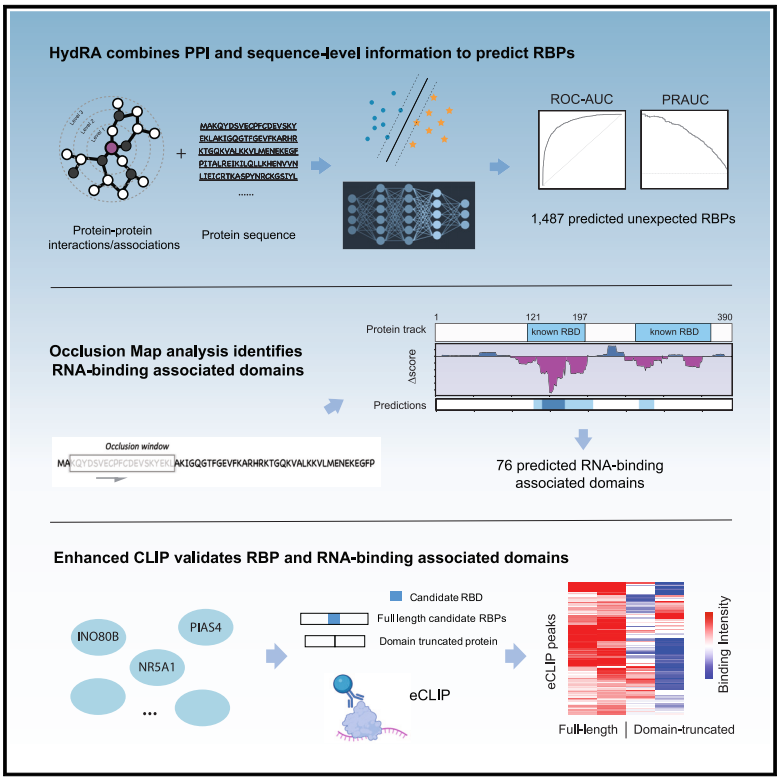
Molecular Cell

Resource

HydRA: Deep-learning models for predicting RNA-binding capacity from protein interaction association context and protein sequence

Graphical abstract



Highlights

- HydRA is a deep-learning model combining PPI and sequence features to predict RBPs
- Occlusion mapping with HydRA enables RBD discovery
- HydRA predicts RNA-binding activity for 1,487 candidate proteins and 76 candidate RBDs
- Enhanced CLIP confirms HydRA RBP predictions with RBD resolution

Authors

Wenhao Jin, Kristopher W. Brannan, Katannya Kapeli, ..., Joy S. Xiang, Limsoon Wong, Gene W. Yeo

Correspondence

geneyeo@ucsd.edu

In brief

Jin et al. developed HydRA, an RBP classifier that leverages protein interactions and sequence patterns. Utilizing machine-learning and deep-learning techniques, HydRA accurately predicts RNA-binding capacity and identifies numerous uncharacterized RNA-binding proteins and domains. eCLIP validation confirms the RNA-binding activity, expanding the catalog of known RNA-binding proteins and domains.

Jin et al., 2023, Molecular Cell *83*, 2595–2611
July 20, 2023
https://doi.org/10.1016/j.molcel.2023.06.019

CellPress

Molecular Cell

CellPress
OPEN ACCESS

Resource

HydRA: Deep-learning models for predicting RNA-binding capacity from protein interaction association context and protein sequence

Wenhao Jin,[1,2,3,7] Kristopher W. Brannan,[1,2,3,6,7] Katannya Kapeli,[4,7] Samuel S. Park,[1,2,3] Hui Qing Tan,[4] Maya L. Gosztyla,[1,2,3] Mayuresh Mujumdar,[1,2,3] Joshua Ahdout,[1,2,3] Bryce Henroid,[1,2,3] Katherine Rothamel,[1,2,3] Joy S. Xiang,[1,2,3] Limsoon Wong,[5] and Gene W. Yeo[1,2,3,8,*]

[1]Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA, USA
[2]Institute for Genomic Medicine and UCSD Stem Cell Program, University of California, San Diego, La Jolla, CA, USA
[3]Stem Cell Program, University of California, San Diego, La Jolla, CA, USA
[4]Department of Physiology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore
[5]Department of Computer Science, National University of Singapore, Singapore, Singapore
[6]Present address: Center for RNA Therapeutics, Department of Cardiovascular Sciences, Houston Methodist Research Institute, Houston, USA
[7]These authors contributed equally
[8]Lead contact
*Correspondence: geneyeo@ucsd.edu
https://doi.org/10.1016/j.molcel.2023.06.019

SUMMARY

RNA-binding proteins (RBPs) control RNA metabolism to orchestrate gene expression and, when dysfunctional, underlie human diseases. Proteome-wide discovery efforts predict thousands of RBP candidates, many of which lack canonical RNA-binding domains (RBDs). Here, we present a hybrid ensemble RBP classifier (HydRA), which leverages information from both intermolecular protein interactions and internal protein sequence patterns to predict RNA-binding capacity with unparalleled specificity and sensitivity using support vector machines (SVMs), convolutional neural networks (CNNs), and Transformer-based protein language models. Occlusion mapping by HydRA robustly detects known RBDs and predicts hundreds of uncharacterized RNA-binding associated domains. Enhanced CLIP (eCLIP) for HydRA-predicted RBP candidates reveals transcriptome-wide RNA targets and confirms RNA-binding activity for HydRA-predicted RNA-binding associated domains. HydRA accelerates construction of a comprehensive RBP catalog and expands the diversity of RNA-binding associated domains.

INTRODUCTION

Whether individually or in ribonucleoparticle protein (RNP) complexes, RNA-binding proteins (RBPs) orchestrate every step of the RNA life cycle.[1] The recognition that an increasing number of human diseases are associated with perturbed RBP function[2–4] has fueled a resurgence in RBP discovery efforts. In the past decade, RNA-interactome capture and organic phase-separation-based technologies have uncovered thousands of proteins that contact RNA.[5–10] These experimental RBP discovery approaches reveal that the RNA-binding proteome is cell-type and context specific, indicating that completing the catalog of RBPs will require resource-intensive RNA-interactome interrogations across many tissues and conditions. Furthermore, approximately half of the recently discovered RNA-interacting proteins are unconventional RBPs that lack both annotated RNA-related biological functions and annotated RNA-binding domains (RBDs),[5,6] which poses problems for prioritizing candidates for validation and characterization.

Computational and machine-learning methods offer an alternative strategy for predicting proteome-wide RNA-binding capacity. Several RBP classifiers that utilize protein sequence information, known as sequence-based classifiers, have been developed, including RNApred,[11] SPOT-seq,[12] catRAPID signature,[13] an RNA-binding protein predictor (RBPPred),[14] and TriPepSVM.[15] All these classifiers except SPOT-seq rely on machine-learning models, such as support vector machines (SVMs), and use features extracted from protein sequences such as amino acid composition (AAC), *k*-mers, and predicted physicochemical properties to distinguish RBPs from non-RBPs. However, each of these algorithms has its own limitations. Machine-learning methods based on human-defined features such as global protein properties (e.g., AAC) and primary subsequence count (e.g., *k*-mer) are not sensitive to relatively small RNA-binding signatures. SPOT-seq employs a template-based strategy that aligns query proteins to known protein-RNA complex structures and evaluates their RNA-binding affinity.[12] However, SPOT-seq's performance depends on the quality and

comprehensiveness of the template set of protein-RNA complex structures. Although these classifiers exhibit some accuracy in recognizing "conventional" RBPs with annotated RBDs, they demonstrate limited predictive ability for the entire catalog of experimentally defined human RBPs, of which >50% are anticipated to be unconventional RBPs lacking RBDs.[15] This limited predictive ability is due to the simplicity of the sequence-based machine-learning models employed by these classifiers and the inherent complexities involved in RBP prediction. The scarcity of positive-training samples further hinders the modeling of implicit and intricate sequence patterns present in unconventional RBPs.

We previously developed a machine-learning algorithm, support vector machine obtained from neighborhood associated RBPs (SONAR), which utilized protein-protein interaction (PPI) networks to identify RBPs.[16] SONAR leveraged our findings from a panel of RNase-treated RBP interactomes, revealing that RBPs often physically interact with one another and identified unconventional RBPs at a high validation rate.[16] However, SONAR failed to classify a portion of conventional RBPs with known RBDs, likely because SONAR's high reliance on experimentally defined PPI networks limits its application for proteins underrepresented in these networks.

State-of-the-art machine-learning algorithms, such as deep learning, have revolutionized the feature-engineering field and have thus improved the precision of predictive models. Convolutional neural network (CNN) and Transformer, widely used in computer vision (CV) and natural-language processing (NLP), respectively, can automatically extract valuable information from raw data without the need for hand-engineered features. These algorithms have shown promise in biological sequence classification, addressing tasks such as protein folding, SNP calling, and protein-family classification.[17] However, these models require large training datasets to avoid overfitting and maintain robustness. To overcome the challenge of small positive-training sizes, pretraining methods like self-supervised learning and transfer learning have been successfully applied in CV, NLP, and protein classification problems.[18] For instance, ProteinBERT, a Transformer-based protein language model, achieved state-of-the-art performance on various protein classification tasks, covering diverse protein properties.[19] Traditional machine-learning models like SVM and random forest are effective for small datasets and mitigating overfitting. Therefore, integrating cutting-edge deep learning with traditional machine-learning models through ensemble learning presents a promising approach to designing precise and robust RBP classifiers.

Here, we introduce a novel machine-learning-based algorithm, hybrid ensemble classifier for RBPs (HydRA) that predicts not only the RNA-binding capacity of proteins but also protein regions involved in RNA-protein interaction. This algorithm applies an ensemble learning method integrating CNN, Transformer, and SVM in RBP prediction by utilizing both intermolecular protein context and sequence-level information. Pretraining methods including self-supervised and transfer learning enhance model robustness and alleviate overfitting. Our findings demonstrate that HydRA achieves state-of-the-art performance in RBP prediction and effectively recognizes RNA-binding-associated domains using the occlusion map approach.[20–22] Our results also emphasize the importance of intermolecular protein context in identifying new RBPs with elusive sequence or structural patterns. HydRA predicted 1,487 previously unclassified RBPs and 76 new RBDs. Using enhanced cross-linking and immunoprecipitation (eCLIP), we validated a subset of HydRA predictions, including previously uncharacterized RBPs HSP90A and the YWHA family of proteins—as well as the novel RBP candidates INO80B, PIAS4, NR5A1, ACTN3, and MCCC1—providing molecular insights into their function. Experimental verification of RNA-binding activity for predicted RBDs supports HydRA's ability to identify protein regions that contribute to RNA-binding capacity. These results establish HydRA as the most accurate RBP classifier to date, capable of predicting RBDs at the amino acid level.

RESULTS

Improved SONAR models penalize low PPI neighborhood and increase network density

We previously developed SONAR (1.0), an SVM-based machine-learning algorithm that leverages large-scale experimental PPI data (without protein sequence information) to predict unannotated RBPs.[16] Employing the BioPlex PPI dataset,[23] SONAR was able to predict RBPs with high specificity and sensitivity. The success of SONAR was based on observations that an unusually high proportion of proteins that associate with RBPs (via RNA-dependent and -independent interactions) are themselves RBPs. As such, SONAR's ability to accurately identify RBPs depended on the density of PPIs centered on an interrogated protein (i.e., the number of "edges") or the completeness of the local PPI network. Basically, known RBPs that were assigned low SONAR scores (false negatives) harbored significantly fewer PPI edges than correctly predicted, known RBPs (true positives) (Figure 1A). In this study, to evaluate the extent to which the density of PPI networks affects the predictive power of SONAR, we constructed a more comprehensive PPI network by combining BioPlex2.0[23] with Mentha,[24] a collection of physical PPIs from MINT,[25] BioGRID,[26] and IntAct[27]—we refer to this ensemble network as Mentha-BioPlex (MB). Next, we incrementally reduced the network density of MB by randomly removing a percentage of edges and determined the receiver operating characteristic (ROC)-area under the curve (AUC) value of SONAR by 10-fold cross-validation. Indeed, both sensitivity and specificity of SONAR improve with more comprehensive PPI networks (Figure 1B). To evaluate the predictive power of SONAR on a corrupted network, we incrementally removed edges from a shuffled MB network such that the number of edges was unchanged, but edges were randomly connected to maintain network topological properties (i.e., node degree distribution). The resulting ROC-AUC values of SONAR models using shuffled MB networks were significantly lower (<0.65) with large standard error of the mean compared with that of the unshuffled network (Figure 1B). This result demonstrates that low-quality PPI data will weaken the predictive power of SONAR. We therefore added an indicator feature in our algorithm to mark proteins with limited PPI information to alleviate the impact of PPI sparsity on SONAR performance. We split the whole protein set into a training/validation (80% of the whole protein set), which is used for model training, hyperparameter tuning and feature selection,

CellPress
OPEN ACCESS

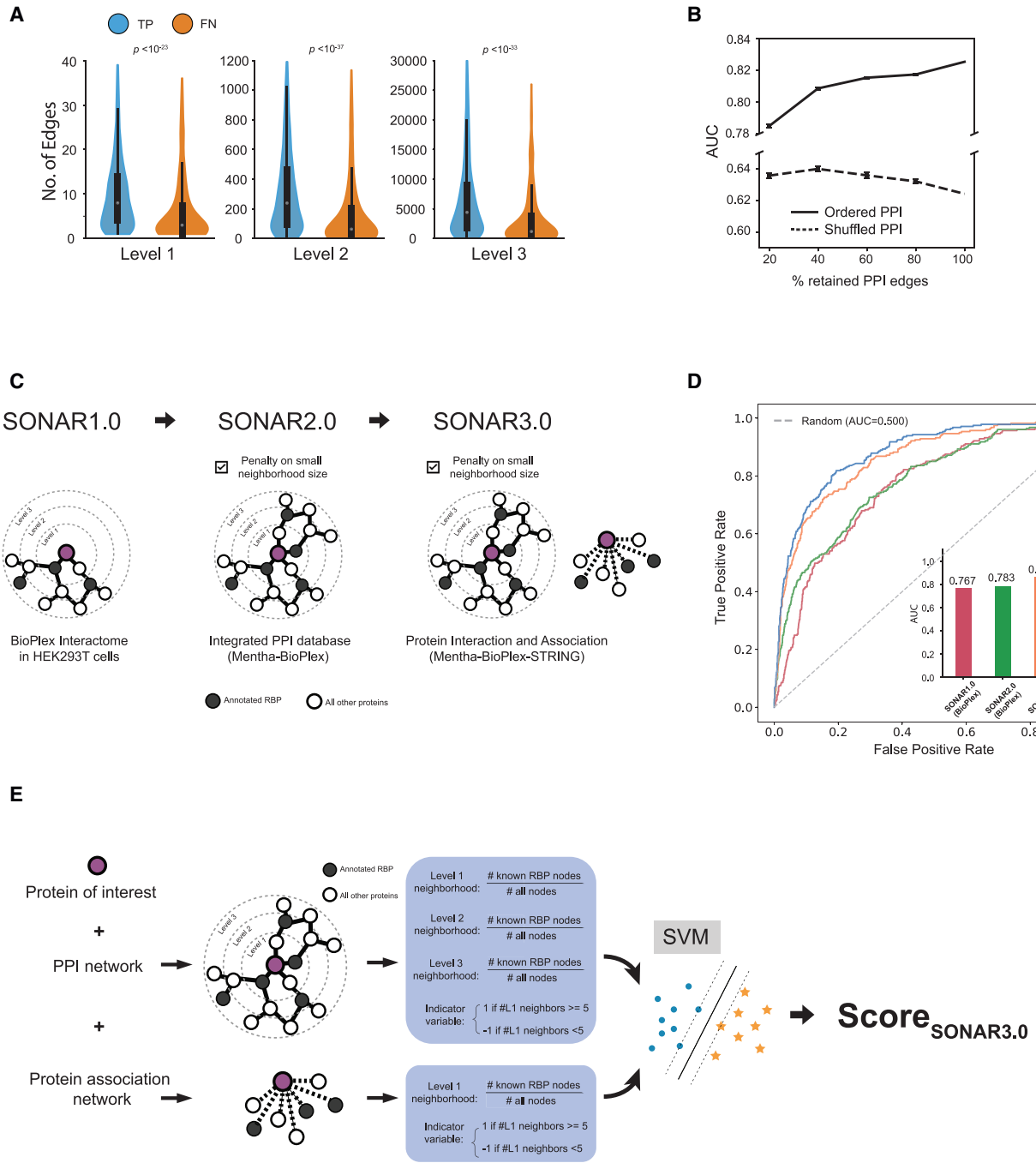


Figure 1. Upgraded RBP classification performance of SONAR3.0 classifier compared with SONAR1.0

(A) Violin plots showing the number of level 1, 2, and 3 neighbors for RBPs that were scored correctly (TP) or incorrectly (FN) by SONAR. Significant differences in neighbor counts exist between TP and FN RBPs for all levels (level 1, $p < 1e{-}23$; level 2, $p < 1e{-}37$; level 3, $p < 1e{-}33$ by one-side Mann-Whitney U test). The medians of each population are shown.

(B) Positive correlation between network density and ROC-AUC value. Random shuffling of edges resulted in no correlation with ROC-AUC value. Error bars represent variability from 10 replicates.

(C) Evolution from SONAR1.0 to SONAR3.0 where new PPI data type was introduced in each version.

(D) SONAR3.0 utilizes extracted features from PPI and protein association networks, employing an SVM with an RBF kernel for mathematical modeling. Features include the percentage of known RBPs in each neighborhood (first three neighborhoods for PPI and direct neighborhood for protein association network), and penalty is applied for cases with a low number of protein neighbors in each network.

