MultiRBP: Multi-task Neural Network for Protein–RNA Binding Prediction

Jonathan Karin*
School of Electrical and Computer
Engineering, Ben-Gurion University
of the Negev
Beer Sheva, Israel
karinjo@post.bgu.ac.il

Hagai Michel*
School of Electrical and Computer
Engineering, Ben-Gurion University
of the Negev
Beer Sheva, Israel
hagaimi@post.bgu.ac.il

Yaron Orenstein†
School of Electrical and Computer
Engineering, Ben-Gurion University
of the Negev
Beer Sheva, Israel
yaronore@bgu.ac.il

ABSTRACT

Protein–RNA binding plays vital roles in post-transcriptional gene regulation. High-throughput *in vitro* binding measurements were generated for more than 200 RNA-binding proteins, enabling the development of computational methods to predict binding to any RNA transcript of interest. In recent years, deep learning-based methods have been developed to predict RNA binding *in vitro* achieving state-of-the-art results. However, all methods train a single model per protein, under-utilizing the similarities in binding preferences shared by multiple RNA-binding proteins. In this work, we developed MultiRBP, a deep learning-based method to predict RNA binding of hundreds of proteins to a given RNA sequence. The innovation of MultiRBP is in its multi-task nature, i.e., predicting binding for hundreds of proteins at the same time. We trained MultiRBP on the RNAcompete dataset, the most comprehensive dataset of *in vitro* binding measurements. Our method outperformed extant methods in both *in vitro* and *in vivo* RNA-binding prediction. Our method achieved an average Pearson correlation of 0.692 ± 0.17 for *in vitro* binding prediction, and a median AUROC of 0.668 ± 0.09 for *in vivo* binding prediction. Moreover, by visualizing the learned binding preferences, MultiRBP provided more interpretable visualization than a single-task model. The code is publicly available at github.com/OrensteinLab/MultiRBP.

CCS CONCEPTS

• Computing methodologies → Supervised learning by regression; Neural networks; • Applied computing → Bioinformatics.

KEYWORDS

deep learning, neural networks, multi-task learning, RNA-binding proteins, protein–RNA binding, RNAcompete, eCLIP

*Both authors contributed equally to this research.

†Corresponding author.

This work is licensed under a Creative Commons Attribution International 4.0 License.
BCB '21, August 1–4, 2021, Gainesville, FL, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8450-6/21/08.
<https://doi.org/10.1145/3459930.3469525>

ACM Reference Format:

Jonathan Karin, Hagai Michel, and Yaron Orenstein. 2021. MultiRBP: Multi-task Neural Network for Protein–RNA Binding Prediction. In *12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '21)*, August 1–4, 2021, Gainesville, FL, USA. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3459930.3469525>

1 INTRODUCTION

RNA-binding proteins (RBPs) are key players in post-transcriptional regulation of gene expression. RBPs coordinate multiple processing and regulatory steps that are related to RNA translation [10]. Better understanding of the mechanisms by which RBPs function may provide essential tools for building models that explore regulatory processes and identify disease variants, e.g., exploring the regulatory mechanisms by which miRNA and RBP binding activity interfere in the development and progression of cancer diseases [21].

Along the years, many experimental techniques were developed to measure the binding of RBPs to RNA both *in vivo* and *in vitro*. eCLIP (and its variants) is an *in vivo* experimental technique that measures protein–RNA binding transcriptome-wide [20]. However, due to interference of various cellular processes, eCLIP experiments exhibit low signal-to-noise ratio. Therefore, the experimental output is a list of binding sites, which can be augmented by unbound RNAs to form a binary classification problem, i.e., predicting if an RNA segment is bound by a specific RBP given an RNA sequence. Moreover, experimental noise and the complex cellular environment prevent accurate learning of the protein–RNA binding preferences from such *in vivo* data [3].

To tackle these problems, *in vitro* techniques were developed to measure protein–RNA binding. RNAcompete is an *in vitro* experimental technique that measures the binding intensity of one RBP to more than 240,000 short synthetic RNA sequences (30–41nt-long) [15]. The data produced by RNAcompete naturally forms a regression problem, i.e., predicting the binding intensity of a specific RBP given an RNA sequence. The RNA sequences in the RNAcompete dataset are divided into two sets, Set A and Set B, which have similar uniform coverage of all 9-mers.

