frontiers | Frontiers in Genetics

ORIGINAL RESEARCH
published: 24 May 2022
doi: 10.3389/fgene.2022.876721

Interpretable Deep Learning Model Reveals Subsequences of Various Functions for Long Non-Coding RNA Identification

Rattaphon Lin[1] *and Duangdao Wichadakul*[1,2]*

[1]Department of Computer Engineering, Faculty of Engineering, Chulalongkorn University, Pathumwan, Thailand, [2]Center of Excellence in Systems Biology, Faculty of Medicine, Chulalongkorn University, Pathumwan, Thailand

Long non-coding RNAs (lncRNAs) play crucial roles in many biological processes and are implicated in several diseases. With the next-generation sequencing technologies, substantial unannotated transcripts have been discovered. Classifying unannotated transcripts using biological experiments are more time-consuming and expensive than computational approaches. Several tools are available for identifying long non-coding RNAs. These tools, however, did not explain the features in their tools that contributed to the prediction results. Here, we present Xlnc1DCNN, a tool for distinguishing long non-coding RNAs (lncRNAs) from protein-coding transcripts (PCTs) using a one-dimensional convolutional neural network with prediction explanations. The evaluation results of the human test set showed that Xlnc1DCNN outperformed other state-of-the-art tools in terms of accuracy and F1-score. The explanation results revealed that lncRNA transcripts were mainly identified as sequences with no conserved regions, short patterns with unknown functions, or only regions of transmembrane helices while protein-coding transcripts were mostly classified by conserved protein domains or families. The explanation results also conveyed the probably inconsistent annotations among the public databases, lncRNA transcripts which contain protein domains, protein families, or intrinsically disordered regions (IDRs). Xlnc1DCNN is freely available at https://github.com/cucpbioinfo/Xlnc1DCNN.

Keywords: long non-coding RNA (lncRNA), one-dimensional convolutional neural network (1D CNN), deep learning, explainable artificial intelligence (XAI), SHAP (SHapley additive exPlanations)

Edited by:
Sarath Chandra Janga,
Indiana University, Purdue University Indianapolis, United States

Reviewed by:
Doaa Salem,
Indiana University, Purdue University Indianapolis, United States
Tsukasa Fukunaga,
Waseda University, Japan

***Correspondence:**
Duangdao Wichadakul
duangdao.w@chula.ac.th

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *15 February 2022*
Accepted: *11 April 2022*
Published: *24 May 2022*

Citation:
Lin R and Wichadakul D (2022) Interpretable Deep Learning Model Reveals Subsequences of Various Functions for Long Non-Coding RNA Identification.
Front. Genet. 13:876721.
doi: 10.3389/fgene.2022.876721

1 INTRODUCTION

Long non-coding RNAs (lncRNAs) are RNAs that are not translated into proteins and are longer than 200 nucleotides. lncRNAs play important roles in many critical biological processes, including gene expression, gene regulation, gene silencing, chromatin remodeling, acting as molecular scaffolds, etc. (Rinn and Chang, 2012; Marchese et al., 2017; Statello et al., 2021), and have been implicated in human diseases such as cancers and diabetes (Morán et al., 2012; Fang and Fullwood, 2016; Chan and Tay, 2018; Jin et al., 2020). The enhancements of next-generation sequencing technology, i.e., RNA sequencing (RNA-Seq) (Wang et al., 2009; Stark et al., 2019) have led to numerous discoveries of unannotated transcripts. However, classifying the innumerable number of unclassified sequences using experimental approaches is time-consuming and expensive. In contrast, computational approaches are faster and more convenient.

Most of the existing computational approaches for classifying lncRNA and protein-coding transcripts used feature extraction methods to obtain training features, e.g., the upgraded version of Coding Potential Calculator (CPC2) (Kang et al., 2017), CNIT (Guo et al., 2019), PLEK (Li et al., 2014), CPAT (Wang et al., 2013), FEELnc (Wucher et al., 2017), RNAsamba (Camargo et al., 2020), LncADeep (Yang et al., 2018), and lncRNA_Mdeep (Fan et al., 2020). Most of them used similar features such as the Fickett and hexamer scores, the ORF length, and then topped up with additional sequence and structural features. Moreover, none of them explained how the features contributed to the model prediction results.

Deep learning algorithms have become very popular, especially for a dataset with a large number of data points and data dimensions as the features will be learned by the algorithms themselves during the training. Many convolutional neural networks (CNNs), the 2D-CNNs, have been widely used for image classification and segmentation applications (Yamashita et al., 2018) because of their great capability for extracting features from input data. Recently, many applications such as speech recognition and ECG monitoring (Kiranyaz et al., 2021) started using 1D-CNN instead of the traditional machine learning approaches. The applications for detecting irregular heartbeats (Acharya et al., 2017; Li et al., 2019; Hsieh et al., 2020) have shown that using only a simple 1D-CNN could achieve high prediction accuracy without explicitly addressing and extracting features as inputs for the models.

While most complex black-box models (e.g., boosting tree algorithms, ensemble models, deep neural networks) typically provide better learning performance, they usually are uninterpretable. To understand how a complex model learns to differentiate things, explainable artificial intelligence (XAI) has recently become one of the popular topics aiming to interpret and explain machine learning or deep learning models (Tjoa and Guan, 2021). Explainable AI is essential for users to understand and trust the model prediction results. It can help illustrate what the models perceive and explain how these perceptions can be mapped with the underlying knowledge of the human. Some of the favored approaches to obtain an explanation from a complex black-box model are LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). LIME builds a local surrogate model to explain individual prediction. SHAP (Shapley Additive exPlanations) introduced SHAP values representing the unified measure of feature importance together with SHAP value estimation methods. DeepSHAP (Chen et al., 2019) was built based on the connection between the original SHAP and DeepLIFT (Shrikumar et al., 2017) to explain the deep learning model and further refined and extended with relative background distributions and stacks of mixed model types.

With still some ambiguities in classifying lncRNA and mRNA sequences based on training features, together with the promising results of 1D-CNN in previous applications, in this paper, we propose Xlnc1DCNN, a 1D-CNN model for classifying lncRNA and mRNA with an explanation. The model solely uses nucleotide sequences as the training set. On the human test set, Xlnc1DCNN outperformed all other models in terms of accuracy and F1-score. For the cross-species dataset, Xlnc1DCNN also had the generalization across testing species. We explained how the Xlnc1DCNN distinguished the lncRNA from mRNA transcript sequences by applying DeepSHAP to generate SHAP values representing how the model captured and visualized the contribution of each nucleotide using an in-house python code. The explanation of true positives (i.e., lncRNA transcript sequences) showed that the model classified a sequence as lncRNA if the sequence did not contain any important regions or contained only an N-terminal signal peptide or transmembrane helices. The explanation of true negatives (i.e., mRNA transcript sequences) showed that the model learned protein domains/families from the input transcript sequences and used them to predict the sequences as mRNAs. The explanation of false positives (i.e., mRNA predicted as lncRNA transcript sequences) showed that the model could not capture any important regions representing protein domains/families or found important regions contributing to both lncRNA and mRNA prediction. A few false positive sequences were also found with inconsistent transcript types among the databases. Lastly, the explanation of false negatives (i.e., lncRNA predicted as mRNA transcript sequences) showed that the model captured protein domains or families within these lncRNA sequences and, hence, misclassified them as mRNAs.

2 MATERIALS AND METHODS

2.1 Data Compilation and Pre-Processing

The human transcript datasets for training the model were obtained from GENCODE (Frankish et al., 2018) and LNCipedia (Volders et al., 2018). GENCODE (release 32) contains 48,351 sequences of lncRNA transcripts and 100,291 sequences of protein-coding transcripts (PCTs). For LNCipedia (version 5.2), only high confidence sequences were selected, which resulted in 107,039 lncRNA transcripts. To remove lncRNA transcript sequences from LNCipedia that are duplicates of GENCODE, we used CD-HIT-EST-2D (Li and Godzik, 2006) to compare lncRNA sequences between LNCipedia and GENCODE and filter out the sequences with more than 95% similarity from the LNCipedia dataset. A total of 72,803 lncRNA sequences from LNCipedia remained. We then pre-processed the sequences used for training the Xlnc1DCNN model by discarding the sequences shorter than 200 bases and longer than 3,000 bases. After filtering, one-hot encoding was used to encode the sequences. The total number of remaining sequences after cleansing was 185,030 with 108,578 lncRNAs and 76,453 PCTs (**Table 1**). The lncRNAs and PCTS were set as the positive and negative classes, respectively. The dataset was stratified split by 80% and 20% into the training and test sets.

