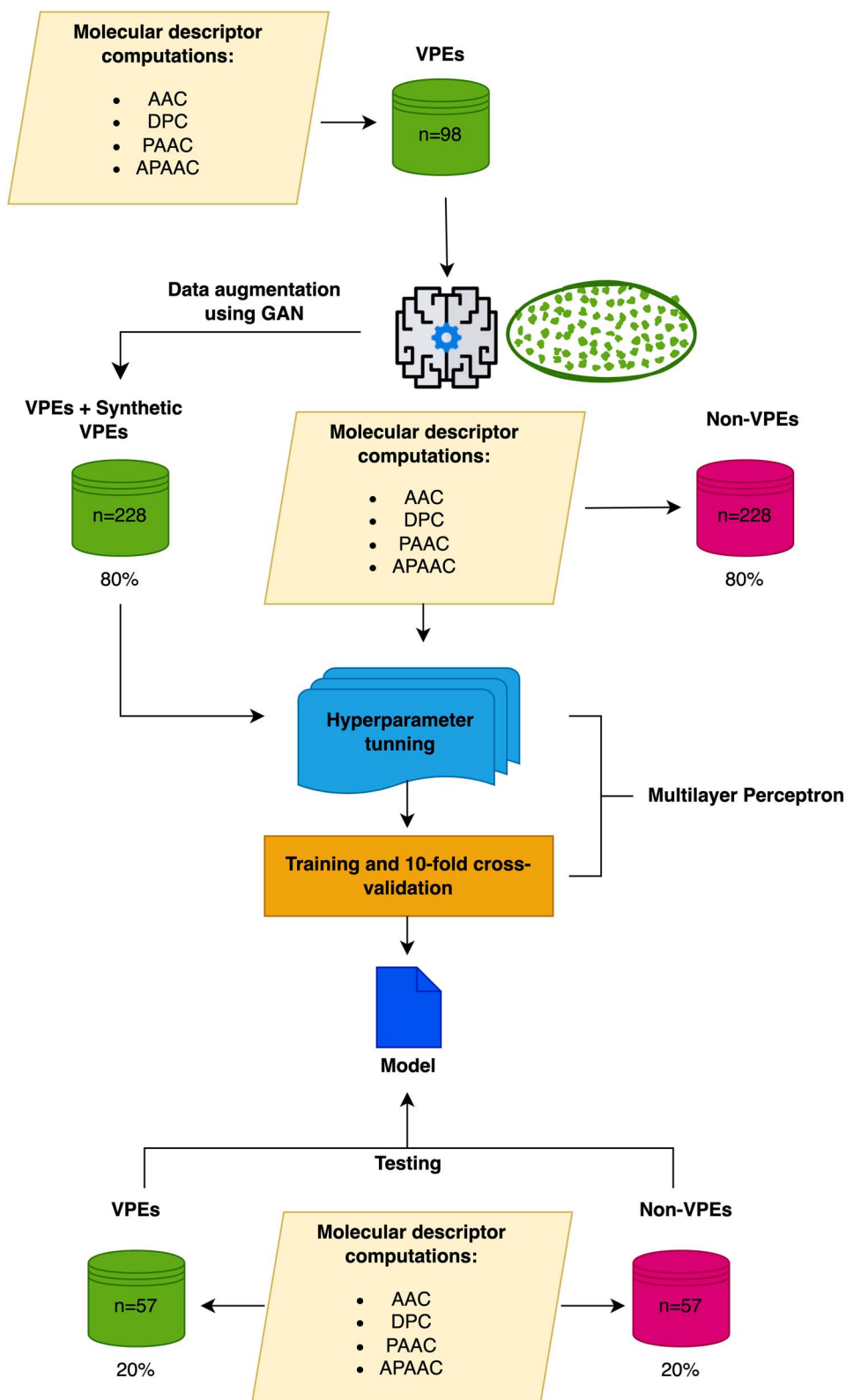


Figure 1. Workflow addressed in the present study. In the first stage, data were extracted from the UniProt platform. Next, the positive datasets were augmented using a GAN, based on four types of calculated molecular descriptors. Finally, the whole datasets (positive + negative) were divided and used to train and test predictive VPE models.

