cDeepbind: A context sensitive deep learning model of RNA-protein binding

Shreshth Gandhi [1,4] Leo J. Lee [1,2] Andrew Delong [1] David Duvenaud[1,3] Brendan J. Frey[1,2,3,4]

[1]Department of Electrical and Computer Engineering [2]Donnely Centre for Cellular and Biomolecular Research

[3]Department of Computer Science, University of Toronto [4]Deep Genomics Inc.

Introduction

- RNA structure has a contextual effect on RBP binding. Local secondary structure restricts access to a large subset of sequence motifs that would otherwise be bound.
- Many RBPs share similar sequence motifs and interact by forming complexes or competing for the same binding sites.

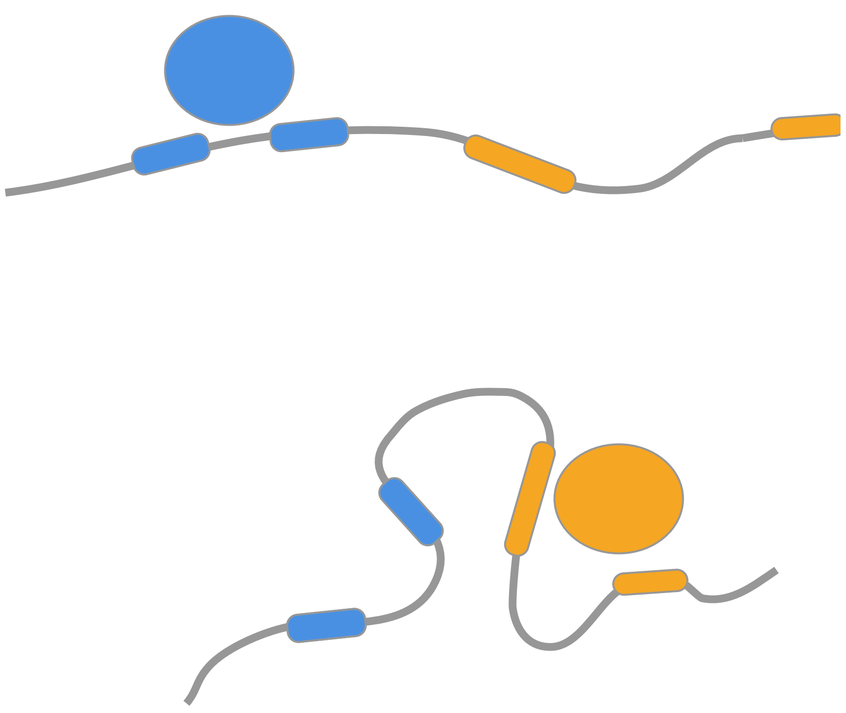


Figure 1. Local structural context can switch RBP binding even in the presence of the same sequence motifs

Key contributions

Inspired by previous approaches, our modelling framework aims to achieve the following advancements :

- **Multitask learning** Jointly predict binding intensities for multiple RBPs (244 probes in RNAcompete)
- **Graph embedding** A fixed size structure encoding that preserves position information from the base pairing matrix.
- **Gated SE-Resnet** Faster alternative to recurrent networks with similar expressive strength.

Preprint and Code

[1] Shreshth Gandhi, Leo J Lee, Andrew Delong, David Duvenaud, and Brendan Frey.
cdeepbind: A context sensitive deep learning model of rna-protein binding.
bioRxiv, page 345140, 2018.

