genes

MDPI

Article

LGFC-CNN: Prediction of lncRNA-Protein Interactions by Using Multiple Types of Features through Deep Learning

Lan Huang [1], Shaoqing Jiao [1,2], Sen Yang [3], Shuangquan Zhang [1], Xiaopeng Zhu [1,2], Rui Guo [1] and Yan Wang [1,4,*]

1 Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China; huanglan@jlu.edu.cn (L.H.); jiaosq19@mails.jlu.edu.cn (S.J.); shuangquan18@mails.jlu.edu.cn (S.Z.); zhuxiaopen11@163.com (X.Z.); sdsbshe@163.com (R.G.)
2 College of Software, Jilin University, Changchun 130012, China
3 School of Computer Science and Artificial Intelligence & Aliyun School of Big Data, Changzhou University, Changzhou 213164, China; ystop2020@gmail.com
4 College of Artificial Intelligence, Jilin University, Changchun 130012, China
* Correspondence: wy6868@jlu.edu.cn

Abstract: Long noncoding RNA (lncRNA) plays a crucial role in many critical biological processes and participates in complex human diseases through interaction with proteins. Considering that identifying lncRNA–protein interactions through experimental methods is expensive and time-consuming, we propose a novel method based on deep learning that combines raw sequence composition features, hand-designed features and structure features, called LGFC-CNN, to predict lncRNA–protein interactions. The two sequence preprocessing methods and CNN modules (GloCNN and LocCNN) are utilized to extract the raw sequence global and local features. Meanwhile, we select hand-designed features by comparing the predictive effect of different lncRNA and protein features combinations. Furthermore, we obtain the structure features and unifying the dimensions through Fourier transform. In the end, the four types of features are integrated to comprehensively predict the lncRNA–protein interactions. Compared with other state-of-the-art methods on three lncRNA–protein interaction datasets, LGFC-CNN achieves the best performance with an accuracy of 94.14%, on RPI21850; an accuracy of 92.94%, on RPI7317; and an accuracy of 98.19% on RPI1847. The results show that our LGFC-CNN can effectively predict the lncRNA–protein interactions by combining raw sequence composition features, hand-designed features and structure features.

Keywords: lncRNA-protein interactions; convolutional neural network; two sequence preprocessing methods; raw sequence features; hand-designed features; structure features

check for updates

Citation: Huang, L.; Jiao, S.; Yang, S.; Zhang, S.; Zhu, X.; Guo, R.; Wang, Y. LGFC-CNN: Prediction of lncRNA-Protein Interactions by Using Multiple Types of Features through Deep Learning. *Genes* **2021**, *12*, 1689. https://doi.org/10.3390/genes12111689

Academic Editor: Piero Fariselli

Received: 25 July 2021
Accepted: 22 October 2021
Published: 24 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Long noncoding RNA (lncRNA) is a type of noncoding RNA with at least 200 nucleotides that plays vital roles in many critical biological processes [1], such as cell differentiation, gene expression, and the display of developmental and tissue-specific expression patterns [2–4]. Some lncRNA regulates a wide range of biological processes through interactions with miRNAs. They compete with mRNA for binding to the same miRNA to create a competitive endogenous RNA (ceRNA) regulatory network through which their modular structure permits their interaction with miRNAs [5,6]. Some lncRNA participates in many complex human diseases by interacting with proteins [7]. For example, silencing lncRNA-5657 inhibits the pneumonia lung inflammatory response via suppressing the expression of spinster homology protein2, thereby reducing sepsis-induced lung injury [8]. FAM83H-AS1 contributes to radioresistance and cell metastasis in ovarian cancer through the stabilizing HuR protein [9]. LncRNA NEAT1 promotes MPTP-induced autophagy in Parkinson's disease through the stabilization of the PINK1 protein [10]. Therefore, predicting potential

lncRNA–protein interactions is a crucial step in understanding the function of lncRNA and creating the conditions for solving complex human diseases. With the development of experimental technology, computational methods have become crucial as a silver-bullet solution for the large-scale capture of lncRNA–protein interactions, which helps to prioritize lncRNA–protein interaction candidates and conduct further experimental verification.

Existing computational methods can be categorized into network-based methods and machine learning-based methods. Network-based methods construct a lncRNA/protein similarity matrix and then use network algorithms to calculate correlation scores to make predictions. For example, Zhang et al. proposed a model called LPLNP, in 2017, which calculated the linear neighborhood similarity between lncRNAs and proteins and the regularized linear neighborhood similarity and predicted the observed lncRNA–protein interactions by a label-propagation process [11]. Zhao et al. introduced a method named LPI-BNPRA in 2018, which used the known lncRNA–protein interactions matrix, lncRNA similarity matrix, and protein similarity matrix to predict lncRNA–protein relationships [12]. Zhu et al. presented a model named ACCBN, in 2019, the model first to use an ant-colony algorithm for data clustering and then constructed a lncRNA–protein bipartite network inference (LPBNI) to predict lncRNA–protein interactions [13]. However, network-based methods require that each node in the network has at least two linkages; the lncRNA–protein interaction network is composed of a few isolated subnetworks, and the imbalance of the degree distribution of each node in the network will also affect its prediction performance [14].

The machine learning-based methods extract manual features from lncRNA and protein sequences to represent lncRNA–protein pairs and then input them into machine learning classifiers to predict lncRNA–protein interaction pairs. For example, Ge et al. proposed a model named RPISeq, in 2015, that input the 4-mer frequency characteristics of RNA sequences and the 3-mer frequency characteristics of proteins into a random forest classifier and support-vector-machine classifiers to identify RNA-protein interactions [15]. Pan et al. developed a method called IPMiner, in 2016, that input raw sequence-composition features, the advanced features extracted by a cascaded autoencoder and the features extracted by fine-tuned cascaded noise reduction auto-encoding into a random forest classifier, then used a cascading ensemble to integrate the output of the above three classifiers to predict lncRNA–protein interactions [16]. In 2019, Fan et al. proposed LPI-BLS; they first combined lncRNA's and protein's features, input these features into five separate extensive learning systems, and finally integrated separate BLS classifiers through a stacking integration strategy to obtain their prediction results [17]. Liu et al. presented a model named LPI-NRLMF in 2017, which mapped the lncRNA–protein interaction matrix to the lncRNA similarity matrix and the protein similarity matrix to predict the possibility of lncRNA–protein interactions [18]. However, the machine learning-based methods have limitations that rely on the quality of hand-designed features [19,20].

In this paper, we propose a new deep-learning model (LGFC-CNN) that combines raw sequence-composition features, hand-designed features, and structure features to comprehensively predict lncRNA–protein interactions. First, we improve the sequences' preprocessing, originally used to predict RNA-protein binding sites, and apply it to transform the sequences into fixed-length sequences [21]. After that, the lncRNA and protein sequences are encoded by using one-hot encoding [22,23] and fed into GloCNN and LocCNN modules to extract the raw sequence's global and local features. Meanwhile, a random forest (RF) classifier [24] is employed to compare various lncRNA's and protein's hand-designed combinations of features, and, of such features, those with the three most-superior predictive effects are fed into an FC module to gain useful information. Furthermore, the secondary structures, hydrogen bonding, and van der Waals interactions of the lncRNA and protein are encoded and fed into an SS module, after unifying their feature dimensions through a Fourier transform. Finally, the four network modules are integrated, to improve predictive performance by analyzing multiple types of features. In addition,

comparing LGFC-CNN with several existing methods, the results show that LGFC-CNN is a competitive method for effectively predicting lncRNA–protein interactions.

2. Materials and Methods

An illustration of LGFC-CNN for predicting lncRNA–protein interactions is shown in Figure 1.

