Briefings in Bioinformatics, 23(1), 2022, 1–16

https://doi.org/10.1093/bib/bbab412
Review

Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2

Balachandran Manavalan†, Shaherin Basith† and Gwang Lee

Corresponding authors: Balachandran Manavalan; Gwang Lee, Department of Physiology, Ajou University School of Medicine, Suwon 16499, Korea. Tel.: +(82)-31-219-4913, Fax: +(82)-31-219-5049. E-mail: bala@ajou.ac.kr (B.M.); Tel.: +(82)-31-219-4555, Fax: +(82)-31-219-5049. E-mail: glee@ajou.ac.kr
†These authors contributed equally to this work.

Abstract

Coronavirus disease 2019 (COVID-19) has impacted public health as well as societal and economic well-being. In the last two decades, various prediction algorithms and tools have been developed for predicting antiviral peptides (AVPs). The current COVID-19 pandemic has underscored the need to develop more efficient and accurate machine learning (ML)-based prediction algorithms for the rapid identification of therapeutic peptides against severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). Several peptide-based ML approaches, including anti-coronavirus peptides (ACVPs), IL-6 inducing epitopes and other epitopes targeting SARS-CoV-2, have been implemented in COVID-19 therapeutics. Owing to the growing interest in the COVID-19 field, it is crucial to systematically compare the existing ML algorithms based on their performances. Accordingly, we comprehensively evaluated the state-of-the-art IL-6 and AVP predictors against coronaviruses in terms of core algorithms, feature encoding schemes, performance evaluation metrics and software usability. A comprehensive performance assessment was then conducted to evaluate the robustness and scalability of the existing predictors using well-constructed independent validation datasets. Additionally, we discussed the advantages and disadvantages of the existing methods, providing useful insights into the development of novel computational tools for characterizing and identifying epitopes or ACVPs. The insights gained from this review are anticipated to provide critical guidance to the scientific community in the rapid design and development of accurate and efficient next-generation *in silico* tools against SARS-CoV-2.

Key words: antiviral peptides; IL-6 inducing peptides; machine learning; performance assessment; bioinformatics; SARS-CoV-2; coronavirus

Balachandran Manavalan is an Assistant Professor at the Department of Physiology, Ajou University School of Medicine, Republic of Korea. He is also an associate member of the Korea Institute for Advanced Study, Republic of Korea. His research interests include artificial intelligence, bioinformatics, machine learning, big data and functional genomics.
Shaherin Basith is a Research Assistant Professor in the Department of Physiology, Ajou University School of Medicine, Republic of Korea. Her research mainly focuses on exploring the structure–function relationships of proteins using state-of-the-art molecular modeling tools, phylogenetic analysis and biomolecular simulations. She is also actively involved in the application of machine learning tools for peptide and small-molecule drug discovery.
Gwang Lee is a Professor at the Department of Physiology, Ajou University School of Medicine, Republic of Korea. His main research interests include the integration of triple omics (transcriptomics, metabolomics and proteomics) for nanotoxicity studies, machine learning, neurodegenerative diseases and biomolecular simulations targeting therapeutically important proteins or enzymes.
Submitted: 1 July 2021; **Received (in revised form):** 27 August 2021

Introduction

The coronavirus disease 2019 (COVID-19) pandemic has impacted the society by affecting the global economy, health services, large- and small-scale industries, travel and tourism industries and so on. The first coronavirus in humans, identified in the 1960s, was zoonotic in origin [1, 2]. Few prevalent human coronaviruses, such as HCoV-229E, HCoV-OC43, HCoV-NL63 and HCoV-HKU1, cause mild upper respiratory symptoms such as common cold [1, 2]. The severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV), which emerged in 2002 and 2012, respectively, are more pathogenic to humans and cause life-threatening diseases in children, the elderly and immunocompromised individuals. Coronaviruses were named by virologists because of the crown-like appearance of spike proteins on their surface, similar to the sun's aura, known as corona. A team of Chinese researchers led by Xu Jianguo found that the mysterious virus identified in humans in December 2019 is a novel coronavirus similar to the known viruses collected from bats [3]. Following the discovery of SARS-CoV-2, its genetic sequence data were made available within weeks and it was found to be similar to SARS-CoV, but less fatal. The high incidence rate, rapid and easy transmission, and long incubation periods (often without symptoms) of SARS-CoV-2 hinder its detection, tracing and eradication. Although the virus is commonly referred as coronavirus globally, it was formally named as novel coronavirus (2019-nCoV) and was subsequently renamed as SARS-CoV-2 [4].

Like MERS-CoV and SARS-CoV, SARS-CoV-2 primarily infects the airways, causing symptoms ranging from mild respiratory infections to severe acute respiratory infections, with the latter leading to multiple organ dysfunction syndrome (MODS) and eventual death [5]. Less common symptoms include aches, conjunctivitis, diarrhoea, headache, loss of smell and taste, skin rashes and sore throat, whereas the most common symptoms are dry cough, fever and fatigue. Serious symptoms include dyspnea, chest pain or pressure, loss of speech or movement, systemic infection and pneumonia. The World Health Organization declared COVID-19 as a pandemic on 11 March 2020, pointing to over 199 million confirmed COVID-19 cases and 4.2 million related deaths (as of 04 August 2021) [6]. Equitable access to safe and effective vaccines is crucial for ending the COVID-19 pandemic [6]. As of 16 June 2021, around 22 COVID-19 vaccines are being tested and going into development. However, fears and concerns surrounding vaccination against SARS-CoV-2 are gripping the global population. Moreover, the long-lasting protective effect of these vaccines against SARS-CoV-2 remains unknown and may not help people who are already infected [7]. Hence, it is crucial to develop appropriate antiviral therapeutic agents against SARS-CoV-2 within a short period.

Currently, the emergence and re-emergence of viruses has imposed a serious threat to humans, as the relevant therapeutic options are quite limited and virus-specific. Few antiviral drugs show effects against a broad spectrum of viral pathogens; however, a few viral agents might still show resistance to antiviral drugs and produce adverse side effects, thereby promoting serious complications. To mitigate these limitations, the design and development of antiviral peptides (AVPs) with minimal or no side effects are necessary. AVPs serve as excellent therapeutic alternatives because of their role in preventing viral attachment and replication. The development of peptides as effective therapeutics against newly emerging viral pathogens, particularly SARS-CoV-2, holds great promise owing to their low molecular weight, safety, specificity, biocompatibility, low toxicity, fewer side effects, rapid elimination and efficacy [8, 9]. Clinical studies of anti-human immunodeficiency virus (HIV) peptide enfuvirtide (T20) and hepatitis-C peptide inhibitors, such as boceprevir and telaprevir, serve as successful peptide models, thus underscoring the importance of AVPs as safe and better alternatives for treating infectious diseases [10, 11]. Computational methods such as machine learning (ML) are instrumental in rapidly capturing meaningful data from ever-growing big biological data. Since 2012, several ML approaches have been developed for predicting AVPs, and they represent a promising field in the identification of novel and effective antiviral therapeutics. Several sequence-based ML methods for AVP prediction, including AVPpred [12], the methods described by Chang *et al.* [13] and Zare *et al.* [14], AntiVPP 1.0 [15], Firm-AVP [16], PEPred-Suite [17], iAMPpred [18], AMPfun [19], AMAP [20], MLAMP [21], PPTPP [22], AVPIden [23], PreAntiCoV [24], ENNAVIA [25] and Meta-iAVP [26], have been developed. Interestingly, a few of the above-mentioned methods have been specifically designed for predicting anti-coronavirus peptides (ACVPs), including PreAntiCoV, ENNAVIA and AVPIden.

The cytokine storm is an intriguing aspect of SARS-CoV-2 infection [27, 28]. Detection and modulation of proinflammatory cytokines are essential in the early stages of viral infection. Recently, proinflammatory cytokines have been found to play a significant role in COVID-19 progression due to the aberrant release of circulating cytokines [2, 29–31]. Abnormal levels of proinflammatory cytokines have been detected in severely affected patients with COVID-19 showing pulmonary inflammation, lung damage and MODS [32]. Previous studies have reported that high levels of proinflammatory cytokines, such as IL-1β, IL-6, IL-10, IFN-γ, IP10, TNF-α, CRP and MCP1, were identified in the serum samples of patients with COVID-19 [33–37]. Notably, IL-6 is considered one of the key players in viral cytokine storms. IL-6 is an important proinflammatory cytokine that regulates inflammation, acute phase responses, haematopoiesis, oncogenesis and specific immune responses [38–40]. Patients with SARS-CoV-2 infection show high levels of IL-6 and low levels of suppressor of cytokine signaling-3, which modulate the negative feedback loop of IL-6, resulting in severe pneumonia and acute respiratory distress syndrome, eventually leading to MODS and ultimately death [41–43]. Immune dysfunction plays a dominant role in COVID-19 and could serve as a possible therapeutic target. Hence, it is necessary to identify IL-6-inducing region or peptides to study their mechanism of action and to develop immunotherapeutic applications. Owing to the rapid transmission of SARS-CoV-2 and difficulties in experimental techniques (such as time consumption, cost effectiveness and labour-intensiveness), computational tools, such as ML, are required to identify IL-6-inducing peptides for rapid screening and validation using experimental techniques. Previously, several ML-based techniques, particularly cytokine-specific approaches, have been developed in the field of immunotherapeutics. IFNepitope was developed for predicting IFN-γ-inducing peptides [44], whereas IL4Pred [45], IL10Pred [46] and IL17eScan [47] were devised for predicting IL-4-, IL-10- and IL-17-inducing peptides, respectively. CytoPred is a method that predicts and further classifies the cytokine family and subfamily [48]. ProInflam [49] and PIP-EL [50] have been developed for predicting peptides that induce a group of proinflammatory cytokines, whereas AntiInflam [51] is a prediction method for identifying anti-inflammatory cytokines. Although several computational approaches have been developed for cytokine-specific methods, prediction of IL-6-inducing peptides using ML is still nascent. In the recent past, only two methods,