Cross-species datasets included the mouse dataset obtained from GENCODE (Frankish et al., 2018) (release M23) and the gorilla, chicken, and cow datasets obtained from Ensembl (Cunningham et al., 2019) (release 102). We pre-processed the cross-species datasets by discarding the sequences shorter than

TABLE 1 | Summary of datasets from GENCODE and LNCipedia.

Sequence Type	Species	Data Source	Dataset Size	<200 bps	>3,000 bps	No.of Transcripts after Cleansing
mRNA	Human	GENCODE (release 32)	100,291	374	23,464	76,453
lncRNA	Human	GENCODE (release 32)	48,351	291	3,486	44,574
lncRNA	Human	LNCipedia (version 5.2)	72,803	0	8,799	64,004

TABLE 2 | Hyperparameters of the proposed 1D-CNN architecture.

Layer	Hyperparameter
Conv 1D	kernel size = 57, stride = 1
Max-Pooling	pool size = 2
Dropout	p = 0.3
Conv 1D	kernel size = 57, stride = 1
Max-Pooling	pool size = 2
Dropout	p = 0.3
Conv 1D	kernel size = 57, stride = 1
Max-Pooling	pool size = 2
Dropout	p = 0.3
Flatten	-
Dense	256
Dropout	p = 0.5
Dense	256
Dropout	p = 0.5
Softmax	2

200 bases and longer than 3,000 bases. We then randomly selected the mRNA and lncRNA sequences for each species. The test transcripts of gorilla, chicken, cow, and mouse contained 8,000, 8,000, 11,000, and 32,000 sequences, respectively, each with an equal number of sequences from each class.

2.2 Model Architecture

In this study, we designed and implemented the Xlnc1DCNN model in Python3 using TensorFlow on NVIDIA GeForce GTX 1080 Ti and Intel Xeon Silver 4112 Processor. The built model could distinguish lncRNAs from the mRNAs (PCTs) and outperformed the existing tools for the human dataset. The model architecture consists of three convolutions with pooling layers, two fully connected layers, and a Softmax layer. We used ReLU as the activation function for convolution and fully connected layers. We also found that adding the dropout layer after the pooling layer made the model perform slightly better.

We used 10% of the data from the training set to perform hyperparameter optimizations over the kernel size, dropout rate, stride size, batch size, and learning rate by using the grid search algorithm. The best kernel size was 57, with the stride size equal to 1. The model performance started to decrease after increasing the stride size for almost every kernel size. For the learning details, the momentum, learning rate, number of epochs, and batch size were 0.9, 0.01, 120, and 128, respectively, with the stochastic gradient descent as an optimizer. The final hypermeters used in the model architecture are shown in **Table 2**.

2.3 Model Interpretation

DeepSHAP was used to interpret how the proposed Xlnc1DCNN model could classify the lncRNAs and mRNAs from the input transcript sequences. As DeepSHAP needs background distributions as references to approximate the SHAPley values on conditional expectation, 175 sequences from each class were randomly selected as the representative background. A total of 350 sequences were used as the backgrounds as it was limited by the available GPU.

The output from DeepSHAP is SHAP values representing each nucleotide's contribution to the model. To obtain SHAP values representing each nucleotide within a sequence, we summed up SHAP values inside the array of one-hot encoding and got a single SHAP value of each nucleotide. To visualize SHAP values from DeepSHAP of the input transcript sequence, we further summed up the SHAP values of three consecutive nucleotides, which probably represented an amino acid, and generated the results in three reading frames. We then plotted a color line for each representative amino acid. The blue and red colors, respectively, indicate the contribution of each amino acid for classifying the sequence as an lncRNA and an mRNA (**Figure 1**).

2.4 Evaluation

2.4.1 Model Evaluation Metrics

To evaluate the performance of the proposed Xlnc1DCNN model with other existing tools, we used the following metrics. True positive (TP) represents the lncRNA transcript sequences that are predicted as lncRNAs. True negative (TN) represents PCTs that are predicted as PCTs. False positive (FP) represents the PCTs that are predicted as lncRNAs. False negative (FN) represents lncRNAs that are predicted as PCTs.

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

$$\boldsymbol{Sensitivity = \frac{TP}{TP + FN}}$$

$$\boldsymbol{Specificity = \frac{TN}{TN + FP}}$$

$$\boldsymbol{Precision = \frac{TP}{TP + FP}}$$

$$\boldsymbol{F1 - Score = \frac{2 \times precision \times sensitivity}{precision + sensitivity}}$$

2.4.2 Interpretation Evaluation Method

To compare the explanation results of Xlnc1DCNN on the human test set with known biological knowledge, we utilized the available bioinformatics tools/databases such as TMHMM

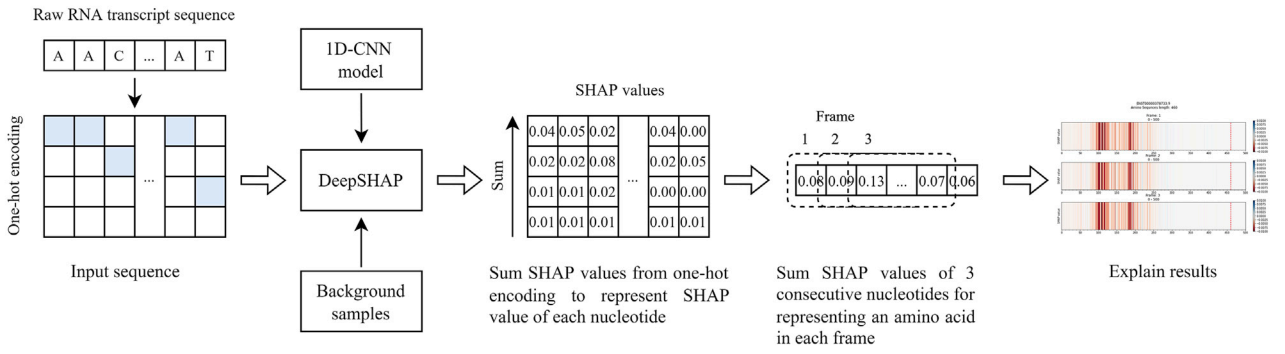


FIGURE 1 | The process to obtain SHAP values for explaining the nucleotide contribution that was captured by the model to differentiate lncRNA from mRNA transcript sequences.

TABLE 3 | Evaluation results of all tools on the human test set.

Model	TP	FP	TN	FN	Accuracy	Sensitivity	Specificity	Precision	F1
Xlnc1DCNN	20,895	1,204	14,087	821	**94.53**	96.22	92.13	94.55	**95.38**
CPC2	21,023	6,457	8,834	693	80.68	96.81	57.77	76.50	85.47
CNIT	21,307	3,580	11,711	409	89.22	**98.12**	76.59	85.61	91.44
PLEK	20,704	6,665	8,626	1,012	79.26	95.34	56.41	75.65	84.36
CPAT	20,646	2,597	12,694	1,070	90.09	95.07	83.02	88.83	91.84
FEELNC	20,023	1,182	14,109	1,693	92.23	92.20	92.27	94.43	93.30
RNASAMBA	20,998	1,795	13,496	718	93.21	96.69	88.26	92.12	94.35
lncRNA_Mdeep	20,813	1,799	13,492	903	92.70	95.84	88.23	92.04	93.90
LncADeep	20,232	1,113	14,178	1,484	92.98	93.17	**92.72**	**94.79**	93.97

The bold values indicate the highest value within each column.

(Krogh et al., 2001) to identify transmembrane helices, Pfam (Mistry et al., 2020), and InterPro (Blum et al., 2020) to identify protein domains or families for all sequences in the test set. From InterPro, we considered InterPro entries, which include InterPro domain, family, homologous superfamily, repeat, and sites (i.e., active site, binding site, conserved site, PTM site). MobiDB (integrated within InterPro) (Piovesan et al., 2021) was also used to identify intrinsically disordered regions within sequences.