(E) ROC-AUC analysis to compare the predictive power of SONAR1.0 with BioPlex network, SONAR2.0 with BioPlex network, SONAR2.0 with Mentha-BioPlex (MB) network, and SONAR3.0 with Mentha-BioPlex and STRING (MBS). Inset: bar graph of ROC-AUC values.

and a hold-out test set for model evaluation, which was used for all the model evaluation in this study. Adding the indicator feature led to an increase in ROC-AUC value by 0.016 (from 0.767 to 0.783) for the SONAR model with BioPlex network (Figures 1C and 1D) and an increase in precision-recall-AUC (PR-AUC) value by 0.108 (from 0.347 to 0.455) (Figure S1B). We refer to this revised model as SONAR2.0. By replacing BioPlex network with the denser MB network (Figures 1C and S1A), we improve SONAR1.0's performance by another 0.078 (from 0.783 to 0.861, Figure 1D) in ROC-AUC and 0.108 (from 0.455 to 0.563) in PR-AUC (Figure S1B).

Next, we evaluated if adding functional protein association data from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database would further improve SONAR2.0 as functional protein association data had previously been successfully applied in other PPI-based protein prediction tasks.[28] Like SONAR2.0, we set an indicator feature to penalize proteins with few interacting neighbors within its local neighborhood in the STRING network. With the extended protein networks, we obtained an improved RBP classifier (SONAR3.0) using an SVM model with a radial basis function (RBF) kernel (Figures 1E and S1A), where the prediction performance increased by 0.020 in ROC-AUC and 0.051 in PR-AUC (Figures 1D and S1B). Specifically, the PR analysis highlights the advantage of incorporating more PPI and functional association information to densify our networks and minimize false positives (Figure S1B). For instance, the precision of the model increased from 0.41 to 0.57, 0.73, and 0.76 at the 0.3 recall level. Setting a cutoff to recover the top 30% of known RBPs ranked by classification scores, we obtain 86 false positive predictions (precision = 0.41) using SONAR1.0, 43 false positives using SONAR1.0 with an updated BioPlex network, 21 false positives using SONAR2.0 (with MB) and only 18 false positives using SONAR3.0 (Figure S1B).

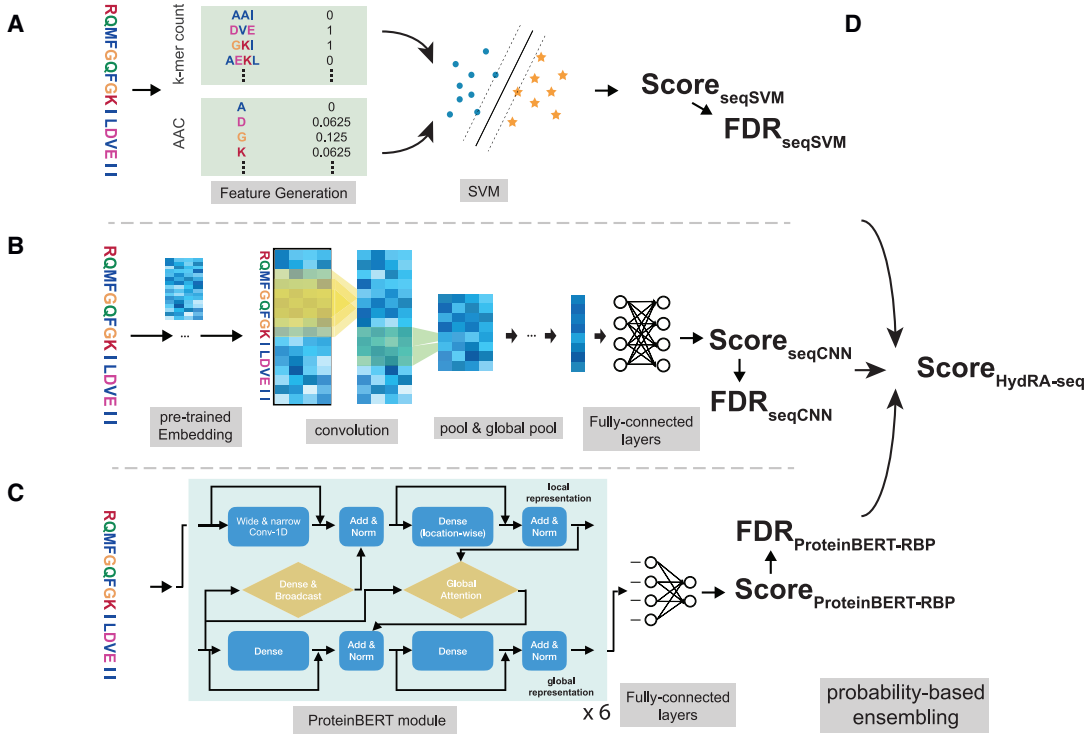
HydRA-seq classifies RBPs from protein sequence by ensemble learning

We next set out to augment the HydRA workflow to enable classification of RBPs and non-RBPs based on the combination of two components: (1) SONAR3.0, which captures the RNA-binding-related intermolecular patterns (Figure 1E), and (2) RNA-binding-related patterns from primary amino acid sequences (referred to as HydRA-seq) (Figures 2A–2D). The primary amino acid sequence of protein domains has been previously leveraged by RBP classifiers with classic machine-learning models to predict RNA-binding capacity of canonical RBPs with well-defined RBDs.[11–15] Expectedly, these sequence-based classifiers perform poorly on experimentally defined RBPs that do not have canonical RBDs (Figures S2A and S2B).

To combine the strength of both traditional machine-learning models and deep-learning models, we constructed three RBP classifiers with state-of-the-art machine-learning technology using classical and novel sequence-based features which are generated from protein sequence. In the first classifier termed seqSVM, *k*-mer (sub-sequence of length *k*) and AAC extracted from the amino acid sequences were used to train an SVM with an RBF kernel (Figure 2A). For the second classifier called seqCNN, we employed a CNN to learn features of RBPs from full-length (FL) amino acid sequences (Figure 2B). Briefly, an embedding layer pre-trained by BioVec[29] was employed to obtain a comprehensive representation of the biophysical and biochemical properties from each protein sequence, followed by convolutional and pooling layers that automatically extract and compress sequence-level features. Fully connected layers with non-linear activation units were added to further process the compressed features and generate final predictions. To avoid convergence of suboptimal solutions and accelerate the training process, seqCNN was initialized with weights from a convolutional autoencoder pre-trained with all the protein sequences in our dataset before model training. For the third classifier called ProteinBERT-RBP, protein sequences are fed into an attention-based Transformer architecture designed for proteins (ProteinBERT)[19] to obtain predicted RNA-binding propensities (Figure 2C). ProteinBERT-RBP was pre-trained with ~106 million UniRef90 protein sequences and their corresponding functional annotations from the gene ontology database in a self-supervised manner and then fine-tuned using our human RBP annotations. As a result, ProteinBERT-RBP digests each input protein sequence in two parallel paths, (1) local features of the sequence and (2) global function of the sequence based on the knowledge obtained in the pretraining stage, while the two paths interact and guide the learning process of each other via attention mechanism and generate predictions on RNA-binding propensity.

ROC-AUC and PR analysis with the held-out test set demonstrates good predictive power of seqCNN, seqSVM, and ProteinBERT-RBP for individually identifying RBPs (Figures 2E and S2C). These classifiers identify RBPs with characterized RNA targets (CLIP-based data, collected in CLIPdb[30]) or known RNA-binding or RNA-processing-related domains (referred to as "characterized RBPs") with approximate ROC-AUC values of 0.80, but performance is diminished for identifying RBPs lacking known RBDs (referred to as "uncharacterized RBPs") (Figures S2D and S2E). To improve the prediction performance for both characterized and uncharacterized RBPs, we designed a false discovery rate (FDR)-based ensemble approach to integrate the predictions from seqCNN, seqSVM, and Protein BERT-RBP (Figure 2D). This strategy utilizes the joint probability of each classifier's predictions of being false discoveries and outputs the complementary probability of this joint probability as the ensemble classification score. We referred to this ensemble classifier as HydRA-seq and find that HydRA-seq outperforms all of its sub-components (i.e., seqCNN, seqSVM, and ProteinBERT-RBP) (ROC-AUC = 0.842 and PR-AUC = 0.643, Figures 2E and S2C). In particular, HydRA-seq exhibits better predictive power for identifying uncharacterized RBPs (ROC-AUC = 0.782 and PR-AUC = 0.347) when compared with individual seqCNN, seqSVM, and ProteinBERT-RBP predictions (Figures S2D and S2E). HydRA-seq, trained with data only from the human proteome, also showed better performance on both characterized and uncharacterized RBPs than the current state-of-the-art sequence-based RBP classifier, TriPepSVM, which was trained with all the RBPs and non-RBPs in *human*, *Salmonella*, and *E. coli* (Figures S2D and S2E, HydRA-seq: ROC-AUC = 0.913 and PR-AUC = 0.693 for characterized RBPs; TriPepSVM: ROC-AUC = 0.890 and PR-AUC = 0.508 for characterized RBPs, ROC-AUC = 0.719 and PR-AUC = 0.293 for uncharacterized RBPs).

CellPress
OPEN ACCESS



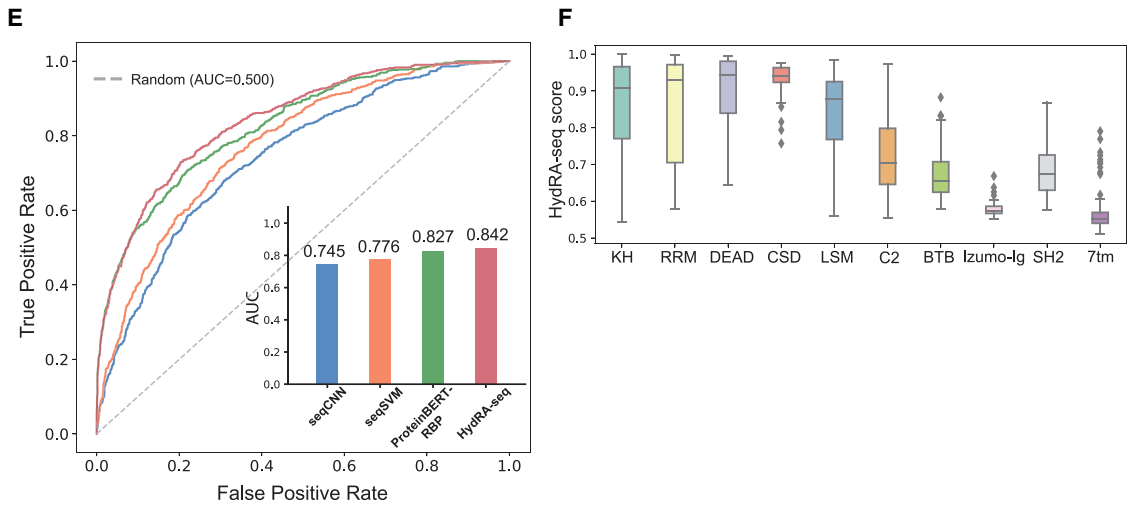


Figure 2. HydRA-seq classifies RNA-binding proteins

(A–D) HydRA-seq model that consists of sub-classifiers including seqSVM, seqCNN, and ProteinBERT.

(A) seqSVM employs SVM with an RBF kernel, utilizing *k*-mer and amino acid composition for prediction.

(B) seqCNN includes an embedding layer, convolutional and pooling layers, a global pooling layer, fully connected layers, and an output layer. It takes protein sequence as input. Only one set of convolutional and pooling layers is shown. Each layer (except embedding and pooling layers) is coupled with a batch-normalization layer.

(C) ProteinBERT-RBP process the protein sequence using six sequential ProteinBERT modules. These modules apply parallel neural networks to extract and process local and global presentations, integrating the information using a cross-attention mechanism. The resulting embedding vector passes through a fully connected neural layer to generate the classification score ($Score_{ProteinBERT\text{-}RBP}$).

(D) Probability-based ensembling converts classification scores from the three RBP classifiers into the overall sequence-based RBP score ($Score_{HydRA\text{-}seq}$). The false discovery rate of each classifier's prediction is used to generate the $Score_{HydRA\text{-}seq}$.

(E) ROC-AUC analysis evaluates the performance of seqCNN, seqSVM, ProteinBERT-RBP, and the ensemble classifier (HydRA-seq). Inset: bar graph comparing ROC-AUC values.

(F) Classification scores that are generated by HydRA-seq for amino acid sequences from five RBDs and five domains unrelated to RNA processing (non-RBDs). Th RBDs include KH, RRM, DEAD box, CSD, and LSM. The non-RBDs include C2, BTB, SH2, immunoglobulin (Ig) domain, and 7 transmembrane (7tm) domain.

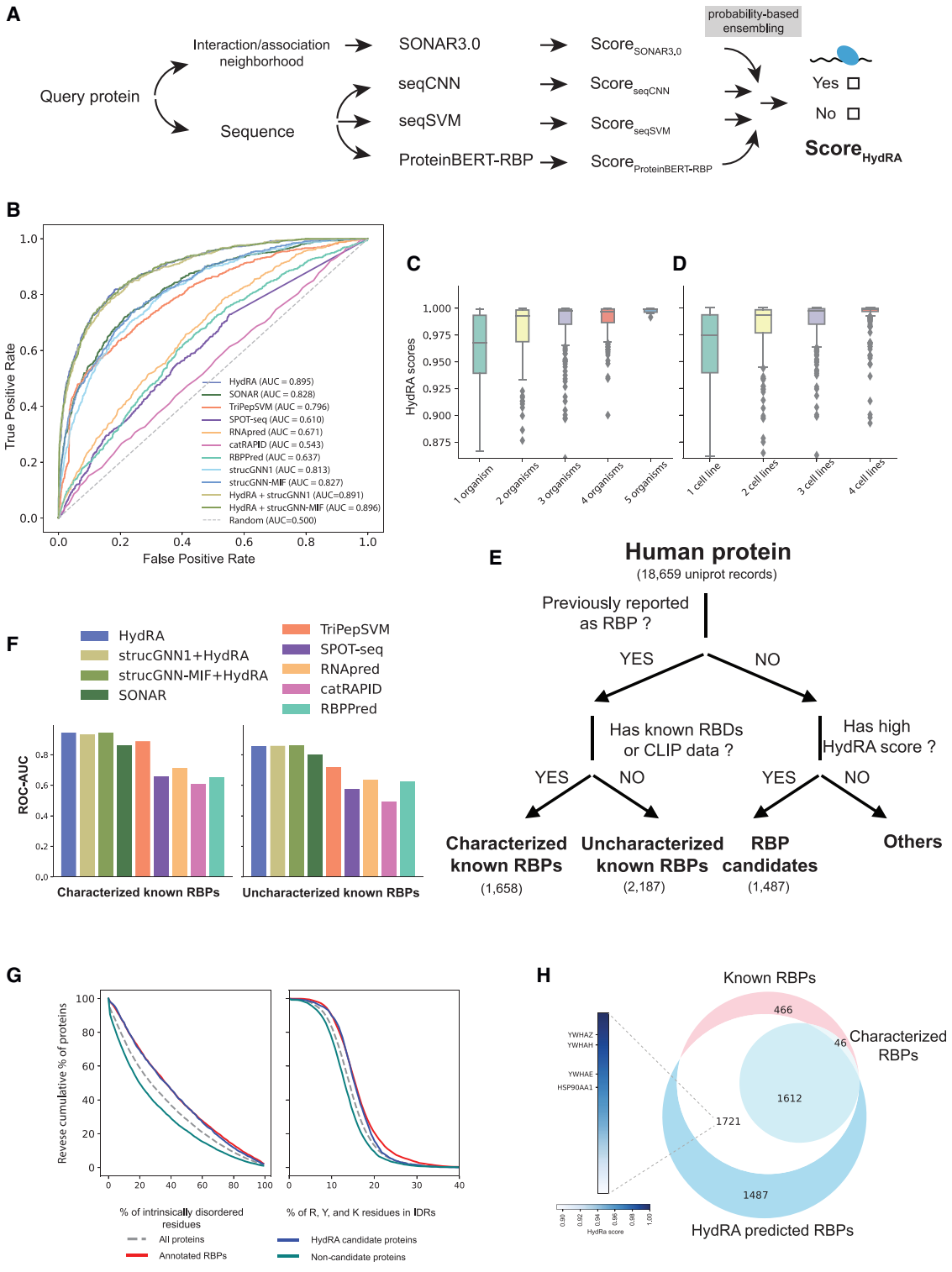


Figure 3. A novel PPI- and sequence-based approach to recognize RBPs

(A) The HydRA architecture integrates four sub-classifiers (SONAR3.0, seqCNN, seqSVM, and ProteinBERT) to predict RBPs. Their output scores are combined using a probability-based ensembling method to generate the final RBP prediction score ($Score_{HydRA}$).

(B) Comparison of available RBP classifiers based on ROC-AUC analysis using the hold-out test set. Area under curve (AUC) values are indicated for each classifier.

(legend continued on next page)

We next tested whether HydRA-seq can distinguish known RBDs from other protein domains. We computed HydRA-seq classification scores for sequences collected from Pfam database corresponding to RBDs (K homology domain [KH], RNA recognition motif [RRM], Asp-Glu-Ala-Asp box [DEAD], cold-shock domain [CSD], and like-SM domain [LSM]) in different species. Expectedly, these sequences generally resulted in high classification scores (Figure 2F). By contrast, sequences from protein domains that are not commonly found in RBPs such as PPI (Broad-Complex, Tramtrack, and Bric à brac [BTB], IgG), membrane associated (transmembrane, C2), and src homology 2 (SH2) domains overall had lower classification scores than the known RBDs (Figure 2F). These results demonstrate the capacity of our sequence-based RBP classifiers to distinguish sequence patterns that contribute to RNA-binding activity with domain-level resolution.