Many computational methods were developed to predict RNA binding based on RNAcompete data. The state of the art in *in vivo* binding prediction, DLPRB, introduced a convolutional neural network (CNN) architecture as well as a recurrent neural network (RNN) architecture [3]. Both the RNA sequence and structure are used in DLPRB, where RNA structure is represented as a probability vector of different RNA structural context (paired, hairpin loop,

inner loop, multi-loop, and external loop). The state of the art in *in vitro* prediction, ThermoNet, predicts both *in vitro* and *in vivo* binding by a CNN architecture [18]. The main innovation of ThermoNet is in using thermodynamic ensemble of RNA secondary structures. The main limitation of all the aforementioned methods is that they were trained to predict protein–RNA binding for a single protein at a time.

Multi-task learning was theoretically proven to improve prediction performance [5], and has been applied successfully to many machine-learning problems, including in the bioinformatics field. The coupling of multi-task learning with deep neural network has been applied in genomics since the early works of Basset and DeepSEA [9, 24]. Only recently, DeepRIPE predicted protein–RNA binding *in vivo* to many RBPs in parallel [6]. These works motivate the application of multi-tasking learning to RNAcompete data. As opposed to previous applications, where the task was to classify genomic sequences to bound or unbound, with almost all sequences being bound by only one protein, RNAcompete data includes a measured binding intensity for each RNA probe sequence out of more than 240,000 RNA probes. To the best of our knowledge, this is the first application of multi-task learning in bioinformatics to a regression problem with a fully labelled intensity vector.

In this work, we propose a new method called MultiRBP. MultiRBP is a multi-task model based on CNNs that is trained simultaneously on all 244 RNAcompete experiments. By predicting binding intensity for multiple RBPs in parallel, MultiRBP outperforms the state of the art in both *in vitro* and *in vivo* binding predictions. Our innovative approach relies on using combined information from similar RBPs that is captured by the multi-task neural network. For predicting binding intensities for all 244 experiments, the multi-task model has lower overall training runtime than a single-task model, as well as more interpretable visualization of RNA-binding preferences.

2 METHODS

2.1 RNAcompete data representation and pre-processing

The dataset we used is comprised of 244 experiments of short synthetic RNA sequences (30–41nt-long) (Figure 1A) [16]. MultiRBP receives as input an RNA sequence and predicted RNA secondary structure for that sequence. An RNA sequence of length L is a string of L characters over the nucleotide alphabet $\{A, C, G, U\}$. In MultiRBP an RNA sequence is represented as one-hot encoded matrix $N \in \{0, 1\}^{4 \times L}$, where each column represents a different nucleotide position in the sequence (Figure 2). In RNAcompete the length of each sequence is between 30 to 41 nucleotides. Thus, to obtain a fixed-length input matrix, every input matrix is zero-padded so that all matrices are 41nt-long.

We predicted the RNA secondary structure of each RNA sequence using a variant of RNAplfold (ViennaRNA Package [12]), which was also used in DLPRB. This variant predicts the probability of each nucleotide being in each one of the following structural contexts: hairpin-loop (H), inner-loop (I), multi-loop (M), and external loop (E). The probability of being in a paired (P) context is calculated by subtracting the sum of all other probabilities from 1. In MultiRBP the probabilities are represented as matrix $S \in \mathbb{R}^{5 \times L}$,

where each column represents the distribution of each nucleotide over the structural contexts. The matrix is zero-padded to the length of 41 and concatenated to the sequence matrix to produce input matrix $M \in \mathbb{R}^{9 \times 41}$.

Similar to previous computational methods (e.g., DeepBind, DLPRB, and ThermoNet), RNAcompete binding intensities greater than the 99.5 percentile were clamped to the 99.5 percentile in each experiment. Approximately 2% of the intensities in the dataset are nan values. Since the nan values are distributed among all the experiments, removal of those sequences will cause significant data loss (175,309 sequences, which are 73% of all the sequences). Instead, we replaced the nan values with the mean binding intensity of each experiment. While there are other approaches to replace nan values in the matrix, since the nan values are very few and uniformly distributed, we believe that the effect of choosing one approach over the other is negligible.