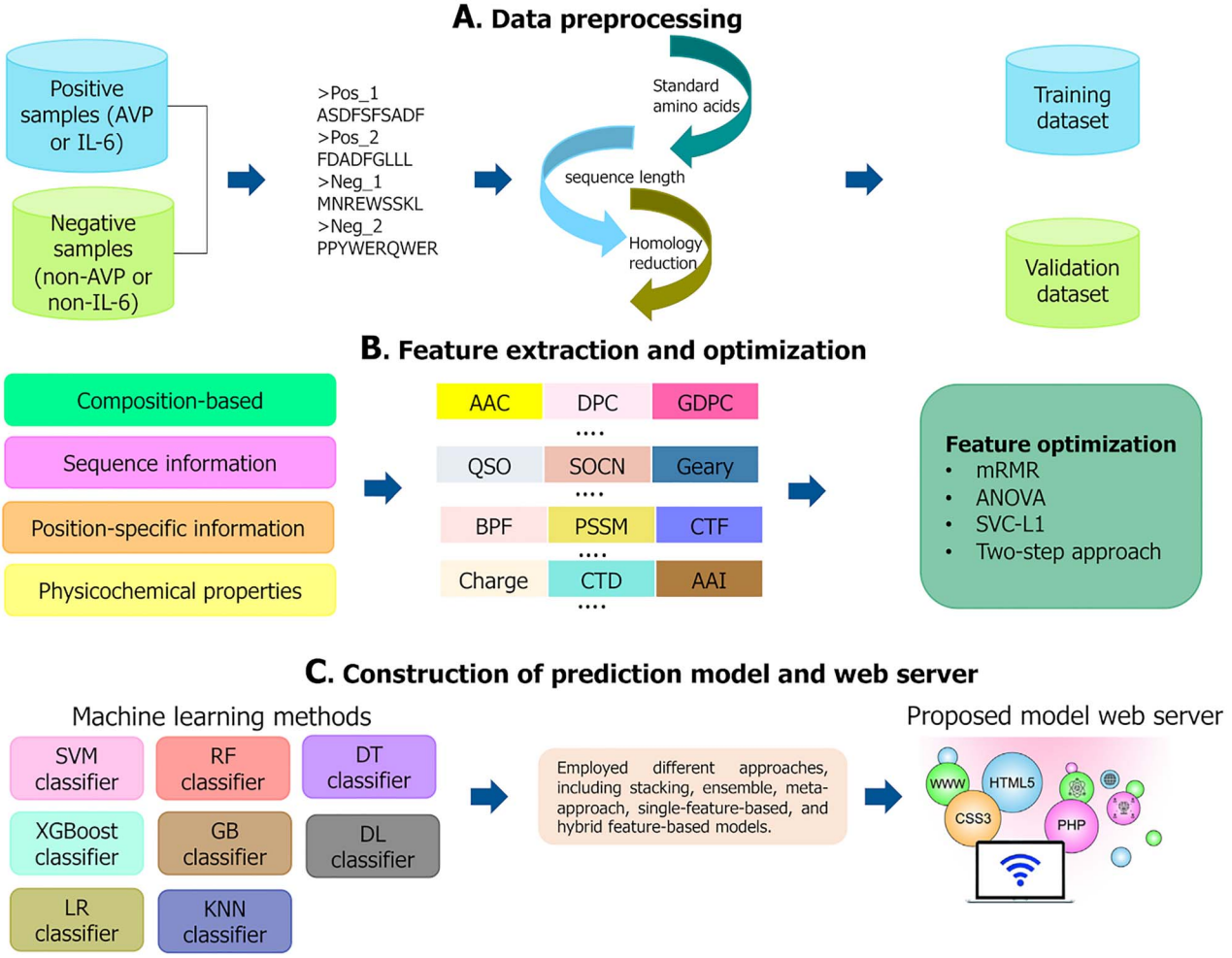

Figure 1. Overview of current computational frameworks for AVP or IL-6 prediction. (**A**) Dataset preprocessing. (**B**) Feature extraction and optimization. (C) Construction of the prediction model and web server.

namely IL-6Pred [52] and StackIL6 [53], have been developed for predicting IL-6-inducing peptides. A general overview of the present computational frameworks for AVPs/IL-6 is shown in Figure 1, and the proposed AVP and IL-6 predictor timelines are shown in Figure 2.

In this paper, we present a comparative analysis of the underlying framework of the state-of-the-art computational methods for anti-coronavirus therapies in terms of the datasets used along with the feature encoding methods, ML algorithms, cross-validation (CV) methods and prediction performances. Importantly, we provide general guidelines for developing robust AVP predictors that show activity against SARS-CoV (hereafter simply referred as AVPs) and IL-6-inducing peptide models to overcome some of the inherent drawbacks associated with the existing models. Notably, we constructed rationalized independent datasets to critically assess the robustness and scalability of existing ML approaches. Furthermore, we discussed the limitations of existing approaches and future viewpoints for improving the current and forthcoming ML tools. We anticipate that this review will contribute to the growth and expansion of this field, assist researchers in the selection of appropriate tools and accelerate the development of peptide-based therapeutics against SARS-CoV-2.

Materials and methods

Construction of an independent test data set

Construction of an independent test dataset is essential to objectively evaluate the predictive performance of state-of-the-art ML tools. In this review, we evaluated prediction methods for IL-6-inducing peptides and AVPs using two separate test datasets as follows.

ACVP and non-AVP dataset construction

Notably, we considered AVPs that showed activity only against coronavirus as positive samples. The detailed procedure for constructing positive samples is as follows: initially, we extracted 185 entries from the DBAASP database [54], which comprises experimentally validated peptides that show activity against coronaviruses. Subsequently, we excluded entries containing N- or C-terminal modifications, non-standard amino acids and multimeric peptides. This filtering process resulted in 98 peptides. Next, we collected peptides through extensive literature searches and identified 87 peptides [55–57]. Finally, a total of 125 unique ACVPs were obtained through the DBAASP and literature searches. In general, if any peptide is present in the existing

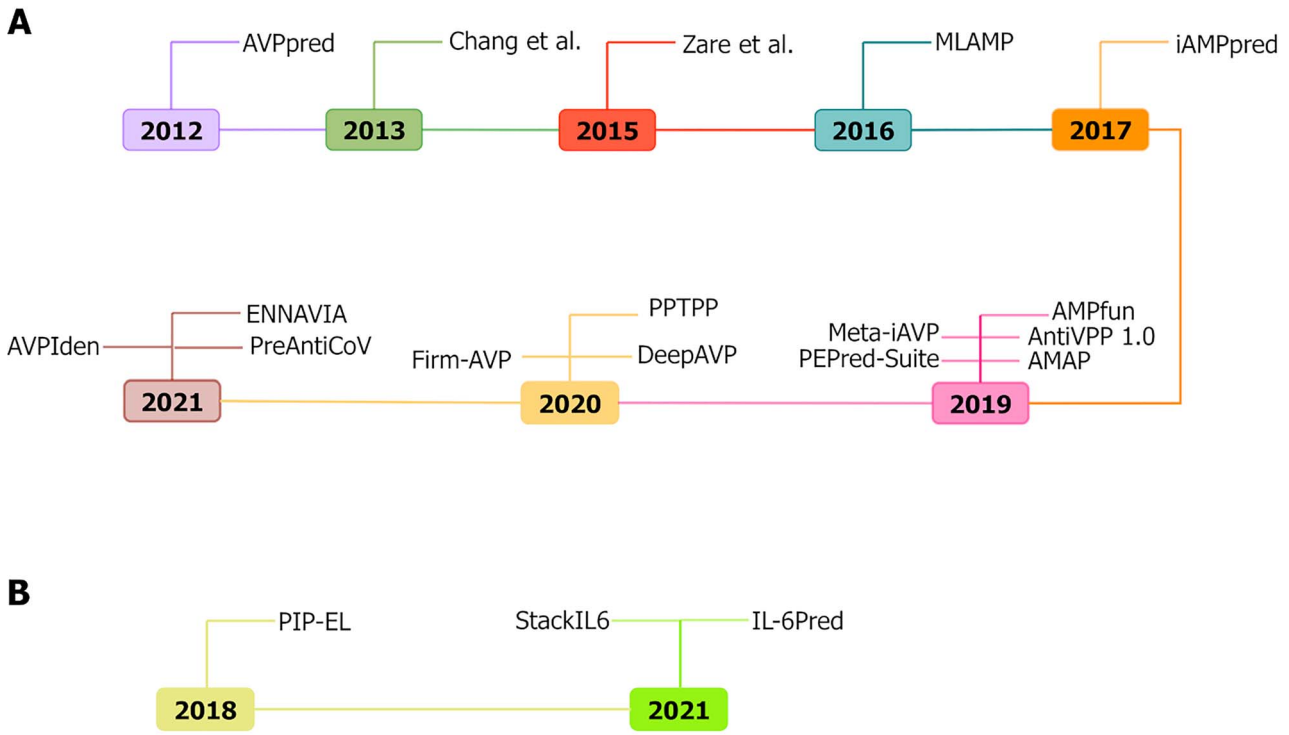

Figure 2. Timeline of the existing computational approaches for AVP prediction (**A**) and IL-6 peptide prediction (**B**).

method training dataset, the generated prediction model will correctly classify the peptide during the evaluation stage. Therefore, to avoid potential bias, we excluded overlapping peptides that shared 100% sequence identity with the existing method training samples. This explains the differences observed in the ACVP dataset sizes for each method during the evaluation.

Generally, a novel peptide with AVP activity can be identified from antimicrobial peptides (AMPs), random peptides (RP) and non-AMPs. Therefore, we carefully constructed non-AVPs by mimicking the real scenario as follows: (i) we considered RPs reported by Wei *et al.* [17, 58] where the authors generated RPs by re-arranging the residual positions of therapeutic peptides. Notably, none of the RPs shared 80% sequence identity with other therapeutic peptides; (ii) recently, Xu *et al.* [59] retrieved the peptide sequences from UniProt and excluded the peptides presenting all possible antimicrobial activities (e.g. anticancer, AVP, antibacterial and so on) and non-standard amino acids. Later, using CD-HIT [60, 61], they excluded peptides that shared >40% sequence identity with AMPs, and the remaining samples were treated as non-AMPs. The same dataset was used for this evaluation. (iii) AMPs other than AVPs [herein referred to as other functional peptides (OFPs)] were extracted from different databases and extensive literature searches that resulted in more than 10,000 peptides. Subsequently, OFPs that shared >80% sequence identity with AVPs were excluded. Finally, we combined RPs, non-AMPs and OFPs, excluded the samples that shared >80% sequence identity with the existing training data and obtained 2341 non-AVPs (negative samples). Of these, 897, 942 and 502 peptides were RPs, non-AMPs and OFPs, respectively.

IL-6 and non-IL-6-inducing peptide dataset construction

We extracted 434 IL-6-inducing linear peptides and 194 non-IL-6-inducing peptides from the IEDB database [62], which were experimentally verified in human and mouse species using any of three different assays (T-cell, B-cell and MHC ligand). Subsequently, we excluded peptides with lengths greater than 25 amino acids and peptides overlapping with the existing training dataset, which resulted in a total of 101 IL-6-inducing peptides. Similar to the non-AVP samples, we mimicked the actual scenario for non-IL-6-inducing peptides. To construct non-IL-6 dataset, we first considered only RPs and OFPs from the above-mentioned dataset. Non-AMPs were excluded because most sequences had a length greater than 25 amino acids. Later, we constructed experimentally validated non-IL-6-inducing peptide sequences (EVS) and peptides that induce proinflammatory cytokines other than IL-6 (OS). For OS, we extracted experimentally characterized peptides that induce other cytokines (IL-1α, IL-1β, TNF-α, IL-8, IL-12, IL-17 and IL-18) tested in human and mouse species. Finally, we obtained more than 12,000 peptide sequences. After redundancy reduction (CD-HIT of 0.7) with the existing method training dataset, we obtained ~1985 sequences. Of these, RP, OFP, EVS and OS contributed 205, 377, 46 and 1,357 peptides, respectively. A statistical summary of length distribution and amino acid composition (ACC) in the AVPs and IL-6 datasets is shown in Figure 3.

