diagnostics

Article

A First Computational Frame for Recognizing Heparin-Binding Protein

Wen Zhu [1,2,3], Shi-Shi Yuan [4], Jian Li [5], Cheng-Bing Huang [6], Hao Lin [4,*] and Bo Liao [1,2,3,*]

[1] Key Laboratory of Computational Science and Application of Hainan Province, Haikou 571158, China; syzhuwen@163.com
[2] Key Laboratory of Data Science and Intelligence Education, Hainan Normal University, Ministry of Education, Haikou 571158, China
[3] School of Mathematics and Statistics, Hainan Normal University, Haikou 571158, China
[4] School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 611731, China; 202211140611@std.uestc.edu.cn
[5] School of Basic Medical Sciences, Chengdu University, Chengdu 610106, China; lijian01@cdu.edu.cn
[6] School of Computer Science and Technology, ABa Teachers University, Chengdu 623002, China; 20049607@abtu.edu.cn
* Correspondence: hlin@uestc.edu.cn (H.L.); dragonbw@163.com (B.L.)

Abstract: Heparin-binding protein (HBP) is a cationic antibacterial protein derived from multinuclear neutrophils and an important biomarker of infectious diseases. The correct identification of HBP is of great significance to the study of infectious diseases. This work provides the first HBP recognition framework based on machine learning to accurately identify HBP. By using four sequence descriptors, HBP and non-HBP samples were represented by discrete numbers. By inputting these features into a support vector machine (SVM) and random forest (RF) algorithm and comparing the prediction performances of these methods on training data and independent test data, it is found that the SVM-based classifier has the greatest potential to identify HBP. The model could produce an auROC of 0.981 ± 0.028 on training data using 10-fold cross-validation and an overall accuracy of 95.0% on independent test data. As the first model for HBP recognition, it will provide some help for infectious diseases and stimulate further research in related fields.

Keywords: heparin-binding protein; amino acid composition; dipeptide composition; dipeptide deviation from expected mean; composition/transition/distribution; support vector machine

check for updates

Citation: Zhu, W.; Yuan, S.-S.; Li, J.; Huang, C.-B.; Lin, H.; Liao, B. A First Computational Frame for Recognizing Heparin-Binding Protein. *Diagnostics* **2023**, *13*, 2465. https://doi.org/10.3390/diagnostics13142465

Academic Editors: Dania Cioni, Cemil Colak and Sami Akbulut

Received: 10 June 2023
Revised: 13 July 2023
Accepted: 21 July 2023
Published: 24 July 2023

1. Introduction

Heparin-binding protein (HBP), also known as azurocidin or CAP-37, is a cationic antimicrobial protein derived from the granulosa protein of polynuclear neutrophils [1–3]. Studies have found that the biosynthetic HBP in neutrophils is rapidly released under bacterial stimulation, leading to increased vascular permeability and edema [4,5], and has a proinflammatory effect on a variety of leukocytes and epithelial cells [6]. Therefore, HBP in plasma can be used as a new diagnostic marker for bacterial skin infection, acute bacterial meningitis, leptospirosis, protozoan parasites, and even some noninfectious diseases [4,7–10]. Especially for sepsis, a systemic inflammatory response syndrome caused by infection, HBP is an effective early and predictive biomarker [11–13]. In fact, it has been found that HBP levels in plasma are elevated in septic patients a few hours before the onset of hypotension or organ dysfunction [14].

The correct recognition of HBP can provide important clues for the study of biomarkers of infectious diseases. Traditional molecular biology methods can provide accurate information to study HBP [15,16]. However, these experiments require a longer cycle, more experimental resources, and more expensive manpower. The continuous accumulation of biological data provides a basis for us to mine potential biological knowledge from these data [17–22]. The continuous progress of various data analysis methods and artificial intelligence technology provides a favorable tool for us to obtain knowledge [23]. In fact, machine learning methods have been widely used in the recognition of special functional proteins [20,24–33], for example, bioluminescent protein [34], hormone-binding protein [35], and transcription factors [36–40]. In these works, several kinds of sequence descriptors, such as amino acid composition (AAC), reduced amino acid composition (RAAC) [41–43], pseudo amino acid composition (PseAAC) [20,44], and dipeptide composition (DC) [35], were developed.

Although research on these special functional proteins has been successful, to our knowledge, there is still no computational prediction work for HBP recognition at present because there was a lack of available datasets in the past, and people had previously paid more attention to the research of molecular biology experiments. Thus, it is urgent to develop an efficient prediction model to identify HBP.

This work aims to build a powerful computational model to identify HBP. At first, a reliable benchmark dataset was collected and constructed for training and testing various computational models. Subsequently, four sequence descriptors were adopted to formulate sequence samples. Two kinds of machine learning methods, namely, support vector machine (SVM) [45] and random forest (RF) [46], were selected as classifiers for executing classification. The following sections provide a detailed description of the workflow (Figure 1).

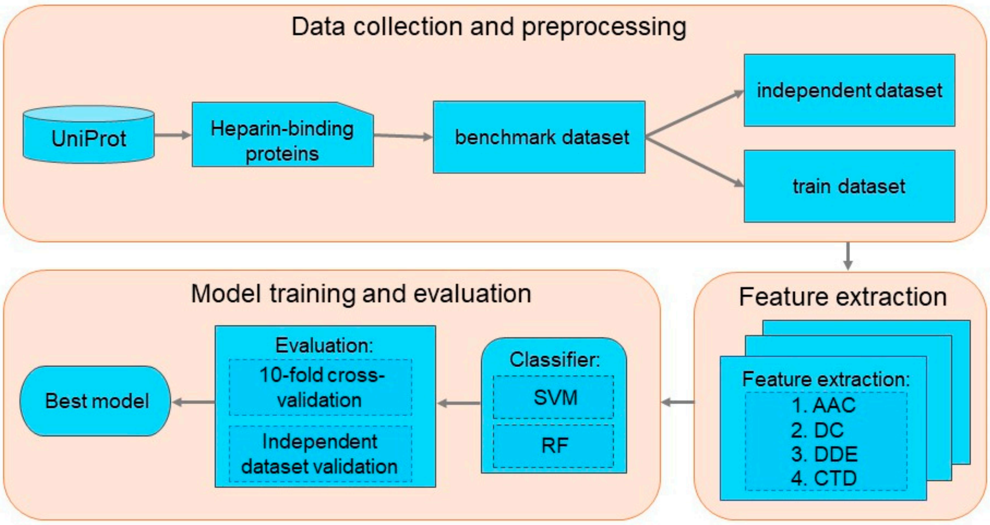


Figure 1. The workflow of the prediction of HBP.

2. Materials and Methods

2.1. Benchmark Dataset Construction

In biological macromolecular classification and recognition, a reliable benchmark dataset is the foundation for constructing a reliable model [47–50]. It is well known that the Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data [51]. This database provides abundant protein information. Therefore, the raw HBP data were collected from UniProt by using heparin binding (KW-0358) as keyword. In UniProt, there are five forms of evidence for protein existence, that is, evidence at the protein level, evidence at the transcriptional level, evidence from homology, predicted, and uncertain. Obviously, the proteins with evidence provided by the first two have higher reliability, so HBPs with the other three kinds of evidence were excluded. In addition, sequences that have ambiguous residues, such as “B”, “J”, “O”, “U”, “X”, and “Z” were checked and excluded. Finally, a total of 391 HBPs were obtained.

Because protein sequences with high similarity will reduce the scalability of the prediction model, those proteins with high similarity must be removed. In general, 40% or 25% is the commonly used threshold of sequence identity when constructing prediction models of special functional proteins. However, due to the limitation of the number of samples, if such sequence identity threshold is used, the number of samples will not be statistically significant, which will lead to the loss of objectivity of the model. Therefore, to balance the number of samples and sequence identity, 80% was adopted as the threshold of sequence identity. The software that performs redundant sequence removal is CD-HIT [52]. As a result, 183 HBPs were kept as positive samples.

