Journal of proteome research

Article

Cite This: J. Proteome Res. XXXX, XXX, XXX−XXX

pubs.acs.org/jpr

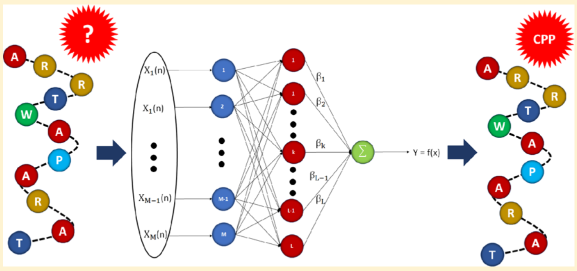
KELM-CPPpred: Kernel Extreme Learning Machine Based Prediction Model for Cell-Penetrating Peptides

Poonam Pandey,[†] Vinal Patel,[‡] Nithin V. George,[‡] and Sairam S. Mallajosyula*[,§]

[†]Department of Biological Engineering, Indian Institute of Technology Gandhinagar, Ahmedabad, Gujarat 382355, India
[‡]Department of Electrical Engineering, Indian Institute of Technology Gandhinagar, Ahmedabad, Gujarat 382355, India
[§]Department of Chemistry, Indian Institute of Technology Gandhinagar, Ahmedabad, Gujarat 382355, India

Supporting Information

ABSTRACT: Cell-penetrating peptides (CPPs) facilitate the transport of pharmacologically active molecules, such as plasmid DNA, short interfering RNA, nanoparticles, and small peptides. The accurate identification of new and unique CPPs is the initial step to gain insight into CPP activity. Experiments can provide detailed insight into the cell-penetration property of CPPs. However, the synthesis and identification of CPPs through wet-lab experiments is both resource- and time-expensive. Therefore, the development of an efficient prediction tool is essential for the identification of unique CPP prior to experiments. To this end, we developed a kernel extreme learning machine (KELM) based CPP prediction model called KELM-CPPpred. The main data set used in this study consists of 408 CPPs and an equal number of non-CPPs. The input features, used to train the proposed prediction model, include amino acid composition, dipeptide amino acid composition, pseudo amino acid composition, and the motif-based hybrid features. We further used an independent data set to validate the proposed model. In addition, we have also tested the prediction accuracy of KELM-CPPpred models with the existing artificial neural network (ANN), random forest (RF), and support vector machine (SVM) approaches on respective benchmark data sets used in the previous studies. Empirical tests showed that KELM-CPPpred outperformed existing prediction approaches based on SVM, RF, and ANN. We developed a web interface named KELM-CPPpred, which is freely available at http://sairam.people.iitgn.ac.in/KELM-CPPpred.html



KEYWORDS: *cell-penetrating peptides, kernel extreme learning machine, prediction server, machine learning, sequence-based prediction, feature vector, amino acid composition, dipeptide amino acid composition, pseudo amino acid composition, hybrid features*

INTRODUCTION

The intracellular delivery of a wide range of cargoes (i.e., small molecules, oligonucleotides, proteins, etc.) offers a great opportunity for future therapeutics.[1,2] Considerable progress has been made to design new techniques to improve the delivery of therapeutic compounds across the cell membrane.[3,4] In the past decade, protein transduction domains (PTDs) or cell-penetrating peptides (CPPs) have attracted substantial attention of the scientific community as potential drug-delivery candidates, facilitating the transport of pharmacologically active molecules, such as oligonucleotides,[5] plasmid DNA,[6] short interfering RNA (siRNA)[7], peptide nucleic acid[8] (PNA), peptides,[9,10] proteins,[11] and nanoparticles,[12] across the membrane.

CPPs are generally short peptides (fewer than 30 amino acids in length), derived from natural or synthetic proteins or chimeric peptides. In recent years, the number of known CPP sequences have increased rapidly, with new modified CPPs being developed to improve their stability and bioavailability. These novel CPPs are most often derived from the existing proteins and further optimized to be the shortest peptides having maximum transportation capability across the cell membrane. The wide occurrence of CPPs in different cell types and organisms depicts their biological significance in living organisms. CPPs are also involved in a wide range of pharmacological applications. Therefore, it is very crucial to understand their function and translocation mechanism. The accurate identification of CPPs is the primary step toward studying CPP translocation activity.

Although the experimental techniques can provide detailed insight into the translocation property of CPPs, these techniques are expensive and time-consuming. Over the years, with the increase in biological data for CPPs, many in silico approaches have emerged as an alternative approach to predict CPPs.[13−17] Here we briefly summarize the current progress in the in silico approaches for CPP prediction. In 2010, Dobchev et al. developed a CPP prediction model using artificial neural network (ANN) and principal component analysis (PCA), with an overall accuracy of 83.16%.[14] Later on,

Received: May 8, 2018
Published: July 22, 2018

DOI: 10.1021/acs.jproteome.8b00322

in 2011, Sanders et al. developed a prediction model using support vector machine (SVM) and a standard benchmark data set having 111 experimentally verified CPPs and 34 known non-CPPs with an overall accuracy of 75.9%.[15] From their study, it was observed that the accuracy obtained from the prediction model using balanced training data set overruled the accuracy obtained from the unbalanced training data set. In 2012, Gautam et al.[16] developed SVM-based prediction methods for CPP prediction, using several sequence-based features like amino acid composition (AAC), dipeptide amino acid composition (DAC), binary profile of patterns, and physicochemical properties, with maximum prediction accuracy of 97.40%.[16] In their study, it was found that a hybrid method, having input feature in combination with motif information, shows better accuracy then traditional methods. In the same year, Holton et al.[17] introduced the N-to-1 neural network-based prediction model with an accuracy of 82.98% for an independent test set. In 2015, Chen et al. developed an RF-based CPP prediction model using a pseudo amino acid composition (PseAAC)-based feature, with an accuracy of 83.5%.[18] Later on, in 2016, Tang et al. constructed an SVM-based prediction model using dipeptide composition optimized with variance-based technique with overall accuracy of 83.6%.[19] Recently, Wei et al. introduced a two layered prediction framework based on random forest (RF) algorithm with maximum accuracy of 91.6%.[20]