State-of-the-art computational predictors for AVPs

In 2012, Thakur *et al.* [12] developed the first predictor AVPpred. They constructed two training datasets (T544p + 407n and T544p + 544n) and two validation datasets (V60p + 45n and V60p + V60n). In their datasets, positive samples were similar between the two training datasets and two validation datasets. However, the first portion of the training and validation sets (T544p + 407n and V60p + 45n) contained experimentally verified non-effective AVPs as negative samples. Similarly, the

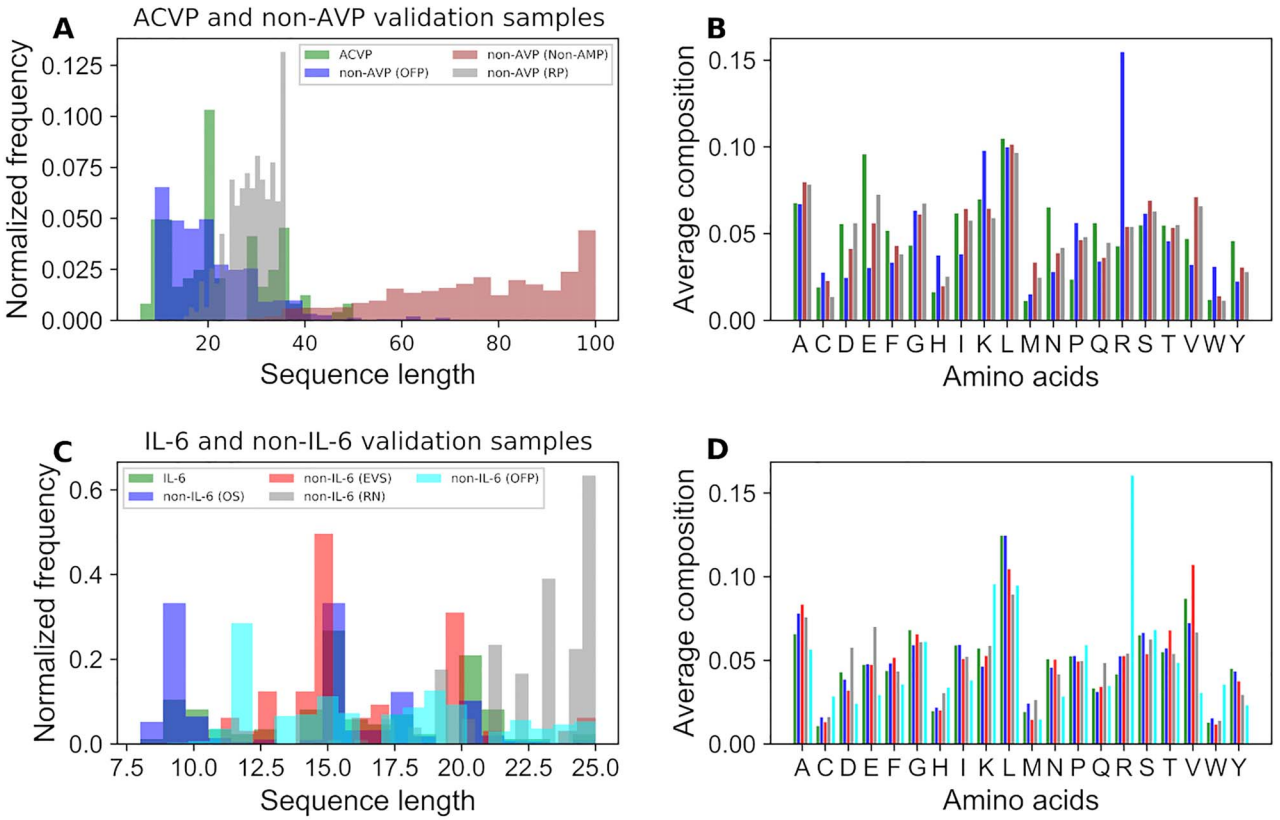

Figure 3. Length distribution and AAC of the constructed validation dataset. Length distribution of ACVP/non-AVP samples and IL-6/non-IL-6 samples is shown in (**A**) and (**C**). AACs of ACVP/non-AVP samples and IL-6/non-IL-6 samples are shown in (**B**) and (**D**). For the sake of clarity, negative samples were partitioned according to their function.

second portion (T544p + 544n and V60p + 60n) of the negative sets included RPs. Several computational predictors have been developed using these datasets, including AntiVPP 1.0 [15], DeepAVP [63], the methods described by Chang *et al.* [13] and Zare *et al.* [64], iAMPpred [65], PEPred-Suite [17], Meta-iAVP [26], FIRM-AVP [66], PPTPP [22] and ENNAVIA [67]. However, only a few methods have been developed using their own datasets, including AMPfun [19], AMAP [20], iAMPpred, PreAntiCoV, AVPIden and MLAMP [21]. In addition to antiviral prediction, almost half of the methods (iAMPpred, AMAP, AMPfun, PEPred-Suite, PPTPP, ENNAVIA and MLAMP) can predict other peptide functional activities. A comprehensive survey of most of these methods, including benchmark datasets, feature encoding schemes, ML classifiers, and training and validation performances, has been presented elsewhere [68]. Table 1 summarizes the current AVP computational approaches according to the tool name, employed ML algorithm, training and independent dataset size, utilized feature encodings, performance evaluation strategy, feature selection methods, reported accuracy (ACC), web server availability and additional function prediction. The following sections highlight the four most recent AVP prediction methods that have included ACVPs in their analysis: PreAntiCoV, AVPIden, AMPfun and ENNAVIA.

PreAntiCoV

Pang *et al.* [24] proposed the first AVP predictor specifically designed to identify ACVPs based on sequence information. First, they constructed a positive sample containing 137 unique ACVPs. Of these, 99 sequences were experimentally verified and were retrieved from the AVPdb database; the remaining 38 sequences were putative functional peptides. Second, they selected 70% of the positive samples (95 peptides) to develop four prediction models with the same positive samples but different negative sets, including 1399 AVPs without anti-coronavirus activity (AVPs_Wo_ACVPs), 3746 peptides with various antimicrobial functions other than antiviral activity (non-AVPs), 3485 RPs without antimicrobial activity (non-AMPs) and a combination of AVPs_Wo_ACVPs + non-AVPs + non-AMPs (8535 peptides). The rationale for constructing four models is to understand how well a classifier can discriminate between positive and negative samples.

Notably, all four models were trained in a similar manner using different optimal feature sets. The authors employed five different feature-encoding schemes, including AAC, dipeptide composition (DPC), pseudo AAC (PseAAC), physicochemical properties (PCPs) and the composition of *k*-space amino acid group pairs (CKSAAGP), to encode diverse peptide sizes into a fixed length of features. All these features were linearly combined to obtain a 527-dimensional feature vector. The optimal feature set was then identified using the Wilcoxon rank-sum test. Specifically, the optimal feature set differed among the four datasets. As the dataset was massively imbalanced between the positive and negative samples, they employed two approaches (NearMiss undersampling technique and balanced RF techniques) using the optimal feature set, trained the model using 5-fold CV and later selected the best model with the highest sensitivity (Sn). Subsequently, the resultant models using the two approaches were compared, and the best model was selected.

Table 1. A comprehensive summary of the reviewed approaches for AVP prediction

Tool, Year	ML algorithm	Dataset size (No. of positive/negative samples)	Feature encoding	Evaluation strategy	Feature selection	Accuracy in % CV/independent assessment (IA)	Webserver/ Software availability	Additional function prediction
AVPpred [12], 2012	Support vector machine (SVM)	Two training and two validation/in-dependent datasets were utilized. Training dataset 1 (TD 1): 544/407 Independent dataset 1 (ID 1): 60/45 TD 2: 544/544 ID 2: 60/60	Sequence motifs, sequence alignment, amino acid composition (AAC) and physicochemical properties (PCPs)	5-Fold cross validation (CV) and independent data set	Not available (N/A)	TD 1: 85.0 (CV) TD 2: 90.0 (CV) ID 1: 85.7 (IA) ID 2: 92.5 (IA)	N	N
Chang *et al.* [13], 2013	Random forest (RF)	AVPpred datasets (TD1, ID1, TD2 and ID2)	AAC, PCP, aggregation and secondary structure	10-Fold CV and independent dataset	minimum Redundancy Maximum Relevance (mRMR)	TD 1: 84.2 (CV) TD 2: 91.1 (CV) ID 1: 86.7 (IA) ID 2: 93.3 (IA)	N	N
Zare *et al.* [14], 2015	Adaboost	Train: 342/312	Pseudo ACC (PSeAAC)	5-Fold CV	N/A	93.26 (CV)	N	N
AntiVPP 1.0 [15], 2019	RF	AVPpred datasets (TD2 and ID2)	PCPs, relative frequency of all 20 natural amino acids and residues composition of peptides	Independent dataset	N/A	TD2: 99.3 (CV) ID2: 93.0 (IA)	Y	N
PEPred-Suite [17], 2019	RF	AVPpred datasets (TD1 and ID1) AD1: 60/2000 (positive/negative)	10 Commonly used feature encoding algorithms such as AAC, dipeptide composition (DPC), G-gap DPC (GGAP), ASDC, Composition-Transition-Distribution (CTD), Twenty-Bit features (BIT20), Twenty-One-Bit features (BIT21), Over-Lapping Property features, Information Theory features and PCPs	10-Fold CV, main and alternative independent datasets	mRMR	TD1: 86.4 (CV) ID1 and AD1: N/A	Y	Y
Meta-iAVP [26], 2019	RF, SVM, *k*-nearest neighbor (*k*-NN), recursive partitioning and regression trees (rpart), generalized linear model (glm) and eXtreme gradient boosting (XGBoost)	AVPpred datasets (TD1, ID1, TD2 and ID2)	AAC, PseAAC, amphiphilic (Am)-PseAAC, DPC and g-gap DPC (GDC)	5-Fold CV and independent dataset	N/A	TD 1: 88.17 (CV) TD 2: 92.31 CV) ID 1: 95.19 (IA) ID 2: 94.92 (IA)	Y	N