In the prediction of special functional proteins, the selection of negative samples is a very difficult task. If all proteins in UniProt that are not annotated as HBP are selected, the data are huge, and the negative samples are almost 3000 times that of the positive samples, which is extremely unfavorable to the construction of the prediction model. In addition, the functional annotation of many of these proteins is not complete. Some proteins may be HBP, but they have not been identified before, which can also lead to bias in a machine learning model. Therefore, to avoid the above two problems as much as possible, the following steps were carried out to select negative samples. First of all, human DNA-binding proteins were chosen as candidate negative samples, because they differ greatly from positive samples in biological functions, which can avoid the problems mentioned above. To improve the reliability of negative samples, those DNA-binding proteins that have structural information, the existence of evidence at the protein level, and a sequence identity of >80% were selected. Despite such stringent criteria, 559 negative sample sequences were obtained, which is more than positive samples. To balance positive and negative samples, 183 of them were randomly selected as the final negative sample dataset.

Based on such benchmark dataset, 50 positive samples and 50 negative samples were randomly selected as test data, and the remaining as training data (133 positives and 133 negatives), formulated as

$$\begin{cases} S_{train} = S_{train}^{positive} \cup S_{train}^{negtive} \\ S_{test} = S_{test}^{positive} \cup S_{test}^{negtive} \end{cases} \tag{1}$$

2.2. Formulation of Protein Sequences

The method for special function protein recognition based on the machine learning method is to classify the samples according to their characteristics in the benchmark dataset [53–57]. Protein is a sequence composed of 20 amino acids with different lengths. However, machine learning requires that every sample should have the same dimension of features. Therefore, how to transform the protein sequence into a discrete digital vector is a key problem for classification model construction. The feature vector should be able to effectively characterize the basic attributes of these samples without losing the original information [58]. In fact, in the past 30 years, scholars have developed a variety of sequence representation methods. Some of these features have good universality, such as pseudo nucleotide composition and position-specific scoring matrix, which have been successful in many protein prediction problems. Additionally, some features have strong specificity, such as amino acid composition, which is very suitable for the recognition of thermophilic proteins. In any previous work, it has not been pointed out what features are used to characterize HBP samples. Therefore, in order to describe the sequence attributes of HBP from multiple perspectives, the following four descriptors were used to extract the features of protein sequences, described as follows.

2.2.1. Amino Acid Composition (AAC)

Although AAC does not perform well in many prediction problems of predicted proteins, it can be used as a supplement to other features as a basic feature [59]. Therefore, this paper also uses this feature to test its prediction performance for HBP.

For any protein sequence expressed as $\boldsymbol{P} = R_1 R_2, \cdots, R_L$, where L is the length of the protein and also is the number of residues. $R_i (i = 1, 2, \cdots, L)$ is the residue at the i-th position in the sequence. R belongs to 1 of 20 amino acids ($R \in (A, C, \ldots, Y)$). Then the AAC is the probability that 20 kinds of amino acids appear in this protein, that is, the number of 20 kinds of amino acids divided by the total number of amino acids in this protein (namely, the sequence length L). The following formula was used to express the AAC:

$$F(R) = \frac{N_R}{\sum_R N_R} = \frac{N_R}{L} \tag{2}$$

where N_R is the total number of amino acids R in the given sequence $\boldsymbol{P}$. Then this protein can be expressed by the feature vector as

$$\boldsymbol{P} = [F(A), F(C), \ldots, F(Y)]_{20} \tag{3}$$

where 20 denotes the dimension of the vector.

2.2.2. Dipeptide Composition (DC)

Studies have found that the linkage between amino acids is not random. A certain amino acid is often followed by another relatively fixed type of amino acid; that is, the arrangement of amino acids in proteins is also a unique feature of proteins. In fact, when people study the information storage of the genome, they also find that the adjacent association of nucleotides is the main way of genetic information storage. Therefore, it can also be speculated that the order of amino acids in protein sequences is one way of protein information storage. The order information of adjacent amino acids, also called DC, in proteins is still an important feature to characterize amino acids, and has been widely used in protein classification.

Since proteins have 20 kinds of normal amino acids, there are 400 kinds of dipeptides [60]. We need to count the numbers of these 400 dipeptides in the protein, and then calculate their frequencies in the whole sequence. The calculation formula is as follows:

$$F(R^f R^b) = \frac{N_{R^f R^b}}{\sum R^f R^b} = \frac{N_{R^f R^b}}{L-1} \tag{4}$$

where $N_{R^f R^b}$ is the total number of the dipeptides $R^f R^b$ in the given sequence $\boldsymbol{P}$. f and b represent the front residue R and back residue R in the dipeptide $R^f R^b$. Then the feature vector for this protein can be expressed by

$$\boldsymbol{P} = [F(AA), F(AC), \ldots, F(YY)]_{400} \tag{5}$$

where 400 denotes the dimension of this vector.

2.2.3. Dipeptide Deviation from Expected Mean (DDE)

There are 4 nucleotides and 20 amino acids in the genome. The 3 nucleotides are connected to form a codon to encode amino acids or become stop codons. Since 20 kinds of amino acids are encoded by 61 codons, there is degeneracy; that is, one kind of amino acid is encoded by multiple codons. Therefore, for any protein sequence, the theoretical frequency of dipeptide appearance can be described by the coding degeneracy of codons, which is defined as follows:

$$TF(R^f R^b) = \frac{C_{R^f}}{C_N} \times \frac{C_{R^b}}{C_N} \tag{6}$$

where C_{R^f} and C_{R^b} are the numbers of codons that code for the first amino acid residue and the second amino acid residue in the given dipeptide "$R^f R^b$". C_N is the total number of possible codons after excluding the three stop codons (C_N = 61).

Then the theoretical variance of the dipeptide "$R^f R^b$" can be defined as

$$TV(R^f R^b) = \frac{TF(R^f R^b)[1 - TF(R^f R^b)]}{L-1} \tag{7}$$

The Z-transform is performed between the observed dipeptide frequency (defined in Equation (4)) and the theoretical dipeptide frequency (defined in Equation (6)) in a sequence, as shown below.

$$DDE(R^f R^b) = \frac{F(R^f R^b) - TF(R^f R^b)}{\sqrt{TV(R^f R^b)}} \tag{8}$$

Equation (8) describes the deviation of the observed dipeptide frequency from the theoretical dipeptide frequency and is thus called DDE. Protein sample vectorization is described as

$$\boldsymbol{P} = [DDE(AA), DDE(AC), \ldots, DDE(YY)]_{400} \tag{9}$$

where 400 denotes the dimension of this vector.

2.2.4. Composition/Transition/Distribution (CTD)

Usually, some fragments in a protein chain will form a specific secondary structure or have some special biological activities. Many attempts have been made to describe these fragment features effectively. Among them, CTD is one of the more effective ways to represent the amino acid distribution patterns of a specific structural or physicochemical property in a protein or peptide sequence. Thus, in this work, it is also used for feature extraction to express protein samples.

Amino acid itself is a chemical molecule with specific physicochemical properties. According to the physicochemical properties of amino acids, the frequency of amino acids in each group of properties for a sample sequence (expressed as C) can be redescribed. Amino acids with certain characteristics may form a fragment, such as a continuous hydrophilic fragment exposed on the protein surface. However, the next few amino acids may have other properties. Therefore, T measures the frequencies of property change of amino acids compared with the immediately adjacent amino acids in the sample sequence. D was proposed to character the distribution patterns of the first 25%, 50%, 75%, and 100% of the sample sequence. Details of features are in the following section.

