frontiers in Cell and Developmental Biology

ORIGINAL RESEARCH
published: 01 February 2021
doi: 10.3389/fcell.2020.614080

DeepRTCP: Predicting ATP-Binding Cassette Transporters Based on 1-Dimensional Convolutional Network

Zhaoxi Zhang[1], *Juan Wang*[1,2]* *and Jiameng Liu*[1]

[1] School of Computer Science, Inner Mongolia University, Hohhot, China, [2] Stage Key Laboratories of Reproductive Regulation & Breeding of Grassland Livestock, Hohhot, China

ATP-binding cassette (ABC) transporters can promote cells to absorb nutrients and excrete harmful substances. It plays a vital role in the transmembrane transport of macromolecules. Therefore, the identification of ABC transporters is of great significance for the biological research. This paper will introduce a novel method called DeepRTCP. DeepRTCP uses the deep convolutional neural network and a feature combined of reduced amino acid alphabet based tripeptide composition and PSSM to recognize ABC transporters. We constructed a dataset named ABC_2020. It contains the latest ABC transporters downloaded from Uniprot. We performed 10-fold cross-validation on DeepRTCP, and the average accuracy of DeepRTCP was 95.96%. Compared with the start-of-the-art method for predicting ABC transporters, DeepRTCP improved the accuracy by 9.29%. It is anticipated that DeepRTCP can be used as an effective ABC transporter classifier which provides a reliable guidance for the research of ABC transporters.

Keywords: ABC transporters, deep convolutional neural network, tripeptide composition, cross validation, PSSM

OPEN ACCESS

Edited by:
Liang Cheng,
Harbin Medical University, China

Reviewed by:
Junwei Luo,
Henan Polytechnic University, China
Ancuta Jurj,
Iuliu Hatieganu University of Medicine and Pharmacy, Romania
Junwei Han,
Harbin Medical University, China

***Correspondence:**
Juan Wang
wangjuan@imu.edu.cn

Specialty section:
This article was submitted to Molecular Medicine, a section of the journal Frontiers in Cell and Developmental Biology

Received: *05 October 2020*
Accepted: *24 December 2020*
Published: *01 February 2021*

Citation:
Zhang Z, Wang J and Liu J (2021) DeepRTCP: Predicting ATP-Binding Cassette Transporters Based on 1-Dimensional Convolutional Network. Front. Cell Dev. Biol. 8:614080. doi: 10.3389/fcell.2020.614080

1. INTRODUCTION

The ABC transporter is a member of ATP-binding protein superfamily. The core structures of ABC transporters are two nucleotide-binding domains and two transmembrane domains (Abbas et al., 2015). The nucleotide-binding domain is a conserved domain. It can help the transmembrane domain to perform the function. Subdomains of the nucleotide-binding domain have some conserved sequence motifs with specific functions. The most important motifs are Walker-A motifs and LSGGQ motifs. In the process of molecular transports, the two nucleotide-binding domains bind together. There will be two ATP-binding sites and two hydrolysis sites between the Walker-A motifs of one nucleotide-binding domain and the other nucleotide-binding domain (Chen et al., 2003). The ABC transporter performs its transport functions based on the nucleotide-binding domain and the transmembrane domain. The transport functions of ABC transporters are divided into two types: inward transport and outward transport. The inward ABC transporter exists not only in prokaryotes but also in eukaryotes. It can promote the transport of nutrients such as amino acids and carbohydrates from the extracellular environment into the intracellular matrix, thereby promoting cell growth. Outward ABC transporters, like inward ABC transporters, coexist in prokaryotes and eukaryotes. They can expel antibiotics, fatty acids and other substances that are not conducive to cell growth. Outward ABC transporters help cells to keep non-essential foreign

substances or secondary metabolites in a low concentration range, thereby reducing the growth pressure of the cells, maintaining the normal growth of the cells, and greatly improving the survival rate of the cells (Gedeon et al., 2006; Davidson et al., 2008; Cui and Davidson, 2011). Based on these physiological characteristics, the identification of ABC transporters is of great significance not only for the development of biomedicine, but also for the crop cultivation and the microbial industry. Effective and accurate ABC identification methods are urgently needed.

Biological experiments are reliable methods for identifying protein functions. But most of these methods require expensive equipments and long experimental cycles. In addition, the experimental method cannot give priority to a function which needs to be identify urgently (Konc et al., 2013). With the development of sequencing technology, a large number of new protein sequences have been discovered. Biological experiments alone are not enough to meet the growing need for protein function identification. Researchers need a fast and accurate method to help identify protein functions. In recent years, predicted methods for protein functions have been widely used. Predicted methods have improved the efficiency of protein identifications, and its accuracy is also high. Protein function predictions usually use machine learning algorithms (Libbrecht. and Noble, 2015) as the classifiers such as support vector machine (SVM) (Suykens and Vandewalle, 1999), random forest (RF) (Vladimir et al., 2003), naive bayes (NB) (Rish, 2001) and artificial neural network (ANN), and obtain good results. The development of deep learning (Lecun et al., 2015) has further improved the performance of predicted methods (Gligorijevic et al., 2018; You et al., 2018). The most commonly used deep learning algorithm is the deep convolutional neural network (DCNN) (Lecun and Bottou, 1998). It has got good results in both the identification of protein functional sites and the protein function prediction (Kulmanov and Robert, 2019; Zhang and Yu, 2019).

A good feature is also crucial for the protein function prediction. In past studies, researchers usually used information extracted from proteins as features, including protein-protein interactions (Haretsugu et al., 2010; Jiang, 2012), structural information (Zhang et al., 2016; Le et al., 2019), physicochemical property (Cai et al., 2003), amino acid composition (Luo et al., 2010), evolutional information (Mundra et al., 2007), and the combinations of different information mentioned above (Chen et al., 2012, 2013; Song et al., 2014; Zou et al., 2016). Among these features, amino acid composition and evolutional information have been widely used. The amino acid composition is divided into peptide composition, dipeptide composition, and tripeptide composition, etc. The tripeptide composition contains more information than the peptide composition and the dipeptide composition. But the dimension of tripeptide composition is large. The tripeptide composition of amino acids sequence is an 8000-dimensional vector. The tripeptide composition of most proteins is a sparse vector, which affects the use of tripeptide composition in protein function predictions. Lin et al. (2017) divided the 20 amino acids into several pseudo-amino-acids called reduced amino acid alphabet (RAAA). By using RAAA to represent protein sequences, the dimension of the tripeptide composition can be reduced, thereby improving the accuracy of protein function predictions. The position specific score matrix (PSSM) (Michael et al., 1987) contains evolutional information of a protein. It is obtained from the multiple sequence alignment (MSA). PSSM contains statistical information about the distribution of residues at different positions in a MSA. The value in PSSM represents the score that a residue at one position will mutate to another residue during evolution. PSSM contains information about the homologous sequence of the query protein, which is not available in other sequence-based protein features. It has achieved good results in protein function predictions (Wang et al., 2018, 2019; Gao et al., 2019).

TABLE 1 | Classification of amino acids based on different types of physicochemical properties.

Physicochemical properties	Class 1	Class 2	Class 3
Hydrophobicity	RKEDQN	GASTPHY	CVLIMFW
Surface tension	GQDNAHR	KTSEC	ILMFPWYV
Solvent solubility	ALFCGIVW	KTSEC	MPSTHY
Charged polarity	LIFWCMVY	PATGS	HQRKEND

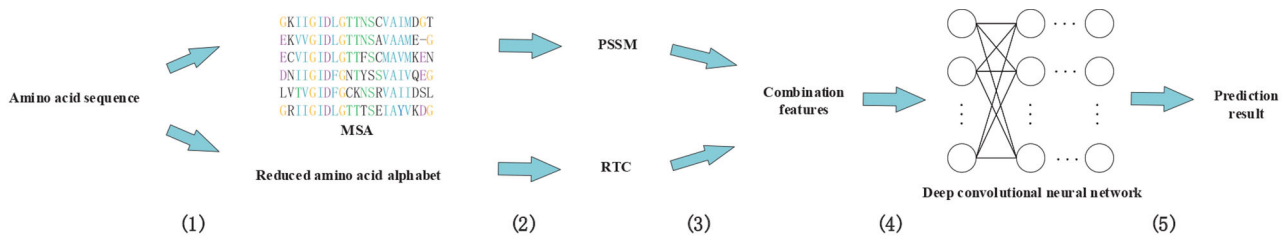

FIGURE 1 | An example using DeepRTCP to predict ABC transporters. (1) Multiple sequence alignments were built for query proteins, and amino acid sequences were replaced with reduced amino acids sequences. (2) The PSSM of the query protein was calculated from multiple sequence alignment, and the RTC was extracted from the reduced amino acids sequence. (3) PSSM and RTC combine into RTCP. (4) Deep convolutional neural network uses RTCP as input to train the classifier. (5) The output of the convolutional neural network passes through a binary classifier to get the prediction result.

