Computational and Structural Biotechnology Journal 23 (2024) 3625–3633

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

ACVPICPred: Inhibitory activity prediction of anti-coronavirus peptides based on artificial neural network

Min Li [a], Yifei Wu [a], Bowen Li [a], Chunying Lu [a], Guifen Jian [a], Xing Shang [a], Heng Chen [a,*], Jian Huang [b,*], Bifang He [a,c,**]

[a] Medical College, Guizhou University, Huaxi District, Guiyang 550025, Guizhou, China
[b] School of Life Science and Technology, University of Electronic Science and Technology of China, No.2006, Xiyuan Ave, West Hi-Tech Zone, Chengdu 6173001, Sichuan, China
[c] State Key Laboratory of Public Big Data, Guizhou University, Huaxi District, Guiyang 550025, Guizhou, China

ARTICLE INFO

Keywords:
Anti-coronavirus peptides
Inhibitory concentration
Regression
Artificial neural network

ABSTRACT

Peptides, as small molecular compounds, exhibit prominent advantages in the inhibition of coronaviruses due to their safety, efficacy, and specificity, holding great promise as drugs against coronaviruses. The rapid and efficient determination of the activity of anti-coronavirus peptides (ACovPs) can greatly accelerate the development of drugs for treating coronavirus-related diseases. Hence, we present ACVPICPred, a computational model designed to predict the inhibitory activity of ACovPs based on their sequences and structural information. By leveraging bioinformatics tools AlphaFold3 for structural predictions and several feature extraction methods, the model integrates both sequence and structural features to enhance prediction accuracy. To address the limitations of existing datasets, we employed data augmentation techniques, including the introduction of noise and the SMOGN, to improve the model robustness. The model's performance was evaluated through five-fold cross-validation, achieving a Pearson correlation coefficient of 0.7668 ($p < 0.05$) and an R^2 of 0.5880 on the training dataset. Overall, in our study, compared to models that only use sequence features, models that combine structural features have achieved more robust results in various evaluation metrics. ACVPICPred is freely accessible at the following URL: http://i.uestc.edu.cn/acvpICPred/main/Main.php.

1. Introduction

Coronaviruses are a group of enveloped, positive-sense, single-stranded RNA viruses that belong to the order Nidovirales and the family *Coronaviridae*. They possess receptor-binding proteins on the surface of their outer shells that engage with specific receptors on the host cell surface, mediating the virus's entry into the host cell [1]. These viruses cause respiratory infections in both animals and humans [2], leading to manifestations such as colds and respiratory ailments [3]. There have been three major human coronavirus outbreaks to date, respectively caused by severe acute respiratory syndrome coronavirus (SARS-CoV), middle east respiratory syndrome coronavirus (MERS-CoV), and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). As of July 21, 2024, the worldwide tally for confirmed COVID-19 cases has surpassed 775 million, with more than 7 million reported deaths [4].

Various approaches for treating coronavirus infections have been investigated. Treatment strategies can generally be divided into two categories based on their therapeutic targets: host-targeted and virus-directed therapies [5]. Host-targeted therapies disrupt host cell mechanisms, enhance immune responses, and reduce inflammation [6]. They combat viruses by enhancing interferon responses, inhibiting host signaling pathways associated with viral replication, disrupting host factors that utilized during viral processes, stimulating the host's defense mechanisms, and modulating pathways disrupted by pathogens that cause excessive inflammation. Examples of such medications are type-I interferon-β [7], entry inhibitors N-(2-aminoethyl)−1-aziridine-ethanamine [8], and convalescent plasma therapy [9]. The virus-directed therapeutic strategies specifically target viral components, including spike glycoproteins, enzymes involved in nucleic acid synthesis, and structural/accessory proteins [5]. Nucleoside/nucleotide reverse transcriptase inhibitors, protease inhibitors, entry/uncoating inhibitors, and polymerase inhibitors, are among the main types of

* Corresponding authors.
** Corresponding author at: Medical College, Guizhou University, Huaxi District, Guiyang 550025, Guizhou, China.
E-mail addresses: hchen13@gzu.edu.cn (H. Chen), hj@uestc.edu.cn (J. Huang), bfhe@gzu.edu.cn (B. He).

https://doi.org/10.1016/j.csbj.2024.09.015
Received 8 June 2024; Received in revised form 18 September 2024; Accepted 24 September 2024
Available online 2 October 2024

antiviral medications [6]. For example, protease inhibitors like lopinavir/ritonavir[10], polymerase inhibitor remdesivir [11], and favipiravir [12], are utilized as post-infection treatments. However, the emergence of viral resistance has led to diminished efficacy of current treatments [13]. Consequently, there is an urgent need for the development of novel antiviral drugs to counteract coronavirus infections effectively.

Peptide drugs are distinguished by their high selectivity, natural composition, potent efficacy, low toxicity, broad target coverage, minimal tissue accumulation, and relatively straightforward synthesis process compared to other biopharmaceuticals [14]. Studies have shown that specific anti-coronavirus peptides (ACovPs) exhibit inhibitory effects against coronaviruses [15–17]. For instance, HR2P, a peptide derived from the S protein of MERS-CoV, has been found to inhibit the fusion and replication of MERS-CoV [18]. Additionally, the peptide OC43-HR2P, originating from the HR2 region of HCoV-OC43, has demonstrated extensive inhibitory effects on the fusion of diverse human coronaviruses. Its optimized form, EK1, has exhibited notably improved inhibitory activity and superior drug properties [19]. Despite the vast potential of peptide drugs, their approval is often hindered by the intricate and lengthy procedures involving peptide selection, identification, preclinical and clinical trials, and ultimate approval [14]. Therefore, developing computational methods for in silico screening of peptides with potent coronavirus-inhibiting abilities would be highly conducive to expediting drug development.

With the advancement of bioinformatics, machine learning and deep learning methods have been extensively utilized in peptide research, providing effective tools for predicting peptides with specific function [20–22], estimating the physicochemical properties of peptides [23] and aiding in peptide design [24,25]. Currently, numerous computational models, such as PreAntiCoV [26] and iACVP [27], have been developed to identify ACovPs. However, these classification models are unable to predict the specific inhibitory activity value or potency of ACovPs. Investigating anti-coronavirus activity of a peptide by traditional experimental methods is time-consuming, resource-intensive and costly. Therefore, an in silico method that directly predicts peptides' anti-coronavirus activity would be advantageous for more accurately discovering peptides with strong anti-coronavirus ability, thereby improving drug development efficiency, avoiding unnecessary experimentation and trial-and-error processes, conserving time and research resources.

In response to these challenges, this study proposes a prediction model that utilizes peptide sequences and structural information to assess the inhibitory potency of ACovPs. The model's predictive accuracy is validated by comparing predicted values with experimentally determined anti-coronavirus activity, using five-fold cross-validation and an independent test dataset. To facilitate broader use, the model has been implemented into a user-friendly web server called ACVPICPred, which is publicly accessible at http://i.uestc.edu.cn/acvpICPred/main/Main.php. This computational model can be used to identify peptides with enhanced anti-coronavirus potential for experimental validation, thereby advancing ACovPs research.