HydRA achieves state-of-the-art RBP classification performance

Given that SONAR3.0 and HydRA-seq individually predict RBPs with high accuracy even though they are trained using different protein features, we reasoned that a combined model would show enhanced performance compared with each individual model. Indeed, we observed that combining prediction scores from SONAR3.0 and HydRA-seq components (i.e., seqCNN, seqSVM, and ProteinBERT-RBP) using the FDR-based probabilistic ensemble approach (Figure 3A), produced a new model (herein termed HydRA) which outperformed its sub-classifiers with higher specificity and sensitivity displayed by AUC of 0.895 in ROC analysis and an AUC of 0.745 in PR analysis within the hold-out test set (Figures S3A and S3B; Table S1). For instance, at the 0.1 false positive rate level, HydRA identifies known RBPs (higher sensitivity) and fewer false positives (reduced FDR) compared with SONAR3.0 and HydRA-seq (Figure S3C; Table S1). To determine HydRA's reliance on sequence homology, we evaluated HydRA sub-components with subsets of the test set consisting of proteins with varying sequence similarity to proteins in the training and validation set. HydRA (and each of its components) shows no significant degeneration in RBP predictive power as sequence similarities decrease, suggesting HydRA detects other deterministic RBP characteristics independent of sequence similarity (Figure S3D). We next compared the performance of HydRA with other published RBP classifiers, namely SONAR, TriPepSVM, SPOT-seq, RNApred, catRAPID signature, and RBPPred, and observed higher specificity and sensitivity than all other classifiers in the hold-out test set as shown in ROC-AUC and PR-AUC analyses (Figures 3B and S3E). High-quality protein structures predicted from amino acid sequences have recently become available for most natural proteins with AlphaFold2.[31] To test if predicted protein structures can further improve our RBP predictions, we constructed a group of structure-based RBP classifiers with AphalFold2 predictions as input using a graph neural network widely used in structure-based models for protein prediction problems.[32–34] Optimal models (referred to as strucGNN1) were chosen after iterative feature selection and model tunning with our validation set (Table S4). We also obtained another structure-based RBP classifier by fine-tuning a pre-trained GNN-based protein model with our training set (see STAR Methods), referred to as strucGNN-MIF. These strucGNNs achieve 0.813–0.827 in ROC-AUC analysis and 0.563–0.607 in PR-AUC analysis in the hold-out test set (Figures 3B and S3E).

Next, we again grouped known RBPs into characterized and uncharacterized sets (Figure 3E) and showed that HydRA had higher recall, specificity, and precision than other classifiers (including HydRA sub-classifiers) for both characterized RBPs and uncharacterized RBPs that lack RNA-binding or -processing related domains and defined RNA targets (ROC-AUC = 0.945 and PR-AUC = 0.784 for characterized RBPs, ROC-AUC = 0.854 and PR-AUC = 0.493 for uncharacterized RBPs, Figures 3F and S3G–S3I). Notably, as shown above, TriPepSVM performed well for characterized RBPs but showed more degeneration for uncharacterized RBPs (Figure 3F). These results demonstrate that HydRA achieves state-of-the-art predictive power for RBPs either with or without annotated RBDs, additionally indicating that utilizing protein interaction and association information increases predictive power for RBPs with unknown RNA-binding signatures. This is also evident when we compare the performance of SONAR3.0 with the combination of HydRA's sub-classifiers for uncharacterized RBPs (Figures S3G and S3H). StrucGNNs displayed slightly better performance than HydRA-seq for recognizing uncharacterized RBPs but showed inferior predictive power for characterized RBPs (Figures S2D and S2E). Similarly, we observed that incorporating predicted protein structure features into HydRA (referred to as "strucGNN1+HydRA" and "strucGNN-MIF+HydRA") leads to a slight improvement for recognizing uncharacterized RBPs over HydRA but a slight degeneration on recognizing characterized RBPs (Figures 3F

(C) Boxplots showing the correlation between the HydRA scores of RBPs and the number of species in which these RBPs have been reported to bind RNA according to the review by Hentze et al.[4] In the boxplots, the lower and upper hinges denote the first and third quartiles, with the whiskers extending to the largest value less than 1.5× the interquartile range. Flier points are those past the ends of whiskers, considered as outliers.

(D) Boxplots showing the correlation between the HydRA scores of RBPs and the number of human cell lines in which these RBPs have been reported to bind RNA according to the review by Hentze et al.[4]

(E) Human proteins were grouped into four categories. Known RBPs were subdivided into "characterized known RBPs" and "uncharacterized known RBPs." Proteins that did not fall into these categories were labeled as "RBP candidates" if their HydRA score exceeded the threshold (with a false positive rate of 0.1), and as "others" otherwise.

(F) The prediction performance of RBP classifiers on characterized and uncharacterized RBPs via ROC-AUC analysis.

(G) Reverse cumulative density curves showing the distribution of physicochemical properties in known RBPs, candidate RBPs, non-candidate proteins, and the entire human protein set. The properties include the relative size of intrinsically disordered regions (IDRs) and the composition of RNA-binding-related amino acids (i.e., R, Y, and K) in the IDRs.

(H) Venn diagram illustrating the relationship among previously reported RBPs ("known RBPs"), HydRA-predicted RBPs, and characterized RBPs. Highly scored uncharacterized RBPs, HSP90AA1, YWHAH, YWHAE, and YWHAZ along with YWHAG, are selected for further exploration in this study.

and S3I). Because these "strucGNNs+HydRA" models do not significantly improve the prediction performance on the overall RBP set (ROC-AUC = 0.891–0.896, PR-AUC = 0.731–0.742) shown in Figure 3B, we chose not to include strucGNNs in the final HydRA model.

Incidentally, although the HydRA model was only trained with human proteome and human RBP annotations, we observe that this model can also recognize RBPs from mouse, yeast, and fly with substantial accuracy producing ROC-AUC values of ~0.8 and a PR-AUC range from 0.433–0.639 (Figures S4A and S4B). This result indicates the model succeeds in capturing the general RNA-binding associated features across species even just with learning from human RBPs.

We next examined the association between HydRA classification scores and RBP conservation. RBP annotations across multiple organisms (i.e., *Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Caenorhabditis elegans*) and human cell lines (i.e., HeLa, HuH7, HEK293, and K562) collected from a previous study[4] revealed that HydRA classifier scores positively correlate with the number of organisms in which a given RBP has been defined (Figure 3C). HydRA classification scores also positively correlated with the number of cell-types with validated RNA-binding activity for a given RBP (Figure 3D). These results indicate that HydRA scores are strong indicators of the reliability of RBP annotations and establish HydRA analysis as a complementary technique for experimental high-throughput RBP discovery approaches.

HydRA predicts 1,487 human RBP candidates out of the 18,659 proteins that exist in the MB network, at a false positive rate of 10% (Figure 3E; Table S2). We observed that HydRA-predicted candidate RBPs are, like known RBPs, significantly enriched in intrinsically disordered regions (IDRs) compared with background, all human proteins ($p < 10^{-23}$, Mann-Whitney U test). We defined negatively scored predictions as proteins with HydRA scores lower than 50% of the human proteins (referred to as "non-candidate proteins") and these were not statistically different from the background (Figure 3G). Furthermore, candidate RBPs demonstrated a significantly higher proportion of the amino acids arginine, tyrosine, and lysine within IDRs, which are biophysical indicators of RNA-binding,[4,35] compared with background ($p < 1e{-}20$, Mann-Whitney U test) and negative predictions ($p < 1e{-}55$, Mann-Whitney U test) (Figure 3G). These observations indicate that HydRA candidate RBPs share similar characteristics with known RBPs and suggest that IDR-based RNA-binding mechanisms may be widespread among known and unconventional RBPs. Gene ontology analysis of HydRA RBP candidates reveals enrichment for functional categories such as histone modification, DNA replication, and actin cytoskeleton organization, implying potential RNA-binding roles in co-transcriptional processes and RNA localization (Figure S3J). Uncharacterized RBPs and candidate RBPs can be prioritized by implied biological functions and HydRA predictive scores to explore new RBP classes and RNA-binding landscapes.

HSP90 is an RBP that binds 3′ UTRs of mRNA encoding other heat shock proteins

HydRA predicts functional RNA-binding for 1,721 uncharacterized RBPs without known RBDs, RNA-binding functionality, or RNA targets (Figure 3H). To illustrate the utility of HydRA to discover novel RBPs with no previous experimental evidence to support a role in RNA binding, we selected the heat-shock protein 90 alpha family class A member 1 (HSP90AA1) as a high-scoring (HydRA score: 0.968), uncharacterized RBP of substantial biological interest with a classical role as a molecular chaperone that controls cellular homeostasis (Figure 3H). Although an HSP90 homolog in tobacco plants was found to interact with the genomic RNA of the Bamboo mosaic virus,[36] there has been no transcriptome-wide characterization of HSP90AA1 as a functional RBP in mammalian systems. We performed eCLIP analysis using V5-tagged HSP90AA1 transiently expressed in HEK293 cells (Figure S5A). We obtained highly reproducible binding profiles when compared with control eCLIP of V5 tag expressed without HSP90AA1(Figure S5B). Combined HSP90AA1 eCLIP replicate libraries identified 480 reproducible binding sites (eCLIP peaks) on 225 target genes. Enriched binding motifs were determined from these binding sites with HOMER[37] (Figure S5E). We found that HSP90 RNA-binding sites were highly enriched in the 5′ and 3′ untranslated regions (UTRs) of target RNAs (Figures S5C and S5D). Gene ontology analysis of the top HSP90AA1 RNA targets revealed an enrichment for genes related to heat response, RNA metabolism, and G2/M phase cell cycle regulation (Figure 4A). Most mRNA-targets related to these terms contained HSP90AA1 binding sites within 3′ UTR regions. Many of the mRNA-targets related to heat response terms, such as transcripts encoding the HSP70 components HSPA1B and HSPA6, are involved in the HSP90 protein folding pathway (Figure 4B). HSP40, HSP70 (composed of subunits that include HSPA1B, HSPA6, and HSPA13), and PTGES3/p23, which itself is more recently discovered as an RBP,[38] are HSP90 chaperones that help recognize unfolded client proteins, deliver the unfolded proteins to HSP90, and facilitate release of the folded client protein (Figure 4C).[39] As an orthogonal validation of these HSP90AA1 mRNA-targets, we performed RNA immunoprecipitation followed by qPCR (Figures S5F and S5G) and observed significant enrichment of mRNAs that encode proteins involved in HSP90 protein folding (HSPA1B, HSP6, HSP40) and RNA metabolism (MBNL1 and CNOT1) but not the mRNAs encoding actin beta (ACTB) or glyceraldehyde-3-phosphate dehydrogenase (GAPDH). These results reveal a highly specific RNA-binding landscape for HSP90AA1 that likely regulates chaperone responses to cellular stress.

Whole-transcriptome characterization of YWHAG/H/E/Z as RBPs

YWHA (14-3-3) is a family of multifunctional proteins that work as phosphor-binding adapters, scaffolding proteins, and protein chaperones. They are known for their involvement in the critical processes of cancers (such as apoptosis, cell cycle progression, and autophagy) and neurological disorders.[40–44] All the members of this family were assigned with high HydRA scores (Figure 3H), and although they have been identified as RBP candidates in one or more high-throughput screens,[5,6,9,10,35] genome-wide RNA targets and RNA-binding functionality have not been confirmed for this protein family.

Here, we present a transcriptome-wide characterization of RNA-binding sites of YWHAH, YWHAG, YWHAE, and YWHAZ

CellPress
OPEN ACCESS

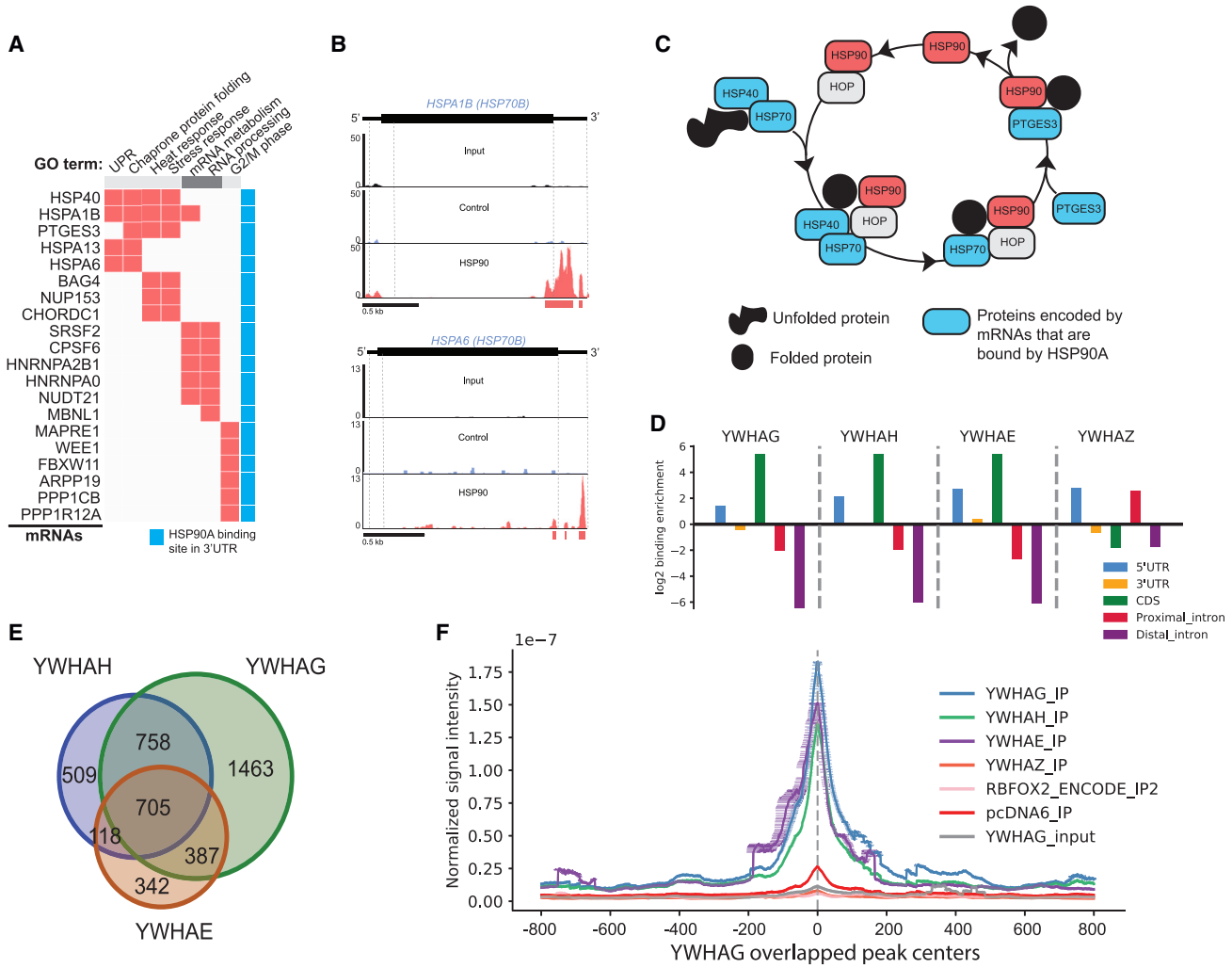


Figure 4. eCLIP analysis discovers HSP90 as an RBP that binds mRNA encoding other heat shock proteins and YWHA (14-3-3) family that binds RNAs with similar binding preference across the transcriptome

(A) Gene ontology enrichment analysis for the 100 top HSP90 RNA targets using the Enrichr Gene Ontology enrichment tool identified GO terms related to heat response, mRNA metabolism/processing, and cell cycle G2/M phase. RNAs related to the GOs are indicated by the orange box. The blue box indicates the RNAs that are bound by HSP90A in their 3′ UTR.

(B) Genome browser shots of HSPA1B (top) and HSPA6 (bottom), which both belong to the HSP70B complex, showing an enrichment of sequencing reads in the 3′ UTRs of both mRNAs in the HSP90A eCLIP-seq library but not in the input library or control eCLIP-seq library.

(C) Schematic of the HSP90 protein folding pathway shows that many of the constituents in this pathway are RNA targets of HSP90A.

(D) Binding preference of YWHAG, YWHAH, YWHAE, and YWHAZ on different genomic regions, including 5′ UTR, 3′ UTR, coding sequence (CDS), proximal introns, and distal intron, measured by $\log_2$-transformed region enrichment E_{region} (see STAR Methods).

(E) The venn diagram showing the overlap of RNA targets of YWHAG, YWHAH, and YWHAE.

(F) Read density around the binding peak center of YWHAG across the transcriptome. The read density of the eCLIP IP samples of YWHAG, YWHAH, YWHAZ, RBFOX2, vector control, and the size-match input samples of YWHAG is shown. Standard error of the mean of of the reads density of each site is shown with horizontal error bar.

proteins with eCLIP using V5-tagged proteins transiently expressed in HEK293 cells (Figure S6B). With the peak-calling algorithm CLIPper, we identified 11,187 binding sites (eCLIP peaks) over the transcriptome with stringent threshold (see STAR Methods) shared by YWHAG eCLIP replicates, 4,506 common eCLIP peaks in YWHAH replicates, 2,939 common eCLIP peaks for YWHAE and 308 common eCLIP peaks for YWHAZ (Figure S6A) with their enriched binding motifs on the RNA targets determined by HOMER[37] (Figure S6E). Three of the four proteins (viz. YWHAG, YWHAH, and YWHAE) have their eCLIP peaks enriched in the protein-coding sequence (CDS) and 5′ UTR regions of mRNA molecules and under-presenting in intronic regions, while the eCLIP peaks of YWHAZ are slightly enriched in 5′ UTR and proximal intronic (defined as the intronic regions that are with the distance <500 bp from exon-intron junction) regions (Figures 4D and S6C). Furthermore, we found high overlap in the

RNA targets of YWHAG, YWHAH, and YWHAE, where about 70% of the RNA targets of YWHAH and YWHAE are shared with YWHAG (Figure 4E). The locations of the binding sites of YWHAH and YWHAE are also highly correlated with YWHAG as we plotted the density of sequence reads around YWHAG eCLIP peaks (Figure 4F), whereas YWHAZ, RBFOX2, and negative controls (i.e., the reads from size-match INPUT sample of YWHAG and empty vector) show no significant binding behaviors around YWHAG's eCLIP peaks. An example of YTDC1, a regulator of alternative splicing, was shown with genome browser tracks (Figure S6D). GO analysis of the common RNA targets of these three proteins shows an enrichment in RNA splicing and translational regulation process (Figure S6F). All of these results suggest that YWHAG, YWHAH, and YWHAE are probably working synergistically or antagonistically on RNAs and may participate in the regulation of RNA splicing and translation processes.