According to the previous studies, 13 physicochemical properties of amino acids are selected for the next characterization. For each property, these 20 amino acids are divided into three categories, such as, for the secondary structure, they can be divided into helix, strand, and coil. Then, 20 amino acids can be divided into 39 total (13 × 3) groups. The percentage of each group in protein sequence is defined as follows:

$$C(i,j) = \frac{n_{i,j}}{\sum_{i,j} n_{i,j}} = \frac{n_{i,j}}{L} \tag{10}$$

where $n_{i,j}$ is the number of residues in the i-th group of the j-th physicochemical property. Therefore, this descriptor, also known as CTDC, is used to describe the protein sequence as

$$\boldsymbol{P} = [C(1,1), C(1,2), \ldots, C(3,13)]_{39} \tag{11}$$

where 39 denotes the dimension of this vector.

CTDT represents the transition probability of two adjacent amino acid residues belonging to two different groups, which can be calculated by the following formula:

$$T(i,j) = \frac{n_{R^f R^b} + n_{R^b R^f}}{L-1} \tag{12}$$

where $n_{R^f R^b}$ and $n_{R^b R^f}$ are the numbers of the dipeptides "$R^f R^b$" and "$R^b R^f$", respectively, while R^f and R^b are amino acids in the i-th group and not. Then, vectorization is used to represent the protein as

$$\boldsymbol{P} = [T(1,1), T(1,2), \ldots, T(3,13)]_{39} \tag{13}$$

where 39 denotes the dimension of this vector.

The relative location in one sequence-represented distribution of residues of given groups can be described by CTDD. Considering that a protein sequence is divided into 5 segments according to percentages of 1%, 25%, 50%, 75%, and 100%, the number of amino acids with the *j*-th physicochemical property in group *i* in each segment can be expressed as

$$n_{i,j}^{p} = \left\lfloor \frac{p}{100} \times n_{i,j} \right\rfloor \tag{14}$$

where p is 1, 25, 50, 75, and 100. When $n_{i,j}^{p}$ is less than 1, it is assigned a value of 1. Then, CTDD can be represented as

$$D_{i,j}^{(1+inter(\frac{p}{25}))} = \frac{loc\left(n_{i,j}^{p}\right)}{L} \times 100 \tag{15}$$

where $loc\left(n_{i,j}^{p}\right)$ denotes the location at the sequence that the occurrence number of residues of a given group reaches $n_{i,j}^{p}$. Then, the feature vector of CTDD can be expressed as

$$\boldsymbol{P} = [D_{1,1}^{1}, D_{1,2}^{1}, \ldots, D_{3,13}^{5}]_{195} \tag{16}$$

where 195 denotes the dimension of this vector.

By combining the three features CTDC, CTDT, and CTDD (Equations (11), (13), and (16)), a protein sample could be formulated as a 39 × (2 + 5) = 273 dimensional vector shown as follows:

$$\boldsymbol{P} = [C(1,1), \ldots, T(1,1), \ldots, D_{3,13}^{5}]_{273} \tag{17}$$

It should be pointed out that although the three CTD features are mixed in many protein prediction works, the three features exist independently of each other, so they can also be used independently for protein prediction.

2.3. Machine Learning Methods

How to find appropriate decision conditions in the feature space, and then distinguish different types of samples, is the third step in the biological macromolecule recognition problem [61–66]. Machine learning methods can provide appropriate classification decision criteria to distinguish different types of samples [67]. They have been widely applied in bioinformatics [68–76]. At present, deep learning has become a popular method. However, it requires a lot of computing resources and needs more experience to search parameters. Thus, in this work, two popular algorithms, namely, support vector machine (SVM) and random forest (RF), which are very suitable for small samples, were only considered.

SVM is a typical representative of machine learning suitable for small-sample learning. Its principle involves utilizing the kernel function to transform low-dimensional samples into high-dimensional feature space, and then find the hyperplane that can distinguish samples in high-dimensional space. Since most problems in biology are nonlinear subproblems, radial basis function (RBF) are most commonly used. For a detailed introduction to SVM, please refer to the literature.

RF is an integrated learning method based on a decision tree, which can be regarded as an upgrade of a decision tree. Its principle is to construct multiple decision trees for classification during training, and its output is the category selected by most trees. RF avoids the overfitting of a decision tree when building a model.

2.4. Evaluation Indexes

After the model is built, the performance of the model needs to be evaluated. In this study, stratified 10-fold cross-validation without shuffle was performed on training data to fine-tune parameters and test models [77–84]. The grid search method was applied to search for the best parameters of the model in search spaces (Table 1). Additionally, the independent data were utilized to test the final model after the best model was established on the training data.

Table 1. Search spaces of SVM and RF.

Parameters	SVM [1]	Parameters	RF
"kernel"	Linear, RBF, sigmoid, poly	"criterion"	Gini, entropy
"C"	$2^x, x \in [-1, 15]$	"max_depth"	[5, 150]
"gamma"	$2^x, x \in [-14, 2]$	"min_samples_split"	[2, 30]
"degree"	[1, 5]	"min_samples_leaf"	[5]
\	\	"max_leaf_nodes"	[100]
\	\	"ccp_alpha"	[0.001]
\	\	"n_estimators"	$10^x, x \in [1, 3]$

[1] When "kernel" is linear, there are no "gamma" and "degree" parameters to be set. When the only "kernel" specifies as poly, the "degree" parameter makes sense.

Tenfold cross-validation and independent set test are evaluation strategies for accessing prediction ability. Models also need specific evaluation indicators to evaluate [85–90]. Here, the prediction ability of the model was evaluated by using sensitivity (*Sn*), specificity (*Sp*), overall accuracy (*OA*), Matthews correlation coefficient (*MCC*), and area under the receiver operating characteristic curve (auROC):

$$Sn = \frac{TP}{TP + FN} \tag{18}$$

$$Sp = \frac{TN}{TN + FP} \tag{19}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TN + FN) \times (TP + FN) \times (TN + FP)}} \tag{20}$$

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{21}$$

where *TP* and *TN* are the numbers of correctly predicted HBPs and non-HBPs, respectively. *FP* denotes the number of non-HBPs that were recognized as HBPs, while *FN* denotes the number of HBPs that were identified as non-HBPs. Additionally, the auROC can quantitatively evaluate the performance of the model. Thus, as there are several metrics, the auROC is the first metric to be considered. The greater the auROC, the better the performance of the model is. If there are models that have the same auROCs, *OA* can be the second metric to be considered. To be more precise, *MCC* can be the next metric that represents the performances of models.