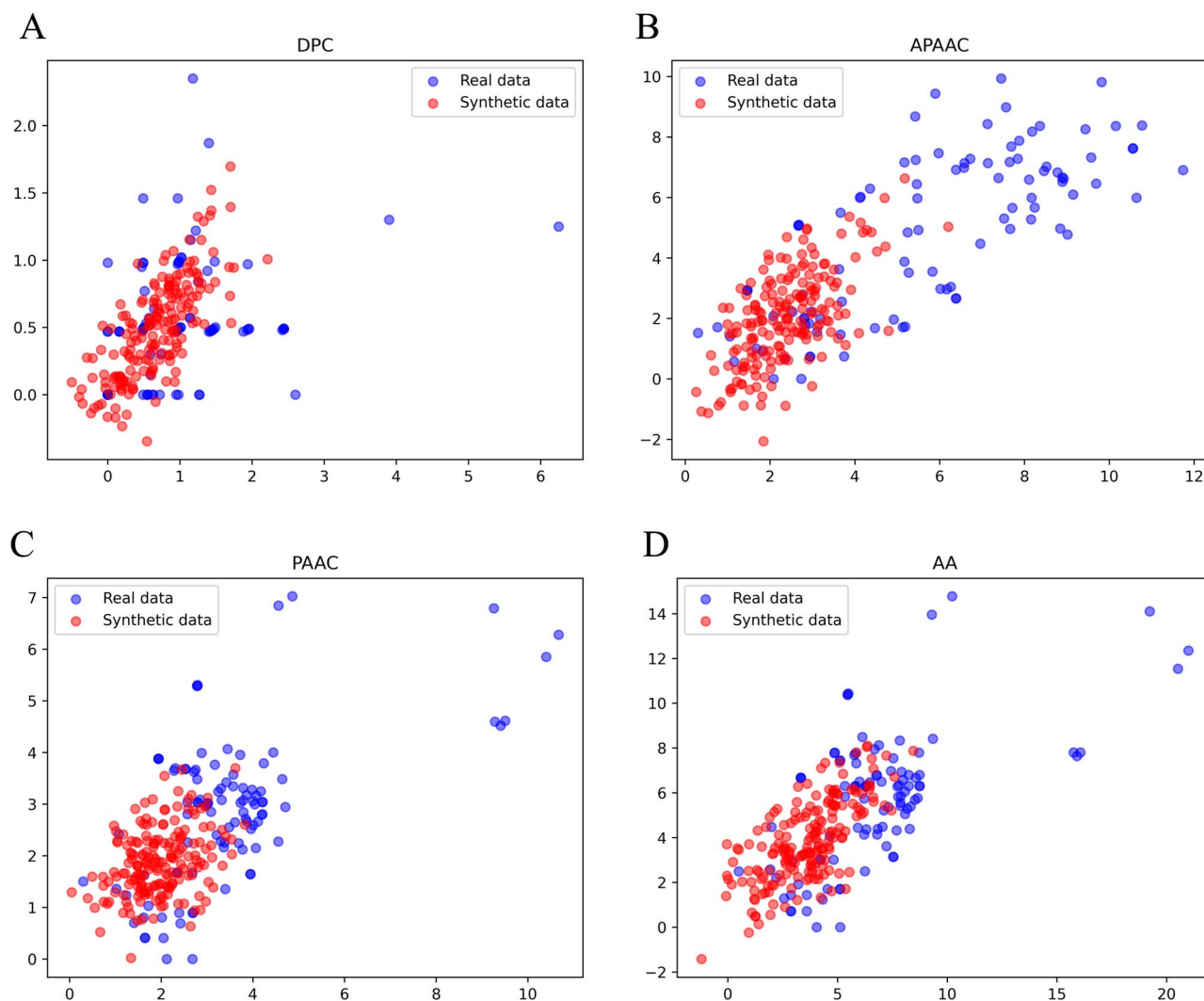


Figure 2. Plots of real values corresponding to each evaluated molecular descriptor and false values generated with the GAN.