2. Materials and methods

2.1. Datasets

Bioinformatic resources, for example ACovPepDB [28] and dbAMP [29], have consolidated extensive information related to ACovPs. ACovPepDB, the most recent manually curated repository, includes 214 distinct ACovPs with information on their amino acid sequence and inhibitory activity, as well as assays to measure their inhibitory effects. We retrieved the complete dataset from this database and performed necessary data preprocessing steps. We initially excluded sequences containing non-natural amino acids. We also discarded entries lacking inhibition values. For entries with ambiguous inhibitory values, such as numerical ranges (for example >500, <10) or with uncertainties (for example 1.2 ± 0.5), we applied the following criteria: (1) Entries with values in the "> range" were excluded; (2) For those in the "< range", we used the maximum value as the final data value. (3) For values with a ± symbol, we considered the average value preceding the ± as the definitive data value.

In addition to peptide sequences, we took into account several other additional factors related to anti-coronavirus activity. These included the specific virus targeted by each peptide, type of antiviral assays conducted to evaluate its inhibitory effect on coronaviruses, inhibitory value type (Half-Maximal Inhibitory Concentration (IC_{50}), 90 % Inhibitory Concentration (IC_{90}), or Half-Maximal Effective Concentration (EC_{50})), and the units of these inhibitory values (micromole (μM) or micrograms per milliliter (μg/ML)). After completing the aforementioned data preprocessing, the dataset comprised a total of 163 data items. Subsequently, 10 % of them (16 items) were randomly selected as an independent testing dataset, while the remaining 90 % (147 items) were allocated as the training dataset.

2.2. Structural feature

Peptide structures were predicted based on their sequences using the AlphaFold3 algorithm [30]. By analyzing peptide structures with DSSP [31], we extracted several key features including secondary structure type, solvent-accessible surface area (ASA), average relative solvent-accessible surface area (avg_rASA), the secondary structure segment length, and residue positions. ASA represents the surface area of a biomolecule that can interact with solvents, which is crucial for comprehending protein interactions, identifying active functional regions, and distinguishing stable internal structures. The avg_rASA metric quantifies the average exposure of local regions to the solvent. Secondary structures of peptides refer to specific spatial conformations, such as helices, strands, and random coils, formed by hydrogen bonds within localized regions of the peptide chain. DSSP categorizes protein secondary structures into eight types: α-helix, 3_{10}-helix, π-helix, parallel β-sheets, antiparallel β-sheets, turns, bends, and random coils. We incorporated secondary structure type, segment length, and residue position into the structural feature matrix to enhance the characterization of these conformations.

2.3. Sequence feature

To comprehensively characterize the differences between various peptide sequences, five feature extraction methods were utilized to encode each peptide: amino acid composition (AAC), dipeptide composition (DPC), composition of k-spaced amino acid group pairs (CKSAAGP), pseudo amino acid composition (PAAC), and physicochemical properties. This process was performed using an internally developed Python script. By consolidating all peptide descriptors, each peptide was transformed into a feature vector with a dimensionality of

Table 1
529-dimensional features.

	Descriptors	Number of features
1	AAC	20
2	DPC	400
3	CKSAAGP	75
4	PAAC	25
5	Isoelectric point Net charge Hydrophobicity Hydrophobic moment Transmembrane propensity Boman index Aliphatic index Alpha helical propensity Solubility	9
6	Total	529

529 (Table 1). Furthermore, we applied One-Hot encoding to convert various factors related to anti-coronavirus activity, including the type of resistant coronaviruses, assay type, inhibitory value type, and the unit of the inhibitory value, into numerical data.

AAC characterizes peptide sequences by calculating the percentage of each amino acid, yielding a 20-dimensional feature vector. In this vector, each dimension corresponds to the relative abundance of its respective amino acid.

$$AAC(i) = \frac{N(i)}{N} \tag{1}$$

where i represents the i-th amino acid, $N(i)$ denotes the number of the i-th amino acid, and N represents the total number of residues in the sequence.

DPC calculates the occurrence frequency of each dipeptide in the sequence, generating a 400-dimensional feature vector.

$$DPC(a, \quad r) = \frac{N_{ar}}{N-1} \tag{2}$$

where a and r represent the a-th and r-th amino acids, respectively, and N_{ar} represents the number of this dipeptide. N is the total number of residues in the sequence.

PAAC compensates for the limitations of traditional AAC methods by incorporating sequence order information [32]. PAAC introduces a set of discretization factors and is defined by two crucial parameters: the weight factor ω and discrete counted-rank correlation factor λ. It generates a feature vector with $20 + \lambda$ dimensions. In this study, we have specifically chosen the values $\omega = 0.4$ and $\lambda = 5$.

CKSAAGP integrates amino acid spacing with the physicochemical properties of amino acids to calculate the occurrence frequency of residue pairs with k-spacing [33]. Specifically, CKSAAGP initially classifies the 20 amino acids into five groups based on their physicochemical properties: aliphatic, aromatic, positively charged, negatively charged, and uncharged amino acids ($g1$-$g5$). For each group of 25 amino acid pairs labeled with physicochemical properties, the occurrence frequencies of these amino acid pairs in peptide sequences are calculated using k-spacing, and the frequencies of these combinations are computed to obtain the final feature vector. In this study, k values of 0, 1, and 2 were selected, thereby resulting in a 75-dimensional feature vector.

$$\left(\frac{N_{g1g1}}{L-(k+1)}, \frac{N_{g1g2}}{L-(k+1)}, \quad \frac{N_{g1g3}}{L-(k+1)}, \ldots, \quad \frac{N_{g5g5}}{L-(k+1)}\right)_{25} \tag{3}$$

where L represents the length of the peptide sequence.

The physicochemical properties of peptides, such as solubility and isoelectric point, are also crucial factors influencing their functions. Referring to the methods utilized in PreAntiCoV [26], we selected the same eight physicochemical property features and additionally incorporated solubility (see Table 1). Since no existing dataset or database directly provides solubility information for peptides, we employed the PeptideBERT [34] tool to predict the solubility from the peptide sequences.

2.4. Data Augmentation

To tackle the challenge of limited sample size, we employed two methods to augment the training dataset. The first method involved extracting features from the peptide sequences and adding noise to the feature vectors. Specifically, we randomly selected 10 % of the standardized features and introduced 5 % noise. The second method, SMOGN [35], combines random undersampling, SmoteR, and Gaussian noise. The core concept of SMOGN is to generate synthetic samples by merging these strategies while using Gaussian noise as a conservative measure to mitigate the potential risks of SmoteR and increase the diversity of the synthetic samples. After applying these two augmentation techniques, we obtained an expanded dataset with 388 ACovPs (Table 2). It should be noted that data augmentation was performed exclusively on the training dataset, not on the testing dataset.