2.2 eCLIP data representation and pre-processing

The *in vivo* dataset, which we used only for testing, was generated by the eCLIP experimental technique [22]. This dataset includes 54 eCLIP experiments of RBPs that were also tested in the RNAcompete dataset (Figure 1B). The *in vitro* sequence length is between 30 to 41nt. The length of the *in vivo* data has much higher variance: average length of $378\text{nt} \pm 23\text{nt}$ (with some sequence lengths over 2000nt). We tested two approaches to overcome the length difference. Denote by L the length the model was trained on. The first approach used the L center nucleotides of each eCLIP peak (as done in RCK [14]). The second approach predicts binding to all possible L -long windows, and takes as the predicted binding intensity the maximum or the mean of the window-specific predictions. The former approach achieved better prediction performance and thus was used in the comparisons to other methods.

2.3 MultiRBP network architecture and training

MultiRBP network receives as input a $9 \times L$ matrix. The network is divided into two paths by implementing two parallel CNNs, one with multiple kernel sizes, and the other with one kernel size (Figure 3). The input is fed in both CNNs to a 1D-convolutional layer with filters of multiple sizes. ReLU activation function is applied on the outputs of the filters. A global max-pooling layer reduces the dimension of the output by selecting the maximum value. The outputs of both CNNs are fed into a fully connected layer with ReLU activation. The multi-kernel CNN path has an additional fully connected layer. The outputs of both paths are concatenated along with two residual dense paths of the multi-kernel CNN path: one from the max-pooling output, and the other from the first fully connected layer. The concatenated vector is fed into a final fully connected layer of 244 neurons with leaky ReLU activation function. The size of the output vector is 244, yielding 244 different binding intensities corresponding to 244 RNAcompete experiments. The filters length, use of ReLU activation function, and CNN architecture were inspired by DLBRP CNN network [3]. The parallel CNN structure was inspired by ThermoNet [18]. The inspiration for using residual dense paths came from computer vision models [7].

A

RNAcompete data example:

Probe Set	RNA sequence	RNCMPT00100	.....	RNCMPT00103
Set A	AGAAGGCACCAAGAAGC	0.449584	.....	-0.067155
Set A	AGAGGUCGGGAGGUAUGA	1.386679	.....	-0.997684
....	....	....	....	....
Set B	AGAAAAUUCGAUACGGCA	0.064375	.....	1.000528
Set B	AGGGCGGGCGAUACGGACA	0.859564	.....	0.161771
Set B	AGGUUAUAUCUAAUGCA	-0.645497	.....	-1.609274

B

eCLIP data example for one experiment:

RNA sequence	Binding
AUGAGGCACUCAAGAAGC	1
AUGAUGGUCGGGAGGAU	0
....	....
AGAAUUUUCGAUACGGCA	1
AUGGCGGGCGAACGGACA	0
ACCCGUUAUUCUAAUGCA	0

Figure 1: Illustration of the experimental datasets used in this study. (A) Illustration of the RNAcompete dataset. More than 240,000 RNA sequences are assigned binding intensities to 244 different experiments. (B) Illustration of the eCLIP dataset. Each experiment reports a set of bound RNA segments (labelled 1). This set is augmented by a set of nearby unbound RNA segments (labelled 0).

Figure 2: Encoding of RNAcompete data. An RNA sequence of length L is one-hot encoded and zero-padded to the length of 41 (the maximum length of an RNAcompete sequence). A probability matrix is calculated by a variant of RNAplfold. The one-hot encoding and structure matrix are concatenated to form a 9×41 input to MultiRBP.

MultiRBP hyper-parameters include layer sizes, number of convolutional filters, and dropout probability (Supplementary Table S1). The hyper-parameters calibration method we used is based on the method that was used in DeepBind [1]. More specifically, the training set included 100,000 sequences chosen randomly from Set A, and the remaining ~20,000 from Set A were used as the validation set. We randomly sampled 120 complete sets of hyper-parameters. For each set of hyper-parameters, a model was trained on the training set, and the one achieving the best results on the validation set was chosen. The final model was trained on all sequences of Set A with the chosen hyper-parameters.