Although sufficient progress has been made to improve the prediction accuracy of existing CPP prediction model, there still exist some issues that need to be addressed. First, the feature representation capability has not been fully explored, which restrains the prediction capability of the existing models. In literature, most of the CPP prediction models have employed AAC, DAC, and physiochemical properties as an input feature. However, in recent studies, it has been demonstrated that classifiers based on hybrid feature showed superior prediction accuracy compared with individual-feature-based classifiers.[16,21] On the contrary, in different prediction models, PseAAC has been used to improve the prediction quality of protein characteristics, such as protein class,[22] protein subcellular localization,[23] enzyme family and subfamily class,[24] and post-translational modification site.[25] Considering these facts, in the current study, we intend to demonstrate the comparative analysis of the prediction model based on AAC, DAC, PseAAC, and their hybrid feature descriptors. Second, traditional classifiers, like ANN and SVM, have been extensively used for designing CPP prediction models.[13−17] The main challenges for practical application of ANN structure are the computational burden and suboptimal solutions due to back-propagation. Furthermore, the optimization of ANN structure and choice of learning parameters require intensive human intervene, as reported by Huang.[26,27] Similarly, in the case of SVM-based prediction models, it also requires excessive human interference for parameter tuning, and its computation time increases quadratically with the size of input feature vector.[27] Therefore, in this paper, we have designed a new prediction model based on extreme learning machine (ELM) structure, which overcomes the limitations of traditional predictors based on ANN and SVM architecture.

ELM has recently emerged as an efficient machine learning approach that has capability as both a universal classifier and an approximator.[28] The learning capability of ELM structures has more resemblance to the brain and requires less human interference compared with SVM and ANN. The ELM structure consists of one hidden layer and one output layer. The main advantages of ELM is that only the weights of the output layer need to be tuned, whereas the weights of the hidden layer can be chosen randomly.[29] Therefore, the training of ELM requires less time and can be achieved in a more efficient way than SVM and ANN. ELM has been successfully applied in diverse areas including pattern recognition,[27,30,31] classification, and regression.[32] In previous studies, it has been reported that ELM outperforms SVM on several standard classification and regression problems.[27,32] Recently, Huang et al.[27,33] proposed a kernel version of ELM called Kernel ELM, which further enhances the performance of ELM without increasing the computational cost. In this paper, we have made an attempt to develop a Kernel-ELM based prediction model for CPP. We call our model KELM-CPPpred, as an abbreviation for "**K**ernel **E**xtreme **L**earning **M**achine based **C**ell **P**enetrating **P**eptides **pred**iction" model. We have also carried out a comparative study for evaluating the performance of the proposed prediction model with existing models on benchmark data sets. Even for an unbalanced data set, which is a difficult task for all of the existing CPP prediction approaches, as reported by Sanders et al.,[15] the proposed prediction model outperforms existing CPP prediction models. On the basis of the study conducted in this paper, a user-friendly web server has been developed to help the researchers for predicting and designing CPPs (Supplementary Figure S2).

■ METHODS

To develop an efficient sequence-based statistical prediction model, the following five steps should be followed: (i) construct a reliable and stringent data set; (ii) map the peptide sequences to the fixed length numerical vector, which can be further used as a input feature vector in the prediction model; (iii) develop an efficient classifier algorithm; (iv) perform cross-validation for checking the reliability of the prediction model; and (v) develop a user-friendly application for the prediction model. In this section, we described steps i, iii, and iv in detail, whereas steps ii and v are described in the Supporting Information (SI).

Data Set Preparation

Main Data Set. A reliable and stringent data set is essential to construct and evaluate a statistical prediction model. In this respect, the positive samples for this study have been derived from the CPPsite-2.0 database.[34] The positive samples consist of 408 nonredundant peptides derived from the CPPsite 2.0 database. The negative samples have been generated from 34 experimentally validated non-CPPs from Sanders et al.,[15] supplemented with nonredundant peptides randomly selected from 374 bioactive peptides of 5−30 amino acids in length retrieved from BIOPEP[35] and CAMP[36] databases. The overall data set can be formulated as

$$D = \overset{+}{D} \cup \overset{-}{D}$$

where $\overset{+}{D}$ and $\overset{-}{D}$ refer to the subsets having CPPs (positive sample) and non-CPPs (negative sample), respectively. The symbol ∪ represents the "union" in the set theory. To construct a high-quality and nonredundant benchmark data set, the protein sequences obtained have been filtered to reduce redundancy using CD-HIT[37] with a threshold value 0.8.

Independent Data Set. To validate the performance of the proposed model, an independent data set was adapted

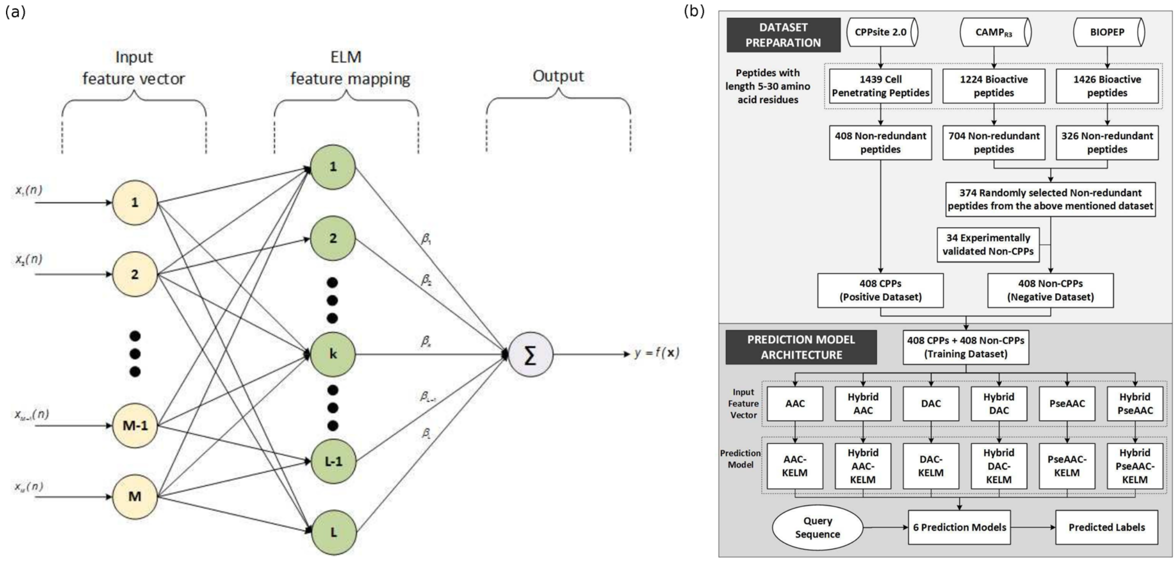


Figure 1. Proposed prediction model architecture. (a) Basic structure of Extreme Learning Machine (ELM) model for binary classification. (b) Overview of ELM-based cell-penetrating peptide prediction protocol.