2.5. Feature selection

The utilization of diverse peptide descriptors enables a more comprehensive extraction of features from peptide sequences. However, not all features contribute effectively to model performance [36]. Our dataset consists of 388 samples with an initial feature dimension of 560. Given the limited sample size and high feature dimension, there exists a significant potential for overfitting. To mitigate this issue, we employed three common feature selection methods, including Pearson correlation coefficient, mutual information (MI), and least absolute shrinkage and selection operator (Lasso), which are also frequently used in the field of bioinformatics [37–39], to identify key features in this study.

Feature selection based on the Pearson correlation coefficient involves evaluating the linear relationship between each feature and the target variable. This is accomplished by computing the respective Pearson correlation coefficients, which quantify the degree of linear association. MI is a metric used to assess the level of interdependence between two random variables [40]. In feature selection, it calculates the mutual information between each feature and the target variable and then selects features with mutual information exceeding a set threshold. Lasso regression is an extension of linear regression that enables feature selection by incorporating an L1 regularization term [41], which imposes constraints on the coefficients of the model. Specifically, Lasso achieves sparsity in features by minimizing the objective function comprising squared loss and the L1 regularization term, leading to some feature coefficients being reduced to zero. By adjusting the regularization parameter, the extent of feature selection can be flexibly modulated.

2.6. Machine learning and deep learning

To construct regression models with robust fitting effects, we employed various traditional machine learning and deep learning methods for model training. Focusing solely on sequence data and utilizing the aforementioned feature extraction and selection methods, we developed models using Ridge regression, Lasso regression, Bayesian Ridge regression, Elastic Net, Gradient Boosting, K-Nearest Neighbors regression, Random Forest regression, Support Vector Regression, and multilayer perceptron (MLP). We utilized the grid search for optimal parameter tuning in all models to achieve a relatively better fitting performance for each regression model. Model construction was performed at a computational server (Sugon I840-G20, Dawning Information Industry Co., LTD., Beijing, China).

Furthermore, we designed a hybrid model that integrates both sequential and structural features to enhance predictive accuracy. The structure of the model is shown in Fig. 1. This model is architected with three principal components: a sequential feature processing module, a structural feature processing module, and a final feature fusion and prediction module. The sequence feature processing module processes the input sequence features through a single fully connected layer, with an input dimension corresponding to the number of features and an output dimension of 64. A rectified linear unit (ReLU) activation function is applied to enhance nonlinear expressiveness. For the structural feature processing module, the input is a two-dimensional matrix with

Table 2
Training Dataset after Augmentation.

Training Dataset	Number of anti-coronavirus peptides
Original training dataset	147
Newly generated dataset (5 % noise)	147
Newly generated dataset (SMOGN)	94
Total	388

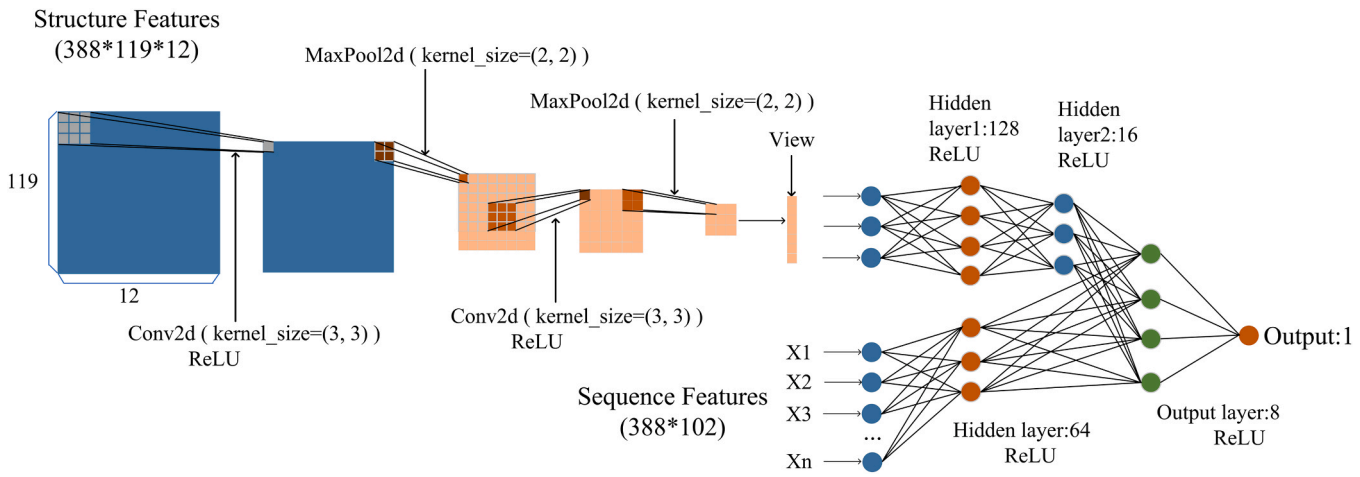


Fig. 1. Schematic diagram of the hybrid model.

dimensions corresponding to the structural feature length of 119 and the feature dimension of 12. It employs two two-dimensional convolutional neural network (2D-CNN) layers. The first layer has one output channel, a kernel size of 3 × 3, and a ReLU activation function, followed by a max-pooling operation with a kernel size of 2 × 2 to reduce the spatial dimensions. The second convolutional layer similarly has one output channel, a 3 × 3 kernel size, ReLU activation, and max-pooling with a 2 × 2 kernel size. After the convolutional layers, the output is flattened, and two fully connected layers are applied to reduce the feature dimensionality. The first fully connected layer maps the features to a 128-dimensional space, and the second layer reduces this to 16 dimensions, both using ReLU activation functions. In the feature fusion and prediction module, the output features from the sequence and structure feature processing modules are concatenated, yielding a unified feature vector of 80 dimensions (64 +16). This vector is then passed through a fully connected layer (8 dimensions, ReLU activation function), followed by a final linear layer that outputs a single predicted value. This predicted value corresponds to the inhibitory potency of ACovPs. The model was trained using Python 3.9 as the coding language, and PyTorch version 1.9.1 with support for CUDA 11.1 was employed to construct the model.

2.7. Performance evaluation

We employed five-fold cross-validation to evaluate the model performance. This approach evenly divides the training dataset into five folds, with four folds used for model training and the remaining one fold is used for testing [42]. Through five iterations, each time a different fold is chose as the testing dataset, and the model is trained with the other four folds and tested on the testing dataset. The final performance of model is determined by calculating the average performance across the five iterations.