Table 1: Performance measures were obtained through 10-fold cross-validation on the training data, with the testing phase conducted using RF

MD/P	ACC	F1	PPV	TNR	TPR	k	MCC	AUC
AA/CVTr	0.942	0.939	0.966	0.970	0.914	0.885	0.886	0.986
AAC/Te	0.956	0.958	1.0 ^{Te}	1.0 ^{Te}	0.920	0.912	0.915	0.998
APAAC/CVTr	0.953 ^{Tr}	0.951 ^{Tr}	0.985	0.987	0.918 ^{Tr}	0.907 ^{Tr}	0.909 ^{Tr}	0.989 ^{Tr}
APAAC/Te	0.947	0.95	1.0 ^{Te}	1.0 ^{Te}	0.904	0.894	0.899	0.995
DPC/CVTr	0.947	0.943	0.994 ^{Tr}	0.995 ^{Tr}	0.896	0.894	0.898	0.989 ^{Tr}
DPC/Te	0.964 ^{Te}	0.967 ^{Te}	1.0 ^{Te}	1.0 ^{Te}	0.936	0.929 ^{Te}	0.931 ^{Te}	1.0 ^{Te}
PAAC/CVTr	0.929	0.926	0.943	0.948	0.909	0.859	0.859	0.984
PAAC/Te	0.964 ^{Te}	0.967 ^{Te}	0.983	0.980	0.952 ^{Te}	0.929 ^{Te}	0.929	0.995

The measures obtained from the 10-fold cross-validation represent the averaged values from each fold. MD/P: molecular descriptor/phase, CVTr: 10-fold cross-validation phase on training data, Te: testing phase, AAC: amino acid composition, APAAC: amphiphilic pseudo amino acid composition, DPC: dipeptide composition, PAAC: pseudo amino acid composition.

number will gradually increase in the future as much as possible. The VirusHound-I tool is freely available at <https://www.biochemintelli.com/VirusHound-I>.

DISCUSSION

Throughout evolution, pathogenic viruses of humans and animals have developed numerous strategies to camouflage themselves from the immune system [12, 24]. One of these strategies is the

coding of proteins that evade the adaptive immune system, which is crucial for the successful replication of these pathogens in host cells [7, 25]. Therefore, identifying such viral proteins is crucial to developing vaccines and therapeutic drugs that can help combat these viruses.

In recent years, machine learning techniques have been key in the development of bioinformatics tools for studying virus proteins like the examples mentioned above [15–18], as well as others for the discovery of peptides with antiviral activity