3. Results

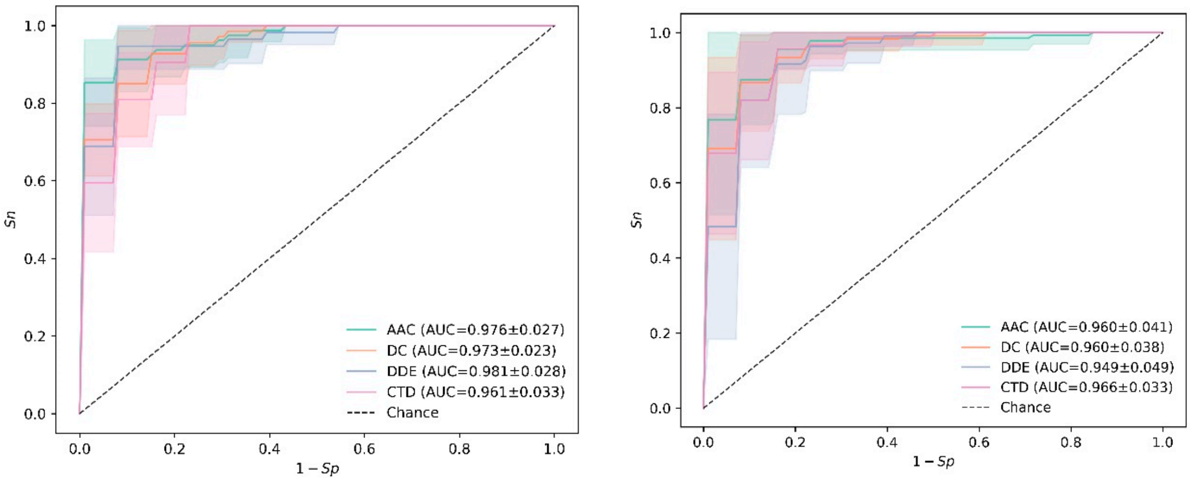
3.1. Experiments on Training Data

Four kinds of sequence feature extraction strategies were introduced in the above section. According to their definition, each protein can be described as 20 dimension, 400 dimension, 400 dimension, and 273 dimension vectors, respectively, for AAC, DC, DDE, and CTD. Analysis of variance (ANOVA) was applied for ranking the top 5 features of four feature descriptors (Table 2). AAC, DC, and DDE have a common high *F*-score and *p*-value about amino acid *S* (Serine). These results also reveal the extremely different composition between HBPs and non-HBPs. Additionally, according to the high score features of CTD, the differences between HBPs and non-HBPs are mainly concentrated on solvent access, hydrophobicity, polarity, and second structure properties.

Table 2. The *F*-scores and corresponding *p*-values for the top 5 features of feature descriptors.

Feature Descriptor	Feature Name	*F*-Score	*p*-Value
AAC	S	105.4221	4.9857×10^{-21}
	C	51.9136	6.0761×10^{-12}
	P	39.1761	1.5583×10^{-9}
	V	28.9828	1.6138×10^{-7}
	W	18.6945	2.1764×10^{-5}
DC	SS	85.6575	7.6231×10^{-18}
	PS	64.2325	3.5827×10^{-14}
	SP	63.8450	4.1972×10^{-14}
	PA	39.7520	1.1720×10^{-9}
	QP	28.6408	4.4865×10^{-8}
DDE	SS	87.9754	3.1559×10^{-18}
	PS	62.1955	8.2516×10^{-14}
	SP	60.4646	1.6840×10^{-13}
	PA	36.5852	4.9772×10^{-9}
	FC	29.5543	1.2376×10^{-7}
CTD	solventaccess.G3	93.4288	4.0577×10^{-19}
	hydrophobicity_ARGP820101.G2	83.7918	1.5572×10^{-17}
	polarity.G3	80.3504	5.8770×10^{-17}
	hydrophobicity_ZIMJ680101.G1	73.3974	8.9800×10^{-16}
	secondarystruct.G1	69.7518	3.8413×10^{-15}

Next, the prediction performance of each kind of feature on training data using SVM and RF was investigated. The ROC curves of 10-fold cross-validation are plotted in Figure 2. From Figure 2a, one may notice that *DDE* could produce a maximum auROC of 0.981 ± 0.028 among the four kinds of features when using SVM as classifier. However, the best feature is CTD for RF, as shown in Figure 2b. The auROC is 0.966 ± 0.033. Especially for RF, DDE is the worst feature, which can only achieve an auROC of 0.949 ± 0.049. For AAC and DC, they could produce similar results no matter what kind of classifier was adopted. Through overall comparison, the results of SVM combined with DDE are better. Therefore, this model has the greatest potential to become the ultimate HBP prediction model.



(**a**) (**b**)

Figure 2. The results on training data using 10-fold cross-validation: (**a**) SVM, (**b**) RF.

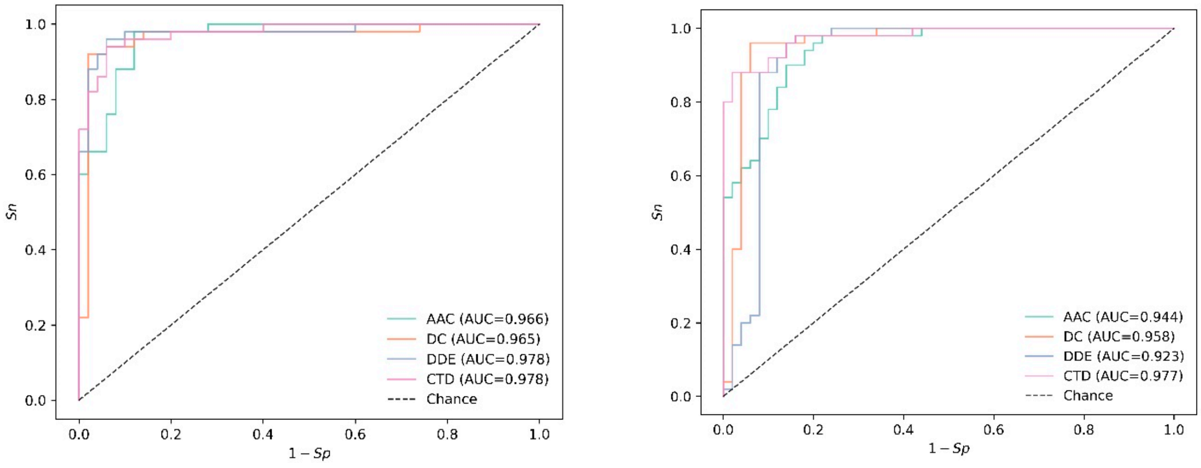
3.2. Experiments on Independent Data

On the training data, the prediction results of two algorithms combined with four features were investigated. To further confirm whether SVM combined with DDE is the best prediction model, eight models were tested on independent data (Table 3). It shows that SVM with DDE also has the best *OA* of 95.0%, with a balanced *Sn* and *Sp* of 96.0% and 94.0%, respectively. Other combinations also have the best OA of 95.0%. However, the final model will be chosen according to the highest auROC.

Table 3. Results of models on the independent data using different algorithms and feature descriptors.

Algorithm	Feature	*Sn* (%)	*Sp* (%)	*MCC*	*OA* (%)
SVM	AAC	98.0	88.0	0.864	93.0
	DC	92.0	98.0	0.902	95.0
	DDE	96.0	94.0	0.900	95.0
	CTD	94.0	94.0	0.880	94.0
RF	AAC	90.0	86.0	0.761	88.0
	DC	96.0	94.0	0.900	95.0
	DDE	96.0	86.0	0.824	91.0
	CTD	88.0	98.0	0.864	93.0

The eight ROC curves are plotted in Figure 3. For SVM-based models, there are two features, DDE and CTD, both of which obtain the maximum auROC. Different from the SVM-based model, the best feature of RF, still CTD, has not changed in the training data and independent data. Additionally, the best parameters for SVM with DDE are shown in Table 4.



(**a**) (**b**)

Figure 3. Results on independent data: (**a**) SVM, (**b**) RF.

Table 4. Best parameters of SVM with DDE.

Parameters	Value
"kernel"	RBF
"C"	4.59479341998814
"gamma"	0.07982260524725553

4. Further Discussion

Based on the above results on training data and independent data, a very good model for HBP recognition was obtained. However, one must realize that the sequence similarity of the benchmark dataset used in this model is relatively high. Generally speaking, building models based on low-similarity datasets has better robustness and scalability. However, the number of current samples is not enough to support us in generating such data for building a model. Therefore, it is our direction to constantly collect new data to expand the samples of the model.