The primary evaluation indicators of the regression model in this study include mean square error (*MSE*), Pearson correlation coefficient (*r*) and coefficient of determination (*R-squared*, R^2). The *MSE* was utilized as the loss function for all model training procedures. This metric effectively captures the discrepancy between the predicted and actual values, as demonstrated by the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(X_i^{act} - X_i^{pred}\right)^2 \quad (4)$$

The final fitting performance of the model was evaluated by employing the *r*, calculated by the following formula:

$$r = \sum_{i=1}^{n}\left(X_i^{act} - \frac{\overline{X^{act}}\left(X_i^{pred} - \overline{X^{pred}}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i^{act} - \overline{X^{act}}\right)^2}\sqrt{\sum_{i=1}^{n}\left(X_i^{pred} - \overline{X^{pred}}\right)^2}}\right) \quad (5)$$

The adequacy of fit for the regression model was assessed by using the R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(X_i^{act} - X_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(X_i^{act} - \overline{X^{act}}\right)^2} \quad (6)$$

In the above formulas, X_i^{pred} represents the predicted values by the model and X_i^{act} represents the actual values. $\overline{X^{act}}$ denotes the mean of the actual values, $\overline{X^{pred}}$ signifies the mean of the predicted values, and *n* represents the sample size.

2.8. Target prediction model construction

Additionally, we trained a simple target prediction model to enhance the practical relevance of ACVPICPred. We reorganized the peptide dataset downloaded from ACovpepDB. After retaining the entries with target information and removing redundancies, 74 ACovPs were obtained. According to the target type of these ACovPs, we divided them into two categories: the S1 class, consisting of 28 ACovPs, and the S2 class, comprising 46 ACovPs. ACovPs in the S1 class target the N-terminal subunit of spike protein (S1), whereas those in the S2 class interact with the C-terminal subunit of spike protein (S2). Employing the same feature extraction and selection methods mentioned above, we selected the Random Forest algorithm, which is suitable for small and imbalanced data sets, to construct the model. The model's efficacy was evaluated by a five-fold cross-validation, utilizing common evaluation indicators, including accuracy, sensitivity and specificity.

3. Result

3.1. Overall workflow

The workflow for the model construction in this study is illustrated in Fig. 2. We first conducted data collection, preprocessing, and dataset splitting. The resulting data comprised two main components: peptide sequences and classification data related to inhibitory values. We then obtained peptide structures that predicted by AlphaFold3 based on their sequences. Subsequently, various feature extraction methods were used to capture key feature information, and data enhancement was performed using the additive noise and SMOGN methods. We then trained the model using the augmented data and evaluated its performance

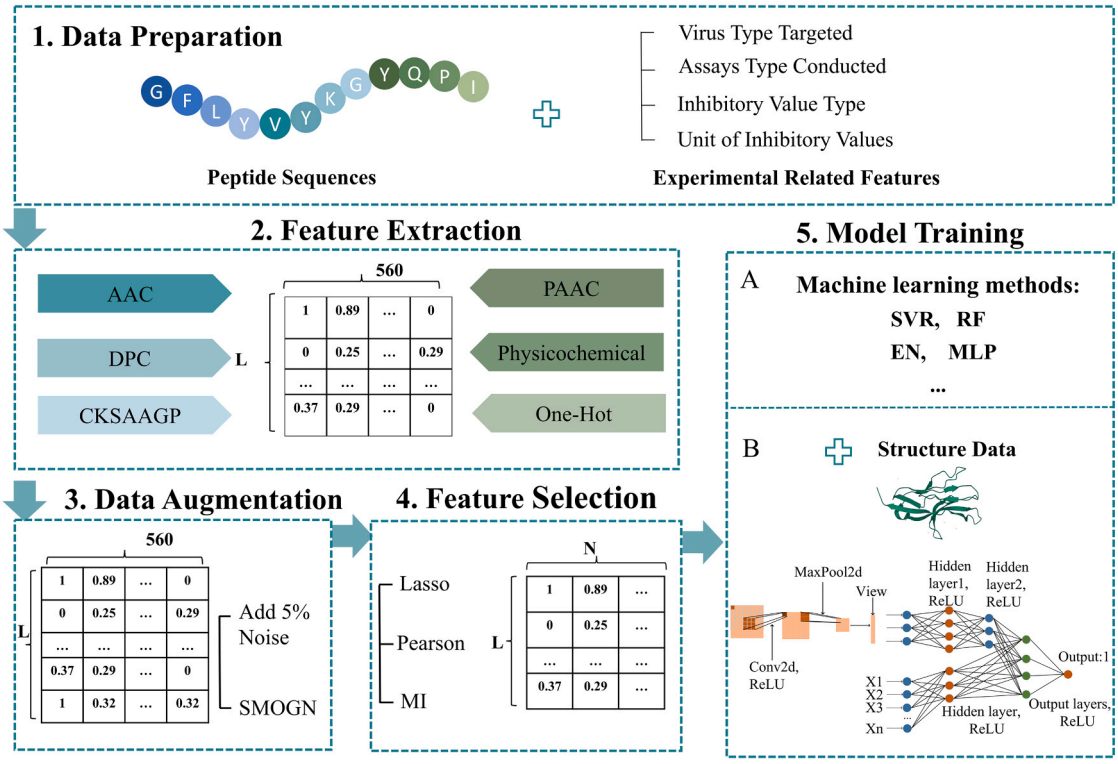


Fig. 2. Workflow of this study. AAC: amino acid composition; DPC: dipeptide composition; PAAC: pseudo amino acid composition; CKSAAGP: composition of k-spaced amino acid group pairs; Pearson: Pearson correlation coefficient; MI: mutual information; Lasso: least absolute shrinkage and selection operator; RF: Random Forest regression; SVR: Support Vector Regression; EN: Elastic Net; MLP: multilayer perceptron; 2D-CNN: Two-Dimensional Convolutional Neural Network.

using evaluation metrics. Finally, we selected the model that demonstrated the best performance to build our web platform.

3.2. Model performance on the training dataset

During the evaluation phase, all models were assessed using a five-fold cross-validation strategy. Among the models that did not incorporate structural data, the combination of MLP and Lasso exhibited the most promising results, achieving a Pearson correlation coefficient of 0.6427 ($p < 0.05$) and an R^2 of 0.4131 (Table 3).

In contrast, models trained with sequence and structure features demonstrated enhanced performance. Specifically, the model with Lasso for feature selection notably outperformed those constructed with features selected by the other two feature selection techniques with a Pearson correlation coefficient of 0.7668 ($p < 0.05$) and an R^2 of 0.5880 (Table 3). The superior performance observed in the model using Lasso during five-fold cross-validation suggests that the selected features effectively capture the key information influencing ACovPs inhibition values. Therefore, features selected by Lasso and the hybrid model were utilized for the construction of the final predictive model.

3.3. Model performance on the independent testing dataset

The model's generalization ability was rigorously evaluated using an independent testing dataset. The hybrid model using Lasso as the feature selection method demonstrated notable performance in independent testing, yielding a Pearson correlation coefficient of 0.8836 ($p < 0.05$) and an R^2 of 0.7807. These results indicate the model's robust generalization capacity on previously unseen data.