from the study by Raghava-2013 et al.[16] In their study, Raghava-2013 et al. used an independent data set consisting of 99 CPPs and an equal number of non-CPPs. The CPP sequences were collected manually from research papers and patents, whereas the non-CPP sequences were extracted from SwissProt.[38] These sequences were not included in their training, feature selection, and model parameter optimization. This data set was compared with our main data set, which we used for developing the KELM-CPPpred model, and any sequences that showed >80% similarity were further excluded. The final independent data set consists of 96 CPPs and an equal number of non-CPPs.

Benchmark Data Sets. To compare the proposed method with existing methods, we extracted seven data sets from the previous studies: Hällbrink-2005,[39] Hansen-2008,[13] Dobchev-2010,[14] Sanders-2011(a,b,c),[15] Raghava-2013,[16] Chen-2015,[18] and Wei-2017.[20] The details of the data sets have been described in the Supporting Information.

Prediction Architecture: KELM-CPPpred Model

ELM has been proven to be a universal classifier that requires less human interference compared with SVM and ANN.[26,40] Here we apply the ELM architecture for the prediction of CPPs. Training of ELM requires fixed-length input feature vectors. Therefore, we employed AAC, DAC, PseACC, and their hybrid features as input vectors, which, in turn, were obtained from the peptide sequences of variable length.

Consider a data set $S = (\boldsymbol{x}_k, t_k)_{k=1}^{K}$, where $\boldsymbol{x}_k$ is the input feature vector of size $M \times 1$ derived from the protein/peptide sequence, t_k is the target vector of size 1×1, here $t_k = 1$ or 0 for CPP or non-CPP, respectively, and K is the number of elements in the data set. In an ELM, the input feature vector is nonlinearly mapped to an ELM feature space given by $\boldsymbol{h}(\boldsymbol{x}_k) = [h_1(\boldsymbol{x}_k), h_2(\boldsymbol{x}_k), ..., h_i(\boldsymbol{x}_k), ..., h_L(\boldsymbol{x}_k)]^{\mathrm{T}}$ of size $L \times 1$, as shown in Figure 1a.

Any nonlinear piece-wise continuous function can be used as the nonlinear mapping function $h_i(\cdot)$.[28] The output of the ELM is given by

$$f(\boldsymbol{x}_k) = \sum_{i=1}^{L} h_i(\boldsymbol{x}_k)\beta_i = \boldsymbol{h}^{\mathrm{T}}(\boldsymbol{p}_k)\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_i, ..., \beta_L]^{\mathrm{T}}$ is the weight vector corresponding to the output layer. In general, we can write

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{H\beta} \tag{2}$$

where $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_k(\boldsymbol{x}), ..., f_K(\boldsymbol{x})]^{\mathrm{T}}$ is the output vector and

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{h}^T(\boldsymbol{x}_1) \\ \boldsymbol{h}^T(\boldsymbol{x}_2) \\ \vdots \\ \boldsymbol{h}^T(\boldsymbol{x}_k) \\ \vdots \\ \boldsymbol{h}^T(\boldsymbol{x}_K) \end{bmatrix} = \begin{bmatrix} h_1(\boldsymbol{x}_1) & \cdots & h_L(\boldsymbol{x}_1) \\ \vdots & \vdots & \vdots \\ h_1(\boldsymbol{x}_k) & \cdots & h_L(\boldsymbol{x}_k) \\ \vdots & \vdots & \vdots \\ h_1(\boldsymbol{x}_K) & \cdots & h_L(\boldsymbol{x}_K) \end{bmatrix} \tag{3}$$

The output weight vector of the ELM can be estimated by minimizing the function given by

$$\xi = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \frac{C}{2}\sum_{k=1}^{K}\|\eta_k\|^2 \tag{4}$$

with respect to $\boldsymbol{\beta}$. In eq 4, $\|\cdot\|_2$ is the l_2 norm of $\boldsymbol{\beta}$, C is the regularization parameter tuned by the user, and $\eta_k = t_k - f(\boldsymbol{x}_k)$ is the learning error. The above-mentioned minimization is carried out subject to the condition that $\boldsymbol{h}(\boldsymbol{x}_k)\boldsymbol{\beta} = t_k - \eta_k$, and the solution is given by

$$\boldsymbol{\beta} = \boldsymbol{H}^T\left(\frac{1}{C}\boldsymbol{I}_K + \boldsymbol{H}\boldsymbol{H}^T\right)^{-1}\boldsymbol{T} \tag{5}$$

where $\boldsymbol{I}_K$ is an identity matrix of size $K \times K$ and $\boldsymbol{T} = [t_1, t_2, ..., t_K]^T$. Similar to SVM, we can use kernel function in ELM in place of feature vector $\boldsymbol{h}(\boldsymbol{x})$; this variant is called Kernel ELM (K-ELM).[33] In a K-ELM, the output is given by

DOI: 10.1021/acs.jproteome.8b00322

Table 1. Performance of KELM-CPPpred Model Prediction Model on Main Data Set[a]

	10-fold cross-validation				
input feature vector	sensitivity	specificity	accuracy (%)	MCC	AUC
AAC	80.60 (90.20)	91.84 (80.39)	86.36 (85.29)	0.73 (0.71)	0.91 (0.90)
hybrid AAC	81.88 (90.20)	91.84 (81.62)	86.98 (85.91)	0.74 (0.72)	0.92 (0.91)
DAC	80.72 (87.50)	89.21 (79.50)	85.20 (83.46)	0.71 (0.67)	0.91 (0.90)
hybrid DAC	81.22 (87.50)	89.21 (79.90)	85.44 (83.70)	0.71 (0.68)	0.91 (0.90)
PseAAC	85.63 (86.27)	87.62 (83.10)	86.64 (84.68)	0.73 (0.69)	0.92 (0.92)
hybrid PseAAC	85.63 (86.27)	87.62 (83.33)	86.64 (84.80)	0.73 (0.70)	0.92 (0.92)
average	82.61 (87.99)	89.56 (81.31)	86.21 (84.64)	0.73 (0.70)	0.92 (0.91)