In this work, four features were adopted to encode samples, and numerous sequence features were developed. However, some relatively simple features were utilized to obtain satisfactory prediction accuracy, demonstrating that HBP sequences have their special characteristics. In addition, feature fusion and feature filtering are also commonly used to improve model accuracy, while not applied in this work since the current features have generated considerable prediction performance. Now, although the data size is small, it is sufficient for small sample learning algorithms such as SVM and RF to build light models with pretty good performances. This is also why this first computational work for HBP recognition comes out. Of course, another possible reason for the absence of computational methods is that researchers always focus on experimental methods and ignore them.

With the data size increasing, feature fusion and feature filtering will be implemented when a single feature cannot fully describe the sample characteristics. If there are enough data, more machine learning algorithms can be considered for comparison. Deep learning is now very popular in bioinformatics. However, this algorithm requires more computational resources and also has certain requirements for sample size and feature dimensions.

The identification HBP in medical plasma must consider that the proteins have to obtain their sequences, since our model was constructed based on sequence features. Once sequences of proteins are obtained, feature extraction and model prediction can be conducted, and the results can be produced within a few seconds. Additionally, this procedure only consumes some computing resources—even on mobile phones—without laboratory resources. Compared with molecular biology methods, for example, enzyme-linked immunosorbent assay [91] is based on antibodies of known HBPs, which takes several hours to complete recognition. Computational methods save time and resources. However, there are some limitations, such as false predictions, and the model cannot predict the affinity of heparin binding, which is essential for medical use. In the future, with the development of computational methods and computing resources, more accurate recognition, more functional prediction, and faster processing speed will be achieved. Then more HBPs from various species can be identified, which aids in the research of infectious disease biomarkers.

HBP is an important biomarker of infectious diseases. The correct identification of HBP is of great significance for the study of infectious diseases. The construction of this model will provide clues for the identification of important biomarkers of infectious diseases and the discovery of potential drug targets. This work also contributes to the wider application of artificial intelligence methods in the field of clinical medicine, especially in the identification of biomarkers.

5. Conclusions

Four kinds of sequence features were extracted for HBP, and two machine learning methods, SVM and RF, were evaluated. Eventually, DDE combined with SVM was chosen to construct the final prediction model. The model shows good prediction results on both the training set and independent set. To our knowledge, this is the only HBP recognition model based on machine learning. This model is slightly rough, but it provides pioneering research on the use of artificial intelligence methods to study HBP. It is hoped that a more in-depth and detailed analysis of HBP can be carried out in the future.

Author Contributions: Conceptualization, C.-B.H., H.L. and B.L.; methodology, W.Z. and S.-S.Y.; software, W.Z. and S.-S.Y.; validation, W.Z. and J.L.; formal analysis, W.Z. and S.-S.Y.; investigation, S.-S.Y. and H.L.; resources, J.L. and H.L.; data curation, W.Z. and C.-B.H.; writing—original draft preparation, W.Z. and H.L.; writing—review and editing, H.L. and B.L.; visualization, S.-S.Y. and C.-B.H.; supervision, H.L. and B.L.; project administration, H.L. and B.L.; funding acquisition, H.L. and B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the National Nature Science Foundation of China (Grant Nos. 62250028, 61863010, 11926205, 11926412, and 61873076), National Key R&D Program of China (Grant No. 2020YFB2104400), and Natural Science Foundation of Hainan, China (Grant Nos. 121RC538, 119MS036, and 120RC588).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed in the current study are available in UniProt. Other data are available upon request to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

auROC	area under the receiver operating characteristic curve
AAC	amino acid composition
CTD	composition/transition/distribution
DC	dipeptide composition
DDE	dipeptide deviation from expected mean
HBP	heparin-binding protein
MCC	Matthews correlation coefficient
OA	overall accuracy
PseAAC	pseudo amino acid composition
RAAC	reduced amino acid composition
RF	random forest
Sn	sensitivity
Sp	specificity
SVM	support vector machine
UniProt	Universal Protein Resource

References

1. Fisher, J.; Kahn, F.; Wiebe, E.; Gustafsson, P.; Kander, T.; Mellhammar, L.; Bentzer, P.; Linder, A. The Dynamics of Circulating Heparin-Binding Protein: Implications for Its Use as a Biomarker. *J. Innate. Immun.* **2022**, *14*, 447–460. [CrossRef] [PubMed]
2. Cheng, L.; Qi, C.; Zhuang, H.; Fu, T.; Zhang, X. gutMDisorder: A comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Res.* **2020**, *48*, D554–D560. [CrossRef] [PubMed]
3. Yu, H.; Shen, Z.-A.; Zhou, Y.-K.; Du, P.-F. Recent advances in predicting protein-lncRNA interactions using machine learning methods. *Curr. Gene Ther.* **2021**, *22*, 228–244.
4. Yang, Y.; Liu, G.; He, Q.; Shen, J.; Xu, L.; Zhu, P.; Zhao, M. A Promising Candidate: Heparin-Binding Protein Steps onto the Stage of Sepsis Prediction. *J. Immunol. Res.* **2019**, *2019*, 7515346. [CrossRef]
5. Cheng, L.; Qi, C.; Yang, H.; Lu, M.; Cai, Y.; Fu, T.; Ren, J.; Jin, Q.; Zhang, X. gutMGene: A comprehensive database for target genes of gut microbes and microbial metabolites. *Nucleic Acids Res.* **2021**, *50*, D795–D800. [CrossRef]
6. Sato, W.; Kadomatsu, K.; Yuzawa, Y.; Muramatsu, H.; Hotta, N.; Matsuo, S.; Muramatsu, T. Midkine is involved in neutrophil infiltration into the tubulointerstitium in ischemic renal injury. *J. Immunol.* **2001**, *167*, 3463–3469. [CrossRef]
7. Ao, C.; Yu, L.; Zou, Q. Prediction of bio-sequence modifications and the associations with diseases. *Brief. Funct. Genom.* **2021**, *20*, 1–18. [CrossRef]
8. Qi, C.; Wang, C.; Zhao, L.; Zhu, Z.; Wang, P.; Zhang, S.; Cheng, L.; Zhang, X. SCovid: Single-cell atlases for exposing molecular characteristics of COVID-19 across 10 human tissues. *Nucleic Acids Res.* **2021**, *50*, D867–D874. [CrossRef]
9. Bascuas, T.; Zedira, H.; Kropp, M.; Harmening, N.; Asrih, M.; Prat-Souteyrand, C.; Tian, S.; Thumann, G. Human Retinal Pigment Epithelial Cells Overexpressing the Neuroprotective Proteins PEDF and GM-CSF to Treat Degeneration of the Neural Retina. *Curr. Gene Ther.* **2021**, *22*, 168–183. [CrossRef]
10. Ning, L.; Abagna, H.B.; Jiang, Q.; Liu, S.; Huang, J. Development and application of therapeutic antibodies against COVID-19. *Int. J. Biol. Sci.* **2021**, *17*, 1486–1496. [CrossRef]