MultiRBP parameters were learned using the Adam optimization algorithm with the log of hyperbolic cosine loss function (Equation 1), as it achieved the best results when compared with other loss functions (Supplementary Table S2):

$$L(y, y^p) = \sum_{i=1}^n \sum_{j=1}^{244} \log(\cosh(y_{i,j}^p - y_{i,j})) \quad (1)$$

where y and y^p are the measured and predicted binding intensities, respectively. Indices i, j denote experiment j of data point i in one training batch. We decided to weigh all experiments equally in the loss function since we used the normalized RNAcompete dataset

Figure 3: MultiRBP network architecture. The input is fed into two parallel CNN paths, one with multiple kernel sizes and the other with one kernel size, followed by fully connected layers. Outputs from all fully connected layers and the output of the global max-pooling layer of the multi-kernel CNN path are concatenated and fed to the final fully connected layer. 244 binding intensities are predicted for each input sequence.

(as in previous studies), which brings all experiment intensities to be on the same scale [16].

We trained MultiRBP on Intel® oneAPI devcloud using Intel® Xeon® processor with 24 cores (48 hyper-threads) and 196GB RAM. Runtime was 2 hours and 47 minutes with maximum memory usage of 15.3GB for training on the *in vitro* RNAcompete dataset with inclusion of RNA secondary structure information in the input.

2.4 Prediction performance evaluation

To evaluate the performance in *in vitro* binding prediction, we used the comprehensive dataset of RNAcompete, which includes 244 experiments. Each RNAcompete experiment contains the binding intensities between a single RBP and more than 240,000 different RNA sequences (Figure 1A). We trained our model on half of the RNA sequences (Set A) and tested it on the other half (Set B). Sets A and B were originally defined by the developers of RNAcompete to serve as training and test sets with similar uniform coverage of all 9-mers. We gauged prediction performance by Pearson correlation of predicted and measured intensities. Results of extant methods were downloaded from the DLPRB, the ThermoNet, and the RDense studies [3, 11, 18].

We formulated the *in vivo* binding prediction problem as binary classification. We used AUROC (Area Under the Receiver Operating Characteristics), a conventional measure of binary classification performance on balanced data sets [4]. We trained MultiRBP on

the *in vitro* data with extended input length L , where L is greater or equal than the maximum *in vitro* sequence length. Following previous studies that showed that RNA structure information does not improve *in vivo* binding prediction [1], RNA secondary structure information was not used. The trained model was tested on the eCLIP dataset [22]. The overlap with RNAcompete dataset covers 21 RBPs, tested in 36 RNAcompete experiments and 54 eCLIP experiments, resulting in 94 experimental pairs testing the same RBP. Results of extant methods were downloaded from the DLPRB, and the RDense studies [3, 22]. ThermoNet *in vitro* models were not available and were not evaluated in *in vivo* binding prediction as part of the ThermoNet study.

2.5 Single-task model

To gauge the benefit of MultiRBP compared to a single-task model, we developed a single-task model with the same architecture as MultiRBP. A similar hyper-parameters calibration procedure was performed, and 10 different hyper-parameters sets were sampled for every experiment (only 10 sets instead of 120 due to runtime limitation). The single-task model was trained and tested in the same manner as MultiRBP. The average training runtime of the single-task model was 6 minutes and 47 seconds. The average memory usage by the training process was 1.03GB. As evidence to the effectiveness of the hyper-parameters search of the single-task model, the performance of the single-task model was on par with

the performance of DLPRB-CNN (average Pearson correlation 0.605, p -value=0.563, Wilcoxon rank-sum test). We hypothesize that this is due to the fact that MultiRBP’s architecture, was inspired by DLPRB-CNN architecture. We did not develop multi-task versions of competing methods due to runtime and technical reasons (failures in data preparation).