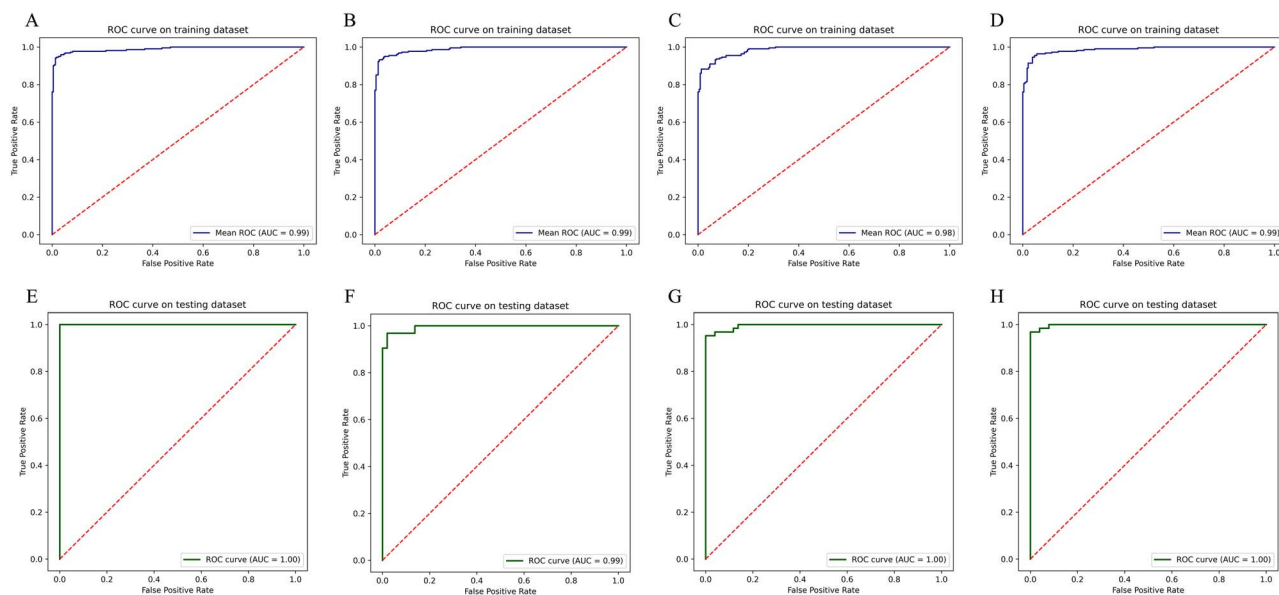


Figure 3. ROC curve for each molecular descriptor evaluated in the training and testing stages. AA (A and E), APAAC (B and F), PAAC (C and G) and DPC (D and H).

[26–34], which have experienced a recent explosion as a result of the lessons learned from the COVID-19 pandemic crisis. Discovering viral proteins that evade the immune system is a challenging task through both *in vivo* and *in vitro* experimental methods [35–37]. Firstly, in *in vivo* studies, which involve observing the interaction between the virus and the immune system within living organisms, numerous ethical and practical limitations arise. The use of animal models or even humans for such research raises ethical dilemmas concerning exposure to potentially harmful viruses [38]. In addition, viruses can evolve rapidly, further complicating the detection of proteins evading the immune system in some scenarios [2, 5, 35–37, 39–41]. On the other hand, *in vitro* studies conducted in controlled laboratory environments with cell cultures also present significant challenges in discovering these proteins. Viruses are inherently complex, and *in vitro* systems often oversimplify viral interactions, which could lead to inaccurate or incomplete results [42, 43]. Ultimately, *in vitro* methods often require sophisticated and costly techniques, limiting their applicability on a larger scale.

While it is true that there has been an increase in research in the field of machine learning applied to the study of pathogenic viruses in recent years, to date, there are no studies focused on the prediction of VPEs. Within the field of deep learning, GANs have been shown to be very useful in the development of bioinformatic predictive models [44]. For example, this type of neural network has been used as a data augmentation technique in the study of protein post-translational modifications [45], antiviral peptides [28] and protein solubility [46], among other cases reported in an excellent review by Wan *et al.* [44], showing excellent results. The results of our study regarding data augmentation using the proposed GAN (Figure 2) allowed us to obtain excellent models with the four molecular descriptors evaluated (Figure 3 and Table 1), corroborating the robustness of this technique to deal with few data. It is important to highlight that, although in this study there were cases where the evaluated molecular descriptors represented low (AAC = 20 features, APAAC = 50 features, APAA = 50 features) and high (DPC = 400 features) dimensionality, in all cases, good performance measures were obtained using synthetic data generated with the GAN. In fact, our study

corresponds to what has been reported by Wan and Jones, who demonstrated that through a GAN-based approach, this type of neural network accurately learns the high-dimensional distributions of biophysical features based on protein sequences, as well as allowing the generation of high-quality synthetic protein feature samples that improve predictive model performance measures [47].

