A novel NLP-based method and algorithm to discover RNA-binding protein (RBP) motifs, contexts, binding preferences, and interactions

Shaimae I. Elhajjajy[1,*] and Zhiping Weng[1,*]

[1]Department of Genomics and Computational Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA

[*]To whom correspondence should be addressed.

sielhajjajy@gmail.com (SIE); zhipingweng@gmail.com (ZW)

Shortened title (max 50 characters): A novel method to discover RBP binding patterns

Keywords (max 6): RNA-binding proteins, RBP interactions, NLP

ABSTRACT

RNA-binding proteins (RBPs) are essential modulators in the regulation of mRNA processing. The binding patterns, interactions, and functions of most RBPs are not well-characterized. Previous studies have shown that motif context is an important contributor to RBP binding specificity, but its precise role remains unclear. Despite recent computational advances to predict RBP binding, existing methods are challenging to interpret and largely lack a categorical focus on RBP motif contexts and RBP-RBP interactions. There remains a need for interpretable predictive models to disambiguate the contextual determinants of RBP binding specificity *in vivo*. Here, we present a novel and comprehensive pipeline to address these knowledge gaps. We devise a Natural Language Processing-based decomposition method to deconstruct sequences into entities consisting of a central target *k*-mer and its flanking regions, then use this representation to formulate the RBP binding prediction task as a weakly supervised Multiple Instance Learning problem. To interpret our predictions, we introduce a deterministic motif discovery algorithm designed to handle our data structure, recapitulating the established motifs of numerous RBPs as validation. Importantly, we characterize the binding motifs and binding contexts for 71 RBPs, with many of them being novel. Finally, through feature integration, transitive inference, and a new cross-prediction approach, we propose novel cooperative and competitive RBP-RBP interaction partners and hypothesize their potential regulatory functions. In summary, we present a complete computational strategy for investigating the contextual determinants of specific RBP binding, and we demonstrate the significance of our findings in delineating RBP binding patterns, interactions, and functions.

INTRODUCTION

RNA-binding proteins (RBPs) extensively govern the biogenesis and maturation of nascent messenger RNAs (mRNAs) in co- and post-transcriptional manners (Gerstberger et al. 2014; Bentley 2014). The human genome encodes approximately 1,500 RBPs, each of which has diverse roles in regulating nearly every step of mRNA processing, including transcription, canonical and alternative splicing, 3' end cleavage and polyadenylation, stability, localization and transport, degradation, and translation (Gerstberger et al. 2014). RBPs accomplish these tasks by binding their mRNA targets to form messenger ribonucleoprotein (mRNP) complexes whose constituents are

dynamically remodeled based on the functional requirements at each stage of mRNA processing (Gerstberger et al. 2014; Müller-McNicoll and Neugebauer 2013; Gehring et al. 2017). Thus, the activities and functions of RBPs are major determinants of the homeostatic expression and heterogeneity of the steady-state transcriptome. The dysregulation of RBPs can manifest into a variety of Mendelian (Castello et al. 2013) and somatic genetic diseases (Gebauer et al. 2021), including, most notably, neurological disorders (Nussbacher et al. 2015; Conlon and Manley 2017) and cancers (Pereira et al. 2017). Because the majority of RBPs catalogued to-date are not well-characterized, an improved understanding of RBP binding patterns, activities, functions, and interactions is needed to delineate their constitutive roles in mediating mRNA metabolism and to discern their causal roles in disease.

RBPs interact with their target mRNAs through RNA-binding domains (RBDs), which confer upon the RBPs their sequence and structural binding preferences (Gerstberger et al. 2014; Müller-McNicoll and Neugebauer 2013; Gehring et al. 2017). The complex and enigmatic binding patterns of RBPs can be attributed in part to their structural composition: they can possess single or multiple RBDs that belong to the same or different domain class (Gerstberger et al. 2014), and these RBDs may be either well-characterized or non-canonical (Ray et al. 2023), together enabling a sizable potential interactome. Due to their short motif length (between 3 to 8 nucleotides) and low nucleotide complexity (often containing only a subset of the 4 canonical nucleotides), RBPs with structurally diverse RBDs can recognize highly similar motifs, but they still assert a remarkable degree of transcriptomic binding specificity (Dominguez et al. 2018). Based on large-scale *in vitro* experiments, previous studies have posited that the flanking sequences of RBP binding sites may play a critical role in enabling RBP binding specificity by fostering conformations conducive to stable binding; however, these findings were based on short, synthetic oligonucleotide sequences in isolated, exogenous conditions lacking a standard cellular environment (Dominguez et al. 2018). Further investigation is necessary to corroborate the contribution of these flanking regions to RBP binding specificity *in vivo*.