2.6 MultiRBP visualization

To explore what a trained MultiRBP model learned about the RNA-binding preferences, we used the Integrated Gradients (IG) method [19]. IG aims to attribute the prediction of a deep neural network to its input features. The main idea behind IG is to compute the effect of a small change of the input on the output. IG computes the impact of each one of the input elements (nucleotides in the case of an RNA-binding site) on each one of the output elements (binding intensity prediction for different experiments). For each RNAcompete experiment, we applied IG with an all-zeros matrix baseline to the sequence for which MultiRBP predicted the highest binding intensity. We then plotted sequence logos based on IG output using “Logomaker” Python package.

3 RESULTS

3.1 Correlation between different experiments

Multi-task learning is highly desirable when true labels are available for many instances over the same pool of data points. Indeed, each RBP in the RNAcompete dataset is tested against the same pool of 241,357 RNA sequences. Thus, to motivate our study, we first computed the correlation between different RNAcompete experiments. The correlation can indicate how much combined information is shared among different experiments and RBPs. Figure 4 shows the correlation between binding intensity vectors of RNAcompete experiments, reordered by hierarchical clustering [2]. Several noticeable clusters (squares on the diagonal) demonstrate correlation within groups of different experiments.

3.2 *In vitro* binding prediction performance

To evaluate the performance in *in vitro* binding prediction, we used the comprehensive dataset of RNAcompete. We trained our model on half of the RNA sequences (Set A) and tested it on the other half (Set B). Performance was gauged using Pearson correlation of predicted and measured intensities.

MultiRBP achieved significantly better results than extant methods for the task of *in vitro* binding prediction (Figure 5A). MultiRBP with RNA secondary structure information achieved an average Pearson correlation of 0.692 ± 0.17 , which is significantly higher than the current state of the art ThermoNet, which achieved an average of 0.671 ± 0.17 (p -value= $3.87 \cdot 10^{-24}$, Wilcoxon rank-sum test). MultiRBP outperformed ThermoNet in 177 out of 244 RNAcompete experiments (Figure 5B). MultiRBP without RNA secondary structure information achieved an average Pearson correlation of 0.686 ± 0.17 , which is also significantly higher than ThermoNet (p -value= $1.64 \cdot 10^{-15}$, Wilcoxon rank-sum test). For complete results, see Supplementary Table S3.

We wondered whether the improvement in prediction performance over ThermoNet was associated with specific RBP families. For this aim, we analyzed the performance gain in each of 7

Figure 4: Correlation matrix of RNAcompete experiments. The matrix was reordered using hierarchical clustering. The yellow squares on the diagonal represent clusters of RBPs with shared binding preferences.

RBP clusters that were clearly defined in the original RNAcompete compendium study [16]. The original study clustered all RBPs by amino acid sequence identity, and focused on these 7 as sharing similar or almost identical binding motifs. The improvement of MultiRBP over ThermoNet in 5 out of these 7 clusters was much higher than the total mean improvement (Figure 5C). For example, the mean advantage in cluster 3 which contains RBPs ASD-1/FOX-1/A2BP1/RBFOX1, which share the consensus binding site GCAUG, is 0.088 compared to the mean advantage over all RBPs, which is 0.021. We speculate that one of the reasons for the great performance of MultiRBP is the ability to learn shared binding information between similar RBPs. For complete results, see Supplementary Table S4.

3.3 *In vivo* binding prediction performance

To evaluate the performance in *in vivo* binding prediction, we used the eCLIP dataset [22]. MultiRBP and competing methods were trained on the *in vitro* data, i.e., RNAcompete, and tested on the *in vivo* data. We formulated the *in vivo* binding prediction as a binary classification problem. We used AUROC to gauge binary classification performance.

MultiRBP achieved state-of-the-art performance in the task of *in vivo* binding prediction (Figure 6A). MultiRBP (without RNA secondary structure information) achieved a median AUROC of 0.668 ± 0.09 , which is better, albeit not statistically significant, than the state of the art DLPRB (without RNA secondary structure as well), which achieved a median AUROC of 0.657 ± 0.10 (p -value=0.347, Wilcoxon rank-sum test, Figure 6B). MultiRBP trained with the secondary structure information achieved a median AUROC of 0.641. A possible explanation for this degradation in performance (also

Figure 5: *In vitro* binding prediction performance. (A) Comparison between MultiRBP and extant methods for the task of predicting *in vitro* RNA binding. MultiRBP outperforms all extant methods for the task. (B) Comparison between MultiRBP and the state of the art ThermoNet for each RNAcompete experiment. MultiRBP performs significantly better than ThermoNet in this task. (C) Comparison between the difference between MultiRBP and ThermoNet performance over 7 different RBP clusters. The clusters were taken from the RNAcompete compendium study [16].

observed by DLPRB developers) is that RNA structure prediction is less accurate *in vivo* compared to RNA structure prediction *in vitro*. Collectively, this suggests that the learned structural preferences may not improve binding prediction *in vivo*. For complete results, see Supplementary Table S5.