Occlusion map interprets sequence-based RBP prediction and highlights functional elements in RBPs

To further interpret the sequence-based component of HydRA for RBP classification, we applied occlusion analysis, a widely used method in CV,[20–22] which we refer to here as occlusion map[20–22] (Figure S7A). Occlusion map involves occluding a window of twenty consecutive amino acids from the protein sequence in the inference step and comparing the model output score with the score obtained using the intact protein sequence. The difference in scores (Δscore), accounting for protein lengths and model-specific score distributions, determines the relevance of the occluded region to the protein's RNA-binding ability (Figure S7A; details in STAR Methods). A negative Δscore indicates that the occluded region positively contributed to the RNA binding. We used the p values generated from the Δscores to determine a region's predicted ability to participate in RNA binding. We define continuous stretches of occlusion windows with p values < 0.05 as "predicted RNA-binding regions" (pRBRs) and with p values < 0.001 as stringent pRBRs.

We present the occlusion maps for eight well-known RBPs involved in various RNA-processing roles ranging from splicing to mRNA decay and translation (RBFOX2, ZFP36, ATXN2, and RPLP0 in Figure 5A; and PUF60, LSM11, EIF4E, and YBX2 in Figure S8A). In RBFOX2 (Figure 5A), we observed a contiguous area of negative Δscores with stringent pRBRs that correspond well with its known RRM. We also found a contiguous area with weaker but significant occlusion signals (pRBRs) ($p < 0.05$, Figure 5A) in a Calcitonin gene-related peptide regulator C-terminal domain (Fox-1_C), which has been described to regulate alternative splicing events by binding to 5′-UGCAUGU-3′ elements.[45] Similarly, we observed two peaks of stringent pRBRs in the two zinc-finger domains (zf-CCCH) described to bind RNA in ZFP36.[46] Since this zinc-finger domain consists of 22 amino acids, this demonstrates that our sequence-based classifier with occlusion map can recognize RBP regions at high resolution. There are also strong correspondences between Δscore peaks within pRBRs/stringent pRBPs and diverse annotated RBDs found in ATXN2, RPLP0, PUF60, LSM11, EIF4E, and YBX2 (Figures 5A and S8A).

In general, we discover that annotated RNA-binding or -processing-related domains are 2.5× more likely to contain pRBRs ($p < 0.05$) than non-RNA-binding regions of the human proteins (Figure 5B), as evaluated by positive likelihood ratios (see STAR Methods). The ratio increases to 4.5× when considering stringent pRBRs ($p < 0.001$) (Figure S8D). Additionally, the high-confidence RBDs, supported by experimentally identified RNA-binding peptides (RBDpeps) studies,[5,47–49] exhibit a higher enrichment of pRBRs and stringent RBRs compared with RBDs or RNA-processing domains without experimental evidence (Figures 5B and S8D), indicating the robustness of occlusion map in detecting bona fide RBDs. Moreover, although HydRA was not explicitly trained with RNA-binding annotations, it achieves slightly better performance on the domain-level characterization of RNA-binding elements than the baseline tools in the field, such as RNABindRPlus[50] and DRNApred,[51] which were originally designed to predict RNA-binding residues in RBPs (Figure 5C). To uncover potential new RNA-binding-associated domains from HydRA's perspective, we obtained occlusion score distributions for each protein domain from the human protein sequences and visualized them as heatmaps (Figures 5D, S8B, and S8C), with domains sorted by the mean of their occlusion score distributions. One-sample t test and Benjamini-Hochberg method for multiple test correction recognized the domains that have Δscores significantly less than 0 (predicted RBDs) or greater than 0 (predicted non-RBDs). We included domains with more than 5 human protein members, resulting in a total of 1,165 domains. We observed the majority of the RNA-binding and RNA-processing-related domains (including LSM, CSD, RRM, KH, zf-CCCH, etc.) appear in the top 100 of the sorted domain list with significant p values (Figure S8B). A large majority of the bottom 100 domains are unrelated to RNA-binding or RNA-processing except RNA polymerase Rpb2 domain 7, which is known for DNA-binding[52] and is not statistically significant (FDR-adjusted p value = 0.102) (Figure S8C). GSEA analysis further confirms that most known RBDs and RNA-processing-related domains are enriched at the top of this sorted list (normalized enrichment score [ES] = 3.237, p value < 0.001, Figure 5E). Furthermore, among the domains with significant p values (76 out of 104 domains, FDR-adjusted $p < 0.05$), many are not canonical RBDs or known to be related to RNA processing, indicating the discovery of 76 new RNA-binding-associated domains through this analysis (Table S3).

Validation of RNA-binding activity and RNA-binding elements of candidate RBPs

We next evaluated HydRA's ability to predict RNA-binding activity at the region level for a selection of candidate RBPs harboring domains with significant occlusion delta scores. We selected candidates with representative domains, such as zinc-finger (zf) and EF hand domains, that received top RBD predictive scores from occlusion map analysis (Figure S8B). Individual occlusion mapping across the 5 candidate RBPs INO80B, PIAS4, NR5A1, ACTN3, and MCCC1 revealed specific domain regions that are predicted to bind RNA (Figure 6A). We performed eCLIP on these V5-tagged RBP candidates, with and without these occlusion-predicted RNA-binding elements, and assessed transcriptome-wide RNA targets (Figure 6B). In total, we obtained 48 eCLIP libraries (including SMInput) for these 5 candidate RBPs. Each library was sequenced to 5.7–57.0 million reads, of which ~0.42–16.0 million reads mapped uniquely to the human genome. We used a stringent peak-calling workflow that included both size-matched

CellPress
OPEN ACCESS

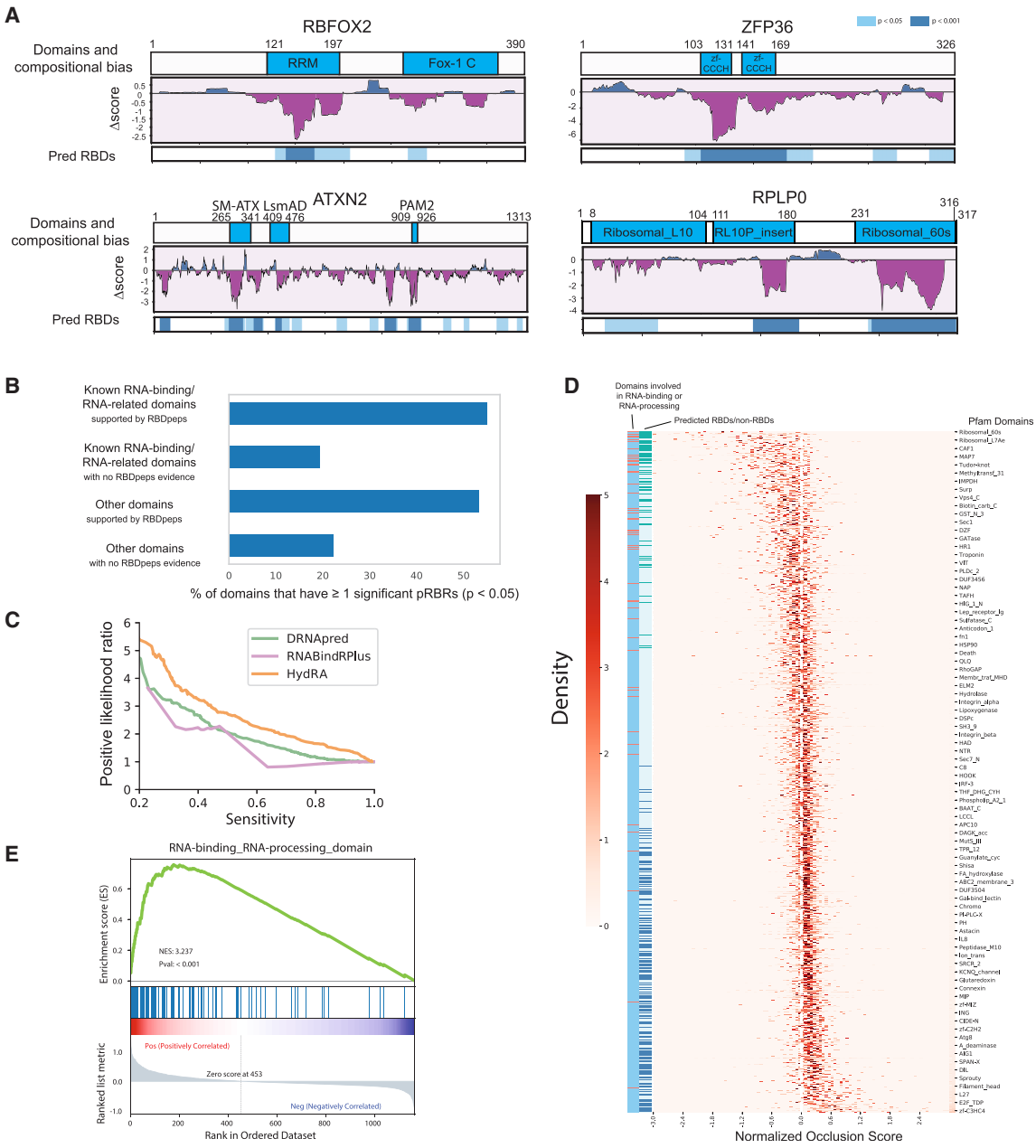


Figure 5. Occlusion map enables model interpretation of HydRA-seq and novel RNA-binding-associated domain discovery

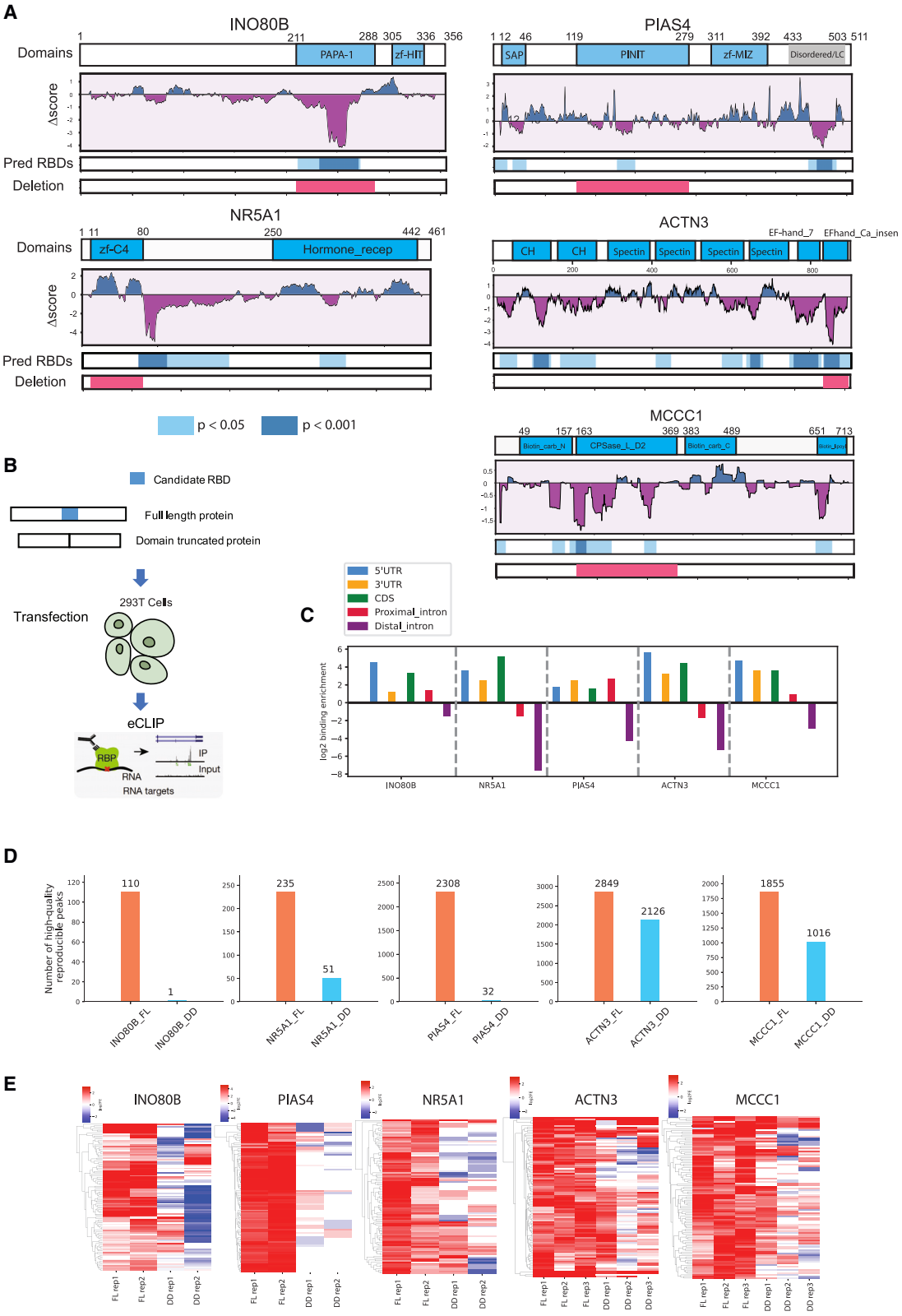
(A) Occlusion maps of known RBPs (RBFOX2, ZFP36, ATXN2, and RPLP0). The coordinates of protein domains, compositional bias, and RBDpeps are shown. Purple regions indicate Δscore < 0, while blue regions indicate Δscore > 0. Significant regions ($p < 0.05$ or $p < 0.001$) are marked in sky blue or dark blue, respectively, on the bottom track.

(B) We grouped the protein domains in known RBPs to four categories: RBDs that are supported by RBDpeps studies, RBDs that lack support from current RBDpeps studies, other domains with and without evidence from RBDpeps studies. Bar plot represents the percentage of domains within each category having occluders with significant negative Δscores (p value < 0.05).

(C) Domain-level RBD detection performance evaluated using positive likelihood ratio at different sensitivity levels for baseline algorithms (DNAPred and RNABindRPlus) and HydRA occlusion map.

(D) Heatmap showing the distribution of occlusion scores for human protein domains with more than 5 human protein members. Known RNA-binding or RNA-processing domains are marked in red on the left annotation bar of the heatmap. The right annotation bar shows the predicted RBD (mean > 0 and FDR-adjusted p value < 0.05, one-sample t test) in green and predicted non-RBD (mean > 0 and FDR-adjusted p value < 0.05) in dark blue.

(E) GSEA analysis shows enrichment of RNA-binding or RNA-processing-related domains in the top-ranked human domains. The panel displays the enrichment score (ES), position of RBD/RNA-processing-related domains in the ranked list, and the ranking metric value as genes are ranked.



(legend on next page)

and V5-tag-only inputs as well as both Clipper[53] and ChIP-R reproducible peak-calling algorithms[54] (Figure S9C), which identified between 110 (INO80B) and 2,849 (ACTN3) RNA-binding peaks per candidate (Figure 6D). Ontology analysis of bound targets revealed that INO80B bound genes are involved in translation and have binding sites enriched in 5′ UTR regions, PIAS4 bound genes are involved in signaling and differentiation and have sites in proximal intronic regions, NR5A1 bound genes are involved in mitosis with CDS binding sites, ACTN3 bound genes are involved in ER-stress with sites in 5′ UTR regions, and MCCC1 bound genes are involved in cell morphology and polarity with sites broadly across 5′ UTR, CDS and 3′ UTR regions (Figures 6C and S9D). These RNA-binding profiles suggest roles for these candidates in regulating the cytoplasmic localization and/or translation of target genes in the cases of UTR and CDS binding, and alternative splicing regulation in the case of intronic binding.

Removal of putative RNA-binding-associated domains with significant delta-occlusion scores for these candidates did not influence expression of V5-fusions as assessed by western blot analysis (Figure S9A). However, absence of these predicted domains did result in substantially higher PCR cycle requirements (eCT values) to obtain eCLIP libraries from immunoprecipitated RNA (Figure S9B), indicating the importance of the removed domains for RNA-binding. By comparing the eCLIP peaks of the full-length (FL) proteins and domain-deleted (DD) proteins (Figure 6D), we observed that these putative RBD deletions coincided with a global loss of 2,779 eCLIP peaks on 2,089 targets (out of 2,849 peaks on 2,123 targets) for ACTN3, and 1,820 peaks on 1,576 targets (out of 1,855 peaks on 1,597 targets) for MCCC1, 110 peaks on 151 targets (out of 110 peaks on 151 targets) for INO80B, 2,308 peaks for eCLIP peaks on 1,187 targets (out of 2,308 peaks on 1,187 targets) for PIAS4, 232 peaks on 235 targets (out of 235 peaks on 238 targets) for NR5A1. Analysis of the top 100 binding peaks for these RBP candidates upon candidate domain deletion revealed highly reproducible reduction in RNA-binding across these high-confidence binding sites (Figure 6E). This drastic loss of RNA-binding upon domain deletion demonstrates the high-resolution accuracy with which HydRA predicts RNA-protein interactions. Together these results validate the power of the HydRA workflow for both RBP (gene-level) and RNA-binding-associated domain (amino acid level) discovery.

DISCUSSION

We developed HydRA, a hybrid machine-learning-based RBP classifier that combines local PPI networks and amino acid sequence signatures. HydRA demonstrates exceptional sensitivity, specificity, and precision in RBP recognition. The correlation of HydRA scores with RBP conservation across species and experiments makes it a valuable tool for reducing FDRs and complementing biochemical high-throughput RBP discovery approaches. Additionally, we employed the occlusion map technique[20–22] on protein sequences to interpret HydRA-seq and identify RNA-binding-associated domains without explicit training using RNA-binding information (such as RBDs and RNA-binding residues). Through a proteome-wide HydRA search, we generated extensive candidate lists of RBPs and RNA-binding-associated domains. We experimentally validated the RNA-binding capability of five HydRA-predicted RBP candidates with occlusion map-predicted domains, revealing novel RBPs along with previously unknown and conserved RNA-binding-associated domains.

The efficacy of occlusion map extends HydRA's applicability to *in silico* perturbation analyses, enabling the evaluation of how protein sequence variations influence RNA-binding capacity. Moreover, HydRA's flexibility and independence from prior knowledge of RBPs (such as RBDs) allow it to be easily transferred to RBP prediction in other species (Figure S4) and even other protein function prediction scenarios with appropriate training datasets.