[a] Sensitivity, specificity, accuracy, MCC, and AUC values obtained from jackknife validation are given in brackets.

$$f(\boldsymbol{x}) = [\kappa(\boldsymbol{x}, \boldsymbol{x}_1), \kappa(\boldsymbol{x}, \boldsymbol{x}_2), ..., \kappa(\boldsymbol{x}, \boldsymbol{x}_k), ..., \kappa(\boldsymbol{x}, \boldsymbol{x}_K)]^T \left(\frac{1}{C}\boldsymbol{I}_K + \boldsymbol{H}\boldsymbol{H}^T\right)^{-1} \boldsymbol{T} \tag{6}$$

where $\kappa(\boldsymbol{x},\boldsymbol{x}_k)$ is the kernel function. We have used a Gaussian kernel $\kappa(\boldsymbol{x},\boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{y}\|)$ in our study, where γ is the width of the Gaussian kernel. The proposed KELM-based CPP prediction model consists of one input layer, one hidden layer, and one output layer. The weights between the input and hidden layer are randomly assigned, whereas the weights (β) for output layer have to be calculated using eq 5. We have employed a grid search approach for finding the optimal value of regularization parameter (C) and the bandwidth (γ) of the Gaussian kernel in the range from [10^{-3} to 1000] to achieve the best performance. The overall proposed framework for CPP prediction is shown in Figure1b. As stated by ELM theory, any nonlinear piece-wise continuous function can be used as the nonlinear mapping function $h(\cdot)$ so that ELM can approximate any continuous target functions. However, in the case of SVM, the feature mapping is usually unknown, and not every feature mapping used by SVM can lead to a universal approximation. Our results show that the KELM model shows the highest prediction accuracy when compared with ANN- and SVM-based prediction models due to it is universal approximation capability.

Evaluation of Prediction Performance

The performance of the proposed prediction model is evaluated using the three most conventional approaches in statistical prediction methods:[41−43] 10-fold cross validation, jackknife test, and independent data set test. In 10-fold cross-validation, peptide sequences are divided into 10 subsets; at each time, 9 subsets are used to train the model and one remaining subset is used to test the model. This process is repeated 10 times so that each fold is used once as the test set. In the jackknife test, for each iteration, a single protein sequence is used as a testing sample, whereas all of the other sequences are used to train the model. In the independent data set test, the prediction models were trained using main data set, and the prediction was made for peptide sequences in the independent data set.

Five quality indices have been used to validate the proposed model

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{7}$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{8}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \times 100 \tag{9}$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \times 100 \tag{10}$$

where TP, TN, FP, FN, and MCC denote true-positives (correctly predicted CPPs), true-negatives (correctly predicted non-CPPs), false-positives (non-CPPs that are incorrectly predicted as CPPs), false-negatives (CPPs that are incorrectly predicted as non-CPPs), and Matthews correlation coefficient, respectively. We have also measured the area under curve (AUC) for the ROC plots to evaluate the performance of prediction models.

■ RESULTS AND DISCUSSION

Amino Acid Composition and Physicochemical Property Analysis

To evaluate the distribution of amino acids in CPPs as compared with non-CPPs, an AAC analysis was performed for both of the peptide classes. The amino acid compositional analysis revealed that the average occurrence of positively charged amino acids (arginine (R), histidine (H), and lysine (K)) is higher in CPPs as compared with non-CPPs (Supplementary Figure S3a). In non-CPPs tiny amino acids (alanine (A), glycine (G), cysteine (C), and serine (S)) and aliphatic amino acids (isoleucine (I), leucine (L), and valine (V)) are frequent (Supplementary Figure S3b). This suggests a preferential occurrence of certain amino acids in CPPs. In the same way, the dipeptide compositional analysis was performed for both the peptide classes, and 75 dipeptides were found to differ significantly in CPPs as compared with non-CPPs (Supplementary Figure S3c). Out of the 75 significantly different dipeptides (p value ≤0.005, Welch's t test), CPPs were found to be rich in leucine−leucine (L−L), leucine−serine (L−S), serine−leucine (S−L), serine−serine (S−S), alanine−serine (A−S), and serine−alanine (S−A) dipeptides.

Motif Analysis

The identification of functional motifs in peptide sequences is a key technique for the functional annotation of proteins. To identify the motifs present in CPPs, the positive training data set was analyzed using MERCI[44] software. The overall coverage of motifs represents the number of CPPs having that particular motif. A total of 13 motifs are identified using the Betts and Russell algorithm[45] (Supplementary Table S2). In the CPP data set, the most frequent amino acid motifs are RRRRRR, RRA, GRRX (where X = R, W, T), RRGRX (X = R, G, T), and KKRK.