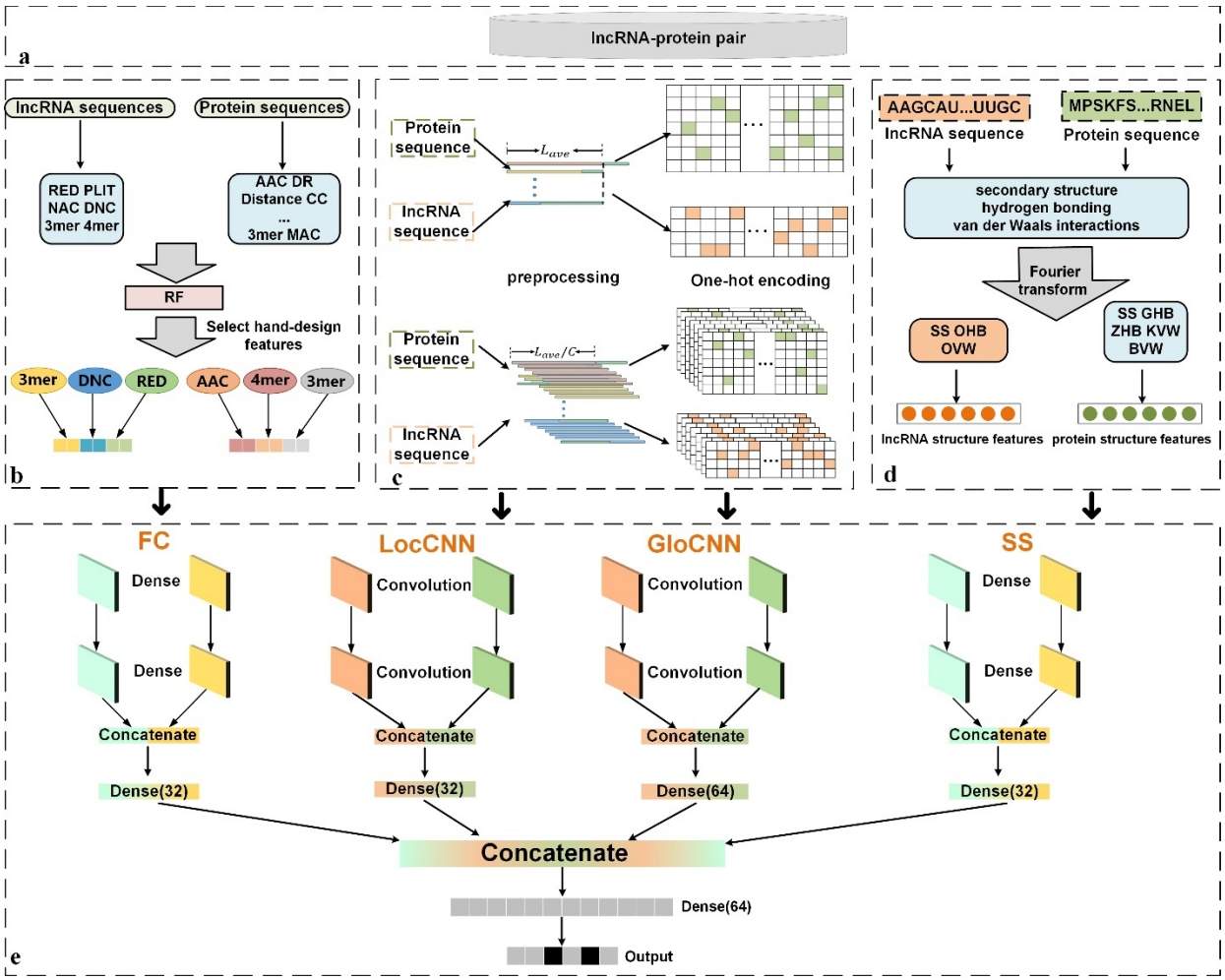

Figure 1. The flowchart of LPI-CNNCP. (**a**) Build lncRNA-protein interactions datasets and obtain lncRNA and protein sequences; (**b**) Feed lncRNA and protein hand-designed feature combinations into RF classifier to select the hand-designed features with superior predictive effect; (**c**) lncRNA and protein sequences are preprocessed by two methods and encoded by using one-hot encoding; (**d**) lncRNA and protein secondary structure, hydrogen bonding propensities, and van der Waals interactions are obtained and unifying the dimensions through Fourier transform. (e) Feed the global and local encoded sequences, hand-designed features and structure features into CNN model to predict the lncRNA-protein interactions.

2.1. Construction of Datasets

To evaluate the performance of LGFC-CNN, we test it on the lncRNA–protein interaction datasets of $D_{RPI21850} = D^+ \cup D^-$ (named RPI21850) constructed from the NPInterv4.0 database [25,26]. To filter lncRNAs and their interacting proteins, the ncRNA sequences whose length less than 200nt and the lncRNA–protein interactions not from Homo sapiens were excluded. Then, we constructed a positive dataset D^+, which contained 21850 pairs of high-confidence lncRNA–protein interactions consisting of 4221 lncRNAs and 701 proteins.

Due to the lack of negative samples in the NPInter4.0 database and the assumption that obtaining the negative dataset by randomly pairing lncRNA and protein is not entirely

reasonable, we adopted the following criteria from FIRE [27] to build the high-quality negative dataset D^-:

For a lncRNA–protein interaction of protein p_1 and RNA r, r is highly possible to interact with any protein, p_2, similar to p_1. Contrarily, if protein p_2 is dissimilar to p_1, there is a low possibility that p_2 interacts r [27]. Therefore, we used the pairwise2 module in Biopython [28] to calculate the global sequence similarity score S_s between all proteins from the positive dataset. Then, we sorted the global sequence similarity scores Ss between the proteins, in ascending order.

To reduce the repeated lncRNA–protein interactions and consider that lncRNA has a certain probability of having a relationship with those proteins with higher scores, instead of selecting the lncRNA–protein pairs with the lowest interaction scores, we divided all lncRNAs into two equal parts. In the first part, the lncRNA–protein pairs were selected for which their lncRNA matched with a random p_a among the 20% of proteins, p_j, with the lowest interaction scores. In the second part, the lncRNA–protein pairs were selected for which their lncRNA matched with a random p_b of the remaining 80% of proteins, p_k. Then, the two parts were combined to build a negative dataset, D^-, that containe 21850 lncRNA–protein pairs. The flowchart of constructing reliable negative samples is shown in Figure 2.

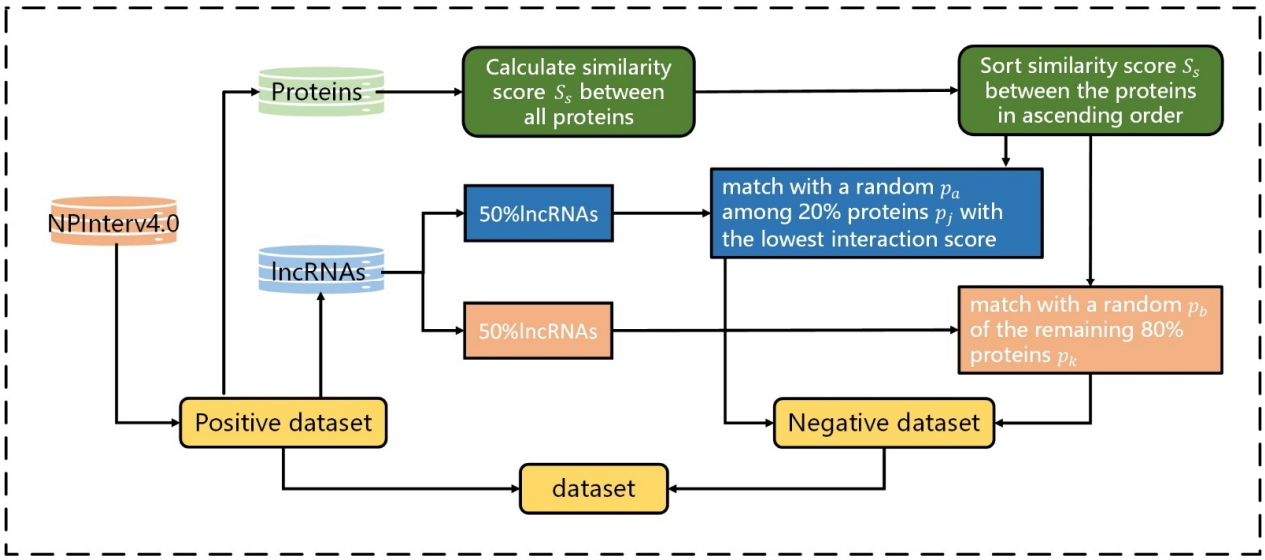

Figure 2. The flowchart of constructing reliable negative samples. The positive dataset is constructed by extracting interactions with NPInterv4.0. We calculated and sorted the similarity scores, S_s, between all proteins from the positive dataset, and divided all lncRNAs into two equal parts before applying different strategies to build the negative dataset.

To further assess the reliability and robustness of LGFC-CNN, RPI7317 and RPI1847 in LPI-BLS [16] were constructed by adopting a similar method as in assessing RPI21850, and the numbers of lncRNA–protein interacting pairs they contained were 7317 and 1847, respectively. There was no overlap between RPI7317, PRI1847, and RPI21850. The corresponding lncRNA sequences were obtained by NONECODE v6.0 [29], and the corresponding protein sequences were obtained by UniProt [30]. Table 1 lists the numeric description of the datasets.

Table 1. Numeric description of the datasets.