In this study, we proposed a novel method called DeepRTCP. It applies the 1-dimensional DCNN and a feature combined of the PSSM and RAAA based tripeptide composition to predict ABC transporters. This experiment used a dataset named ABC_2020. Through experimental comparison, we chose the feature based on surface tension and solvent solubility, and chose a 7-layers DCNN as the classifier. Finally, we compared DeepRTCP with the state-of-the-art method for predicting ABC transporters. The results show that DeepRTCP is better than the existing method in all evaluation indicators used in this article. The overall process of using DeepRTCP to predict ABC transporters is shown in **Figure 1**.

TABLE 2 | RAAAs based on different types of physicochemical properties.

Combination of different classes	HP_ST	HP_SS	HP_CP	ST_SS	ST_CP	SS_CP
Class-1-1	RQND				GA	ICWFLV
Class-1-2	EK	NEQRDK		AG	NQDR	AG
Class-1-3			RKDQEN	RNDQH	H	
Class-2-1	AGH	AG	Y	C	C	
Class-2-2	ST		GPSTA	ST	EK	
Class-2-3	PY	SPYHT	H	KE	TS	RDENQK
Class-3-1		CFIVWL	LFVCIWM	MYVLFWI	FLWIV	MY
Class-3-2	C			P		SPT
Class-3-3	WIFLMV	M			MYP	LH

Class-i-j represents a class of reduced amino acid that determined by the intersection set between Class-i of one physicochemical property and Class-j of another physicochemical property.

TABLE 3 | Comparison among methods based on different types of RTCP.

Type of RTCP	Acc	Spec	Sens	F-score	Mcc
HP_ST	93.54%	92.21%	94.76%	0.9343	0.8716
HP_SS	93.44%	92.26%	94.28%	0.9334	0.8668
HP_CP	93.23%	93.23%	93.33%	0.9421	0.8619
ST_SS	93.94%	92.45%	94.76%	0.9383	0.8810
ST_CP	93.68%	92.30%	94.28%	0.9357	0.8715
SS_CP	93.32%	92.26%	94.28%	0.9323	0.8691

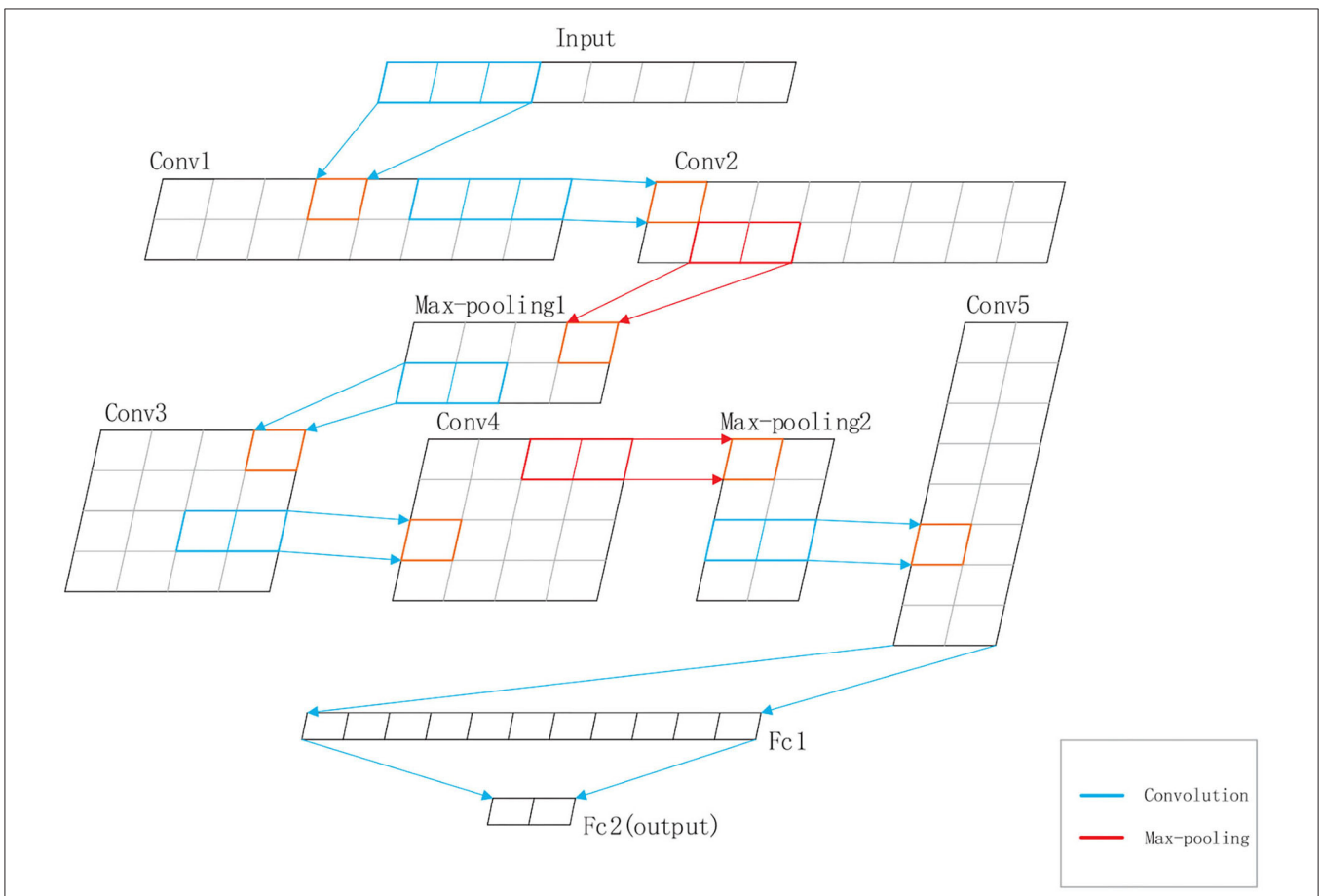

FIGURE 2 | The structure of the DCNN used in DeepRTCP. The DCNN has five convolutional layers: Conv1, Conv2, Conv3, Conv4, and Conv5, five batchnormalization layers, two max-pooling layers: Max-pooling1 and Max-pooling2, and two fully connected layers: Fc1 and Fc2. Fc2 is the output layer. The blue line represents the convolution. The red line represents the max-pooling.

2. MATERIALS AND METHODS

2.1. Dataset

This experiment used a dataset named ABC_2020. It includes 2,105 positives and 2,105 negatives. The positives were download from the Swiss-prot database of Uniprot (Amos et al., 2009) by using a key word "ABC transporter." We used CD-HIT to remove redundant sequences in the downloaded data. We hope to keep a large number of samples while reducing the impact of redundant sequences on the model, so that our model can be fully trained. So we chose 0.6 as the E-value of CD-HIT. We selected the protein families in Pfam (Finn et al., 2014), which do not contain the proteins in positives. Then we took the longest protein sequence in each protein family as a negative. We got 9,736 negatives and randomly selected 2,105 sequences from these negatives as the negative set of ABC_2020.

TABLE 4 | Comparison between RTCP based method and TCP based methods.