The current occlusion map approach leverages solely the sequence-based component of HydRA, HydRA-seq, without PPI and association features. Although the structure-based models, strucGNNs, were not included in HydRA due to their limited impact on overall classification performance, they exhibited improved ability in predicting RNA-binding by RBPs with uncharacterized RBDs (Figures S2D and S2E). This indicates the potential of strucGNNs to enhance the modeling of novel RBDs using the predicted protein structures. Further research on the enhanced integration of PPI, sequence, and structure information in the design and interpretation of models will be valuable for better understanding RBPs and RNA-binding-associated domains.

The HydRA algorithm is a powerful protein context- and sequence-based method for the *de novo* identification of candidate RBPs and RNA-binding regions. Future efforts to validate and characterize the extensive list of HydRA-predicted RBP candidates and novel RNA-binding-associated domains will deepen our understanding of RBP roles in diverse biological processes. HydRA serves as a framework for predicting RBPs and potentially other protein classes where both protein context and sequence features contribute substantially to functional activity. Additionally, HydRA's generative component (i.e., ProteinBERT-RBP), provides an opportunity for novel RBP design tailored for specific biological tasks.

Limitations of the study

We want to caution that the occlusion map-predicted domains may participate in the interaction with RNA molecules indirectly

Figure 6. Novel RBPs and functional regions related to RNA-binding were predicted using HydRA and experimentally validated

(A) Occlusion maps of INO80B, NR5A1, PIAS4, ACTN3, and MCCC1. The candidate domain to be deleted were colored red on the bottom track of each map.

(B) Workflow to experimental validation of candidate RBPs and their predicted RNA-binding-associated domain. Original and truncated protein (with predicted functional domain deleted) were expressed in HEK293T cells followed by eCLIP experiments to investigate their RNA-binding behaviors.

(C) Binding preference of INO80B, NR5A1, PIAS4, ACTN3, and MCCC1 on different genomic regions measured by $\log_2$-transformed region enrichment E_{region} (see STAR Methods). Genomic regions shown include 5′ UTR, 3′ UTR, coding sequence (CDS), proximal introns, and distal intron.

(D) Number of reproducible eCLIP peaks for full-length and truncated protein of INO80B, NR5A1, PIAS4, ACTN3, and MCCC1.

(E) Heatmap displaying binding strength (measured by log-transformed fold-enrichment) across full-length (FL) and truncated (DD) protein samples on the 100 most significant peaks found in the full-length proteins.

since training of the HydRA models does not involve detailed RBD and residue information. New methods that integrate known RNA-binding residues or protein-RNA interfaces into the RBP prediction will be required to give higher-resolution predictions of domains that directly contact RNA molecules.

Our domain deletion experiments, performed using overexpressed V5-tagged proteins, have shown functional association between the predicted domains and RNA-binding in HEK293T cells. We note that V5-tagged RBP-overexpression eCLIPs may not fully represent physiological RNA-binding profiles that may be more accurately detected with antibodies targeting endogenous cellular proteins. We utilized this tagging strategy due to limited availability of eCLIP-grade antibodies for these targets, and the need to delete large regions specifically for RBD assessment. However, further studies using approaches such as saturating CRISPR-Cas9 screens on endogenous proteins in multiple cell lines will likely obtain a more comprehensive and detailed characterization of the occlusion map-predicted RNA-binding associated domains.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Mentha-BioPlex network
  - The functional protein association network
  - Protein dataset and Known RBPs collection
  - Network properties' effects on SONAR
  - Rationale for HydRA
  - Dataset for modeling
  - Protein interaction and association-based classifier
  - Sequence-based SVM classifier
  - Convolutional neural network for sequence-based RBP classification
  - Attention-based protein language model for RBP classification
  - Graph neural network modelling on protein structures
  - Finetuning pretrained graph neural network on RNA-binding protein structures
  - Meta ensembling of classifiers
  - Software comparisons
  - RNA-binding domain sequences and RNA-binding peptide
  - Protein domain annotation
  - Occlusion map
  - Criteria of RBP candidacy
  - Biophysical properties of proteins
  - Antibodies for western blot and immunoprecipitation
  - Plasmid construction
  - Generation of cell lines
  - RNA immunoprecipitation and qPCR
  - eCLIP Library Preparation for YWHAG/H/E/Z and HSP90AA1
  - eCLIP Library Preparation for INO80B, PIAS4, NR5A1, ACTN3, and MCCC1
  - eCLIP Library Preparation for V5 tag control
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Evaluating domain-level RNA-binding element prediction
  - eCLIP data processing and analysis
  - Peak enrichment analysis
  - GO enrichment for candidate RBPs
  - Other statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.molcel.2023.06.019.

ACKNOWLEDGMENTS

G.W.Y. is supported by NIH R01 HG004659, NIH U24 HG009889, an Allen Distinguished Investigator Award, and a Paul G. Allen Frontiers Group advised grant of the Paul G. Allen Foundation. K.W.B. is supported by NIH/NINDS K22 K22NS112678 and CPRIT Award RR220017.

AUTHOR CONTRIBUTIONS

W.J. and G.W.Y. conceived the study. W.J. designed HydRA algorithm, developed the software, tested the software, collected data, analyzed data, and visualized results. K.W.B., K.K., G.W.Y., W.J., and J.S.X. designed wet-lab validation experiments. K.W.B., K.K., S.S.P., H.Q.T., M.L.G., M.M., and J.A. carried out the experimental work. W.J., G.W.Y., and L.W. designed computational validation experiments. W.J. and B.H. carried out the computational validation experiments. W.J., K.W.B., and K.K. wrote the original manuscript draft. W.J., K.W.B., K.K., G.W.Y., S.S.P., and K.R. reviewed and edited the manuscript. G.W.Y. acquired the funding and supervised the study.

DECLARATION OF INTERESTS

G.W.Y. is a co-founder, member of the Board of Directors, on the SAB, equity holder, and paid consultant for Locanabio and Eclipse BioInnovations. G.W.Y. is a visiting professor at the National University of Singapore. G.W.Y.'s interests have been reviewed and approved by the University of California, San Diego in accordance with its conflict-of-interest policies.

Received: January 3, 2023
Revised: March 20, 2023
Accepted: June 13, 2023
Published: July 7, 2023

REFERENCES

1. Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. Nat. Rev. Genet. *15*, 829–845. https://doi.org/10.1038/nrg3813.
2. Lukong, K.E., Chang, K.W., Khandjian, E.W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. Trends Genet. *24*, 416–425. https://doi.org/10.1016/j.tig.2008.05.004.
3. Castello, A., Fischer, B., Hentze, M.W., and Preiss, T. (2013). RNA-binding proteins in Mendelian disease. Trends Genet. *29*, 318–327. https://doi.org/10.1016/J.TIG.2013.01.004.
4. Hentze, M.W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. Nat. Rev. Mol. Cell Biol. *19*, 327–341. https://doi.org/10.1038/nrm.2017.130.

5. Castello, A., Fischer, B., Frese, C.K., Horos, R., Alleaume, A.M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M.W. (2016). Comprehensive identification of RNA-binding domains in human cells. Mol. Cell *63*, 696–710. https://doi.org/10.1016/J.MOLCEL.2016.06.029.

6. Beckmann, B.M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.M., Schwarzl, T., Curk, T., Foehr, S., Huber, W., et al. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat. Commun. *6*, 10127. https://doi.org/10.1038/ncomms10127.

7. Baltz, A.G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., et al. (2012). The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol. Cell *46*, 674–690. https://doi.org/10.1016/j.molcel.2012.05.021.

8. Conrad, T., Albrecht, A.-S., de Melo Costa, V.R., Sauer, S., Meierhofer, D., and Ørom, U.A. (2016). Serial interactome capture of the human cell nucleus. Nat. Commun. *7*, 11212. https://doi.org/10.1038/ncomms11212.

9. Queiroz, R.M.L., Smith, T., Villanueva, E., Marti-Solano, M., Monti, M., Pizzinga, M., Mirea, D.-M., Ramakrishna, M., Harvey, R.F., Dezi, V., et al. (2019). Comprehensive identification of RNA–protein interactions in any organism using orthogonal organic phase separation (OOPS). Nat. Biotechnol. *37*, 169–178. https://doi.org/10.1038/s41587-018-0001-2.

10. Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M.W., and Krijgsveld, J. (2019). The human RNA-binding proteome and its dynamics during translational arrest. Cell *176*, 391–403.e19. https://doi.org/10.1016/j.cell.2018.11.004.

11. Kumar, M., Gromiha, M.M., and Raghava, G.P.S. (2011). SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J. Mol. Recognit. *24*, 303–313. https://doi.org/10.1002/jmr.1061.

12. Zhao, H., Yang, Y., Janga, S.C., Kao, C.C., and Zhou, Y. (2014). Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome. Proteins *82*, 640–647. https://doi.org/10.1002/prot.24441.

13. Livi, C.M., Klus, P., Delli Ponti, R., and Tartaglia, G.G. (2016). catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics *32*, 773–775. https://doi.org/10.1093/bioinformatics/btv629.

14. Zhang, X., and Liu, S. (2017). RBPPred: predicting RNA-binding proteins from sequence using SVM. Bioinformatics *33*, 854–862. https://doi.org/10.1093/bioinformatics/btw730.

15. Bressin, A., Schulte-Sasse, R., Figini, D., Urdaneta, E.C., Beckmann, B.M., and Marsico, A. (2019). TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. Nucleic Acids Res. *47*, 4406–4417. https://doi.org/10.1093/nar/gkz203.

16. Brannan, K.W., Jin, W., Huelga, S.C., Banks, C.A.S., Gilmore, J.M., Florens, L., Washburn, M.P., Van Nostrand, E.L., Pratt, G.A., Schwinn, M.K., et al. (2016). SONAR discovers RNA-binding proteins from analysis of large-scale protein-protein interactomes. Mol. Cell *64*, 282–293. https://doi.org/10.1016/j.molcel.2016.09.003.

17. Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. (2020). Transforming the language of life: transformer neural networks for protein prediction tasks ACM reference format. In Proceedings of the 11th ACM International Conference on Bioinformatics (Computational Biology and Health Informatics). https://doi.org/10.1145/3388440.

18. Yang, X., He, X., Liang, Y., Yang, Y., Zhang, S., and Xie, P. (2020). Transfer learning or self-supervised learning? A tale of two pretraining paradigms. https://doi.org/10.36227/techrxiv.12502298.

19. Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics *38*, 2102–2110. https://doi.org/10.1093/BIOINFORMATICS/BTAC020.

20. Zeiler, M.D., and Fergus, R. (2014). Visualizing and understanding convolutional networks arXiv:1311.2901v3. Comput. Vis. *8689*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53.

21. Brunetti, A., Buongiorno, D., Trotta, G.F., and Bevilacqua, V. (2018). Computer vision and deep learning techniques for pedestrian detection and tracking: A survey. Neurocomputing *300*, 17–33. https://doi.org/10.1016/J.NEUCOM.2018.01.092.

22. Sáez Trigueros, D., Meng, L., and Hartnett, M. (2018). Enhancing convolutional neural networks for face recognition with occlusion maps and batch triplet loss. Image Vis. Comput. *79*, 99–108. https://doi.org/10.1016/J.IMAVIS.2018.09.011.

23. Huttlin, E.L., Ting, L., Bruckner, R.J., Gebreab, F., Gygi, M.P., Szpyt, J., Tam, S., Zarraga, G., Colby, G., Baltier, K., et al. (2015). The BioPlex network: A systematic exploration of the human interactome. Cell *162*, 425–440. https://doi.org/10.1016/j.cell.2015.06.043.

24. Calderone, A., Castagnoli, L., and Cesareni, G. (2013). mentha: a resource for browsing integrated protein-interaction networks. Nat. Methods *10*, 690–691. https://doi.org/10.1038/nmeth.2561.

25. Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857–D861. https://doi.org/10.1093/nar/gkr930.

26. Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). BioGRID: a general repository for interaction datasets. Nucleic Acids Res. *34*, D535–D539. https://doi.org/10.1093/nar/gkj109.

27. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. *42*, D358–D363. https://doi.org/10.1093/nar/gkt1115.

28. Yong, C.H., Liu, G., Chua, H.N., and Wong, L. (2012). Supervised maximum-likelihood weighting of composite protein networks for complex prediction. BMC Syst. Biol. *6*, S13. https://doi.org/10.1186/1752-0509-6-S2-S13.

29. Asgari, E., and Mofrad, M.R.K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS One *10*, e0141287. https://doi.org/10.1371/journal.pone.0141287.

30. Zhao, W., Zhang, S., Zhu, Y., Xi, X., Bao, P., Ma, Z., Kapral, T.H., Chen, S., Zagrovic, B., Yang, Y.T., et al. (2022). POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res. *50*, D287–D294. https://doi.org/10.1093/NAR/GKAB702.

31. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

32. Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nat. Commun. *12*, 3168. https://doi.org/10.1038/s41467-021-23303-9.

33. Jha, K., Saha, S., and Singh, H. (2022). Prediction of protein–protein interaction using graph neural networks. Sci. Rep. *12*, 8360. https://doi.org/10.1038/s41598-022-12201-9.

34. Strokach, A., Becerra, D., Corbi-Verge, C., Perez-Riba, A., and Kim, P.M. (2020). Fast and flexible protein design using deep graph neural networks. Cell Syst. *11*, 402–411.e4. https://doi.org/10.1016/J.CELS.2020.08.016.

35. Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B.M., Strein, C., Davey, N.E., Humphreys, D.T., Preiss, T., Steinmetz, L.M., et al. (2012). Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell *149*, 1393–1406. https://doi.org/10.1016/j.cell.2012.04.031.

36. Huang, Y.W., Hu, C.C., Liou, M.R., Chang, B.Y., Tsai, C.H., Meng, M., Lin, N.S., and Hsu, Y.H. (2012). Hsp90 interacts specifically with viral RNA and differentially regulates replication initiation of Bamboo mosaic virus and associated satellite RNA. PLoS Pathog. *8*, e1002726. https://doi.org/10.1371/journal.ppat.1002726.

37. Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004.

38. Liepelt, A., Naarmann-de Vries, I.S., Simons, N., Eichelbaum, K., Föhr, S., Archer, S.K., Castello, A., Usadel, B., Krijgsveld, J., Preiss, T., et al. (2016). Identification of RNA-binding proteins in macrophages by interactome capture. Mol. Cell. Proteomics *15*, 2699–2714. https://doi.org/10.1074/mcp.M115.056564.

39. Genest, O., Wickner, S., and Doyle, S.M. (2019). Hsp90 and Hsp70 chaperones: collaborators in protein remodeling. J. Biol. Chem. *294*, 2109–2120. https://doi.org/10.1074/jbc.REV118.002806.

40. Fu, H., Subramanian, R.R., and Masters, S.C. (2000). 14-3-3 proteins: structure, function, and regulation. Annu. Rev. Pharmacol. Toxicol. *40*, 617–647. https://doi.org/10.1146/annurev.pharmtox.40.1.617.

41. Pennington, K.L., Chan, T.Y., Torres, M.P., and Andersen, J.L. (2018). The dynamic and stress-adaptive signaling hub of 14–3-3: emerging mechanisms of regulation and context-dependent protein–protein interactions. Oncogene *37*, 5587–5604. https://doi.org/10.1038/s41388-018-0348-3.

42. Wang, B., Underwood, R., Kamath, A., Britain, C., McFerrin, M.B., McLean, P.J., Volpicelli-Daley, L.A., Whitaker, R.H., Placzek, W.J., Becker, K., et al. (2018). 14-3-3 proteins reduce cell-to-cell transfer and propagation of pathogenic α-synuclein. J. Neurosci. *38*, 8211–8232. https://doi.org/10.1523/JNEUROSCI.1134-18.2018.

43. Zhang, J., and Zhou, Y. (2018). 14-3-3 proteins in glutamatergic synapses. Neural Plast. *2018*, 8407609. https://doi.org/10.1155/2018/8407609.

44. Yuan, L., Barbash, S., Kongsamut, S., Eishingdrelo, A., Sakmar, T.P., and Eishingdrelo, H. (2019). 14-3-3 signal adaptor and scaffold proteins mediate GPCR trafficking. Sci. Rep. *9*, 11156. https://doi.org/10.1038/s41598-019-47478-w.

45. Ponthier, J.L., Schluepen, C., Chen, W., Lersch, R.A., Gee, S.L., Hou, V.C., Lo, A.J., Short, S.A., Chasis, J.A., Winkelmann, J.C., et al. (2006). Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. J. Biol. Chem. *281*, 12468–12474. https://doi.org/10.1074/jbc.M511556200.

46. Fu, M., and Blackshear, P.J. (2016). RNA-binding proteins in immune regulation: a focus on CCCH zinc finger proteins. Nat. Rev. Immunol. *17*, 130–143. https://doi.org/10.1038/nri.2016.129.

47. Liao, Y., Castello, A., Fischer, B., Leicht, S., Föehr, S., Frese, C.K., Ragan, C., Kurscheid, S., Pagler, E., Yang, H., et al. (2016). The cardiomyocyte RNA-binding proteome: links to intermediary metabolism and heart disease. Cell Rep. *16*, 1456–1469. https://doi.org/10.1016/J.CELREP.2016.06.084.

48. Mullari, M., Lyon, D., Jensen, L.J., and Nielsen, M.L. (2017). Specifying RNA-binding regions in proteins by peptide cross-linking and affinity purification. J. Proteome Res. *16*, 2762–2772. https://doi.org/10.1021/acs.jproteome.7b00042.

49. Kramer, K., Sachsenberg, T., Beckmann, B.M., Qamar, S., Boon, K.-L., Hentze, M.W., Kohlbacher, O., and Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. Nat. Methods *11*, 1064–1070. https://doi.org/10.1038/nmeth.3092.

50. Walia, R.R., Xue, L.C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2014). RNABindRPlus: A predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLOS One. https://doi.org/10.1371/journal.pone.0097725.

51. Yan, J., and Kurgan, L. (2017). DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res *45*, e84.

52. Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. *49*, D344–D354. https://doi.org/10.1093/NAR/GKAA977.