Table 3
Prediction performance of models (five-fold cross-validation).

Methods	*r* (*p*)				R^2			
	No feature selection	**LASSO**	**Pearson Cor**	**MI**	**No feature selection**	**LASSO**	**Pearson Cor**	**MI**
BR	0.4327 (<0.05)	0.5488 (<0.05)	0.4703 (<0.05)	0.5102 (<0.05)	0.1872	0.3012	0.2212	0.2603
EN	0.4951 (>0.05)	0.5354 (<0.05)	0.5174 (<0.05)	0.5276 (<0.05)	0.2451	0.2867	0.2677	0.2784
GB	0.5628 (<0.05)	0.5791 (<0.05)	0.6154 (<0.05)	0.6310 (<0.05)	0.3167	0.3354	0.3787	0.3982
KNN	0.5788 (<0.05)	0.5950 (<0.05)	0.6016 (<0.05)	0.5804 (<0.05)	0.3350	0.3540	0.3619	0.3369
Ridge	0.4770 (>0.05)	0.5582 (<0.05)	0.5038 (>0.05)	0.4964 (>0.05)	0.2275	0.3116	0.2538	0.2464
Lasso	0.4508 (>0.05)	0.5362 (<0.05)	0.4729 (>0.05)	0.4793 (<0.05)	0.2032	0.2875	0.2236	0.2297
RF	0.5343 (<0.05)	0.6096 (<0.05)	0.5862 (<0.05)	0.5924 (<0.05)	0.2855	0.3716	0.3436	0.3509
SVR	0.4979 (<0.05)	0.5838 (<0.05)	0.5108 (<0.05)	0.5128 (<0.05)	0.2479	0.3408	0.2609	0.2630
MLP	0.6000 (<0.05)	0.6427 (<0.05)	0.6116 (<0.05)	0.6077 (<0.05)	0.3600	0.4131	0.3741	0.3693
Hybrid (MLP, 2D-CNN)	0.6212 (<0.05)	**0.7668 (<0.05)**	0.7407 (<0.05)	0.7430 (<0.05)	0.3859	**0.5880**	0.5486	0.5520

Notes: *r*(*p*): Pearson correlation coefficient (*p*_value); BR: Bayesian Ridge regression; EN: Elastic Net; GB: Gradient Boosting; KNN: K-Nearest Neighbors; Ridge: Ridge regression; Lasso: Least Absolute Shrinkage and Selection Operator; RF: Random Forest; SVR: Support Vector Regression; MLP: multilayer perceptron; 2D-CNN: Two-Dimensional Convolutional Neural Network.

3.4. Performance of the target prediction model

The target prediction model exhibited commendable performance in the five-fold cross-validation, achieving an average accuracy of 86.38 %, an average sensitivity of 86.67 %, and an average specificity of 85.33 %. While this model is rudimentary in its design, it serves as a valuable adjunct to the network service, enhancing the interpretability of the predicted activities to a moderate degree. Its integration provides users with supplementary insights into potential targets of ACovPs.

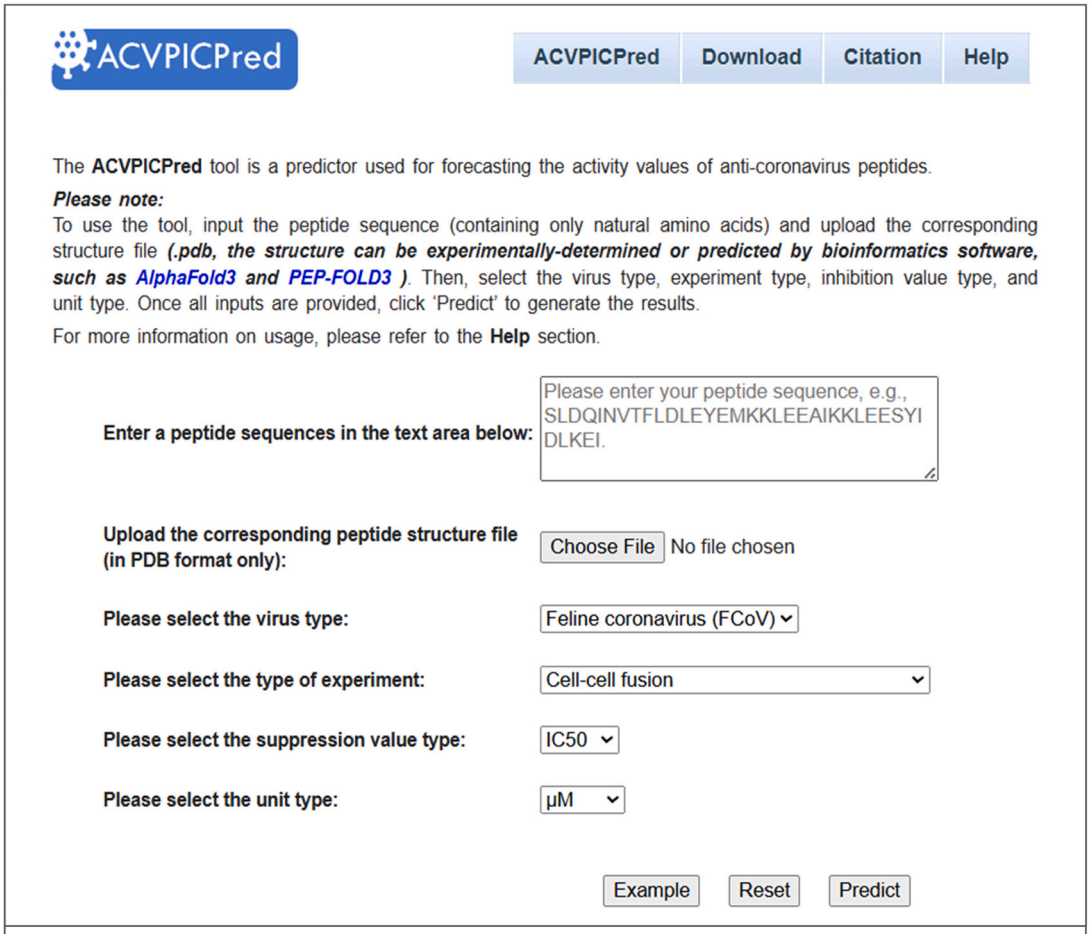
3.5. ACVPICPred web service

To enhance the accessibility of our model in predicting the activity value of ACovPs, we have developed a user-friendly web service that can be accessed for free at http://i.uestc.edu.cn/acvpICPred/main/Main.php. This webpage requires users to input the peptide sequence, select relevant experimental information, and upload peptide structure files in PDB format to estimate the anti-coronavirus activity values of the peptides (Fig. 3). The result interface provides the predicted anti-coronavirus efficacy values as well as resistant virus, type of the assay conducted, inhibitory value type, unit and the predicted target.