All the molecular descriptors in general allowed the development of predictive models of VPEs with good performance (Table 1 and Figure 3). However, we determined that models based on the molecular descriptors APAAC and DPC showed the best performances in predicting VPEs according to the metrics calculated during cross-validation and the testing stage (Table 1 and Figure 3). It is important to note that although both models performed well, they do exhibit differences. The APAAC-based model performed better in the training phase than the DPC-based model; however, the latter showed better performance on the independent test set. In this regard, we consider the DPC-based model to be a better choice for predicting VPEs, taking into account its better ability to generalize to new data. On the other hand, the computational resources needed to carry out predictions based on DPC are significantly lower than those required for APAAC. The utility of the DPC molecular descriptor has been widely demonstrated in many studies focused on the development of predictive models. For example, it has been used for the prediction of subcellular localization of eukaryotic proteins [48], antioxidant proteins [49], multiple subcellular localization of viral proteins [17], phage virion proteins [50], protein–protein interactions [51], thermophilic proteins [52], druggable proteins [53] and antifreeze proteins [54], among other studies. Consequently, considering the excellent performance of this molecular descriptor in our work, as well as the background of its successful use in the development of models based on machine learning, the DPC-based model was selected for the prediction of VPEs with the VirusHound-I tool.

Studying the proteins that allow viruses to evade the adaptive immune system is key to understanding how these pathogens infect host cells. Currently, there are some specific tools for studying viruses; however, none of them allow the prediction of VPEs. Therefore, we propose a machine learning-based tool called

VirusHound-I for VPE predictions. VirusHound-I is a powerful tool based on a model developed with the DPC molecular descriptor, which showed good performance during the training and testing stages. The primary innovation of VirusHound-I is to expedite the large-scale discovery of VPEs, which would otherwise be unattainable using conventional *in vitro* and *in vivo* methods without significant resource and time expenditure. This tool can be employed as a preliminary step prior to laboratory experiments to explore and reduce the number of VPE candidates. We believe that VirusHound-I can be very useful in understanding the molecular mechanisms by which pathogenic viruses evade the adaptive immune response, as well as in discovering new therapeutic targets.

Key Points

- Pathogenic viruses have evolved diverse strategies to evade the adaptive immune system by encoding proteins that manipulate host cell mechanisms.
- Existing bioinformatics tools for virus research do not address the prediction of viral proteins that evade the adaptive immune system.
- The study introduces VirusHound-I, a novel machine learning-based tool using the Composition + Transition + Distribution molecular descriptor and generative antagonistic networks for data augmentation.
- VirusHound-I demonstrates high accuracy, precision and specificity in predicting viral proteins evading the host's adaptive immune system, providing insights into the molecular mechanisms of pathogenic virus evasion and facilitating the discovery of potential therapeutic targets.