3 RESULTS

3.1 Model Evaluation Results

We compared the performance of Xlnc1DCNN with eight existing tools: CPC2, CPAT, CNIT, PLEK, FEELnc, RNAsamba, LncADeep, and lncRNA_Mdeep (Wang et al., 2013; Li et al., 2014; Kang et al., 2017; Wucher et al., 2017; Yang et al., 2018; Guo et al., 2019; Camargo et al., 2020; Fan et al., 2020) with the version listed in **Supplementary Table S1**. To have a fair and unbiased evaluation, we retrained CPAT, FEELnc, and RNAsamba that provided a training option using our human training dataset and used the pre-trained models of CPC2, CNIT, and LncADeep that did not provide a training option. Although PLEK and lncRNA_Mdeep came with a training option, retraining PLEK and lncRNA_Mdeep was very time-consuming, so we skipped retraining both and used their default pre-trained models.

3.1.1 Performance Evaluation on the Human Test Set

The results on the human test set (**Table 3**) show that Xlnc1DCNN achieved the highest accuracy (94.53) and F1-Score (95.38), the second-highest precision (94.55) slightly lower than LncADeep, and the third-highest specificity (92.13) slightly lower than LncADeep and FEELnc. CPC2, CNIT, and CPAT achieved high sensitivity but much lower specificity. While FEELnc, RNAsamba, LncADeep, and lncRNA_Mdeep performed well on the average of every metric but overall, still lower than Xlnc1DCNN. We then analyzed the classification power of each tool by plotting a receiver operating characteristic curve (ROC) and measuring the area under the curve (AUC) as shown in **Figure 2A**, where Xlnc1DCNN achieved the highest AUC (0.9825) on the human test set. **Figure 2B** shows that Xlnc1DCNN also outperformed all tools on any range of sequence lengths of the human test set (**Supplementary Table S2**).

3.1.2 Performance Evaluation on Cross-Species Datasets

To evaluate the generalization of Xlnc1DCNN with cross-species datasets, we compared the model with other tools using the mouse, gorilla, chicken, and cow datasets. The evaluation

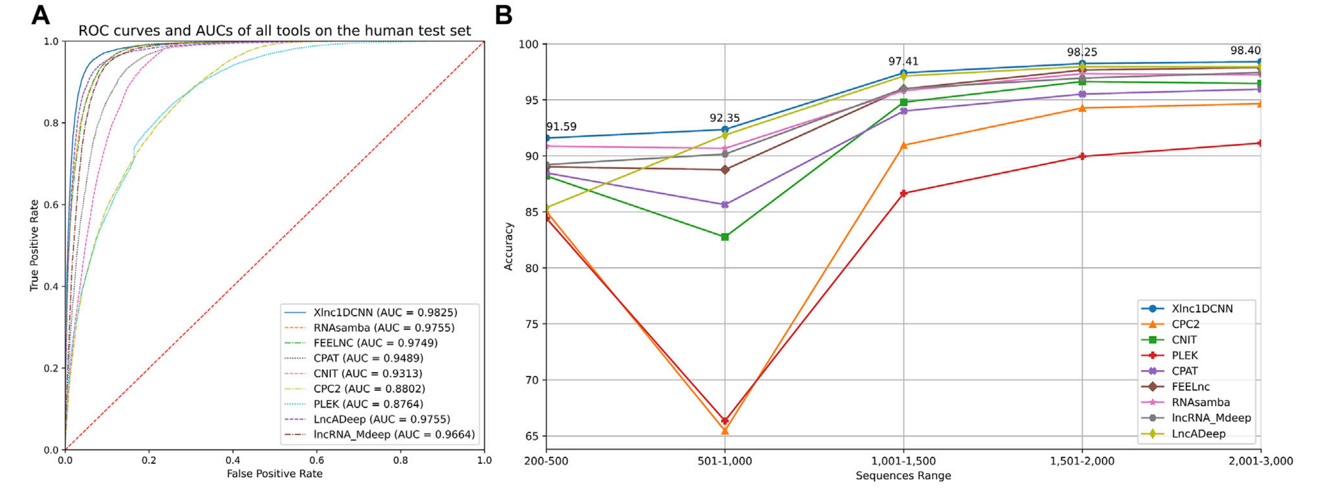


FIGURE 2 | (A) ROC curves of all tools and their AUCs on the human test set. **(B)** Accuracy of all tools for any range of sequence lengths of the human test set.

TABLE 4 | Accuracy of the nine models on cross-species datasets.

Model	Mouse	Gorilla	Chicken	Cow
Xlnc1DCNN	92.58	**96.06**	92.35	95.92
CPC2	80.06	94.96	93.51	94.48
CNIT	87.68	94.00	92.94	95.18
PLEK	73.62	89.53	79.54	86.22
CPAT	89.46	95.1	93.70	95.52
FEELnc	90.51	94.8	92.75	93.97
RNAsamba	91.91	**96.06**	**93.98**	96.39
LncADeep	**94.95**	96.05	93.46	**96.70**
lncRNA_Mdeep	91.38	95.58	92.59	95.63

The bold values indicate the highest value within each column.

results show that Xlnc1DCNN, which was trained on the human dataset, has a generalization for classifying lncRNAs and mRNAs on other species (**Table 4** and **Supplementary Tables S3–S6**). Xlnc1DCNN achieved the highest accuracy on the gorilla dataset together with RNAsamba and the second highest accuracy on the mouse dataset while LncADeep achieved the highest accuracy on mouse and cow datasets. **Figure 3** shows that Xlnc1DCNN has the ROC curves and AUCs close to other tools on cross-species datasets. Overall, based on AUCs, LncADeep got the best generalization performance on cross-species datasets.

3.2 Model Interpretation Results

As Xlnc1DCNN outperformed other tools on the human test set, we assumed that 1D-CNN captured patterns within sequences that could be used to distinguish lncRNAs from mRNAs. To explain the model, we used DeepSHAP to describe the contribution of each nucleotide to the prediction results. The explanation output from DeepSHAP was SHAP values for all nucleotides of the entire sequence. This explanation result was then visualized based on the summed SHAP values of each three consecutive nucleotides, with important representative amino acids highlighted in the sequence.

In the following subsections, we present the explanation results of Xlnc1DCNN focusing on the true positive, true negative, false positive, and false negative sequences predicted by Xlnc1DCNN on the human test set.

3.2.1 True Positive Sequences

The explanation results of Xlnc1DCNN highlighted the important regions that contributed to the correct classification of an input lncRNA transcript sequence as a lncRNA with blue color. From **Figures 4A,B**, the explanation results of the ENST00000658844.1 and lnc-REXO4-2:1 suggested that Xlnc1DCNN classified a transcript sequence as a lncRNA if it did not capture any important regions or specific patterns within the sequence. Additional explanation results of the TP sequences are shown in **Supplementary Figures S1–S8**.

3.2.2 True Negative Sequences

The explanation results of Xlnc1DCNN highlighted the important regions of a protein-coding transcript (i.e., mRNA) as red, as shown in **Figures 5A–C**. **Figure 5D** shows the transmembrane helix regions of the ENST00000528724.5 transcript predicted by TMHMM, corresponding to the important regions captured by Xlnc1DCNN. The prediction results of TMHMM and the explanation results of Xlnc1DCNN have similar patterns in several other mRNA transcripts within the test set (**Supplementary Figures S9 and S10**). **Figure 5E** shows the KRAB box (Krüppel associated box) identified by Pfam within the transcript ENST00000593088.5, which mostly overlapped with the important region captured by Xlnc1DCNN as shown in **Figure 5B**. **Figure 5F** shows the FAM32A family (family with sequence similarity 32 member A) identified by InterPro within the ENST00000589852.5 transcript, which corresponds to the important region of the ENST00000589852.5 identified by Xlnc1DCNN as shown in **Figure 5C**. This transcript has been linked to an ovarian tumor-associated gene (Chen et al., 2011). Additional explanation results of the TN sequences are shown in **Supplementary Figures S11–S14**.