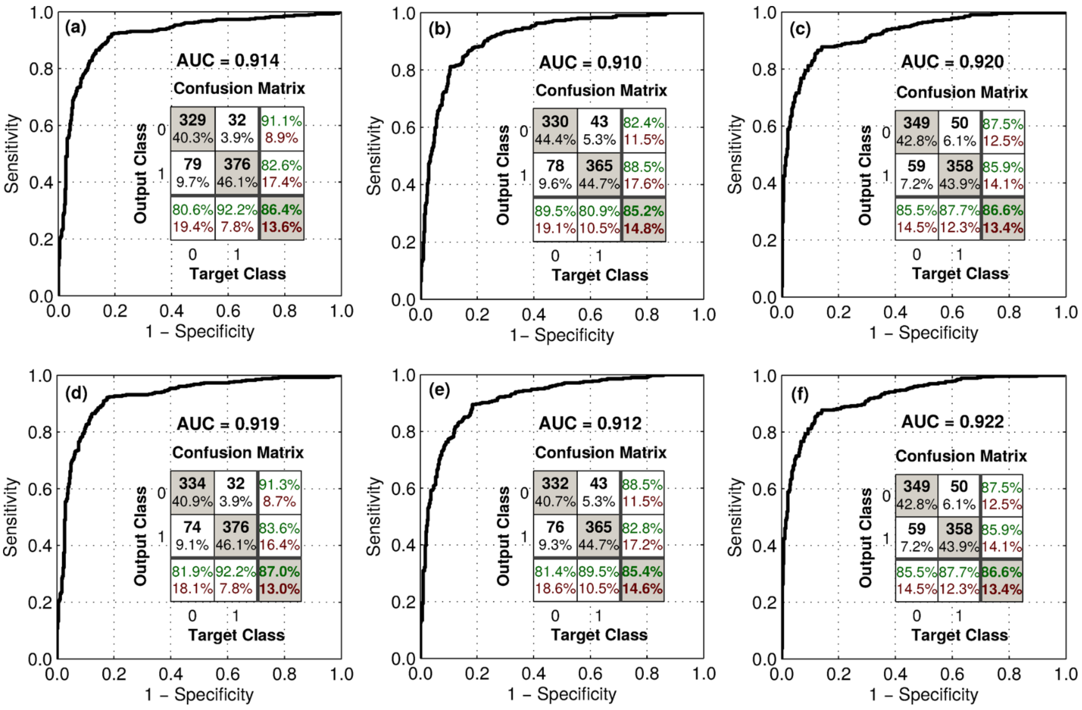


Figure 2. ROC curve for the proposed model based on (a) AAC, (b) DAC, (c) PseAAc, (d) hybrid AAC, (e) hybrid DAC, and (f) hybrid PseAAC method.

Machine-Learning-Based Classification Model

On the basis of compositional analysis, it is clear that CPPs and non-CPPs differ in AAC-based features. Hence sequence-driven features can be further exploited to develop a machine-learning-based classifier. Several machine-learning-based classifiers such as SVM,[15,16] RF,[18] and NN[17,46] have been employed for CPP prediction. The current study is divided into two parts. In the first part, we used the data set prepared by us to classify a given sequence as CPP or non-CPP based on various feature vectors and their hybrid implementation (Table 1). We also validate the prediction accuracy of the proposed model using the independent data set (Table 2). In the second part of the study, we compared the KELM-CPPpred model with previous classification approaches (Table 3). For the comparative study, we used the benchmark data sets as described in the SI (Supplementary Table S1).

AAC-Based KELM-CPPpred Model. AAC analysis showed that CPPs and non-CPPs have significant compositional differences. Hence these features can be utilized to predict a given peptide to be CPP or non-CPP using the machine-learning approach. The KELM-CPPpred model with C (4.5) and γ (0.009) parameters produces the best prediction model in the AAC-based prediction approach. It showed mean accuracy of 86.36% (MCC = 0.73, AUC = 0.91) (Table 1).

DAC-Based KELM-CPPpred Model. On the basis of compositional analysis, DAC was also used to build the predictive model by using a 400 sized vector having all of the possible combinations of dipeptides. The KELM model with C (9.9) and γ (0.06) parameters produces the best prediction model in a DAC-based prediction approach. It showed accuracy of 85.20% (MCC = 0.71, AUC = 0.91) (Table 1).

PseAAC-Based KELM-CPPpred Model. To include the effect of physicochemical properties along the local sequence order, pseudo amino acid composition (PseAAC) was used to build the prediction model. The KELM-CPPpred model with C (1.8) and γ (0.07) parameters produces the best prediction model in the PseAAc-based prediction approach. It showed maximum accuracy of 86.64% (MCC = 0.73, AUC = 0.92) (Table 1).

Hybrid Prediction Model. To improve the performance of KELM-CPPpred model, the above given features (AAC, DAC, and PseAAC) were further used in the combination with unique motif features to build a hybrid predictive model. The hybrid AAC-based KELM-CPPpred model showed an accuracy of 86.98% (MCC = 0.74, AUC = 0.92), whereas hybrid DAC-based KELM-CPPpred model showed an accuracy of 85.44% (MCC = 0.71, AUC = 0.91) (Table 1). The PseAAC-based hybrid KELM-CPPpred model performed similar to the normal PseAAC-based KELM-CPPpred model (Table 1).

Receiver Operating Characteristic Curve Analysis. To provide visual comparison of the proposed model for different input feature vectors, we further performed a graphical analysis using the receiver operating characteristic (ROC) curve.[47] ROC curve is a plot of "1-specificity" (X axis) versus "sensitivity" (Y axis). It represents a reasonable trade-off between false-positive and true-positive rates corresponding to a particular threshold value. Figure 2 shows the ROC curve for the proposed model based on (a) AAC, (b) hybrid AAC, (c) DAC, (d) hybrid DAC, (e) PseAAC, and (f) hybrid PseAAC method. It can be noted from Figure 2 that AUC of AAC (Figure 2a) is nearly equal to the DAC (Figure 2b), while AUC for PseAAC (Figure 2c) is significantly better then AAC and DAC. The same trend is observed for the hybrid method (Figure 2d–f). It can also be observed that the performance of hybrid models (Figure 2d–f) is slightly better than the traditional feature-based model (Figure 2a–c).

Performance of KELM-CPPpred Model on Independent Data Set

To validate the proposed approach, the KELM-CPPpred prediction model was also evaluated on an independent data set for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC (Table 2). The ROC analysis for the same is illustrated in Supplementary Figure S4. For AAC-based features, KELM-CPPpred showed accuracy of 84.40% (MCC = 0.69, AUC = 0.89). For DAC-based features, KELM-CPPpred showed accuracy of 79.20% (MCC = 0.61, AUC = 0.88). For PseAAC-based features, KELM-CPPpred showed accuracy of 82.30% (MCC = 0.66, AUC = 0.92). For Hybrid-AAC-based features, KELM-CPPpred showed accuracy of 87.00% (MCC = 0.73, AUC = 0.92). For Hybrid-DAC-based features, KELM-CPPpred showed accuracy of 81.80% (MCC = 0.66, AUC = 0.90). For Hybrid-PseAAC-based features, KELM-CPPpred showed accuracy of 83.90% (MCC = 0.68, AUC = 0.93). To test the quality of proposed prediction model, jackknife test is also performed to evaluate the five quality indices, and we found the comparable results to 10-fold cross-validation, as given in Table 2. Therefore, for the further studies on CPP prediction, which include comparison with state-of-the-art methods, we have used 10-fold cross-validation only because it is less expensive in terms of computational cost as compared with the jackknife test.