53. van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514. https://doi.org/10.1038/nmeth.3810.

54. Newell, R., Pienaar, R., Balderson, B., Piper, M., Essebier, A., and Bodén, M. (2021). ChIP-R: assembling reproducible sets of ChIP-seq and ATAC-seq peaks from multiple replicates. Genomics *113*, 1855–1866. https://doi.org/10.1016/J.YGENO.2021.04.026.

55. Chollet, F. (2015). Keras. GitHub. https://github.com/fchollet/keras.

56. Van Nostrand, E.L., Pratt, G.A., Shishkin, A.A., Gelboin-Burkhart, C., Fang, M.Y., Sundararaman, B., Blue, S.M., Nguyen, T.B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat. Methods *13*, 508–514.

57. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: machine learning in python. J. Mach. Learn. Res. *12*, 2825–2830.

58. Fey, M., and Lenssen, J.E. (2019). Fast graph representation learning with PyTorch Geometric. https://doi.org/10.48550/arxiv.1903.02428.

59. Hagberg, A.A., Schult Daniel, A., and Swart, P.J. (2008). Exploring network structure, dynamics, and function using NetworkX. In Proceedings of the 7th Python in Science Conference, *11–15*. https://conference.scipy.org/proceedings/SciPy2008/paper_2/.

60. Huttlin, E.L., Bruckner, R.J., Paulo, J.A., Cannon, J.R., Ting, L., Baltier, K., Colby, G., Gebreab, F., Gygi, M.P., Parzen, H., et al. (2017). Architecture of the human interactome defines protein communities and disease networks. Nature *545*, 505–509. https://doi.org/10.1038/nature22366.

61. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. *32*, D449–D451. https://doi.org/10.1093/nar/gkh086.

62. Launay, G., Salza, R., Multedo, D., Thierry-Mieg, N., and Ricard-Blum, S. (2015). MatrixDB, the extracellular matrix interaction database: updated content, a new navigator and expanded functionalities. Nucleic Acids Res. *43*, D321–D327. https://doi.org/10.1093/nar/gku1091.

63. Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., et al. (2017). The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic Acids Res. *45*, D362–D368. https://doi.org/10.1093/nar/gkw937.

64. Sundararaman, B., Zhan, L., Blue, S.M., Stanton, R., Elkins, K., Olson, S., Wei, X., Van Nostrand, E.L., Pratt, G.A., Huelga, S.C., et al. (2016). Resources for the comprehensive discovery of functional RNA elements. Mol. Cell *61*, 903–913. https://doi.org/10.1016/j.molcel.2016.02.012.

65. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. (2000). Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29. https://doi.org/10.1038/75556.

66. Gene Ontology Consortium (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res. *43*, D1049–D1056. https://doi.org/10.1093/nar/gku1179.

67. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

68. Cao, D.-S., Xu, Q.-S., and Liang, Y.-Z. (2013). propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics *29*, 960–962. https://doi.org/10.1093/bioinformatics/btt072.

69. Liou, C.-Y., Cheng, W.-C., Liou, J.-W., and Liou, D.-R. (2014). Autoencoder for words. Neurocomputing *139*, 84–96. https://doi.org/10.1016/J.NEUCOM.2013.09.055.

70. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., et al. (2016). TensorFlow: large-scale machine learning on heterogeneous distributed systems. https://doi.org/10.48550/arXiv.1603.04467.

71. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282–1288. https://doi.org/10.1093/BIOINFORMATICS/BTM098.

72. van Kempen, M., Kim, S.S., Tumescheit, C., Mirdita, M., Gilchrist, C.L.M., Söding, J., and Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. https://doi.org/10.1101/2022.02.07.479398.

73. Zhao, H., Jiang, L., Jia, J., Torr, P., and Koltun, V. (2020). Point transformer. In Proceedings of the IEEE International Conference on Computer Vision, pp. 16239–16248. https://doi.org/10.1109/ICCV48922.2021.01595.

74. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: an imperative style, high-performance deep learning library. https://doi.org/10.48550/arxiv.1912.01703.

75. Yang, K.K., Eleutherai, N.Z., and Yeh, H. (2022). Masked inverse folding with sequence transfer for protein representation learning. https://doi.org/10.1101/2022.05.25.493516.

76. Sillitoe, I., Bordin, N., Dawson, N., Waman, V.P., Ashford, P., Scholes, H.M., Pang, C.S.M., Woodridge, L., Rauer, C., Sen, N., et al. (2021). CATH: increased structural coverage of functional space. Nucleic Acids Res. *49*, D266–D273. https://doi.org/10.1093/NAR/GKAA1079.

77. He, C., Sidoli, S., Warneford-Thomson, R., Tatomer, D.C., Wilusz, J.E., Garcia, B.A., and Bonasio, R. (2016). High-resolution mapping of RNA-binding regions in the nuclear proteome of embryonic stem cells. Mol. Cell *64*, 416–430. https://doi.org/10.1016/J.MOLCEL.2016.09.034.

78. Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., Duvaud, S., Flegel, V., Fortier, A., Gasteiger, E., et al. (2012). Expasy: SIB bioinformatics resource portal. Nucleic Acids Res. *40*, W597–W603. https://doi.org/10.1093/nar/gks400.

79. Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434. https://doi.org/10.1093/bioinformatics/bti541.

80. Lovci, M.T., Ghanem, D., Marr, H., Arnold, J., Gee, S., Parra, M., Liang, T.Y., Stark, T.J., Gehman, L.T., Hoon, S., et al. (2013). Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nat. Struct. Mol. Biol. *20*, 1434–1442. https://doi.org/10.1038/nsmb.2699.

81. Van Nostrand, E.L., Freese, P., Pratt, G.A., Wang, X., Wei, X., Xiao, R., Blue, S.M., Chen, J.Y., Cody, N.A.L., Dominguez, D., et al. (2020). A large-scale binding and functional map of human RNA-binding proteins. Nature *583*, 711–719. https://doi.org/10.1038/s41586-020-2077-3.

82. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

83. Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P.D. (2017). PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. *45*, D183–D189. https://doi.org/10.1093/nar/gkw1138.

84. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., and McDermott, M.G. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic acids research *44*, W90–W97.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
anti-V5	Proteintech	66007-1-Ig
anti-His	Proteintech	66005-1-Ig
anti-His	Invitrogen	MA1-21315
anti-a-Tubulin	Sigma	Clone B-5-1-2
anti-HSP90B	Sigma	SAB4501463
anti-HNRNPA2B1	Proteintech	14813-1-AP
Bethyl anti-V5	Bethyl	Bethyl A190-120A
Bacterial and virus strains		
One Shot *E. Coli* TOP10	Thermo Scientific	Cat# C404003
DH5α Competent Cells	Thermo Scientific	Cat# 18265017
Lentivirus	Lenti-X HEK293T	Cat# 632180
Chemicals, peptides, and recombinant proteins		
Puromycin	Sigma Aldrich	P8833-10MG
Lipofectamine 3000	Invitrogen	Cat# L3000008
Doxycycline	Sigma Aldrich	Cat# D9891-5G
Polybrene	Millipore Sigma	Cat# TR-1003-G
Critical commercial assays		
Pierce BCA Protein Assay Kit	ThermoFisher Scientific	Cat# 23225
QIAprep Spin Miniprep Kit	Qiagen	Cat# 27106
QIAquick PCR Purification Kit	Qiagen	Cat# 28106
Gateway BP Clonase II	Invitrogen	Cat# 11789020
Gateway LR Clonase II	Invitrogen	Cat# 11791020
Deposited data		
eCLIP sequencing data for the over-expressed full-length protein HSP90A, YWHAH, YWHAG, YWHAE, YWHAZ, INO80B, PIAS4, NR5A1, ACTN3, and MCCC1, and domain deleted protein INO80B (labeled as INO80B_DD), PIAS4 (PIAS4_DD), NR5A1 (NR5A1_DD), ACTN3 (ACTN3_DD), and MCCC1 (MCCC1_DD).	This study	GSE221870
Experimental models: Cell lines		
Lenti-X HEK293T	Takara Bio	Cat# 632180
Oligonucleotides		
qPCR primers (table)	This study	Table S5
Deletion cloning primers (table)	This study	Table S6
Recombinant DNA		
pEF5/FRT/HSP90AA1-V5-DEST	This study	N/A
pLIX403_V5_De	This study	N/A
pCDNA6-YWHAH-myc-6xHis	This study	N/A
pCDNA6-YWHAG-myc-6xHis	This study	N/A
pCDNA6-YWHAE-myc-6xHis	This study	N/A
pCDNA6-YWHAZ-myc-6xHis	This study	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MCCC1_pLIX403_V5_FL	This study	N/A
ACTN3_pLIX403_V5_FL	This study	N/A
INO80B_pLIX403_V5_FL	This study	N/A
PIAS4_pLIX403_V5_FL	This study	N/A
NR5A1_pLIX403_V5_FL	This study	N/A
MCCC1_pLIX403_V5_Del	This study	N/A
ACTN3_pLIX403_V5_Del	This study	N/A
INO80B_pLIX403_V5_Del	This study	N/A
PIAS4_pLIX403_V5_Del	This study	N/A
NR5A1_pLIX403_V5_Del	This study	N/A
Software and algorithms		
HydRA	This study	https://doi.org/10.5281/zenodo.7754206
Keras v2.6	Chollet et al.[55]	https://keras.io/
eCLIP analysis pipeline	Van Nostrand et al.[56]	https://github.com/YeoLab/eclip
ChIP-R	Newell et al.[54]	https://github.com/rhysnewell/ChIP-R
Scikit-learn v0.22.1	Pedregosa et al.[57]	https://scikit-learn.org
PyTorch-Geometric v2	Fey et al.[58]	https://pytorch-geometric.readthedocs.io/
NetworkX	Hagberg et al.[59]	https://networkx.org/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Gene Yeo (geneyeo@ucsd.edu).

Materials availability

Materials generated by the authors in this study will be distributed upon request.

Data and code availability

- All eCLIP sequencing data at GEO under the accession number GSE221870 All data are publicly available as of the date of publication. Accession numbers and DOI are listed in the key resources table
- All custom code and model weights associated with HydRA software and Occlusion Map have been deposited to GitHub (https://github.com/YeoLab/HydRA) and Zenodo. DOI are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Mentha-BioPlex network

We obtained a comprehensive protein-protein interaction (PPI) network, i.e. Mentha-BioPlex, by combining BioPlex2.0[60] a high-quality PPI dataset generated by robust affinity purification–mass spectrometry, with a meta-database Mentha[24] that collects experimental evidence from different PPI database such as MINT,[25] IntAct,[27] DIP,[61] MatrixDB[62] and BioGRID.[26] Each interaction in the network was assigned a reliability score taking into account all the aggregated experimental evidence, according to the formula defined in MINT database and Mentha:

$$S = 1 - a^{-x}$$

$\boldsymbol{a}$ is a constant. It determines the growth rate of the curve. To keep the scores in a convenient range, we chose $\boldsymbol{a}$=1.4 as Mentha did. $\boldsymbol{x}$ is calculated by adding up all the experimental evidence considering type and size of the experiments and the number of publications that support the interaction:

$$x = \sum_i d_i e_i + n/10$$

i is the index of all experimental evidence supporting the interaction, while ***d*** takes into consideration the experiment size (d=0.5 for large scale experiment reporting more than 50 interactions, otherwise d=1) and ***e*** reflects the type of information the experiment provides (e=1.0 for evidence of direct interaction and e=0.5 for evidence that only support and association). Additionally, ***x*** also takes into account the number (***n***) of experiments (publications) that support this interaction.

We then presented this network, consisting of only experimentally demonstrated physical protein-protein interactions, as an undirected graph with each protein as the node and the interaction between proteins as the edge. In total, this network contains 19,066 nodes and 309,653 edges.

The functional protein association network

To supplement current PPI network with extra extrinsic association information of proteins, we introduced functional protein association data into our modelling. We take STRING[63] as our functional protein association resource in this study. STRING is a database consisting of protein-protein association data from several sources, i.e. genomic context predictions, high-throughput lab experiments, conserved co-expression, automated text-mining and other PPI databases. To construct the functional protein association network, we kept all the STRING interactions that has at least one protein existing in Mentha-BioPlex network, while the experimentally demonstrated physical protein-protein interactions from STRING are removed to avoid redundancies.

Protein dataset and Known RBPs collection

All proteins in the Mentha-BioPlex network constituted our protein dataset and the corresponding protein sequences were retrieved from Uniprot database. We further collected currently known RBPs from published studies,[1,5–10,35,64] including high-throughput studies poly-A capture-based and organic phase separation-based RBP discovery approaches a comprehensive RBPome survey, and merged these with proteins in the GO term of ''RNA-binding'' (GO:0003723) and its descendent terms from AmiGO 2 database (version released in July 2016),[65,66] after filtering proteins assigned by automated methods without curatorial judgement. Each protein in our protein dataset was assigned a class label, viz. known RBP or not known for RBP (we referred to it as non-RBP for simplicity), based on this RBP list. In all, our dataset contained 18,951 UniProt protein IDs with sequences retrieved, among which 3,762 are known RBPs (also referred to as annotated RBPs).

Network properties' effects on SONAR

We studied the effects of PPI network's properties on the performance of SONAR, such as network density (also referred to as completeness) and the reliability of PPI edges. In this analysis, we first randomly removed certain numbers of edges from the original network that led to a couple of induced networks with retained edges in percentage of 20%, 40%, 60% and 80% respectively from the whole network. For each percentage, we generated 10 different networks as replicates. SONAR was then run with each of these induced networks and the performance was evaluated by ROC-AUC analysis with the output from SONAR software which was obtained by a cross-validation-like (10-fold) approach.[16] To test if the reliability of PPI edges in the network matters, we also generated a group of induced networks in the same way as above but from a random network that came from an edges-shuffle operation with the original network. ROC-AUC analysis was also done to evaluate the performance of SONAR with these induced networks.

On the other hand, to determine if the degree of reliability of the experimental PPI data matters, we sort the edges in the network based on their corresponding reliability scores (from high to low) and generated a couple of subnetworks from the original network by keeping the top 20%, 30%, 40%, 60% and 80% edges. Similar to the analysis of network density, those induced subnetworks were then run with SONAR and the performance of SONAR was evaluated by ROC-AUC analysis.

Rationale for HydRA

HydRA is a classification model (also called classifier) aimed at recognizing the RNA-binding capacity of a given protein. It's the revised version of SONAR[16] by introducing new protein features and new machine learning techniques into RBP prediction. HydRA consists of two main components: 1) the first component, referred to as extrinsic classifier, utilizes information describing the extrinsic context of proteins to predict RNA-binding potential from experimental protein interaction data and predicted protein association data (i.e. STRING[63]); 2) the second component, referred to as HydRA-seq, measures proteins' RNA-binding potential from their intrinsic properties with their amino acid sequence -based information. Specifically, the second component comprises three sub-classifiers, including a SVM model (seqSVM), a convolutional neural network (seqCNN) model and an attention-based model following Transformer architecture (named ProteinBERT-RBP). The outputs of all the classifiers from the two main components are then combined with a simple probabilistic model to give the final prediction of the input proteins.

Dataset for modeling

Under the criterion where the class ratio (i.e. known RBPs versus other proteins) in each set keeps same, we first randomly took out 20% of the data in our protein dataset as the hold-out test set, which is only used for the final evaluation of the final RBP classifier and the comparison among different RBP classifiers/software. The rest 80% of data constitutes the training/validation set that is used to train and select model, select features and tune the hyper-parameters of the RBP classifiers. In the evaluation and comparisons of classifiers that utilize proteins sequences, we discarded the proteins in the test dataset whose sequences are highly similar to any

protein's sequence (>90% similarity and >90% coverage) in the training/validation dataset. The similarity between protein sequences was calculated by CD-HIT-2D program in CD-HIT package[67] with default parameters.

Protein interaction and association-based classifier

A classifier exploiting the data from physical protein-protein interaction network (i.e. Mentha-BioPlex) and functional protein association network (i.e. STRING), which described the extrinsic context of the protein, was constructed using support vector machine (SVM) model with radial basis function as its kernel. We referred to this classifier as Protein Interaction and Association based (PIA) classifier (also referred to as SONAR3.0). The feature extraction and classifier construction process are modified from that of SONAR.

Firstly, we extracted the local network around the given protein and split the local network into different neighborhoods (level-1, level-2, etc.) according to the length of the path from the protein to the given protein (more details can be found in SONAR.[16] Specifically, we found all the paths with length k that start from the given protein and collect all the end nodes of these paths into level-k neighborhoods. Notably, a protein can be in both level-1 and level-2 neighborhoods if there are paths in length 1 and 2 between this protein and the given protein. In addition, a protein node can appear more than once in the level-k neighborhoods if there are more than one path of length k between this protein and the given protein.

Secondly, the proportion of nodes that are annotated RBPs in each neighborhood represents the feature of that neighborhood. In details, for a given protein, its feature F_k of the kth level neighborhood is calculated as

$$F_k = \frac{\sum_{p_{k,i} \in P_k} \left[N(p_{k,i}) \in known\ RBPs \right]}{|P_k|}.$$

P_k is the set of paths with length = k in the network that start from the given protein, $p_{k,i}$ denotes a specific path in P_k while i is the index of the Path. $N(p)$ denotes the end node of path p. Here […] are the Iverson brackets. [P] is defined to be 1 if P is true, and 0 if it is false.

The first three levels of neighborhoods in physical PPI network and the first level neighborhood in functional protein association network are considered in this classifier. Different from SONAR, we also added an indicator feature for each local network (i.e. physical PPI local network and functional protein association local network) to penalize those proteins with few level 1 neighbors, whose PPI information is not adequate for confident predictions. We set the indicator feature as 1 if the query protein has no less than θ level-1 neighbors, otherwise the feature was set as 0. By optimal threshold searching, we set $\theta = 5$.

Lastly, we took the six network-based features of each given protein as the input of SVM model. Cross-validation (10-fold) method was used to search for the optimal hyper-parameters and threshold (such as the number of PPI neighborhoods included and the indicator threshold θ) of this model within our training/validation set. Additionally, to handle the class imbalance issue in our training process, we chose oversampling out of a couple of widely used approaches, where we oversampled the number of our RBP training samples by equally replicating current RBP samples in the training set.