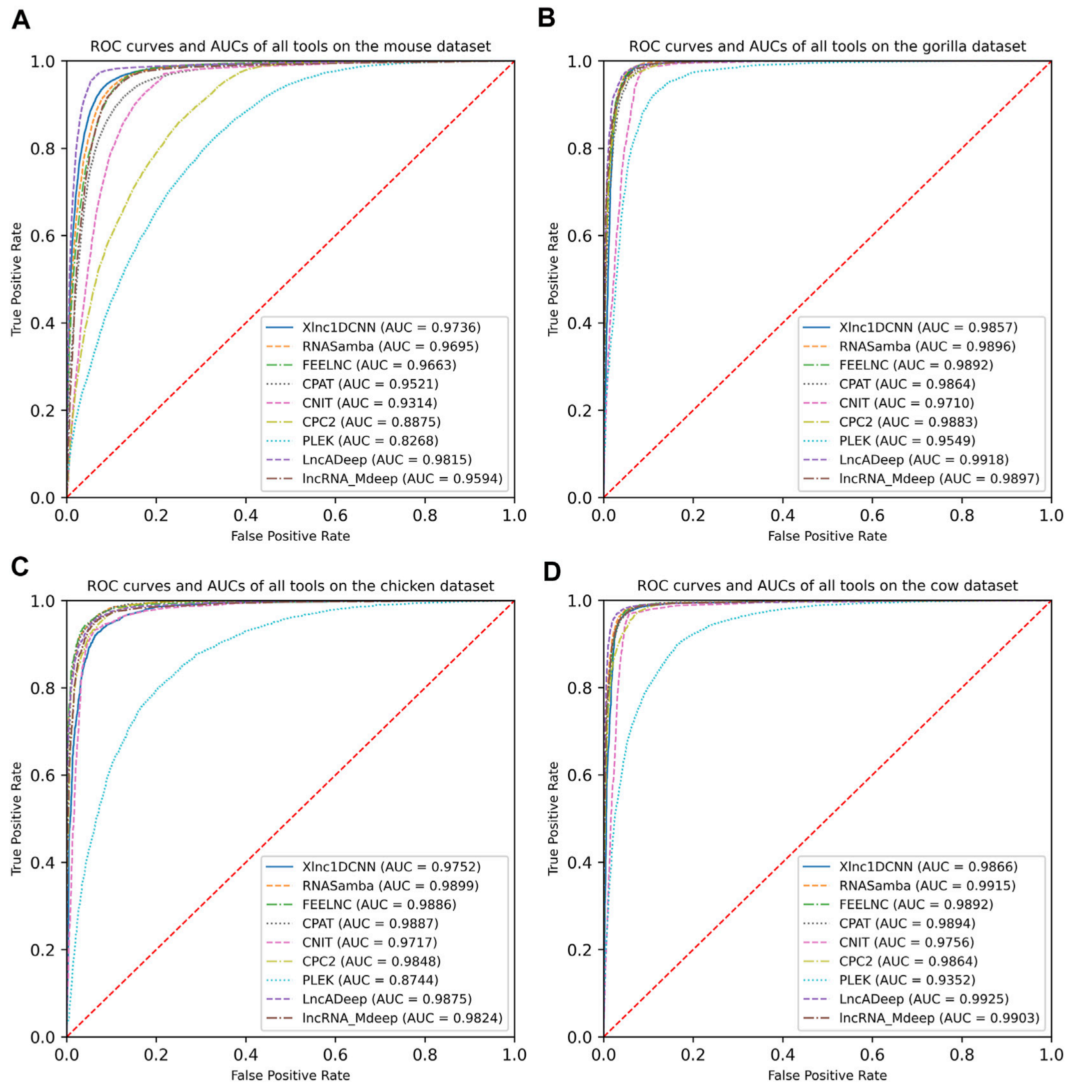


FIGURE 3 | Receiver operating characteristic curves and AUCs of nine models on the datasets of **(A)** mouse, **(B)** gorilla, **(C)** cow, and **(D)** chicken.

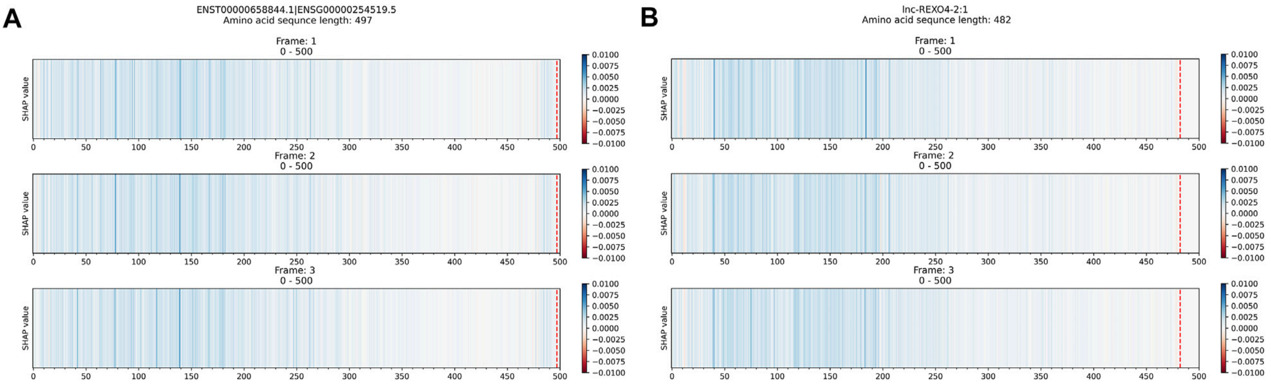
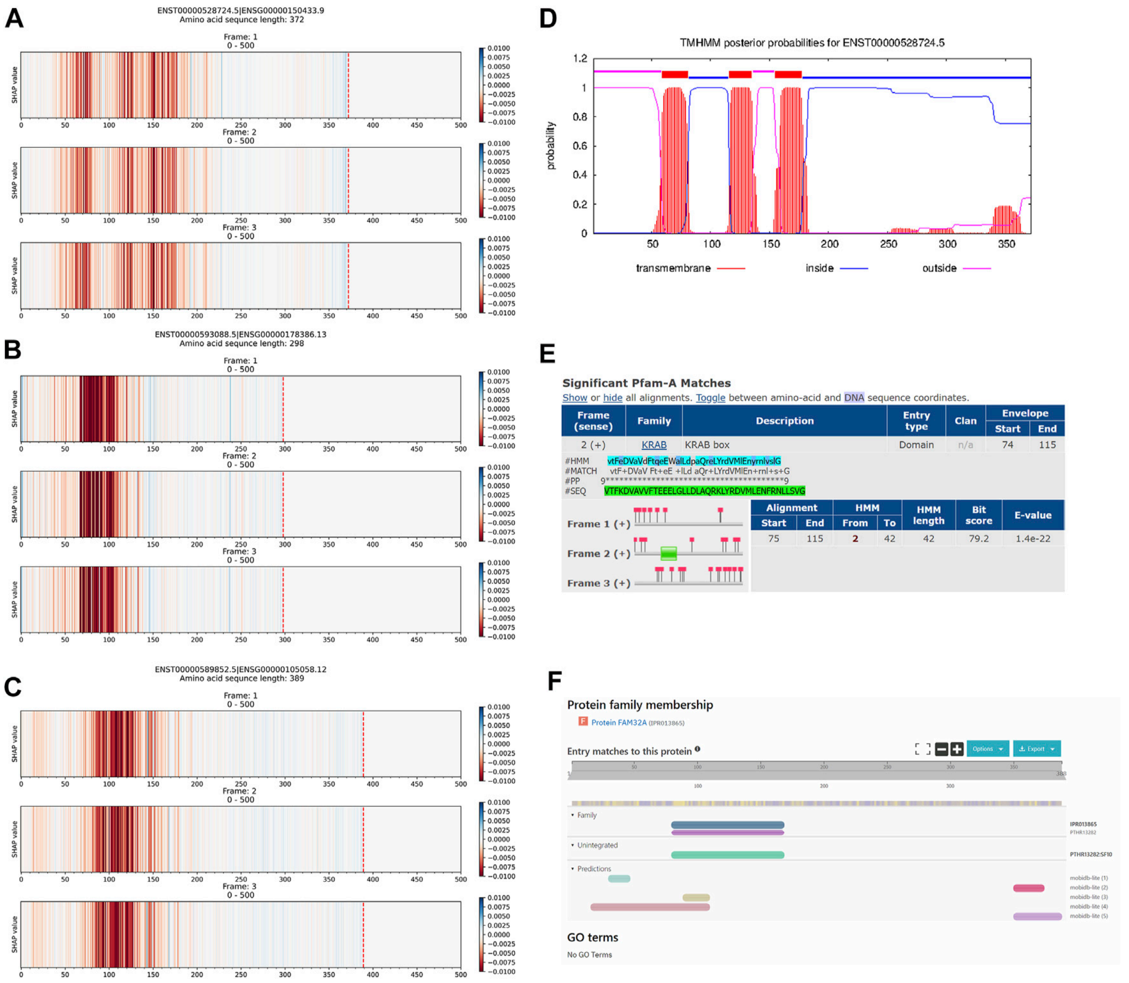


FIGURE 4 | Explanation results of Xlnc1DCNN on TP sequences **(A)** ENST00000658844.1, a lncRNA sequence obtained from GENCODE and **(B)** lnc-REXO4-2:1, a lncRNA sequence obtained from LNCipedia.