11. Neumann, A. Rapid release of sepsis markers heparin-binding protein and calprotectin triggered by anaerobic cocci poses an underestimated threat. *Anaerobe* **2022**, *75*, 102584. [CrossRef] [PubMed]
12. Ning, L.; Liu, M.; Gou, Y.; Yang, Y.; He, B.; Huang, J. Development and application of ribonucleic acid therapy strategies against COVID-19. *Int. J. Biol. Sci.* **2022**, *18*, 5070–5085. [CrossRef] [PubMed]
13. Ren, L.; Xu, Y.; Ning, L.; Pan, X.; Li, Y.; Zhao, Q.; Pang, B.; Huang, J.; Deng, K.; Zhang, Y. TCM2COVID: A resource of anti-COVID-19 traditional Chinese medicine with effects and mechanisms. *iMETA* **2022**, *1*, e42. [CrossRef] [PubMed]
14. Fisher, J.; Linder, A. Heparin-binding protein: A key player in the pathophysiology of organ dysfunction in sepsis. *J. Intern. Med.* **2017**, *281*, 562–574. [CrossRef]
15. Wu, Y.L.; Yo, C.H.; Hsu, W.T.; Qian, F.; Wu, B.S.; Dou, Q.L.; Lee, C.C. Accuracy of Heparin-Binding Protein in Diagnosing Sepsis: A Systematic Review and Meta-Analysis. *Crit. Care Med.* **2021**, *49*, e80–e90. [CrossRef]
16. Zhang, Y.; Liu, T.; Wang, J.; Zou, B.; Li, L.; Yao, L.; Chen, K.; Ning, L.; Wu, B.; Zhao, X.; et al. Cellinker: A platform of ligand-receptor interactions for intercellular communication analysis. *Bioinformatics* **2021**, *37*, 2025–2032. [CrossRef]
17. Su, W.; Liu, M.L.; Yang, Y.H.; Wang, J.S.; Li, S.H.; Lv, H.; Dao, F.Y.; Yang, H.; Lin, H. PPD: A Manually Curated Database for Experimentally Verified Prokaryotic Promoters. *J. Mol. Biol.* **2021**, *433*, 166860. [CrossRef]
18. Ning, L.; Cui, T.; Zheng, B.; Wang, N.; Luo, J.; Yang, B.; Du, M.; Cheng, J.; Dou, Y.; Wang, D. MNDR v3.0: Mammal ncRNA-disease repository with increased coverage and annotation. *Nucleic Acids Res.* **2021**, *49*, D160–D164. [CrossRef]
19. Cheng, L.; Hu, Y.; Sun, J.; Zhou, M.; Jiang, Q. DincRNA: A comprehensive web-based bioinformatics toolkit for exploring disease associations and ncRNA function. *Bioinformatics* **2018**, *34*, 1953–1956. [CrossRef]
20. Ahmed, Z.; Zulfiqar, H.; Khan, A.A.; Gul, I.; Dao, F.Y.; Zhang, Z.Y.; Yu, X.L.; Tang, L. iThermo: A Sequence-Based Model for Identifying Thermophilic Proteins Using a Multi-Feature Fusion Strategy. *Front. Microbiol.* **2022**, *13*, 790063. [CrossRef]
21. Hu, Y.; Sun, J.Y.; Zhang, Y.; Zhang, H.; Gao, S.; Wang, T.; Han, Z.; Wang, L.; Sun, B.L.; Liu, G. rs1990622 variant associates with Alzheimer's disease and regulates TMEM106B expression in human brain tissues. *BMC Med.* **2021**, *19*, 11. [CrossRef] [PubMed]
22. Hu, Y.; Zhang, H.; Liu, B.; Gao, S.; Wang, T.; Han, Z.; International Genomics of Alzheimer's, P.; Ji, X.; Liu, G. rs34331204 regulates TSPAN13 expression and contributes to Alzheimer's disease with sex differences. *Brain* **2020**, *143*, e95. [CrossRef]
23. Dao, F.Y.; Lv, H.; Zhang, Z.Y.; Lin, H. BDselect: A Package for k-mer Selection Based on the Binomial Distribution. *Curr. Bioinform.* **2022**, *17*, 238–244.
24. Sanami, S.; Alizadeh, M.; Nosrati, M.; Dehkordi, K.A.; Azadegan-Dehkordi, F.; Tahmasebian, S.; Nosrati, H.; Arjmand, M.-H.; Ghasemi-Dehnoo, M.; Rafiei, A.; et al. Exploring SARS-CoV-2 structural proteins to design a multi-epitope vaccine using immunoinformatics approach: An in silico study. *Comput. Biol. Med.* **2021**, *133*, 104390. [CrossRef]
25. Wu, X.; Yu, L. EPSOL: Sequence-based protein solubility prediction using multidimensional embedding. *Bioinformatics* **2021**, *37*, 4314–4320. [CrossRef] [PubMed]
26. Liu, Q.; Wan, J.; Wang, G. A survey on computational methods in discovering protein inhibitors of SARS-CoV-2. *Brief. Bioinform.* **2022**, *23*, bbab416. [CrossRef]
27. Zhao, X.; Wang, H.; Li, H.; Wu, Y.; Wang, G. Identifying Plant Pentatricopeptide Repeat Proteins Using a Variable Selection Method. *Front. Plant. Sci.* **2021**, *12*, 506681. [CrossRef]
28. Teng, Z.; Zhang, Z.; Tian, Z.; Li, Y.; Wang, G. ReRF-Pred: Predicting amyloidogenic regions of proteins based on their pseudo amino acid composition and tripeptide composition. *BMC Bioinform.* **2021**, *22*, 545. [CrossRef]
29. Zhai, Y.; Chen, Y.; Teng, Z.; Zhao, Y. Identifying Antioxidant Proteins by Using Amino Acid Composition and Protein-Protein Interactions. *Front. Cell. Dev. Biol.* **2020**, *8*, 591487. [CrossRef]
30. Tao, Z.; Li, Y.; Teng, Z.; Zhao, Y. A Method for Identifying Vesicle Transport Proteins Based on LibSVM and MRMD. *Comput. Math. Methods Med.* **2020**, *2020*, 8926750. [CrossRef]
31. Hu, Y.; Qiu, S.; Cheng, L. Integration of Multiple-Omics Data to Analyze the Population-Specific Differences for Coronary Artery Disease. *Comput. Math. Methods Med.* **2021**, *2021*, 7036592. [CrossRef]
32. Hu, Y.; Zhang, Y.; Zhang, H.; Gao, S.; Wang, L.; Wang, T.; Han, Z.; International Genomics of Alzheimer's, P.; Liu, G. Mendelian randomization highlights causal association between genetically increased C-reactive protein levels and reduced Alzheimer's disease risk. *Alzheimers Dement.* **2022**, *18*, 2003–2006. [CrossRef]
33. Hu, Y.; Zhang, Y.; Zhang, H.; Gao, S.; Wang, L.; Wang, T.; Han, Z.; Sun, B.L.; Liu, G. Cognitive performance protects against Alzheimer's disease independently of educational attainment and intelligence. *Mol. Psychiatry* **2022**, *27*, 4297–4306. [CrossRef]
34. Zhang, D.; Chen, H.-D.; Zulfiqar, H.; Yuan, S.-S.; Huang, Q.-L.; Zhang, Z.-Y.; Deng, K.-J. iBLP: An XGBoost-Based Predictor for Identifying Bioluminescent Proteins. *Comput. Math. Methods Med.* **2021**, *2021*, 6664362. [CrossRef] [PubMed]
35. Tang, H.; Zhao, Y.W.; Zou, P.; Zhang, C.M.; Chen, R.; Huang, P.; Lin, H. HBPred: A tool to identify growth hormone-binding proteins. *Int. J. Biol. Sci.* **2018**, *14*, 957–964. [CrossRef] [PubMed]
36. Zhang, L.; Yang, Y.; Chai, L.; Li, Q.; Liu, J.; Lin, H.; Liu, L. A deep learning model to identify gene expression level using cobinding transcription factor signals. *Brief. Bioinform.* **2022**, *23*, bbab501. [CrossRef]
37. Li, H.; Gong, Y.; Liu, Y.; Lin, H.; Wang, G. Detection of transcription factors binding to methylated DNA by deep recurrent neural network. *Brief. Bioinform.* **2022**, *23*, bbab533. [CrossRef]
38. Wang, H.; Liu, Y.; Guan, H.; Fan, G.-L. The Regulation of Target Genes by Co-occupancy of Transcription Factors, c-Myc and Mxi1 with Max in the Mouse Cell Line. *Curr. Bioinform.* **2020**, *15*, 581–588. [CrossRef]