Feature	Acc	Spec	Sens	F-score	Mcc
TCP_8020	93.25%	91.92%	94.00%	0.9314	0.8507
TCP_1000	93.39%	92.11%	94.59%	0.9330	0.8645
TCP_500	93.49%	92.21%	95.50%	0.9339	0.8738
TCP_200	93.34%	92.01%	94.59%	0.9325	0.8645
TCP_80	93.16%	92.40%	93.69%	0.9309	0.8552
RTCP	93.94%	92.45%	94.76%	0.9383	0.8810

The number behind TCP represents the dimension of TCP.

2.2. Feature Extraction

We proposed a novel RAAA based on the physicochemical property of amino acids. The research used four physicochemical properties, including hydrophobicity (HP), surface tension (ST), solvent solubility (SS), and charged polarity (CP). Each physicochemical property divides amino acids into three classes (as shown in **Table 1**). We regarded the intersection set of a class of amino acid based on a type of physicochemical property and a class of amino acid based on another type of physicochemical property as a reduced amino acid. Therefore, every two types of physicochemical properties can determine an expression of RAAA. **Table 2** shows six RAAA representations: HP_ST, HP_SS, HP_CP, ST_SS, ST_CP, and SS_CP. We replaced the amino acid sequence with a RAAA sequence. The RAAAs of HP_ST, HP_SS, HP_CP, ST_SS, ST_CP, and SS_CP are composed of 7,

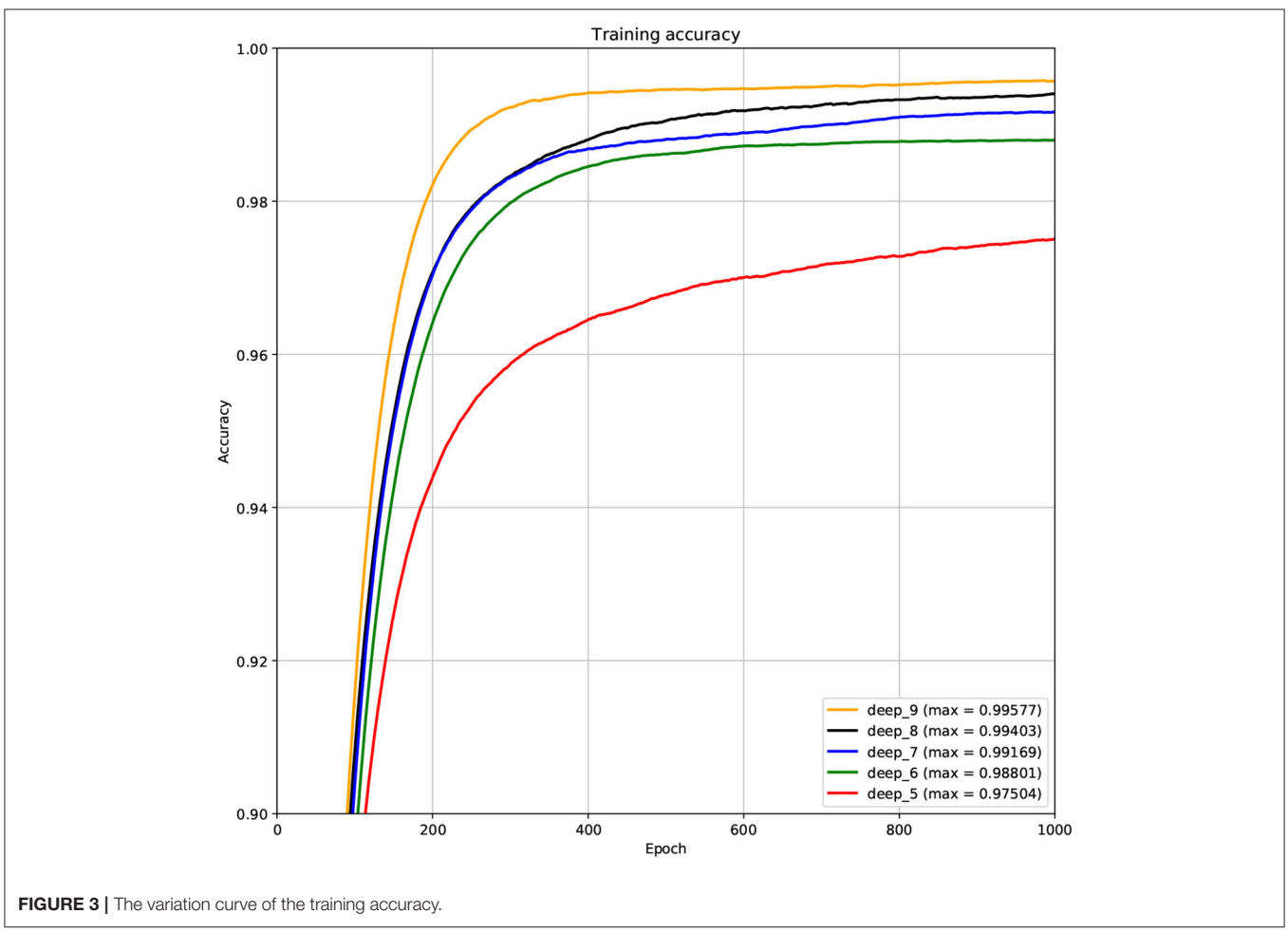

FIGURE 3 | The variation curve of the training accuracy.

5, 5, 7, 8, and 6 reduced amino acids, respectively. We counted the frequency of different tripeptides in the RAAA sequence to obtain the RAAA based tripeptide composition (RTC). The RTC used in this experiment is a 1-dimensional vector. Comparing with amino acid based tripeptide composition (TC), RTC further adds the information of the physicochemical property of the amino acid.

For a protein P, the PSSM of P was obtained by PSI-BLAST (Altschul et al., 1997). The database used in PSI-BLAST is Swissprot which can be download from ftp://ftp.ncbi.nih.gov/blast/db/FASTA/. The PSSM contains the evolutional information of a protein. It is a L*20 matrix, where L is the length of P. The matrix is shown as follows:

$$PSSM = \begin{bmatrix} a_{1,1} & \cdots & a_{1,L} \\ \vdots & \ddots & \vdots \\ a_{20,1} & \cdots & a_{20,L} \end{bmatrix}_{20\times L} \quad (1)$$

where $a_{i,j}$ represents the score that the i-th residue in P evolves into an amino acid j. We used the following formulas to convert PSSM into a 20-dimensional vector:

$$A_{i,j} = \frac{1}{1+e^{-a_{i,j}}} \quad (2)$$

$$PSSM - feature = \left\{ \frac{\sum_{i=1}^{L} A_{i,j}}{L} \mid j = 1, 2 \cdots 20 \right\} \quad (3)$$

We spliced the feature vectors of RTC and PSSM together to form a new feature called RTCP, and used RTCP as the input of the classifier.

2.3. Classifier

DeepRTCP uses SVM and DCNN as classifiers. SVM is used for choosing the optimal RTCP. DCNN uses the optimal RTCP as input to determine whether a protein is an ABC transporter. SVM is a powerful and effective machine learning algorithm. We used the RBF kernel SVM in this study. The penalty parameter and the gamma were set into 10^5 and "auto," respectively. DCNN is a heuristic algorithm that imitates the local receptive field of biological neurons. We used a 1-dimensional DCNN in this work. **Figure 2** shows the architecture of the DCNN, which consists of 5 convolutional layers, 5 batchnormalization layers, 2 max-pooling layers and 2 fully connected layers. The convolutional layer extracts important local information from the input features. The fully connected layer is equivalent to a classifier. It uses the information extracted by the convolutional layer to classify the input protein. If we have an input x of