Adding another layer of complexity is the high propensity for RBPs to interact with one another, working in tandem to jointly regulate common targets (Dassi 2017). In order to account for the vastness of the transcriptome and the

diversity of associated regulatory functions, RBPs are further believed to interact combinatorially, forming complexes of varying sizes and composed of varying components (Khoroshkin et al. 2024); this is yet another attribute that may enable RBPs to achieve their observed degree of binding specificity (Lang et al. 2021; Street et al. 2024). These RBP-RBP interactions may occur via direct (i.e., requiring RNA binding and physical interaction between RBPs) or indirect (i.e., requiring RNA binding but not physical interaction between RBPs) binding modes (Street et al. 2024), and they may operate in cooperative, competitive, or mutual manners (Dassi 2017). Such interactions, which may involve as few as 2 RBPs or as many as required for large macromolecular complexes like the spliceosome, are vital to the fidelity of mRNA regulatory processes (Lang et al. 2021). Previous work has provided experimental evidence of an extensive RBP-RBP interactive network (Dassi 2017; Quattrone and Dassi 2019; Lang et al. 2021; Street et al. 2024), but the biological mechanisms for the majority of these interactions have not yet been determined. In short, the exponentially large landscape and functional implications of the RBP-RBP interactome remain underexplored (Khoroshkin et al. 2024).

Machine learning and deep learning have become increasingly prevalent in genomics applications, especially for the prediction of RBP binding (Alipanahi et al. 2015; Ghanbari and Ohler 2020; Sun et al. 2021; Sharma et al. 2021). While such applications at large have contributed to our current understanding of RBP binding, there remain significant areas that have been under- or never before investigated, such as (1) the structure, composition, and role of motifs to RBP binding and interactions, (2) the salient features within the regions flanking RBP binding motifs, (3) the *in vivo* preferences of RBP motifs and flanking regions, (4) the role of flanking regions in facilitating RBP binding and interactions, and (5) the similarities in flanking regions across RBPs. Furthermore, current motif discovery methods rely on enumerative or probabilistic approaches (D'haeseleer 2006a) but do not consider the contributions of *k*-mers in the surrounding flanking regions when constructing motifs, which may cause noise or inaccuracies to be introduced into the discovered motif. Therefore, we elected to focus on two areas of great import to mRNA regulation: the contribution of motifs and their flanking regions to RBP binding and RBP-RBP interactions, as well as the combinatorial capacity of RBP-RBP interactions. We defined domain requirements to guide our formulation and construction of a comprehensive pipeline to address these RBP knowledge gaps.

Namely, we aimed to develop an architecture-independent solution – to maximize flexibility and generalizability – that balances the performance-interpretability tradeoff intrinsic to neural networks, whose "black box" nature makes their decision-making rationales ambiguous to the user (Hwang et al. 2024). We strove to design an interpretable strategy that achieves high accuracy in RBP binding prediction and motif discovery, both of which are integral to the investigation of motif flanking regions and RBP-RBP interactions. As opposed to commonly-practiced top-down approaches that search for significant regions within a full-length, intact sequence, we instead adopt a bottom-up approach in which our goal is to deconstruct a sequence into its constituent *k*-mers and regions, discover the role and importance of each constituent, and then reconstruct the sequence based on the significant elements that were discovered.

Given that RBPs only bind motif instances that are overrepresented and that are located in a favorable environment with significant flanking regions, we sought to understand the composition, structure, and contribution of these flanking regions to sites of RBP binding events. To this end, we required a method by which to identify the overrepresented *k*-mers that constitute the motif and its proximal flanking regions, the similarities between *k*-mer instances of the binding motif, the relationships between *k*-mers comprising the motif and *k*-mers comprising the motif's flanking regions, and the relationships between *k*-mers within the flanking regions themselves. The genomics and computational biology fields have speculated that DNA and RNA possess an underlying language and linguistic structure, whose components and inner workings remain elusive (Durbin et al. 1998). In our endeavor to understand the *k*-mers contributing to RBP binding, we regarded RNA as possessing features and structures akin to those found in natural languages and were thus motivated to model RNA using concepts from natural language processing (NLP). While natural language has a defined lexicon, syntax, and semantics, as well as established delimiters which enable text parsing, analogous structures and representations of genomic sequences are either missing, difficult to formalize, or not yet discovered. For our work, we decided to model sentences as sequences, phrases as significant regions, and words as *k*-mers. Furthermore, a crucial similitude between natural language and genomic language is that context confers meaning, which makes NLP particularly apt for this study.

