

Article

3Dloop-FPSSM: Predicting Protein–Protein Interactions by Fusing 3D Local Optimal Oriented Pattern and Folded Position-Specific Scoring Matrix

Fangfang Bai ^{1,2} , Dan Liu ^{1,2}, Guangxian Wang ^{1,2}, Rao Mo ^{1,2} and Peng Zhang ^{1,2,*}

¹ Faculty of Science, Kunming University of Science and Technology, Kunming 650500, China; baifangfang@stu.kust.edu.cn (F.B.); 20232111016@stu.kust.edu.cn (D.L.); 20242111021@stu.kust.edu.cn (G.W.); morao@stu.kust.edu.cn (R.M.)

² Center for Applied Statistics, Kunming University of Science and Technology, Kunming 650500, China

* Correspondence: zhangpeng@kust.edu.cn

Abstract

Protein–protein interactions (PPIs) are fundamental to cellular processes, and understanding their mechanisms aids in disease diagnosis, drug target identification, and therapeutic development. Traditional experimental methods for PPI detection are costly and time-consuming, highlighting the need for efficient computational tools. In this study, we introduce a novel sequence-based framework for PPI prediction, which combines position-specific scoring matrices (PSSMs), 3D local optimal orientation patterns (3Dloop), and histogram gradient boosting (HistGB). Protein sequences are first transformed into PSSMs to capture evolutionary conservation, which are then processed into folded PSSMs (FPSSMs) to reveal hidden relationships among discontinuous amino acids. High-dimensional features are extracted using 3Dloop and classified with HistGB. We demonstrate the superiority of this method over random forest (RF) and support vector machine (SVM) models, achieving accuracies of 95.61% on the *yeast* dataset and 89.93% on the *Helicobacter pylori* dataset. Ablation studies confirm the effectiveness of each component in the framework. The results show that our approach provides a reliable and efficient solution for PPI prediction.

Keywords: histogram gradient boosting; position-specific scoring matrix; random forest; support vector machine; 3D local optimal oriented pattern

1. Introduction

Protein–Protein Interactions (PPIs) play a crucial role in almost all cellular processes and functions, such as homeostasis regulation, signal transduction, organogenesis, and gene expression [1–5]. Accurate detection and characterization of PPI interaction modes are key prerequisites for deciphering complex biological networks and elucidating the essence of life activities.

Although high-throughput experimental techniques such as affinity purification-mass spectrometry [3] and *yeast* two-hybrid [6,7] have been widely applied in PPI identification, these methods generally suffer from drawbacks including cumbersome operational procedures, high experimental costs, and long detection cycles. Moreover, they are susceptible to interference from experimental conditions, leading to relatively high false-positive results, which makes it difficult to meet the demand for large-scale and high-precision PPI analysis. Therefore, the development of efficient computational prediction methods based on pro-



Academic Editor: Alicia Cordero

Received: 6 May 2026

Revised: 19 May 2026

Accepted: 25 May 2026

Published: 3 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

tein sequence information has become a critical need to make up for the shortcomings of experimental techniques and accelerate the progress of PPI research.

In recent years, researchers have proposed various PPI prediction methods based on different information sources, including tertiary structure, genomic information, phylogenetic profiles, protein domains, and secondary structure [8–11]. However, prior knowledge is only available for a small subset of target proteins, making the aforementioned methods inapplicable in most cases. With the advancement of sequencing technologies, an increasing number of genomic sequences have been obtained [12–14], leading to the successive proposal of a series of computational models based on protein sequence features for PPI prediction, such as those utilizing transition descriptors, Zernike moment descriptors, and combined triplet (CT) local descriptors [15–18]. Nevertheless, these methods that rely solely on protein primary sequence information often have limited prediction accuracy due to issues such as single feature dimension and lack of evolutionary conservation information. Against this backdrop, prediction strategies that integrate the Position-Specific Scoring Matrix (PSSM) to capture protein evolutionary features have gradually become a research focus in this field. Numerous studies have confirmed that PSSM can effectively integrate the evolutionary conservation information of homologous sequences, providing more biologically meaningful feature support for PPI prediction.

Currently, such PSSM-based PPI prediction methods can be categorized into three core technical branches: The first branch involves the integration of deep learning and mathematical transformation. It takes deep neural networks as the core and combines mathematical transformations to mine high-dimensional nonlinear features from PSSM. For instance, Pan et al. combined deep neural networks (DNN) with discrete Hilbert transform (DHT) and singular value decomposition (SVD) to achieve accurate PPI identification [19]; An et al. proposed the RRN-SIFT model, which integrates recurrent neural networks (RNN) with scale-invariant feature transform (SIFT) for PPI prediction [20]. The second branch focuses on traditional machine learning and feature enhancement. It adopts classical machine learning models and enhances the information representation capability of the PSSM through moment descriptors, feature extraction algorithm optimization, or multi-feature fusion. For example, Wang constructed an identification system by combining probabilistic classification vector machines (PCVMs) with Legendre moment descriptors [21]; Li et al. used the MatFLDA feature extraction algorithm to mine evolutionary information [22]; An et al. proposed the LCPSSMMF method and improved PSSM representation capability through multi-feature fusion [23]; An et al. developed the BIGP algorithm and combined it with relevance vector machines (RVM) to achieve efficient classification of PSSM features [24]. The third branch adopts ensemble frameworks represented by Rotation Forest (ROF). Previous studies have employed ROF as a classifier to handle high-dimensional features and achieve satisfactory robustness and anti-overfitting performance [25,26]. For example, Pan et al. constructed the DSI-ROF model by integrating discrete sine transform (DST), the PSSM, and the ROF classifier [25]; Zhan et al. extracted features from PSSMs using locality preserving projection (LPP) and fed them into the ROF classifier for prediction [26].

Although the aforementioned PSSM-based methods have significantly improved PPI prediction performance, existing strategies generally share a key limitation: they all fail to fully consider the potential correlations between discontinuous amino acids in protein sequences. From the perspective of protein spatial conformation, amino acid residues that are far apart in the primary sequence can approach each other in three-dimensional space through the folding and coiling of polypeptide chains and participate in the protein interaction process collaboratively [27]. Such spatially adjacent amino acid residue combinations often contain key information closely related to PPIs, and the neglect

of this type of information by existing methods has become a major bottleneck for further improving prediction accuracy.

To address this issue, this study proposes a novel PSSM-based PPI prediction framework, with its core workflow outlined as follows: ① Convert raw protein sequences into PSSMs that capture evolutionary conservation information; ② Fold the 2D PSSMs into 3D Folded Position-Specific Scoring Matrices (FPSSMs) using fixed rules, to capture spatial correlation information of discontinuous amino acids; ③ Extract high-dimensional feature vectors from FPSSMs using the 3Dloop descriptor, to mine PPI-related features hidden in spatial conformations; ④ Train a robust Histogram Gradient Boosting (HistGB) classifier on the extracted features, to identify potential PPIs.

The technical designs and methodological innovations of this work are summarized as follows:

1. Introduction of a 3D folding operation for PSSMs.

Unlike conventional schemes that extract features only from adjacent amino acids, the proposed operation projects the 2D position-specific scoring matrix (PSSM) into a three-dimensional folded structure (FPSSM). This design enables the simultaneous representation of both adjacent and non-consecutive amino acid correlations within the same feature input space, thereby incorporating discontinuous evolutionary information into the subsequent feature extraction process.

2. Development of a 3Dloop feature extraction module.

A feature extraction workflow tailored to the folded PSSM (FPSSM) is constructed. This workflow builds a vertex-center pixel matrix within a $3 \times 3 \times 3$ neighborhood and employs a loop-based encoding scheme to jointly extract continuous and discontinuous structural information. The module outputs a high-dimensional feature vector that serves as input to the downstream classifier.

3. Construction of a multi-species experimental pipeline.

The proposed framework is evaluated on multiple benchmark datasets covering different species (*yeast*, *Helicobacter pylori*, human, and mouse) using a consistent nested cross-validation protocol. This pipeline is designed to assess both intra-species prediction performance and cross-species generalizability under controlled experimental conditions.

2. Materials and Methods

2.1. Datasets

The selection of appropriate datasets is crucial for reliably evaluating the performance of the proposed prediction framework. In this study, four distinct protein–protein interaction (PPI) datasets are employed to assess the validity and generalization capability of our model. Two standard benchmark datasets are initially used for quantitative evaluation. The *yeast* PPI dataset is obtained from the DIP database [28]. To ensure data quality and minimize bias, protein pairs with sequences shorter than 50 amino acids and those exhibiting more than 40% sequence identity are filtered out during preprocessing. The refined *yeast* dataset ultimately consists of 11,888 protein pairs, evenly split into 5594 interacting pairs and 5594 non-interacting pairs. The second benchmark dataset, focused on *Helicobacter pylori*, was originally reported by Martin et al. [29], and includes 2916 protein pairs, consisting of 1458 positive samples and 1458 negative samples.

To further evaluate the practical prediction capacity of the proposed method, two additional independent datasets derived from *Homo sapiens* and *Mus musculus* are also incorporated. Notably, these datasets contain only experimentally verified positive interactions: 1412 pairs for humans and 313 pairs for mice, with no corresponding negative samples [28]. Given the absence of negative non-interaction samples, these datasets are

used exclusively for cross-species positive sample verification rather than for standard binary classification tasks. The key statistical characteristics of all datasets utilized in this study are summarized in Table 1.

Table 1. Protein–Protein Interaction Datasets Across Four Species.