39. Cheng, S.; Li, D.; Zhang, R.-Z.; Zhu, J.; Wang, L.; Liu, Q.; Chen, R.-H.; Liu, X.-M. Characterization of Induced Pluripotent Stem Cells from Human Epidermal Melanocytes by Transduction with Two Combinations of Transcription Factors. *Curr. Gene Ther.* **2020**, *19*, 395–403. [CrossRef] [PubMed]
40. Zhang, Y.; Liu, T.; Hu, X.; Wang, M.; Wang, J.; Zou, B.; Tan, P.; Cui, T.; Dou, Y.; Ning, L.; et al. CellCall: Integrating paired ligand-receptor and transcription factor activities for cell-cell communication. *Nucleic Acids Res.* **2021**, *49*, 8520–8534. [CrossRef]
41. Zuo, Y.; Li, Y.; Chen, Y.; Li, G.; Yan, Z.; Yang, L. PseKRAAC: A flexible web server for generating pseudo K-tuple reduced amino acids composition. *Bioinformatics* **2017**, *33*, 122–124. [CrossRef] [PubMed]
42. Zheng, L.; Liu, D.Y.; Yang, W.; Yang, L.; Zuo, Y.C. RaacLogo: A new sequence logo generator by using reduced amino acid clusters. *Brief. Bioinform.* **2021**, *22*, bbaa096. [CrossRef]
43. Zheng, L.; Liu, D.Y.; Li, Y.A.; Yang, S.Q.; Liang, Y.C.; Xing, Y.Q.; Zuo, Y.C. RaacFold: A webserver for 3D visualization and analysis of protein structure by using reduced amino acid alphabets. *Nucleic Acids Res.* **2022**, *50*, W633–W638. [CrossRef] [PubMed]
44. Ni, Y.-H.; Zhao, X.; Wang, W. CD24, A Review of its Role in Tumor Diagnosis, Progression and Therapy. *Curr. Gene Ther.* **2020**, *20*, 109–126. [CrossRef]
45. Zhang, Z.Y.; Yang, Y.H.; Ding, H.; Wang, D.; Chen, W.; Lin, H. Design powerful predictor for mRNA subcellular location prediction in Homo sapiens. *Brief. Bioinform.* **2021**, *22*, 526–535. [CrossRef]
46. Hunt, C.; Montgomery, S.; Berkenpas, J.W.; Sigafoos, N.; Oakley, J.C.; Espinosa, J.; Justice, N.; Kishaba, K.; Hippe, K.; Si, D.; et al. Recent Progress of Machine Learning in Gene Therapy. *Curr. Gene Ther.* **2021**, *22*, 132–143. [CrossRef]
47. Lv, H.; Dao, F.-Y.; Lin, H. DeepKla: An attention mechanism-based deep neural network for protein lysine lactylation site prediction. *iMeta* **2022**, *1*, e11. [CrossRef]
48. Wei, L.; Liao, M.; Gao, Y.; Ji, R.; He, Z.; Zou, Q. Improved and Promising Identification of Human MicroRNAs by Incorporating a High-Quality Negative Set. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 192–201. [CrossRef] [PubMed]
49. Wei, L.; Wan, S.; Guo, J.; Wong, K.K.L. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif. Intell. Med.* **2017**, *83*, 82–90. [CrossRef]
50. Jeon, Y.J.; Hasan, M.M.; Park, H.W.; Lee, K.W.; Manavalan, B. TACOS: A novel approach for accurate prediction of cell-specific long noncoding RNAs subcellular localization. *Brief. Bioinform.* **2022**, *23*, bbac243. [CrossRef]
51. UniProt, C. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489.
52. 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. [CrossRef] [PubMed]
53. Zhang, Z.Y.; Ning, L.; Ye, X.; Yang, Y.H.; Futamura, Y.; Sakurai, T.; Lin, H. iLoc-miRNA: Extracellular/intracellular miRNA prediction using deep BiLSTM with attention mechanism. *Brief. Bioinform.* **2022**, *23*, bbac395. [CrossRef]
54. Dao, F.Y.; Lv, H.; Zhang, D.; Zhang, Z.M.; Liu, L.; Lin, H. DeepYY1: A deep learning approach to identify YY1-mediated chromatin loops. *Brief. Bioinform.* **2021**, *22*, bbaa356. [CrossRef]
55. Basith, S.; Hasan, M.M.; Lee, G.; Wei, L.; Manavalan, B. Integrative machine learning framework for the identification of cell-specific enhancers from the human genome. *Brief. Bioinform.* **2021**, *22*, bbab252. [CrossRef] [PubMed]
56. Manavalan, B.; Patra, M.C. MLCPP 2.0: An Updated Cell-penetrating Peptides and Their Uptake Efficiency Predictor. *J. Mol. Biol.* **2022**, *434*, 167604. [CrossRef]
57. Thi Phan, L.; Woo Park, H.; Pitti, T.; Madhavan, T.; Jeon, Y.J.; Manavalan, B. MLACP 2.0: An updated machine learning tool for anticancer peptide prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4473–4480. [CrossRef]
58. Zheng, L.; Huang, S.; Mu, N.; Zhang, H.; Zhang, J.; Chang, Y.; Yang, L.; Zuo, Y. RAACBook: A web server of reduced amino acid alphabet for sequence-dependent inference by using Chou's five-step rule. *Database* **2019**, *2019*, baz131. [CrossRef]
59. Alim, A.; Rafay, A.; Naseem, I. PoGB-pred: Prediction of Antifreeze Proteins Sequences Using Amino Acid Composition with Feature Selection Followed by a Sequential-based Ensemble Approach. *Curr. Bioinform.* **2021**, *16*, 446–456. [CrossRef]
60. Yuan, S.S.; Gao, D.; Xie, X.Q.; Ma, C.Y.; Su, W.; Zhang, Z.Y.; Zheng, Y.; Ding, H. IBPred: A sequence-based predictor for identifying ion binding protein in phage. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 4942–4951. [CrossRef]
61. Zhang, Q.; Li, H.; Liu, Y.; Li, J.; Wu, C.; Tang, H. Exosomal Non-Coding RNAs: New Insights into the Biology of Hepatocellular Carcinoma. *Curr. Oncol.* **2022**, *29*, 5383–5406. [CrossRef]
62. Ye, Z.X.; Zhang, Y.X.; Liang, Y.B.; Lang, J.D.; Zhang, X.L.; Zang, G.L.; Yuan, D.W.; Tian, G.; Xiao, M.S.; Yang, J.L. Cervical Cancer Metastasis and Recurrence Risk Prediction Based on Deep Convolutional Neural Network. *Curr. Bioinform.* **2022**, *17*, 164–173.
63. Liu, T.; Chen, J.M.; Zhang, D.; Zhang, Q.; Peng, B.; Xu, L.; Tang, H. ApoPred: Identification of Apolipoproteins and Their Subfamilies With Multifarious Features. *Front. Cell. Dev. Biol.* **2020**, *8*, 621144. [CrossRef]
64. Zulfira, F.Z.; Suyanto, S.; Septiarini, A. Segmentation technique and dynamic ensemble selection to enhance glaucoma severity detection. *Comput. Biol. Med.* **2021**, *139*, 104951. [CrossRef] [PubMed]
65. Tang, H.; Cao, R.Z.; Wang, W.; Liu, T.S.; Wang, L.M.; He, C.M. A two-step discriminated method to identify thermophilic proteins. *Int. J. Biomath.* **2017**, *10*, 1750050. [CrossRef]
66. Zou, Y.; Wu, H.; Guo, X.; Peng, L.; Ding, Y.; Tang, J.; Guo, F. MK-FSVM-SVDD: A Multiple Kernel-based Fuzzy SVM Model for Predicting DNA-binding Proteins via Support Vector Data Description. *Curr. Bioinform.* **2021**, *16*, 274–283. [CrossRef]
67. Wang, H.; Liang, P.F.; Zheng, L.; Long, C.S.; Li, H.S.; Zuo, Y.C. eHSCPr discriminating the cell identity involved in endothelial to hematopoietic transition. *Bioinformatics* **2021**, *37*, 2157–2164. [CrossRef] [PubMed]