3.4 Handling variable-length sequences in *in vivo* binding prediction

The average sequence length in eCLIP data is 378nt, while in RNAcompete data the maximum length is 41nt. We compared two main approaches to handle variable-length sequences in binding prediction. The first approach trains a model on L -long zero-padded sequences, and then uses the L center nucleotides of each eCLIP peak (Figure 7A). The second approach predicts binding to all possible L -long windows, and takes maximum or mean of the predictions (Figure 7B).

Increasing the input size was found to be much more beneficial for binding prediction than predicting binding to all possible windows. Taking 75nt from the center of each eCLIP peak was the most effective method (Figure 7C). Thus, we used this method for any *in vivo* binding comparison reported in this study. We speculate the prediction performance declined as we increased the input size as the binding site is most likely in the center of the sequence due to eCLIP experimental protocol [22]. Therefore, increasing the input size likely adds more noise than signal. A possible explanation for the advantage of taking only the center nucleotides over predicting binding for all windows is that predicting binding for all windows does not take into account the position of the window in the RNA sequence. Thus, the all-windows approach is more sensitive to noise along the whole sequence. A solution might be to assign different weights to different parts of the sequence leading to more accurate predictions, which might be achieved by an additional neural network. For complete results, see Supplementary Table S6.

Figure 6: *In vivo* binding prediction performance. (A) Comparison between MultiRBP and extant methods trained on RNAcompete in predicting *in vivo* binding. MultiRBP outperforms all extant methods for the task. (B) Comparison between MultiRBP and DLPRB for each pair of RNAcompete and eCLIP experiments testing the same RBP. MultiRBP is on par with DLPRB-CNN in this task.

Figure 7: Approaches to handle length differences between *in vitro* and *in vivo* data. (A) The first method uses the L center nucleotides of each eCLIP peak. (B) The second method predicts binding to all possible L -long windows, and takes maximum or mean of them. (C) Comparison between different methods for handling different sequence lengths in *in vivo* binding prediction.

3.5 Comparison of MultiRBP and a single-task model

To gauge the benefit of MultiRBP compared to a single-task model, we developed a single-task model with the same architecture as MultiRBP. Results show that MultiRBP benefits from learning on all RBPs together. The single-task model achieved an average Pearson correlation between the measurements and the predictions of *in vitro* binding intensities of 0.610 ± 0.16 compared to 0.692 ± 0.17 obtained by MultiRBP ($p\text{-value} = 1.22 \cdot 10^{-48}$, Wilcoxon rank-sum test). MultiRBP outperformed the single-task model in 238 out of 244 of RNAcompete experiments (Figure 8A). The median AUROC achieved by the single-task model in *in vivo* binding prediction

is 0.637 ± 0.09 compared to MultiRBP, which achieved a median AUROC of 0.668 ± 0.09 ($p\text{-value} = 1.01 \cdot 10^{-3}$, Wilcoxon rank-sum test) (Figure 8B). This comparison supports the conclusion that MultiRBP significantly improves prediction performance for both *in vitro* data and *in vivo* data compared to a single-task model.

3.6 Visualization and interpretability

We applied Integrated Gradients on multiple sequences to interrogate what the model learned as well as compare it to visualizations derived from a single-task model. For each RNAcompete experiment, we used the sequence for which MultiRBP predicted the highest binding intensity. We plotted the derived attribution maps

Figure 8: Comparison of MultiRBP and a single-task model. (A) Comparison between MultiRBP and a single-task model in *in vitro* binding prediction. (B) Comparison between MultiRBP and a single-task model in *in vivo* binding prediction. In both tasks, MultiRBP outperforms the single-task model significantly.

as sequence logos, where the importance of each input nucleotides is represented by its height. Positive or negative height represents positive or negative influence on the binding intensity.