Significant Pfam-A Matches

Show or hide all alignments. Toggle between amino-acid and DNA sequence coordinates.

Frame (sense)	Family	Description	Entry type	Clan	Envelope	
					Start	End
2 (+)	KRAB	KRAB box	Domain	n/a	74	115

```
#HMM     vtFeDVaVdFtqeEWalLdpaQreLYrdVMlEnyrnlvslG
#MATCH    vtF+DVaV Ft+eE +lLd aQr+LYrdVMlEn+rnl+s+G
#PP     9***************************************9
#SEQ    VTFKDVAVVFTEEELGLLDLAQRKLYRDVMLENFRNLLSVG
```

Frame 1 (+)
Frame 2 (+)
Frame 3 (+)

Alignment		HMM		HMM length	Bit score	E-value
Start	End	From	To			
75	115	2	42	42	79.2	1.4e-22

FIGURE 5 | Comparison between the explanation results of Xlnc1DCNN on TN sequences **(A)** ENST00000528724.5 **(B)** ENST00000593088.5, and **(C)** ENST00000589852.5 protein-coding transcripts; and **(D)** prediction result of the TMHMM program on the ENST00000528724.5, **(E)** KRAB (Krüppel associated box) domain identified by Pfam within the ENST00000593088.5, and **(F)** FAM32A family identified by InterPro within the ENST00000589852.5 transcripts.

3.2.3 False Positive Sequences

False positive sequences are mRNA transcript sequences that are predicted as lncRNAs. **Figure 6A** shows the explanation result of ENST00000408930.6, which did not contain any important regions with red color contributing to the prediction as an mRNA. **Figures 6B,C** show Pfam and InterPro's results that both could not identify any protein domains or families within the ENST00000408930.6 protein-coding transcript. While the Ensembl database reports the ENST00000408930.6 as a protein-coding transcript of the HEPN1 (ENSG00000221932) gene, the Gene database at NCBI reports HEPN1 as the ncRNA gene (https://www.ncbi.nlm.nih.gov/gene/641654) and the RefSeq database reports the NR_170,124.1 (ENST00000408930.6) as a long non-coding RNA (https://www.ncbi.nlm.nih.gov/nuccore/NR_170124.1). Based on our evaluation, the top five long non-coding RNA identification (our Xlnc1DCNN, RNAsamba, LncADeep, lncRNA_Mdeep, FEELnc) predicted this sequence as lncRNA. This sequence highlights an example of inconsistent annotations among public databases that affect the model performance and evaluation. Additional explanation results of the FP sequences are shown in **Supplementary Figures S15–S19**.

3.2.4 False Negative Sequences

False negative sequences are lncRNA transcript sequences that are predicted as mRNAs. **Figures 7A,B** show the explanation results of lncRNAs: LNC-SIGIRR-2:1 and ENST00000616537.4 with important regions that contributed to the wrong prediction as mRNA transcripts. These regions correspond to the identified Anoctamin and the Taxilin InterPro families identified by InterPro, as shown in **Figures 7C,D**. Additional explanation results of the FN sequences are shown in **Supplementary Figures S20–S25**.

4 DISCUSSION

The explanation results of Xlnc1DCNN on the true positive sequences (TPs) show that most of the lncRNAs were found

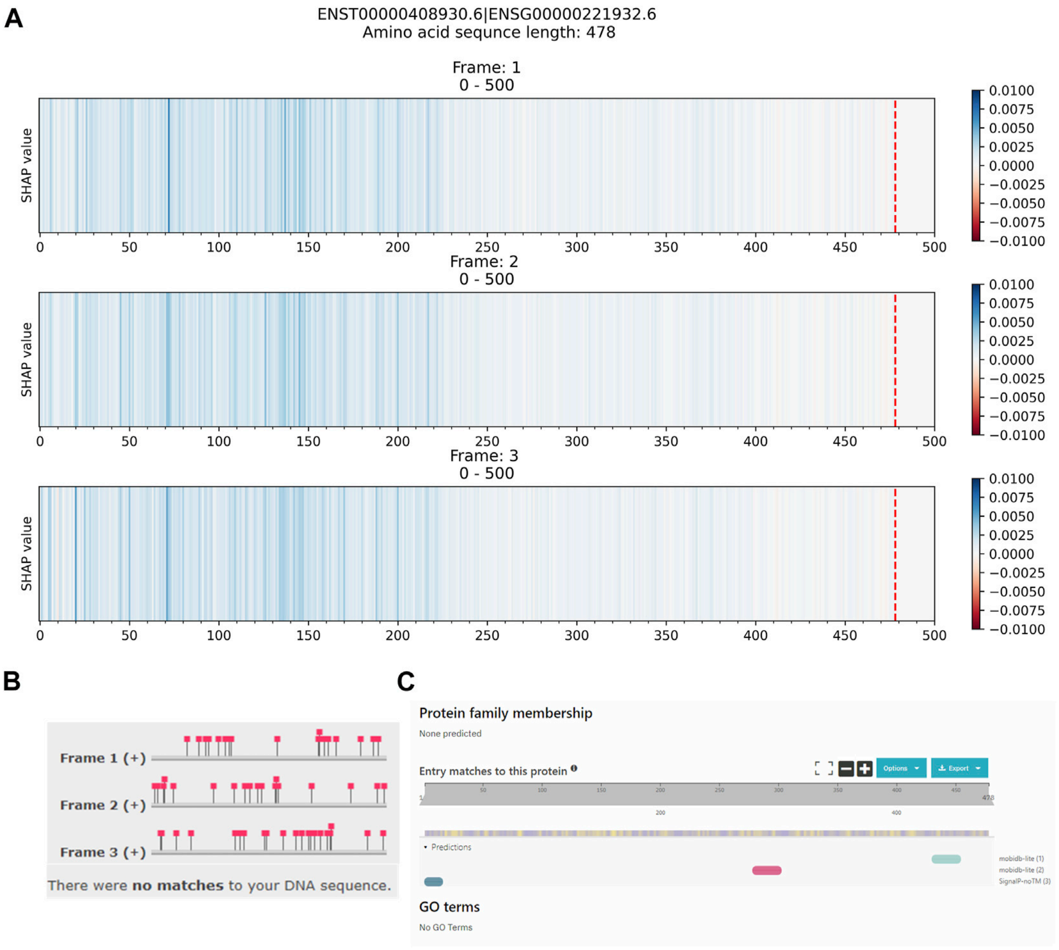


FIGURE 6 | Comparison between **(A)** the explanation result of Xlnc1DCNN on the ENST00000408930.6 protein-coding transcript, predicted as a lncRNA, **(B)** identification result from Pfam, and **(C)** identification result from InterPro.

with no conserved regions or patterns in short regions with unknown functions, i.e., the highlighted regions do not correspond to any InterPro entries (**Supplementary Figures S1, S2**). The important regions of some other lncRNA sequences highlighted transmembrane helices (**Supplementary Figures S3–S5**) or signal peptides (**Supplementary Figures S6–S8**). Over recent years, some studies also found a transmembrane helix inside lncRNAs (Anderson et al., 2015; Makarewich, 2020) and hidden peptides encoded within non-coding RNAs (Matsumoto and Nakayama, 2018). These findings correspond to what Xlnc1DCNN has learned and highlighted via the explanation result as important regions for classifying a sequence as lncRNA. Out of 20,895 TPs, only 1,692 (8.10%) TPs were found with InterPro entries, 9,833 (47.06%) TPs were found with only intrinsically disordered regions (IDRs), and 11,490 (36.91%) TPs were found with transmembrane helices identified by TMHMM without any InterPro entries. Although 8.10% of TPs were found with InterPro entries, top protein domains and families of the TPs were found in only a few TNs (≤5) on the test set (**Supplementary Tables S7, S8**).