Table 2. Performance of KELM-CPPpred Model Prediction Model on Independent Data Set

	independent data set				
input feature vector	sensitivity	specificity	accuracy (%)	MCC	AUC
AAC	74.06	93.54	84.40	0.69	0.89
hybrid AAC	79.17	93.54	87.00	0.73	0.92
DAC	75.31	85.52	79.20	0.61	0.88
hybrid DAC	79.90	85.52	81.80	0.66	0.90
PseAAC	80.63	85.10	82.30	0.66	0.92
hybrid PseAAC	83.23	85.10	83.90	0.68	0.93
average	78.72	88.05	83.10	0.67	0.91

Independent Data Set Validation Test. The KELM-CPPpred model exhibits an average accuracy of 86.21 and 83.10% for the 10-fold cross-validation and for the independent data set, respectively, demonstrating the efficiency of the proposed model.

Comparison with Existing Methods

To evaluate the performance of the proposed KELM-CPPpred prediction model, a comparative study has been carried out with the existing prediction structure for benchmark data sets and enumerated in Table 3. In 2005, Hällbrink et al. used bulk property values and descriptor scales to predict CPPs. The prediction accuracy of KELM-CPPpred model on the Hällbrink data set[39] is 15.4, 13.4, 16.06, 16.06, 13.73, and 16.06% higher than Hällbrink's method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. Later on, in 2008, Hansen et al. used z-scale values, previously published by Sandberg et al. to predict CPPs. The prediction accuracy of KELM-CPPpred model on the Hansen et al. data set[13] showed 14.48, 18.04, 16.98, 14.48, 18.04, and 16.98% improvement over the Hansen et al. method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. In 2010, Dobchev et al. employed ANN-based prediction model based on molecular descriptors. Feature selection was performed using principle component analysis (PCA). The prediction accuracy of KELM-CPPpred model on the Dobchev et al. data set[14] showed 1.55, 2.13, 2.72, 1.55, 2.13, and 2.72% improvement over the Dobchev et al. method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. Sanders et al. used the SVM-based classification model for three types of data set based on a set of physiochemical properties of amino acids (e.g., charge, molecular weight, secondary structure, etc.). Comparison results, as shown in Table 3, depict that the prediction performance of our proposed model is more superior to the Sanders et al. method for both the balanced and unbalanced data sets. For the Sanders-2011(a) data set,[15] which consists of 111 CPPs and 34 known non-CPPs, the prediction accuracy of our proposed model is 0.5, 1.37, 1.33, 1.83, 1.37, and 1.91% higher than Sander's method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively.

For the Sanders-2011(b) data set,[15] which consists of 111 CPPs and 111 non-CPPs, the prediction accuracy of our proposed model is 7.94, 7.79, 9.33, 7.94, 7.78, and 10.04% higher than Sander's method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. For the Sanders-2011(c) data set,[15] which consists of 111 CPPs and 111 randomly selected non-CPPs from 34 experimentally validated non-CPPs, the prediction

Table 3. Comparison of KELM-CPPpred Prediction Model with Existing Prediction Methods

	accuracy (%)						
		proposed model					
benchmark data set	existing method	AAC	DAC	PseAAC	hybrid AAC	hybrid DAC	hybrid PseAAC
Hällbrink-2005[39]	77.27	92.67	90.67	93.33	93.33	91.00	93.33
Hansen-2008[13]	67.44	81.92	85.48	84.42	81.92	85.48	84.42
Dobchev-2010[14]	83.16	84.71	85.29	85.88	84.71	85.29	85.88
Sanders-2011(a)[15]	95.94	96.44	97.31	97.27	97.77	97.31	97.85
Sanders-2011(b)[15]	75.86	83.80	83.65	85.19	83.80	83.65	85.90
Sanders-2011(c)[15]	88.73	90.53	92.31	93.71	90.08	92.31	93.71
Raghava-2013(a)[16]	90.75	90.66	91.37	91.60	91.02	94.62	95.73
Raghava-2013(b)[16]	92.98	92.51	94.35	93.30	92.78	94.62	93.30
Raghava-2013(c)[16]	68.98	69.80	67.67	70.36	70.33	68.21	70.64
Chen-2015[18]	83.45	83.80	83.65	85.19	83.80	83.65	85.90
Wei-2017[20]	90.60	91.03	89.84	91.66	91.03	89.84	91.66

accuracy of our proposed model is 1.80, 3.58, 4.98, 1.35, 3.58, and 4.98% higher than Sander's method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. Raghava et al. also used SVM-based classification model for the three types of data sets (mentioned in Supporting material S1 of Supporting Information), using AAC, DAC, binary profiles, and physicochemical properties as input feature vector. For the Raghava-2013(a) data set,[16] which consists of 708 CPPs and 708 non-CPPs, the prediction accuracy of our proposed model is comparable for the AAC feature vector and is showing 0.62, 0.85, 0.27, 3.87, and 4.98% improvement over Raghava's method for DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. For the Raghava-2013(b) data set,[16] which consists of 187 CPPs having high uptake efficiency and 187 non-CPPs, the prediction accuracy of our proposed model is comparable for AAC and hybrid-AAC input feature vector, whereas the accuracy of KELM-CPPpred is 1.37, 0.32, 0.27, and 3.87% higher than Raghava et al.'s method for DAC, PseAAC, hybrid-DAC, and hybrid-PseAAC feature input vectors, respectively. For the Raghava-2013(c) data set,[16] which consists of 187 CPPs having high uptake efficiency as positive data set and 187 CPPs having low uptake efficiency as negative data set, the prediction accuracy of our proposed model is comparable for DAC and hybrid-DAC input feature vector. However, the prediction accuracy of proposed KELM-CPPpred model is 0.82, 1.38, 1.35, and 1.68% higher than Raghava's method for AAC, PseAAC, hybrid-AAC, and hybrid-PseAAC input feature vectors, respectively. For the Chen-2015 data set,[18] which consists of 111 CPPs and 34 non-CPPs, the prediction accuracy of our model is 0.35, 0.2, 1.74, 0.35, 0.2, and 2.45% higher than Chen's method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. For the Wei-2017 data set,[20] which consist of 462 CPPs and 462 non-CPPs, the prediction accuracy of our model is 0.43, 1.06, 0.43, and 1.06% higher than Wei's method for AAC, DAC, PseAAC, hybrid-AAC, hybrid-DAC, and hybrid-PseAAC feature vectors, respectively. However, the prediction accuracy of proposed KELM-CPPpred model is comparable for DAC and hybrid-DAC input feature vector.