Species	Positive Parts	Negative Pairs
<i>Yeast</i>	5594	5594
<i>H. pylori</i>	1458	1458
<i>M. musculus</i>	313	0
<i>H. sapiens</i>	1412	0

2.2. Position-Specific Scoring Matrix (PSSM)

The Position-Specific Scoring Matrix (PSSM) was originally developed for detecting distantly related proteins and can be generated from a set of protein sequences [30–32]. For a given protein sequence, the PSSM is defined as an $N \times 20$ matrix, as shown in Equation (1):

$$D = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \dots & \alpha_{1,20} \\ \alpha_{2,1} & \alpha_{2,2} & \dots & \alpha_{2,20} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N,1} & \alpha_{N,2} & \dots & \alpha_{N,20} \end{bmatrix}, \quad (1)$$

where N denotes the length of the protein sequence, and 20 corresponds to the 20 types of amino acids. The score of the j -th amino acid at the i -th position in the query protein sequence, denoted as α_{ij} , is given by the PSSM, as shown in Equation (2):

$$\alpha_{ij} = \sum_{k=1}^{20} m(i, k) \times n(j, k), \quad (2)$$

Here, $m(i, k)$ represents the frequency ratio of the k -th amino acid at the i -th position in the probe sequence, and $n(j, k)$ is the value of the j -th and k -th amino acids in the *Dayhoff* mutation matrix. A higher score indicates stronger conservation at this position, while a lower score indicates weaker conservation.

In our experiments, we employed the position-specific iterative PSI-BLAST (NCBI BLAST+ 2.17.0, URL: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>, accessed on 23 September 2025) and the Swiss-Prot database to construct the PSSM for each protein amino acid sequence [33–35]. The code for generating the PSSM is available in Code S1 of the Supplementary Materials. To ensure the high reliability of homologous sequences, we set the number of iterations to 3 and the e -value to 0.001, while all other parameters were set to their default values.

2.3. Folded Position-Specific Scoring Matrix (FPSSM)

Protein folding refers to the process by which a linear amino acid sequence adopts a specific three-dimensional conformation, driven by intermolecular forces such as hydrophobic interactions, hydrogen bonds, and ionic bonds. Folding patterns are primarily classified based on the arrangement of secondary structure elements:

- α -helix: Stabilized by intra-chain hydrogen bonds, forming a right-handed helical structure;
- β -sheet: Composed of multiple extended peptide chains linked by inter-chain hydrogen bonds, forming a zigzag sheet;
- β -turn: Enables a 180° rapid reversal of the polypeptide chain [36], commonly found on protein surfaces;

- Random coil: Lacks a fixed secondary structure and often acts as a linker between structural elements [37].

These folding patterns collectively determine the spatial structure of proteins, which is directly linked to their biological activity. For example, keratin, which is enriched in α -helices, exhibits exceptional mechanical strength; membrane proteins with β -barrel structures form pores for substance transport; β -turns assist globular proteins in folding into compact 3D structures and are often involved in protein–ligand binding processes.

To simplify the model and focus on core structural features, this study selects three typical protein folding patterns— α -helices, β -sheets, and β -turns—as the primary research focus. Random coils are excluded due to their loose structures, highly dynamic conformations, and lack of stable sequence correlation rules, making their characteristics challenging to capture accurately through sequence information. Furthermore, random coils contribute relatively little to the core functional conformations of proteins. Previous studies have demonstrated that the formation of these three folding patterns is sequence-dependent [38–41]. When characteristic amino acid fragments (illustrated in Figure 1) appear in the protein sequence, they fold into specific spatial conformations following defined rules, which are standardized as the folding modes shown in Figure 2.

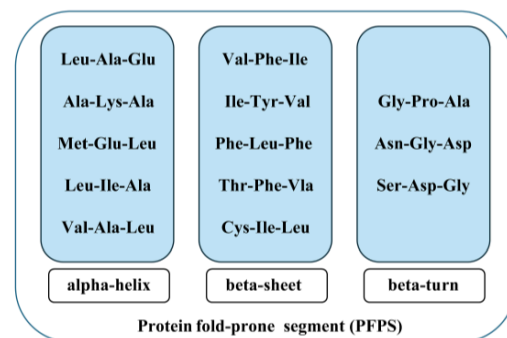


Figure 1. Correspondence between Protein Fold-Prone Segments (PFPSs) and three types of protein secondary structures including α -helix, β -sheet, and β -turn.

Subsequently, this folding-induced structural information is integrated with the PSSM features described earlier, following the workflow depicted in Figure 3. This integration ensures that each amino acid residue retains its 20-dimensional PSSM evolutionary conservation score while incorporating structural folding constraints, ultimately generating the Folded Position-Specific Scoring Matrix (FPSSM). For the detailed algorithmic procedures, refer to Algorithm 1.

It is important to note that the proposed FPSSM folding process is a heuristic, sequence-based construction, rather than a prediction or reconstruction of the actual protein 3D structure. Specifically, the 13 tripeptide fragments used to identify fold-prone segments (5 for α -helix, 5 for β -sheet, and 3 for β -turn) were curated exclusively from established biochemical literature, based on their known secondary structure propensities. These fragments are not derived from, nor statistically optimized on, our training data. The odd-even folding rule follows a fixed heuristic encoding strategy, similar to the fixed-radius neighborhood in Local Binary Patterns, rather than being a data-driven parameter. As a result, the same folding procedure is independently applied to each protein sample, ensuring no cross-sample information exchange. This design guarantees that the FPSSM generation process does not introduce data leakage between the training and test sets.

It should be emphasized that this heuristic feature construction strategy differs fundamentally from experimental protein structure determination and does not aim to reproduce biologically native folding conformations. The FPSSM representation is a computational

feature-engineering design for capturing discontinuous residue correlations, rather than a structural model intended for biological or structural inference. Therefore, the proposed approach remains a sequence-based computational predictor and should not be overinterpreted as a method for characterizing real protein three-dimensional structures.

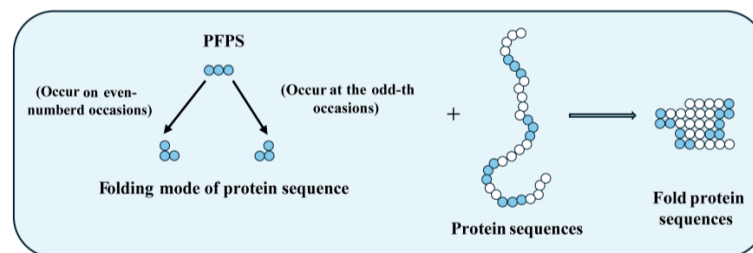


Figure 2. Odd-even folding rule used to convert 2D position-specific scoring matrix (PSSM) into 3D folded position-specific scoring matrix (FPSSM) based on protein fold-prone segments. Odd-even folding rule used to convert 2D position-specific scoring matrix (PSSM) into 3D folded position-specific scoring matrix (FPSSM) based on protein fold-prone segments.

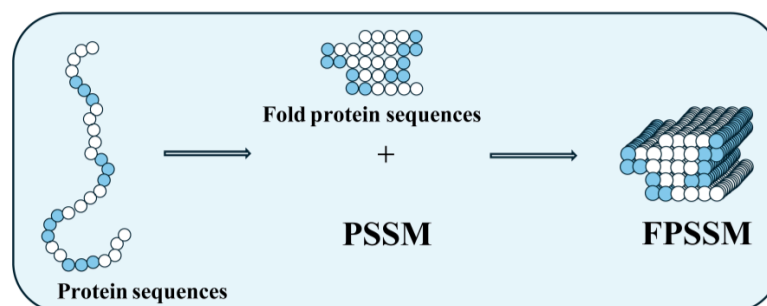


Figure 3. Schematic illustration of mapping a 2D PSSM profile into a 3D folded FPSSM structure for capturing discontinuous amino acid relationships. Odd-even folding rule used to convert 2D position-specific scoring matrix (PSSM) into 3D folded position-specific scoring matrix (FPSSM) based on protein fold-prone segments.

Algorithm 1: Folded Position-Specific Scoring Matrix (FPSSM)

Input: $pssm \in \mathbb{R}^{(sequence_length \times feature_dim)}$, sequence (length = sequence_length)
Output: $grid \in \mathbb{R}^{(height \times width \times feature_dim)}$, coordinate list coords

- 1: target_triplet_positions $\leftarrow$ find_target_triplet_positions(sequence)
- 2: coords $\leftarrow [(0,0)]$, direction $\leftarrow (0,1)$, index $\leftarrow 1$, count $\leftarrow 0$
- 3: while index < sequence_length do
- 4: if index $\in$ target_triplet_positions then
- 5: count $\leftarrow$ count + 1
- 6: append coords[-1] + direction to coords; index $\leftarrow$ index + 1
- 7: direction $\leftarrow (1,0)$; append coords[-1] + direction to coords; index $\leftarrow$ index + 1
- 8: direction $\leftarrow (0,-1)$ if count is odd else (0,1)
- 9: append coords[-1] + direction to coords; index $\leftarrow$ index + 1
- 10: else
- 11: append coords[-1] + direction to coords; index $\leftarrow$ index + 1
- 12: end
- 13: end
- 14: h_min, h_max $\leftarrow$ min/max of row in coords; w_min, w_max $\leftarrow$ min/max of col in coords
- 15: height $\leftarrow$ h_max - h_min + 1, width $\leftarrow$ w_max - w_min + 1

Algorithm 1: *Cont.*

```

16: grid ← zeros(height, width, feature_dim)
17: for i ← 0 to sequence_length − 1 do
18:   (h,w) ← coords[i]; grid[h − h_min][w − w_min][:] ← pssm[i][:]
19: return grid, coords

```

2.4. 3D Local Optimal Orientation Pattern (3Dloop)

Tapabrata et al. proposed the Local Optimal Orientation Pattern (LOOP), a novel binary local pattern descriptor that integrates rotation invariance into the core design logic of binary local descriptors [42]. The LOOP descriptor is an improved design based on the Local Binary Pattern (LBP) [43,44] and Local Directional Pattern (LDP) [45].

