PaRPI predicts RNA-Protein interactions from cross-protocol and cross-batch RNA-binding protein datasets

Liangchen Peng[1,♯], Lijun Quan[1,2,3,♯,*], Lingkun Meng[4], Zhihong Zhang[4], Shengju Zhang[4], Zhijun Zhang[1], Yi Zhang[1], Qiufeng Chen[1], Bei Zhang[1], Lexin Cao[1], Tingfang Wu[1,2,3], and Qiang Lyu[1,2,3,*]

[1]School of Computer Science and Technology, Soochow University, Suzhou 215006, China
[2]Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing 210000, China
[3]Jiangsu Key Laboratory of Drug Discovery and Translational Research for Brain Diseases, Soochow University, 215006, Suzhou, China
[4]China Mobile (Suzhou) Software Technology Company Limited, Suzhou 215163, China

ABSTRACT

RNA-binding proteins (RBPs) play a pivotal role in the regulation of gene expression, with their interactions with RNA reflecting the biological functions and regulatory mechanisms. However, current computational methods are typically tailored to specific RBPs and depend on specific protocols and batches of biological experiments. To overcome these challenges, we propose a method called PaRPI, which aims to predict RNA-protein binding sites in a bidirectional RBP-RNA selection manner. PaRPI groups all RBP datasets based on cell lines, integrating experimental data from different protocols and batches, thereby enabling the development of a unified computational model that effectively captures both shared and distinct interaction patterns among different proteins. Our results demonstrate that PaRPI achieves exceptional performance in accurately identifying binding sites, surpassing state-of-the-art models on 261 RBP datasets from eCLIP and CLIP-seq experiments. Furthermore, PaRPI stands out for its robust generalization capabilities, uniquely able to predict interactions with previously unseen RNA and protein receptors. We also investigate the impact of disease-associated variants on RBP binding and evaluate PaRPI's components and semantic embeddings, demonstrating its capability to dissect complex interaction networks. PaRPI enables large-scale exploration of RNA-protein interactions, facilitating future studies on gene regulation and disease mechanisms.

Introduction

Protein-RNA interactions form the cornerstone of key biological processes such as mRNA splicing, localization, degradation, and translation [1–3]. Given that RBPs constitute nearly 10% of the human proteome, understanding their interactions with RNA is crucial for elucidating their biological functions

♯ These authors contributed equally: Liangchen Peng, Lijun Quan.
∗e-mail: ljquan@suda.edu.cn and qiang@suda.edu.cn.

and regulatory mechanisms. To this end, numerous high-throughput technologies have been developed to profile and predict RNA-protein interactions (RPI). Techniques such as systematic evolution of ligands by exponential selection (SELEX) [4], RNAcompete [5], and RNA Bind-n-Seq [6] have significantly advanced the characterization of RBP sequence preferences in vitro. Complementing these in vitro methods, approaches such as RNA immunoprecipitation (RIP) [7] and UV crosslinking-based methods (e.g., HITS-CLIP, PAR-CLIP, and eCLIP) have enabled the identification of RBP binding sites in vivo. [8–10]

Despite these advancements, experimental methods often come with significant costs, time-consuming procedures, and potential biases, highlighting the need for computational approaches to efficiently and accurately predict RBP binding profiles under diverse cellular conditions. With the emergence of CLIP-seq technologies [11, 12], numerous RBP-RNA binding targets have been uncovered, facilitating the development of effective data-driven computational methods [13,14]. For instance, DeepBind [15] employs deep neural network (DNN) to identify RBP binding preferences in RNA sequence data. Similarly, Ilan et al. developed DLPRB [16] combines convolutional neural network (CNN) and recurrent neural networks (RNNs) to learn intrinsic RBP binding preferences and predict interactions. Other methods, such as iDeep [17], integrate convolutional networks and deep belief networks to predict RBP binding sites and motifs on RNA, while GraphProt [18] utilizes graph-based kernels to integrate RNA sequence and secondary structure features for RBP binding prediction. Additionally, PRIESSTESS [19] uses a LASSO-regularized logistic regression model to identify enriched RNA sequence and/or structural motifs. While these methods have advanced our understanding of RBP binding, most focus exclusively on specific cellular conditions and are limited in their ability to predict protein-RNA interactions across dynamic contexts. In fact, studies have demonstrated that RBP-RNA interaction patterns are influenced by diverse cellular and tissue environments. Recently, new deep learning approaches have been developed to address these dynamic contexts. For example, PrismNet [20] integrates in vivo experimental RNA structure information to predict dynamic RBP binding across various cellular conditions. HDRNet [21] builds on PrismNet by utilizing Bidirectional Encoder Representations from Transformers (BERT) [22] to learn contextual RNA sequence information and employs hierarchical multi-scale residual networks to capture nucleotide-level dependencies.

In the aforementioned literature, computational models for predicting RBPs evolved from relying on small datasets derived from biological experiments. These early models were trained on individual datasets specific to particular RBPs, effectively modeling the unidirectional selection of RNA by RBPs without sensing the existence of RBP as a necessary counterpart of RBP-RNA complex. This implies that the binding preferences of each specific RBP for all RNAs are treated as isolated and independent, rendering them ineffective for RBPs not covered by experimental datasets. Moreover, these binding preferences also depend on specific protocols and batches of biological experiments.The RBP-RNA complex consists of two counterparts that are essential to each other. This paper proposes that the RBP-RNA complex, consisting of two mutually dependent components, emerges from bidirectional selection. Specifically, while the RBP selects RNA, as traditionally understood, the RNA reciprocally selects the RBP, a characteristic we define as "protein-aware." Therefore, the formation of the RBP-RNA complex may follow a similar rule but is conditioned on a specific type of cell line. This rule sounds also to be followed no matter what protocols and batches of biological experiments are adopted when targeting on the investigations of RBP-RNA binding preferences.

In this study, we propose a method to overcome these limitations, and refer to it as RBP-aware interaction prediction (PaRPI for short). We group all experimental datasets based on cell lines, integrating data from different experimental protocols and batches, thereby encompassing both the RNA selection preferences of RBPs and the RBP selection preferences of RNAs. PaRPI models the binding preferences of RBPs and RNAs on such cross-protocol, cross-batch, and RBP-aware experimental datasets. A direct advantage of this approach is its potential to study the possibility of RBP-RNA

complexes formed by RBPs not covered by experimental datasets. Specifically, PaRPI utilizes the large language model ESM-2 [23] to obtain protein representations, enhancing the ability to mine interactions between proteins and ligands, which aids in the formation of binding sites. Next, PaRPI learns RNA representations by combining GNNs and the Transformer architecture, revealing underlying relationships between nucleotides based on sequence and structural information. Additionally, the interaction module integrates protein and RNA representations, enabling the model to accurately identify binding patterns between RNAs and RBPs. We trained and evaluated PaRPI on 261 RBP datasets, comparing it with other baseline methods. The results show that PaRPI outperforms competing methods on the majority of RBP datasets, achieving the best performance on 209 RBP datasets. Notably, PaRPI exhibits excellent performance in predicting interactions of homologous proteins and achieves effective cross-cell predictions for the same protein receptor. The model's capabilities extend to predicting interactions of novel RNA and RBPs and provide a degree of cross-species and cross-protocol and batch prediction capabilities. Furthermore, through motifs identified by PaRPI, we gain insights into the mechanisms of RNA-protein binding recognition, thereby offering biological explanations for RBP-RNA interactions. We also investigated the impact of disease-associated genetic variants on RNA-protein binding. Finally, we evaluated the influence of individual components within PaRPI and conducted a visualization analysis of its semantic representations, highlighting the model's robustness and versatility in understanding the complex landscape of RNA-protein interactions.