On the true negative sequences (TNs), the explanation results of Xlnc1DCNN show that the model could capture the regions representing the protein domains or families in the transcript sequences. Out of 14,087 TNs, 13,079 (92.86%), 882 (5.84%), and 289 (2.05%) TNs were found with InterPro entries, only IDRs, and transmembrane helices were identified by TMHMM without any InterPro entries. Hence, it could classify most of the input mRNA sequences correctly as the protein-coding transcripts.

The explanation results of false positive sequences (FPs) typically do not contain the important regions (red color) that contributed to the model prediction as mRNAs. Out of 1,204 FPs, 500 (42.53%) FPs were found without any InterPro entries, 359 (29.81%) FPs were found with only IDRs, and 161 (13.37%) FPs were found with transmembrane helices without any InterPro entries.

For false negative sequences (FNs), from a total of 821 FNs, there were 463 (56.39%) FNs found with InterPro entries, and the explanation results of FNs also correspond to these entries as shown in **Figure 7**, and **Supplementary Figures S20–S25**, 264 (32.16%) FNs were found with only

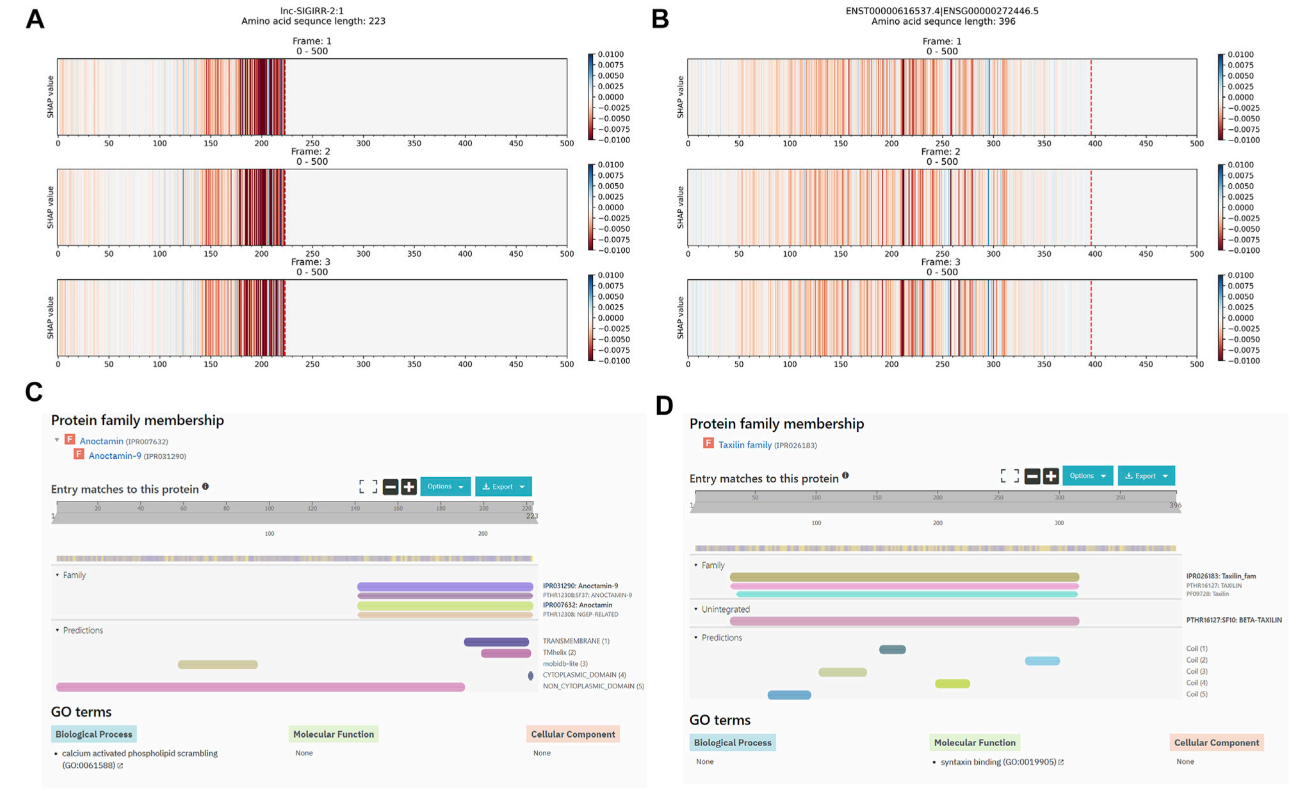


FIGURE 7 | Comparison between the explanation result of Xlnc1DCNN on the long non-coding RNA transcripts **(A)** lnc-SIGIRR-2:1 and **(B)** ENST00000616537.4, predicted as mRNAs; **(C)** Anoctamin family within the lnc-SIGIRR-2:1 transcript and **(D)** Taxilin family within the ENST00000616537.4 transcript identified by InterPro.

TABLE 5 | Summary of test set sequences annotated with InterPro entries.

Metrics	Amount	Found with InterPro Entries	Found without InterPro Entries	Contain IDRs without InterPro Entries	Contain Transmembrane Helices without InterPro Entries
TP	20,895	1,692 (8.10%)	19,203 (91.9%)	9,833 (47.06%)	7,713 (36.91%)
TN	14,087	13,085 (92.89%)	1,002 (7.11%)	822 (5.84%)	289 (2.05%)
FP	1,204	704 (58.47%)	500 (41.53%)	359 (29.82%)	161 (13.37%)
FN	821	463 (56.39%)	358 (43.61%)	264 (32.16%)	104 (12.67%)
All missed FP	344	93 (27.03%)	251 (72.97%)	164 (47.67%)	94 (27.33%)
All missed FN	105	90 (85.71%)	15 (14.92%)	5 (4.76%)	4 (3.81%)

IDRs and 104 (12.67%) FNs were found with transmembrane helices.

We summarized the TP, TN, FP, and FN sequences of the test set annotated with InterPro entries in **Table 5**. For TPs, most of the sequences were found without InterPro entries, in contrast with TNs. The number of TPs annotated with only IDRs, or transmembrane helices also highlighted the contributions of these regions to the predicted sequences as lncRNAs. The 704 out of 1,204 (58.47%) and 358 out of 821 (43.61%) annotated FPs and FNs with and without InterPro entries indicated the limitations of Xlnc1DCNN. We then further analyzed the misclassified FPs and FNs by top tools (Xlnc1DCNN, RNAsamba, LncADeep lncRNA_Mdeep, FEELnc). The 93 out of 344 (27.03%) and 15 out of 105 (14.92%) annotated FPs and FNs with and without InterPro entries misclassified by all top tools suggested sequences that were difficult to identify. Finally, the 251 out of 344 (72.97%) and 90 out of 105 (85.71%) annotated FPs and FNs without and with InterPro entries misclassified by all top tools suggested the possible limitations of all top tools or inconsistent annotations across the public databases.