For a given image I , let i_c denote the grayscale value of the pixel at (x_c, y_c) . Assume $i_n (n = 0, 1, \dots, 7)$ represents the grayscale value of a pixel (excluding the central pixel i_c) within the 3×3 neighborhood of (x_c, y_c) . Figure 4 illustrates the eight-directional Kirsch edge detector centered at (x_c, y_c) . Let $m_n (n = 0, 1, \dots, 7)$ be the eight response values of the Kirsch mask, corresponding to the pixels with grayscale values $i_n (n = 0, 1, \dots, 7)$. Suppose m_k is the m -th largest Kirsch activation value; an exponential weight ω_n is assigned to the n -th pixel based on the ranking of m_n amplitudes among the eight Kirsch mask outputs. Finally, the LOOP value of the pixel (x_c, y_c) is calculated as Equation (3):

$$LOOP(x_c, y_c) = \sum_{n=0}^7 s(i_n - i_c) 2^{\omega_n}. \quad (3)$$

where

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

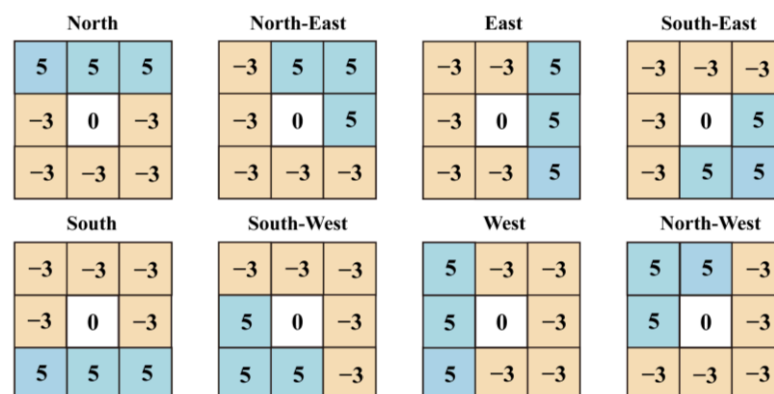


Figure 4. Eight directional Kirsch edge detection masks applied for local feature extraction in the LOOP and 3Dloop descriptors.

In this study, the input PSSM is a $20 \times M$ matrix; thus, each protein sequence is represented as a 256-dimensional feature vector after processing with the LOOP descriptor.

To handle three-dimensional data, we extended the LOOP descriptor to develop the 3D Local Optimal Orientation Pattern (3Dloop), whose core is to extract dual features of discrete vertices and continuous regions from a $3 \times 3 \times 3$ neighborhood. For a 3D image I , let i_c denote the grayscale value of the pixel at (x_c, y_c, z_c) . Assume $i_n (n = 0, 1, \dots, 26)$ represents the grayscale value of a pixel (excluding the central pixel i_c) within the $3 \times 3 \times 3$ neighborhood of (x_c, y_c, z_c) . The detailed implementation procedure of the dual feature extraction is summarized in Algorithm 2. We first constructed a new 3×3 matrix using the eight vertex pixels from the $3 \times 3 \times 3$ neighborhood of i_c and the central pixel i_c . Then, we

applied the LOOP method to extract features from these non-adjacent data points, generating a new 256-dimensional feature vector. This vector was concatenated with the 256-dimensional feature vector derived from the original continuous data, resulting in a 512-dimensional feature vector. In this way, the features capture both continuous local information and discrete geometric information. The detailed process is illustrated in Figure 5.

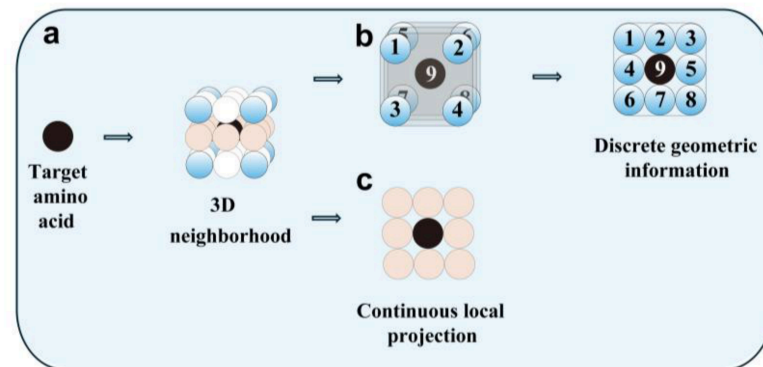


Figure 5. Dual feature extraction mechanism of the 3D Local Optimal Orientation Pattern (3Dloop). (a) 3D neighborhood construction around the target amino acid (black circle). (b) Discrete geometric information extraction via mapping to a 3×3 grid, where blue circles represent neighboring amino acids and the black circle denotes the target amino acid. (c) Continuous local projection of the 3D neighborhood, where light-colored circles represent neighboring amino acids and the black circle denotes the target amino acid. The mechanism captures both continuous local and discrete geometric information from 3D neighborhood data.

Traditional 3D feature methods, such as 3D-LBP, have a limitation: they perform uniform encoding over the entire $3 \times 3 \times 3$ neighborhood [44], failing to capture the “continuous local information” and “discrete geometric information” inherent in the FPSSM structure. In contrast, 3Dloop isolates a discrete subset of “eight-vertex pixels” and extracts their features using the LOOP descriptor independently. These features are then concatenated with those from the full neighborhood. This “dual-feature concatenation” simultaneously captures both continuous local and discrete geometric information, offering a more accurate representation of the feature distribution in protein 3D structures.

Moreover, traditional 3D-LBP relies on a fixed traversal order for the neighborhood (e.g., traversing voxels in an $X \rightarrow Y \rightarrow Z$ sequence). When rotated, this order is disrupted, directly altering the binary code. In contrast, 3Dloop employs “intra-class rank-based weighting,” assigning weights by sorting the 3D Kirsch response intensities. After rotation, the azimuth priority and response rank relationships of the neighborhood remain unchanged, ensuring that the binary features are inherently invariant to rotation. This eliminates the need for additional post-processing (such as reordering binary bits) and theoretically resolves rotational variance, highlighting the advantage of rotation invariance.

As another example, the LNBP (Local Normal Binary Pattern) [46] achieves rotation invariance by utilizing the normal directions (geometric poses) of 3D voxels. It classifies the normal directions of neighboring voxels into “angular intervals relative to the central voxel’s normal” and encodes these intervals into binary features. However, this approach discards voxel “intensity (grayscale value)” and “texture (local grayscale variation)” information entirely [46,47], as further detailed in Table 2.

In summary, 3Dloop extracts and concatenates features in a partitioned manner using the “eight-vertex pixel + full neighborhood” strategy, effectively capturing both continuous local information and discrete geometric information. It incorporates inherent rotation invariance through “intra-class rank-based weighting,” ensuring robustness against rotational transformations. Additionally, unlike the LNBP, which discards intensity and texture

information, 3Dloop preserves these subtle feature differences within the FPSSM structure. As a result, 3Dloop offers a more comprehensive representation of information, improved rotational robustness, and enhanced feature discriminability.

Table 2. Qualitative comparison of 3D descriptors for FPSSM-based feature extraction.

Comparison Metric	3Dloop	3D-LBP	LNBP
Captured information	Captures two complementary types of information from the $3 \times 3 \times 3$ neighborhood: continuous local patterns from all voxels, and discrete geometric patterns from the 8-vertex subset.	Encodes the entire $3 \times 3 \times 3$ neighborhood uniformly, without explicitly separating continuous and geometric components.	Focuses only on geometric pose (normal direction), discarding voxel intensity and texture information.
Rotation handling	Achieves invariance via intra-class rank-based weighting of 3D Kirsch responses. Features remain invariant after rotation without additional post-processing.	Uses a fixed neighborhood traversal order (e.g., $X \rightarrow Y \rightarrow Z$). Rotation disrupts this order, altering the binary code and introducing rotational variance.	Achieves invariance by classifying normal angle intervals relative to the central voxel, but requires additional computation of normal directions.
Potential relevance to protein structural representation	Dual-information concatenation may better align with protein 3D structural patterns.	Single-neighborhood encoding has limited alignment with protein structural feature distributions.	Loss of intensity/texture information weakens correlation with protein 3D structure features.

Algorithm 2: PSSM-LOOP Feature Extraction

Information Capture Capability

Rotation Invariance

Feature Relevance to Protein Structure

Input: PSSM $P \in \mathbb{R}^{L \times F}$, sequence seq

Output: feature vector $\in \mathbb{R}^{512}$

```

1: hist_orig, hist_imp  $\leftarrow$  zeros(256)
2: // 2D LOOP on original PSSM
3: for each valid element in P (with symmetric padding) do
4:    $M_2 \leftarrow 3 \times 3$  neighborhood patch
5:   ComputeLOOP( $M_2$ , hist_orig)
6: // 3D folded LOOP
7:  $V \leftarrow \text{fold\_pssm\_to\_3d\_irregular}(P, \text{seq})$ 
8: for each valid grid point & channel in V (with symmetric padding) do
9:   cube  $\leftarrow 3 \times 3 \times 3$  neighborhood
10:   $M_2 \leftarrow$  map 8 corners + center of cube into a  $3 \times 3$  matrix
11:  ComputeLOOP( $M_2$ , hist_imp)
12: feat_orig  $\leftarrow \text{hist\_orig} / \Sigma(\text{hist\_orig})$ 
13: feat_imp  $\leftarrow \text{hist\_imp} / \Sigma(\text{hist\_imp})$ 
14: return concat(feat_orig, feat_imp)
Subroutine ComputeLOOP(mat  $\in \mathbb{R}^{(3 \times 3)}$ , hist):
1: c  $\leftarrow$  mat[1,1] // center pixel
2: m  $\leftarrow$  convolution responses of mat with 8 Kirsch masks
3: weights  $\leftarrow$  ranks (7..0) assigned to directions sorted by |m| descending
4: loop_val  $\leftarrow$  0
5: for  $n = 1.8$  (fixed spatial order) do
6:   b  $\leftarrow$  1 if neighbor[n]  $\geq$  c else 0
7:   loop_val += b << weights[direction_of_neighbor[n]]
8: hist[loop_val] += 1

```

2.5. Histogram Gradient Boosting

Histogram Gradient Boosting (HistGB) is an efficient ensemble learning algorithm based on Gradient Boosting Decision Trees (GBDTs) [47]. By binning continuous features into histograms, it significantly improves training speed and memory efficiency [48] while maintaining competitive predictive performance, making it particularly suitable for large-scale datasets.