Continued

Table 1. Continued

Tool, Year	ML algorithm	Dataset size (No. of positive/negative samples)	Feature encoding	Evaluation strategy	Feature selection	Accuracy in % CV/independent assessment (IA)	Webserver/ Software availability	Additional function prediction
FIRM-AVP [16], 2020	SVM	AVPpred datasets (TD1 and ID1)	AAC, DPC, PseAAC, Am-PseAAC, CTD and secondary structure sequence	10-Fold CV and independent dataset	Recursive feature elimination (RFE)	TD1: N/A ID1: 92.4 (IA)	Y	N
AMPfun [19], 2019	RF	TD: 1400/2451 ID: 601/1374	Binary profiling of amino acid position, AAC and PCPs	10-Fold CV and independent dataset	Forward feature selection algorithm	TD: 92.47 (CV) ID: 86.13 (IA)	Y	Y
AMAP [20], 2019	SVM and XGBoost	TD: 180/5156 (Multiclass predictor)	AAC, 3-mer composition	Leave-one-cluster-out (LOCO) CV, 5-fold CV, independent dataset	N/A	N/A	Y	Y
MLAMP [21], 2016	RF	TD: 879/2405 (Multiclass predictor)	PseAAC	Jackknife CV, independent dataset	mRMR	TD: 89.9 (CV) ID: 94.7 (IA)	Y	Y
iAMPpred [18], 2017	SVM	TD: 738/738	compositional (AAC, PseAAC and normalized ACC-NAAC), secondary structure and PCPs	10-Fold CV, independent dataset	Fast correlation-based feature selection (FCBF)	TD: 90.09 (CV)	Y	Y
PPTPP [22], 2020	RF	AVPpred and PEPred-Suite datasets (TD1, ID1 and AD1)	PCPs	10-Fold CV, independent dataset	IPP and maximum-relevance-maximum-distance (MRMD)	N/A	Y	Y
PreAntiCoV [24], 2021	RF	Stage1: TD1: 5145/3485 ID1: 2208/1494 Stage2: TD2: 95/5050 ID2: 42/2166	AAC, DPC, PseAAC, PCPs and CKSAAGP	5-Fold CV and independent dataset	Wilcoxon rank-sum test	TD1: 91.33 ID1: N/A TD2: N/A ID2: N/A	Y	N
ENNAVIA [25], 2021	Neural network	ENNAVIA-A: AVPpred dataset (TD1 and ID1) ENNAVIA-B: AVPpred dataset (TD2 and ID2) ENNAVIA-C: 109/407 ENNAVIA-D: 109/544	AAC, DPC, TPC, g-gap DPC, g-gap TPC, conjoint triad, CTD, PseAAC, PCP and AAI	10-Fold CV, independent dataset	N/A	ENNAVIA-A: 91.24 (CV) and 93.88 (IA) ENNAVIA-B: 95.9 (CV) and 95.65 (IA) ENNAVIA-C: 94.95 (CV) ENNAVIA-D: 97.29 (CV)	Y	N
AVPIden [23], 2021	RF	16 datasets. Please refer Table 6 of [23]	AAC, DPC, CKSAAGP, PseAAC and PCPs	4-Fold CV, independent dataset	Shapley value	NA	Y	Y

PreAntiCoV was developed as a two-stage approach based on the observations for four prediction model performances. The first stage identifies whether a given peptide belongs to an AMP or a non-AMP, whereas the second stage identifies ACVPs from the predicted AMP. To execute the two-stage approach, two models were trained using entirely different datasets: (i) non-AMPs were considered negative samples, and peptides belonging to ACVPs, AVPs_Wo_ACVPs and non-AVP were considered positive samples (AMPs). (ii) ACVPs were considered positive samples and two datasets (AVPs_Wo_ACVPs and non-AVP) were considered negative samples. PreAntiCoV achieved Matthews Correlation Coefficient (MCC) and AUC values of 0.822 and 0.972, respectively, during an independent test on stage 1. Similarly, Sn, specificity (Sp) and MCC of 0.738, 0.855 and 0.223, respectively, were achieved during stage 2 independent tests. The standalone version of PreAntiCoV is publicly available at https://github.com/poncey/PreAntiCoV.

The group that developed PreAntiCoV proposed a novel scheme for identifying AVPs using the ML approach, namely AVPIden [23]. In this method, the authors proposed a double-stage ML prediction scheme for predicting AVPs and characterizing their functional activities against different viruses in parallel at the levels of both family (Coronaviridae, Retroviridae, Herpesviridae, Paramyxoviridae, Orthomyxoviridae and Flaviviridae) and species (feline immunodeficiency virus, HIV), hepatitis C virus, human parainfluenza virus type 3 (HPIV3), herpes simplex virus type 1 (HSV1), influenza A virus (INFVA), respiratory syncytial virus and SARS-CoV. Initially, AVP sequences were retrieved from various databases such as AVPdb [69], dbAMP [70], DBAASP [54], DRAMP [71] and HIPdb [72]. The positive dataset comprised 2662 sequences targeting different viruses. The negative dataset comprised both non-AMPs and non-AVPs (AMPs without antiviral activities) with 5116 and 4979 sequences, respectively. The sequences were vectorized using peptide descriptors, including AAC, DPC, CKSAAGP, PseAAC and physicochemical features, which were further used to construct a double-stage classifier. In the first-stage classifier, AVPs are distinguished from non-AVPs and then fed into the second classifier for functional characterization of several viruses. To overcome the limitation of class imbalance, an imbalanced learning strategy was implemented to improve prediction performance. Furthermore, the Shapley value explanation was employed to identify critical peptide features for antiviral functions. In the first-stage classifier, two models were developed with the same positive samples and different negative samples (non-AMP _only and non-AVPs_included). In the non-AMP only dataset, RF achieved the best performance with an ACC of 98.58 ± 0.18, whereas the RF classifier constructed with another negative set achieved an ACC of 94.44 ± 0.26. In the second-stage classifier, performance metrics at the family level with balanced RF were above 80% for all viral families except Orthomyxoviridae Sn metric. Similarly, performance metrics at the species level, except for the Sn metrics of INFVA and SARS-CoV, were above 80% for all viral species. A user-friendly web interface is publicly accessible at https://awi.cuhk.edu.cn/AVPIden/#/.

ENNAVIA

Timmons *et al.* [25] proposed a novel predictor, ENNAVIA. It contains four prediction models (ENNAVIA-A, ENNAVIA-B, ENNAVIA-C and ENNAVIA-D), of which two (ENNAVIA-A and ENNAVIA-B) predict AVPs and the remaining two (ENNAVIA-C and ENNAVIA-D) predict ACVPs. Here, we briefly discuss the implementation protocol of ENNAVIA. First, the authors employed AVPpred datasets (Table 1) and named them as ENNAVIA_A and ENNAVIA_B. Second, they extracted 137 ACVPs from PreAntiCoV and further considered them as positive samples for ENNAVIA-C and ENNAVIA-D. The negative samples for ENNAVIA-C and ENNAVIA-D were the same as those used in ENNAVIA-A and ENNAVIA-B. Notably, they excluded peptides with <7 and >40 residue lengths and only considered peptides with a range of 7–40 amino acid residues for each dataset. Using these four datasets, they explored different feature encoding schemes, including AAC, DPC, tripeptide composition (TPC), g-gap DPC (with g set to 1–3), g-gap TPC (with g set to 3 and 4), conjoint triad, composition transition and distribution (CTD), PAAC, PCP and amino acid index (AAI). All these features were combined linearly and trained using a neural network classifier for the final prediction. ENNAVIA-A and ENNAVIA-B achieved ACCs of 0.913 and 0.959, respectively. ENNAVIA-C achieved ACC, Sn, Sp and MCC of 0.950, 0.916, 0.960 and 0.87, respectively. The corresponding performances for the ENNAVIA-D dataset were 0.973, 0.898, 0.988 and 0.910, respectively. This webserver is publicly accessible at https://research.timmons.eu/ennavia.

AMPfun

AMPfun is a two-stage ML framework designed to identify AMPs and their functional activities [19]. Three steps were involved in each stage of its implementation: feature calculation, feature selection and application of ML algorithms. A sequential forward selection algorithm was utilized for feature selection. Training, testing and independent datasets were constructed based on seven AMP functional activities, including antiparasitic, antiviral, anticancer, targeting mammals, antifungal, targeting gram-positive bacteria and targeting gram-negative bacteria. The number of sequences for antiviral functional activity was collected from APD3, ADAM and AVPdb databases, with the following statistics: training set, positive (1400) and negative (2451) samples; testing set, positive (601) and negative (1374) samples; and independent testing set, positive (601) and negative (1374) samples. Features were divided into three types: binary profiling of amino acid position [n-gram binary profiling of position as determined by counting (NCB), n-gram binary profiling of position as determined by t-test (NTB) and motif-based binary profiling of position (MB)], AAC and physical–chemical properties (PSeAAC and CTD). For antiviral activity prediction, 4075 features were used. Three ML classifiers, including RF, SVM and decision tree (DT), were assessed based on 10-fold CV, and the RF classifier was selected as the best ML prediction method. A sequential forward selection algorithm was used to extract the key informative features linked with the functional activities of AMPs. In total, 1130 features [NTC, n-gram composition determined by t-test (918) + AAC (20) + MB (186) + NCB (6)] were selected for antiviral activity prediction. The performance of the RF classifier in terms of AUC with the selected features using 10-fold CV for antiviral activity predictions was 0.9692 and 0.9404 for the training and testing datasets, respectively. Similarly, the independent testing results showed an AUC of 0.940 with respect to the antiviral model. The proposed framework was implemented as a web server that is publicly available at http://fdblab.csie.ncu.edu.tw/AMPfun/index.html.

State-of-the-art computational methods for IL-6-inducing peptide prediction

Focusing only on IL-6-inducing peptides, two prediction models have been reported recently, namely IL-6Pred [52] and StackIL6

[53]. A brief description of these two methodologies is provided below.

IL-6Pred

Dhall *et al.* [52] proposed the first IL-6 prediction method, IL-6Pred, in 2020. First, the authors constructed benchmarking datasets using information collected from the IEDB database, where IL-6-inducing peptides (experimentally verified in human and mouse species) were considered as positive samples, and experimentally validated peptides inducing proinflammatory cytokines other than IL-6, such as IL-1α, IL-1β, TNF-α, IL-8, IL-12, IL-17 and IL-18, were considered as negative samples. Subsequently, they used 80% of the data (292 IL-6-inducing and 2393 non-IL-6-inducing peptides) to train the prediction model and the remaining data (73 IL-6 and 598 non-IL-6) for external validation. Using the training data, they extracted 15 different feature encodings (AAC, DPC, TPC, atomic and bond composition, residue repeat information, distance distribution of residue, Shannon entropy of protein, amino acids and PCP, conjoint triad calculation, CTD, PAAC, amphiphilic PAAC, quasi-sequence order and sequence order coupling number), integrated them in a linear fashion and obtained 9149 features. Subsequently, a feature selection technique called SVC-L1 was applied to identify 186 critical features, ∼2% of the original features. Finally, they employed six different classifiers and trained their respective models using 186 features. The DT, random forest (RF), logistic regression (LR), k-nearest neighbors, Gaussian naïve Bayes (GNB) and extreme gradient boosting (XGB) achieved ACCs of 84.47, 75.79, 77.13, 58.62, 85.48 and 86.29%, respectively, during 5-fold CV. The corresponding performances for independent validations were 84.20, 73.23, 75.26, 55.73, 84.50 and 84.65%, respectively. Notably, the performances of the three classifiers (DT, GNB and XGB) were similar regardless of the validation and were significantly better than the remaining classifiers. Although the reported metrics for IL-6Pred were excellent, we observed that most models were biased toward the larger class. Hence, it achieved a higher Sp and lower Sn. Furthermore, the authors reported six different prediction models with the top 10 features and slightly lower ACC than those mentioned above. All six models are available on their webserver at https://webs.iiitd.edu.in/raghava/il6pred/.