68. Yang, H.; Luo, Y.; Ren, X.; Wu, M.; He, X.; Peng, B.; Deng, K.; Yan, D.; Tang, H.; Lin, H. Risk Prediction of Diabetes: Big data mining with fusion of multifarious physical examination indicators. *Inf. Fusion.* **2021**, *75*, 140–149. [CrossRef]
69. Wang, X.; Wang, S.; Fu, H.; Ruan, X.; Tang, X. DeepFusion-RBP: Using Deep Learning to Fuse Multiple Features to Identify RNA-binding Protein Sequences. *Curr. Bioinform.* **2021**, *16*, 1089–1100. [CrossRef]
70. Wang, D.; Zhang, Z.; Jiang, Y.; Mao, Z.; Wang, D.; Lin, H.; Xu, D. DM3Loc: Multi-label mRNA subcellular localization prediction and analysis based on multi-head self-attention mechanism. *Nucleic Acids Res.* **2021**, *49*, e46. [CrossRef]
71. Lv, H.; Shi, L.; Berkenpas, J.W.; Dao, F.Y.; Zulfiqar, H.; Ding, H.; Zhang, Y.; Yang, L.; Cao, R. Application of artificial intelligence and machine learning for COVID-19 drug discovery and vaccine design. *Brief. Bioinform.* **2021**, *22*, bbab320. [CrossRef]
72. Berahmand, K.; Nasiri, E.; Mohammadiani, R.P.; Li, Y. Spectral clustering on protein-protein interaction networks via constructing affinity matrix using attributed graph embedding. *Comput. Biol. Med.* **2021**, *138*, 104933. [CrossRef] [PubMed]
73. Ali, F.; Akbar, S.; Ghulam, A.; Maher, Z.A.; Unar, A.; Talpur, D.B. AFP-CMBPred: Computational identification of antifreeze proteins by extending consensus sequences into multi-blocks evolutionary information. *Comput. Biol. Med.* **2021**, *139*, 105006. [CrossRef]
74. Yu, L.; Xia, M.; An, Q. A network embedding framework based on integrating multiplex network for drug combination prediction. *Brief. Bioinform.* **2022**, *23*, bbab364. [CrossRef] [PubMed]
75. Zhang, S.; Jiang, H.; Gao, B.; Yang, W.; Wang, G. Identification of Diagnostic Markers for Breast Cancer Based on Differential Gene Expression and Pathway Network. *Front. Cell. Dev. Biol.* **2021**, *9*, 811585. [CrossRef]
76. Zhang, S.; Wang, Y.; Gu, Y.; Zhu, J.; Ci, C.; Guo, Z.; Chen, C.; Wei, Y.; Lv, W.; Liu, H.; et al. Specific breast cancer prognosis-subtype distinctions based on DNA methylation patterns. *Mol. Oncol.* **2018**, *12*, 1047–1060. [CrossRef]
77. Lv, H.; Zhang, Y.; Wang, J.S.; Yuan, S.S.; Sun, Z.J.; Dao, F.Y.; Guan, Z.X.; Lin, H.; Deng, K.J. iRice-MS: An integrated XGBoost model for detecting multitype post-translational modification sites in rice. *Brief. Bioinform.* **2022**, *23*, bbab486. [CrossRef]
78. Naseer, S.; Hussain, W.; Khan, Y.D.; Rasool, N. NPalmitoylDeep-pseaac: A predictor of N-Palmitoylation Sites in Proteins Using Deep Representations of Proteins and PseAAC via Modified 5-Steps Rule. *Curr. Bioinform.* **2021**, *16*, 294–305. [CrossRef]
79. Ao, C.; Zou, Q.; Yu, L. NmRF: Identification of multispecies RNA 2′-O-methylation modification sites from RNA sequences. *Brief. Bioinform.* **2022**, *23*, bbab480. [CrossRef]
80. Jin, Q.; Meng, Z.; Tuan, D.P.; Chen, Q.; Wei, L.; Su, R. DUNet: A deformable network for retinal vessel segmentation. *Knowl. -Based Syst.* **2019**, *178*, 149–162. [CrossRef]
81. Manavalan, B.; Basith, S.; Shin, T.H.; Wei, L.; Lee, G. Meta-4mCpred: A Sequence-Based Meta-Predictor for Accurate DNA 4mC Site Prediction Using Effective Feature Representation. *Mol. Ther. Nucleic Acids* **2019**, *16*, 733–744. [CrossRef]
82. Su, R.; Liu, X.; Wei, L.; Zou, Q. Deep-Resp-Forest: A deep forest model to predict anti-cancer drug response. *Methods* **2019**, *166*, 91–102. [CrossRef]
83. Wei, L.; Xing, P.; Zeng, J.; Chen, J.; Su, R.; Guo, F. Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artif. Intell. Med.* **2017**, *83*, 67–74. [CrossRef] [PubMed]
84. An, Q.; Yu, L. A heterogeneous network embedding framework for predicting similarity-based drug-target interactions. *Brief. Bioinform.* **2021**, *22*, bbab275. [CrossRef] [PubMed]
85. Zulfiqar, H.; Yuan, S.S.; Huang, Q.L.; Sun, Z.J.; Dao, F.Y.; Yu, X.L.; Lin, H. Identification of cyclin protein using gradient boost decision tree algorithm. *Comput. Struct. Biotechnol. J.* **2021**, *19*, 4123–4131. [CrossRef]
86. Zunair, H.; Ben Hamza, A. Sharp U-Net: Depthwise convolutional network for biomedical image segmentation. *Comput. Biol. Med.* **2021**, *136*, 104699. [CrossRef]
87. Zhu, Q.; Fan, Y.; Pan, X. Fusing Multiple Biological Networks to Effectively Predict miRNA-disease Associations. *Curr. Bioinform.* **2021**, *16*, 371–384. [CrossRef]
88. Yu, L.; Wang, M.; Yang, Y.; Xu, F.; Zhang, X.; Xie, F.; Gao, L.; Li, X. Predicting therapeutic drugs for hepatocellular carcinoma based on tissue-specific pathways. *PLoS Comput. Biol.* **2021**, *17*, e1008696. [CrossRef] [PubMed]
89. Wang, X.; Yang, Y.; Liu, J.; Wang, G. The stacking strategy-based hybrid framework for identifying non-coding RNAs. *Brief. Bioinform.* **2021**, *22*, bbab023. [CrossRef]
90. Jiang, Q.H.; Wang, G.H.; Jin, S.L.; Li, Y.; Wang, Y.D. Predicting human microRNA-disease associations based on support vector machine. *Int. J. Data Min. Bioinform.* **2013**, *8*, 282–293. [CrossRef] [PubMed]
91. Linder, A.; Arnold, R.; Boyd, J.H.; Zindovic, M.; Zindovic, I.; Lange, A.; Paulsson, M.; Nyberg, P.; Russell, J.A.; Pritchard, D.; et al. Heparin-Binding Protein Measurement Improves the Prediction of Severe Infection With Organ Dysfunction in the Emergency Department. *Crit. Care Med.* **2015**, *43*, 2378–2386. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.