Results

Overview of PaRPI

The PaRPI framework leverages robust multi-source biological features and interaction modules to accurately predict RNA-binding events across diverse cell lines. As illustrated in Figure 1, PaRPI exhibits strong generalization capabilities, including the ability to predict interactions involving previously unseen RBPs. Unlike previous methods such as PrismNet and HDRNet, which rely solely on RNA sequence and structural features, PaRPI explicitly incorporates protein sequence information by encoding receptors using the ESM-2 pre-trained language model. This allows the model to capture rich contextual and evolutionary signals from protein sequences. This receptor-aware design stands in contrast to traditional approaches that either omit protein features altogether. As a result, PaRPI can model RNA-binding preferences across both known and novel RBPs.

For RNA, PaRPI employs the k-mer method for encoding RNA sequences, which are subsequently input into a pre-trained BERT model. This provides feature vectors that capture nucleotide context information and long-range dependencies. Concurrently, RNA secondary structure features are extracted using the icSHAPE pipeline [24] and RNAplfold [25], providing valuable insights into RBP structural binding preferences.PaRPI then uses two CNN modules to harmonize the feature dimensions derived from BERT and icSHAPE. Following this, an RNA graph is constructed through a graph conversion module, where node features combine the features derived from BERT and icSHAPE, and spatial features are defined by sequence adjacency edges and secondary structure data from RNAplfold.

In the Learning Interaction Module, multimodal feature fusion is performed based on both internal RNA nucleotide interactions and interactions between RNA and protein receptors. A GraphConv module aggregates node information, and the updated node features are processed through a Transformer to effectively capture long-range dependencies within the RNA structure. Subsequently, the deep protein-RNA binding predictor (DPRBP) [21] module incrementally reduces the dimensionality of the sequence length by selecting key nucleotide tokens at each layer. Ultimately, interaction module is applied to integrate processed RNA features with protein representations. The fused features are then fed into a

multi-layer perceptron (MLP) classifier to predict the binding affinity between RNA and protein.

To emphasize, our model is fundamentally based on cell line-specific grouping, integrating diverse experimental protocols and batches. This framework comprehensively captures both the RNA-binding preferences of RBPs and the protein-binding preferences of RNAs, representing a bidirectional selection paradigm. This approach significantly advances our understanding of RNA-protein interactions, enabling cross-protein/cell line predictions, transfer learning applications, motif identification, and even mutation impact analysis.

PaRPI outperformance over existing methods on RBP binding events

We evaluated the performance of our proposed PaRPI by comparing it with six state-of-the-art computational methods (HDRNet [21], PrismNet [20], PRIESSTESS [19], DMSK [26], iDeep [27], DeepBind [28], and GraphProt [29], detailed links for all the aforementioned comparative methods are provided in Supplementary Information 1 on a database of 261 RBP binding site datasets.These datasets from a unified pipeline were used [20], ensuring both accuracy and consistency in the evaluation, and allowing for a fair comparison. It is important to note that there are six PaRPI prediction models, each trained on ensemble cell line-specific datasets, including K562, HepG2, HEK293, HEK293T, HeLa, and H9 cell lines. In contrast, other methods constructed 261 models based on both protein and cell line specific data. For further details, please refer to the dataset construction section in the Methods.The area under the receiver operating characteristic curve (AUC) was used as the performance metric for all methods. For each RBP dataset, we partitioned the binding sites into training and testing sets, then merged them for model evaluation. The trained integrated model was subsequently used to benchmark each RBP test set.

The Circos plot in Figure 2a illustrates that PaRPI outperforms the competing methods on the majority of the RBP datasets, securing the top position in 209 RBP datasets, while the other top 3 baseline methods rank first in 49, 3 and 1 RBP datasets, respectively. The violin plot in Figure 2b further emphasizes PaRPI's consistent performance. Across the 261 RBP datasets, PaRPI surpasses HDRNet, the best-performing baseline, with an average AUC increase of 1.6%. As shown in Figure 2c, the CPSF6_K562 and KHDRBS1_K562 datasets exemplify where PaRPI achieves superior AUC values. However, PaRPI exhibited suboptimal performance in the SF3B4_HepG2 and SRRM4_HEK293T datasets, achieving slightly lower AUC scores than HDRNet but outperforming other methods. This discrepancy may be attributed to the specific binding characteristics of RBPs, which the integrated model fails to capture effectively. These results clearly demonstrate the strong competitiveness of PaRPI in most RBP-related data evaluation scenarios, indicating its higher sensitivity in identifying RBP binding sites.

To further evaluate the robustness of PaRPI across all 261 datasets, we incorporated additional performance metrics to compare PaRPI with HDRNet and PrismNet. Specifically, we summarized the distributions of Accuracy (ACC), F1 score, and Matthews Correlation Coefficient (MCC) using box plots. As shown in Figure 2d, PaRPI consistently outperformed both baseline methods across all three metrics. Specifically, PaRPI achieved median values of approximately 0.825 for ACC, 0.761 for F1, and 0.635 for MCC, demonstrating high stability and reliable prediction performance. In comparison, HDRNet and PrismNet yielded lower median scores. Notably, PaRPI improved upon HDRNet by approximately 1.3% in ACC, 3.9% in F1, and 5.1% in MCC, underscoring its superior predictive consistency and generalization ability. We also present the Receiver Operating Characteristic (ROC) and Precision-Recall (PRC) curves for two randomly selected datasets, with detailed analyses provided in Supplementary Information 2.

To assess PaRPI under more stringent conditions, we applied MMseqs2 [30] to remove test RNA sequences with greater than 80% identity to any training RNA sequence. As shown in Figure 2e, PaRPI

continued to demonstrate strong performance across all four evaluation metrics. It achieved a median AUC of 0.902, ACC of 0.825, F1 score of 0.723, and MCC of 0.613, indicating robustness and predictive reliability. Compared to HDRNet, PaRPI showed relative improvements of approximately 5.1% in AUC, 5.9% in ACC, 11.6% in F1, and 19.3% in MCC. These gains were even more pronounced when compared to PrismNet, especially in F1 and MCC, indicating that PaRPI retains robust predictive performance even under stringent, redundancy-reduced conditions.

Transfer predicting binding preference of homologous proteins

To evaluate PaRPI's performance across varying levels of protein sequence similarity, we performed sequence similarity clustering using MMseqs2 for each cell line and integrated the classification results. Proteins were categorized into three groups based on pairwise sequence identity: low similarity (< 0.3), medium similarity (0.3–0.6), and high similarity (> 0.6). This comprehensive analysis aimed to assess PaRPI's predictive capability across a wide range of homologous and non-homologous protein datasets. As shown in Figure 3a, PaRPI consistently outperformed HDRNet across all similarity levels, achieving an average AUC improvement of 1.7% in the high-similarity group. These results highlight PaRPI's robustness in predicting interactions not only among homologous proteins but also in non-homologous scenarios, with particularly notable advantages in the high-similarity group.

In addition, we performed a detailed analysis on the K562 cell line by extracting the 112 RBP datasets belonging to this cell line from the previously defined similarity groups. The results further confirm PaRPI's superior performance in identifying interactions among closely related proteins. Specifically, PaRPI achieved an average AUC improvement of 2.8% over HDRNet in the high-similarity group and 1.7% in the medium-similarity group, demonstrating consistent gains across all similarity levels. Notably, the performance advantage was most pronounced among highly similar proteins, highlighting PaRPI's enhanced generalization and sensitivity in handling homologous sequences.

Focusing on the high-similarity group, we identified through the STRING database [31] that FMR1, FXR1, and FXR2 are closely related members of the fragile X mental retardation protein (FMRP) family. As shown in Figure 3b, PaRPI achieved AUC scores of 0.918, 0.925, and 0.938 for FMR1, FXR1, and FXR2, respectively, outperforming other models across all three proteins. These proteins are co-expressed in the cytoplasm of specifically differentiated neurons [32] and exhibit significant structural and functional similarities. Their motif-binding sites were found to be highly conserved in the mCross database [33], as illustrated in Figure 3c. This strong sequence and functional similarity likely contributes to the improved performance of PaRPI, highlighting its ability to effectively capture and utilize shared features among homologous proteins in RNA-protein interaction prediction. In contrast to conventional baseline methodologies, our approach involves encoding proteins in a way that allows the model to identify and leverage the structural and functional similarities among homologous proteins. By doing so, we significantly enhance the model's ability to generalize across different protein families. This architectural design is particularly advantageous in improving predictive accuracy during joint training on datasets comprising homologous proteins, as it effectively captures the shared characteristics and evolutionary relationships within protein families.