Methods

Model Architecture

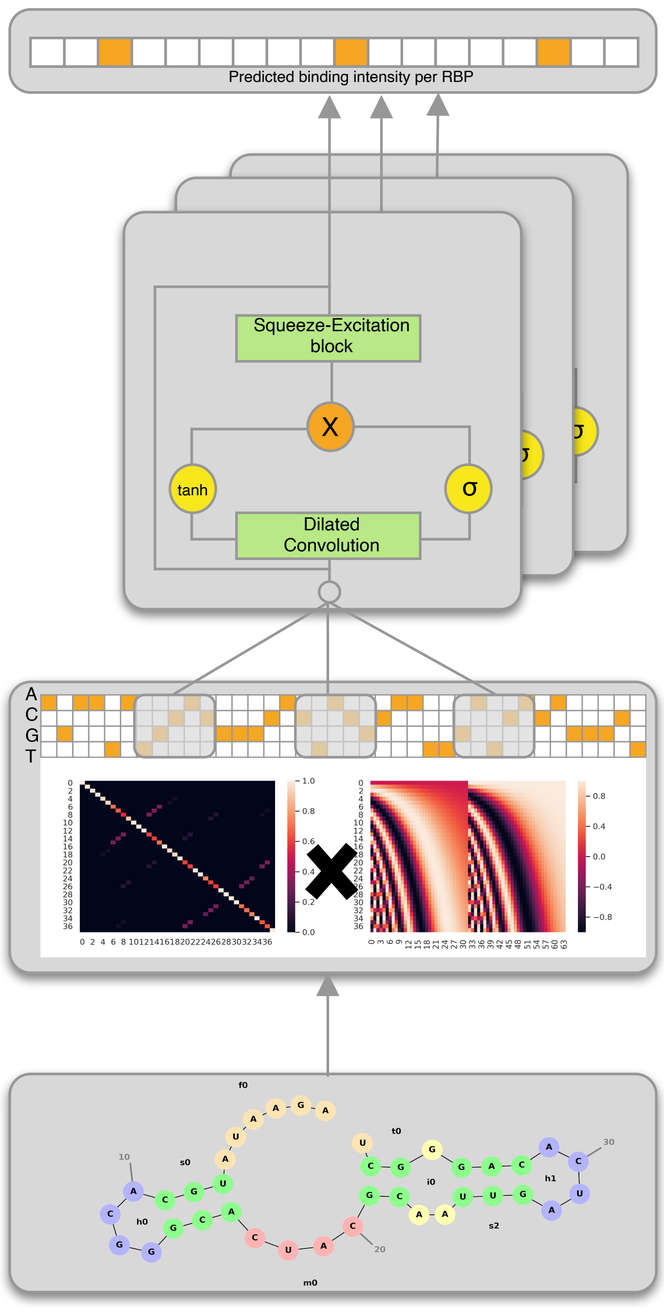


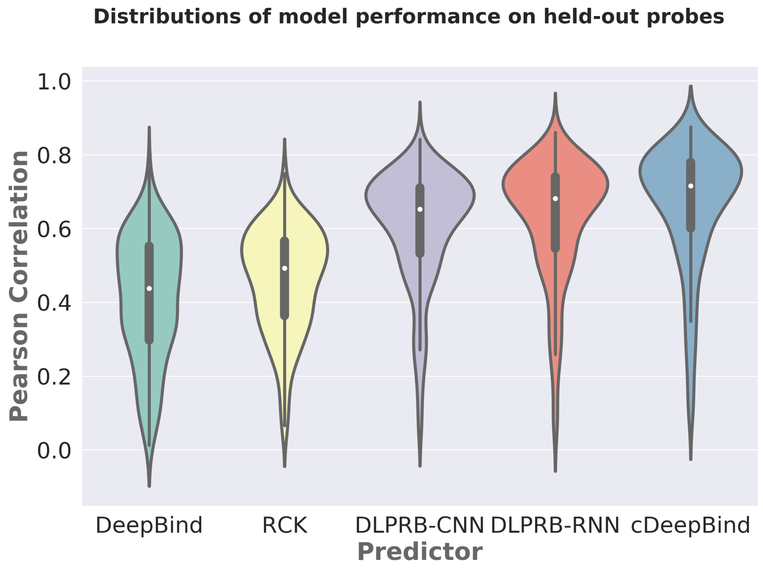
Figure 2. The cDeepbind model takes in the encoding of an RNA sequence and outputs binding intensities for target RBPs

Design Choices

1. To encode the structural context we sample an ensemble of 10 structure graphs using **boltzman sampling** using the *Forgi package*. We encode the graph into a pairing matrix by first setting all diagonal values to 1. Then for paired bases, we set the value of the nucleotide they are paired with to 1, and normalize the row. Finally, we average the pairing matrices obtained for each graph. We then multiply the pairing matrix with the sinusoidal positional encoding matrix to obtain the final structure representation.
2. We encode the RNA sequence as a one-hot-encoded vector and concatenate it with the **transformed graph embedding** representing the structure. The target RNAcompete probe intensities are **clamped at the 99.95^{th} percentile** and normalized to zero mean and unit variance, as done by other benchmarks on this dataset.
3. We used random sampling to generate hyperparameters defining our model. We use a reduction factor of 16 in the Squeeze-Excitation unit and batch-normalization before the activation. To make our training robust to outliers, we used the **Huber loss** function, which we confirmed empirically to work better than mean squared error. We use the **mean loss from the 244 probes** for each input while masking the contribution that would arise from targets with missing values.