We also analyzed the contribution of each nucleotide by plotting the mean of absolute SHAP values on the test set for a single nucleotide, dinucleotide, and trinucleotide (codon). The higher mean of absolute SHAP values indicates the higher impact of that genetic code (**Figure 8**). For

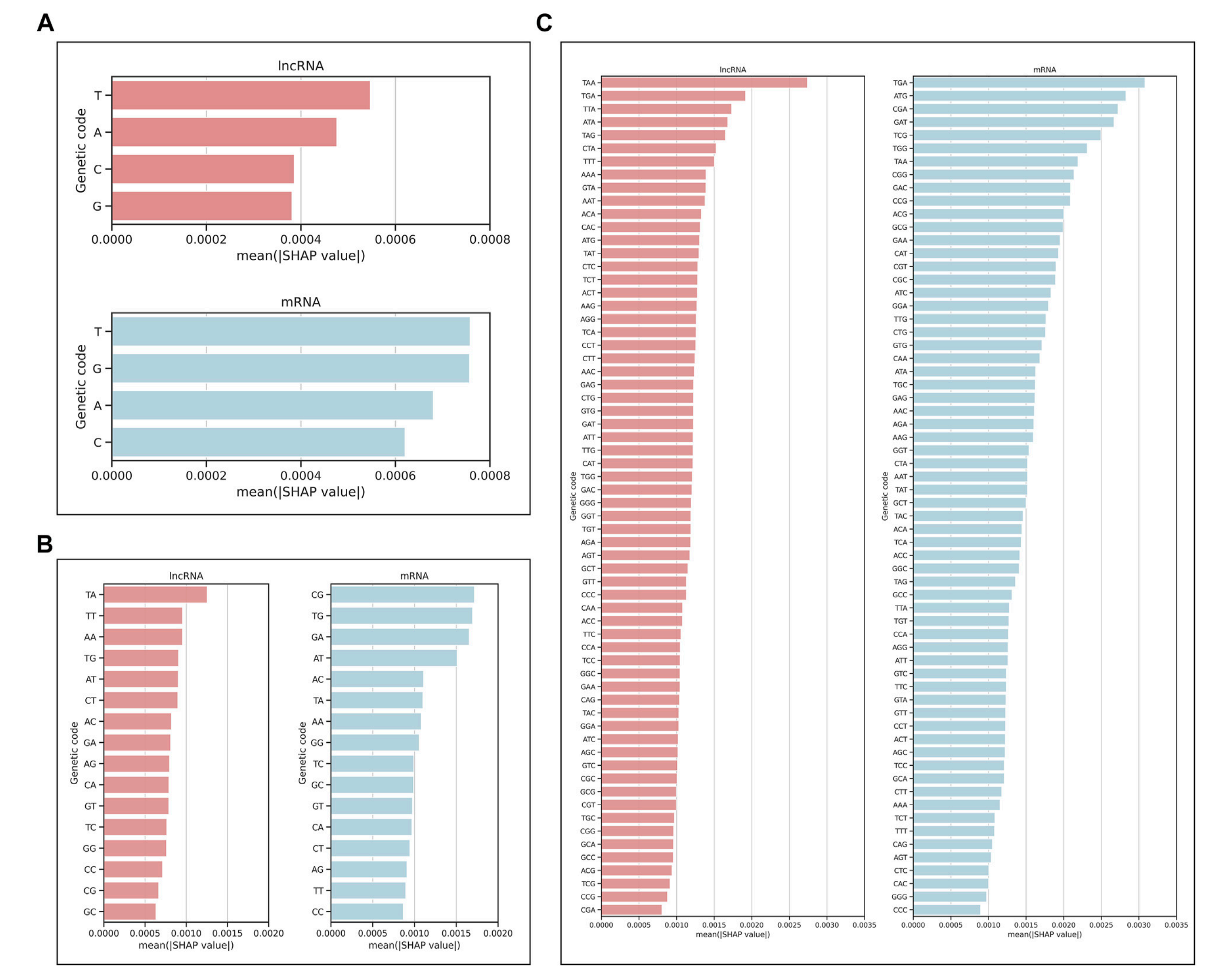


FIGURE 8 | Mean of absolute SHAP values for **(A)** single nucleotide, **(B)** dinucleotide, and **(C)** trinucleotide, indicating the impact of each genetic code on the model prediction as lncRNA or mRNA.

lncRNA, we found that the top three codons with the highest contribution were all stop codons (TAA, TGA, TTA), and for mRNA, the top three were the stop codon, start codon, and arginine (TGA, ATG, CGA). For dinucleotide, CG has the highest mean of absolute SHAP values for classifying as mRNA, which is consistent with those of (Ulveling et al., 2014).

As recent studies found that some putative lncRNAs contain a short open reading frame (sORF) (Hartford and Lal, 2020), we further analyzed the association of lncRNAs and sORF using the explanation results of Xlnc1DCNN. Some false negative sequences were randomly selected and checked if they contained sORF using MetamORF (Choteau et al., 2021). While MetamORF found sORFs in some of these sequences, the reported regions of these sORFs did not correspond to the important regions highlighted by the explanation results.

5 CONCLUSION

In this study, we proposed Xlnc1DCNN, a simple but effective 1D-CNN model for classifying and explaining lncRNA and protein-coding transcripts. We have shown that using 1D-CNN as a feature extractor can lead to a better prediction performance than other existing tools using traditional feature extraction methods. The explanation results provided insights into what the model learned to distinguish the lncRNA from protein-coding transcripts. The transmembrane helix region highlighted by the explanation results of several true positive lncRNA transcripts agreed with the recent findings of transmembrane microproteins within lncRNAs. Disordered proteins without any important regions highlighted in the explanation results were misclassified as lncRNAs. Several explanation results of lncRNA misclassified as protein-coding transcripts contained important regions that correspond to protein domains or families in Pfam and/or InterPro. These insights revealed the complexity of long non-coding RNAs and the need to evaluate cross-referenced gene annotation among public databases periodically.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

RL and DW: conceptualization, collecting resources, validation, investigation, writing—review and editing. RL: methodology, software, formal analysis, data curation, writing—original draft preparation, and visualization. DW: supervision, project administration, and funding acquisition. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

The Ratchadaphiseksomphot Endowment Fund Part of the “Research Grant for New Scholar CU Researcher’s Project”, grant number (RGN_2559_025_06_21).

ACKNOWLEDGMENTS

A preprint of this article was previously deposited to bioRxiv and can be found here with this doi: https://doi.org/10.1101/2022.02.11.479495

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2022.876721/full#supplementary-material

REFERENCES

Acharya, U. R., Fujita, H., Lih, O. S., Hagiwara, Y., Tan, J. H., and Adam, M. (2017). Automated Detection of Arrhythmias Using Different Intervals of Tachycardia ECG Segments with Convolutional Neural Network. *Inf. Sci.* 405, 81–90. doi:10.1016/j.ins.2017.04.012

Anderson, D. M., Anderson, K. M., Chang, C.-L., Makarewich, C. A., Nelson, B. R., McAnally, J. R., et al. (2015). A Micropeptide Encoded by a Putative Long Noncoding RNA Regulates Muscle Performance. *Cell* 160 (4), 595–606. doi:10.1016/j.cell.2015.01.009

Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., et al. (2020). The InterPro Protein Families and Domains Database: 20 Years on. *Nucleic Acids Res.* 49 (D1), D344–D354. doi:10.1093/nar/gkaa977

Camargo, A. P., Sourkov, V., Pereira, G. A. G., and Carazzolle, M. F. (2020). RNAsamba: Neural Network-Based Assessment of the Protein-Coding Potential of RNA Sequences. *NAR Genomics and Bioinformatics* 2 (1), lqz024. doi:10.1093/nargab/lqz024

Chan, J., and Tay, Y. (2018). Noncoding RNA:RNA Regulatory Networks in Cancer. *Int. J. Mol. Sci.* 19 (5), 1310. doi:10.3390/ijms19051310

Chen, X., Zhang, H., Aravindakshan, J. P., Gotlieb, W. H., and Sairam, M. R. (2011). Anti-proliferative and Pro-apoptotic Actions of a Novel Human and Mouse Ovarian Tumor-Associated Gene OTAG-12: Downregulation, Alternative Splicing and Drug Sensitization. *Oncogene* 30 (25), 2874–2887. doi:10.1038/onc.2011.11

Chen, H., Lundberg, S., and Lee, S.-I. (2019). *Explaining Models by Propagating Shapley Values of Local Components*.