Transfer predicting binding preference of identical RBP cross cell lines

RBP binding behavior is influenced not only by RNA and protein but also by the cellular context. To investigate PaRPI's ability to predict RNA-protein interactions for specific proteins across different cell lines, we obtained 62 shared RBP datasets from both the K562 and HepG2 cell lines and evaluated PaRPI's performance on this cross-cell prediction task. Specifically, we applied the PaRPI model, which was trained using data from the K562 cell line, to predict RNA-protein interaction profiles across 62

RBP datasets derived from the HepG2 cell line. Following this, we reversed the approach by using a PaRPI model trained on the HepG2 cell line to predict RNA-protein interactions within 62 RBP datasets obtained from the K562 cell line. We compared PaRPI's cross-cell prediction capability with other state-of-the-art deep learning methods, including HDRNet, PrismNet, PRIESSTESS, iDeep, DMSK, and GraphProt.

Figure 4 presents a heatmap of the dynamic prediction results, where we can observe that the proposed PaRPI outperforms other methods in both K562 and HepG2 predictions. In particular, when predicting RBP binding sites in K562 cells after training on HepG2 data, PaRPI achieved an average AUC of 0.84, outperforming the best-performing HDRNet by 5%. Similarly, when predicting RBP binding sites in HepG2 cells after training on K562 data, PaRPI achieved a 3% improvement over HDRNet's average AUC. PaRPI performs well in cross-cell line predictions, which may be attributed to its incorporation of precise protein representations, enabling the model to be trained on large-scale ensemble datasets spanning multiple proteins. This approach helps mitigate overfitting issues that other baseline methods encounter when relying on small, protein-specific and cell-line-specific datasets, thus enhancing the model's tolerance to variation. By training on extensive ensmble dataset, PaRPI is able to learn more generalized representations of RNA-protein interactions, allowing the model to not only perform well on specific cell lines but also make effective predictions across different cell lines. For example, it can capture shared structural and functional features within protein families, facilitating the transfer and application of knowledge across cell lines. This cross-cell line generalization capability makes PaRPI more accurate and reliable in predicting RNA-protein interactions.

Transfer predicting binding preference of novel RNA and RBPs

We further explored the generalization capability of PaRPI by collecting newly added binding datasets from the ENCODE database [34]. Specifically, we collected additional protein receptors ADAT1, DDX47, ELAVL1, MBNL1, METTL1, and RPS10 from the K562 cell line as independent test sets to evaluate its generalization performance. To assess the novelty of these proteins with respect to the training set, we further evaluated their sequence similarity to the training proteins using MMseqs2. Among these six proteins, ADAT1, MBNL1, METTL1, and RPS10 showed less than 30% sequence similarity to any proteins in the training set, indicating minimal homology with previously seen RBPs. Therefore, these four low-similarity proteins were used to more accurately assess the model's generalization ability on truly novel RBPs. Building upon this, to explore whether PaRPI is capable of predicting entirely novel RNA–protein interactions, we further integrated these four low-similarity RBP datasets and performed sequence clustering using MMseqs2 to remove redundancy. Sequences sharing over 80% identity with those in the training set were excluded. The resulting non-redundant dataset was used to evaluate PaRPI's performance in predicting entirely novel RNA–protein interaction events.

Since the binding data obtained were all positive samples, we chose ACC as the evaluation metric for the model's predictive performance, as shown in Figure 5a. The results show that PaRPI achieved good performance on four out of the six test datasets, with its generalization performance being relatively weaker on the remaining two. Notably, on the four low-similarity proteins (ADAT1, DDX47, METTL1, and MBNL1), three still achieved relatively high accuracy, while its performance was relatively weaker on MBNL1. PaRPI achieved an ACC of 0.67 on the entirely novel RNA–RBP dataset. Although this result is not highly optimal, it suggests that the model retains some capacity to generalize to previously unseen interaction pairs. Overall, these findings demonstrate that PaRPI is capable of capturing essential features from the dataset, indicating its potential in predicting RNA–protein interactions for novel RBPs.

We assessed the performance of HDRNet on DDX47 dataset. Given that HDRNet is trained on individual RBP datasets and lacks the ability to generalize to unseen proteins, we trained HDRNet

on other RBP datasets from the DDX family to predict the DDX47 protein. As shown in Figure 5b, when trained only on a single homologous protein dataset, HDRNet exhibited significantly lower ACC in predicting DDX47 binding sites. These results highlight PaRPI's ability to leverage shared structural and functional features within protein families, enabling generalization to unseen protein receptor related datasets.

Transfer predicting binding preference of RBPs cross species

To further assess the generalization capability of PaRPI in predicting RBP binding sites across different species, we collected RBFOX2 (RNA binding Fox-1 Homolog 2) binding data from the POSTAR database (GEO accession: GSE54794) [35]. We compared the performance of the PaRPI model, trained on the K562 cell line specific ensemble dataset, with the HDRNet model, trained on RBFOX2_K562 and RBFOX2_HepG2 datasets, in predicting RBFOX2 RNA-binding sites in *Mus musculus*. The results presented in Figure 5c demonstrate that PaRPI achieved an accuracy of 0.7386, whereas the accuracies of HDRNet models trained on single datasets were 0.6817 and 0.6535, respectively.

Additionally, we collected more *Mus musculus* binding data for the U2AF2 (GEO accession: GSM1378378) and SRRM4 (GEO accession: GSM1486175) proteins from the POSTAR database. We also obtained two additional binding datasets for *Drosophila melanogaster*, FMR1 (GEO accession: GSM927208) and RPS3 (GEO accession: GSM927213), from the modENCODE repository [36]. As illustrated in Figure 5d, the PaRPI model demonstrated robust performance across the U2AF2, SRRM4, and RPS3 datasets. However, the model's accuracy on the FMR1 dataset was notably lower, achieving an accuracy rate of only 0.3642. While certain conserved mechanisms underlie RBP-RNA interactions, species-specific variations can significantly impact the predictive power of computational models when applied across different organisms. If more RBP data from diverse species becomes available in the future, it would further enable the investigation of how species-specific differences influence RBP–RNA interactions.

Transfer predicting binding preference of RBPs cross protocol

Our approach integrates elements from multiple high-throughput sequencing methodologies, including CLIP-seq and eCLIP, to provide a comprehensive analysis platform. Our method enables the simultaneous assessment of RNA-protein interactions across different biological samples. As shown in the Figure 5e, we collected SLBP_K562 binding data obtained via RIP-chip [37] sequencing from the ENCODE database. Predictions were made using PaRPI, which was trained on datasets specific to the K562 cell line. The results demonstrate that PaRPI achieved higher accuracy compared to HDRNet, which was trained solely on the original SLBP_K562 dataset. This highlights PaRPI's stability and robustness in predicting experimental data derived from different sequencing techniques.

Identifying RNA motifs for RBP-RNA complexes

Similar to other deep learning approaches for RBP prediction, our model is capable of capturing key RNA sequence fragments. We analyzed the output from the final layer of GraphConv in PaRPI, where each nucleotide's score reflects the strength of its interaction with the protein within the RNA sequence. To minimize noise in the binding data, we retained only sequence samples with predicted binding probabilities greater than 0.8 to construct a comprehensive motif for each RBP. High-score fragments were defined as segments where the attention scores exceeded the average score of the sequence, and a uniform 6-mer sliding window was applied to capture continuous high-score fragments to identify binding patterns for each RBP.