The core idea of HistGradientBoosting is to use histogram approximation to replace the precise sorting and splitting calculations for each feature value in traditional GBDTs within the gradient boosting framework, thereby substantially reducing computational complexity. The model framework and training process are described as follows:

Let X be an $N \times n$ matrix of training samples, where N denotes the number of samples and n denotes the number of features. Let the target vector $Y = [y_1, y_2, \dots, y_N]^T$ represent the corresponding target values (continuous values for regression tasks, class labels for classification tasks). Let M be the number of base decision trees in the ensemble model, denoted as $T_1, T_2, \dots, T_M$. It should be noted that the number of base trees M , learning rate η , maximum depth d , and number of histogram bins B are important hyperparameters for tuning HistGradientBoosting. To ensure robustness and optimality, a systematic sensitivity analysis was performed over learning rates (0.01, 0.05, 0.1), maximum tree depths (4, 6, 8, 10, 12), histogram bins (32, 64, 128, 255), and max iterations (100, 300, 500, 700). Based on the highest mean accuracy, AUC, and minimal performance variance observed in the sensitivity analysis, we set these hyperparameters in our experiments as follows: the number of base trees M (max_iter) was set to 500, the learning rate η was set to 0.05, the maximum depth d of each tree was set to 8, and the number of histogram bins B was set to 255. The choice of 500 iterations avoids redundant computation while achieving performance saturation; a learning rate of 0.05 balances convergence speed and generalization; a maximum depth of 8 ensures sufficient model capacity without overfitting; and 255 bins provide sufficient resolution without excessive memory overhead. All other hyperparameters were kept at their default settings to ensure reproducibility.

The training process for a single base tree T_m is as follows:

1. Model initialization

For regression tasks, the initial prediction value is set to the mean of the target values; for classification tasks, the initial prediction value is set to an initial logit (e.g., 0), denoted as $F_0(x)$.

2. Iterative training of base trees

For the m iteration ($m = 1, 2, \dots, M$):

- (1) Calculate the pseudo-residuals as shown in Equation (5):

$$r_{i,m} = - \left[\frac{\partial \mathcal{L}(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)} \right] \quad (5)$$

where $\mathcal{L}$ is the loss function (e.g., squared loss, log loss).

- (2) Divide the value range of each feature into B histogram bins of equal width or frequency, map the samples to the corresponding bins, and count the number of samples and the sum of pseudo-residuals within each bin.
- (3) Based on the histogram statistics, find the optimal split point for each feature to minimize the sum of squared pseudo-residuals in the child nodes after splitting, and construct the decision tree $T_m(x)$.
- (4) Update the model with the learning rate η , and the update formula is shown in Equation (6):

$$F_m(x) = F_{m-1}(x) + \eta \cdot T_m(x). \quad (6)$$

3. Feature importance calculation

Feature importance ranking is obtained by counting the number of times each feature is used as a splitting feature and the split gain across all base trees, which is used for feature selection and model interpretation.

In the prediction phase, given a test sample x , the prediction value of the ensemble model is the weighted sum of the predictions from all base trees, as shown in Equation (7):

$$F(x) = F_0(x) + \eta \cdot T_m(x). \quad (7)$$

For classification tasks, the prediction values are converted into class probabilities via the Sigmoid or Softmax function, and the test sample is finally assigned to the class with the highest probability.

2.6. Model Validation Strategy

To prevent data leakage and ensure unbiased evaluation, we adopted a nested 5-fold Group cross-validation strategy. Each protein pair was assigned a unique group label based on its sample identifier, ensuring that the same protein pair never appeared in both the training and test sets simultaneously. An inner 3-fold Group cross-validation was used for hyperparameter optimization.

Critically, all feature extraction steps—including PSSM loading, FPSSM folding, and 3Dloop descriptor computation—were performed strictly within each outer fold, as detailed below:

1. For each outer fold, the data were split into training and test sets based on group labels.
2. Training phase: Protein names from the training set were used to load the corresponding PSSM files. LOOP features (256-dimensional) were extracted from the original PSSM, and 3Dloop features (256-dimensional) were extracted from the FPSSM, independently for each training sample. These features were then concatenated into 512-dimensional feature vectors.
3. A StandardScaler was fitted on the training features and used to transform the training set.
4. Inner 3-fold Group cross-validation was performed on the training set for hyperparameter tuning.
5. Test phase: Protein names from the test set were used to independently load PSSM files and extract LOOP + 3Dloop features following the same procedure. The scaler, fitted only on the training data, was applied to transform the test features.
6. Predictions were made using the best model obtained from the inner cross-validation.

This strict separation ensures that no information from the test set—whether raw PSSM values, feature statistics, or folding patterns—affects the training phase. As a result, the reported performance metrics accurately reflect the model's genuine generalization ability, rather than being inflated due to data leakage.

2.7. Ablation Study

To evaluate the contribution of each feature component in our framework, an ablation study was conducted using the *yeast* dataset. We constructed three feature configurations:

1. Only LOOP: using only the original 2D LOOP features;
2. Only 3Dloop: using only the 3Dloop features extracted from the FPSSM;
3. LOOP + 3Dloop (full model): fusing LOOP and 3Dloop features as the final input.

All other settings, including classifier parameters, cross-validation strategy, and data preprocessing, were kept consistent to ensure a fair comparison.

2.8. Statistical Analysis

To assess the statistical reliability of performance differences, paired *t*-tests were performed on the 5-fold cross-validation results for both classifier comparisons and ablation studies. All tests were conducted using identical data splits to ensure fairness. The significance threshold was set at $p < 0.05$, and the Bonferroni correction was applied to control the family-wise error rate in multiple comparisons.

3. Results

3.1. Subsection

The predictive performance of the proposed 3Dloop-FPSSM framework was evaluated using the nested 5-fold group cross-validation strategy described in Section 2.6. Unlike standard random K-fold cross-validation, this strategy groups samples by protein-pair identity to reduce data leakage caused by repeated protein-pair information and to ensure a more unbiased assessment of generalization ability. In each outer fold, the training and test samples were strictly separated, while the inner cross-validation loop was used exclusively for hyperparameter optimization within the training set.

The predictions generated by the proposed method were assessed using five evaluation metrics: accuracy (*Accu.*), sensitivity (*Sen.*), precision (*Prec.*), specificity (*Spec.*), and Matthews correlation coefficient (*MCC*). The formulas for these metrics are as follows:

$$Accu. = \frac{TN + TP}{FP + FN + TP + TN} \quad (8)$$

$$Sen. = \frac{TP}{FN + TP} \quad (9)$$

$$Prec. = \frac{TP}{FP + TP} \quad (10)$$

$$Spec. = \frac{TN}{FP + TN} \quad (11)$$

$$MCC. = \frac{TN \times TP - FP \times FN}{\sqrt{(FP + TP)(FP + TN)(TP + FN)(TN + FN)}} \quad (12)$$

Additionally, receiver operating characteristic (ROC) curves were used to visualize and compare the performance of the prediction models. The vertical axis of the ROC curve represents sensitivity, also known as the true positive rate (TPR), while the horizontal axis represents $1 - \text{specificity}$, also known as the false positive rate (FPR). The area under the ROC curve (AUC) is a widely used metric for evaluating classification performance, where an AUC value of 0.5 indicates random classification and an AUC value of 1.0 indicates perfect classification.

For clarity, the detailed workflow of the proposed 3Dloop-FPSSM framework for identifying protein–protein interactions (PPIs) is illustrated in Figure 6, and the full workflow code is provided in Code S2 of the Supplementary Materials.

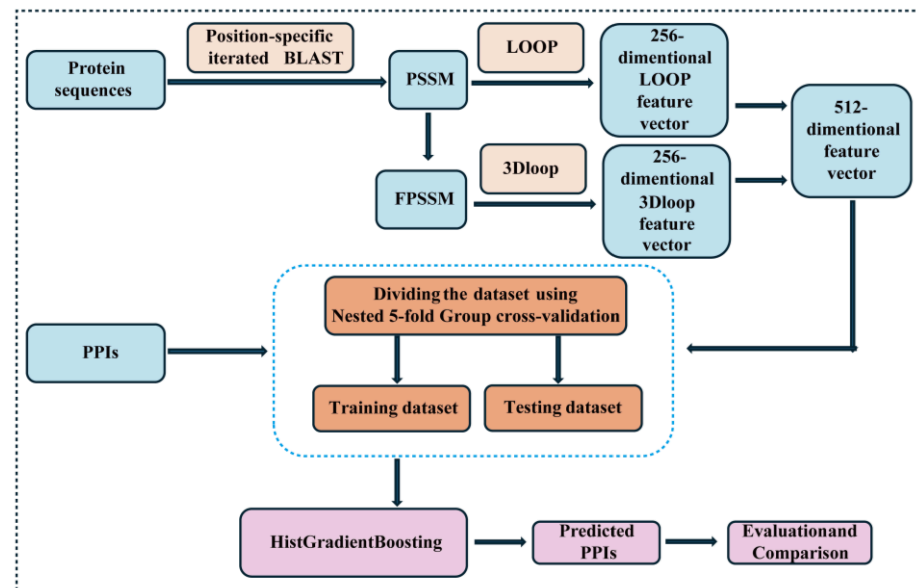


Figure 6. Complete workflow of the proposed 3Dloop-FPSSM framework for protein–protein interaction prediction based on sequence-derived evolutionary and structural features.