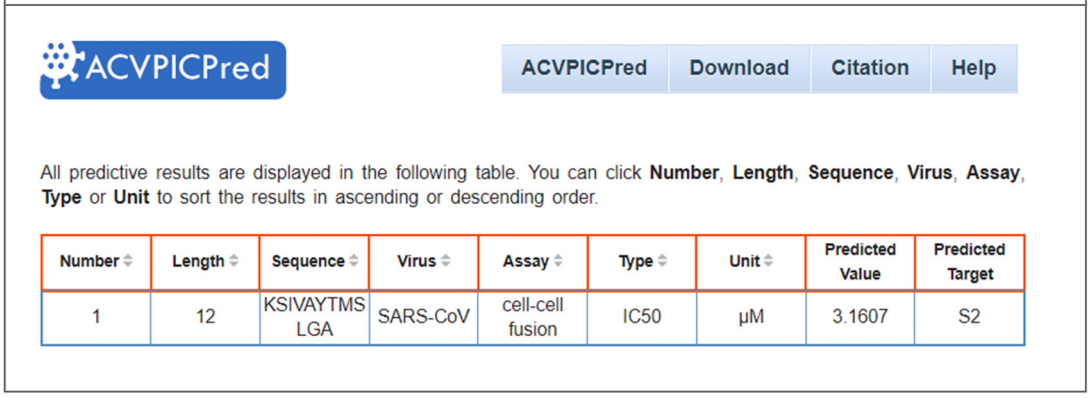
4. Discussion

AcovPs are expected to be peptide drug candidates for the treatment of diseases caused by coronaviruses. The inhibitory activity of AcovPs is one of the most critical parameters that determine its probability of being successfully developed into a therapeutic peptide. Computational

A



B



Number	Length	Sequence	Virus	Assay	Type	Unit	Predicted Value	Predicted Target
1	12	KSIVAYTMS LGA	SARS-CoV	cell-cell fusion	IC50	μM	3.1607	S2

Fig. 3. Web service interface. (A) Input interface. Users can input the peptide sequence, upload the corresponding structure file (.pdb) and select relevant inhibition value features to get the prediction. (B) Result interface. The result table provides the predicted anti-coronavirus efficacy values as well as resistant virus, type of the assay, inhibitory value type, unit, and the predicted target.

methods for predicting the inhibitory activity of AcovPs with high throughput and low costs are highly beneficial. They enable fast and efficient screenings. In this study, we developed an artificial neural network (ANN) based model for predicting the activity values of AcovPs based on sequences and structural information that extracted from three-dimensional structures. Independent testing results suggest that ACVPICPred achieves a decent prediction performance. ACVPICPred has the potential to reduce the time and resource costs associated with measuring the inhibition concentration of AcovPs.

Several computational tools have been developed to identify AcovPs. For example, PreAntiCoV employs a two-stage approach with balanced random forest to distinguish AcovPs from other peptides [26]. ENNAVIA-C and ENNAVIA-D differentiate AcovPs from non-antiviral peptides and random peptides using deep neural networks [43]. Iacvp leverages word2vec embedding methods to enhance feature encoding and combines with the random forest for model training [27]. ACP-Dnnel is an ensemble model that utilizes bi-directional LSTM as the base model for pre-training and employs deep convolutional neural networks for model construction [44]. The aforementioned models are binary classification models used to predict whether a given peptide exhibits anti-coronavirus function. ACVPICPred is a regression model that capable of estimating the inhibitory activity of AcovPs based on peptide sequences, experimentally relevant features and their structures. By evaluating the inhibition concentration, it becomes possible to preliminarily identify peptides with potent anti-coronaviral activity. Hence, the classification model can identify AcovPs and regression model can pinpoint more effective AcovPs within the identified set, facilitating a reduction in the experimentation scope and costing.

Currently, there is no predictor tailored for the prediction of the inhibitory effect of ACovPs. In 2015, a model known as AVP-IC_{50}Pred was developed to predict the antiviral activity of peptides [45]. While AVP-IC_{50}Pred was designed as a broad-spectrum predictor of antiviral peptide activity, its efficacy in predicting the specific inhibitory effects of ACovPs was found to be inferior to our proposed ACVPICPred, with a Pearson correlation coefficient of 0.3942 ($p > 0.05$) on our independent testing set. This disparity is primarily attributed to the variations in the training datasets. Human-infecting coronaviruses are notably constrained in number, with only seven known types to date [46]. Prior to the emergence of SARS-CoV-2 in 2019, the diversity and abundance of ACovPs are indeed limited. Consequently, the development of the AVP-IC_{50}Pred model in 2015 was constrained by a limited dataset of ACovPs, which is likely a primary factor contributing to its reduced effectiveness. Furthermore, our model offers an additional capability of predicting the targets of ACovPs, a feature that AVP-IC_{50}Pred lacks. There are also differences between the two models in terms of the features and machine learning methods employed, as delineated in Table 4. The pursuit of more comprehensive datasets and advanced machine learning techniques will be pivotal in enhancing the predictive accuracy and interpretability of future models.

The function of a peptide is determined by its structure. Peptides with similar sequences may exhibit significant structural variations, which may confer distinct functions. Therefore, relying solely on sequence-derived features is considered insufficient for comprehensive peptide characterization. In constructing our predictive model, we employed 2D-CNN to extract features from the structural information. This approach ensures a more thorough analysis of peptide characteristics and enhances the predictive capability of the model.

In the present study, we developed ACVPICPred, a tool designed to predict the inhibitory activity of ACovPs based on their sequences and structures. This model, however, exhibits certain limitations. First, we used an ANN to train the model. While ANNs have shown decent predictive performance in many fields [47], they operate as black-box models, which hinders the understanding of their internal mechanisms. Consequently, ACVPICPred predicts the activity of ACovPs without revealing the underlying virological processes. The development of interpretable ANNs remains an ongoing challenge, and we aim to investigate more transparent models in subsequent studies. Second, although we performed data augmentation, the size of the training data remains relatively limited, which may not be adequate for the training of more extensive networks. Future endeavors should prioritize expanding the dataset. By incorporating more experimental data and structural insights, ACVPICPred will be better equipped to capture the characteristics and mechanisms of peptides more accurately.

5. Conclusion

In this study, we developed an online computational tool called ACVPICPred (http://i.uestc.edu.cn/acvpICPred/main/Main.php) to predict the inhibitory activity of ACovPs. It is based on peptide sequences and structural information, trained using MLP and 2D-CNN, and achieves relatively high performance through cross-validation and independent test datasets. With this tool, researchers can quickly obtain predictions of ACovPs' inhibitory activity, which can reduce the time and resource investment in drug discovery for coronavirus-induced diseases and improve the initial screening efficiency of potential drug candidates for these diseases.