As presented in Figure 6a, we identified key fragments from an RBP dataset, aligned them, and converted them into a position weight matrix (PWM) for comparison with known motifs retrieved from the mCross database. Notably, some of the PWMs were found to closely resemble the known motifs, which effectively highlights the potential of PaRPI in motif identification tasks. For instence, as demonstrated in the first line of Figure 6b, PaRPI successfully captured the poly-A binding motif with a preference for single-stranded structures, where the secondary structure diagram of RNA was generated using the ViennaRNA Web Services [38]. In addition, as depicted in Figure 6b, for a given RNA sequence within the KHDRBS1_K562 RBP dataset, we were able to obtain the attention score of each nucleotide through the GraphConv module in PaRPI. The corresponding key fragment was found to be close to the known binding motifs, further validating the model's effectiveness in accurately discerning and analyzing relevant sequence features.

Exploring the impact of disease-associated genetic variations by PaRPI

Single nucleotide polymorphisms (SNPs) are a form of genetic variation, referring to permanent changes in a single nucleotide within a DNA sequence. SNPs can significantly affect gene function, particularly the binding specificity of RBPs, as each RBP recognizes specific RNA sequences. These SNPs may alter the properties of the RNA, consequently affecting the protein receptor's affinity or specificity for RNA, leading to allele-specific functional changes that can result in a variety of severe diseases.

To investigate the impact of genetic variation on RBP binding sites, we retrieved a large number of SNP datasets from the dbSNP database [39] and selected disease-associated SNP data through the ClinVar database [40]. We employed PaRPI to scan RNA fragments before and after mutation, observing changes in RBP binding activity following allele conversion. As shown in Figure 7a, in the rs1053748363 sample, research indicates that the repeated nucleotide A in this fragment may contribute to Inborn Genetic Diseases (ClinVar accession: RCV002563641.2), Cerebroretinal Microangiopathy with Calcifications and Cysts 1 (ClinVar accession: RCV003388067.1), and Dyskeratosis Congenita (ClinVar accession: RCV001962938.4). We observed an increased likelihood of RBP binding in the localized region due to the insertion of a single nucleotide A. In the rs138332774 sample, the study reports that this allele conversion might be benign (ClinVar accession: RCV002148766.6). From the data, we also noted that this mutation did not lead to significant changes in the binding activity within the region.

Otherwise, we observed significant alterations in binding properties within high-attention regions in part of the transcripts due to substitutions, insertions, and deletions. for instance, in the K562 cell line as shown in the first row in Figure 7b, the insertion of a single nucleotide A in transcript ENST00000422207 caused a marked increase in binding within the local region; in contrast, the deletion of a nucleotide fragment in transcript ENST00000366607 in HEK293 cells (the second row of Figure 7b) resulted in a significant reduction of binding within the corresponding region. These insights reveal that genetic variations could potentially affect gene regulation. These findings demonstrate the biological interpretability of PaRPI in identifying potential genetic variations and elucidating the relationship between RBP binding sites and human diseases.

Visualization of learned representations of RNA and its interaction with RBP

Furthermore, we explored the contribution of different stages of the model to RNA-RBP interaction recognition. We extracted the feature matrices from each stage of PaRPI and projected them into 2D space using t-SNE [41] for better interpretation of the learning process.

Learned representations of RNA

PaRPI acquires semantically - rich RNA representations through learning. As shown in the visualization cases of GNL3_K562 presented in Figure 8a, the first subplot displays the original distribution of RNA features for positive and negative samples. The second subplot shows the distribution of RNA representation after being processed by graph sample and aggregation (GraphSAGE) [42] and Transformer. The final subplot presents the updated ultimately RNA representation after being processed by DPRBP. Evidently, it reveals that the clustering patterns of positive and negative samples become increasingly prominent, demonstrating the effectiveness of PaRPI in learning and refining RNA representations.

Learned interaction representations between protein and RNA

PaRPI also learns to obtain semantically - rich interaction representation between RNA and protein. As depicted in Figure 8b, the first subplot exhibits the original distribution of RNA and protein features for these five RBP datasets. The second subplot, on the other hand, presents the the distribution of the learned RNA-protein interaction representation after the application of the interaction module. It is evident that the model has successfully clustered the five different RBP datasets. Moreover, the distribution boundaries between positive and negative samples within each RBP dataset are clearly defined, indicating the model's effectiveness in discerning and organizing relevant features.

Ablation study of PaRPI

Evaluating the impact of model components

To evaluate the contribution of each component in the proposed deep network architecture, we conducted an ablation study on PaRPI. The experiments were carried out using 17 RBP datasets from the Hela cell line. Specifically, the following ablation configurations for the architectural design were applied: (1) we first removed the GraphConv graph network component, resulting in PaRPING; (2) we removed the Transformer component, resulting in PaRPINT; (3) we replaced the hierarchical pooling strategy of DPRBP with a max-pooling layer, resulting in PaRPIND; (4) we replaced the interaction module by simply concatenating the RNA and protein features, resulting in PaRPINC. As shown in Figure 9a, we introduced an additional evaluation metric, True Positive Rate (TPR), to more comprehensively illustrate the contribution of each component. The results indicate that PaRPI outperforms all ablation models, achieving optimal AUC, ACC, and TPR scores of 0.9378, 0.8614, and 0.8822, respectively.

Evaluating the impact of multiplex features

In the same time, we assessed the contribution of different RNA and protein features using another ablation study, again using the 17 RBP datasets from the Hela cell line. In this study: (1) we replaced the RNA features processed by BERT with one-hot encoding, resulting in PaRPIRO; (2) we removed the icSHAPE features, resulting in PaRPINIC; (3) during RNA graph construction, we removed the RNAplfold features and only used adjacent sequences for graph building, resulting in PaRPISeq; (4) during RNA graph construction, instead of using adjacent sequences, we randomly built edges to ensure graph connectivity, resulting in PaRPIRan; (5) we replaced the protein features processed by ESM-2 with one-hot encoding, resulting in PaRPIPO. As shown in Figure 9b, PaRPI consistently outperformed all the ablation models. Although PaRPIPO also achieved satisfactory results, it no longer possesses generalization capability for predictive tasks due to the use of one-hot encoding for protein representation. In addition, to further evaluate the potential of advanced RNA representation learning methods, we conducted comparative experiments using several pre-trained RNA language models, the results of which are provided in the Supplementary Information 3.

The robust performance of PaRPI in various ablation studies indicates that it is reasonable and effective in feature extraction and architectural design, contributing significantly to the model's ability to learn and generalize from data.

Discussion

RBPs play a pivotal role in post-transcriptional gene regulation by recognizing specific RNA targets and influencing their stability, localization, splicing, and translation. Recent advances, such as CLIP-seq technologies, have enabled high-throughput measurements of RBP binding patterns in vivo, providing critical insights into the interplay between RBPs and RNAs under diverse cellular conditions.Despite these advancements, existing computational methods are often constrained to predicting binding patterns for individual RBP datasets, showing limitations in addressing the diversity of RBPs. This poses a significant challenge in predicting interactions between RNA sequences and novel proteins.