Dataset	lncRNAs	Proteins	Interaction Pairs	Non-Interaction Pairs
RPI21850	4221	701	21850	21850
RPI7317	1874	118	7317	7317
RPI1847	1939	60	1847	1847

2.2. Sequence Encoding

As the CNN model requires fixed-length sequence inputs, whereas different lncRNA(or protein) sequences vary significantly in their lengths, we improved the sequences preprocessing in iDeepE [21] to transform the sequences into the fixed-length sequences. Considering that some lncRNA sequences are extremely long (more than 80,000 bp), we set the average sequence length, L_{lnc} and L_{pro}, to represent the lncRNA and protein sequences' fixed lengths, respectively. Since the local structure of a sequence allows us to understand its protein–RNA binding nature, in terms of structural fragments [31]—which can supplement the lack of global structure—we performed two preprocessing procedures on the raw sequence.

For the GloCNN module, if the lncRNA sequence length was greater than L_{lnc}, the sequence was cropped to the fixed length; when lesser, it was extended to the fixed length with nucleotide N.

For the LocCNN module, a lncRNA sequence was divided into subsequences of W windows, in which each subsequence is regarded as a channel and where each window has S overlapping shifts; the size of each window was L_{lnc}/W. Here, we calculated the maximum number of channels, C, according to the sequence length. If the number of channels for one sequence was greater than C, the sequence was cropped to C.; when lesser, it was extended by channels derived from sequences with all nucleotide Ns to C.

For a given protein sequence, we adopted the same preprocessing to transform it into the fixed-length sequence with L_{pro}. After that, the lncRNA and protein sequences were encoded by using one-hot encoding [22,23]. Given a lncRNA sequence $S = (s_1, s_2, \ldots, s_n)$ with L_{lnc} nucleotides, conversion of the matrix, M, by one-hot encoding, can be expressed as [21]:

$$M_{i,j} = \begin{cases} 1/4 \; if \; s_{i-m+1} = N \; or \; i\langle m \; or \; i\rangle n - m \\ 1 \; if \; s_{i-m+1} \; in \; \{A, C, G, U\} \\ 0 \; otherwise \end{cases} \quad (1)$$

where i is the index of nucleotide, j is the index of A, C, G, U in the matrix M. For the padded nucleotide at the start and end of sequences, we assumed four nucleotides were equally distributed. Thus, [0.25, 0.25, 0.25, 0.25] was used in the padded nucleotides and N in the one-hot matrix.

For a given protein sequence $P = (p_1, p_2, \ldots, p_n)$ with L_{pro} amino acids, the sequence is composed of 20 natural amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y). When using one-hot encoding to encode the protein sequence, the encoding matrix can be vast and sparse. Thus, we compressed the 20 amino acid alphabets into seven groups, based on their dipole moments and side chains [32]: $R_1 = \{A, G, V\}$, $R_2 = \{I, L, F, P\}$, $R_3 = \{Y, M, T, S\}$, $R_4 = \{H, N, Q, W\}$, $R_5 = \{R, K\}$, $R_6 = \{D, E\}$, $R_7 = \{C\}$. The matrix R, converted by one-hot encoding, can be expressed as:

$$R_{i,j} = \begin{cases} 1/7 if \; p_{i-m+1} = N \; or \; i\langle m \; or \; i\rangle n - m \\ 1 if \; p_{i-m+1} \; in \; \{R_1, R_2, R_3, R_4, R_5, R_6, R_7\} \\ 0 \; otherwise \end{cases} \quad (2)$$

where i is the index of the amino acid, j is the index of $R_1, R_2, \ldots, R_7$ in matrix R. For the padded at the start and end of sequences, we assume 7 groups are equally distributed. Thus, $[1/7, \ldots, 1/7]$ for the padded amino acid and N in the one-hot matrix. The flowchart of lncRNA sequence encoding is shown in Figure 3.

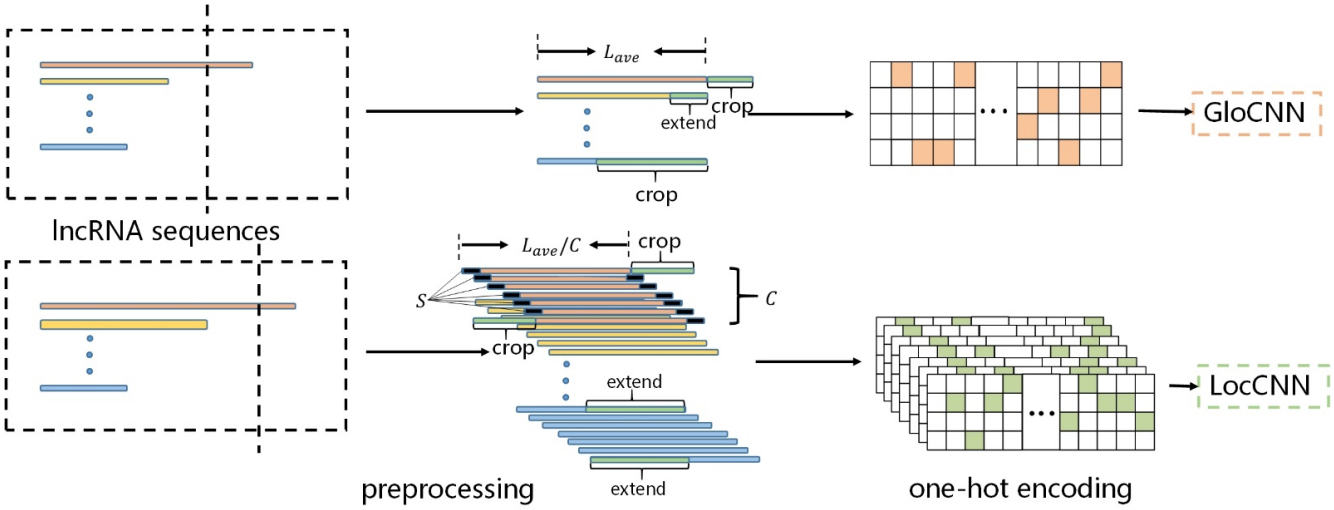

Figure 3. The flowchart of lncRNA sequence encoding. The two preprocessing methods are applied to transform the lncRNA sequences into fixed-length sequences. After that, the sequences are encoded by using one-hot encoding and fed into GloCNN and LocCNN. The protein sequences apply the same sequence encoding.

2.3. Hand-Designed Features

In this work, six hand-designed features of lncRNA are combined with ten hand-designed features of the protein. Each feature combination is ranked according to their average performance in the random forest classifier. Then the top three features with superior predictive effect were selected to represent the lncRNA and protein, respectively. For the lncRNA, the top three features were RNA-coding potential characteristics, L_{RED}, di-nucleotide composition, L_{DNC}, and lncRNA 3-mer frequency, L_{3mer}. For the protein, the top three features were amino acid composition, P_{AAC}, protein 3-mer frequency, P_{3mer}, and protein 4-mer frequency, P_{4mer}. For lncRNA–protein pairs, we concatenated the three lncRNA and protein feature vectors to form two feature vectors, $A_1 = [L_{RED}, L_{DNC}, L_{3mer}]$, $A_2 = [P_{AAC}, P_{3mer}, P_{4mer}]$. The following subsections explain the six feature encodings we used (other feature encodings are explained in Supplementary File S1).

2.3.1. lncRNA Feature RED

CPPred is a tool developed by Xiaoxue Tong et al. to predict coding potential based on the global description of an RNA sequence [33]. It is based on SVM to distinguish ncRNAs from coding RNAs using sequence features, such as ORF length, ORF coverage, ORF integrity, Fickett score, Hexamer score, PI, Gravy, Instability index, and CTD features. Therefore, we use L_{RED} to represent the features generated by CPPred.

$$L_{RED} = [ORF - integrity, ORF - civerage, Instability, \ldots, CTD] \quad (3)$$

2.3.2. lncRNA Feature DNC

L_{DNC} describes the A, G, C, and T to represent the trinucleotides by generating a 16-dimensional vector [34,35]. L_{DNC} can reflect the chemical properties of the accumulated energy of di-nucleotide and reflect the evolutionary information of lncRNA sequences. It can be computed as follows:

$$L_{DNC}(r,s) = \frac{N_{rs}}{N-1} \quad (4)$$

where N_{rs} is the number of di-nucleotide represented by nucleic acid types r and s, N is the length of a nucleotide sequence.