Choteau, S. A., Wagner, A., Pierre, P., Spinelli, L., and Brun, C. (2021). MetamORF: a Repository of Unique Short Open reading Frames Identified by Both Experimental and Computational Approaches for Gene and Metagene Analyses. *Database* 2021, baab032. doi:10.1093/database/baab032

Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., et al. (2019). Ensembl 2019. *Nucleic Acids Res.* 47 (D1), D745–D688. doi:10.1093/nar/gky1113

Fan, X.-N., Zhang, S.-W., Zhang, S.-Y., and Ni, J.-J. (2020). lncRNA_Mdeep: An Alignment-free Predictor for Distinguishing Long Non-coding RNAs from Protein-Coding Transcripts by Multimodal Deep Learning. *Int. J. Mol. Sci.* 21 (15), 5222. doi:10.3390/ijms21155222

Fang, Y., and Fullwood, M. J. (2016). Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics & Bioinformatics* 14 (1), 42–54. doi:10.1016/j.gpb.2015.09.006

Frankish, A., Diekhans, M., Ferreira, A.-M., Johnson, R., Jungreis, I., Loveland, J., et al. (2018). GENCODE Reference Annotation for the Human and Mouse Genomes. *Nucleic Acids Res.* 47 (D1), D766–D773. doi:10.1093/nar/gky955

Guo, J.-C., Fang, S.-S., Wu, Y., Zhang, J.-H., Chen, Y., Liu, J., et al. (2019). CNIT: a Fast and Accurate Web Tool for Identifying Protein-Coding and Long Non-coding Transcripts Based on Intrinsic Sequence Composition. *Nucleic Acids Res.* 47 (W1), W516–W522. doi:10.1093/nar/gkz400

Hartford, C. C. R., and Lal, A. (2020). When Long Noncoding Becomes Protein Coding. *Mol. Cel Biol* 40 (6), e00528–00519. doi:10.1128/MCB.00528-19

Hsieh, C.-H., Li, Y.-S., Hwang, B.-J., and Hsiao, C.-H. (2020). Detection of Atrial Fibrillation Using 1D Convolutional Neural Network. *Sensors* 20 (7), 2136. doi:10.3390/s20072136

Jin, K.-T., Yao, J.-Y., Fang, X.-L., Di, H., and Ma, Y.-Y. (2020). Roles of lncRNAs in Cancer: Focusing on Angiogenesis. *Life Sci.* 252, 117647. doi:10.1016/j.lfs.2020.117647

Kang, Y.-J., Yang, D.-C., Kong, L., Hou, M., Meng, Y.-Q., Wei, L., et al. (2017). CPC2: a Fast and Accurate Coding Potential Calculator Based on Sequence Intrinsic Features. *Nucleic Acids Res.* 45 (W1), W12–W16. doi:10.1093/nar/gkx428

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., and Inman, D. J. (2021). 1D Convolutional Neural Networks and Applications: A Survey. *Mech. Syst. Signal Process.* 151, 107398. doi:10.1016/j.ymssp.2020.107398

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting Transmembrane Protein Topology With a Hidden Markov Model: Application to Complete Genomes. *J. Mol. Biol.* 305 (3), 567–580. doi:10.1006/jmbi.2000.4315

Li, A., Zhang, J., and Zhou, Z. (2014). PLEK: a Tool for Predicting Long Non-coding RNAs and Messenger RNAs Based on an Improved K-Mer Scheme. *BMC Bioinformatics* 15 (1), 311. doi:10.1186/1471-2105-15-311

Li, F., Liu, M., Zhao, Y., Kong, L., Dong, L., Liu, X., et al. (2019). Feature Extraction and Classification of Heart Sound Using 1D Convolutional Neural Networks. *EURASIP J. Adv. Signal. Process.* 2019 (1), 59. doi:10.1186/s13634-019-0651-3

Li, W., and Godzik, A. (2006). Cd-hit: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics* 22 (13), 1658–1659. doi:10.1093/bioinformatics/btl158

Lundberg, S. M., and Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*, 4765–4774.

Makarewich, C. A. (2020). The Hidden World of Membrane Microproteins. *Exp. Cel Res.* 388 (2), 111853. doi:10.1016/j.yexcr.2020.111853

Marchese, F. P., Raimondi, I., and Huarte, M. (2017). The Multidimensional Mechanisms of Long Noncoding RNA Function. *Genome Biol.* 18 (1), 206. doi:10.1186/s13059-017-1348-2

Matsumoto, A., and Nakayama, K. I. (2018). Hidden Peptides Encoded by Putative Noncoding RNAs. *Cell Struct. Funct.* 43 (1), 75–83. doi:10.1247/csf.18005

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, Gustavo, A., et al. (2020). Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* 49 (D1), D412–D419. doi:10.1093/nar/gkaa913

Morán, I., Akerman, İ., van de Bunt, M., Xie, R., Benazra, M., Nammo, T., et al. (2012). Human β Cell Transcriptome Analysis Uncovers lncRNAs that Are Tissue-specific, Dynamically Regulated, and Abnormally Expressed in Type 2 Diabetes. *Cel Metab.* 16 (4), 435–448. doi:10.1016/j.cmet.2012.08.010

Piovesan, D., Necci, M., Escobedo, N., Monzon, A. M., Hatos, A., Mičetić, I., et al. (2021). MobiDB: Intrinsically Disordered Proteins in 2021. *Nucleic Acids Res.* 49 (D1), D361–d367. doi:10.1093/nar/gkaa1058

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). “"Why Should I Trust You?": Explaining the Predictions of Any Classifier,” in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA (Association for Computing Machinery). doi:10.1145/2939672.2939778

Rinn, J. L., and Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annu. Rev. Biochem.* 81 (1), 145–166. doi:10.1146/annurev-biochem-051410-092902

Shrikumar, A., Greenside, P., and Kundaje, A. (2017). “Learning Important Features through Propagating Activation Differences,” in Proceedings of the 34th International Conference on Machine Learning - Volume 70, Sydney, NSW, Australia (JMLR.org). doi:10.48550/arXiv.1704.02685

Stark, R., Grzelak, M., and Hadfield, J. (2019). RNA Sequencing: the Teenage Years. *Nat. Rev. Genet.* 20 (11), 631–656. doi:10.1038/s41576-019-0150-2

Statello, L., Guo, C.-J., Chen, L.-L., and Huarte, M. (2021). Gene Regulation by Long Non-coding RNAs and its Biological Functions. *Nat. Rev. Mol. Cel Biol* 22 (2), 96–118. doi:10.1038/s41580-020-00315-9

Tjoa, E., and Guan, C. (2021). A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* 32 (11), 4793–4813. doi:10.1109/tnnls.2020.3027314

Ulveling, D., Dinger, M. E., Francastel, C., and HubÃ©, F. (2014). Identification of a Dinucleotide Signature that Discriminates Coding from Non-coding Long RNAs. *Front. Genet.* 5, 316. doi:10.3389/fgene.2014.00316

Volders, P.-J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P., et al. (2018). LNCipedia 5: towards a Reference Set of Human Long Non-coding RNAs. *Nucleic Acids Res.* 47 (D1), D135–D139. doi:10.1093/nar/gky1031

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-seq: a Revolutionary Tool for Transcriptomics. *Nat. Rev. Genet.* 10 (1), 57–63. doi:10.1038/nrg2484

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J.-P., and Li, W. (2013). CPAT: Coding-Potential Assessment Tool Using an Alignment-free Logistic Regression Model. *Nucleic Acids Res.* 41 (6), e74. doi:10.1093/nar/gkt006

Wucher, V., Legeai, F., Hédan, B., Rizk, G., Lagoutte, L., Leeb, T., et al. (2017). FEELnc: a Tool for Long Non-coding RNA Annotation and its Application to the Dog Transcriptome. *Nucleic Acids Res.* 45 (8), gkw1306–e57. doi:10.1093/nar/gkw1306

Yamashita, R., Nishio, M., Do, R. K. G., and Togashi, K. (2018). Convolutional Neural Networks: an Overview and Application in Radiology. *Insights Imaging* 9 (4), 611–629. doi:10.1007/s13244-018-0639-9

Yang, C., Yang, L., Zhou, M., Xie, H., Zhang, C., Wang, M. D., et al. (2018). LncADeep: Anab initiolncRNA Identification and Functional Annotation Tool Based on Deep Learning. *Bioinformatics* 34 (22), 3825–3834. doi:10.1093/bioinformatics/bty428

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.

Publisher’s Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.