To address this gap, we propose a graph network-based approach, PaRPI, for accurate prediction of multiple RBP binding profiles across cell lines. Specifically, we grouped all experimental datasets based on cell lines, integrating data across different experimental protocols and batches. The PaRPI model was then applied to this cross-protocol, cross-batch, and RBP-aware integrated datasets to model the binding preferences between RBPs and RNAs. A direct advantage of this approach is its capability to investigate the likelihood of RBP-RNA complex formation for RBPs that are not covered by the experimental datasets. To comprehensively evaluate PaRPI's performance, we compared it against seven state-of-the-art RBP prediction methods for identifying cell-specific protein-RNA interactions, and our experimental results revealed that PaRPI not only achieves superior predictive performance but also consistently outperforms all baseline methods across most benchmark datasets; moreover, PaRPI demonstrates generalizability in predicting novel RNA and RBPs with enhanced accuracy on datasets containing homologous proteins. To further assess its applicability across different cell lines, we trained PaRPI on RBP data from the K562 and HepG2 cell lines and tested its performance on independent datasets from other cell lines, where PaRPI maintained a competitive edge over six baseline approaches, showing superior performance on numerous datasets from these cell lines. In addition to this, to evaluate the generalization capability of PaRPI, we incorporated newly added RBP datasets from the ENCODE database, mouse RBP datasets from the POSTAR database, Drosophila melanogast RBP datasets from the modENCODE database, and RBP binding data collected through the RIP-chip pipeline. This comprehensive approach allowed us to rigorously assess the model's predictive performance across diverse datasets and experimental conditions. Despite the increased challenge, PaRPI achieved high accuracy, highlighting its precision and robustness compared to models trained on single RBP datasets. Furthermore, through analyzing high-attention segments identified by PaRPI, we gained insights into RNA-protein binding recognition mechanisms as well as the effects of mutations on RBP binding sites. By leveraging a large SNP datasets from dbSNP, we explored the impact of disease-associated genetic variations on RNA-protein binding, revealing potential connections between RBP binding sites and human diseases. Lastly, we extensive ablation studies and visualizations of feature characterizations were conducted to uncover the mechanisms behind PaRPI's outstanding performance, underscoring its effective feature extraction and architecture design, which significantly enhance the model's learning and generalization capabilities from the data.

The enhanced performance of our model can be attributed to three key aspects. Firstly, aggregating cell-line-specific datasets enables the model to effectively capture common patterns of protein-RNA interactions across diverse proteins. Secondly, the protein encoding methodology facilitates in-depth exploration of the principles governing RNA-protein interactions and enhances our understanding of

the bidirectional selection process between RBPs and RNAs. Lastly, the proposed network architecture demonstrates superior capability in capturing both the sequence and structural features of RNAs and proteins, thereby significantly improving the model's ability to discern complex interaction mechanisms and enhancing its overall predictive accuracy and robustness.

In summary, as a deep learning approach, PaRPI enables comprehensive training and prediction across multiple RBP datasets within cell lines, while accurately predicting interactions involving unseen proteins. Currently, our framework does not incorporate cell line-specific modeling. In future work, we plan to extend our methodology by integrating cell line-specific features into the model architecture, thereby enabling generalized prediction of RBP binding profiles across diverse cell lines. This enhancement will provide a more robust and comprehensive computational tool to support biological research.

Methods

Dataset construction

We adopted the dataset provided by HDRNet and these data have been deposited in [43]. The dataset gathered a total of 261 RBP binding site datasets from various cell lines across multiple databases. This collection includes 172 RBPs that were identified cross different protocols and batches in K562, HepG2, HEK293, HEK293T, HeLa, and H9 cell lines. Specifically, it comprises 61 RBPs from 65 CLIP-seq datasets provided by the POSTAR database [44], and 111 RBPs from 196 eCLIP datasets available through the ENCODE project [45].

For each RBP dataset, peaks identified from eCLIP and CLIP-seq data (e.g., from ENCODE and POSTAR3) were defined as high-confidence binding sites [46, 47], and the top 5,000 peaks ranked by signal enrichment were selected as positive samples. These regions represent reproducible crosslinking signals and are widely used as proxies for true binding sites in large-scale RBP studies [20, 21]. However, as they are not individually validated at biochemical or structural resolution, they should be interpreted as high-confidence putative binding sites rather than definitive interaction sites. Each binding site is fixed at a length of 101 nt. For regions shorter than 101 nt, the sequence is extended symmetrically from the center, while regions longer than 101 nt are truncated from both ends to maintain the fixed length. This uniform-length strategy provides a standardized format that facilitates efficient RBP model training and comparative analyses, while preserving both the central binding peak region and adjacent contextual information encoded in flanking sequences. To construct the negative set, 10,000 negative samples are generated by randomly selecting 101 nt sequences from the entire transcriptome while strictly excluding all genomic intervals overlapping with high-confidence binding peaks of the target RBP based on ENCODE and POSTAR3 data. The positive samples are labeled as 1' and the negative samples as 0'.

For the RBP datasets, we integrated the RBP datasets based on cell lines. The six PaRPI models are built and trained on these cell specific datasets. During the training process of each model, 20% of the samples are randomly selected as an independent test set for evaluating model performance. From the remaining 80% of the training samples, 20% are randomly chosen as a validation set for optimizing model parameters, while the remaining samples are used for training. To further ensure a more rigorous evaluation and reduce potential RNA sequence-based bias, we applied MMseqs2 to remove test RNA sequences with greater than 80% identity to any training RNA sequence, based on criteria established in prior works [48, 49]. This step minimized the risk of data leakage.

RNA graph construction

The constructed RNA graph can be represented as a graph G=(V,E,A), where:

- The node set $V = \{h_i \mid i = 1, 2, \ldots, n\}$ represents the nucleotide residue embeddings for each RNA nucleotide, with $|V| = n$ nodes;
- where V is a finite set of $|V| = n$ nodes,
- E represents the set of edges, which determined by the sequence adjacency and secondary structure interactions
- $A = \{A_{ij}\} \in \mathbb{R}^{n \times n}$ is the adjacency matrix encoding the connection weights between two nodes.

This graph representation captures the interactions and structural information between nucleotides in the RNA molecule, making it suitable for further analysis and prediction using graph-based neural network methods.

Nucleotide residue embeddings (nodes)

The representation of nucleotide residues consists primarily of two components: RNA Sequence Embedding and Secondary Structure.

1) RNA sequence embedding: The RNA sequence embeddings are generated using a BERT, which provides dynamic global contextual information for RNA sequences. The process involves preprocessing the RNA sequence by converting each nucleotide into a token. Then, the BERT model is used to map these tokens into a low-dimensional embedding space, where each nucleotide residue r_i is represented as a vector $BERT(r_i)$. This vector captures the contextual and syntactic features of the RNA sequence.

2) Secondary structure embedding: The secondary structure embedding of RNA under different cellular conditions is predicted using icSHAPE, a chemical modification-based method renowned for providing insights into the native conformation of RNA folding [24]. The icSHAPE structure score is computed based on reverse transcription signal counts and polymeric sequencing coverage at individual nucleotide positions. Here, each nucleotide residue's secondary structure information is represented by the pairing probability $icSHAPE(r_i, cell)$. Please refer to Supplementary Information 4 for the detailed calculation formulas.

Therefore, Each node represents an RNA nucleotide residue r_i (where $i = 1, 2, \ldots, n$), and its representation is given by:

$$\mathbf{h}_i = \mathrm{BERT}(r_i) \,||\, \mathrm{icSHAPE}(r_i, \mathrm{cell})$$

Where $||$ is used to represent vertical concatenation.

Interactions embeddings between nucleotides (edges)

The edges represent the interactions between nucleotide residues in the RNA sequence and structure. Two types of interactions determine the existence and weight of edges:

1) Sequence adjacency: An edge is present between nucleotide residues r_i and r_i if they are adjacent in the RNA sequence.

$$\mathrm{Adj}(r_i, r_j) = 1 \quad \text{if } |i - j| = 1$$

2) Secondary structure interaction: An edge is present between nucleotide residues and if they are paired in the RNA secondary structure. Here, the base-pairing probability are determined using

RNAplfold, which returns the secondary structure associated with the minimum free energy for the RNA sequence, using the thermodynamic nearest-neighbor approach. The returned structure is in bracket notation, that is a vector of dots and brackets. Each dot represents an unpaired base, while a pair of equally nested, opening and closing brackets represents a base pair [25].

$$\mathrm{Sec}(r_i, r_j) = p_{ij}^2$$

Therefore, The weight of the edge between nodes i and j is defined as the sum of the sequence adjacency and secondary structure interaction embeddings:

$$A_{ij} = \mathrm{Adj}(r_i, r_j) + \mathrm{Sec}(r_i, r_j)$$

where A_ij is the weight of the edge between nodes i and j, and the dimensionality is $N \times N$.