Sequence-based SVM classifier

In line with many previous studies on sequence-based RBP prediction, a support vector machine (SVM) model with radial basis function (RBF) kernel was used to construct one of the sequence-based RBP sub-classifiers from protein sequence. We referred to this classifier as seqSVM.

Firstly, protein sequences were preprocessed to generate two categories of protein features:

(1) *k-mer*: all the possible substrings of length k (k=*3, 4*) that are contained in a protein's amino acid sequence were counted. The counts of all k-mers were subsequently encoded into an integer vector. To alleviate the computational cost, for each protein, we only look at the 1000 most frequent *k-mers* that appeared in the known RNA-binding proteins.
(2) Amino acid composition (AAC): We calculated AAC for each protein with *propy*[68] package. A vector of 20 descriptors per protein was gained.[69]

Secondly, feature selection was done in order to reduce the dimension of feature vectors. Chi-square test was used to filter out unimportant features within each feature category respectively via their p values controlled for false discovery rate (FDR) at level 0.01. For *k-mer* and *SS-kmer*, features were further selected by linear SVM model by keeping the features the absolute value of whose weights are larger than the average of those of all input features.

Thirdly, the seqSVM classifier was built with the features selected from last step utilizing SVM model and RBF kernel. We only use proteins shorter than 1500 amino acids in our training step to alleviate the involvement of unrelated information carried by proteins, particularly those long ones.

We used the training/validation set with cross-validation approach to search for optimal values for k (in *k-mer* and *SS-kmer*) and parameters or hyper-parameters used in feature selection pipeline and SVM model construction. The same oversampling approach as mentioned in SONAR3.0 construction was employed to overcome the class imbalance issue in the training process.

Convolutional neural network for sequence-based RBP classification

Parallel to seqSVM, we also created a feed forward convolutional neural network (CNN) to recognize RBPs by automatically learning the pattern underlying the protein sequence with convolutional layers. We referred to this CNN model as seqCNN. The seqCNN model consisted of 8 different hidden layers of neuron-like computational units that are stacked up in a certain order (shown in Figure 2A). The first layer was an embedding layer that mapped amino acids to dense vectors encoded with biophysical and biochemical properties of the amino acids using a pre-trained weight matrix called ProtVec.[29] This layer not only enriches protein sequence information with the knowledge from available database, but also alleviates the difficulties on deep learning caused by sparse one-hot encoded matrix of the protein sequence. Following this layer were two 1D convolutional layers (kernel size = 5) with a pooling layer for each to extract patterns from sub-sequences and condense the information for downstream layers. The output was then followed by a global pooling layer for further condensation. Next, the condensed information was conveyed to last 2 hidden layers, which are regular fully-connected neural network layers (also called dense layers), for more processing. An output layer in the form of fully connected neurons after the last hidden layers is used to encode all learned protein information into the form of probability (i.e. the classification score), showing how likely the input protein is a RNA-binding protein. The binary-cross-entropy loss was used to measure the agreement between the output predictions and the true class labels the on the training set. Besides, regularization methods such as dropout and batch normalization are applied to avoid overfitting. We choose this model structure from a broad range of CNN topologies (i.e. the number of CNN layers, the number of kernels in each layer and the) and alternative CNN architectures, such as residual convolutional neural network.

This model was trained with backpropagation algorithm that adjusts and optimizes the weight matrix in each layer during the training process to minimize the binary-cross-entropy loss on the training set. Specifically, Adam optimization was used in the process. To prevent the training process from getting stuck in undesired local optima, we pre-trained the hidden layers with auto-encoder technique[69] where the convolutional layers and fully-connected layers were pre-trained separately. Similar to seq SVM, only proteins with length shorter than 1500 amino acids are used to train the model and all the optimal hyper-parameters of this model were obtained using cross-validation within our training/validation set. The seqCNN classifier was implemented with TensorFlow kernel[70] utilizing its high-level API, Keras.[55] The class imbalance issue in the training process was handled by adjusting the class weight argument in Keras model.

Attention-based protein language model for RBP classification

ProteinBERT is an attention-based Transformer/BERT architecture specifically designed for protein classification problems.[19] It originally takes protein sequence and protein GO annotations as input. The key component in this architecture is a the transformer block with two parallel neural networks: (1) one takes the representation matrix from the protein sequence (denoted by local representations) and uses convolutional layers with skip connections and layer normalizations to extract and compress location-wise signals followed by (location-wise) fully connected layer; (2) the other consists of two fully connected layers with skip connections and layer normalizations and takes a global representation vector encoded from the GO annotations of the protein as input. In each transformer block, information between local and global representations flows to each other via two attention layers respectively allowing the local and global signals to guide the learning process of each other easily.[19] We employed the same ProteinBERT architecture as proposed in their original paper which includes 6 transformer blocks with 4 global attention heads in each block.

This ProteinBERT model was pretrained following the original paper via a self-supervised training strategy with ~106M proteins and their corresponding GO annotations from UniRef90.[71] We next fine-tune the model with our training set with RNA-binding annotations (mentioned above) allowing the model to exclusively focus on the supervised learning of RNA-binding related sequence patterns and to classify RBP and non-RBPs. In the fine-tuning, all layers of the pretrained model were first frozen except a newly added fully connected layer in the end of the model, which was used to connect the pretrained model to RBP classification output layer. Similar to seqCNN, binary-cross-entropy loss was used to measure the agreement between the output predictions and the true class labels and Adam was used in the optimization process. The model was trained for up to 40 epochs using training proteins with sequence lengths of 512 tokens (i.e. the proteins with no longer than 512 amino acids). Next, we unfroze all the layers and trained the model for up to 40 additional epochs with the same training proteins. Lastly, following the original paper, we did one final epoch of training with the model using longer training proteins of 1,024 tokens, which was introduced to encourage the model to generalize to different sequence lengths.[19] We evaluate the model using the same strategy of the classifiers above with training and test set while 10% of the training set was used as "validation set". The model was fine-tuned using Adam optimizer for the backpropagation, and we reduced the learning rate on plateau and used early-stopping base on the "validation set" to avoid overtraining. This ProteinBERT model for RBP classification (referred to as ProteinBERT-RBP) was implemented using Keras[55] with Tensorflow kernel.[70]

Graph neural network modelling on protein structures

The protein structure backbone is represented as graphs G = (V, E), where each node $v \in V$ is an amino acid connected by edge $\varepsilon \in E$ to its amino acid neighbors in the 3D space whose corresponding Ca atoms are less than 10Å distant from the C_α atom of v or within the k-neareast neighbor set of the C_α atom of v. To obtain the optimal features to represent each amino acid and their spatial relationship to their neighbors, we select node features from various types of amino acid features including (1) one-hot representation, (2) main chain torsion angles (Phi and Psi), (3) the structural features encoded by FoldSeek,[72] and the latent representations learned by large pre-trained protein language models, such as (4) DeepFRI,[32] (5) ProteinBERT,[19] and (6) HydRA's ProteinBERT-RBP.

Similarly, we select edge features from an edge attribute pool including (1) using the same attributes for each edge (i.e. no edge features) (2) distance between C_α atom of the two amino acids, (3) the structural features encoded by FoldSeek between the two amino acids, (4) the average confidence score (pLDDT) from AlphaFold2 prediction of the two amino acids. The feature selection was done using the forementioned training set and validation set where the models with different combination of the node and edge features were trained on the protein structures with RBP labels in the training set and evaluated on the validation set. The optimal graph neural network architecture was also selected in this step from the widely used architectures, e.g. graph convolutional network (GCN), graph attention network (GAT) and PointTransformer.[73] The optimal model was selected and denoted as strucGNN1 which employs PointTransformer as the main architecture. The local presentation output for amino acid (node v_i) by the last transformer block in ProteinBERT-RBP model was selected as the node features in strucGNN1, defined as $x1_i = \text{ProteinBERT\_RBP}(v_i)$, while no edge feature was used as PointTransformer architecture does not support edge feature input.

To update the node features of each amino acid according to the protein structure graph, strucGNN1 use 3 PointTransformer blocks. Each PointTransformer block firstly down-samples the nodes in the graph using farthest point sampling and then aggregates information from the neighboring nodes for each node using self-attention mechanism.[73] In details, the node-wise aggregation operation in kth PointTransformer block is defined as:

$$x_i^k = \sum_{x_j \in N(i)} \rho\left(\gamma\left(\varphi\left(x_i^{k-1}\right) - \psi\left(x_j^{k-1}\right) + \delta\right)\right) \odot \left(\alpha\left(x_j^{k-1}\right) + \delta\right),$$

where $N(i)$denotes the neighbors of v_i, and v_i is denoted as i here for simplicity.and $x_i^0 = x1_i$. As defined in the original paper for PointTransformer,[73] $\varphi, \psi,$ and α are pointwise feature transformations using fully connected layers, δ is a position encoding function used to encode the relative position between two nodes. δ is defined as $\delta = \theta(p_i - p_j)$, where p_i and p_j are the 3D coordinates of the Cα atom for each node and the encoding function θ is two linear layers with one ReLU nonlinearity. ρ is a normalization function *softmax*, while the mapping function γ is also two linear layers and one ReLU nonlinearity. In this way, each block applies self-attention locally, within a local neighborhood around each amino acid.

The global features of the protein are then calculated by taking the channel-wise average across the node dimension output by the PointTransformer blocks using a global mean pool layer, followed by two fully connected layers (with 64 neurons each) with ReLU non-linear activation function to further process the global features. The final output is computed as a linear mapping of the fully connected layer's output followed by Sigmoid transformation. Binary-cross-entropy loss was used to measure the agreement between the output predictions and the true class labels and Adam was used in the optimization process. The strucGNN1 were implemented using PyTorch[74] and PyTorch-Geometric.[58]

Finetuning pretrained graph neural network on RNA-binding protein structures

To exploit the recent advancement in the structure-based protein pretraining models, we adjusted and finetuned Masked Inverse Folding (MIF) model for RBP classification task and referred this new model as strucGNN-MIF. MIF is a pretrained graph neural network model which achieved state-of-the-art performance in many protein prediction tasks,[75] MIF was pre-trained with the CATH4.2 database,[76] and is built on structured graph neural network architecture with the pretraining task of reconstructing a corrupted protein sequence conditioned on its backbone structure.

Similar to the graph construction in strucGNN1 and strucGNN2, the protein structure backbone is represented as graphs G = (V, E) with each amino acid as a node $v \in$ V. But instead of taking the other amino acids with the distance less than 10 Å to the given amino acid as neighbors, each given amino acid node is connected by edge $\varepsilon \in$ E to its k-nearest amino-acid neighbors in the structure.

To adjust MIF to binary RBP classification task, we took the encoder layers of MIF and connected it to a global mean pool layer to average node features across the node dimension. The pooled features were then fed to three fully connected layers with ReLU nonlinear activation before connecting to the output layer with sigmoid activation function.

We finetuned this model with our RBP training set and evaluate the performance on the validation and test set. Similar to strucGNN1 and strucGNN2, binary-cross-entropy loss was used to measure the agreement between the output predictions and the true class labels. Adam was used to adjusts and optimizes the weight matrix in each layer and to minimize the binary-cross-entropy loss on the training set.

Meta ensembling of classifiers

For simplicity, we denoted the aforementioned three classifiers (i.e. SONAR3.0, seqSVM and seqDNN) as primary classifiers. A simple probability-based ensemble approach was then derived to integrate predictions from these primary classifiers. This approach consists of three steps:

(1) We collected the classification scores generated by each primary classifier during the cross-validation step and took these scores and corresponding class labels as reference lists *S* and *L*:

$$S_i = (s_{i,1}, s_{i,2}, \ldots, s_{i,n}) : s_{i,j} \in [0,1]$$

$$L_i = (l_{i,1}, l_{i,2}, \ldots, l_{i,n}) : l_{i,j} \in \{0,1\}$$

where *i* is the index of the primary classifier, *j* is the index of the proteins and *n* represents the total number of protein scores collected from the cross-validation step.

(2) For a query protein, given a classification score *x* from a primary classifier, we calculated the probability of this protein being a false discovery (i.e. false discovery rate) for this primary classifier, based on *x* and the reference lists *S* and *L*:

$$FP_i(x_i) = \sum_{j=1}^{n} [s_{i,j} > x_i \text{ and } l_{i,j} = 0]$$

$$P_i(x_i) = \sum_{j=1}^{n} [s_{i,j} > x_i]$$

$$FDR_i(x_i) = \frac{FP_i(x_i)}{P_i(x_i)}$$

where *FDR* represents false discovery rate, *FP* represents the number of false positives and *P* represents the number of all the positive prediction. Here […] are the Iverson brackets. [*P*] is defined to be 1 if *P* is true, and 0 if it is false.

(3) For this query protein, we then got its probability of being a false discovery for all these primary classifiers as follows

$$\prod_{i \in \text{primary classifiers}} FDR_i(x_i)$$

(4) Thus, the probability of this query protein not to be a false discovery for at least one of the primary classifiers was got as follows. This is taken as the output score of this ensemble approaches.

$$Score = 1 - \prod_{i \in \text{primary classifiers}} FDR_i(x_i)$$

Besides constructing the final RBP classifier (i.e. HydRA), this ensemble approach was also used to integrate the results from intrinsic classifiers (i.e. seqSVM and seqDNN)

This ensemble approach is chosen from a couple of other classifier ensemble methods, including other variant of this probability-based approach (i.e. the false positive rate based version) and machine learning based approaches such as SVM and logistic regression.

Software comparisons

We applied other publicly available RBP classifiers, i.e. RNApred,[11] SPOT-seq,[12] catRAPID signature[13] and RBPPred[14] to predict all the proteins in our test dataset. The predicted scores and class labels were then collected to calculate model evaluation metrics such as accuracy, sensitivity, specificity, precision, ROC-AUC, F1 score, Matthews Correlation Coefficient and balanced accuracy.

RNA-binding domain sequences and RNA-binding peptide

Amino acid sequences corresponding to different protein domains (i.e. KH, RRM, DEAD, CSD, LSM, BTB, SH2, Ig, UBA and 7 transmembrane domains)) are downloaded from Pfam database. RNA-binding peptides are collected from five recent studies of four

technologies that experimentally identify the regions that interact with RNA molecules in RNA-binding proteins: RBDmap,[5,47] pCLAP,[48] RBR-ID[77] and RNP^{xl}.[52]

Protein domain annotation

The domain annotations in each protein are obtained by homology search using hmmscan (HMMER 3.1b1) against Pfam-A.hmm from Pfam database (v27.0). Overlapped and redundant domain hits are merged using hmmscan-parser.sh (from https://github.com/carden24/Bioinformatics_scripts/blob/master/hmmscan-parser.sh). Only domain annotations with E-value < 0.1 were kept for downstream analysis.

Occlusion map

Occlusion map (also referred to as occlusion analysis) aims at interpreting the foci of the machine learning classifier on their target object by detecting the changes in predictions when a part of the object is deliberately ''occluded''. This analysis has been widely used in computer vision.[20–22] Here, we use it to interpret which parts of the protein our intrinsic-feature classifier is ''paying attention'' to. Specifically, shown in Figure S7A, we occluded a subsequence of fixed length k (k=20 in this study) within the protein sequences by setting a window of consecutive sites to be zero or null (referred to as ''occlusion window"). The occluded protein then went through all our feature generation stages and was fed to seqCNN, seqSVM and ProteinBERT-RBP to generate new classification scores (noted as $Score_{occ,\ seqCNN}$, $Score_{occ,\ seqSVM}$ and $Score_{occ,\ ProteinBERT}$). We defined the difference between the new classification scores and original scores without occlusion (denoted by $Score_{orig,\ seqSVM}$, $Score_{orig,\ seqDNN}$ and $Score_{orig,\ ProteinBERT}$) as $\Delta score_{model\,j} = Score_{occluded,model\,j} - Score_{original,model}$, where j indicates the index of the occluded part and *model* is either seqCNN, seqSVM or ProteinBERT-RBP. In this case, negative Δ *score* represents a negative effect on categorizing this protein to RBP, indicating the occluded part are considered as important for RNA-binding by our classifier. To integrate occlusion information from seqCNN, seqSVM and ProeinBERT-RBP, $\Delta score_{seqDNN}$, $\Delta score_{seqSVM}$ and $\Delta score_{ProteinBERT}$ were first standardized (i.e. z-score transformed) respectively as $\Delta zscore_{model\,j} = \frac{\Delta score_{model\,j} - \mu_{model}}{\sigma_{model}}$, where μ and σ represents the mean and standard deviation of all $\Delta scores$ we have obtained from all the training proteins for specific model (i.e. seqCNN or seqSVM). The normality of $\Delta score_{seqDNN}$, $\Delta score_{seqSVM}$ and $\Delta score_{ProteinBERT}$ populations were confirmed with D'Agostino and Pearson's test. The integrated $\Delta score$ was then generated by averaging the standardized $\Delta score$ from seqCNN and seqSVM: $\Delta score_j = \frac{\Delta zscore_{seqDNN\,j} + \Delta zscore_{seqSVM\,j} + \Delta zscore_{ProteinBERT\,j}}{3}$. To get a map of Δ *score* for the whole protein, we slid the occlusion window within each protein from N-terminal to C-terminal and repeated the calculations. P-value was obtained for each occlusion window based on the Δ *score* population which follows a normal distribution. We defined a predicted RNA-binding region (pRBR) as the occlusion window with significantly lower $\Delta score$with p-value < 0.05 and a strong predicted RNA-binding region (strong pRBP) with p-value < 0.001. Notably, $\Delta score_{model}$ tends to get smaller as the protein length increases (see Figures S7B and S7C) so that occlusion windows in short proteins are more likely to be called as pRBR or strong pRBR than those in long proteins. To alleviate this bias, we grouped the proteins based on their protein lengths and do the standardization within each group in the standardization step and this results in a more balanced $\Delta score_{model}$ distribution (Figures S7D and S7E).