StackIL6

Phasit *et al.* proposed StackIL6 using the same IL-6Pred dataset but with different approaches. In StackIL6, the authors employed five different encodings, including AAC, DPC, CTD, gap DPC (CGAP) and PCPs. CGAP was used with five different parameters to retrieve five features. The CTD was split into three individual components to obtain three distinct features. Notably, AAC, DPC and CTD were already employed in IL-6Pred, whereas CGAP and PCP were applied for the first time in IL-6-inducing peptide prediction. Moreover, they used an undersampling approach to handle the imbalanced training dataset. Specifically, 10 different training datasets were generated with an equal number of the same positive but different negative samples. For each balanced dataset, 60 baseline models were generated (12 feature descriptors × 5 ML classifiers). ML classifiers included an extra tree, LR, multilayer perceptron, RF and SVM. In total, 600 baseline models were generated, and nine essential baseline models were identified using the GA-SAR algorithm, whose predicted probability values were trained with RF for the final prediction. StackIL6 achieved balanced accuracy (BACC), Sn, Sp and MCC of 0.942, 0.887, 0.997 and 0.890, respectively, during training. The corresponding metrics for the independent validations were 0.795, 0.849, 0.741 and 0.393, respectively. The reported metrics were significantly better than those of IL-6Pred. However, we observed that StackIL6 could not reproduce the training performance in independent validation (MCC 0.890 versus MCC 0.393), thus questioning its robustness. This webserver is publicly accessible at http://camt.pythonanywhere.com/StackIL6.

Performance evaluation

We applied five commonly used metrics to evaluate model performance comprehensively [73–75], including Sn, Sp, ACC, BACC and MCC. The definition of each metric is as follows:

$$\begin{cases} \mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{BACC} = \frac{\mathrm{Sn}+\mathrm{Sp}}{2} \\ \mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})\times(\mathrm{TP}+\mathrm{FN})\times(\mathrm{TN}+\mathrm{FP})\times(\mathrm{TN}+\mathrm{FN})}} \end{cases},$$

where TP, FP, TN and FN denote the number of true positives, false positives, true negatives and false negatives, respectively.

Results and discussion

Performance evaluation of AVP predictors based on an independent validation dataset

We considered only publicly accessible tools for comparative assessment of existing ML-based AVP methods. These tools included PEPred-Suite, Meta-iAVP, FIRM-AVP, ENNAVIA, AMPfun, AVPIden and iAMPpred. Notably, the ENNAVIA method includes four prediction models (ENNAVIA-A, ENNAVIA-B, ENNAVIA-C and ENNAVIA-D), and all of these were considered for evaluation. We also included the PreAntiCoV standalone version in the comparative evaluation owing to its earliest development. However, we faced difficulties in installing the program; hence, the values reported for PreAntiCOV were based on personal communication (Yuxuan Pang, The Chinese University of Hong Kong). Notably, most of the training datasets for existing AVP methods include only a small portion of ACVPs while developing their respective models. Therefore, it would be interesting to determine whether a general AVP method can predict ACVPs.

As mentioned in Methods section, we excluded validation samples that overlapped with the existing methods. Thus, the number of positive samples varied among the methods, but the number of negative samples remained constant for most methods (except AMPfun). Furthermore, the ENNAVIA method can only handle sequences with 7–40 amino acid residues. Notably, none of the sequences in our validation set was <7 residues, and several sequences were >40 residues. Therefore, in this case, we considered only the first 40 residues to evaluate the ENNAVIA prediction models. During our evaluation, each tool parameter (predicted probability value cut-off) was set to the corresponding recommended configurations or to the default cut-off if no recommendations were provided.

As the dataset was imbalanced, we ranked the AVP prediction methods according to MCC, as recommended [76]. Table 2 shows that ENNAVIA-D achieved the best performance, with an MCC, ACC, Sn, Sp and BACC of 0.350, 0896, 0.639, 0.908 and 0.773, respectively. Surprisingly, the second-best method, AMPfun, showed the same BACC as ENNAVIA-D. However, the MCC was

Table 2. Performance of various methods with the ACVP/non-AVP validation dataset

S. No	Methods (positive samples)	MCC	ACC	Sn	Sp	BACC
1	ENNAVIA-D	0.350	0.896	0.639	0.908	0.773
2	AMPfun	0.236	0.768	0.779	0.767	0.773
3	ENNAVIA-B	0.219	0.864	0.481	0.882	0.682
4	AVPIden	0.107	0.611	0.800	0.607	0.704
5	PreAntiCoV	0.067	0.727	0.452	0.735	0.593
6	iAMPpred	0.004	0.656	0.339	0.670	0.505
7	ENNAVIA-C	−0.017	0.329	0.648	0.314	0.481
8	Meta-iAVP	−0.049	0.477	0.400	0.481	0.440
9	PEPred-Suite	−0.141	0.403	0.255	0.410	0.332
10	ENNAVIA-A	−0.183	0.343	0.222	0.348	0.285
11	Firm-AVP	−0.220	0.354	0.118	0.365	0.241

~12.0% lower than that of ENNAVIA-D. ENNAVIA-B, the third-best method, achieved a good performance similar to AMPfun. The remaining eight methods (AVPIden, PreAntiCoV, iAMPpred, ENNAVIA-C, ENNAVIA-D, Meta-iAVP, PEPred-Suite and Firm-AVP) achieved an MCC of less than 0.11, indicating that the practical application of these methods is quite limited. The PreAntiCoV method was anticipated to perform better because it was designed explicitly for ACVP prediction using the two-step framework. Unfortunately, it did not perform well in our evaluation, incorrectly classifying both ACVPs and non-AVPs. AVPIden, the most recent method, employed a spectacular and novel approach for identifying AVPs activities. Notably, our reported value for AVPIden was based on AVPs functional activities. Unexpectedly, this method also failed in evaluation.

Notably, the top three methods included one method (ENNAVIA-D) specifically designed to predict ACVPs and two methods (AMPfun and ENNAVIA-B) that predict AVPs with different activities. Surprisingly, two general AVP prediction methods significantly outperformed PreAntiCoV, which was specifically designed to identify ACVPs. Generally, a program that reduces false positive rates can be useful to experimental researchers by saving time and costs. Considering this viewpoint, we included diverse negative samples in the validation dataset. Table S1, see Supplementary Data available online at http://bib.oxfordjournals.org/, shows that ENNAVIA-B and ENNAVIA-D consistently achieved superior performance (less false positive rate) regardless of a different dataset compared with AMPfun. We anticipated that AMPfun might perform better in reducing the false-positive rates, because it was trained using a vast training dataset (10-fold larger than ENNAVIA-D) and a two-stage approach. However, it ranked third, indicating that model robustness does not depend on the size of the training dataset.

Next, we compared the ENNAVIA-B and ENNAVIA-D methods in terms of data size and strategies. The training samples of ENNAVIA-B and ENNAVIA-D employed the same negative samples (non-secretory peptides), but different positive samples, where ENNAVIA-B and ENNAVIA-D employed AVP activity against different viruses and ACVPs, respectively. However, the model development strategy was similar between the two methods. Notably, the other three methods (PEPred-Suite, Firm-AVP and Meta-iAVP) employed in this evaluation used the same ENNAVIA-B training dataset. Unexpectedly, all these methods failed in our evaluation because of over-optimization or limited feature usage.

Comparison of validation samples with ENNAVIA-D training samples

To determine the reason for the superior performance of ENNAVIA-D, we used the two-sample logo and computed the statistically significant amino acid residues representing relative abundance in the sequence between positive ACVPs and negative non-AVPs. Notably, we combined the first 10 residues from the N-terminal and the last 10 from the C-terminal for ENNAVIA-D (Figure 4A) and the validation set (Figure 4B). We observed that 'N' at five positions (3rd, 7th, 8th, 10th and 12th), Y at the 18th position and F at the 19th position were enriched and shared a similar location between the validation set and ENNAVIA-D dataset. Similarly, 'P' at six positions (2nd, 5th, 12th, 15th, 17th and 18th) and 'R' at the 4th position was depleted and shared between the validation set and ENNAVIA-D datasets. Overall, we observed the position-specific amino acid information at several highly varied positions and found that only few positions were overlapped. For instance, 'E' was enriched at several validation dataset positions, but this residue was not enriched in the ENNAVIA-D training dataset at any position. Previously, we evaluated the 4mC site and 6mA site prediction methods using a standard validation dataset [77, 78] and observed that a stretch of nucleic acids shared a similar location between validation and training samples, leading to the superior performance of the method. In contrast to previous studies, we did not observe any similarity in motif characteristics between the validation and ENNAVIA-D datasets. Therefore, we cannot conclude that the superior performance of ENNAVIA-D is due to partial overlapping information.

The other possible reason could be the implementation of a novel computational approach. In ENNAVIA-D, the authors employed a small high-quality training dataset and generated a 6397-dimensional feature vector by integrating multiple feature encodings capturing the local and global information of the peptides. Subsequently, the NN was trained with a 6397-dimensional feature vector, and the final model was developed. Notably, the feature dimension employed in ENNAVIA-D was nine times larger than the size of the dataset. We anticipate that a high-quality training dataset and a vast feature dimension with a greater discriminative ability to capture the characteristics between ACVPs and non-AVPs resulted in the improved performance.

Performance evaluation of existing IL-6-inducing peptide predictors based on an independent dataset

We considered two publicly available IL-6-inducing peptide predictors (StackIL6 and IL-6Pred) and one general proinflammatory cytokine predictor (PIP-EL) for this evaluation. IL-6Pred contains five models (IL-6Pred_RF, IL-6Pred_DT, IL-6Pred_XGB, IL-6Pred_LR and IL-6Pred_GNB), and all models were used for the evaluation. Table 3 shows that IL-6Pred_XGB managed to achieve the top spot, with MCC, ACC, Sn, Sp and BACC of 0.094, 0.266, 0.950, 0.231 and 0.591, respectively. IL-6Pred_DT ranked second with the corresponding scores of 0.077, 0.420, 0.772, 0.403 and 0.587, respectively. In contrast, StackIL6 and PIP-EL showed poor performance, with MCC scores well below 0.025. Overall, it is evident from Table 3 that all the existing methods showed a mediocre performance, with MCC scores well below 0.1. It is worth mentioning that all predictors showed an excellent ability to identify IL-6 peptides but failed to predict non-IL-6 peptides correctly (Table S2, see Supplementary Data available online at