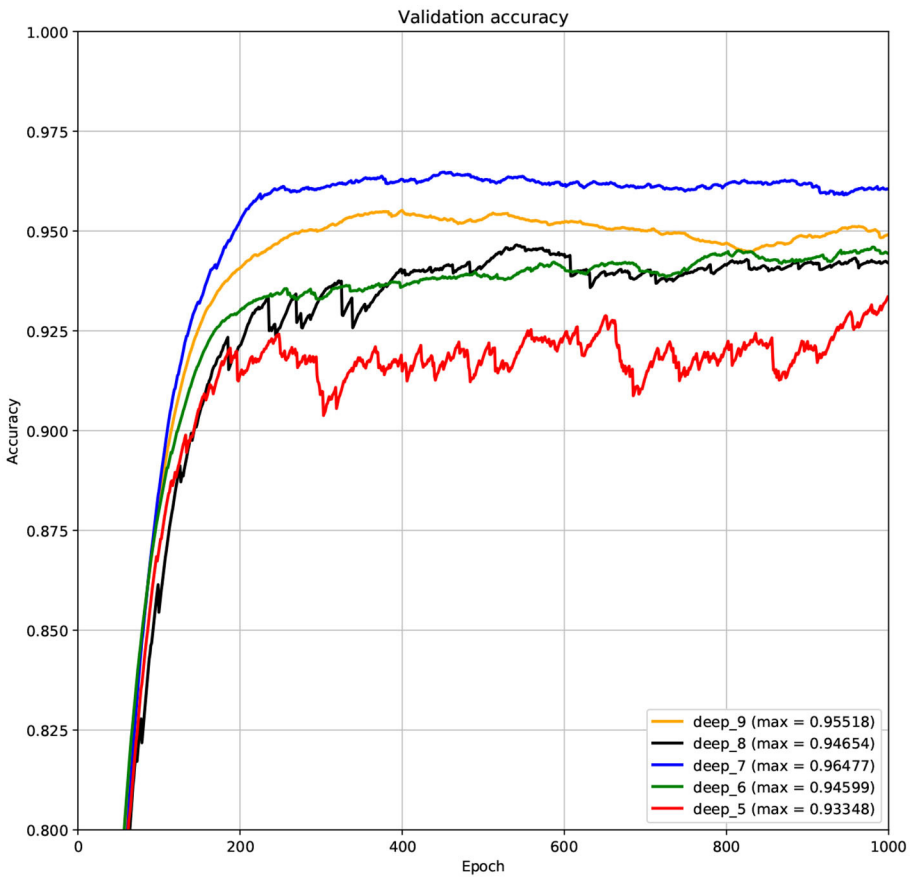

FIGURE 4 | The variation curve of the validation accuracy.

length L and a kernel function $f(x)$, the output of the convolution operation is defined as:

$$y = [a_l, a_2, a_3 \ldots\ldots, a_k], k = \frac{L-F}{S} \tag{4}$$

$$a_i = \sum_{j=1}^{F} f(j) \times x(i+j-1) \tag{5}$$

where F is the length of the filter, and S is the stride. The i-th value of the output vector is obtained by the convolution summation of the $x[i : i+F-1]$ and the convolution kernel $f(x)$.

In order to accelerate the convergence rate of the model, we inserted a batchnormalization layer after each convolutional layer. The batchnormalization layer ensures that the distribution of features in each batch will not change much. We also added max-pooling layers to the network to reduce the redundant information contained in the output of the convolutional layer.

2.4. Training

DeepRTCP takes the RTCP as input. The RTCP is a 532-dimensional vector. We used PCA (Belhumeur et al., 1997) to reduce the dimension of RTCP to 80 to optimize the performance of DeepRTCP. In order to fit the model better, we used a learning rate decay strategy. We set the initial learning rate to 10^{-3}, and then reduced the learning rate to one-tenth of the original value every 100 epochs. We had trained the model for 1,000 epochs, and the learning rate will continue to decrease until the end of training. DeepRTCP uses Relu as the activation function of hidden layers, Sigmoid as the activation function of the output layer, Adam as the optimizer and binary cross-entropy as the loss function. The strides of convolution and pooling are 2. This experiment used Tensorflow (Rampasek and Goldenberg, 2016) to build the deep learning model and Tensorboard to record the loss and the accuracy of the model. DeepRTCP ran on Nvidia RTX 2070(8G) GPU card. Training time of DeepRTCP was greatly reduced by using the CUDA (Nickolls et al., 2008) framework and the GPU card.

2.5. Evaluation Methods

This experiment uses some evaluation indicators that are widely used in protein function predictions, including accuracy (Acc), specificity (Spec), sensitivity (Sens), F-score, and Matthews' correlation coefficient (Mcc) (Matthews, 1975; Shan et al., 2019; Zhang et al., 2019). The formulas of these indicators are

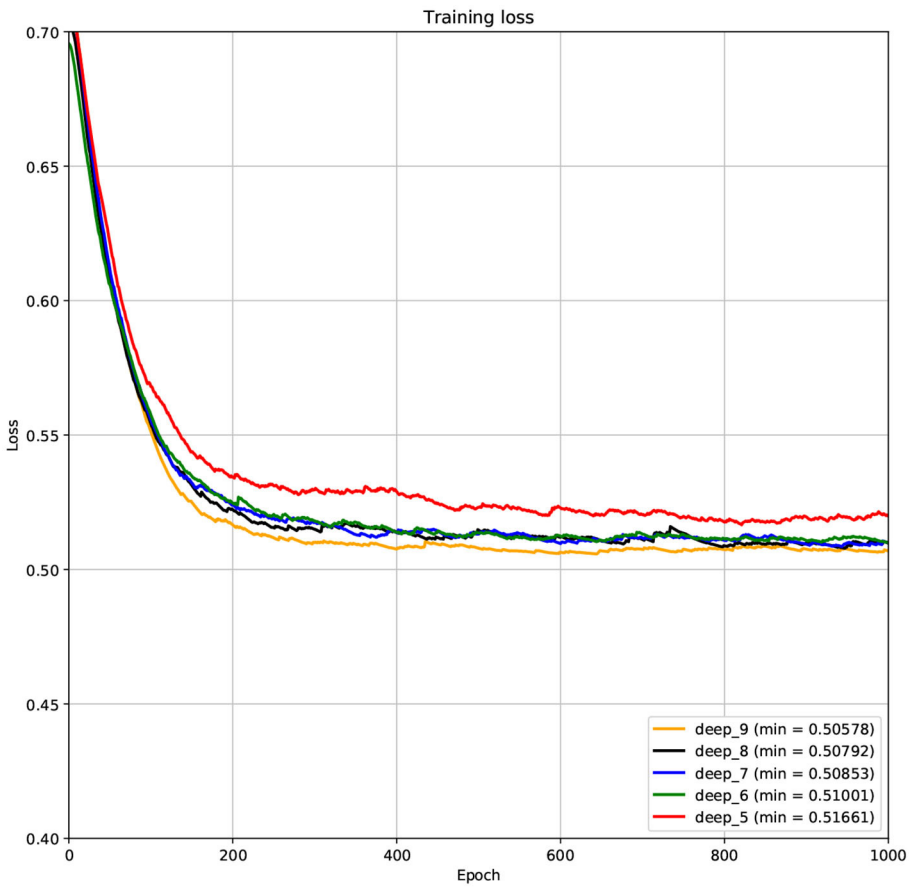

FIGURE 5 | The variation curve of the training loss.

as follows:

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{6}$$

$$Spec = \frac{TN}{TN + FP} \tag{7}$$

$$Sens = \frac{TP}{TP + FN} \tag{8}$$

$$Pre = \frac{TP}{TP + FP} \tag{9}$$

$$F - score = \left(1 + \beta^2\right) \frac{Pre \times Sens}{\beta^2 \times (Pre + Sens)} \tag{10}$$

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

where TP, FP, TN, and FN represent the rates of true positives, false positives, true negatives, and false negatives, respectively. In the formula of F-score, β measures the importance between Pre and Sens. We set β to 1 which means that Pre is as important as Sens.

We also used ROC curve (Hanley and Mcneil, 1982) and the area under the curve (AUC) to evaluate the performance of different classifiers. If the ROC curve of one method is covered by the ROC curve of another method, the latter is better. But the ROC curves of different methods are usually intersecting. So it is difficult to judge which method is better. So we need to compare AUCs of these methods. The larger the AUC, the better the method.

TABLE 5 | Performance comparison among different classifiers.