Table 4
Comparison of our model with AVP-IC_{50}Pred.

Method	Function	Training dataset	Features	Machine learning methods	Performance (Pearson correlation coefficient)[a]
AVP-IC_{50}Pred	Predict the inhibitory activity of antiviral peptides in terms of IC_{50} values (μM).	Antiviral peptides (683)	AAC, DPC, C8 Bin, N8 Bin, Physico, SA, SS	SVM, RF, IBk, K*	0.3942 ($p > 0.05$)
Our model (ACVPICPred)	(1) Predict the inhibitory activity of ACovPs in terms of IC_{50}, IC_{90} and EC_{50}. (2) Predict the target of ACovPs.	ACovPs (388 after data augmentation)	AAC, DPC, CKSAAGP, PAAC, Physico, SS, ASA, avg_rASA	hybrid model based on convolutional and fully connected layers	0.8836 ($p < 0.05$)

[a] Performance on the same independent testing dataset. ACovPs: anti-coronavirus peptides; IC_{50}: half-maximal inhibitory concentration; IC_{90}: 90 % inhibitory concentration; EC_{50}: half-maximal effective concentration; AAC: amino acid composition; DPC: dipeptide composition; C8 Bin: C8 Binary profile; N8 Bin: N8 Binary profile; CKSAAGP: composition of k-spaced amino acid group pairs; PAAC: pseudo amino acid composition; Physico: physicochemical properties; SA: solvent accessibility; SS: secondary structure; ASA: solvent-accessible surface area; avg_rASA: average relative solvent-accessible surface area; SVM: support vector machine; RF: random forest; IBk: instance-based classifier; K* : KStar.

CRediT authorship contribution statement

Jian Huang: Writing – review & editing, Supervision, Funding acquisition. **Bifang He:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Xing Shang:** Validation, Data curation. **Heng Chen:** Writing – review & editing, Supervision, Methodology, Funding acquisition, Conceptualization. **Chunying Lu:** Validation, Software. **Guifen Jian:** Validation, Investigation, Data curation. **Bowen Li:** Validation, Software, Formal analysis, Data curation. **Min Li:** Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Yifei Wu:** Validation, Data curation.

Declaration of Competing Interest

The authors declare that the research was conducted without any commercial or financial relationships that could be perceived as potential conflicts of interest.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant numbers: 62261006, 62263003, 62071099 and 82360340), Science and Technology Department of Guizhou Province (Grant numbers: ZK [2022]-general-056, and ZK [2022]-general-038) and Guizhou University (Grant number: [2020]5). At the same time, thanks for the computing support of the State Key Laboratory of Public Big Data, Guizhou University.