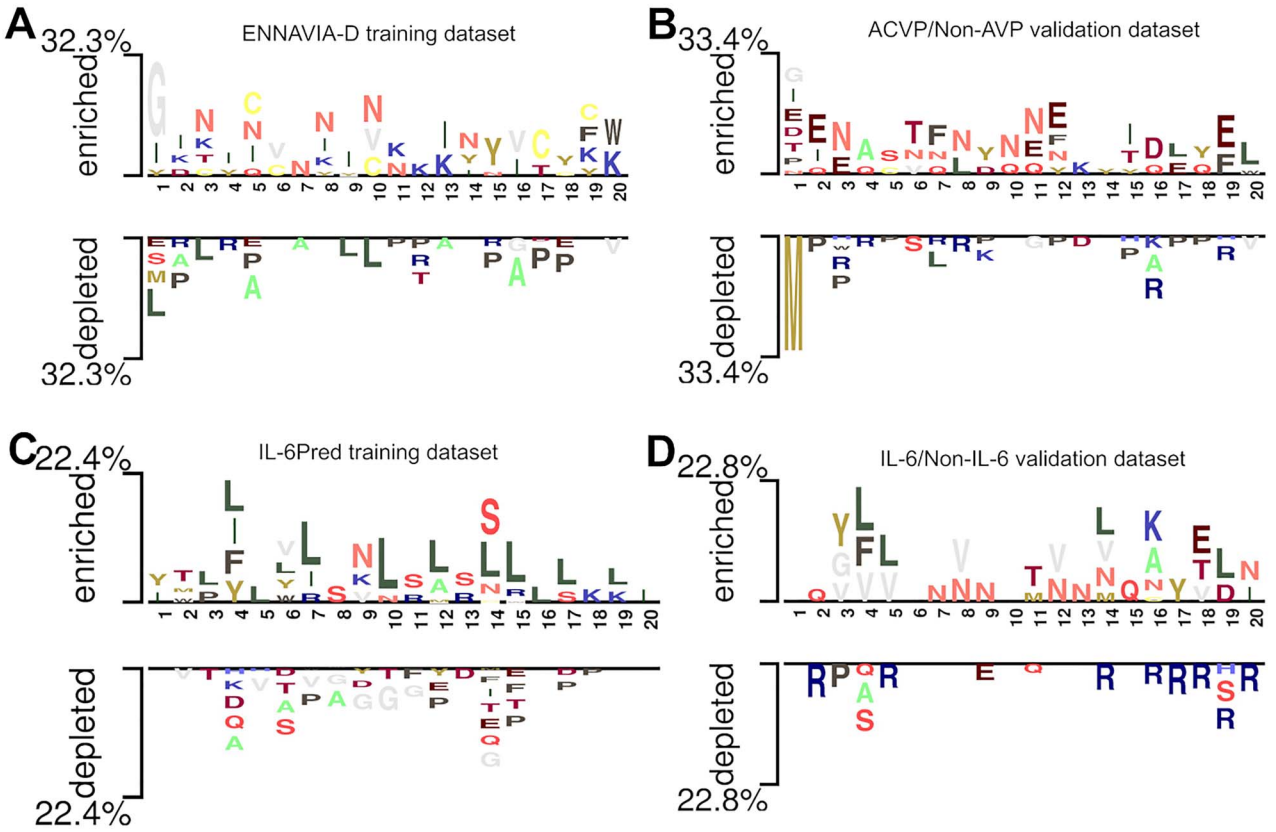

Figure 4. Statistically significant position-specific composition between positive and negative samples. **A** and **B**, respectively, represent the ENNAVIA-D training and validation datasets. **C** and **D**, respectively, represent the IL-6pred training and validation datasets.

Table 3. Performance of various methods with the IL-6/non-IL-6 validation dataset

S. No	Methods	MCC	ACC	Sn	Sp	BACC
1	IL-6Pred_XGB	0.094	0.266	0.950	0.231	0.591
2	IL-6Pred_DT	0.077	0.420	0.772	0.403	0.587
3	IL-6Pred_RF	0.072	0.250	0.921	0.216	0.568
4	StackIL6	0.023	0.212	0.861	0.179	0.520
5	IL-6Pred_LR	0.017	0.407	0.644	0.394	0.519
6	PIP-EL	0.011	0.154	0.901	0.116	0.508
7	IL-6Pred_GNB	−0.009	0.298	0.703	0.278	0.490

http://bib.oxfordjournals.org/). This kind of biased class prediction will mislead experimentalists when identifying putative IL-6-inducing sites, thereby limiting the practicality of these methods.

Furthermore, we compared the IL-6pred training dataset with our validation set in terms of position-specific amino acid positions to understand the problem present in the existing dataset. Figure 4C and D show that 'L' at four locations (4th, 5th, 14th and 19th), and 'N' at two positions (9th and 14th) was enriched at similar locations between the two datasets. However, none of the amino acids shared similar sites in non-IL-6-inducing peptides between the two datasets. This is not surprising because the existing method employed a negative dataset containing only partial proinflammatory cytokine information and excluded experimentally validated non-IL-6-inducing peptides. This indicates that all current approaches suffer a high false-positive rate when testing using diverse non-IL-6 samples. This analysis demonstrates that the negative samples of training datasets in the existing methods have a severe problem that needs to be addressed promptly.

Shortcomings of existing predictors and future perspectives for improving the prediction performance of therapeutic peptides against SARS-CoV

ENNAVIA-D showed the best performance with the correct classification of ACVPs from non-AVPs. Unfortunately, we encountered several limitations during our evaluation, which are as follows: (i) it can only handle sequences within 7–40 amino acid residues in length; (ii) the ENNAVIA-D web server cannot handle more than 100 sequences in a single run and sometimes returns an error which makes it difficult to use. In the case of IL-6 prediction methods, existing predictors showed underperformance during our validation, which is directly related to issues in the training dataset. Thus, the existing methods are not suitable for mapping IL-6-inducing peptide regions from different SARS-CoV-2 proteins.

Addressing the shortcomings of existing methods and developing novel computational tools with increased robustness and practical applicability is a rather challenging task. However,

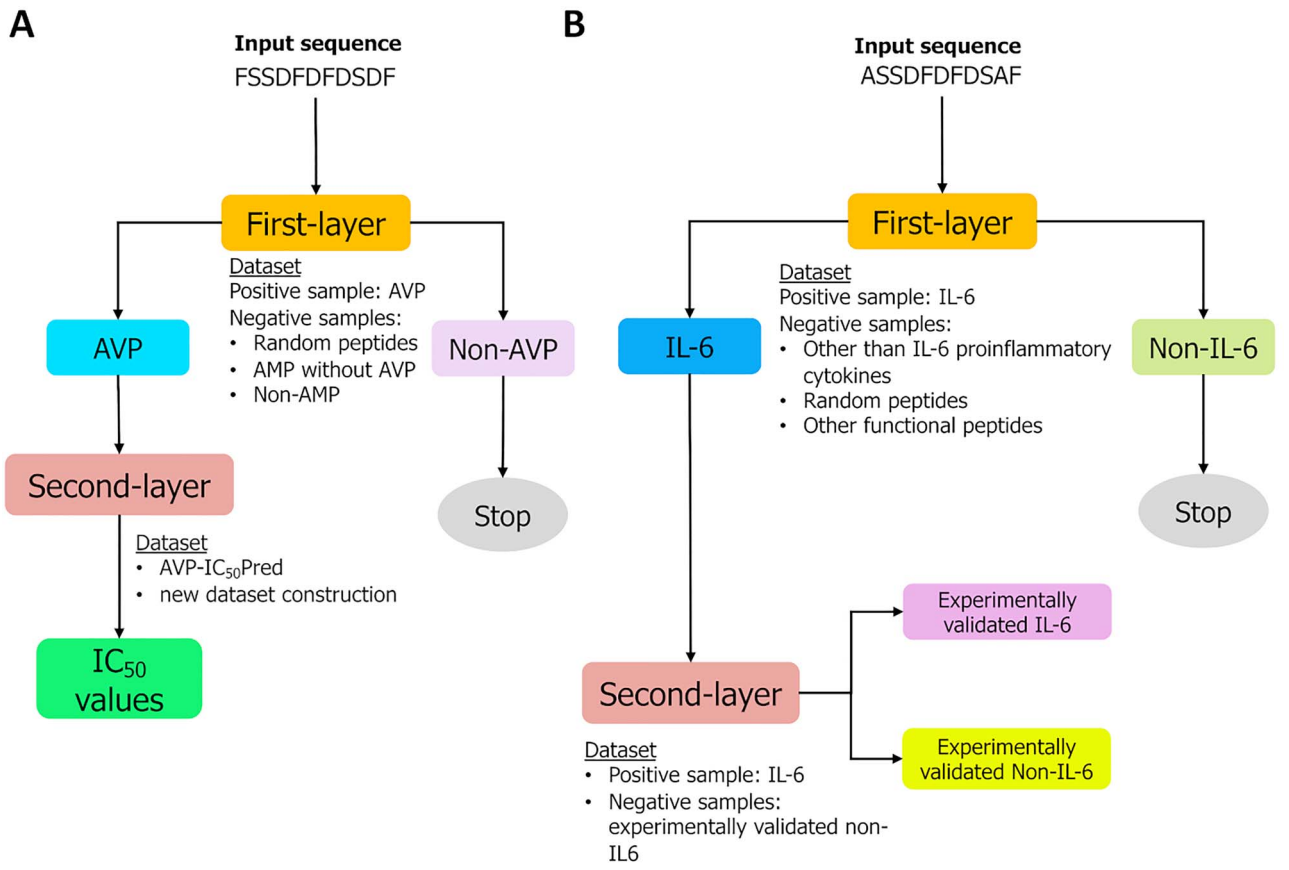

Figure 5. A feasible computational framework designed for the future development of (**A**) AVP and (**B**) IL-6-inducing peptide predictors.

we provide directions that may be helpful to researchers while creating a new model. Several models (Meta-iAVP, PTPP and ENNAVIA) have recently utilized the same AVPpred dataset constructed in 2012 [12]. Since 2012, AMP databases, including DBAASB [54], APD [79] and DRAMP [71], have been regularly updated, resulting in several hundreds of AVPs. Initially, these data should be pre-processed to construct non-redundant positive samples. While creating negative samples, researchers should consider peptides from different sources, including non-secretory peptides, RPs and OFPs. This tedious data collection and pre-processing will facilitate the development of a golden standard benchmarking dataset. Second, the development of a prediction model by utilizing high-quality training data is essential to identify whether a given peptide is an AVP or a non-AVP. Apart from merely classifying AVPs and non-AVPs, additional information may attract experimental researchers. Recently, Qureshi *et al.* [80] reported AVP-IC50Pred, which predicts AVP activity at the half-maximal inhibitory concentration (IC50). Thus, a new computational method that focuses on integrating these two predictors, i.e. AVP (classification) and IC50 (regression), into a single framework (Figure 5A) using different approaches would add more value to AVP research.

Unfortunately, none of the IL-6-inducing peptide prediction methods matched the practical applicability. The dataset used for model building has the following major problems. (i) When considering specific proinflammatory cytokines (such as IL-6 or IL-10 inducing peptides), both experimentally verified positive and negative samples were available in IEDB [62, 81]. However, the existing methods employed only specific cytokines (IL-6) as positive samples and considered a small number of cytokines as negative samples. In future studies, we recommend considering all possible cytokines (other than IL-6) as negative samples along with the experimentally validated non-IL-6 samples. (ii) Generally, negative samples are much larger than positive samples. Therefore, the class imbalance problem should be solved using different approaches, as reported previously [24, 50, 82] rather than by excluding a portion of the negative samples, as employed in StackIL6.