Protein representation

To capture the rich contextual and evolutionary information of proteins, we utilized ESM-2, a state-of-the-art protein language model, for feature extraction. ESM-2 leverages a transformer-based architecture pre-trained on a massive corpus of protein sequences to learn deep representations that encode both sequence-level and structural characteristics.

Specifically, for each protein sequence, we fed the raw amino acid sequence into ESM-2 to generate embeddings at the residue level. These embeddings encapsulate hierarchical information, ranging from local sequence patterns to long-range dependencies critical for understanding protein functionality. The residue-level embeddings were then aggregated into a fixed-length vector to represent the global characteristics of the protein.

The resulting protein embeddings were subsequently integrated into our model pipeline, enabling efficient and accurate modeling of RNA-protein interactions. This approach ensured that the protein features retained biologically meaningful information essential for downstream predictions.

Learning interaction module

After obtaining multi-source biological features from RNA graph and protein representation, we designed an end-to-end deep neural network named learning interaction module, as illustrated in Figure 1b. The framework highlights two critical components: the GraphConv module and the interaction module, which are responsible for learning RNA representations and capturing interaction-relevant protein information, respectively.

Learning RNA representation by Transformer enhanced GraphConv module

The GraphConv module have demonstrated exceptional capability in learning structured data representations by leveraging neighborhood information through message-passing mechanisms. However, capturing long-range dependencies and global context in complex graphs remains a challenge. To address this, we propose a Transformer-enhanced Graph Neural Network, which combines the localized feature aggregation strength of GNNs with the global attention mechanism of Transformers.

GraphConv Module: The GraphSAGE algorithm is an end-to-end network designed to learn hierarchical graph representations and node embeddings. In this study, we leverage GraphSAGE to aggregate information from neighboring nodes based on nucleotide interactions, facilitating the integration of multi-source features from RNA structure and sequence. This module ensures an effectively captures interactions between nucleotides, which is critical for RNA representation learning.

We define the graph convolution operation under the general "message-passing" framework as follows:

$$H^{(l)} = \mathrm{Gh}\left(A, H^{(l-1)}; \theta^{(l)}\right)$$

where $H^{(l)} \in R^{n \times d}$ represents the node embeddings after l steps of graph convolution, $\mathrm{Gh}(\cdot)$ denotes the graph convolution operation (also known as the message propagation function), $H^{(l-1)}$ is the output of the previous convolution layer, and $\theta^{(l)}$ are the learnable parameters. The initial node embeddings $H^{(0)}$ are derived from the features of RNA secondary structures.

Various implementations of the message propagation function have been proposed to perform graph convolution. Among these, the graph convolutional network (GCN) is widely used and is defined as:

$$H^{(l)} = \mathrm{ReLU}\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}X^{(l-1)}W^{(l)}\right)$$

where $\tilde{A} = A + I_N$, $\tilde{D}_{ij} = \sum_j \tilde{A}_{ij}$ and $W^{(l)}$ is a trainable matrix.$\tilde{D}$ is the degree matrix of $\tilde{A}$. The normalization term $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is introduced to address numerical instability and mitigate issues related to exploding or vanishing gradients during training.

Transformer module: To further process the node features extracted by GraphSAGE, we employed a Transformer module to capture both global dependencies and hierarchical representations. The Transformer employs a multi-head self-attention mechanism to model long-range relationships between nodes, which is defined as:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

for each position in an RNA sequence, the query matrix Q represents the information that position seeks to attend to; the key matrix K is used to compare against queries from other positions to establish relevance between them; and the value matrix V contains the actual information to be aggregated based on the attention weights computed,In this setup, the dimensionality of the key vectors d_k is set to 32 dimensions for each attention head, This mechanism allows the model to focus on relevant parts of the input sequence, enhancing the contextual understanding of the RNA sequence. The refined feature representations are then passed through a feed-forward network to capture additional structural information, thereby enhancing the RNA representation.

DPRBP module: After incorporating structural information and sequence information into GraphConv module and Transformer module, the limitation of sequence length still poses a challenge for accurate prediction. To address this issue, we referred to a protein-RNA binding predictor model based on deep pyramid convolutional neural networks (DPCNN) [50], known as DPRBP. This model is capable of capturing long-range nucleotide global dependencies. DPRBP captures long-range nucleotide dependencies by halving the sequence length through max-pooling after each layer, selecting maximum values to preserve key information while reducing complexity. It innovates upon the original DPCNN by utilizing two separate CNN modules to learn global context dependencies without sharing parameters, thereby enhancing its ability to capture global sequence information, it can be represented as follows:

$$t(\mathbf{x}) = F_{C_1}\left(F_{C_2}(\mathbf{x})\right)$$

$$r(\mathbf{x}) = \mathrm{ReLU}(x + t(\mathbf{x}))$$

where x represents the refined representation of the RNA sequence after it has been processed through the Transformer module.t denotes two consecutive distinct CNN blocks C_1 and C_2, which are capable

of dynamically learning the spatial relationships between sequences and structures. A residual connection r is added to enhance the perception of DPRBP in high-attention regions of sequences and structures.Subsequently, a max-pooling layer of size 3 with a step size of 2 is utilized to identify specific binding sites of RBPs:

$$g(\mathbf{x}) = \text{MaxPool}(r(\mathbf{x}))$$

Through this max-pooling layer, the sequence length is halved, retaining the most significant nucleotides. After $\log_2(L)$ layers, where L represents the length of the RNA representation, the sequence length is reduced to 1, generating a feature vector $\mathbf{f}$ enriched with structure-sequence space relationships and containing global information of the whole sequence:

$$\mathbf{f} = \text{ReLU}\left(\prod_{\log_2 l}(g_j(\mathbf{x}))\right)$$

Learning interactions representation between RNA and proteins by interaction module

In order to achieve information interaction between RNA representations and protein representations, inspired by the cross-attention mechanism [51], we adopt a similar approach to facilitate effective information exchange. Formally, the definition of the interaction module between RNA representations and protein representations is as follows:

$$h''^{(l)} = h'^{(l)} + \alpha_x^{(l)} W_v^{(l)} x^{(l)}$$

$$\alpha_x^{(l)} = \text{softmax}\left(\frac{\left(W_q^{(l)} h'^{(l)}\right)\left(W_k^{(l)} x^{(l)}\right)^T}{\sqrt{d}}\right)$$

Here, x represents the protein representation, $h''^{(l)}$ represents the updated RNA representations, with the rest of the meanings being similar to those in the neighbor-attention mechanism. In this work, we employ the multi-head cross-attention mechanism with multiple attention heads, so the outputs of multiple parallel cross-attention layers are concatenated together. The architecture and calculation of multi-head cross-attention is shown in Figure 1d.

RNA-protein binding probability predicted by MLP classifier

Finally, the MLP module processes the feature representation from the previous layers through a fully connected neural network and generates the final output prediction.

$$\text{MLP}(x) = W^{(L)}\text{ReLU}(W^{(L-1)}\text{ReLU}(\cdots\text{ReLU}(W^{(0)}x + b^{(0)})\cdots) + b^{(L-1)}) + b^{(L)}$$

where $W^{(l)}$ and $b^{(l)}$ represent the learnable weights and biases for layer l, x represents the refined representation of the RNA and protein representations after they have been processed through the interaction module and $\text{ReLU}(\cdot)$ is the Rectified Linear Unit activation function.

The model was optimized using the binary cross-entropy loss function, ensuring robust predictions of RNA-protein binding events.

$$\text{Loss}(\mathbf{P}, \mathbf{Y}) = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log p_i + (1 - y_i) \cdot \log(1 - p_i)] + \lambda\|W\|_2$$

where N represents the mini batch size, y_i denotes the ground truth label, and p_i is the predicted binding probability. $\lambda\|W\|_2^2$ corresponds to the L2 regularization applied to all model parameters, serving as a weight decay mechanism to mitigate overfitting during training. Furthermore, to comprehensively evaluate the model's performance, the AUC was employed as the primary metric, reflecting the model's ability to distinguish between binding and non-binding events effectively.