3.2. Performance on the Yeast and *H. pylori* Dataset

3.2.1. Baseline Performance of 3Dloop

To evaluate the high predictive performance of the 3Dloop model, experiments were conducted on the *yeast* (*Saccharomyces cerevisiae*) and *Helicobacter pylori* protein–protein interaction (PPI) datasets. To ensure the stability of the prediction results, the nested 5-fold Group cross-validation framework, as described in Section 2.6, was employed to assess the generalization ability of the proposed model.

The experimental results for the *yeast* and *H. pylori* datasets are presented in Table 3, respectively. As shown in the table, the proposed method achieved average accuracies of 95.61% and 89.93%, respectively. To further quantify the predictive performance, additional evaluation metrics were computed, including sensitivity, precision, specificity, Matthews correlation coefficient (MCC), and the area under the receiver operating characteristic curve (AUC).

Table 3. Performance metrics of the proposed method on the *yeast* (Y) and *H. pylori* (H) dataset.

Fold	Accu. (%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC. (%)	AUC
1 (Y)	95.47	95.72	91.56	97.05	88.63	98.29
2 (Y)	95.82	96.15	90.42	97.36	89.17	98.45
3 (Y)	95.53	95.86	91.15	97.12	88.85	98.33
4 (Y)	95.75	96.28	90.74	97.43	89.06	98.51
5 (Y)	95.49	95.64	91.29	96.98	88.79	98.32
Mean ± SD	95.61 ± 0.40	95.93 ± 0.78	91.03 ± 1.26	97.19 ± 0.59	88.90 ± 0.81	98.38 ± 0.27
1 (H)	90.21	87.26	88.97	87.12	86.15	94.82
2 (H)	89.65	87.61	90.12	86.85	86.23	94.36
3 (H)	90.43	87.09	89.35	87.31	87.68	95.01
4 (H)	89.27	87.83	90.64	86.62	85.89	94.15
5 (H)	90.08	87.47	88.43	86.98	85.67	94.59
Mean ± SD	89.93 ± 0.13	87.45 ± 0.91	89.50 ± 0.32	86.98 ± 0.82	86.52 ± 0.65	94.59 ± 0.79

Y: Yeast dataset; H: *Helicobacter pylori* dataset. Accu. = Accuracy, Prec. = Precision, Sen. = Sensitivity, Spec. = Specificity, MCC. = Matthews Correlation Coefficient, AUC = Area Under the ROC Curve. Data are presented as mean ± SD (5-fold nested cross-validation).

Figures 7a and 8a display the ROC curves generated by the 3Dloop method on the *yeast* and *H. pylori* datasets, respectively. The experimental results demonstrate that the 3Dloop method effectively predicts protein–protein interactions (PPIs). Its superior predictive performance can be attributed primarily to the discriminative 3Dloop descriptor and the robust HistGB classifier. Specifically, the position-specific scoring matrix (PSSM) not only encodes protein sequences into matrices but also captures significant evolutionary information, which enhances prediction accuracy. As a widely used ensemble classifier, the HistGB model exhibits exceptional predictive power in identifying potential PPIs, further confirming that the proposed 3Dloop method can serve as a practical tool for PPI prediction.

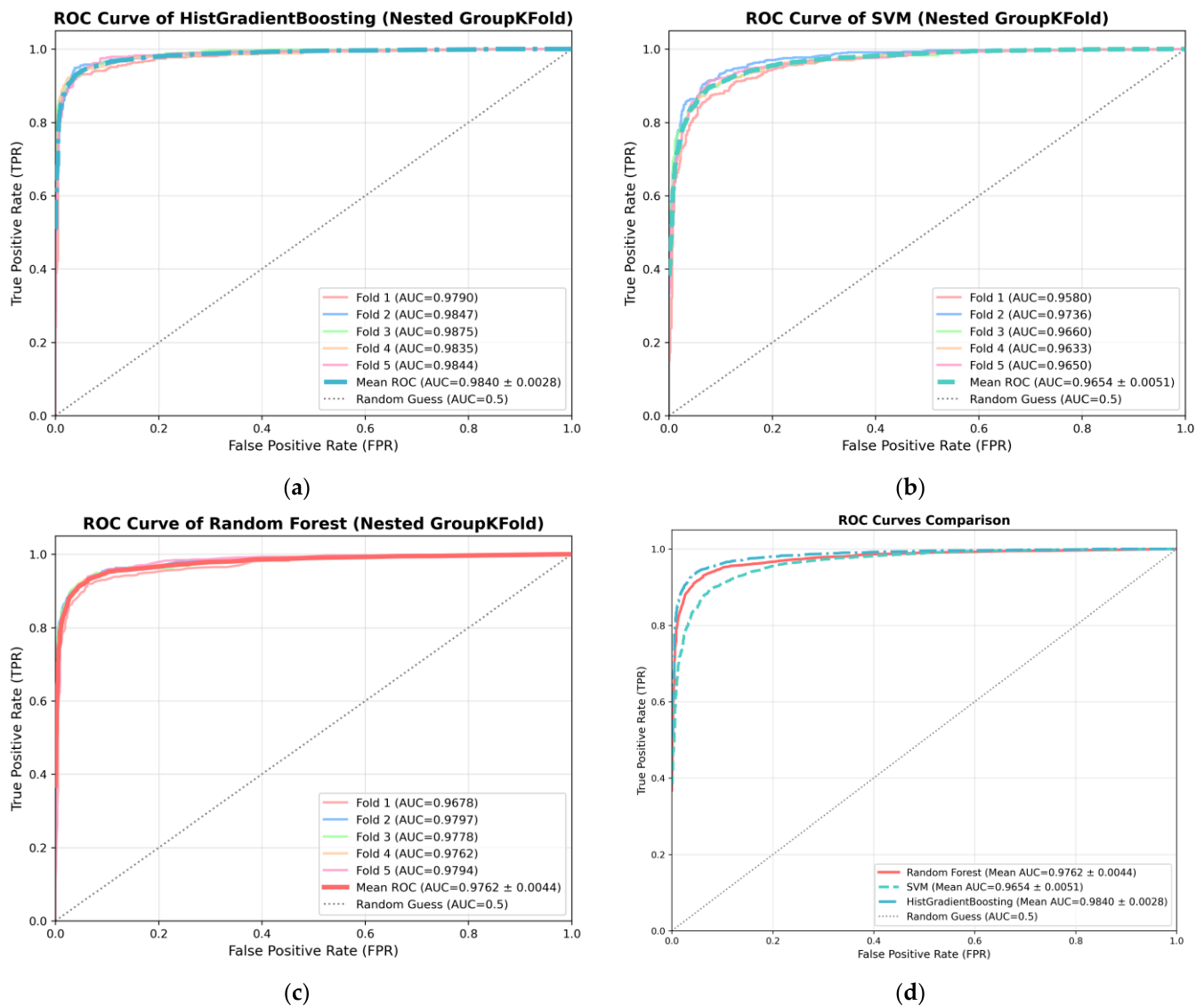


Figure 7. Receiver operating characteristic (ROC) curves for three classifiers evaluated on the *yeast* protein–protein interaction dataset using nested 5-fold group cross-validation. (a) ROC curves of the HistGradientBoosting classifier across five folds; (b) ROC curves of the SVM classifier across five folds; (c) ROC curves of the Random Forest classifier across five folds; (d) Combined comparison of the mean ROC curves of the three classifiers, including the corresponding mean AUC values and their standard deviations.