2.3.3. lncRNA Feature 3-mer

L_{3mer} represents the normalized occurrence frequencies of three neighboring base pairs in the RNA sequence [36], which has been successfully applied to human gene regulatory sequence prediction and enhancer identification [37]. It can be computed as follows:

$$L_{3mer}(t) = \frac{M(t)}{N}, t \in \{AAA, AAT, AAC, \ldots, GGG\} \quad (5)$$

where $M(t)$ is the number of k-mer type t, N is the length of a nucleotide sequence.

2.3.4. Protein Feature AAC

The protein sequence is composed of 20 kinds of amino acids. P_{ACC} provides information regarding the percentage of each residue present in the protein [38]. P_{ACC} can measure the correlation of two properties or the same properties (hydrophobicity, hydrophilicity, van der Waals normalized volume, polarity etc.) along the protein sequence and convert the matrix to a fixed-length vector [34,39]. It can be computed as follows:

$$f(t) = \frac{N(t)}{N} * 100 \quad (6)$$

where $N(t)$ is the number of amino acids type t, N is the length of the protein sequence.

2.3.5. Protein Features 3-mer and 4-mer

For P_{kmer}, amino acids are divided into seven groups according to the dipole moment and side-chain volume of the protein [32]: $R_1 = \{A, G, V\}$, $R_2 = \{I, L, F, P\}$, $R_3 = \{Y, M, T, S\}$, $R_4 = \{H, N, Q, W\}$, $R_5 = \{R, K\}$, $R_6 = \{D, E\}$, $R_7 = \{C\}$. Then, P_{3mer} (the frequency of occurrence of three adjacent coincidences in the protein sequence) and P_{4mer} (the frequency of occurrence of four adjacent symbols in the protein sequence) are obtained. It can be computed as follows:

$$fP_{3mer}(t) = \frac{Q(t)}{N}, t \in \{R_1R_1R_1, R_1R_1R_2, \ldots, R_7R_7R_7\} \quad (7)$$

$$P_{4mer}(t) = \frac{Q(t)}{N}, t \in \{R_1R_1R_1R_1, R_1R_1R_1R_2, \ldots, R_7R_7R_7R_7\} \quad (8)$$

where $Q(t)$ is the number of k-mer type t, N is the length of the protein sequence.

2.4. Structural Features

Molecular features that rely on lncRNA and protein structure information play a significant role in their interactions. Therefore, we used the secondary structure, hydrogen bonding propensities, and van der Waals interactions to represent the lncRNA's and protein's structure information.

For the lncRNA, its secondary structure was obtained through RNAfold [40] based on the minimum free energy algorithm and encoded by replacing each bracket with one and each dot with zero. Meanwhile, we adopted purine and pyrimidine contact information from a set of 41 RNA-protein complexes [41] in lncPro [42] to encode their hydrogen bonding propensities and van der Waals interactions. Each lncRNA structure is represented in these three numerical feature vectors.

For the protein, its secondary structure was obtained through Predator [43], based on its amino acid sequence, and was encoded by replacing each amino acid with the corresponding Chou–Fasman [44] propensity in LncADeep [45]. The hydrogen bonding propensities were encoded by using Grantham propensities [46] and Zimmerman propensities [47]. The Van der Waals interaction was encoded by using the Kyte–Doolittle [48] and Bull–Breese propensities [49]. Each protein structure is represented in these five numerical feature vectors.

However, each lncRNA and protein feature vector is of different dimension, which depends on the length of the corresponding RNA or protein sequence, and the CNN model requires fixed matrix inputs. We adopted the Fourier transform to unify the dimension. The formula of the Fourier series can be expressed as:

$$X'_k = \sqrt{\frac{2}{L}} \sum_{n=0}^{L} X_n \cos\left[\frac{\pi}{L}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right], k = 0, 1, \ldots, 9 \quad (9)$$

where L is the length of the original feature vector.

Here, those criteria in lncPro were adopted that use the first ten terms of the Fourier series as the new numerical feature vector. In this way, we obtain the lncRNA structure feature vector $B_1 = [L_{SS}, L_{OHB}, L_{OVW}]$ of dimension 30 and the protein structure feature vector $B_2 = [P_{SS}, P_{GHB}, P_{ZHB}, P_{KVW}, P_{BVW}]$ of dimension 50.

2.5. Convolutional Neural Networks

Convolutional neural network (CNN) is an effective tool in the field of predictive lncRNA–protein interaction [50,51]. Therefore, we introduce CNN as an algorithm to analyze the input raw sequence composition features, hand-designed features and structure features. In this work, the CNN model consists of four modules, including GloCNN, LocCNN, FC, and SS. The architecture of the CNN model and its detailed hyperparameters are shown in Supplementary Figure S1.

For the GloCNN module, the lncRNA and protein encoding matrices were fed into two-layer 1-channel CNNs (convolutional layer, batch normalization, max pooling layer) to extract raw sequence global features, where all the sequences were transformed into fixed-length equivalents. By experimentation, the prediction accuracy did not grow significantly when the layer number was larger than 2, and a larger hidden layer number brought more computation. In addition, a dropout layer was used to accelerate the training process and avoid overfitting. Finally, the lncRNA and protein feature maps were concatenated and their dimensionality reduced to 64 through a fully connected layer.

The LocCNN module used two-layer multi-channel CNNs to extract raw sequence local features, wherein all the sequences had multiple subsequences. The number of channels was determined by the local sequence encoding. After a dropout layer and fully connected layer, the lncRNA and protein feature maps were concatenated and their dimensionality reduced to 32.

For the FC module, the lncRNA and protein hand-designed feature vectors, A_1 and A_2, were fed into a two-layer fully connected layer to extract high-level features, followed concatenating the two feature maps, from which we extracted useful information through a fully connected layer.

The SS module used a two-layer fully connected layer to analyze the lncRNA and protein structure feature vectors, B_1 and B_2. Then, the feature maps were concatenated and further fed into a two-layer fully connected layer to extract useful information and reduce therr dimensionality to 32.

Finally, the feature maps of four basic module outputs were concatenated and further fed into the two fully connected layers to predict the probabilities of lncRNA–protein interactions.

2.6. Evaluation Metrics

In this study, we used the metrics of accuracy (ACC), Matthew's correlation coefficient (MCC), F1_score (F1), sensitivity (SN), specificity (SP), and positive predictive value (PPV) to measure the performance of LGFC-CNN. The formulas of the six measurements are as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (11)$$

$$F1 = 2 \times \frac{TP}{2TP + FP + FN} \tag{12}$$

$$SN = \frac{TP}{TP + FN} \tag{13}$$

$$SP = \frac{TN}{TN + FP} \tag{14}$$

$$PPV = \frac{TP}{TP + FP} \tag{15}$$

where TP, FP, TN, FN represent true positive, false positive, true negative, and false negative. Furthermore, we drew the area under the ROC curve (AUC) and the precision-recall curve (PRC) to measure the performance of LGFC-CNN.

3. Results

In this section, we first downloaded and ran the algorithms RPISeq-RF [15], RPISeq-SVM [15], LPI-BLS [17], IPMiner [16] following their respective papers and compared the performance of LGFC-CNN with these methods on the benchmark dataset RPI21850. We then compared LGFC-CNN with other methods on the datasets RPI7317 and RPI1847 to test the reliability and robustness of LGFC-CNN. We further used the random forest classifier [24] to compare various lncRNA and protein hand-designed feature combinations and analyze our negative sample strategy's effect. Finally, we verified the effectiveness of the proposed multi-type feature-combination method and the effect of the hyper-parameters in the CNN model. During the experiment process, we selected 70% of samples from these datasets as the training set, and then randomly selected 50% of the remaining data as the fixed-validation set and the assigned the remaining samples to the test set.

3.1. Performance of LGFC-CNN in Predicting lncRNA–Protein Interactions

To assess the performance of our LGFC-CNN, we first compared LGFC-CNN with RPISeq-RF, RPISeq-SVM, LPI-BLS, and IPMiner on the benchmark dataset RPI21850. The results of LGFC-CNN and the other four methods are shown in Table 2, from which we can see that performance of our LGFC-CNN was superior to the other four methods on the RPI21850 dataset. On the RPI21850 dataset, LGFC-CNN yielded an accuracy of 94.14%, which was 1.8%, 2.04%, 2.73, and 1.84% higher than that of RPISeq-RF, RPISeq-SVM, LPI-BLS, and IPMiner, respectively. The MCC of LGFC-CNN was 0.8853, which was 3.72%, 4.28%, 5.67%, and 3.92% higher than RPISeq-RF, RPISeq-SVM, LPI-BLS, and IPMiner, respectively. The F1-score of LGFC-CNN was 0.9435, which was 1.8%, 2.11%, 2.82%, and 1.97% higher than RPISeq-RF, RPISeq-SVM, LPI-BLS, and IPMiner, respectively. The SN of LGFC-CNN was 97.9%, which was 2.75%, 4%, 5.07%, and 4.55% higher than RPISeq-RF, RPISeq-SVM, LPI-BLS, and IPMiner, respectively.