We found that for some binding sites MultiRBP identified the binding site in a precise manner, i.e., high influence is associated with the nucleotides of the binding site whereas no influence is associated with the flanking nucleotides compared to a single-task model which failed to identify the binding site. For example, in Figure 9A for PABPC1 RBP, the known binding site ARAAAAM is identified only by MultiRBP. In Figure 9B for eIF-2 α RBP, which has the known binding site WCAUGK, MultiRBP identified the binding site more precisely than the single-task model. Another example is shown in Figure 9C for CG17838 RBP with the binding site BAAAUUD. The known binding sites were taken from CISBP-RNA database [16]. Thus, we conclude that MultiRBP gains improved interpretability compared to a single-task model thanks to its multi-task nature (without any additional enhancements to its interpretability). For visualization of all 244 experiments, see Supplementary Figure S1.

4 DISCUSSION

In this study, we developed MultiRBP, a new deep learning-based method for predicting protein–RNA binding. MultiRBP utilizes the shared binding preferences of multiple RBPs, and outputs the binding predictions of 244 RNAcompete experiments in parallel. We trained MultiRBP on all RNAcompete experiments simultaneously instead of training a neural network on each experiment separately. As such, we developed MultiRBP based on previous single-task state-of-the-art methods with architecture modifications that increased the number of model parameters compared to previous single-task models. The improvement of MultiRBP was significant in predicting *in vitro* binding compared to extant methods for the task. In predicting *in vivo* binding, the achieved results were better than the current state of art, albeit without statistical significance.

Comparison Between MultiRBP and a single task model using Integrated Gradients

Figure 9: Comparison between MultiRBP and a single-task model in interpretability via Integrated Gradients for important features visualization. Sub-panels A and B show more precise identification of known binding sites by MultiRBP. Sub-panel C shows a binding site that only MultiRBP identified.

There is a significant improvement in MultiRBP compared to a single-task model. A possible explanation is the fact that in the RNA-compete dataset there are multiple experiments for the same RBP. In addition, there are shared binding preferences among different RBPs [8] (Figures 4 and 5C). We speculate that MultiRBP identifies those binding sites more precisely due to increased shared information. We hope that in the future, when additional RNAcompete experimental datasets will be produced, MultiRBP will be trained on them to improve its prediction performance. Since MultiRBP's training runtime and memory usage is linear in the size of the data, such an extension is easily feasible.

We identify two major limitations in our work. First, since MultiRBP is trained on more *in vitro* data than a single-task model, we speculate that MultiRBP overfits the RNAcompete dataset more than any of the single-task models. This limitation may explain why the improvement of the prediction score is smaller in the *in vivo* binding prediction task. Second, the utilization of MultiRBP for binding prediction is limited to RBPs that were tested in an RNAcompete experiment. In order to use a similar network for other RBPs that have available experimental data, there is a need for a non-trivial pre-processing and standardization of datasets derived from different experimental techniques. In addition, the neural network architecture will have to be modified, as well as re-optimized and re-trained on such an augmented dataset.

There are several potential extensions to this work. First, we chose to represent RNA structure by structural context probabilities, but there are many other ways. In the future, we may use different methods to predict and represent RNA secondary structure, e.g., as graphs as in the GraphProt study [13]. Moreover, we may use experimentally measured RNA structure to improve binding prediction. This approach may even lead to improved *in vivo* binding prediction [14]. Second, instead of training MultiRBP over all RNAcompete experiments, we may train several different models. Each model will be trained on a different subset of the partition of all experiments. The partition can be done according to the correlation clusters (Figure 4) or according to known protein families [2, 15]. Third, we plan to learn predictive motifs by applying TF-MoDisco to cluster similar patterns in all sequences generated by Integrated Gradients [17]. Last, another extension to our work could be a method to find the most likely RBP to bind a given RNA sequence, as was solved for transcription factors binding DNA in BindSpace [23].