Here, we present a novel pipeline that integrates state-of-the-art concepts from natural language processing (NLP), weak supervision, and deep learning, as well as a novel motif discovery algorithm, to bridge the knowledge gaps pertaining to RBP binding patterns. To this end, we devised an NLP-based sequence decomposition strategy that establishes a new representation for candidate RBP binding sites and their flanking regions, which we then used to train neural networks to predict RBP binding under a Multiple Instance Learning (MIL) framework. We also developed a consensus-based motif discovery algorithm to interpret our predictions, demonstrating its ability to accurately recapitulate the motifs of several well-characterized RBPs. Consequently, and importantly, we systematically identified and profiled the flanking regions and binding preferences for 71 RBPs in the HepG2 cell line, the majority of which are novel, and we conducted a thorough analysis of the similarities and differences in binding patterns between RBPs. Finally, we integrated our discovered motifs and flanking regions with several analysis metrics and a new cross-prediction approach to propose novel cooperative and competitive RBP-RBP interactions, for which we additionally hypothesize underlying mechanisms and functions. Taken together, we present a comprehensive computational approach that enables characterization of the features contributing to RBP binding specificity, as well as investigation into the participants and functions of RBP interactions.

RESULTS

A novel NLP-based sequence decomposition method to represent RBP motifs and contexts

We drew on theoretical concepts from natural language processing (NLP) to design an appropriate method of sequence representation for RBP binding sites and their flanking regions. We utilized merged replicate enhanced cross-linking and immunoprecipitation (eCLIP) sequencing data (Van Nostrand et al. 2016) from the ENCODE consortium (Van Nostrand et al. 2020a) due to its coverage of a large array of RBPs, its standardized, uniformly processed nature, and its use of input controls to ensure high-confident, high-quality peaks (**Supplemental Table S1**). To construct our training dataset, we defined positive sequences as centered eCLIP peaks resized to a uniform length of 101 bp; for our negative set, we performed random sampling of sequences with equivalent length from matched genomic regions (**Figure 1**; **Materials and Methods**). We operated under the assumptions that (1) each

peak sequence represents a putative RBP binding site, (2) each binding site has a significantly greater abundance of enriched *k*-mers with respect to corresponding control sequences, and (3) these enriched *k*-mers contribute to the RBP-RNA interaction either as the motif instance being recognized by the RBP or as a contextual element within the flanking RNA sequence which promotes the binding event. To gain insight into RBP binding patterns, we sought to identify which of these enriched *k*-mers constitute the RBP binding site, and which comprise the binding site's flanking regions.

With this paradigm, we designed a strategy to deconstruct our sequences into a series of regions, which we term "contexts," each consisting of three elements: a central *k*-mer called the "target," representing a candidate RBP binding site, and two sets of multiple *k*-mers each representing the target's flanks, called the "left context" and "right context," respectively (**Materials and Methods**). We achieve this decomposition by employing a double sliding window that first deconstructs a sequence into a series of potential targets in the first stage (**Figure 1A**), then constructs the context for each target in the second stage (**Figure 1B**). Our approach introduces a new representation structure that clearly demarcates the significant entities constituting an RBP binding event (i.e., target and context), and it does so in an unbiased manner by treating every *k*-mer as a candidate target for RBP binding. It is also highly flexible in implementation, with customizable parameters that enable exploration of targets and contexts independently having different *k*-mer lengths, window strides, region sizes, and gap presence. We decided to conduct our study using *5*-mers for both the target and context elements due to its demonstration as a robust and representative RBP motif length (Dominguez et al. 2018), a window stride of 1 to consider every target as a potential binding site, and 5 *k*-mers each in the left and right context elements to allow inspection of a respectably-sized region for contextual effects.

Formulation of RBP binding prediction as a Multiple Instance Learning (MIL) problem

Our main objective is to increase the resolution of RBP binding prediction by identifying the precise context(s) per positive sequence that are necessary for an RBP binding event to occur. Multiple Instance Learning (MIL) – a branch of weakly supervised machine learning characterized by its arrangement of observations as "bags" and

subsets of those observations as the bags' "instances" (Dietterich et al. 1997; Herrera et al. 2016; Carbonneau et al. 2018) – is a particularly apt framework due to its natural ability to model the context data structure. In our formulation, we denote a sequence as a bag and the sequence's constituent contexts as the bag's instances (**Figure 1C**). We adhere to the collective MIL assumption that a bag's instances constitute a random representative sample from the parent bag's inherent probability distribution, such that each instance has an equal contribution to the label of the bag (Foulds and Frank 2010), but the instance labels themselves are unknown (Herrera et al. 2016).