Table 2. The result of LGFC-CNN and other four methods on the RPI21850 dataset.

Methods	ACC	MCC	F1-Score	SN	SP	PPV
LGFC-CNN	**0.9414**	**0.8853**	**0.9435**	**0.979**	0.9039	0.9106
RPISeq-RF	0.9234	0.8481	0.9255	0.9515	0.8954	0.9009
RPISeq-SVM	0.921	0.8425	0.9224	0.939	0.903	0.9063
LPI-BLS	0.9141	0.8286	0.9153	0.9283	0.8999	0.9027
IPMiner	0.923	0.8461	0.9238	0.9335	**0.9124**	**0.9142**

The ROC curves and the PRC curves of LGFC-CNN and the other four methods on RPI21850 are shown in Figure 4. From the figure, it can be seen that the AUC score of LGFC-CNN reached 0.9761, which is higher than that of RPISeq-SVM, RPISeq-RF, and IPMiner, respectively. The PRC score of LGFC-CNN reached 0.9697, and its curve can wrap the curves of other methods. These results indicate that LGRF-CNN performs well in predicting lncRNA–protein interactions.

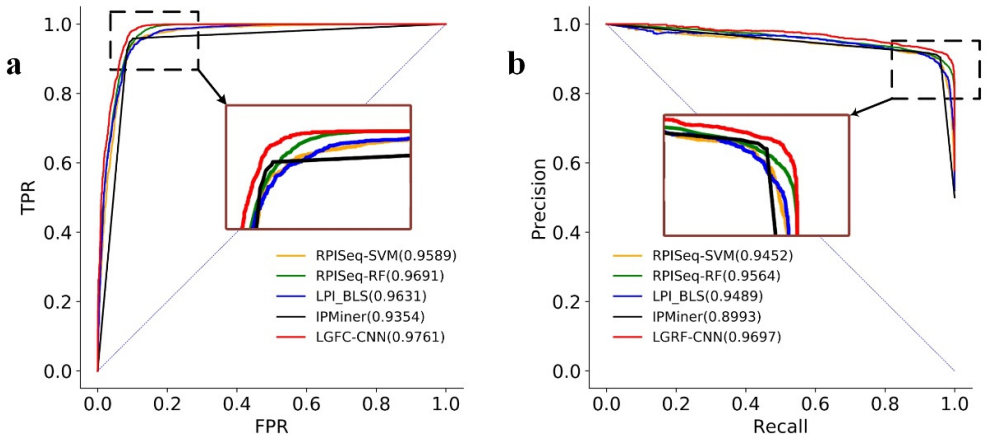

Figure 4. (**a**) The ROC curves and (**b**) the PRC curves of LGFC-CNN and other four methods on the RPI21850 dataset.

To further test the reliability and robustness of LGRF-CNN, we also comparde LGRF-CNN with four comparison methods on the RPI7317 and RPI1847 datasets. Table 3 lists the results of LGRF-CNN, RPISeq-RF, RPISeq-SVM, LPI-BLS, and IPMiner on these two datasets. On dataset RPI7317, the ACC of LGRF-CNN was 92.94%, which was better than RPISeq-RF (ACC: 90.98%), RPISeq-SVM (ACC: 91.53%), LPI-BLS (ACC: 91.44%), IPMiner (ACC: 91.34%), and showed specific improvement in the other five indicators. On dataset RPI1847, the ACC of LGRF-CNN was 98.19%, which is better than RPISeq-RF (ACC: 96.21%), RPISeq-SVM (ACC: 95.85%), LPI-BLS (ACC: 96.75%), IPMiner (ACC: 96.39%).

Table 3. The result of LGFC-CNN and other four methods on the RPI7317 and RPI1847 datasets.

RPI7317						
Methods	**ACC**	**MCC**	**F1-Score**	**SN**	**SP**	**PPV**
LGFC-CNN	**0.9294**	**0.8589**	**0.9299**	**0.9371**	**0.9217**	**0.922**
RPISeq-RF	0.9098	0.8202	0.9116	0.9299	0.8897	0.894
RPISeq-SVM	0.9153	0.8311	0.9169	0.9344	0.8961	0.9
LPI-BLS	0.9144	0.8288	0.9151	0.9226	0.9061	0.9077
IPMiner	0.9134	0.8269	0.9139	0.918	0.9088	0.9097
RPI1847						
LGFC-CNN	**0.9819**	**0.964**	**0.9818**	**0.9747**	0.9856	0.989
RPISeq-RF	0.9621	0.9243	0.9617	0.9531	0.9711	0.9706
RPISeq-SVM	0.9585	0.9191	0.957	0.9242	**0.9928**	**0.9922**
LPI-BLS	0.9675	0.9352	0.9672	0.9567	0.9783	0.9779
IPMiner	0.9639	0.9287	0.9631	0.9422	0.9856	0.9849

Figure 5 shows the ROC curves and PRC curves of LGRF-CNN and other methods on RPI7317 and RPI1847. The figure shows that the AUC scores of LGRF-CNN on RPI7317 and RPI1847 are 0.9785 and 0.9981, respectively, and the PRC scores reach 0.9781 and 0.9981, respectively. The curves of LGFC-CNN can wrap the curves of other methods. The above results show that different data sources will affect the performance of LGFC-CNN, but it can still maintain superior performance.

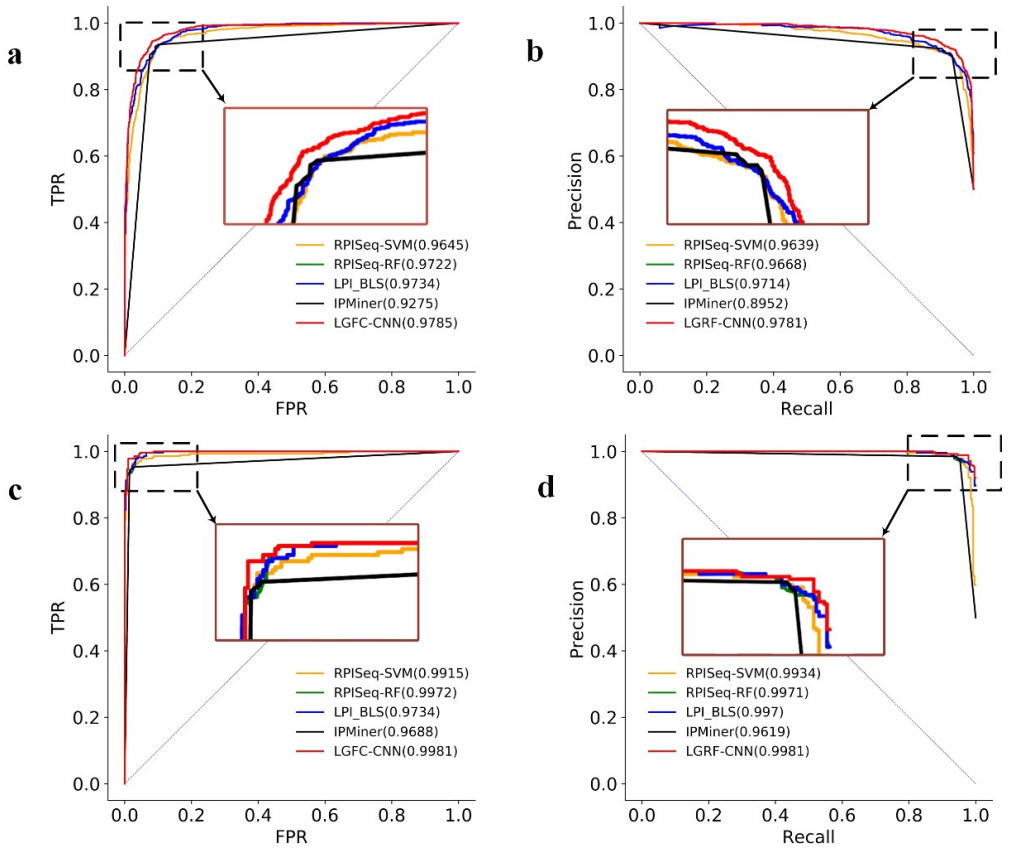

Figure 5. (**a**) The ROC curves and (**b**) the PRC curves of LGFC-CNN and other four methods on RPI7317. (**c**) The ROC curves and (**d**) the PRC curves of LGFC-CNN and other four methods on RPI1847.