Classifier	Acc	Spec	Sens	F-score	Mcc
SVM	93.94%	92.45%	94.76%	0.9383	0.8810
NB	83.13%	79.52%	86.73%	0.8375	0.6643
ANN	91.69%	91.43%	91.94%	0.9172	0.8337
RF	90.97%	95.71%	86.26%	0.9055	0.8025
DCNN	95.96%	97.14%	94.81%	0.9593	0.9195

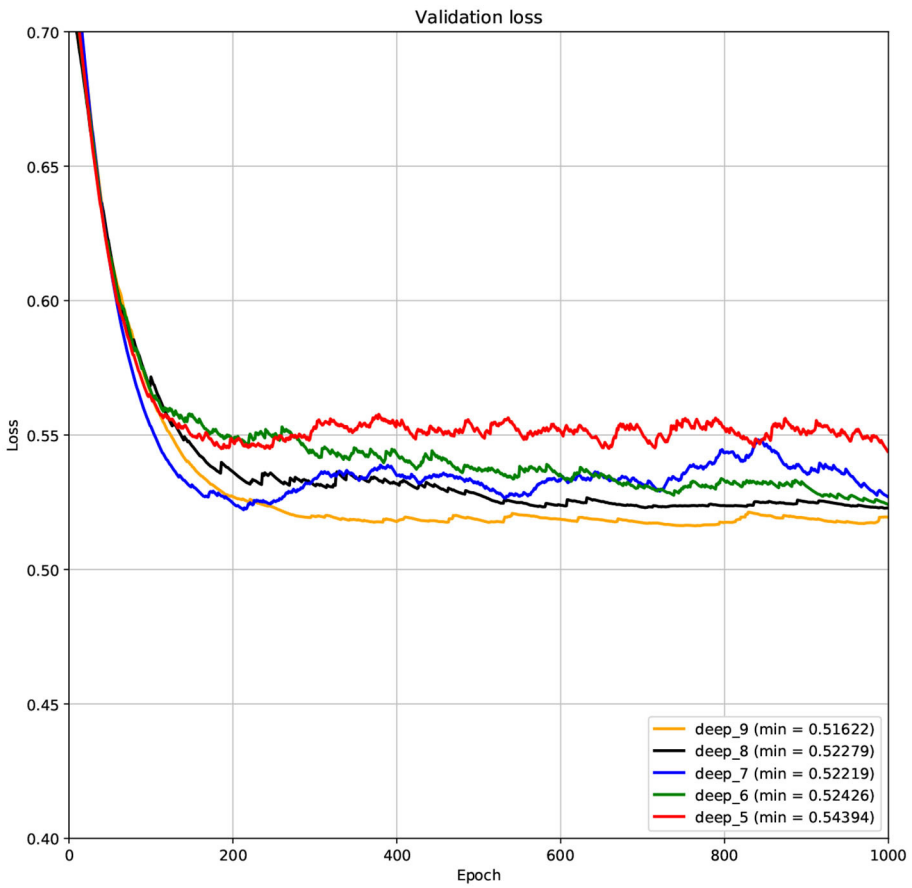

FIGURE 6 | The variation curve of the validation loss.

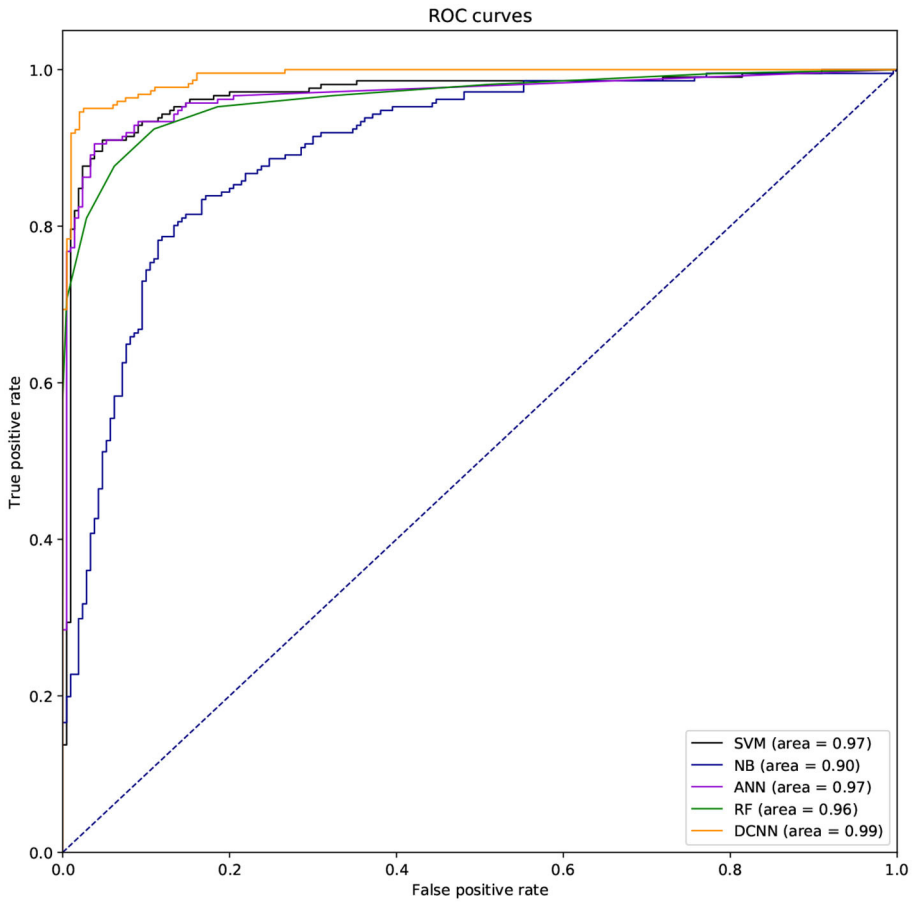

FIGURE 7 | The ROC curves of different classifiers.

3. RESULTS AND DISCUSSION

Firstly, we compared the performances of DeepRTCP when using different RTCPs. Secondly, we compared the RTCP based method with the methods based on TC and PSSM (TCP). Thirdly, we compared the performance among DCNNs with different structures and selected the optimal DCNN for predicting ABC transporters. Fourthly, we compared the performance among different classifiers. Fifthly, we analyzed the predicted false negatives and false positives. Finally, we compared DeepRTCP with the existing method.

3.1. Comparison Among Different Types of RTCPs

In order to select the most suitable features, we used SVM to test the performance among six types of RTCPs. The reason for choosing SVM is that it can get results fast. We used PCA to reduce the dimension of the RTCP. We tested the performance among RTCPs with different dimensions. The results were presented in **Supplementary Tables 1–6**. We compared the performance among different RTCPs. **Table 3** shows the comparison results. The ST_SS based RTCP achieved the best performance which mainly related to the nature of the ABC transporters. ABC transporters perform functions on both sides of the cell membrane, which may cause the difference in surface tension and solvent solubility with other proteins.

3.2. Comparison With the TCP Based Method

We used PCA to reduce the dimension of the TCP, and then compared its performance with the RTCP's. **Table 4** shows that even if PCA is used for dimension reduction, the performance of the TCP based method is still not as good as that of the RTCP based method. Comparing with TC, RTC contains information of the physicochemical property, which makes RTC more efficient than TC for predicting protein functions. The TCP based method achieved the best performance after using PCA to reduce the dimension to 500. This shows that by combining the physicochemical property, RTCP is not only better than TCP in performance, but also has lower dimensions than TCP. This makes the RTCP based method faster and less expensive than TCP based methods.