ACKNOWLEDGMENTS

We gratefully acknowledge the support of Intel Corporation for giving access to the Intel® AI DevCloud platform used for part of this work.

REFERENCES

- [1] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology* 33, 8 (2015), 831–838.
- [2] Michael Behrisch, Benjamin Bach, Nathalie Henry Riche, Tobias Schreck, and Jean-Daniel Fekete. 2016. Matrix reordering methods for table and network visualization. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 693–716.
- [3] Ilan Ben-Bassat, Benny Chor, and Yaron Orenstein. 2018. A deep neural network approach for learning intrinsic protein–RNA binding preferences. *Bioinformatics* 34, 17 (2018), i638–i646.
- [4] Andrew P Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145–1159.
- [5] Rich Caruana. 1997. Multitask learning. *Machine Learning* 28, 1 (1997), 41–75.
- [6] Mahsa Ghanbari and Uwe Ohler. 2020. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Research* 30, 2 (2020), 214–226.
- [7] Liqin Huang, Jianjia Zhang, Yifan Zuo, and Qiang Wu. 2019. Pyramid-structured depth map super-resolution based on deep dense-residual network. *IEEE Signal Processing Letters* 26, 12 (2019), 1723–1727.
- [8] Hilal Kazan, Debashish Ray, Esther T Chan, Timothy R Hughes, and Quaid Morris. 2010. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Computational Biology* 6, 7 (2010), e1000832.
- [9] David R Kelley, Jasper Snoek, and John L Rinn. 2016. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research* 26, 7 (2016), 990–999.
- [10] Julian König, Kathi Zarnack, Nicholas M Luscombe, and Jernej Ule. 2012. Protein–RNA interactions: new genomic technologies and perspectives. *Nature Reviews Genetics* 13, 2 (2012), 77–83.
- [11] Zhong Li, Jiapeng Zhu, Xiaojiang Xu, and Yuhua Yao. 2019. RDense: a protein–RNA binding prediction model based on bidirectional recurrent neural network and densely connected convolutional networks. *IEEE Access* 8 (2019), 14588–14605.
- [12] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdisen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. 2011. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 6, 1 (2011), 1–14.
- [13] Daniel Maticzka, Sita J Lange, Fabrizio Costa, and Rolf Backofen. 2014. GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology* 15, 1 (2014), 1–18.
- [14] Yaron Orenstein, Yuhao Wang, and Bonnie Berger. 2016. RCK: accurate and efficient inference of sequence- and structure-based protein–RNA binding models from RNAcompete data. *Bioinformatics* 32, 12 (2016), i351–i359.
- [15] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Pena Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. 2009. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology* 27, 7 (2009), 667–670.
- [16] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499, 7457 (2013), 172–177.
- [17] Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, and Anshul Kundaje. 2018. Technical note on transcription factor motif discovery from importance scores (TF-MoDisco) version 0.5. 6.5. *arXiv preprint arXiv:1811.00416* (2018).
- [18] Yufeng Su, Yunan Luo, Xiaoming Zhao, Yang Liu, and Jian Peng. 2019. Integrating thermodynamic and sequence contexts improves protein–RNA binding prediction. *PLoS Computational Biology* 15, 9 (2019), e1007283.
- [19] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*. PMLR, 3319–3328.
- [20] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaž Ule, and Robert B Darnell. 2003. CLIP identifies Nova-regulated RNA networks in the brain. *Science* 302, 5648 (2003), 1212–1215.
- [21] Marieke Van Kouwenhove, Martijn Kedde, and Reuven Agami. 2011. MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nature Reviews Cancer* 11, 9 (2011), 644–656.
- [22] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, et al. 2016. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods* 13, 6 (2016), 508–514.
- [23] Han Yuan, Meghana Kshirsagar, Lee Zamparo, Yuheng Lu, and Christina S Leslie. 2019. BindSpace decodes transcription factor binding signals by large-scale sequence embedding. *Nature Methods* 16, 9 (2019), 858–861.
- [24] Jian Zhou and Olga G Troyanskaya. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods* 12, 10 (2015), 931–934.

A ONLINE RESOURCES

The code and supplementary tables and figures are publicly available at github.com/OrensteinLab/MultiRBP.