3.2. Performance Comparison between Different Feature Combinations in Predicting lncRNA–Protein Interactions

Before feeding the hand-designed feature vectors into the FC module, we needed to select the top three features of superior predictive effect to represent the lncRNA and protein. Six features of the lncRNA were combined with ten features of the protein and each feature combination was ranked according to their individual average performances in the random forest classifier. For ranking the feature combinations, we compared the ACC of each combination to predict the lncRNA–protein interactions in RPI21850, and the results obtained through experiments are shown in supplementary Tables S1–S10. Figure 6 shows the results, visually, through a heat map.

As shown in our tables and figures, the average prediction accuracies of the combinations of L_{RED}, L_{PLIT}, L_{NAC}, L_{DNC}, L_{3mer}, L_{4mer} and ten protein features were 93.1%, 93.01%, 92.67%, 93.09%, 93.07%, 92.9%, respectively. The average prediction accuracies of the combinations of P_{AAC}, P_{Dis}, P_{DR}, P_{CC}, $P_{PC-PseAAC}$, P_{MAC}, $P_{SC-PseAAC}$, $P_{PseKRAAC}$, P_{3mer}, P_{4mer} and six lncRNA features were 93.07, 92.92%, 92.86%, 92.95%, 92.98%, 92.96%, 93.03%, 93.06%, 93.1%, 93.06%, also respectively. Accordingly, the three lncRNA features L_{red}, L_{dnc}, and L_{3mer} were selected to represent the lncRNA, and the three protein features P_{AAC}, P_{3mer}, and P_{4mer} were selected to represent the protein. These results indicate that the features related to coding potentials and physicochemical properties appear to be more suitable for predicting lncRNA–protein interactions.

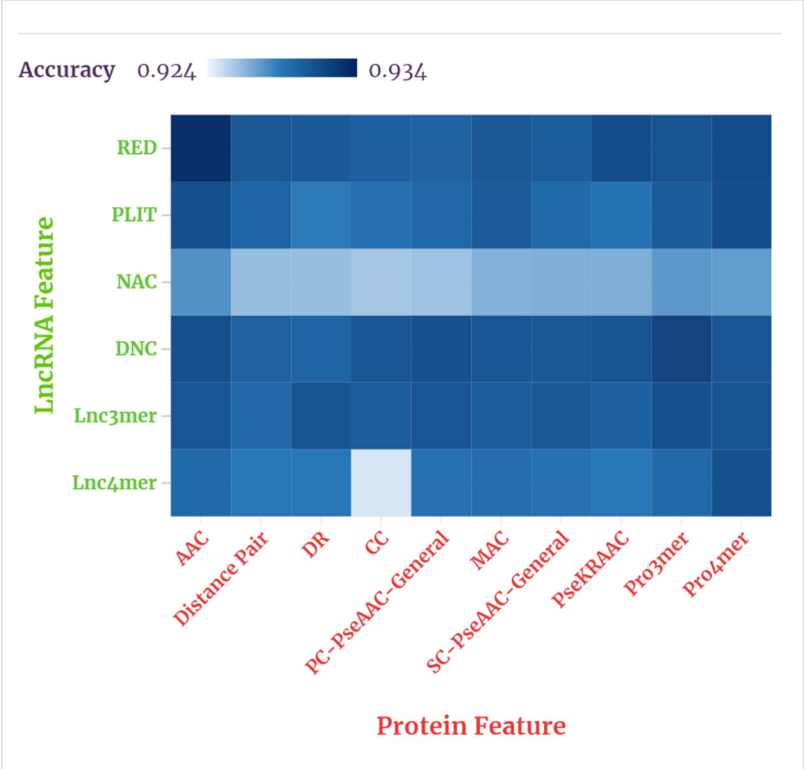

Figure 6. The heat map generated from the feature combinations in RPI21850.

3.3. Comparison between Four Modules of LGFC-CNN

Our LGFC-CNN model contains four fundamental modules: a LocCNN module, a GloCNN module, an FC module, and an SS module. To investigate the superiority of LGFC-CNN, we compared LGFC-CNN with four basic modules on datasets RPI21850, RPI7317, and RPI1847. The performance of LGFC-CNN and four different basic modules are shown in a histogram Figure 7.

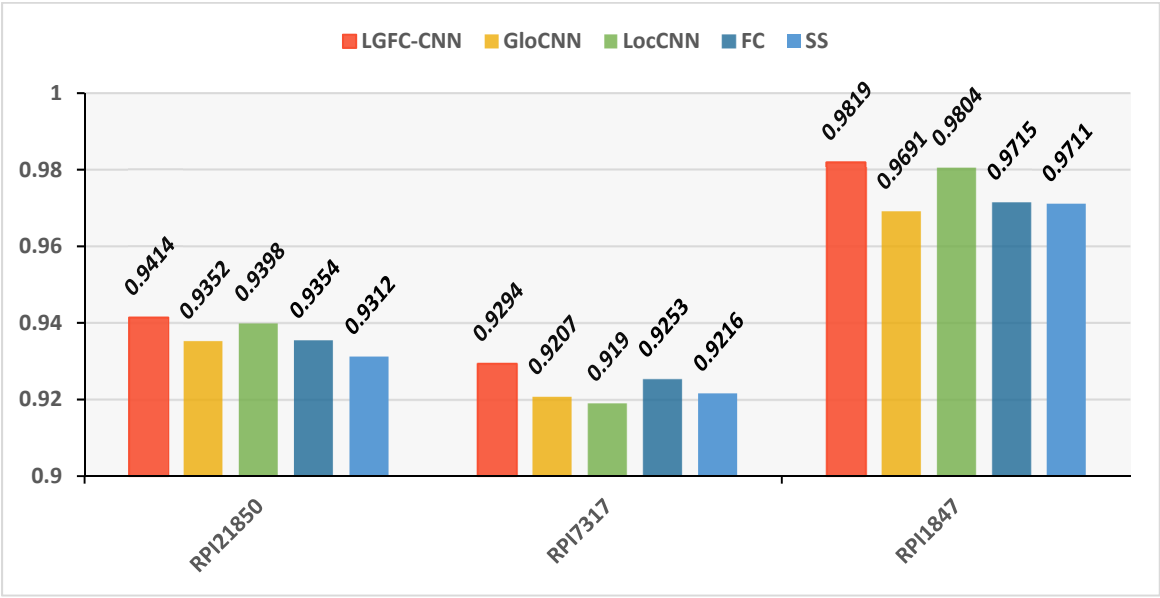

Figure 7. The performance of LGFC-CNN and four different basic modules.

The compared results show that, on the benchmark dataset RPI21850, the prediction accuracy of LocCNN module was 93.98%, which was higher than the 93.52% of the GloCNN module, the 93.54% of the FC module, and the 93.12% of the SS module. On PRI7317, the prediction accuracy of the FC module was 92.53%, which was higher than the 92.07% of the GloCNN module, the 91.9% of the LocCNN module, and the 92.16% of the SS module. On the dataset PRI1847, the prediction accuracy of LocCNN module was 98.04%, which was higher than the 96.91% of GloCNN module, the 97.15% of the FC module, and the 97.11% of the SS module. It can be seen from the results that the prediction accuracy of any single module could always exceed that of the other modules, and none of them was as good as the overall LGFC-CNN. It shows that the LGFC-CNN model we have proposed has the advantages of the four basic prediction modules. The combination of global sequence features, local sequence features, hand-designed features, and structural features can provide more comprehensive lncRNA–protein prediction results.

3.4. Effectiveness of Selecting a Negative Sample Strategy

To show the effectiveness of our strategy for selecting negative samples, we first constructed three new datasets, ranRPI21850, ranRPI7317, and ranRPI1847, by randomly pairing the lncRNA and protein and removing the duplicate interaction pairs. ranRPI21850 contained 21,850 lncRNA–protein interaction pairs and 21,850 pairs of lncRNA–protein non-interaction pairs; ranRPI7317 contained 7317 lncRNA–protein interaction pairs and 7317 pairs of lncRNA–protein non-interaction pairs; and ranRPI1847 contained 1847 lncRNA–protein pairs. Then we predicted the lncRNA–protein interactions in RPI21850, RPI7317, RPI1847, ranRPI21850, ranRPI7317, and ranRPI1847 under the same conditions. Table 4 shows the effect of LGFC-CNN on three datasets generated by our negative sample-generation strategy and random pair generation strategy.