Parameter settings

The implementation details of PaRPI are as follows. For RNA sequences, dimensionality reduction was applied to the features extracted from BERT using 128 one-dimensional convolutional kernels with a kernel size of 1. To upsample RNA secondary structure features, 128 one-dimensional convolutional kernels with a kernel size of 3 were employed. These processed sequence and structural features were subsequently passed through a 4-layer GraphSAGE network and a single Transformer layer for feature enrichment. Further refinement was achieved using DPRBP, where each view incorporated one-dimensional convolution layers with kernel sizes of 1, 2, 3, 5, and 7. For protein sequences, features extracted from ESM-2 were reduced using 256 one-dimensional convolutional kernels with a kernel size of 1. Additionally, the processed features from both RNA and protein sequences are further analyzed through a MLP with 2 hidden layers, progressively reducing the feature dimensions from an input dimension of 256 to an output dimension of 1. All convolutional layers followed a Conv-BN-ReLU sequence without bias terms. To prevent overfitting, a dropout rate of 0.4 was applied. The model was trained end-to-end using the Adam optimizer [52] with a batch size of 512, an initial learning rate of 0.001, and a weight decay of 1×10^8. A linear scaled warm-up scheme was adopted to adjust the learning rate dynamically, mitigating optimization challenges during early training stages. A random search strategy was employed for hyperparameter tuning, and the detailed optimization process is provided in the Supplementary Information 5. The PaRPI model was implemented using Python v3.8 and PyTorch v2.2.1 [53]. Graph construction was performed using the Deep Graph Library (DGL) [54]. Training was conducted on two NVIDIA A100 GPUs, each with 60GB of memory. We trained the model with 200 epochs and employed early stopping if there was no improvement in validation performance for 20 epochs.

High attention region visualization

We extracted the output of the final layer of GraphSAGE in PaRPI, where the score of each nucleotide reflects the strength of interaction between each base in the RNA sequence and the protein. We then selected the maximum value for each label as the final attention score, resulting in an attention vector Att of length 99. Subsequently, we uniformly distributed the attention score of each label across the nucleotides as follows:

$$\text{Score}_i = \begin{cases} \text{Score}_i, & i = 1, \\ \frac{\text{Score}_{i-1}+\text{Score}_i}{2}, & i = 2, \\ \frac{\text{Score}_{i-1}+\text{Score}_{i-2}}{2}, & i = 100, \\ \text{Score}_{i-1}, & i = 101, \\ \frac{\text{Score}_i+\text{Score}_{i-1}+\text{Score}_{i-2}}{3}, & \text{else.} \end{cases}$$

Where Score represents the attention score of the k-th nucleotide, with k in the range [1, 101].

Statistical Analysis and Reproducibility

Detailed statistical tests for each analysis are specified in the corresponding figure legends. Sample data were obtained from public repositories. The sample size was not predetermined and represents the maximum number of samples available for each dataset. No data were excluded from the analyses. Our study does not involve experimental group assignments, nor does it require blinding for group allocation. All results are fully reproducible with the provided code, dependencies, and data resources.

Data availability

Unless otherwise stated, all data supporting the results of this study are available within the article and its supplementary materials. Specifically, the original data for all figures are provided in Supplementary Data 1. The 261 RBP binding sites datasets for cell lines across multiple databases have been deposited on `https://doi.org/10.5281/zenodo.14878562` [55].

Code availability

Our code is available on the GitHub repository `https://github.com/ljquanlab/PaRPI` or the Zenodo repository `https://doi.org/10.5281/zenodo.16794270` [56].

References

[1] Alfredo Castello and et al. Insights into rna biology from an atlas of mammalian mrna-binding proteins. *Cell*, page 1393–1406, Jun 2012.

[2] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human rna-binding proteins. *Nature Reviews Genetics*, page 829–845, Dec 2014.

[3] Fátima Gebauer and et al. Rna-binding proteins in human genetic disease. *Nature Reviews Genetics*, page 185–198, Mar 2021.

[4] Andrew D. Ellington and Jack W. Szostak. In vitro selection of rna molecules that bind specific ligands. *Nature*, page 818–822, Jul 1990.

[5] Debashish Ray and et al. Rapid and systematic analysis of the rna recognition specificities of rna-binding proteins. *Nature Biotechnology*, page 667–670, Jul 2009.

[6] Nicole Lambert and et al. Rna bind-n-seq: Quantitative assessment of the sequence and structural binding specificity of rna binding proteins. *BMC Genomics*, May 2014.

[7] Chris Gilbert and Jesper Q. Svejstrup. Rna immunoprecipitation for determining rna-protein associations in vivo. *Current Protocols in Molecular Biology*, 75(1), Jul 2006.

[8] Markus Hafner and et al. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, page 129–141, Apr 2010.

[9] Donny D. Licatalosi and et al. Hits-clip yields genome-wide insights into brain alternative rna processing. *Nature*, 456(7221):464–469, Nov 2008.

[10] Eric L Van Nostrand. Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nat. Methods*, 13:1, 2016.

[11] Eric L Van Nostrand and et al. Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nature Methods*, page 508–514, Jun 2016.

[12] Jernej Ule and et al. Clip identifies nova-regulated rna networks in the brain. *Science*, 302(5648):1212–1215, Nov 2003.

[13] Jun-Hao Li, Shun Liu, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. starbase v2.0: decoding mirna-cerna, mirna-ncrna and protein–rna interaction networks from large-scale clip-seq data. *Nucleic Acids Research*, page D92–D97, Jan 2014.

[14] Gerd Anders and et al. dorina: a database of rna interactions in post-transcriptional regulation. *Nucleic Acids Research*, page D180–D186, Jan 2012.

[15] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna- and rna-binding proteins by deep learning. *Nature Biotechnology*, page 831–838, Aug 2015.

[16] Ilan Ben-Bassat, Benny Chor, and Yaron Orenstein. A deep neural network approach for learning intrinsic protein-rna binding preferences. *Bioinformatics*, page i638–i646, Sep 2018.

[17] Xiaoyong Pan and Hong-Bin Shen. Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. Nov 2016.

[18] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. Graphprot: modeling binding preferences of rna-binding proteins. *Genome Biology*, page R17, Jan 2014.

[19] Kaitlin U Laverty and et al. Priesstess: interpretable, high-performing models of the sequence and structure preferences of rna-binding proteins. *Nucleic Acids Research*, 50(19):e111–e111, Oct 2022.

[20] Lei Sun and et al. Predicting dynamic cellular protein–rna interactions by deep learning using in vivo rna structures. *Cell research*, 31(5):495–516, 2021.

[21] Haoran Zhu and et al. Dynamic characterization and interpretation for protein-rna interactions across diverse cellular conditions using HDRNet. *Nature Communications*, 14(1):6824, 10 2023.

[22] Jacob Devlin and et al. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North*, Jan 2019.

[23] Zeming Lin and et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

[24] Ryan A Flynn and et al. Transcriptome-wide interrogation of rna secondary structure in living cells with icshape. *Nature Protocols*, page 273–290, Feb 2016.

[25] Ronny Lorenz and et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[26] Hui Li and et al. circrna-binding protein site prediction based on multi-view deep learning, subspace learning and multi-view classifier. *Briefings in Bioinformatics*, 23(1):bbab394, 2022.

[27] Xiaoyong Pan and Hong-Bin Shen. Rna-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18:1–14, 2017.

[28] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature biotechnology*, 33(8):831–838, 2015.

[29] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. Graphprot: modeling binding preferences of rna-binding proteins. *Genome biology*, 15:1–18, 2014.

[30] Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

[31] Damian Szklarczyk and et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, page D605–D612, Jan 2021.

[32] F Tamanini. Differential expression of fmr1, fxr1 and fxr2 proteins in human brain and testis. *Human Molecular Genetics*, 6(8):1315–1322, Aug 1997.