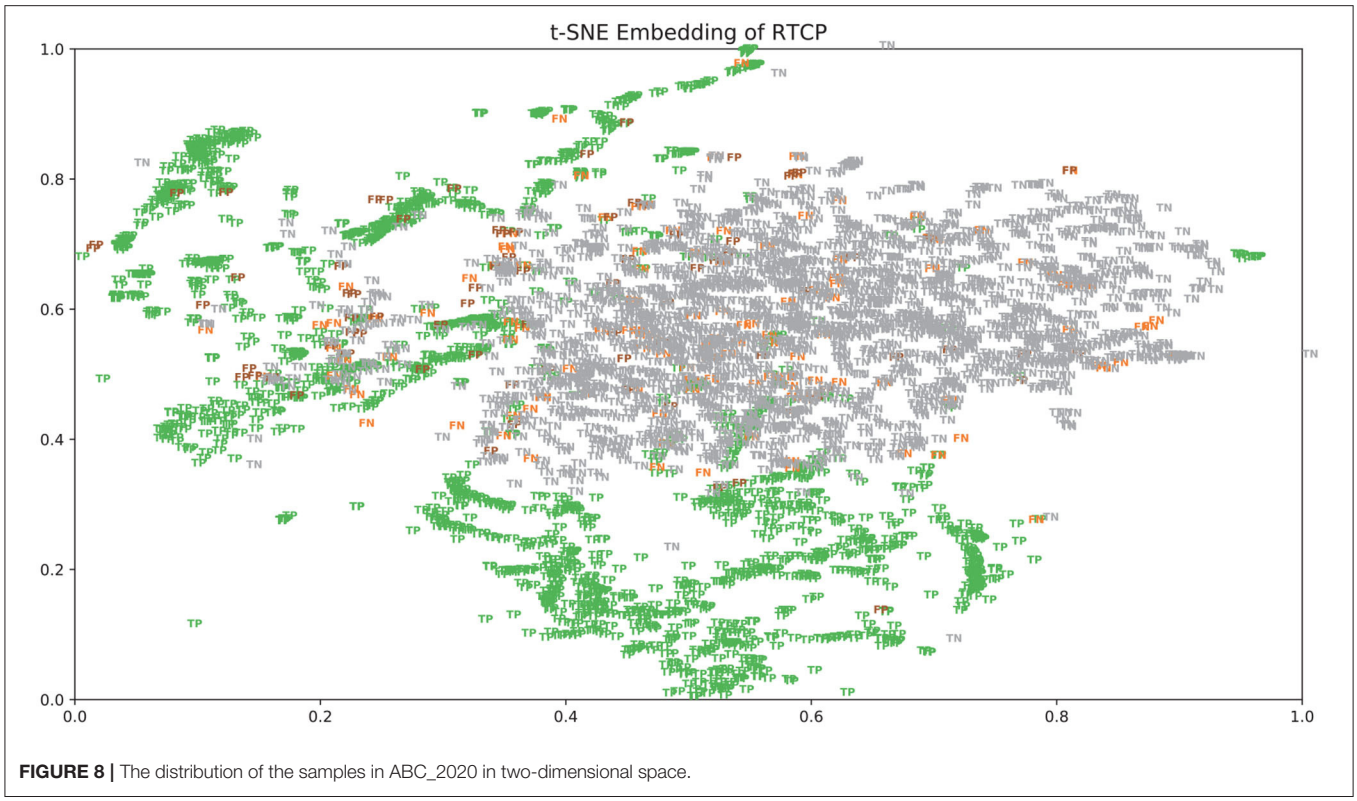

FIGURE 8 | The distribution of the samples in ABC_2020 in two-dimensional space.

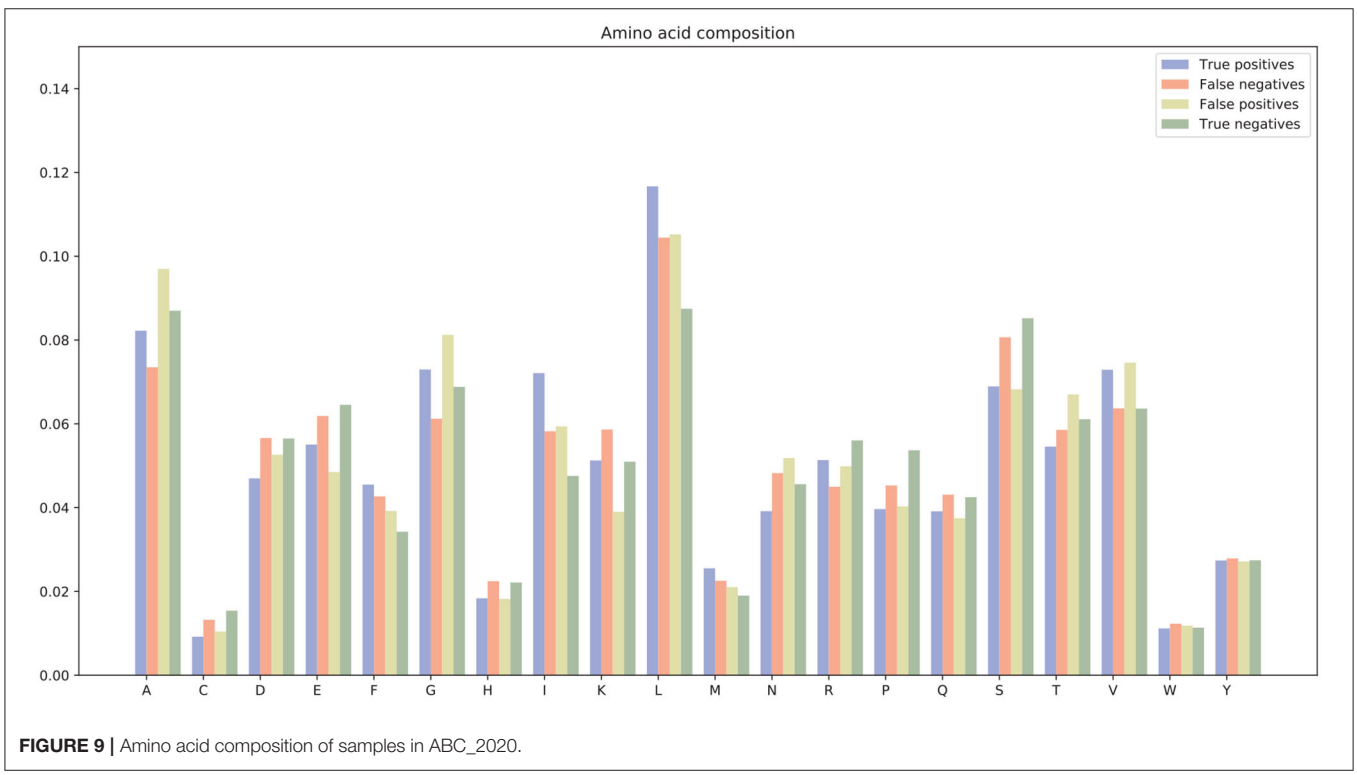

FIGURE 9 | Amino acid composition of samples in ABC_2020.