It can be seen from Table 4 that the accuracy of LGFC-CNN on RPI21850 was 94.14%, which was 2.59% higher than that on ranRPI21850. The accuracy on RPI7317 was 92.94%, which was 2.96% higher than that on ranRPI7317. The accuracy on RPI1847 was 98.19%, which was 1.44% higher than that on ranPRI1847. These results show that the strategy of selecting negative samples used in this work is effective and can improve the prediction performance for lncRNA–protein interactions.

Table 4. Results of LGFC-CNN on six datasets generated by different negative sample generation strategy.

Datasets	ACC	MCC	F1-Score	SN	SP	PPV
RPI21850	**0.9414**	**0.8853**	**0.9435**	0.979	**0.9039**	**0.9106**
ranRPI21850	0.9155	0.8381	0.9207	**0.9805**	0.8505	0.8677
RPI7317	**0.9294**	**0.8589**	**0.9299**	**0.9371**	**0.9217**	**0.9228**
ranRPI7317	0.8998	0.7996	0.9003	0.9052	0.8944	0.8954
RPI1847	**0.9819**	**0.964**	**0.9818**	0.9747	**0.9856**	**0.989**
ranRPI1847	0.9675	0.9359	0.9682	**0.9892**	0.9458	0.9481

3.5. Effects of Hyper-Parameters in LGFC-CNN

Our LGFC-CNN model consists of four basic modules for analyzing global sequence features, local sequence features, hand-designed features, and structure features. To investigate how the hyper-parameters of the convolutional layer kernel number and the fully connected layer neuron number affect the performance of LGFC-CNN, we change one pa-rameter value at a time by fixing other parameters on the validation set of RPI21850 to im-plement our LGFC-CNN. For the GloCNN module, Kernel-G of the convolutional layer was set as n $\times$ 10, n $\times$ 20, n $\times$ 30, n $\times$ 40, n $\times$ 50 (n for lncRNA is 4, n for protein is 7). For the LocCNN module, Kernel-L of the convolutional layer was set as n $\times$ 10, n $\times$ 20, n $\times$ 30, n $\times$ 40, n $\times$ 50 (n for lncRNA is 4, n for protein is 7). The neuron number of the four fully connected layers Dense-G, Dense-L, Dense-FC, and Dense-SS was set to 16, 32, 48, and 64, respectively.

Supplementary Tables S11–16 shows the results of LGFC-CNN using different hyper-parameters on the validation set of RPI21850, from which we can find that the hy-per-parameters have some influence on the prediction results. When setting the kernel-G as n × 40, kernel-L as n × 30, Dense-G as 64, Dense-L as 32, Dense-FC as 32, and Dense-SS as 32, LGFC-CNN achieves the best performance.

4. Discussion

In this study, we proposed a novel method based on deep learning and using multiple types of features to predict lncRNA–protein interactions. On the benchmark dataset we constructed, LGFC-CNN achieved an accuracy of 94.14%, an MCC of 0.8853, an F1-score of 0.9435, an SN of 97.9%, an SP of 90.39%, and a PPV of 91.06%. The experimental results on RPI7317 and RPI1847 also showed the effectiveness of LGFC-CNN.

LGFC-CNN had superior performance in predicting lncRNA–protein interactions; we believe that there are several reasons why. Firstly, for the structure of the negative sample strategy, our method considers that if protein p' is dissimilar to q, there is a low possibility that p' interacts r [27], and we also consider that there is a certain probability that lncRNA is related to those proteins with higher scores. The experimental results verify that our negative sample strategy is more effective than the commonly used random pairing method. Secondly, in the raw sequence features, we used both global sequence features and local sequence features. Our method considers the global sequence's overall characteristics and considers the critical role of the local sequence in the lncRNA–protein interaction. Thirdly, we compare the combinations of various types of lncRNA features and protein features in terms of hand-designed features and use secondary features, hydrogen bonding propensities and van der Waals interactions as structure features. Finally, the combination of global sequence features, local sequence features, hand-designed features, and structure features can provide more comprehensive lncRNA-protein prediction results.

Although LGFC-CNN achieves better performance in predicting lncRNA–protein interactions, there are still limitations. On the one hand, since most of the high-quality experimentally verified human lncRNA–protein pairs are mainly derived from the NPInter dataset [18], our method can still only be trained on a few datasets. When there are multiple data sources, deep learning can play a more significant role, so more data sources are needed to cover more possible situations. On the other hand, there are many types of hand-designed and structural features, and what we have here-explored is only part of them. Finding better hand-designed and structural features and exploring better network structures to improve the hand-designing of feature and structure modules' performance will be the focus of our future work. In future work, we will explore the effect of LGFC-CNN in the interaction between lncRNA with miRNA and hope to find a suitable method that can integrate the LPI similarity network, so that more types of features can be used to improve the classification effect.

5. Conclusions

To understand the various regulatory mechanisms and pathogenic mechanisms involved in lncRNA through lncRNA–protein interactions, some computed methods were developed for predicting lncRNA–protein interactions [11–13]. However, there is currently no method to combine raw sequence features, hand-designed features and structural features to predict lncRNA–protein interactions. In this work, we presented a novel deep learning method, LGFC-CNN, to predict lncRNA–protein interactions by using multiple types of features. We introduced two sequence preprocessing methods to transform arbitrarily long sequences into fixed-length sequences and feed them into two modules to gain raw global and local sequence features. Meanwhile, we selected hand-designed features by comparing the predictive effects of different lncRNA and protein feature combinations. Furthermore, we obtained lncRNA and protein structural features and unified their dimensions through a Fourier transform. The experimental results show that our LGFC-CNN had a better performance than other latest methods. All in all, LGFC-CNN is a feasible and

effective tool for predicting human lncRNA–protein interactions. Our LGFC-CNN model combines raw sequence features, hand-designed features and structure features is also a potential tool for the other bioinformatics classification tasks.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/genes12111689/s1. The explanations of the other three lncRNA and seven protein feature encodings we use for comparison. Figure S1: The architecture of the CNN model; Tables S1–S10: The tables of six lncRNA feature and ten protein feature combinations; Tables S11–S16: The tables of LGFC-CNN using different hyper-parameters on validation set of RPI21850.

Author Contributions: Conceptualization— S.J., L.H. and S.Y.; methodology—S.J., L.H. and Y.W.; data curation—S.J. and Y.W.; writing, original draft preparation—S.J. and S.Y.; writing, review, and editing—S.J., S.Z., R.G. and X.Z. All authors have read and approved final manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (No. 62072212), the Development Project of Jilin Province of China (Nos. 20200401083GX, 2020C003), and Guangdong Key Project for Applied Fundamental Research (No. 2018KZDXM076). This work was also supported by Jilin Province Key Laboratory of Big Data Intelligent Computing (No. 20180622002JC).

Data Availability Statement: The source code and datasets for this work can be obtained from https://github.com/consen1/LGFC-CNN (accessed on 20 June 2021).