Criteria of RBP candidacy

After all the aforementioned steps of model selection and optimization, we got the final RBP classifier (i.e. HydRA) and trained it with all the samples in our protein dataset. With this trained classifier, we got a classification score to each protein in the protein dataset. We set a cutoff as 0.8927 for the classification scores by fixing the false positive rate to 10% within the test dataset. Thus, all those proteins with classification scores higher than the cutoff but with negative class labels (i.e. previously not known for RNA-binding) were collected in the list of candidate RBPs.

Biophysical properties of proteins

Within each RBP candidate, isoelectric points were calculated with ExPASy web server.[78] A score describing intrinsic disorder was assigned to each amino acid position using IUPred.[79] Intrinsic disordered region (IDR) or sites were defined as positions with score < 0.4. Complexity of each amino acid within the protein was calculated as Shannon entropy of the neighbor amino acids within a window of size 21 centering the given amino acid's position. Positions with entropy < 3 were regarded as low complexity region or sites. One-side Mann-Whitney U test was used to statistically examine the difference between different protein populations in the distributions of isoelectric points, amino acid composition and the proportion of disordered and low-complexity regions.

Antibodies for western blot and immunoprecipitation

The primary antibodies used are as follows: anti-V5 (Proteintech, 66007-1-Ig), anti-V5 (Bethyl A190-120A), anti-His (Proteintech, 66005-1-Ig), anti-His (Invitrogen, MA1-21315), anti-a-Tubulin (Sigma, Clone B-5-1-2), anti-HSP90B (Sigma, SAB4501463) and anti-HNRNPA2B1 (Proteintech, 14813-1-AP)

Plasmid construction

All RBP mammalian expression constructs were in one of two lentiviral Gateway (Invitrogen) destination vector backbones: (1) pLIX403_V5_mRuby or (2) pLIX403_V5. The pLIX403 inducible lentiviral expression vector was adapted from pLIX_403 (deposited by D. Root; Addgene plasmid no. 41395) to contain TRE-gateway-mRuby and PGK-puro-2A-rtTA upstream of mRuby by Gibson assembly reaction of PCR products (Cloneamp, Takara Bio). RBP open-reading frames (ORFs) were obtained from human Orfeome 8.1 (2016 release) donor plasmids (pDONR223) when available, or amplified (Cloneamp, Takara Bio) from cDNA obtained by SuperScript III (Invitrogen) RT–PCR of HEK293XT cell purified RNA (Direct-zol, Zymogen) and inserted into pDONR223 by Gateway BP Clonase II reactions (Invitrogen). Donor ORFs were inserted in frame upstream of V5 and mRuby fusion cassettes by gateway LR Clonase II reactions (Invitrogen).

Putative RNA binding domains (as determined by occlusion analysis) were removed via site-directed mutagenisis (NEB E0554S) by amplifying entire entry clone plasmids (candidate RBP ORFs within pDONR221 or pDONR223 plasmids) using standard, non-mutagenic forward and reverse primers flanking the deletion region, followed by PCR product ligation. The resulting deletion clones were used for Gateway insertion into V5 or V5-mRuby for stable expression and eCLIP as was done for deletion or full-length expression clones.

Generation of cell lines

Stable lines were made in human lenti-X HEK293T cells (HEK293XT, Takara Bio), maintained in DMEM (4.5 g l−1 D-glucose) supplemented with 10% FBS (Gibco) at 37 °C with 5% CO2. Cells were passaged at 70%–90% confluency by rinsing cells gently with DPBS without calcium and magnesium (Corning) then dissociating with TrypLE Express Enzyme (Gibco) at a ratio of 1:10. The stable HEK293XT cell lines expressing RBP full-length and domain-deleted RBP candidates and were generated as described below by transducing ~1 million cells with 8 μg m l−1 polybrene and 1 ml viral supernatant in DMEM + 10% FBS at 37 °C for 24 h, followed by subsequent puromycin resistance selection (2 μg ml−1).

Lentivirus was packaged using HEK293XT cells seeded approximately 24 h before transfection at 30%–40% in antibiotic-free DMEM and incubation at 37 °C and 5% CO2 to 70%–90% confluency along with the Lipofectamine 3000 reagent (ThermoFisher Scientific) following the manufactures protocol (https://tools.thermofisher.com/content/sfs/manuals/lipofectamine3000_protocol.pdf). DNA ratios used for transfection is as follows: 10:1:10 proportion of lentiviral vector:pMD.2g:psPAX2 packaging plasmids. Six hours following transfection, medium was replaced with fresh DMEM + 10% FBS. At 48 h after medium replacement, virus-containing medium was filtered through a 0.45-μm low-protein binding membrane. Filtered viral supernatant was then used directly for line generation by transducing ~1 million cells (one well of a six-well dish) with 8 μg ml−1 polybrene and 1 ml viral supernatant in DMEM + 10% FBS at 37 °C for 24 h. After 24 h of viral transduction, cells were split into 2 g l−1 puromycin and selected for 72 h before passaging for storage and downstream validation and experimentation.

RNA immunoprecipitation and qPCR

HEK293 cells were transfected with plasmids to express V5-tagged HSP90AA1 (pcDNA6-HSP90AA1-V5) or empty vector. 24 hours later, cells were lysed in lysis buffer (50 mM Tris pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS and 0.5% sodium deoxycholate) supplemented with 1 Protease Inhibitor cocktail (Roche) and 80 U of RNAse Inhibitor (Roche). Clarified lysates were pre-cleared with Protein G agarose beads (Life Technologies). Aliquots of the supernatant (equivalent to 5% of supernatant) were saved as input protein and RNA. The remainder of the supernatant was incubated with 2 mg of V5 antibody at 4 C for 4 h. The protein–RNA–antibody complex was precipitated by incubation with Protein G magnetic beads overnight at 4 C. Beads were washed twice with lysis buffer and three times with wash buffer (5 mM Tris pH 7.5, 150 mM NaCl, 0.1% Triton X-100). Ten per cent of the bead slurry was reserved for western blot analysis. The remaining bead slurry was resuspended in TRIzol (Life Technologies), and RNA was extracted as per the manufacturer's instructions. Input and immunoprecipitated RNA was converted into cDNA and gene expression was measured with qPCR in technical triplicates. RNA immunoprecipitation qPCR studies were performed in biological duplicates.

eCLIP Library Preparation for YWHAG/H/E/Z and HSP90AA1

The open reading frame of human HSP90AA1 was cloned into pEF5/FRT/V5-DEST (Invitrogen) by Gateway cloning, positioning the gene in frame with a 3′ V5 tag. The open reading frames of human YWHAG, YWHAH, YWHAE, YWHAZ, and HSP90AA1 were cloned into pcDNA6/myc-His B (Invitrogen) by restriction digest, positioning the gene in frame with a 3′ myc-his tag. HEK293T cells transfected with the protein of interest were cultured to confluency in 10 cm dishes. The cells were UV crosslinked (254 nm, 400 mJ/cm2 constant energy) then pelleted and frozen on dry ice. eCLIP procedure was performed as described.[53] Briefly, cell pellets were lysed in eCLIP lysis buffer and sonicated (BioRuptor). Lysates were treated with RNase I (Ambion) to fragment the RNA and incubated with antibodies against V5-tag or His-tag (both from Proteintech) to immunoprecipitate RBP-RNA complexes. 2% of the lysate:antibody mixture was saved as input sample. The remaining immunoprecipitated (IP) samples were stringently washed, followed by dephosphorylation of RNA by FastAP (ThermoFisher Scientific) and T4 Polynucleotide Kinase (NEB), and ligation of a 3′ RNA adapter with T4 RNA Ligase (NEB). Immunoprecipitates and input samples were resolved by SDS-PAGE, transferred to nitrocellulose membrane, and then the region of membrane corresponding to the molecular weight of the protein of interest up to 75 kDa above it was excised for each immunoprecipitate and input sample. RNA was isolated from the membrane, reverse transcribed with AffinityScript (Agilent), free primers were removed (ExoSap-IT, Affymetrix), and a 5′ DNA adapter was ligated onto the cDNA product with T4 RNA ligase

(NEB). Libraries were then amplified with Q5 PCR mix (NEB), size selected using AMPure XP beads (Beckman Coulter, Inc.) and on a 3% agarose gel, and then quantified with a Bioanalyzer instrument (Agilent). Paired-end (50 base pair) sequencing of the eCLIP libraries were performed on Illumina HiSeq 3500.

eCLIP Library Preparation for INO80B, PIAS4, NR5A1, ACTN3, and MCCC1

All eCLIPs were conducted following induction or transient transfections and IP was conducted using anti-V5 tag (Bethyl A190-120A). eCLIP experiments were performed as previously described in a detailed standard operating procedure[53] which is provided as associated documentation with each eCLIP experiment on the ENCODE portal (https://www.encodeproject.org/documents/fa2a3246-6039-46ba-b960-17fe06e7876a/download/attachment/CLIP_SOP_v1.0.pdf/). In brief, 20 million cross-linked cells were lysed and sonicated, followed by treatment with RNase I (Thermo Fisher) to fragment RNA. Antibodies were precoupled to species-specific (anti-rabbit IgG) dynabeads (Thermo Fisher), added to lysate and incubated overnight at 4 °C. Before IP washes, 2% of sample was removed to serve as the paired-input sample. For IP samples, high-salt and low-salt washes were performed, after which RNA was dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3′ RNA adapter was ligated with T4 RNA ligase (NEB). Ten percent of IP and input samples were run on an analytical PAGE Bis-Tris protein gel, transferred to PVDF membrane, blocked in 5% dry milk in TBST, incubated with the same primary antibody used for IP (typically at 1:4,000 dilution), washed, incubated with secondary horseradish peroxidase-conjugated species-specific TrueBlot antibody (Rockland) and visualized with standard enhanced chemiluminescence imaging to validate successful IP. Ninety percent of IP and input samples were run on an analytical PAGE Bis-Tris protein gel and transferred to nitrocellulose membranes, after which the region from the protein size to 75 kDa above protein size was excised from the membrane, treated with proteinase K (NEB) to release RNA and concentrated by column purification (Zymo). Input samples were then dephosphorylated with FastAP (Thermo Fisher) and T4 PNK (NEB) at low pH, and a 3′ RNA adapter was ligated with T4 RNA ligase (NEB) to synchronize with IP samples. Reverse transcription was then performed with AffinityScript (Agilent), followed by ExoSAP-IT (Affymetrix) treatment to remove unincorporated primer. RNA was then degraded by alkaline hydrolysis, and a 3′ DNA adapter was ligated with T4 RNA ligase (NEB). Quantitative PCR was then used to determine the required amplification, followed by PCR with Q5 (NEB) and gel electrophoresis for size selection of the final library. Libraries were sequenced on the HiSeq 2000, 2500 or 4000 platform (Illumina). Each ENCODE eCLIP experiment consisted of IP from two independent biosamples, along with one paired size-matched input (sampled from one of the two IP lysates before IP washes).

eCLIP Library Preparation for V5 tag control

A plasmid encoding the V5 peptide was transfected in HEK293XT cells using Lipofectamine 3000 in biological duplicates (10 cm plates). After 72hr, cells were UV-crosslinked and subjected to the eCLIP protocol using anti-V5 antibody (Bethyl A190-120A, 10 ug per sample), as previously described. At the SDS-PAGE step, each sample was divided equally between three lanes. Membrane regions corresponding to each possible 75 kDa window, each separated by 25 kDa, were excised and prepared separately for sequencing. Samples were compared against a size-matched input without V5 IP for each size range. We referred to this set of data to as "V5 tag only" samples.

QUANTIFICATION AND STATISTICAL ANALYSIS

Evaluating domain-level RNA-binding element prediction

To evaluate HydRA's capability of recognizing RNA-binding elements on domain level, we ran HydRA-powered Occlusion Map with all the known RBPs and measure the performance of recognizing high-confidence RNA-binding domains (RBDs) in the RBPs over other domains using positive likelihood ratio (discussed below). To obtain the annotation of high-confidence RNA binding domains, we first retrieved the homology-annotated domains (see protein domain annotation above) for each RBP and retained the domain hits that exist in the list of RNA-binding and -processing domains from Gerstberger et al.[1] Next, those homology-annotated RBDs that were also supported by the experimentally identified RNA-binding peptides (referred to as RBDpeps) studies[5,47–49] were annotated as high-confidence RBDs. Similarly, high-confidence non-RNA-binding domains were those homology-annotated domains in the RBPs that are not in the RNA-binding and -processing domain list and not supported by RBDpeps studies. Within the occlusion map output, we observed the percentage of the high-confidence RBDs that have at least one pRBR (p-value < 0.05) or stringent pRBRs (p-value < 0.001) (akin to sensitivity), and also the percentage of the high-confidence non-RBDs have at least one pRBR or stringent pRBRs (akin to false positive rate (FPR)). Positive likelihood ratio (also called fold-enrichment) was then calculated as the ratio of sensitivity versus FPR, to demonstrate how well the pRBRs are enriched in high-confidence RBDs over high-confidence non-RBDs in the proteins. Similarly, we ran the baseline tools RNABindPlus and DRNApred, which are designed to predict RNA-binding residues in RBPs, on the same set of RBPs. We obtained positive likelihood ratio by observing the percentage of the high-confidence RBDs that have at least one predicted RNA-binding residue (i.e. sensitivity) and the percentage of the high-confidence non-RBDs that have at least one predicted RNA-binding residue (i.e. FPR). By using different p-value/ score threshold when defining pRBR and predicted RNA-binding residues, we got the positive likelihood ratio changes at different sensitivity level.

eCLIP data processing and analysis

eCLIP sequence reads were processed and analyzed as previously described.[53] For each protein of YWHAG/H/E/Z and HSP90AA1, a total of six eCLIP libraries, 2 replicates for each of the IP, SMInput and control group, were sequenced, while YWHA family shares the same control libraries. For each protein of INO80B, PIAS4, NR5A1, ACTN3, and MCCC1, 2 or 3 replicates for each of the IP, SMInput were sequenced, Reads were first mapped to repetitive elements and only the unmapped reads were retained and were next mapped to human genome (hg19). The uniquely mapped reads were kept for downstream analysis. PCR duplicates were further removed to obtain 'usable reads' with use of a randomer (N10) sequence positioned one of the adapter oligos. Next, usable reads were counted for both eCLIP (immunoprecipitation) and size-matched input (SMInput) samples across different genomic regions (i.e. 5′UTR, 3′UTR, coding exons (CDS), and intronic regions) for all coding genes annotated in UCSC Known Gene table (hg19). To identify the enrichment of binding signals in CLIP samples above SMInput in certain genomic region of certain gene, fold-enrichment was calculated as the ratio of read counts within this region in CLIP versus SMInput. Only regions with at least 10 reads in both of eCLIP and SMInput samples were considered. Sequencing read peaks were identified using CLIPper algorithm[80] for both eCLIP and SMInput libraries. CLIPper-defined peaks were then normalized to size-matched input (SMInput) by comparing the read density in immunoprecipitation (IP) and SMInput samples (referred to as INPUT normalization). Significant peaks were defined as the peaks whose number of reads from the IP sample were 8-fold greater than the number of reads from the SMInput sample with a p-value < 0.001. P-value was calculated by Yates' Chi-Square test (Fisher Exact Test was performed when the observed or expected read number was below five). For HSP90AA1 and YWHA G/H/E/Z, reproducible peak across both biological replicates were identified using an irreproducible discovery rate (IDR) approach used in our previous study.[81] To validate the new RBPs (i.e. INO80B, PIAS4, NR5A1, ACTN3, and MCCC1) and their predicted functional domains, we applied a more stringent strategy to identify the reliable reproducible peaks using eCLIP output from size-matched "V5 tag only" experiments (shown in Figure S9C). In this strategy, the sequencing reads are processed in the same way above except we did two sets of normalization in the INPUT normalization step: (1) the standard normalization, i.e. protein IP against protein SMinput; (2) the normalization against V5 tag only output, i.e. protein IP against the IP sample from the size-matched V5 tag experiment. ChIP-R[54] algorithm was then used to identify the reproducible peaks for each of these outputs because of its capability of handling more than 2 replicates. The shared reproducible peaks from these two normalization ways form the final list of reproducible peaks where at least 5 mapped reads are found in the peak for each IP sample of the protein are also required.

Peak enrichment analysis

Significantly enriched eCLIP peaks (described above) were assigned to different transcriptomic regions (i.e. 5′UTR, 3′UTR, coding exons (CDS), and intronic regions) with BEDTools[82] using annotations from GENCODE. For peaks whose coordinates overlap with multiple transcriptomic regions, the following priority were applied to assign these peaks each with a single region: CDS, 5′UTR, 3′UTR, then proximal (as defined as less than 500 bp from an exon–intron boundary) or distal introns (as defined as 500 bp or greater from an exon–intron boundary). The fraction of peaks in a specific transcriptomic region (P_{CLIP}) was obtained by dividing the number of peaks in this region by the total number of peaks in all regions (5′UTR, 3′UTR, CDS, proximal intron and distal intron). Fold-enrichment (E) was then calculated as $E_{region} = log_2(P_{region} / S_{region})$ to measure the binding preference of the tested RBP over this transcriptomic region, where S_{region} is the fractional region size derived by dividing the total number of base pairs in that region relative to the total number of base pairs in all regions. E_{region} with positive values indicate overrepresentation of peaks in this region and negative values indicate the underrepresentation. To be noticed, if a base pair is associated with multiple transcriptomic regions, this base pair was assigned to regions in the following priority: CDS, 5′UTR, 3′UTR, then proximal or distal introns.

GO enrichment for candidate RBPs

We retrieved the enriched Gene Ontology (GO) terms for the candidate RBPs (YWHA family, INO80B, PIAS4, NR5A1, ACTN3, and MCCC1) with overrepresentation test implemented by PANTHER classification system.[83] To simplify the output, we only observed the enrichment in GO terms from level 5 of the GO hierarchy system.

Gene ontology enrichment analysis for HSP90A was performed using the Enrichr Gene Ontology enrichment tool.[84] Results were ranked by the "combined score", which combines p-value and z-score by multiplication, i.e. combined score = log(p) * z-score.

Other statistical analysis

Other Statistical analysis was performed using Scipy package in Python 3.