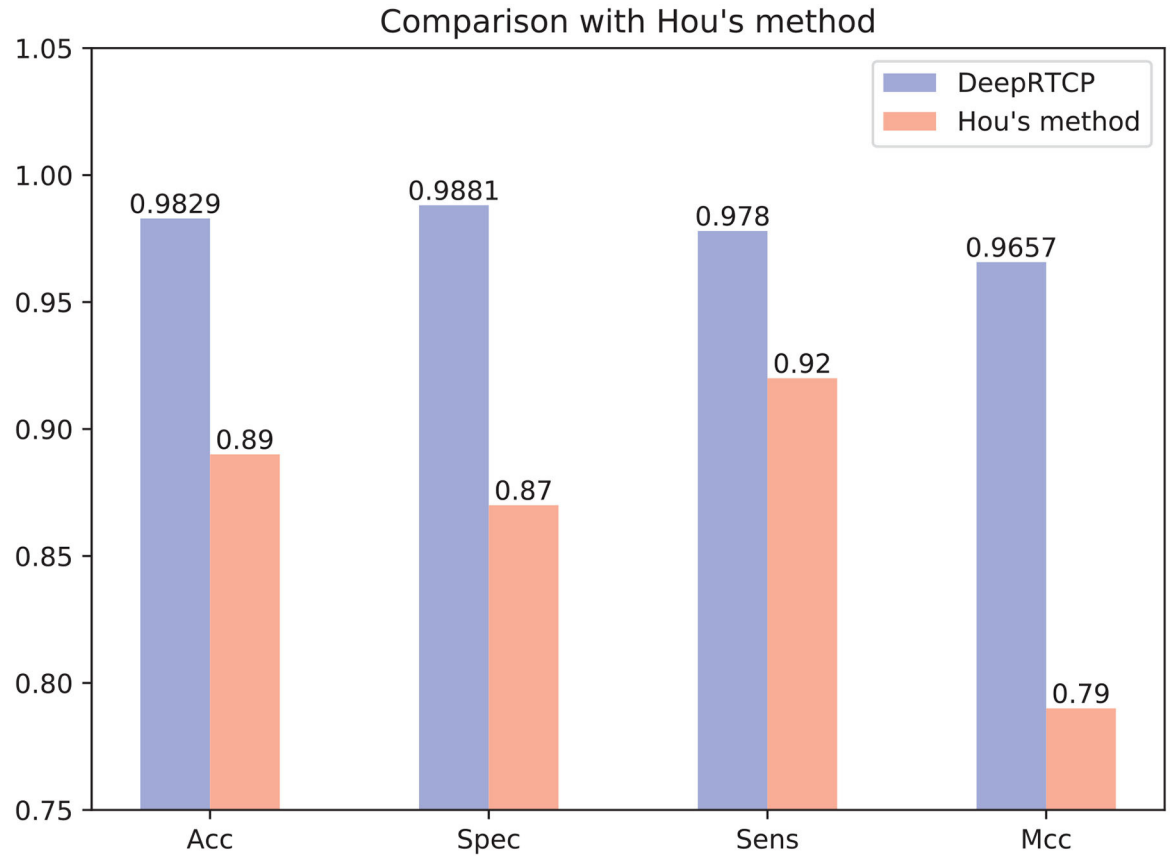

FIGURE 10 | Comparison between DeepRTCP and other method.

3.3. Comparison Among DCNNs With Different Structures

The structure of DCNN has a great influence on its performance. We compared DCNNs with different depths. The layer numbers of these DCNNs are 5, 6, 7, 8, and 9, respectively. The filter number of these DCNNs is 64. From **Figures 3**, **4**, we can see that the 7-layers DCNN achieve the highest accuracy on the validation set. Comparing with other DCNNs, the 7-layers DCNN has the smallest difference between the training accuracy and the validation accuracy. Therefore, the 7-layers DCNN is more suitable for predicting ABC transporters than DCNNs with other depth. From **Figures 5**, **6**, we can see that using RTCP as the feature, these DCNNs fit around the 400th epoch. The training loss and the validation loss of each DCNN are similar, which indicates the robustness of RTCP. We also compared the performance of 7-layers DCNNs with different filter numbers of 8, 16, 24, 32, 40, 48, 56, 64, 72, and 80, respectively. The result was presented in **Supplementary Table 7**. After the filters number is greater than 32, the accuracy of the model no longer changes significantly. The 7-layers-32-filters DCNN and the 7-layers-64-filters DCNN had achieved the best validation accuracy. Comparing with the 7-layers-64-filters DCNN, the 7-layers-32-filters DCNN is simpler and less computationally expensive. So the 7-layers-32-filters DCNN was selected as the classifier for this experiment.

3.4. Comparison Among Methods Based on Different Classifiers

We used different classifiers to predict ABC transporters, including SVM, NB, ANN, RF, and DCNN. We used 10-fold cross-validation to evaluate the performance of these classifiers. **Table 5** shows the results of the test. Except for NB, the validation accuracies of other classifiers exceeds 90%, and the validation accuracy of DCNN is as high as 95.96%. ANN has the minimum difference of 0.54% between Spec and Sens. DCNN achieved the best F-score and Mcc of 0.9593 and 0.9125, respectively. Then we analyzed the ROC curves of these methods (as shown in **Figure 7**). It is worth noting that the AUC of DCNN is as high as 0.99. The true positive rate of DCNN is much higher than that of other classifiers when the false positive rate is 0. These results show that DCNN is more suitable for the identification of ABC transporters than other classifiers.

3.5. Analysis of TP and FP

There were many false positives and false negatives in the prediction results of DeepRTCP. We tried to analyze the reasons

why these data were predicted incorrectly. We trained 100 models for each cross-validation, and selected the proteins that were predicted incorrectly more than 50 times as false negatives and false positives. The false negatives and false positives can be obtained from https://github.com/zhichunlizzx/DeepRTCP. We used t-Distributed Stochastic Neighbor Embedding (T-SNE) (Shao et al., 2018) to map the features of the samples in the dataset to two-dimensional space. We founded that in the two-dimensional space, false positives (false negatives) were distributed in the area where the positives were clustered (as shown in **Figure 8**). We counted the amino acid composition of the samples in ABC_2020 (as shown in **Figure 9**). We found that the content of histidine, glutamine, and valine in positives (negatives) and false positives (false negatives) are similar, but there are significant differences in the content of histidine, glutamine, and valine in positives (negatives) and false negatives (false positives). This may result in a positive (negative) being incorrectly predicted as a false negative (false positive).

3.6. Comparison With Other Method

In the past study, Hou et al. (2020) used RF and a feature of 188 dimension to predict ABC transporters. We downloaded the dataset provided by Hou, which included 875 positives and 875 negatives. We performed 10-fold cross-validation on DeepRTCP on this dataset and compared the results with Hou's. Since the number of samples in Hou's dataset is smaller than that of ABC_2020, we use a simple DCNN on this dataset. The DCNN includes 4 convolutional layers and 2 fully connected layers. The filter numbers in the four convolutional layers are 16, 16, 32, and 32, respectively. The two fully connected layers contain 13 and 2 neurons, respectively. **Figure 10** shows that the average validation accuracy of DeepRTCP is as high as 98.29%. Compared with Hou's method, DeepRTCP improved the Acc by 9.29%, Spec by 12.81%, Sens by 5.8%, and Mcc by 0.1757. In addition, the difference between the Spec and Sens of Hou's method is as high as 5%, which makes Hou's method not well applied in practices. The difference between Spec and Sens of DeepRTCP is about 1%, which is a great improvement to Hou's method. This is mainly due to the effective classifier and feature.

4. CONCLUSION

In this study, we propose a novel method for ABC transporter prediction called DeepRTCP. It uses the DCNN as the classifier. The classifier uses a feature named RTCP which composed of TCP and PSSM. We tested the performance of six types of RTCPs. The results show that the ST_SS based RTCP has the best performance. In the comparison of different classifiers, DCNN achieved the best results that Acc, Spec, Sens, F-score and Mcc were 95.96%, 97.14%, 94.81%, 0.9593 and 0.9195, respectively. Compared with the state-of-the-art method, DeepRTCP improved Acc by 9.29%,Spec by 11.81%, Sens by 5.8%, and Mcc by 0.1757. DeepRTCP can label the ABC transporters faster than traditional biological experiments, and the accuracy of DeepRTCP is also high. DeepRTCP provides a reliable guide for the further research of ABC transporters.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

ZZ and JW conceived and designed the project. ZZ and JL performed the experiments. ZZ, JL, and JW wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the National Natural Science Foundations of China (62002181, 62061035, and 61661040) and the Inner Mongolia Science & Technology Plan (2020GG0186).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcell.2020.614080/full#supplementary-material

REFERENCES

Abbas, M., Horler, R. S. P., Axel, M., Wilkinson, A. J., and Thomas, G. H. (2015). The substrate-binding protein in bacterial ABC transporters: dissecting roles in the evolution of substrate specificity. *Biochem. Soc. Trans.* 43, 1011–1017. doi: 10.1042/BST20150135

Altschul, S. F., Madden, T. L., Schffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Amos, B., Lydie, B., Severine, A., Valeria, A., and Zhang, J. (2009). The universal protein resource (uniprot). *Nucleic Acids Res.* 37, 169–174. doi: 10.1093/nar/gkn664

Belhumeur, P. N., Hespanha, J. P., and Kriegman, D. J. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 711–720. doi: 10.1109/34.598228

Cai, C. Z., Han, L. Y., Ji, Z. L., Chen, X., and Chen, Y. Z. (2003). Svm-prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.* 31, 3692–3697. doi: 10.1093/nar/gkg600