REFERENCES

- Bonneaud C, Longdon B. Emerging pathogen evolution. *EMBO Rep* 2020;**21**:21.
- Vossen MT, Westerhout EM, Söderberg-Nauclér C, Wiertz EJ. Viral immune evasion: a masterpiece of evolution. *Immunogenetics* 2002;**54**:527–42.
- Roetman JJ, Apostolova MKI, Philip M. Viral and cellular oncogenes promote immune evasion. *Oncogene* 2022;**41**:921–9.
- Forsyth KS, Eisenlohr LC. Giving CD4+ T cells the slip: viral interference with MHC class II-restricted antigen processing and presentation. *Curr Opin Immunol* 2016;**40**:123–9.
- Hewitt EW. The MHC class I antigen presentation pathway: strategies for viral immune evasion. *Immunology* 2003;**110**:163–9.
- van de Weijer ML, Luteijn RD, Wiertz EJHJ. Viral immune evasion: lessons in MHC class I antigen presentation. *Semin Immunol* 2015;**27**:125–37.
- Simmons RA, Willberg CB, Paul K. Immune evasion by viruses. *eLS* 2013.
- Ashrafi GH, Haghshenas M, Marchetti B, Campo MS. E5 protein of human papillomavirus 16 downregulates HLA class I and interacts with the heavy chain *via* its first hydrophobic domain. *Int J Cancer* 2006;**119**:2105–12.
- Cortese MS, Ashrafi GH, Campo MS. All 4 di-leucine motifs in the first hydrophobic domain of the E5 oncoprotein of human papillomavirus type 16 are essential for surface MHC class I downregulation activity and E5 endomembrane localization. *Int J Cancer* 2010;**126**:1675–82.
- Hegde NR, Tomazin RA, Wisner TW, et al. Inhibition of HLA-DR assembly, transport, and loading by human cytomegalovirus glycoprotein US3: a novel mechanism for evading major histocompatibility complex class II antigen presentation. *J Virol* 2002;**76**:10929–41.
- Zhang M, Coffino P. Repeat sequence of Epstein-Barr virus-encoded nuclear antigen 1 protein interrupts proteasome substrate processing. *J Biol Chem* 2004;**279**:8635–41.
- Hilleman MR. Strategies and mechanisms for host and pathogen survival in acute and persistent viral infections. *Proc Natl Acad Sci* 2004;**101**:14560–6.
- Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J Biomed Inform* 2015;**53**:405–14.
- Raoufi E, Hemmati M, Eftekhari S, et al. Epitope prediction by novel immunoinformatics approach: a state-of-the-art review. *Int J Pept Res Ther* 2020;**26**:1155–63.
- Herrera-Bravo J, Fariás JG, Contreras FP, et al. VirVACPRED: a web server for prediction of protective viral antigens. *Int J Pept Res Ther* 2022;**28**:35.
- Doytchinova IA, Flower DR. Vaxi Jen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 2007;**8**:4.
- Thakur A, Rajput A, Kumar M. MSLVP: prediction of multiple subcellular localization of viral proteins using a support vector machine. *Mol Biosyst* 2016;**12**:2572–86.
- Shen H-B, Chou K-C. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *J Biomol Struct Dyn* 2010;**28**:175–86.
- Shao Y, Chou K-C. pLoc_deep-mVirus: a CNN model for predicting subcellular localization of virus proteins by deep learning. *Nat Sci (Irvine)* 2020;**12**:388–99.
- Kumar S, Kumar GS, Maitra SS, et al. Viral informatics: bioinformatics-based solution for managing viral infections. *Brief Bioinform* 2022;**23**.
- Bateman A, Martin M-J, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res* 2021;**49**:D480–9.
- Wong T-T. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognit* 2015;**48**:2839–46.
- Schaffer C. Selecting a classification method by cross-validation. *Mach Learn* 1993;**13**:135–43.
- Iannello A, Debbeche O, Martin E, et al. Viral strategies for evading antiviral cellular immune responses of the host. *J Leukoc Biol* 2005;**79**:16–35.
- Bussey KA, Brinkmann MM. Strategies for immune evasion by human tumor viruses. *Curr Opin Virol* 2018;**32**:30–9.
- Beltrán Lissabet JF, Belén LH, Fariás JG. AntiVPP 1.0: a portable tool for prediction of antiviral peptides. *Comput Biol Med* 2019;**107**:127–30.
- Thakur N, Qureshi A, Kumar M. AVPPred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res* 2012;**40**:W199–204.
- Lin T-T, Sun Y-Y, Wang C-T, et al. AI4AVP: an antiviral peptides predictor in deep learning approach with generative adversarial network data augmentation. *Bioinformatics*. *Advances* 2022;**2**:2.
- Chowdhury AS, Reehl SM, Kehn-Hall K, et al. Better understanding and prediction of antiviral peptides through primary and secondary structure feature importance. *Sci Rep* 2020;**10**:19260.
- Zare M, Mohabatkar H, Faramarzi FK, et al. Using Chou's pseudo amino acid composition and machine learning method to predict the antiviral peptides. *Open Bioinform J* 2015;**9**:13–9.