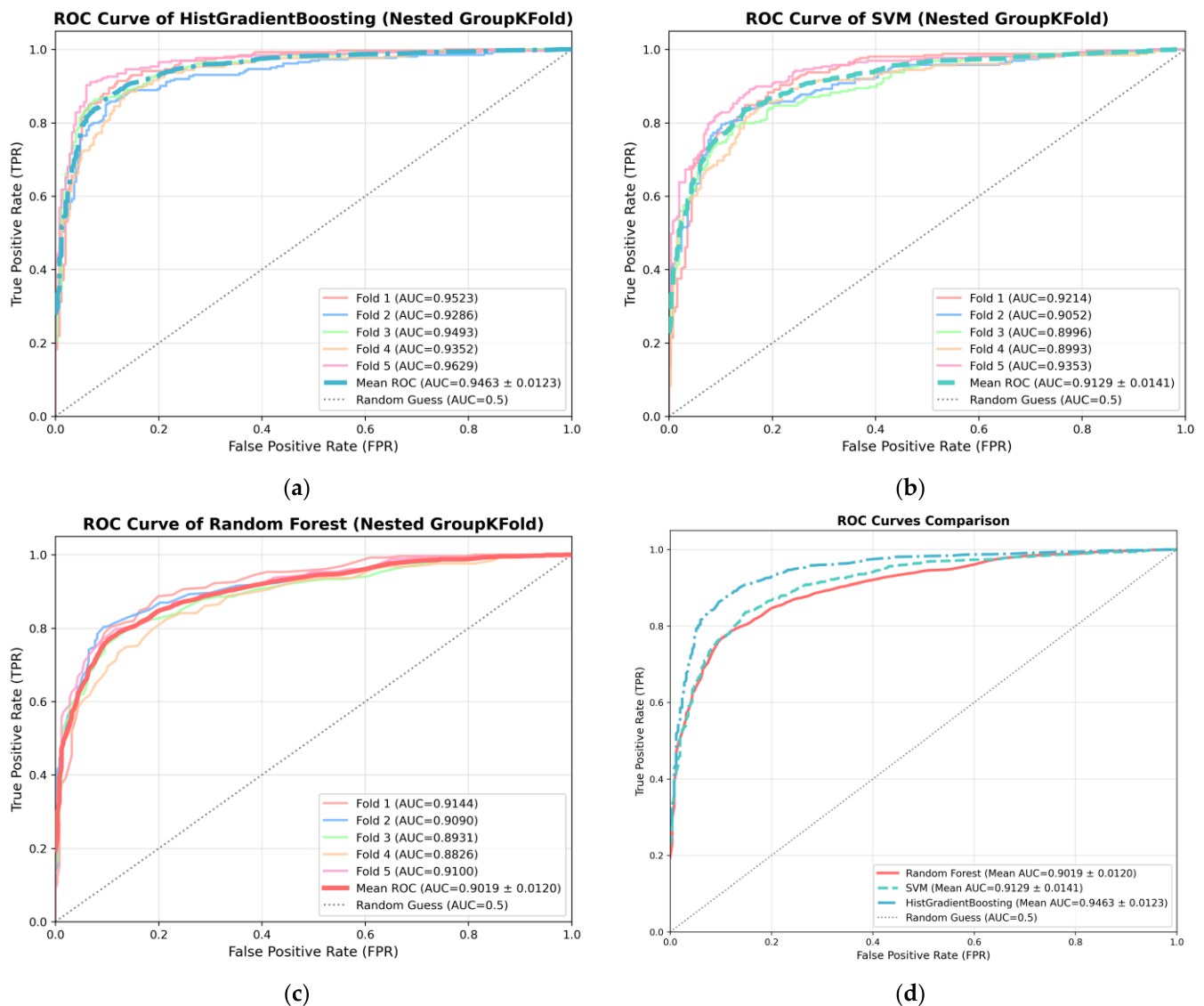


Figure 8. Receiver operating characteristic (ROC) curves for three classifiers evaluated on the *H. pylori* protein-protein interaction dataset using nested 5-fold group cross-validation. (a) ROC curves of the HistGradientBoosting classifier across five folds; (b) ROC curves of the SVM classifier across five folds; (c) ROC curves of the Random Forest classifier across five folds; (d) Combined comparison of the mean ROC curves of the three classifiers, including the corresponding mean AUC values and their standard deviations.

3.2.2. Comparison with Different Classifiers

In this section, we experimentally compared the predictive performance of two widely used classifiers: Support Vector Machine (SVM) and Random Forest (RF). The experimental results of these classifiers combined with the 3Dloop descriptor are summarized in Table 4. As shown, on the *yeast* dataset, the average accuracies of the SVM and RF classifiers are 90.72% and 93.43%, respectively. On the *H. pylori* dataset, the average accuracies of the SVM and RF are 83.59% and 82.68%, respectively. To further assess the predictive capability of the proposed 3Dloop method, we also calculated additional metrics, including sensitivity, precision, MCC, specificity, and AUC. It is evident that the accuracy of the SVM is lower than that of RF and HistGB.

Table 4. Performance comparison of different classifiers combined with the 3Dloop descriptor on two datasets.

Dataset	Model	Accu. (%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUC
Yeast	RF	93.43 ± 0.49	92.44 ± 0.44	90.79 ± 1.02	94.61 ± 0.34	86.49 ± 1.01	97.62 ± 0.44
	SVM	90.72 ± 0.55	92.14 ± 0.74	85.06 ± 1.08	94.79 ± 0.51	80.90 ± 1.13	96.52 ± 0.51
	HistGB	95.61 ± 0.40	95.93 ± 0.78	91.03 ± 1.26	97.19 ± 0.59	88.90 ± 0.81	98.38 ± 0.27
<i>H. pylori</i>	RF	82.68 ± 0.87	84.03 ± 1.47	81.03 ± 1.04	84.33 ± 1.56	65.44 ± 2.96	90.18 ± 1.20
	SVM	83.59 ± 1.57	81.03 ± 3.04	84.57 ± 2.40	82.57 ± 2.38	67.21 ± 3.16	91.22 ± 1.41
	HistGB	89.93 ± 0.13	87.45 ± 0.91	89.50 ± 0.32	86.98 ± 0.82	86.52 ± 0.65	94.59 ± 0.79

Values are mean ± SD from 5-fold nested cross-validation. Abbreviations: *Accu.*, Accuracy; *Prec.*, Precision; *Sen.*, Sensitivity; *Spec.*, Specificity; *MCC*, Matthews correlation coefficient; *AUC*, Area Under the ROC Curve. $p < 0.05$ vs. the best baseline model on the same dataset.

For a more intuitive comparison of classifier performance, we plotted the ROC curves. On the *yeast* PPI dataset, Figures 7b and 7c display the ROC curves of the SVM and RF classifiers combined with the 3Dloop descriptor, respectively, while Figure 7d compares the ROC curves of the three classifiers: SVM, RF, and HistGB. On the *H. pylori* PPI dataset, Figures 8b and 8c present the ROC curves of the SVM and RF, respectively, and Figure 8d compares the ROC curves of all three classifiers. The AUC values of these ROC curves, shown in Table 5, further confirm the superior performance of the HistGB classifier in PPI prediction tasks.

Table 5. Performance Comparison of Different PPI Prediction Methods on *Yeast* (Y) and *H. pylori* (H) dataset.

Model	Method	Accu. (%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUC
(Y) Liu et al. [49]	Boosting	79.52 ± 0.05	81.69 ± 0.09	80.37 ± 0.08	N/A	70.64 ± 0.31	N/A
(Y) Nanni et al. [50]	EnsembleELM	87.50 ± 0.66	88.95 ± 0.95	86.15 ± 0.68	N/A	78.13 ± 1.15	N/A
(Y) An et al. [23]	LCPSSMMF	93.14 ± 0.60	93.90 ± 0.34	92.50 ± 1.00	N/A	87.41 ± 1.06	N/A
(Y) Li et al. [22]	Mntflade	95.03 ± 0.25	99.14 ± 0.26	90.84 ± 0.47	94.33 ± 1.56	90.52 ± 0.45	94.27 ± 0.63
(Y) Li et al. [42]	Loop	87.37 ± 1.46	88.43 ± 2.36	78.63 ± 1.62	95.57 ± 1.38	78.34 ± 1.95	95.22 ± 1.41
(Y) proposed method	3Dloop + loop	95.61 ± 0.40	95.93 ± 0.78	91.03 ± 1.26	97.19 ± 0.59	88.90 ± 0.81	98.38 ± 0.27
(H) An et al. [24]	BiGP-RVM	86.57 ± 0.01	83.55 ± 0.01	81.88 ± 0.01	N/A	82.91 ± 0.02	N/A
(H) Zhou et al. [51]	SVM + LD	88.56 ± 0.33	89.50 ± 0.60	87.37 ± 0.22	N/A	77.15 ± 0.68	N/A
(H) You et al. [52]	PCA + EELM	87.00 ± 0.29	87.59 ± 0.32	86.15 ± 0.43	N/A	77.36 ± 0.44	N/A
(H) Li et al. [22]	Mntflade	85.35 ± 0.64	79.27 ± 0.81	95.72 ± 0.92	93.27 ± 1.46	74.41 ± 1.14	93.67 ± 0.38
(H) Li et al. [42]	Loop	85.29 ± 1.27	85.40 ± 3.30	75.78 ± 2.52	86.35 ± 0.88	68.97 ± 2.94	91.67 ± 0.92
(H) proposed method	3Dloop + loop	89.93 ± 0.13	87.45 ± 0.91	89.50 ± 0.32	86.98 ± 0.82	86.52 ± 0.65	94.59 ± 0.79

Values are reported as mean ± standard deviation (SD) over 5-fold nested cross-validation. Abbreviations: *Accu.*, Accuracy; *Prec.*, Precision; *Sen.*, Sensitivity; *Spec.*, Specificity; *MCC*, Matthews correlation coefficient; *AUC*, Area Under the Receiver Operating Characteristic Curve; N/A, not available. $p < 0.05$ compared with the best baseline method in the same dataset.

The paired *t*-test results indicate that HistGB significantly outperforms both RF and the SVM on the *yeast* and *H. pylori* dataset (adjusted $p < 0.01$).

3.3. Comparison with State-of-the-Art Methods

Performance comparison between 3Dloop and five existing PPI prediction methods (based on PSSM) on the *Yeast* and *H. pylori* dataset is shown in Table 5.

3.4. Performance on Independent Datasets

To further validate the generalization ability and assess its superiority over state-of-the-art methods, the proposed 3Dloop-FPSSM model was tested on two independent datasets (*H. sapiens* and *M. musculus*) without any parameter fine-tuning (Since the independent

human and mouse datasets include only positive samples without negatives, the evaluation here is limited to positive sample prediction/ranking rather than full binary classification). The optimal parameters determined from the *yeast* dataset were directly applied to ensure an unbiased evaluation. As shown in Table 6, our method achieved an accuracy of 94.36% on the human dataset and 92.12% on the mouse dataset. Additionally, we compared our model with several representative existing methods under the same test protocol. The comparison results demonstrate that the proposed framework consistently outperforms other state-of-the-art approaches on independent cross-species datasets, further confirming its robustness and practical applicability.