Chen, J., Lu, G., Lin, J., Davidson, A. L., and Quiocho, F. A. (2003). A tweezers-like motion of the ATP-binding cassette dimer in an ABC transport cycle. *Mol. Cell* 12, 651–661. doi: 10.1016/j.molcel.2003.08.004

Chen, L., Zou, Y., Qin, J., Liu, X., Jiang, Y., Ke, C., et al. (2013). Hierarchical classification of protein folds using a novel ensemble classifier. *PLoS ONE* 8:e56499. doi: 10.1371/journal.pone.0056499

Chen, W., Liu, X., Huang, Y., Jiang, Y., and Lin, C. (2012). Improved method for predicting protein fold patterns with ensemble classifiers. *Genet. Mol. Res.* 11, 174–181. doi: 10.4238/2012.January.27.4

Cui, J., and Davidson, A. L. (2011). Abc solute importers in bacteria. *Essays Biochem.* 50, 85–99. doi: 10.1042/bse0500085

Davidson, A., Dassa, E., Orelle, C., and Chen, J. (2008). Structure, function, and evolution of bacterial ATP-binding cassette systems. *Microbiol. Mol. Biol. Rev.* 72, 317–364. doi: 10.1128/MMBR.00031-07

Finn, R. D., Alex, B., Jody, C., Penelope, C., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, 222–230. doi: 10.1093/nar/gkt1223

Gao, X., Wang, D., Zhang, J., Liao, Q., and Liu, B. (2019). IRBP-Motif-PSSM: identification of RNA-binding proteins based on collaborative learning. *IEEE Access* 7, 168956–168962. doi: 10.1109/ACCESS.2019.2952621

Gedeon, C., Behravan, J., Koren, G., and Micheline, P. M. (2006). Transport of glyburide by placental ABC transporters: implications in fetal drug exposure. *Placenta* 27, 1096–1102. doi: 10.1016/j.placenta.2005.11.012

Gligorijevic, V., Barot, M., and Bonneau, R. (2018). Deepnf: deep network fusion for protein function prediction. *Bioinformatics* 34, 3873–3881. doi: 10.1093/bioinformatics/bty440

Hanley, J. A., and Mcneil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 143, 29–37. doi: 10.1148/radiology.143.1.7063747

Haretsugu, H., Kenta, N., Toshihide, O., Akira, T., and Toshihisa, T. (2010). Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast* 18, 523–531. doi: 10.1002/yea.706

Hou, R., Wang, L., and Wu, Y. J. (2020). Predicting ATP-binding cassette transporters using the random forest method. *Front. Genet.* 11, 156–167. doi: 10.3389/fgene.2020.00156

Jiang, J. Q. (2012). Predicting protein function by multi-label correlated semi-supervised learning. *IEEE ACM Trans. Comput. Biol. Bioinform.* 9, 1059–1069. doi: 10.1109/TCBB.2011.156

Konc, J., HodoEk, M., Ogrizek, M., Trykowska, K. J., JaneI, D., and Mackerell, A. D. (2013). Structure-based function prediction of uncharacterized protein using binding sites comparison. *PLoS Comput. Biol.* 9:e1003341. doi: 10.1371/journal.pcbi.1003341

Kulmanov, M., and Robert, H. (2019). Deepgoplus: improved protein function prediction from sequence. *Bioinformatics* 36, 422–429. doi: 10.1093/bioinformatics/btz595

Le, N. Q. K., Huynh, T. T., Yapp, E. K. Y., and Yeh, H. Y. (2019). Identification of clathrin proteins by incorporating hyperparameter optimization in deep learning and pssm profiles. *Comput. Methods Programs Biomed.* 177, 81–88. doi: 10.1016/j.cmpb.2019.05.016

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–445. doi: 10.1038/nature14539

Lecun, Y., and Bottou, L. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Libbrecht, M. W., and Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* 16, 321–332. doi: 10.1038/nrg3920

Lin, H., Feng, P., Chen, W., and Zuo, Y. (2017). Predicting the types of j-proteins using clustered amino acids. *BioMed. Res. Int.* 2014, 1–8. doi: 10.1155/2014/935719

Luo, R. Y., Feng, Z. P., and Liu, J. K. (2010). Prediction of protein structural class by amino acid and polypeptide composition. *Eur. J. Biochem.* 269, 4219–4225. doi: 10.1046/j.1432-1033.2002.03115.x

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Biophys. Acta* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Michael, G., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.* 84, 4355–4358. doi: 10.1073/pnas.84.13.4355

Mundra, P., Kumar, M., Kumar, K. K., Jayaraman, V. K., and Kulkarni, B. D. (2007). Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. *Pattern Recogn. Lett.* 28, 1610–1615. doi: 10.1016/j.patrec.2007.04.001

Nickolls, J., Buck, I., Garland, M., and Skadron, K. (2008). Scalable parallel programming with CUDA. *Queue* 6, 40–53. doi: 10.1145/1401132.1401152

Rampasek, L., and Goldenberg, A. (2016). Tensorflow: biology's gateway to deep learning. *Cell Syst.* 2, 12–14. doi: 10.1016/j.cels.2016.01.009

Rish, I. (2001). An empirical study of the naive bayes classifier. *J. Univ. Comput. Sci.* 1, 41–46. doi: 10.1039/b104835j

Shan, X., Wang, X., Li, C. D., Chu, Y., and Wei, D. Q. (2019). Prediction of cyp450 enzyme-substrate selectivity based on the network-based label space division method. *J. Chem. Inform. Model.* 59, 4577–4586. doi: 10.1021/acs.jcim.9b00749

Shao, L., Gao, H., Liu, Z., Feng, J., Tang, L., and Lin, H. (2018). Identification of antioxidant proteins with deep learning from sequence information. *Front. Pharmacol.* 9:1036. doi: 10.3389/fphar.2018.01036

Song, L., Li, D., Zeng, X., Wu, Y., Guo, L., and Zou, Q. (2014). Ndna-prot: identification of dna-binding proteins based on unbalanced classification. *BMC Bioinformatics* 15, 298–308. doi: 10.1186/1471-2105-15-298

Suykens, J. A. K., and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Process. Lett.* 9, 293–300. doi: 10.1023/A:1018628609742

Vladimir, S., Andy, L., Christopher, T. J., Christopher, C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and qsar modeling. *J. Chem. Inform. Model.* 43, 1947–1958. doi: 10.1021/ci034160g

Wang, L., You, Z. H., Chen, X., Yan, X., Liu, G., and Zhang, W. (2018). RFDT: a rotation forest-based predictor for predicting drug-target interactions using drug structure and protein sequence information. *Curr. Protein Pept. Sci.* 19, 445–454. doi: 10.2174/1389203718666161114111656

Wang, S., Li, M., Guo, L., Cao, Z., and Fei, Y. (2019). Efficient utilization on pssm combining with recurrent neural network for membrane protein types prediction. *Comput. Biol. Chem.* 81, 9–15. doi: 10.1016/j.compbiolchem.2019.107094

You, R., Zhang, Z., Xiong, Y., Sun, F., and Mamitsuka, H. (2018). Golabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics* 34, 2465–2473. doi: 10.1101/145763

Zhang, L., Kong, L., Han, X., and Lv, J. (2016). Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure. *J. Theor. Biol.* 400, 1–10. doi: 10.1016/j.jtbi.2016.04.011

Zhang, W., Jing, K., Huang, F., Chen, Y., Li, B., Li, J., and Gong, J. (2019). Sflln: a sparse feature learning ensemble method with linear neighborhood regularization for predicting drug-drug interactions. *Inform. Sci.* 497, 189–201. doi: 10.1016/j.ins.2019.05.017

Zhang, Y., and Yu, D. (2019). Protein-atp binding site prediction based on 1d-convolutional neural network. *J. Comput. Appl.* 39, 3146–3150.

Zou, Q., Wan, S., Ju, Y., Tang, J., and Zeng, X. (2016). Pretata: predicting tata binding proteins with novel features and dimensionality reduction strategy. *Bmc Syst. Biol.* 10, 114–116. doi: 10.1186/s12918-016-0353-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.