To further demonstrate the effectiveness of the KELM-CPPpred model, 33 known CPPs extracted from the independent data set were tested on three existing web servers, CellPPD,[16] CPPpred,[17] and SkipCPP-Pred,[20] and KELM-CPPpred. The sequence information and prediction results are presented in Supplementary Table S3. We observe that statistically, KELM-CPPpred and SkipCPP-Pred perform better than others, highlighting the improvements made to the prediction models.

CONCLUSIONS

The development of a computational prediction model for CPPs is highly challenging due to the following three reasons: (i) high variation in the length of CPPs (5 to 30 amino acids), (ii) small number of experimentally verified non-CPPs (34 experimentally verified non-CPPs), and (iii) variable experimental condition (concentration, cell lines, etc.) for experimentally validated CPPs and non-CPPs. Hence in our prediction model we have used a larger data set that consists of 408 nonredundant CPPs obtained from the CPPsite 2.0 database[34] and an equal number of non-CPPs generated from 34 experimentally validated non-CPPs supplemented with nonredundant peptides randomly selected from BIOPEP[35] and CAMP[36] databases.

In this paper, we demonstrated a KELM-based CPP prediction model, which offers higher prediction accuracy compared with the other existing prediction model. In this prediction approach, AAC, DAC, PseAAC, and their CPP motif-based hybrid features were used to map the amino acid sequences to the respective numeric feature vector, which were further used as an input in KELM-CPPpred model. In this study, the proposed prediction model achieved better prediction accuracy and required less tuning of parameter as compared with ANN-,[14,17] random-forest-,[18,20] and SVM-[15,16] based prediction methods. Even for the unbalanced data set, the KELM-CPPpred model outperformed the existing prediction models. To serve the research community, we have developed a web application for CPP prediction using the proposed KELM-CPPpred prediction model. The application is freely available for users at http://sairam.people.iitgn.ac.in/KELM-CPPpred.html. This application will help researchers in designing and predicting CPPs with much ease and better accuracy.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.8b00322.

Supporting Material S1: Benchmark data set information. Supporting Material S2: Feature extraction information. Supporting Material S3: Web application development. Supporting Figure S1: Schematic representation of sequence order correlation mode used in PseAAC feature vector. Supporting Figure S2: Developed web application KELM-CPPpred for CPP prediction. Supporting Figure S3: Amino acid composition and physicochemical analysis of CPPs and non-CPPs. Supporting Figure S4: ROC curve analysis for the proposed model for independent data set. Supporting Table S1: Summary of five benchmark data sets used for comparative analysis. Supporting Table S2: Annotation and coverage of MERCI motifs extracted from CPPs. Supporting Table S3: Comparison of the efficiency of the KELM-CPPpred server with existing servers. (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: msairam@iitgn.ac.in. Tel: +91-79-32454998. Fax: +91-79-2397 2324.

ORCID

Sairam S. Mallajosyula: 0000-0002-6825-0378

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by Indian Institute of Technology Gandhinagar. P.P. and V.P. acknowledge fellowship from Ministry of Human Resource Development (MHRD), Government of India.

REFERENCES

(1) Gao, X.; Kim, K.-S.; Liu, D. Nonviral gene delivery: what we know and what is next. *AAPS J.* **2007**, *9*, E92−E104.

(2) Walther, W.; Stein, U. Viral vectors for gene transfer a review of their use in the treatment of human disease. *Drugs* **2000**, *60*, 249−71.

(3) Järver, P.; Langel, Ü. The use of cell-penetrating peptides as a tool for gene regulation. *Drug Discovery Today* **2004**, *9*, 395−402.

(4) Glover, D. J.; Lipps, H. J.; Jans, D. A. Towards safe, non-viral therapeutic gene expression in humans. *Nat. Rev. Genet.* **2005**, *6*, 299−310.

(5) Margus, H.; Padari, K.; Pooga, M. Cell-penetrating peptides as versatile vehicles for oligonucleotide delivery. *Mol. Ther.* **2012**, *20*, 525−533.

(6) Lehto, T.; Kurrikoff, K.; Langel, Ü. Cell-penetrating peptides for the delivery of nucleic acids. *Expert Opin. Drug Delivery* **2012**, *9*, 823−836.

(7) Presente, A.; Dowdy, S. F. PTD/CPP peptide-mediated delivery of siRNAs. *Curr. Pharm. Des.* **2013**, *19*, 2943−2947.

(8) Bendifallah, N.; Rasmussen, F. W.; Zachar, V.; Ebbesen, P.; Nielsen, P. E.; Koppelhus, U. Evaluation of cell-penetrating peptides (CPPs) as vehicles for intracellular delivery of antisense peptide nucleic acid (PNA). *Bioconjugate Chem.* **2006**, *17*, 750−758.

(9) Hansen, A.; Schäfer, I.; Knappe, D.; Seibel, P.; Hoffmann, R. Intracellular toxicity of proline-rich antimicrobial peptides shuttled into mammalian cells by the cell-penetrating peptide penetratin. *Antimicrob. Agents Chemother.* **2012**, *56*, 5194−5201.

(10) Boisguerin, P.; Giorgi, J.-M.; Barrère-Lemaire, S. CPP-conjugated anti-apoptotic peptides as therapeutic tools of ischemiar-eperfusion injuries. *Curr. Pharm. Des.* **2013**, *19*, 2970−2978.

(11) Nasrollahi, S. A.; Fouladdel, S.; Taghibiglou, C.; Azizi, E.; Farboud, E. S. A peptide carrier for the delivery of elastin into fibroblast cells. *Int. J. Dermatol.* **2012**, *51*, 923−929.

(12) Xia, H.; Gao, X.; Gu, G.; Liu, Z.; Hu, Q.; Tu, Y.; Song, Q.; Yao, L.; Pang, Z.; Jiang, X.; et al. Penetratin-functionalized PEG-PLA nanoparticles for brain drug delivery. *Int. J. Pharm.* **2012**, *436*, 840−850.

(13) Hansen, M.; Kilk, K.; Langel, Ü. Predicting cell-penetrating peptides. *Adv. Drug Delivery Rev.* **2008**, *60*, 572−579.

(14) Dobchev, D. A.; Mager, I.; Tulp, I.; Karelson, G.; Tamm, T.; Tamm, K.; Janes, J.; Langel, U.; Karelson, M. Prediction of cell-penetrating peptides using artificial neural networks. *Curr. Comput.-Aided Drug Des.* **2010**, *6*, 79−89.