Table 6. Cross-species independent validation results on positive sample prediction.

Model	Positive Prediction Rate on <i>H. sapiens</i> Dataset (%)	Positive Prediction Rate on <i>M. musculus</i> Dataset (%)
Liu et al. [49]	83.36	84.54
An et al. [23]	92.69	89.75
Zhou et al. [51]	87.37	84.74
Li et al. [22]	89.87	90.16
Proposemethod	94.36	92.12

3.5. Analysis of Ablation Studies

To validate the contribution of LOOP and 3Dloop features, we conducted ablation experiments using the same nested cross-validation procedure. The quantitative results are presented in Table 7.

Table 7. Performance comparison of different feature sets on the Yeast and *H. pylori* datasets.

Dataset	Model	Accu. (%)	Prec. (%)	Sen. (%)	Spec. (%)	MCC (%)	AUC
Yeast	loop	92.90 ± 0.78	93.40 ± 1.03	89.32 ± 0.98	95.52 ± 0.56	85.44 ± 0.76	96.62 ± 0.77
	3Dloop	92.34 ± 0.65	93.26 ± 0.94	87.99 ± 1.68	95.45 ± 0.81	84.21 ± 0.78	97.51 ± 0.58
	Loop + 3Dloop	95.61 ± 0.40	95.93 ± 0.78	91.03 ± 1.26	97.19 ± 0.59	88.90 ± 0.81	98.38 ± 0.27
<i>H. pylori</i>	Loop	87.48 ± 0.67	86.93 ± 0.47	88.30 ± 0.94	86.33 ± 1.86	74.44 ± 2.96	92.18 ± 0.80
	3Dloop	86.83 ± 0.97	86.37 ± 1.04	87.75 ± 1.40	85.80 ± 1.38	73.60 ± 1.16	91.22 ± 1.41
	Loop + 3Dloop	89.93 ± 0.13	87.45 ± 0.91	89.50 ± 0.32	86.98 ± 0.82	86.52 ± 0.65	94.59 ± 0.79

Values are reported as mean ± SD over 5-fold nested cross-validation. Abbreviations: *Accu.*, Accuracy; *Prec.*, Precision; *Sen.*, Sensitivity; *Spec.*, Specificity; *MCC*, Matthews correlation coefficient; *AUC*, Area Under the ROC Curve. For *H. pylori*: $p < 0.05$ vs. both single feature sets (LOOP-only and 3Dloop-only) (paired *t*-test, two-tailed). For *yeast*: $p < 0.01$ vs. both single feature sets (LOOP-only and 3Dloop-only) (paired *t*-test, two-tailed).

The model using only LOOP or only 3Dloop features achieves moderate performance. In contrast, the full model (LOOP + 3Dloop) achieves the highest accuracy (95.61%), sensitivity, and AUC (0.9838), confirming that both feature modules are essential and complementary.

Statistical analysis through paired *t*-tests reveals no significant difference between the 3Dloop-only feature and the original LOOP feature ($p > 0.05$) for both datasets. However, the combined LOOP + 3Dloop feature significantly outperforms both the 3Dloop-only and LOOP-only features ($p < 0.05$ for *H. pylori*, $p < 0.01$ for *yeast*), highlighting the strong complementary effect of the two components.

4. Discussion

4.1. Interpretation of Experimental Results

4.1.1. Efficacy of 3Dloop's Feature Extraction

The 3Dloop descriptor extracts both continuous local and discrete geometric features from the 3D FPSSM structure. On the *yeast* dataset, the combined LOOP + 3Dloop model

achieved $95.61 \pm 0.40\%$ accuracy, higher than LOOP-only ($92.90 \pm 0.78\%$) and 3Dloop-only ($92.34 \pm 0.65\%$). On the *H. pylori* dataset, the fused feature reached $89.93 \pm 0.13\%$ accuracy. These results confirm that the dual-feature fusion improves feature representation for PPI prediction by modeling discontinuous amino acid correlations via FPSSM.

4.1.2. Advantage of the HistGB Classifier

HistGB exhibited the best overall performance among the three classifiers tested on both datasets. On *yeast*, HistGB attained $95.61 \pm 0.40\%$ accuracy, exceeding RF ($93.43 \pm 0.49\%$) and SVM ($90.72 \pm 0.55\%$). On *H. pylori*, HistGB achieved $89.93 \pm 0.13\%$ accuracy, outperforming RF and the SVM. Paired *t*-tests verified statistically significant superiority ($p < 0.01$) [47]. These results show that HistGB better handles 512-dimensional features and reduces overfitting for this PPI prediction framework [48].

4.1.3. Cross-Species Generalization

The model was evaluated on independent human and mouse datasets using parameters optimized on *yeast*. The method obtained positive prediction rates of 94.36% and 92.12%, respectively, and outperformed the compared methods. These results show that the extracted evolutionary and structural features are transferable across species within the experimental setting tested.

4.2. Comparison with State-of-the-Art Research

Statistical tests confirm that the advantages of HistGB and the fused feature strategy are consistent across experimental settings. On *yeast* and *H. pylori* datasets, the 3Dloop-FPSSM framework performs competitively against existing PSSM-based methods. The improvements arise from three components: FPSSMs for modeling discontinuous residue relationships, 3Dloop for extracting complementary local and geometric features, and HistGB for robust high-dimensional feature integration. The ablation study validates the complementary roles of LOOP and 3Dloop features.

4.3. Limitations

Several limitations of the proposed method should be acknowledged:

1. Heuristic nature of the FPSSM folding strategy.

The FPSSM folding process uses fixed rules based on 13 tripeptide fragments from existing literature, rather than experimental or predicted protein tertiary structures. As such, it provides only a heuristic approximation of residue spatial relationships and may not fully represent authentic protein conformations.

2. Incomplete cross-species validation.

Independent human and mouse datasets contain only positive samples, so standard binary classification metrics cannot be applied. The reported positive prediction rates therefore do not fully reflect the model's ability to distinguish real interactions in cross-species settings.

3. Restricted application scope.

The framework is validated only for protein–protein interaction prediction; its performance on other protein-related tasks remains untested.

4. Low adaptability of fixed folding rules.

Although fixed folding rules help avoid data leakage and ensure reproducibility, they cannot flexibly adapt to diverse protein structures. Further comparisons between fixed and adaptive folding strategies are needed.

5. Conclusions

In this study, we proposed the 3Dloop-FPSSM framework, which combines Folded Position-Specific Scoring Matrix (FPSSM) representation and 3D Local Optimal Orientation Pattern (3Dloop) feature extraction, followed by Histogram Gradient Boosting (HistGB) classification for protein–protein interaction (PPI) prediction. Experimental results demonstrate that 3Dloop-FPSSM achieves accuracies of 95.61% on the *yeast* dataset and 89.93% on the *Helicobacter pylori* dataset, outperforming models using only LOOP or 3Dloop features.

The novelty of this method lies in capturing the spatial correlation between discontinuous amino acids through the FPSSM and extracting both continuous and discrete geometric features via 3Dloop. This enables a more comprehensive representation of protein interactions, providing a substantial improvement in sequence-based PPI prediction. Thus, 3Dloop-FPSSM not only enhances prediction accuracy but also introduces a meaningful advancement in the PPI prediction methodology, particularly beneficial for high-throughput screening of proteins from species with limited structural data.

This framework is especially useful in scenarios where 3D structural annotations are unavailable, enabling efficient PPI prediction based on sequence information alone.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/a19080642/s1>, Code S1: PSSM Matrix Construction Script for Protein Sequences; Code S2: HistGB Classifier Training and PPI Prediction Pipeline.

Author Contributions: Conceptualization, F.B.; methodology, F.B.; software, D.L. and R.M.; validation, F.B., R.M. and G.W.; formal analysis, D.L.; investigation, F.B.; resources, D.L.; data curation, G.W.; writing—original draft preparation, F.B.; writing—review and editing, F.B.; visualization, P.Z.; supervision, P.Z. All authors have read and agreed to the published version of the manuscript.

Funding: Research on the Application of Machine Learning Methods in Environmental Governance and Rural Economic Revitalization in Plateau Areas (KUST-xk2026010). Research on kernel density estimation of interval-observed data with right-censored observations.

Data Availability Statement: Data will be made available on request. The original contributions presented in this study are included in the article/Supplementary Materials. Further inquiries can be directed to the corresponding author.