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

References

1. Khalil, A.M.; Rinn, J.L. RNA–protein interactions in human health and disease. *Semin. Cell Dev. Biol.* **2011**, *22*, 359–365. [CrossRef]
2. Li, C.H.; Chen, Y. Targeting long non-coding RNAs in cancers: Progress and prospects. *Int. J. Biochem. Cell Biol.* **2013**, *45*, 1895–1910. [CrossRef] [PubMed]
3. Statello, L.; Guo, C.-J.; Chen, L.-L.; Huarte, M. Gene regulation by long non-coding RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **2021**, *22*, 96–118. [CrossRef]
4. Derrigo, M.; Cestelli, A.; Savettieri, G.; Di Liegro, I. RNA-protein interactions in the control of stability and localization of messenger RNA (review). *Int. J. Mol. Med.* **2000**, *5*, 111–123. [CrossRef]
5. Barbagallo, C.; Di Maria, A.; Alecci, A.; Barbagallo, D.; Alaimo, S.; Colarossi, L.; Ferro, A.; Di Pietro, C.; Purrello, M.; Pulvirenti, A.; et al. VECTOR: An Integrated Correlation Network Database for the Identification of CeRNA Axes in Uveal Melanoma. *Genes* **2021**, *12*, 1004. [CrossRef]
6. Sardina, D.S.; Alaimo, S.; Ferro, A.; Pulvirenti, A.; Giugno, R. A novel computational method for inferring competing endogenous interactions. *Briefiengs Bioinform.* **2017**, *18*, 1071–1081. [CrossRef]
7. Pan, X.; Shen, H.-B. OUGENE: A disease associated over-expressed and under-expressed gene database. *Sci. Bull.* **2016**, *61*, 752–754. [CrossRef]
8. Liu, F.; Hu, S.; Zhao, N.; Shao, Q.; Li, Y.; Jiang, R.; Chen, J.; Peng, W.; Qian, K. LncRNA-5657 silencing alleviates sepsis-induced lung injury by suppressing the expression of spinster homology protein 2. *Int. Immunopharmacol.* **2020**, *88*, 106875. [CrossRef] [PubMed]
9. Dou, Q.; Xu, Y.; Zhu, Y.; Hu, Y.; Yan, Y.; Yan, H. LncRNA FAM83H-AS1 contributes to the radioresistance, proliferation, and metastasis in ovarian cancer through stabilizing HuR protein. *Eur. J. Pharmacol.* **2019**, *852*, 134–141. [CrossRef]
10. Yan, W.; Chen, Z.-Y.; Chen, J.-Q.; Chen, H.-M. LncRNA NEAT1 promotes autophagy in MPTP-induced Parkinson's disease through stabilizing PINK1 protein. *Biochem. Biophys. Res. Commun.* **2018**, *496*, 1019–1024. [CrossRef] [PubMed]
11. Zhang, W.; Qu, Q.; Zhang, Y.; Wang, W. The linear neighborhood propagation method for predicting long non-coding RNA–protein interactions. *Neurocomputing* **2018**, *273*, 526–534. [CrossRef]
12. Zhao, Q.; Yu, H.; Ming, Z.; Hu, H.; Ren, G.; Liu, H. The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. *Mol. Ther. Nucleic Acids* **2018**, *13*, 464–471. [CrossRef]
13. Zhu, R.; Li, G.; Liu, J.-X.; Dai, L.-Y.; Guo, Y. ACCBN: Ant-Colony-clustering-based bipartite network method for predicting long non-coding RNA–protein interactions. *BMC Bioinform.* **2019**, *20*, 16. [CrossRef]
14. Ge, M.; Li, A.; Wang, M. A Bipartite Network-based Method for Prediction of Long Non-coding RNA–protein Interactions. *Genom. Proteom. Bioinform.* **2016**, *14*, 62–71. [CrossRef]
15. Muppirala, U.K.; Honavar, V.G.; Dobbs, D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinform.* **2011**, *12*, 489. [CrossRef]
16. Pan, X.; Fan, Y.-X.; Yan, J.; Shen, H.-B. IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genom.* **2016**, *17*, 582. [CrossRef]
17. Fan, X.-N.; Zhang, S.-W. LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier. *Neurocomputing* **2019**, *370*, 88–93. [CrossRef]

18. Liu, H.; Ren, G.; Hu, H.; Zhang, L.; Ai, H.; Zhang, W.; Zhao, Q. LPI-NRLMF: lncRNA–protein interaction prediction by neighborhood regularized logistic matrix factorization. *Oncotarget* **2017**, *8*, 103975–103984. [CrossRef]
19. Peng, L.; Liu, F.; Yang, J.; Liu, X.; Meng, Y.; Deng, X.; Peng, C.; Tian, G.; Zhou, L. Probing lncRNA–Protein Interactions: Data Repositories, Models, and Algorithms. *Front. Genet.* **2020**, *10*, 1346. [CrossRef]
20. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
21. Pan, X.; Shen, H.-B. Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics* **2018**, *34*, 3427–3436. [CrossRef]
22. Xiang, X.; Duan, S.; Pan, H.; Han, P.; Cao, J.; Liu, C. From One-Hot Encoding to Privacy-Preserving Synthetic Electronic Health Records Embedding. In Proceedings of the 2020 International Conference on Cyberspace Innovation of Advanced Technologies, Guangzhou, China, 4–6 December 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 407–413.
23. Quang, D.; Xie, X. DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Res.* **2016**, *44*, e107. [CrossRef]
24. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
25. Teng, X.; Chen, X.; Xue, H.; Tang, Y.; Zhang, P.; Kang, Q.; Hao, Y.; Chen, R.; Zhao, Y.; He, S. NPInter v4.0: An integrated database of ncRNA interactions. *Nucleic Acids Res.* **2020**, *48*, D160–D165. [CrossRef] [PubMed]
26. Hao, Y.; Wu, W.; Li, H.; Yuan, J.; Luo, J.; Zhao, Y.; Chen, R. NPInter v3.0: An upgraded database of noncoding RNA-associated interactions. *Database* **2016**, *2016*, baw057. [CrossRef]
27. Cheng, Z.; Huang, K.; Wang, Y.; Liu, H.; Guan, J.; Zhou, S. Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst. Biol.* **2017**, *11*, 9. [CrossRef]
28. Cock, P.J.A.; Antao, T.; Chang, J.T.; Chapman, B.A.; Cox, C.J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; et al. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **2009**, *25*, 1422–1423. [CrossRef] [PubMed]
29. Zhao, L.; Wang, J.; Li, Y.; Song, T.; Wu, Y.; Fang, S.; Bu, D.; Li, H.; Sun, L.; Pei, D.; et al. NONCODEV6: An updated database dedicated to long non-coding RNA annotation in both animals and plants. *Nucleic Acids Res.* **2021**, *49*, D165–D171. [CrossRef]
30. UniProt Consortium, T. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **2018**, *46*, 2699. [CrossRef]
31. Luo, J.; Liu, L.; Venkateswaran, S.; Song, Q.; Zhou, X. RPI-Bind: A structure-based method for accurate identification of RNA-protein binding sites. *Sci. Rep.* **2017**, *7*, 614. [CrossRef]
32. 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]
33. Tong, X.; Liu, S. CPPred: Coding potential prediction based on the global description of RNA sequence. *Nucleic Acids Res.* **2019**, *47*, e43. [CrossRef]
34. Liu, B.; Gao, X.; Zhang, H. BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.* **2019**, *47*, e127. [CrossRef]
35. Li, J.; Zhang, L.; He, S.; Guo, F.; Zou, Q. SubLocEP: A novel ensemble predictor of subcellular localization of eukaryotic mRNA based on machine learning. *Briefings Bioinform.* **2021**, *22*, bbaa401. [CrossRef]
36. Manavalan, B.; Basith, S.; Shin, T.H.; Lee, G. Computational prediction of species-specific yeast DNA replication origin via iterative feature representation. *Briefings Bioinform.* **2020**, *22*, bbaa304. [CrossRef]
37. Lee, D.; Karchin, R.; Beer, M.A. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.* **2011**, *21*, 2167–2180. [CrossRef]
38. Agrawal, P.; Bhagat, D.; Mahalwal, M.; Sharma, N.; Raghava, G.P.S. AntiCP 2.0: An updated model for predicting anticancer peptides. *Briefings Bioinform.* **2020**. [CrossRef]
39. Dong, Q.; Zhou, S.; Guan, J. A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics* **2009**, *25*, 2655–2662. [CrossRef]
40. Lorenz, R.; Bernhart, S.H.; Höner Zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.F.; Hofacker, I.L. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **2011**, *6*, 26. [CrossRef]
41. Morozova, N.; Allers, J.; Myers, J.; Shamoo, Y. Protein–RNA interactions: Exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics* **2006**, *22*, 2746–2752. [CrossRef]
42. Lu, Q.; Ren, S.; Lu, M.; Zhang, Y.; Zhu, D.; Zhang, X.; Li, T. Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genom.* **2013**, *14*, 651. [CrossRef]
43. Frishman, D.; Argos, P. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng. Des. Sel.* **1996**, *9*, 133–142. [CrossRef]
44. Chou, P.Y.; Fasman, G.D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **1978**, *47*, 45–148. [CrossRef] [PubMed]
45. Yang, C.; Yang, L.; Zhou, M.; Xie, H.; Zhang, C.; Wang, M.D.; Zhu, H. LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics* **2018**, *34*, 3825–3834. [CrossRef]
46. Grantham, R. Amino acid difference formula to help explain protein evolution. *Science* **1974**, *185*, 862–864. [CrossRef]
47. Zimmerman, J.M.; Eliezer, N.; Simha, R. The characterization of amino acid sequences in proteins by statistical methods. *J. Theor. Biol.* **1968**, *21*, 170–201. [CrossRef]

48. Kyte, J.; Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **1982**, *157*, 105–132. [CrossRef]
49. Bull, H.B.; Breese, K. Surface tension of amino acid solutions: A hydrophobicity scale of the amino acid residues. *Arch. Biochem. Biophys.* **1974**, *161*, 665–670. [CrossRef]
50. Wang, L.; You, Z.; Huang, D.; Zhou, F. Combining high speed ELM learning with a deep convolutional neural network feature encoding for predicting protein-RNA interactions. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2020**, *17*, 972–980. [CrossRef]
51. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *Insights Imaging* **2018**, *9*, 611–629. [CrossRef]