[33] Huijuan Feng and et al. Modeling rna-binding protein specificity in vivo by precisely registering protein-rna crosslink sites. *Molecular cell*, 74(6):1189–1204, 2019.

[34] Y. Luo and et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.*, 48(D1):D882–D889, Jan 2020.

[35] Yumin Zhu and et al. Postar2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Research*, 47(D1):D203–D211, Jan 2019.

[36] modENCODE Consortium, Sushmita Roy, Jason Ernst, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

[37] Jack D. Keene, Jordan M. Komisarow, and Matthew B Friedersdorf. Rip-chip: the isolation and identification of mrnas, micrornas and protein components of ribonucleoprotein complexes from cell extracts. *Nature protocols*, 1(1):302–307, 2006.

[38] P. Kerpedjiev, S. Hammer, and I. L. Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31(20):3377–3379, 2015.

[39] S. T. Sherry. dbsnp: the ncbi database of genetic variation. *Nucleic Acids Research*, page 308–311, Jan 2001.

[40] Melissa J. Landrum and et al. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46(D1):D1062–D1067, 2018.

[41] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008.

[42] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[43] Lei Sun and et al. Predicting dynamic cellular protein-rna interactions using deep learning and in vivo rna structure. May 2020.

[44] Wei Zhao and et al. Postar3: an updated platform for exploring post-transcriptional regulation coordinated by rna-binding proteins. *Nucleic Acids Research*, 50(D1):D287–D294, 2022.

[45] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

[46] Eric L Van Nostrand and et al. Robust transcriptome-wide discovery of rna-binding protein binding sites with enhanced clip (eclip). *Nature methods*, 13(6):508–514, 2016.

[47] Y. C. T. Yang and et al. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, 16:1–8, 2015.

[48] Kaili Wang, Renyi Zhou, Yifan Wu, and Min Li. Rlbind: a deep learning method to predict rna–ligand binding sites. *Briefings in Bioinformatics*, 24(1):bbac486, 2023.

[49] Nabila Shahnaz Khan and et al. Rna-nrd: a non-redundant rna structural dataset for benchmarking and functional analysis. *NAR Genomics and Bioinformatics*, 5(2):lqad040, 04 2023.

[50] Rie Johnson and Tong Zhang. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Jan 2017.

[51] Mozhdeh Gheini and et al. Cross-attention is all you need: Adapting pretrained transformers for machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[52] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[53] Adam Paszke and et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[54] Minjie Wang and et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*, 2019.

[55] Liangchen Peng and et al. Parpi predicts rna-protein interactions from cross-protocol and cross-batch rna-binding protein datasets[Data set]. Zenodo https://doi.org/10.5281/zenodo.14878562 (2025).

[56] Liangchen Peng and et al. Parpi predicts rna-protein interactions from cross-protocol and cross-batch rna-binding protein datasets[Code]. Zenodo https://doi.org/10.5281/zenodo.16794270 (2025).

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China Fund under Grants (62272335 to L.Q. and 62002251 to T.W.), in part by Jiangsu Colleges and Universities QingLan project (L.Q.), in part by University-Industry Collaborative Education Program (L.Q.), in part by Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions, and in part by the Collaborative Innovation Center of Novel Software Technology and Industrialization (L.Q., T.W., and Q.L.).

Author contributions

LC.P. was responsible for the development of the methodology, programming, performing experiments, and writing the initial draft of the manuscript. LJ.Q. was involved in conceptualizing the study, conducting investigations, establishing the methodology, and revising the manuscript. ZJ.Z. and Y.Z. provided assistance in model design and methodology development. B.Z., LX.C. and QF.C. contributed to the acquisition of additional datasets. LK.M., ZH.Z., and SJ.Z. handled project visualization and managed software. TF.W. undertook the review of the manuscript. Q.Lyu. contributed to the conceptualization of the work, offered computational support and guidance, and participated in the editing of the manuscript.

Competing interests

The authors declare that they have no competing financial interests.

Figure Caption

Figure 1. The architecture of PaRPI. a. The overall framework of PaRPI. After tokenization, RNA sequences are input into a multi-head self-attention layer to generate dynamic contextual embeddings. The icSHAPE pipeline is used to generate RNA secondary structure profiles, while RNAplfold provides predicted RNA secondary structure information. RNA features are then structured into a graph network. The input protein sequences are processed by the pre-trained ESM-2 model to extract protein features. **b.**Learning interactions module. RNA representations are learned through GraphConv and Transformer to capture both local and global information, with the DPRBP network module performing stepwise dimensionality reduction on sequence length. Interaction module facilitates the interaction between RNA sequence representations and protein representations. **c.**Graphconv module. Node spatial features consist of sequence adjacency edges and secondary structure features from RNAplfold. GraphSAGE is used to aggregate information from neighboring nodes. **d.**Interaction module. Learning the interactions between RNA representations and protein representations through element-wise multiplication. **e.**DPRBP module. A multi-layer feature selector module is developed to capture context-sensitive information by gradually reducing the dimensionality of the sequence.

Figure 2. PaRPI predicts RBP binding events more accurately than baseline methods. a. The Circos heatmap displays the AUC scores for PaRPI compared to other methods, including HDRNet, PrismNet, PRIESSTESS, DMSK, iDeep, GraphProt, and DeepBind(n = 261 original datasets). **b.**Violin plot of the overall AUC scores of PaRPI against those of other methods(n = 261 original datasets). **c.**The AUC scores of the original datasets, comparing performances between PaRPI and other methods. **d.**The box plots summarizing the overall ACC, F1 score, and MCC of the PaRPI method compared to other methods(n = 261 original datasets). **e.**The box plots summarizing the overall AUC, ACC, F1 score, and MCC of the PaRPI method compared to other methods(n = 261 de-redundant datasets) Source data are provided as a Source Data file.

Figure 3. Identifying Homologous Proteins within K562 Cell Line. a. Comparative analysis of PaRPI and baseline models across low, medium, and high protein similarity groups(n = 261 original datasets in first figure and n = 112 K562 cell-line datasets in second figure). **b.**The AUC scores for FMR1, FXR1 and FXR2 in K562 cell line. **c.**Similar motifs retrieved from the mCross database. Source data are provided as a Source Data file.

Figure 4. The heatmap of the predicted AUC scores for PaRPI and other baseline methods, including PrismNet, PRIESSTESS, iDeep, DMSK, GraphProt, and DeepBind, across the 62 shared original datasets from the K562 and HepG2 cell lines. Source data are provided as a Source Data file.

Figure 5. Generalization prediction of the PaRPI. a. PaRPI effectively predicts unseen RBP datasets. **b.**PaRPI performs better compared to HDRNet trained on a single RBP dataset. **c.**Comparison of PaRPI with HDRNet for predicting RBFOX2 RNA-binding sites in *Mus musculus*. **d.**PaRPI performance on both *Mus musculus* and *Drosophila melanogaster* RBP datasets. **e.**PaRPI demonstrates superior stability when predicting data cross protocol. Source data are provided as a Source Data file.

Figure 6. Integrative motifs and saliency maps derived from PaRPI. a. Compares the detected motif samples with known motifs retrieved from the mCross database. **b.**The significance map for KHDRBS1_K562 in a prediction task, where the bound regions are successfully identified and highlighted by PaRPI.

Figure 7. Mutation analysis reveals potential genetic impacts on disease mechanisms. a. Utilizing PaRPI to predict disease-associated SNP data and observe changes in RNA fragment binding properties. **b.**Utilizing PaRPI to predict alterations in binding properties within transcripts, thereby uncovering the biological interpretability of RBP binding sites in relation to human diseases.

Figure 8. The advantages of PaRPI a. Performance of PaRPI tasks using different features based on AUC, ACC and TPR. **b.**Performance of PaRPI tasks using different structural information based on AUC, ACC and TPR.

Figure 9. Visualization of PaRPI a. The t-SNE clustering results of PaRPI output features at different stages Using the GNL3_K562 dataset. **b.**The t-SNE clustering results of the interaction module output of PaRPI on five RBP datasets.