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

References

1. Morsy, M.; Gouthu, S.; Orchard, S.; Thorneycroft, D.; Harper, J.F.; Mittler, R.; Cushman, J.C. Charting plant interactomes: Possibilities and challenges. *Trends Plant Sci.* **2008**, *13*, 183–191. [\[CrossRef\]](#)
2. Yuan, J.S.; Galbraith, D.W.; Dai, S.Y.; Griffin, P.; Stewart, C.N., Jr. Plant systems biology comes of age. *Trends Plant Sci.* **2008**, *13*, 165–171. [\[CrossRef\]](#)
3. Fukao, Y. Protein-protein interactions in plants. *Plant Cell Physiol.* **2012**, *53*, 617–625. [\[CrossRef\]](#)
4. Sheth, B.P.; Thaker, V.S. Plant systems biology: Insights, advances and challenges. *Planta* **2014**, *240*, 33–54. [\[CrossRef\]](#)
5. Braun, P.; Gingras, A.C. History of protein-protein interactions: From egg-white to complex networks. *Proteomics* **2012**, *12*, 1478–1498. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Causier, B.; Davies, B. Analysing protein-protein interactions with the yeast two-hybrid system. *Plant Mol. Biol.* **2002**, *50*, 855–870. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Fang, Y.; Macool, D.J.; Xue, Z.; Heppard, E.P.; Hailey, C.F.; Tingey, S.V.; Miao, G.H. Development of a high-throughput yeast two-hybrid screening system to study protein-protein interactions in plants. *Mol. Genet. Genom.* **2002**, *267*, 142–153. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Hayashi, T.; Matsuzaki, Y.; Yanagisawa, K.; Ohue, M.; Akiyama, Y. MEGADOCK-Web: An integrated database of high-throughput structure-based protein-protein interaction predictions. *BMC Bioinform.* **2018**, *19*, 62. [\[CrossRef\]](#)
9. Zahiri, J.; Mohammad-Noori, M.; Ebrahimpour, R.; Saadat, S.; Bozorgmehr, J.H.; Goldberg, T.; Masoudi-Nejad, A. LocFuse: Human protein-protein interaction prediction via classifier fusion using protein localization information. *Genomics* **2014**, *104*, 496–503. [\[CrossRef\]](#)

10. Huang, D.S.; Du, J.X. A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. *IEEE Trans. Neural Netw.* **2008**, *19*, 2099–2115. [\[CrossRef\]](#)
11. Huang, D.S.; Zheng, C.H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. *Bioinformatics* **2006**, *22*, 1855–1862. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Wang, Y.B.; You, Z.H.; Li, L.P.; Huang, D.S.; Zhou, F.F.; Yang, S. Improving Prediction of Self-interacting Proteins Using Stacked Sparse Auto-Encoder with PSSM profiles. *Int. J. Biol. Sci.* **2018**, *14*, 983–991. [\[CrossRef\]](#)
13. Deng, S.P.; Huang, D.S. SFAPS: An R package for structure/function analysis of protein sequences based on informational spectrum method. *Methods* **2014**, *69*, 207–212. [\[CrossRef\]](#)
14. Huang, D.S.; Zhang, L.; Han, K.; Deng, S.; Yang, K.; Zhang, H. Prediction of protein-protein interactions based on protein-protein correlation using least squares regression. *Curr. Protein Pept. Sci.* **2014**, *15*, 553–560. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Yang, L.; Xia, J.F.; Gui, J. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept. Lett.* **2010**, *17*, 1085–1090. [\[CrossRef\]](#)
16. Wang, Y.; You, Z.; Li, X.; Chen, X.; Jiang, T.; Zhang, J. PCVMZM: Using the Probabilistic Classification Vector Machines Model Combined with a Zernike Moments Descriptor to Predict Protein-Protein Interactions from Protein Sequences. *Int. J. Mol. Sci.* **2017**, *18*, 1029. [\[CrossRef\]](#)
17. Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 4337–4341. [\[CrossRef\]](#)
18. Davies, M.N.; Secker, A.; Freitas, A.A.; Clark, E.; Timmis, J.; Flower, D.R. Optimizing amino acid groupings for GPCR classification. *Bioinformatics* **2008**, *24*, 1980–1986. [\[CrossRef\]](#)
19. Pan, J.; Li, L.P.; You, Z.H.; Yu, C.Q.; Ren, Z.H.; Guan, Y.J. Prediction of Protein-Protein Interactions in Arabidopsis, Maize, and Rice by Combining Deep Neural Network With Discrete Hilbert Transform. *Front. Genet.* **2021**, *12*, 745228. [\[CrossRef\]](#)
20. An, J.Y.; Zhou, Y.; Yan, Z.J.; Zhao, Y.J. Predicting Self-Interacting Proteins Using a Recurrent Neural Network and Protein Evolutionary Information. *Evol. Bioinform. Online* **2020**, *16*, 1176934320924674. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Wang, Y.B.; You, Z.H.; Li, L.P.; Huang, Y.A.; Yi, H.C. Detection of Interactions between Proteins by Using Legendre Moments Descriptor to Extract Discriminatory Information Embedded in PSSM. *Molecules* **2017**, *22*, 1366. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Li, Y.; Wang, Z.; You, Z.H.; Li, L.P.; Hu, X. Predicting Protein-Protein Interactions via Random Ferns with Evolutionary Matrix Representation. *Comput. Math. Methods Med.* **2022**, *2022*, 7191684. [\[CrossRef\]](#)
23. An, J.Y.; Zhou, Y.; Zhao, Y.J.; Yan, Z.J. An Efficient Feature Extraction Technique Based on Local Coding PSSM and Multifeatures Fusion for Predicting Protein-Protein Interactions. *Evol. Bioinform. Online* **2019**, *15*, 1176934319879920. [\[CrossRef\]](#)
24. An, J.Y.; Meng, F.R.; You, Z.H.; Chen, X.; Yan, G.Y.; Hu, J.P. Improving protein-protein interactions prediction accuracy using protein evolutionary information and relevance vector machine model. *Protein Sci.* **2016**, *25*, 1825–1833. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Pan, J.; Li, L.P.; Yu, C.Q.; You, Z.H.; Guan, Y.J.; Ren, Z.H. Sequence-Based Prediction of Plant Protein-Protein Interactions by Combining Discrete Sine Transformation With Rotation Forest. *Evol. Bioinform. Online* **2021**, *17*, 11769343211050067. [\[CrossRef\]](#)
26. Zhan, X.; Xiao, M.; You, Z.; Yan, C.; Guo, J.; Wang, L.; Sun, Y.; Shang, B. Predicting Protein-Protein Interactions Based on Ensemble Learning-Based Model from Protein Sequence. *Biology* **2022**, *11*, 995. [\[CrossRef\]](#) [\[PubMed\]](#)
27. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science* **1973**, *181*, 223–230. [\[CrossRef\]](#)
28. Xenarios, I.; Salwinski, L.; Duan, X.J.; Higney, P.; Kim, S.M.; Eisenberg, D. DIP, the Database of Interacting Proteins: A research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **2002**, *30*, 303–305. [\[CrossRef\]](#)
29. Martin, S.; Roe, D.; Faulon, J.L. Predicting protein-protein interactions using signature products. *Bioinformatics* **2005**, *21*, 218–226. [\[CrossRef\]](#)
30. Gribskov, M.; McLachlan, A.D.; Eisenberg, D. Profile analysis: Detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **1987**, *84*, 4355–4358. [\[CrossRef\]](#)
31. Huang, C.; Yuan, J. Using radial basis function on the general form of Chou's pseudo amino acid composition and PSSM to predict subcellular locations of proteins with both single and multiple sites. *Biosystems* **2013**, *113*, 50–57. [\[CrossRef\]](#)
32. Verma, R.; Varshney, G.C.; Raghava, G.P. Prediction of mitochondrial proteins of malaria parasite using split amino acid composition and PSSM profile. *Amino Acids* **2010**, *39*, 101–110. [\[CrossRef\]](#)
33. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [\[CrossRef\]](#)
34. Altschul, S.F.; Koonin, E.V. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* **1998**, *23*, 444–447. [\[CrossRef\]](#) [\[PubMed\]](#)
35. Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* **1997**, *25*, 31–36. [\[CrossRef\]](#)
36. Pauling, L.; Corey, R.B.; Branson, H.R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* **1951**, *37*, 205–211. [\[CrossRef\]](#)

37. Ramakrishnan, C.; Ramachandran, G.N. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophys. J.* **1965**, *5*, 909–933. [[CrossRef](#)]
38. Hassan, M.; Coutsiyas, E.A. Protein secondary structure motifs: A kinematic construction. *J. Comput. Chem.* **2021**, *42*, 271–292. [[CrossRef](#)]
39. Ding, Y.S.; Zhang, T.L.; Chou, K.C. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein Pept. Lett.* **2007**, *14*, 811–815. [[CrossRef](#)] [[PubMed](#)]
40. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
41. Swanson, C.J.; Sivaramakrishnan, S. Harnessing the unique structural properties of isolated α -helices. *J. Biol. Chem.* **2014**, *289*, 25460–25467. [[CrossRef](#)]
42. Chakraborti, T.; McCane, B.; Mills, S.; Pal, U. LOOP Descriptor: Local Optimal-Oriented Pattern. *IEEE Signal Process. Lett.* **2018**, *25*, 635–639. [[CrossRef](#)]
43. Ojala, T.; Pietikainen, M.; Harwood, D. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 9–13 October 1994.
44. Guo, Z.; Zhang, L.; Zhang, D. A Completed Modeling of Local Binary Pattern Operator for Texture Classification. *IEEE Trans. Image Process.* **2010**, *19*, 1657–1663. [[CrossRef](#)]
45. Jabid, T.; Kabir, M.H.; Chae, O. Gender Classification Using Local Directional Pattern (LDP). In Proceedings of the 2010 20th International Conference on Pattern Recognition, Washington, DC, USA, 23–26 August 2010.
46. Sandbach, G.; Zafeiriou, S.; Pantic, M. Local normal binary patterns for 3D facial action unit detection. In Proceedings of the 2012 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012.
47. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [[CrossRef](#)]
48. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
49. Liu, B.; Yi, J.; Aishwarya, S.V.; Lan, Y.; Ma, Y.; Huang, T.H.; Leone, G.; Jin, V.X. QChIPat: A quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. *BMC Genom.* **2013**, *14*, S3. [[CrossRef](#)]
50. Nanni, L. Fusion of classifiers for predicting protein-protein interactions. *Neurocomputing* **2005**, *68*, 289–296. [[CrossRef](#)]
51. Zhou, Y.Z.; Gao, Y.; Zheng, Y.Y. Prediction of protein–protein interactions using local description of amino acid sequence. In *Advances in Computer Science and Education Applications*; Springer: Berlin/Heidelberg, Germany, 2011.
52. You, Z.; Lei, Y.; Zhu, L.; Xia, J.; Wang, B. Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis. *BMC Bioinform.* **2013**, *14*, 69–75. [[CrossRef](#)]

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.