References

[1] Shahrajabian MH, Sun W, Cheng Q. Product of natural evolution (SARS, MERS, and SARS-CoV-2); deadly diseases, from SARS to SARS-CoV-2. Hum Vaccin Immunother 2021;17(1):62–83. https://doi.org/10.1080/21645515.2020.1797369.
[2] Wang Y, Grunewald M, Perlman S. Coronaviruses: an updated overview of their replication and pathogenesis. Methods Mol Biol 2020;2203:1–29. https://doi.org/10.1007/978-1-0716-0900-2_1.
[3] Hasoksuz M, Kilic S, Sarac F. Coronaviruses and SARS-COV-2. Turk J Med Sci 2020;50(SI-1):549–56. https://doi.org/10.3906/sag-2004-127.
[4] Organization W.H. (2024) COVID-19 epidemiological update – 13 August 2024. https://www.who.int/publications/m/item/covid-19-epidemiological-update-edition-170.
[5] Zumla A, Chan JF, Azhar EI, Hui DS, Yuen KY. Coronaviruses - drug discovery and therapeutic options. Nat Rev Drug Discov 2016;15(5):327–47. https://doi.org/10.1038/nrd.2015.37.
[6] Yuan Y, Jiao B, Qu L, Yang D, Liu R. The development of COVID-19 treatment. Front Immunol 2023;14:1125246. https://doi.org/10.3389/fimmu.2023.1125246.
[7] Lei X, Dong X, Ma R, Wang W, Xiao X, et al. Activation and evasion of type I interferon responses by SARS-CoV-2. Nat Commun 2020;11(1):3810. https://doi.org/10.1038/s41467-020-17665-9.
[8] Tong TR. Therapies for coronaviruses. Part I of II – viral entry inhibitors. Expert Opin Ther Pat 2009;19(3):357–67. https://doi.org/10.1517/13543770802609384.
[9] Li L, Zhang W, Hu Y, Tong X, Zheng S, et al. Effect of convalescent plasma therapy on time to clinical improvement in patients with severe and life-threatening COVID-19: a randomized clinical trial. JAMA 2020;324(5):460–70. https://doi.org/10.1001/jama.2020.10044.
[10] Yao TT, Qian JD, Zhu WY, Wang Y, Wang GQ. A systematic review of lopinavir therapy for SARS coronavirus and MERS coronavirus-a possible reference for coronavirus disease-19 treatment option. J Med Virol 2020;92(6):556–63. https://doi.org/10.1002/jmv.25729.
[11] Williamson BN, Feldmann F, Schwarz B, Meade-White K, Porter DP, et al. Clinical benefit of remdesivir in rhesus macaques infected with SARS-CoV-2. Nature 2020;585(7824):273–6. https://doi.org/10.1038/s41586-020-2423-5.
[12] Udwadia ZF, Singh P, Barkate H, Patil S, Rangwala S, et al. Efficacy and safety of favipiravir, an oral RNA-dependent RNA polymerase inhibitor, in mild-to-moderate COVID-19: a randomized, comparative, open-label, multicenter, phase 3 clinical trial. Int J Infect Dis 2021;103:62–71. https://doi.org/10.1016/j.ijid.2020.11.142.
[13] Lan Q, Yan Y, Zhang G, Xia S, Zhou J, et al. Clinical development of antivirals against SARS-CoV-2 and its variants. Curr Res Micro Sci 2024;6:100208. https://doi.org/10.1016/j.crmicr.2023.100208.
[14] Luo X, Chen H, Song Y, Qin Z, Xu L, et al. Advancements, challenges and future perspectives on peptide-based drugs: focus on antimicrobial peptides. Eur J Pharm Sci 2023;181:106363. https://doi.org/10.1016/j.ejps.2022.106363.
[15] Case JB, Chen RE, Cao L, Ying B, Winkler ES, et al. Ultrapotent miniproteins targeting the SARS-CoV-2 receptor-binding domain protect against infection and disease. e5 Cell Host Microbe 2021;29(7):1151–61. https://doi.org/10.1016/j.chom.2021.06.008.
[16] Aloul KM, Nielsen JE, Defensor EB, Lin JS, Fortkort JA, et al. Upregulating human cathelicidin antimicrobial peptide LL-37 expression may prevent severe COVID-19 inflammatory responses and reduce microthrombosis. Front Immunol 2022;13:880961. https://doi.org/10.3389/fimmu.2022.880961.
[17] Li Q, Zhao Z, Zhou D, Chen Y, Hong W, et al. Virucidal activity of a scorpion venom peptide variant mucroporin-M1 against measles, SARS-CoV and influenza H5N1 viruses. Peptides 2011;32(7):1518–25. https://doi.org/10.1016/j.peptides.2011.05.015.
[18] Lu L, Liu Q, Zhu Y, Chan KH, Qin L, et al. Structure-based discovery of Middle East respiratory syndrome coronavirus fusion inhibitor. Nat Commun 2014;5:3067. https://doi.org/10.1038/ncomms4067.
[19] Xia S, Yan L, Xu W, Agrawal AS, Algaissi A, et al. A pan-coronavirus fusion inhibitor targeting the HR1 domain of human coronavirus spike. Sci Adv 2019;5(4):eaav4580. https://doi.org/10.1126/sciadv.aav4580.
[20] Li B, Chen H, Huang J, He B. CD47Binder: identify CD47 binding peptides by combining next-generation phage display data and multiple peptide descriptors. Inter Sci 2023;15(4):578–89. https://doi.org/10.1007/s12539-023-00575-x.
[21] He B, Li B, Chen X, Zhang Q, Lu C, et al. PDL1Binder: identifying programmed cell death ligand 1 binding peptides by incorporating next-generation phage display data and different peptide descriptors. Front Microbiol 2022;13:928774. https://doi.org/10.3389/fmicb.2022.928774.
[22] Chen X, Huang J, He B. AntiDMPpred: a web service for identifying anti-diabetic peptides. PeerJ 2022;10:e13581. https://doi.org/10.7717/peerj.13581.
[23] Oeller M, Kang RJD, Bolt HL, Gomes Dos Santos AL, Weinmann AL, et al. Sequence-based prediction of the intrinsic solubility of peptides containing non-natural amino acids. Nat Commun 2023;14(1):7475. https://doi.org/10.1038/s41467-023-42940-w.
[24] Tallorin L, Wang J, Kim WE, Sahu S, Kosa NM, et al. Discovering de novo peptide substrates for enzymes using machine learning. Nat Commun 2018;9(1):5253. https://doi.org/10.1038/s41467-018-07717-6.
[25] Bhardwaj G, O'Connor J, Rettie S, Huang YH, Ramelot TA, et al. Accurate de novo design of membrane-traversing macrocycles. e26 Cell 2022;185(19):3520–32. https://doi.org/10.1016/j.cell.2022.07.019.
[26] Pang Y, Wang Z, Jhong JH, Lee TY. Identifying anti-coronavirus peptides by incorporating different negative datasets and imbalanced learning strategies. Brief Bioinform 2021;22(2):1085–95. https://doi.org/10.1093/bib/bbaa423.
[27] Kurata H, Tsukiyama S, Manavalan B. iACVP: markedly enhanced identification of anti-coronavirus peptides using a dataset-specific word2vec model. Brief Bioinform 2022;23(4). https://doi.org/10.1093/bib/bbac265.
[28] Zhang Q, Chen X, Li B, Lu C, Yang S, et al. A database of anti-coronavirus peptides. Sci Data 2022;9(1):294. https://doi.org/10.1038/s41597-022-01394-3.
[29] Jhong JH, Yao L, Pang Y, Li Z, Chung CR, et al. dbAMP 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. Nucleic Acids Res 2022;50(D1):D460–70. https://doi.org/10.1093/nar/gkab1080.
[30] Abramson J, Adler J, Dunger J, Evans R, Green T, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 2024;630. https://doi.org/10.1038/s41586-024-07487-w (Jun 13 Tn 8016).
[31] Kabsch WS C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22(12):2577–637. https://doi.org/10.1002/bip.360221211.
[32] Kuo-Chen C. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr Proteom 2009;6(4):262–74. https://doi.org/10.2174/157016409789973707.
[33] Chen Z, Zhao P, Li F, Leier A, Marquez-Lago TT, et al. iFeature: a python package and web server for features extraction and selection from protein and peptide sequences. Bioinformatics 2018;34(14):2499–502. https://doi.org/10.1093/bioinformatics/bty140.
[34] Guntuboina C., Das A., Mollaei P., Kim S., Farimani A.B. PeptideBERT: A Language Model Based on Transformers for Peptide Property Prediction.
[35] Branco P., Torgo L., Ribeiro R.P. (2017) SMOGN: a Pre-processing Approach for Imbalanced Regression.
[36] Liang X, Li F, Chen J, Li J, Wu H, et al. Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. Brief Bioinform 2021;22(4). https://doi.org/10.1093/bib/bbaa312.
[37] Zhang S, Zhu F, Yu Q, Zhu X. Identifying DNA-binding proteins based on multi-features and LASSO feature selection. Biopolymers 2021;112(2):e23419. https://doi.org/10.1002/bip.23419.
[38] Kerr WT, Anderson A, Xia H, Braun ES, Lau EP, et al. Parameter selection in mutual information-based feature selection in automated diagnosis of multiple epilepsies using scalp EEG. Int Workshop Pattern Recognit Neuroimaging 2012:45–8. https://doi.org/10.1109/PRNI.2012.27.
[39] Montesinos-Lopez OA, Crespo-Herrera L, Saint Pierre C, Bentley AR, de la Rosa-Santamaria R, et al. Do feature selection methods for selecting environmental covariables enhance genomic prediction accuracy? Front Genet 2023;14:1209275. https://doi.org/10.3389/fgene.2023.1209275.
[40] Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 2005;27(8):1226–38. https://doi.org/10.1109/TPAMI.2005.159.
[41] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc: Ser B (Methodol) 2018;58(1):267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[42] Parvandeh S, Yeh HW, Paulus MP, McKinney BA. Consensus features nested cross-validation. Bioinformatics 2020;36(10):3093–8. https://doi.org/10.1093/bioinformatics/btaa046.
[43] 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(6). https://doi.org/10.1093/bib/bbab258.
[44] Liu M, Liu H, Wu T, Zhu Y, Zhou Y, et al. ACP-dnnel: anti-coronavirus peptides' prediction based on deep neural network ensemble learning. Amino Acids 2023;55(9):1121–36. https://doi.org/10.1007/s00726-023-03300-6.
[45] 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(6):753–63. https://doi.org/10.1002/bip.22703.
[46] Tang G, Liu Z, Chen D. Human coronaviruses: origin, host and receptor. J Clin Virol 2022;155:105246. https://doi.org/10.1016/j.jcv.2022.105246.
[47] Talaei Khoei T, Ould Slimane H, Kaabouch N. Deep learning: systematic review, models, challenges, and research directions. Neural Comput Appl 2023;35(31):23103–24. https://doi.org/10.1007/s00521-023-08957-4.