Results

In-vitro evaluation



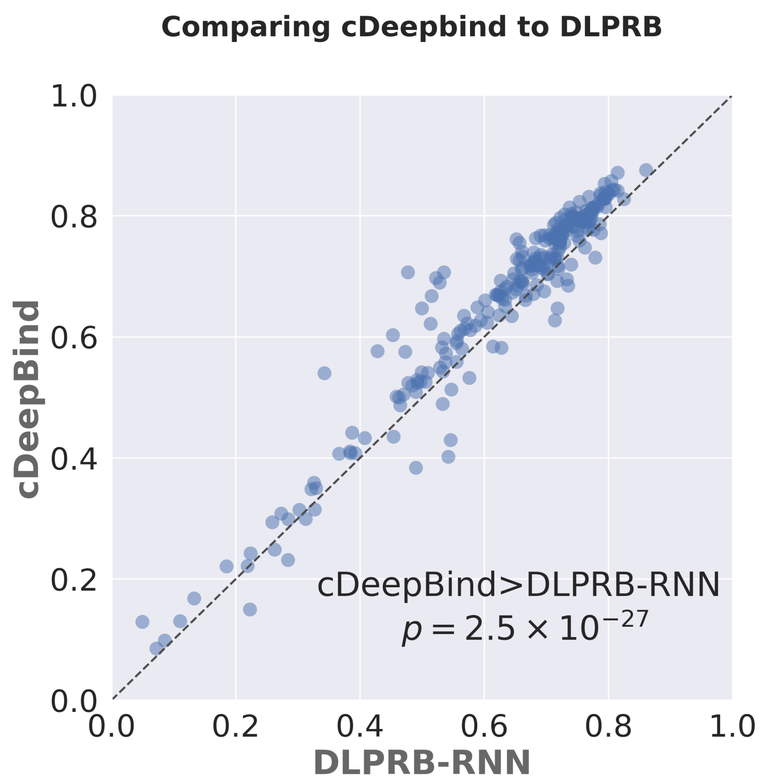


Figure 3. cDeepbind predictions have a higher correlation (0.663 vs 0.628 for DLPRB-RNN) on held out probes in RNAcompete

Predicting the effect of splicing mutations

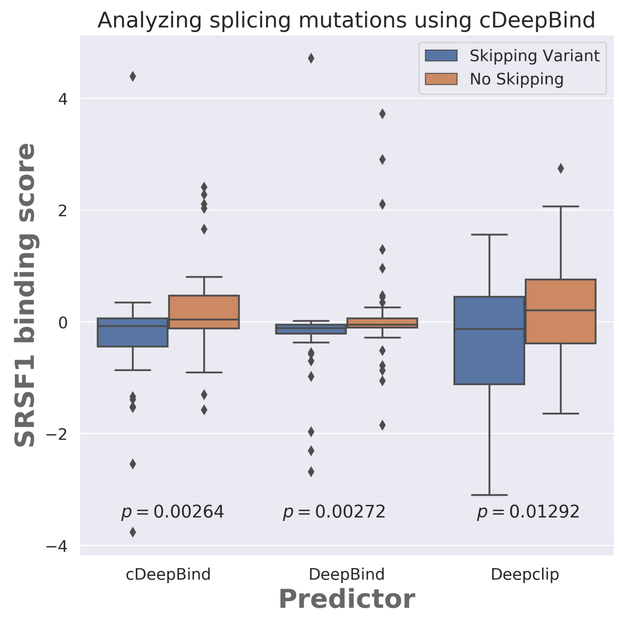


Figure 4. cDeepbind has a significant change in SRSF1 score for exon-skipping variants reported in *Gronning et al*

Discussion

We have presented a new approach for modelling RBP binding that addresses some of the limitations of prior work. Due to the extensible nature of our method, we would like to explore training on other high-throughput binding datasets such as RNA-bind-n-seq and eCLIP. We believe that our multi-task and structure aware framework would be well-suited for modelling the competitive interactions of multiple RBPs in-vivo. We would also like to explore integrating deep learning based methods for secondary structure prediction into our framework.