Based on our observations, we suggest a two-layer approach for identifying IL-6 or any cytokine-specific inducing peptides to overcome the limitations of the existing methods (Figure 5B). In the first layer, a prediction model should be developed using the dataset containing IL-6-inducing peptides as positive samples and peptides inducing proinflammatory cytokines other than IL-6, RPs and OFPs as negative samples. Such dataset preparation will help predict whether a given peptide has the potential to induce IL-6, even if the query sequence is from multiple resources, such as AMPs and RPs. The second layer develops a prediction model using the same first-layer positive sample and experimentally validated non-IL-6. It predicts whether the first layer-predicted IL-6 has been experimentally validated. This two-layer approach can significantly reduce the false positive rate and enhance model robustness and practical applicability. Besides the IL-6-inducing region, computational researchers should focus on mapping all possible proinflammatory cytokines inducing regions from viral proteins. Such predictors would be more exciting and beneficial

for experimentalists to map all possible cytokine regions, even in the scenario of emergence of a new virus in the future.

The application of rigorous computational approaches is essential for the development of a prediction model. Here, we found that several existing ML predictors (Meta-iAVP, Firm-AVP and PEPred-Suit) use the same AVPpred dataset, but different approaches showed poor performance. Nevertheless, ENNAVIA-B (developed with the AVPpred dataset) achieved a comparatively good performance, which could be attributed to its novel computational approach. Recently, several rigorous and classical computational frameworks have been proposed to identify different sequence-based functions [83–85]. Utilizing such strategies may improve the prediction model (AVP or IL-6) performance. First, deep learning (DL) has recently emerged as a robust ML algorithm that automatically learns suitable feature representations from the training data [86–90]. However, larger training samples are required for a reliable and robust performance. In future studies, computational biologists may consider DL approaches based on the size of the training dataset. Second, employing multiple feature encodings and classifiers is highly recommended compared with single feature-based models [23, 91–96]. Third, integrating multiple classifiers using an ensemble approach, a stacking approach and a meta-predictor can combine the strengths of individual classifiers and improve the prediction performance [74, 75, 97, 98]. Finally, exploration and comparison of different computational approaches on the same dataset are necessary for selecting the best method [99]. Furthermore, providing the best model as a user-friendly web server will help experimentalists minimize the time and cost involved in the experimental approaches.

Conclusions

In the last two decades, the emergence and re-emergence of viruses have significantly affected socioeconomic welfare, thus forewarning us about our incompetence in dealing with viral pandemics. Furthermore, new viral pathogens are highly likely to emerge, thus necessitating novel countermeasures. The COVID-19 crisis has rekindled antiviral research, as antiviral drugs serve as essential therapeutic alternatives [57]. It is already evident that SARS-CoV-2 has a greater mutation frequency, which makes it resistant to the current antiviral drugs. AVPs are one of the most promising anti-SARS-CoV-2 therapeutic options owing to their ability to overcome drug resistance by causing minimal sequence alterations in pace with viral mutations or by combining with conventional therapeutics [57, 100]. Furthermore, the cytokine storm syndrome is considered the main cause of mortality in patients with severe COVID-19 due to the triggering of a strong immune response [29]. IL-6 serves as the best biomarker for studying the severity of COVID-19 disease [101–104]. Similar to experimental methods, computational tools require a series of external validations and refinements. In this review, we comprehensively assessed state-of-the-art ML prediction algorithms, that is AVP prediction tools and IL-6-inducing peptide prediction tools. To perform an unbiased evaluation, we constructed two independent validation datasets to benchmark all existing ML tools. Our comparative analysis showed that ENNAVIA-D is the best ML predictor tool for identifying ACVPs among all 11 existing predictors. Although IL-6Pred_XGB was ranked at the top among IL-6 prediction methods, its performance was marginally better than random prediction, thus limiting its practical usage. This study will guide computational biologists in the design and development of robust and improvised ML prediction algorithms against SARS-CoV-2 and other viral infections. We also anticipate that it will provide directions for a wider research group in the field of immunotherapy and vaccine development targeting SARS-CoV.

Key Points

- We systematically evaluated the performance of existing AVPs and IL-6-inducing peptide predictors using their respective standard validation/independent datasets.
- ENNAVIA-D is the best method for ACVP identification. However, the existing tools are unsuitable for identifying IL-6-inducing peptides.
- This study provides valuable guidance to researchers interested in developing cutting-edge bioinformatics tools for identifying ACVPs and IL-6-inducing peptides.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Data availability

Data are available from the authors upon reasonable request.

Acknowledgements

The authors thank Prof. Tzong-Yi Lee and Yuxuan Pang for performing the PreAntiCoV calculations and Dr Patrick Timmons for performing the ENNAVIA calculations for our validation set. We would also like to thank the late Dr Hemant Kumar Srivastava, Associate Dean, NIPER Guwahati, for his scientific insights and contributions to this field.

Funding

National Research Foundation of Korea (NRF) funded by the Korean government (MSIT) (2021R1A2C1014338, 2019R1I1A1A01062260, 2020R1A4A4079722).

References

1. Liu YC, Kuo RL, Shih SR. COVID-19: the first documented coronavirus pandemic in history. *Biom J* 2020;**43**:328–33.
2. Tay MZ, Poh CM, Renia L, *et al.* The trinity of COVID-19: immunity, inflammation and intervention. *Nat Rev Immunol* 2020;**20**:363–74.
3. Mystery virus found in Wuhan resembles bat viruses but not SARS. *Chinese scientist says* . https://www.sciencemag.org/news/2020/01/mystery-virus-found-wuhan-resembles-bat-viruses-not-sars-chinese-scientist-says.
4. Yu K, Zhang Q, Liu Z, *et al.* Deep learning based prediction of reversible HAT/HDAC-specific lysine acetylation. *Brief Bioinform* 2020;**21**:1798–805.
5. Shah M, Woo HG. Molecular perspectives of SARS-CoV-2: pathology, immune evasion, and therapeutic interventions. *Mol Cells* 2021;**44**:408.
6. https://www.who.int/data/stories/world-health-statistics-2021-a-visual-summary.

7. Tannock GA, Kim H, Xue L. Why are vaccines against many human viral diseases still unavailable; an historic perspective? *J Med Virol* 2020;**92**:129–38.
8. Marqus S, Pirogova E, Piva TJ. Evaluation of the use of therapeutic peptides for cancer treatment. *J Biomed Sci* 2017;**24**:21.
9. Craik DJ, Fairlie DP, Liras S, *et al*. The future of peptide-based drugs. *Chem Biol Drug Des* 2013;**81**:136–47.
10. Eggink D, Bontjer I, de Taeye SW, *et al*. HIV-1 anchor inhibitors and membrane fusion inhibitors target distinct but overlapping steps in virus entry. *J Biol Chem* 2019;**294**:5736–46.
11. Ding X, Zhang X, Chong H, *et al*. Enfuvirtide (T20)-based lipopeptide is a potent HIV-1 cell fusion inhibitor: implications for viral entry and inhibition. *J Virol* 2017;**91**.
12. Thakur N, Qureshi A, Kumar M. AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res* 2012;**40**:W199–204.
13. Chang KY, Yang JR. Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS One* 2013;**8**:e70166.
14. Zare MMH, Faramarzi FK, Beigi MM, *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.
15. Beltran Lissabet JF, Belen LH, Farias JG. AntiVPP 1.0: a portable tool for prediction of antiviral peptides. *Comput Biol Med* 2019;**107**:127–30.
16. 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.
17. Wei L, Zhou C, Su R, *et al*. PEPred-suite: improved and robust prediction of therapeutic peptides using adaptive feature representation learning. *Bioinformatics* 2019;**35**:4272–80.
18. Meher PK, Sahu TK, Saini V, *et al*. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci Rep* 2017;**7**:42362.
19. Chung CR, Kuo TR, Wu LC, *et al*. Characterization and identification of antimicrobial peptides with different functional activities. *Brief Bioinform* 2019;**21**:1098–1114.
20. Gull S, Shamim N, Minhas F. AMAP: hierarchical multi-label prediction of biologically active and antimicrobial peptides. *Comput Biol Med* 2019;**107**:172–81.
21. Lin W, Xu D. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types. *Bioinformatics* 2016;**32**:3745–52.
22. Zhang YP, Zou Q. PPTPP: a novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning. *Bioinformatics* 2020;**36**:3982–7.
23. Pang Y, Yao L, Jhong JH, *et al*. AVPIden: a new scheme for identification and functional prediction of antiviral peptides based on machine learning approaches. *Brief Bioinform* 2021.
24. Pang Y, Wang Z, Jhong JH, *et al*. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. *Brief Bioinform* 2021;**22**:1085–95.
25. 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.
26. Schaduangrat N, Nantasenamat C, Prachayasittikul V, *et al*. Meta-iAVP: a sequence-based meta-predictor for improving the prediction of antiviral peptides using effective feature representation. *Int J Mol Sci* 2019;**20**.
27. Cheng L, Han X, Zhu Z, *et al*. Functional alterations caused by mutations reflect evolutionary trends of SARS-CoV-2. *Brief Bioinform* 2021;**22**:1442–50.
28. Cheng L, Zhu Z, Wang C, *et al*. COVID-19 induces lower levels of IL-8, IL-10, and MCP-1 than other acute CRS-inducing diseases. *Proc Natl Acad Sci* 2021;**118**.
29. Mehta P, McAuley DF, Brown M, *et al*. COVID-19: consider cytokine storm syndromes and immunosuppression. *Lancet* 2020;**395**:1033–4.
30. Angriman F, Ferreyro BL, Burry L, *et al*. Interleukin-6 receptor blockade in patients with COVID-19: placing clinical trials into context. *Lancet Respir Med* 2021;**9**:655–64.
31. Cavalli G, Larcher A, Tomelleri A, *et al*. Interleukin-1 and interleukin-6 inhibition compared with standard management in patients with COVID-19 and hyperinflammation: a cohort study. *Lancet Rheumatol* 2021;**3**:e253–61.
32. Channappanavar R, Perlman S. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. *Semin Immunopathol* 2017;**39**:529–39.
33. Diao B, Wang C, Tan Y, *et al*. Reduction and functional exhaustion of T cells in patients with coronavirus disease 2019 (COVID-19). *Front Immunol* 2020;**11**:827.
34. Huang C, Wang Y, Li X, *et al*. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* 2020;**395**:497–506.
35. Zhang W, Li L, Liu J, *et al*. The characteristics and predictive role of lymphocyte subsets in COVID-19 patients. *Int J Infect Dis* 2020;**99**:92–9.
36. Liu F, Li L, Xu M, *et al*. Prognostic value of interleukin-6, C-reactive protein, and procalcitonin in patients with COVID-19. *J Clin Virol* 2020;**127**:104370.
37. Han H, Ma Q, Li C, *et al*. Profiling serum cytokines in COVID-19 patients reveals IL-6 and IL-10 are disease severity predictors. *Emerg Microbes Infect* 2020;**9**:1123–30.
38. Gubernatorova EO, Gorshkova EA, Polinova AI, *et al*. IL-6: relevance for immunopathology of SARS-CoV-2. *Cytokine Growth Factor Rev* 2020;**53**:13–24.
39. Schmidt-Arras D, Rose-John S. IL-6 pathway in the liver: from physiopathology to therapy. *J Hepatol* 2016;**64**:1403–15.
40. Kishimoto T. IL-6: from its discovery to clinical applications. *Int Immunol* 2010;**22**:347–52.
41. Lu R, Zhao X, Li J, *et al*. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* 2020;**395**:565–74.
42. Okabayashi T, Kariwa H, Yokota S, *et al*. Cytokine regulation in SARS coronavirus infection compared to other respiratory virus infections. *J Med Virol* 2006;**78**:417–24.
43. Notz Q, Schmalzing M, Wedekink F, *et al*. Pro- and anti-inflammatory responses in severe COVID-19-induced acute respiratory distress syndrome-an observational pilot study. *Front Immunol* 2020;**11**:581338.
44. Dhanda SK, Vir P, Raghava GP. Designing of interferon-gamma inducing MHC class-II binders. *Biol Direct* 2013;**8**:30.
45. Dhanda SK, Gupta S, Vir P, *et al*. Prediction of IL4 inducing peptides. *Clin Dev Immunol* 2013;**2013**:263952.
46. Nagpal G, Usmani SS, Dhanda SK, *et al*. Computer-aided designing of immunosuppressive peptides based on IL-10 inducing potential. *Sci Rep* 2017;**7**:42851.