31. Pang Y, Yao L, Jhong J-H, et al. AVPIden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. *Brief Bioinform* 2021;**22**:1–10.
32. Timmons PB, Hewage CM. ENNAVIA is a novel method which employs neural networks for antiviral and anti-coronavirus activity prediction for therapeutic peptides. *Brief Bioinform* 2021;**22**:1–17.
33. Qureshi A, Tandon H, Kumar M. AVP-IC₅₀ Pred: multiple machine learning techniques-based prediction of peptide antiviral activity in terms of half maximal inhibitory concentration (IC₅₀). *Biopolymers* 2015;**104**:753–63.
34. Schaduagratt N, Nantasenamat C, Prachayasittikul V, Shoombuatong W. Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int J Mol Sci* 2019;**20**:5743.
35. Alcamí A, Koszinowski UH. Viral mechanisms of immune evasion. *Immunol Today* 2000;**21**:447–55.
36. Beachboard DC, Horner SM. Innate immune evasion strategies of DNA and RNA viruses. *Curr Opin Microbiol* 2016;**32**:113–9.
37. Rubio-Casillas A, Redwan EM, Uversky VN. SARS-CoV-2: a master of immune evasion. *Biomedicine* 2022;**10**:1339.
38. Cleary SJ, Pitchford SC, Amison RT, et al. Animal models of mechanisms of SARS-CoV-2 infection and COVID-19 pathology. *Br J Pharmacol* 2020;**177**:4851–65.
39. Bravo IG, Féllez-Sánchez M. Papillomaviruses. *Evol Med Public Health* 2015;**2015**:32–51.
40. Carabelli AM, Peacock TP, Thorne LG, et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* 2023;**21**:162–77.
41. Donaldson EF, Lindesmith LC, Lobue AD, Baric RS. Norovirus pathogenesis: mechanisms of persistence and immune evasion in human populations. *Immunol Rev* 2008;**225**:190–211.
42. Chua SCJH, Tan HQ, Engelberg D, Lim LHK. Alternative experimental models for studying influenza proteins, host–virus interactions and anti-influenza drugs. *Pharmaceuticals* 2019;**12**:147.
43. Rosa RB, Dantas WM, JCF DN, et al. In vitro and in vivo models for studying SARS-CoV-2, the etiological agent responsible for COVID-19 pandemic. *Viruses* 2021;**13**:379.
44. Wan F, Kontogiorgos-Heintz D, de la Fuente-Nunez C. Deep generative models for peptide design. *Digital Discovery* 2022;**1**:195–208.
45. Yang Y, Wang H, Li W, et al. Prediction and analysis of multiple protein lysine modified sites based on conditional Wasserstein generative adversarial networks. *BMC Bioinformatics* 2021;**22**:171.
46. Han X, Zhang L, Zhou K, Wang X. ProGAN: protein solubility generative adversarial nets for data augmentation in DNN framework. *Comput Chem Eng* 2019;**131**:106533.
47. Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat Mach Intell* 2020;**2**:540–50.
48. Bhasin M, Raghava GPS. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res* 2004;**32**:W414–9.
49. Feng P, Chen W, Lin H. Identifying antioxidant proteins by using optimal dipeptide compositions. *Interdiscip Sci* 2016;**8**:186–91.
50. Charoenkwan P, Kanthawong S, Schaduagratt N, et al. PVPred-SCM: improved prediction and analysis of phage Virion proteins using a scoring card method. *Cell* 2020;**9**:353.
51. Du X, Sun S, Hu C, et al. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J Chem Inf Model* 2017;**57**:1499–510.
52. Charoenkwan P, Chotpatiwetchkul W, Lee VS, et al. A novel sequence-based predictor for identifying and characterizing thermophilic proteins using estimated propensity scores of dipeptides. *Sci Rep* 2021;**11**:23782.
53. Sikander R, Ghulam A, Ali F. XGB-DrugPred: computational prediction of druggable proteins using eXtreme gradient boosting and optimized features set. *Sci Rep* 2022;**12**:5505.
54. Khan A, Uddin J, Ali F, et al. Prediction of antifreeze proteins using machine learning. *Sci Rep* 2022;**12**:20672.