(15) Sanders, W. S.; Johnston, C. I.; Bridges, S. M.; Burgess, S. C.; Willeford, K. O. Prediction of cell penetrating peptides by support vector machines. *PLoS Comput. Biol.* **2011**, *7*, e1002101.

(16) Gautam, A.; Chaudhary, K.; Kumar, R.; Sharma, A.; Kapoor, P.; Tyagi, A.; Raghava, G. P. In silico approaches for designing highly effective cell penetrating peptides. *J. Transl. Med.* **2013**, *11*, 74.

(17) Holton, T. A.; Pollastri, G.; Shields, D. C.; Mooney, C. CPPpred: prediction of cell penetrating peptides. *Bioinformatics* **2013**, *29*, 3094.

(18) Chen, L.; Chu, C.; Huang, T.; Kong, X.; Cai, Y.-D. Prediction and analysis of cell-penetrating peptides using pseudo-amino acid composition and random forest models. *Amino Acids* **2015**, *47*, 1485−1493.

(19) Tang, H.; Su, Z.-D.; Wei, H.-H.; Chen, W.; Lin, H. Prediction of cell-penetrating peptides with feature selection techniques. *Biochem. Biophys. Res. Commun.* **2016**, *477*, 150−154.

(20) Wei, L.; Tang, J.; Zou, Q. SkipCPP-Pred: an improved and promising sequence-based predictor for predicting cell-penetrating peptides. *BMC Genomics* **2017**, *18*, 1.

(21) Sui, T.; Yang, Y.; Wang, X. Sequence-based feature extraction for type III effector prediction. *Int. J. Biosci., Biochem. Bioinf.* **2013**, *3*, 246.

(22) Chou, K.-C.; Shen, H.-B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. *Biochem. Biophys. Res. Commun.* **2007**, *360*, 339−345.

(23) Gao, Y.; Shao, S.; Xiao, X.; Ding, Y.; Huang, Y.; Huang, Z.; Chou, K.-C. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. *Amino Acids* **2005**, *28*, 373−376.

(24) Chou, K.-C. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics* **2005**, *21*, 10−19.

(25) Xu, Y.; Ding, Y.-X.; Ding, J.; Lei, Y.-H.; Wu, L.-Y.; Deng, N.-Y. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. *Sci. Rep.* **2015**, *5*, 10184.

(26) Huang, G.-B. What are extreme learning machines? Filling the gap between Frank Rosenblattas dream and John von Neumannas puzzle. *Cognitive Computation* **2015**, *7*, 263−278.

(27) Huang, G.-B.; Zhou, H.; Ding, X.; Zhang, R. Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* **2012**, *42*, 513−529.

(28) Huang, G.-B.; Zhu, Q.-Y.; Siew, C.-K. Extreme learning machine: theory and applications. *Neurocomputing* **2006**, *70*, 489−501.

(29) Tang, J.; Deng, C.; Huang, G.-B. Extreme learning machine for multilayer perceptron. *IEEE transactions on neural networks and learning systems* **2016**, *27*, 809−821.

(30) Mohammed, A. A.; Minhas, R.; Wu, Q.-M. J.; Sid-Ahmed, M. A. Human face recognition based on multidimensional PCA and extreme learning machine. *Pattern Recognition* **2011**, *44*, 2588−2597.

(31) Zong, W.; Huang, G.-B. Face recognition based on extreme learning machine. *Neurocomputing* **2011**, *74*, 2541−2551.

(32) Rong, H.-J.; Ong, Y.-S.; Tan, A.-H.; Zhu, Z. A fast pruned-extreme learning machine for classification problem. *Neurocomputing* **2008**, *72*, 359−366.

(33) Scardapane, S.; Comminiello, D.; Scarpiniti, M.; Uncini, A. Online sequential extreme learning machine with kernels. *IEEE transactions on neural networks and learning systems* **2015**, *26*, 2214−2220.

(34) Agrawal, P.; Bhalla, S.; Usmani, S. S.; Singh, S.; Chaudhary, K.; Raghava, G. P.; Gautam, A. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. *Nucleic Acids Res.* **2016**, *44*, D1098−D1103.

(35) Iwaniak, A.; Minkiewicz, P.; Darewicz, M.; Sieniawski, K.; Starowicz, P. BIOPEP database of sensory peptides and amino acids. *Food Res. Int.* **2016**, *85*, 155−161.

(36) Waghu, F. H.; Gopi, L.; Barai, R. S.; Ramteke, P.; Nizami, B.; Idicula-Thomas, S. CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.* **2014**, *42*, D1154−D1158.

(37) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150−3152.

(38) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **2000**, *28*, 45−48.

(39) Hällbrink, M.; Kilk, K.; Elmquist, A.; Lundberg, P.; Lindgren, M.; Jiang, Y.; Pooga, M.; Soomets, U.; Langel, Ü. Prediction of cell-penetrating peptides. *Int. J. Pept. Res. Ther.* **2005**, *11*, 249−259.

(40) Şahin, M. Comparison of modelling ANN and ELM to estimate solar radiation over Turkey using NOAA satellite data. *International journal of remote sensing* **2013**, *34*, 7508−7533.

(41) Chou, K.-C.; Zhang, C.-T. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275−349.

(42) Chou, K.-C.; Cai, Y.-D. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* **2002**, *277*, 45765−45769.

(43) Xing, P.; Su, R.; Guo, F.; Wei, L. Identifying N 6-methyladenosine sites using multi-interval nucleotide pair position specificity and support vector machine. *Sci. Rep.* **2017**, *7*, 46757.

(44) Vens, C.; Rosso, M.-N.; Danchin, E. G. Identifying discriminative classification-based motifs in biological sequences. *Bioinformatics* **2011**, *27*, 1231−1238.

DOI: 10.1021/acs.jproteome.8b00322

(45) Betts, M. J.; Russell, R. B. Amino acid properties and consequences of substitutions. *Bioinformatics for geneticists* **2003**, *317*, 289.

(46) Karelson, M.; Dobchev, D. Using artificial neural networks to predict cell-penetrating compounds. *Expert Opin. Drug Discovery* **2011**, *6*, 783−796.

(47) Hanley, J. A.; McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29−36.

DOI: 10.1021/acs.jproteome.8b00322