47. Gupta S, Mittal P, Madhu MK, *et al*. IL17eScan: A tool for the identification of peptides inducing IL-17 response. *Front Immunol* 2017;**8**:1430.
48. Lata S, Raghava GP. CytoPred: a server for prediction and classification of cytokines. *Protein Eng Des Sel* 2008;**21**:279–82.
49. Gupta S, Madhu MK, Sharma AK, *et al*. ProInflam: a webserver for the prediction of proinflammatory antigenicity of peptides and proteins. *J Transl Med* 2016;**14**:178.
50. Manavalan B, Shin TH, Kim MO, *et al*. PIP-EL: a new ensemble learning method for improved proinflammatory peptide predictions. *Front Immunol* 2018;**9**:1783.
51. Gupta S, Sharma AK, Shastri V, *et al*. Prediction of anti-inflammatory proteins/peptides: an insilico approach. *J Transl Med* 2017;**15**:7.
52. Dhall A, Patiyal S, Sharma N, *et al*. Computer-aided prediction and design of IL-6 inducing peptides: IL-6 plays a crucial role in COVID-19. *Brief Bioinform* 2021;**22**:936–45.
53. Charoenkwan P, Chiangjong W, Nantasenamat C, *et al*. StackIL6: a stacking ensemble model for improving the prediction of IL-6 inducing peptides. *Brief Bioinform* 2021.
54. Pirtskhalava M, Amstrong AA, Grigolava M, *et al*. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res* 2021;**49**:D288–97.
55. Chowdhury SM, Talukder SA, Khan AM, *et al*. Antiviral peptides as promising therapeutics against SARS-CoV-2. *J Phys Chem B* 2020;**124**:9785–92.
56. Schutz D, Ruiz-Blanco YB, Munch J, *et al*. Peptide and peptide-based inhibitors of SARS-CoV-2 entry. *Adv Drug Deliv Rev* 2020;**167**:47–65.
57. Tonk M, Ruzek D, Vilcinskas A. Compelling evidence for the activity of antiviral peptides against SARS-CoV-2. *Viruses* 2021;**13**.
58. Rao B, Zhou C, Zhang G, *et al*. ACPred-fuse: fusing multi-view information improves the prediction of anticancer peptides. *Brief Bioinform* 2020;**21**:1846–55.
59. Xu J, Li F, Leier A, *et al*. Comprehensive assessment of machine learning-based methods for predicting antimicrobial peptides. *Brief Bioinform* 2021;**22**:1–22.
60. Fu L, Niu B, Zhu Z, *et al*. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
61. Huang Y, Niu B, Gao Y, *et al*. CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010;**26**:680–2.
62. Vita R, Mahajan S, Overton JA, *et al*. The immune epitope database (IEDB): 2018 update. *Nucleic Acids Res* 2019;**47**:D339–43.
63. Li J, Pu Y, Tang J, *et al*. DeepAVP: a dual-channel deep neural network for identifying variable-length antiviral peptides. *IEEE J Biomed Health Inform* 2020;**24**:3012–9.
64. 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**:13–19.
65. Meher PK, Sahu TK, Saini V, *et al*. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci Rep* 2017;**7**:1–12.
66. 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**:1–8.
67. Timmons PB, Hewage CM. ENNAACT is a novel tool which employs neural networks for anticancer activity classification for therapeutic peptides. *Biomed Pharmacother* 2021;**133**:111051.
68. Charoenkwan P, Anuwongcharoen N, Nantasenamat C, *et al*. In silico approaches for the prediction and analysis of antiviral peptides: a review. *Curr Pharm Des* 2020;**27**:2180–88.
69. Qureshi A, Thakur N, Tandon H, *et al*. AVPdb: a database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Res* 2014;**42**:D1147–53.
70. Jhong JH, Chi YH, Li WC, *et al*. dbAMP: an integrated resource for exploring antimicrobial peptides with functional activities and physicochemical properties on transcriptome and proteome data. *Nucleic Acids Res* 2019;**47**:D285–97.
71. Kang X, Dong F, Shi C, *et al*. DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci Data* 2019;**6**:148.
72. Qureshi A, Thakur N, Kumar M. HIPdb: a database of experimentally validated HIV inhibiting peptides. *PLoS One* 2013;**8**:e54908.
73. Govindaraj RG, Subramaniyam S, Manavalan B. Extremely-randomized-tree-based prediction of N(6)-methyladenosine sites in Saccharomyces cerevisiae. *Curr Genomics* 2020;**21**:26–33.
74. Wei L, He W, Malik A, *et al*. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Brief Bioinform* 2020;**22**:1–15.
75. Hasan MM, Alam MA, Shoombuatong W, *et al*. NeuroPred-FRL: an interpretable prediction model for identifying neuropeptide using feature representation learning. *Brief Bioinform* 2021.
76. 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.
77. Hasan MM, Shoombuatong W, Kurata H, *et al*. Critical evaluation of web-based DNA N6-methyladenine site prediction tools. *Brief Funct Genomics* 2021;**20**:258–72.
78. Manavalan B, Hasan MM, Basith S, *et al*. Empirical comparison and analysis of web-based DNA N4-methylcytosine site prediction tools. *Molecular Therapy-Nucleic Acids* 2020;**22**:406–20.
79. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res* 2016;**44**:D1087–93.
80. 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.
81. Dhanda SK, Mahajan S, Paul S, *et al*. IEDB-AR: immune epitope database-analysis resource in 2019. *Nucleic Acids Res* 2019;**47**:W502–6.
82. Manavalan B, Govindaraj RG, Shin TH, *et al*. iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction. *Front Immunol* 2018;**9**:1695.
83. Liu K, Cao L, Du P, *et al*. im6A-TS-CNN: identifying the N(6)-methyladenine site in multiple tissues by using the convolutional neural network. *Mol Ther Nucleic Acids* 2020;**21**:1044–9.
84. Liu K, Chen W. iMRM: a platform for simultaneously identifying multiple kinds of RNA modifications. *Bioinformatics* 2020;**36**:3336–42.

85. Tang Q, Nie F, Kang J, *et al*. mRNALocater: enhance the prediction accuracy of eukaryotic mRNA subcellular localization by using model fusion strategy. *Mol Ther* 2021;**29**:2617–23.
86. Lv H, Dao FY, Guan ZX, *et al*. Deep-Kcr: accurate detection of lysine crotonylation sites using deep learning method. *Brief Bioinform* 2020;**22**:1–10.
87. Dao FY, Lv H, Su W, *et al*. iDHS-deep: an integrated tool for predicting DNase I hypersensitive sites by deep neural network. *Brief Bioinform* 2021;**22**:1–8.
88. Lv H, Dao FY, Zulfiqar H, *et al*. DeepIPs: comprehensive assessment and computational identification of phosphorylation sites of SARS-CoV-2 infection using a deep learning-based approach. *Brief Bioinform* 2021.
89. Song Z, Huang D, Song B, *et al*. Attention-based multi-label neural networks for integrated prediction and interpretation of twelve widely occurring RNA modifications. *Nat Commun* 2021;**12**:4011.
90. Xie R, Li J, Wang J, *et al*. DeepVF: a deep learning-based hybrid framework for identifying virulence factors using the stacking strategy. *Brief Bioinform* 2021;**22**:1–15.
91. Wei L, Hu J, Li F, *et al*. Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms. *Brief Bioinform* 2018;**21**:106–19.
92. Lv H, Dao FY, Zhang D, *et al*. iDNA-MS: an integrated computational tool for detecting DNA modification sites in multiple genomes. *iScience* 2020;**23**:100991.
93. Lv H, Zhang ZM, Li SH, *et al*. Evaluation of different computational methods on 5-methylcytosine sites identification. *Brief Bioinform* 2020;**21**:982–95.
94. Yang H, Yang W, Dao FY, *et al*. A comparison and assessment of computational method for identifying recombination hotspots in Saccharomyces cerevisiae. *Brief Bioinform* 2020;**21**:1568–80.
95. Liang X, Li F, Chen J, *et al*. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. *Brief Bioinform* 2021;**22**:1–17.
96. Zhang ZY, Yang YH, Ding H, *et al*. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief Bioinform* 2021;**22**:526–35.
97. Basith S, Manavalan B, Shin TH, *et al*. SDM6A: a web-based integrative machine-learning framework for predicting 6mA sites in the rice genome. *Mol Ther Nucleic Acids* 2019;**18**:131–41.
98. Manavalan B, Basith S, Shin TH, *et al*. Meta-4mCpred: a sequence-based meta-predictor for accurate DNA 4mC site prediction using effective feature representation. *Mol Ther Nucleic Acids* 2019;**16**:733–44.
99. Hasan MM, Shoombuatong W, Kurata H, *et al*. Critical evaluation of web-based DNA N6-methyladenine site prediction tools. *Brief Funct Genomics* 2021;**20**:258–72.
100. Elnagdy S, AlKhazindar M. The potential of antimicrobial peptides as an antiviral therapy against COVID-19. *ACS Pharmacol Transl Sci* 2020;**3**:780–2.
101. Kermali M, Khalsa RK, Pillai K, *et al*. The role of biomarkers in diagnosis of COVID-19 - a systematic review. *Life Sci* 2020;**254**:117788.
102. Wang C, Fei D, Li X, *et al*. IL-6 may be a good biomarker for earlier detection of COVID-19 progression. *Intensive Care Med* 2020;**46**:1475–6.
103. Santa Cruz A, Mendes-Frias A, Oliveira AI, *et al*. IL-6 is a biomarker for the development of fatal SARS-CoV-2 pneumonia. *Front Immunol* 2021;**12**:263.
104. Sabaka P, Koščálová A, Straka I, *et al*. Role of interleukin 6 as a predictive factor for a severe course of Covid-19: retrospective data analysis of patients from a long-term care facility during Covid-19 outbreak. *BMC Infect Dis* 2021;**21**:1–8.