Within this MIL framework, we sought to determine the identity of each instance label in order to discern which region(s) within a sequence significantly contribute to the associated RBP binding event. To this end, we trained neural networks to predict RBP binding using standard encoding, embedding, and hyperparameter techniques from NLP and deep learning (**Figure 1D**; **Materials and Methods**). We initialized instance labels for training using label inheritance from the parent bag (Carbonneau et al. 2018). We then performed binary classification at the instance-level to determine the identity of positive context(s) per sequence – which signify a contribution to RBP binding – and at the bag-level using an averaging aggregation technique, which allows assessment of model performance with respect to the ground-truth sequence labels (Herrera et al. 2016; Foulds and Frank 2010).

To filter out noise that is intrinsic to label inheritance, we also implemented a strategy that we term iterative relabeling, in which each training iteration's instance-level predictions are used to reassign context labels prior to the subsequent iteration of training (**Figure 1D**; **Materials and Methods**). Our models generally performed well across an array of 71 RBPs, with a final median bag AUPR of 86.5% (**Figure 1E; Supplemental Table S2**). Despite an apparent dispersion in the distribution of post-relabeling context performance, we attribute this phenomenon to the natural discrepancy between instance labels of two successive iterations. Notably, after relabeling, the bag performance was largely preserved even as the witness rate (i.e., the number of positive instances) (Carbonneau et al. 2018) substantially decreased by about 1.4-fold at the global level and 1.8-fold at the local level (**Supplemental Table S3**; **Materials and Methods**), indicating that our trained models could successfully achieve the removal of labeling noise from positive sequences while simultaneously maintaining bag-level prediction accuracy.

A novel algorithm for the accurate discovery of RBP binding motifs and contexts

In order to derive RBP motifs and, importantly, RBP binding contexts directly from the MIL predictions, we required a method of interpretation for our models that could accommodate the context data structure. To this end, we developed a deterministic, consensus-based motif discovery algorithm (**Figure 2A**) that relies on empirically-derived measurements of three theoretically significant *k*-mer properties to discover motifs.

The algorithm can be regarded as having six phases (**Materials and Methods**). In the first phase, in accordance with our initial assumptions that RBP binding sites consist of overrepresented patterns, we used *k*-mer enrichments to construct a set of strong candidate consensuses, such that the set consisted of all *k*-mers that were enriched as targets of positively-predicted contexts with respect to negatively-predicted contexts (**Figure 2A**, top left). By only considering enriched *k*-mers as candidate consensuses, we simultaneously reduce the motif search space. In the second phase, to model the relative sequence similarity observed between a motif's consensus and its constituent instances (D'haeseleer 2006a, 2006b), and to account for the degenerate nature of RBP motifs (Müller-McNicoll and Neugebauer 2013), for each candidate consensus we assembled a preliminary list termed a "partition" containing all enriched *k*-mers with restricted sequence similarity to that consensus (**Figure 2A**, top right; **Materials and Methods**).

The *k*-mer enrichment and similarity requirements proved insufficient for accurate motif discovery. We thus required an additional constraint for fine-tuning each initial partition to contain only the most highly-probable instances for a given motif. We had previously observed in our data an intriguing relationship between a motif consensus and any putative motif instance, whereby the two *k*-mers occurred within the same sequence at some frequency; we termed this feature the *k*-mer co-occurrence. We also noticed that by enforcing a *k*-mer co-occurrence threshold, we could filter out *k*-mers that were less likely to be part of the true motif. In the third phase, to arrive at a final, filtered partition for each RBP, we developed a programmatic, convergence-based minimization approach to tune the *k*-mer co-occurrence threshold, such that only those *k*-mers belonging to the initial partition

that met this threshold would be retained in the final partition (**Figure 2A**, bottom left; **Materials and Methods**). Taken together, *k*-mer enrichment, *k*-mer similarity, and *k*-mer co-occurrence are the three necessary and sufficient conditions for our algorithm's successful motif discovery.

In the fourth phase, we constructed the final motifs by obtaining every instance of the final partition's constituents that occurred as targets of positive context predictions. For cases in which more than one motif instance was present within the same sequence, we selected only one instance to incorporate into the final motif, prioritizing those instances with higher enrichments and prediction scores (**Figure 2A**, bottom middle). Our algorithm discovers all possible motifs for a given RBP dataset. In the fifth phase, we select a primary motif by using a multi-metric, iterative sorting strategy based on the enrichment and statistical significance of the motif's consensus as well as the weighted relative entropy (Tompa 1999) of the final motif (**Materials and Methods**). In the sixth and final phase, we derived the motif's context by extracting the flanking regions surrounding each motif instance using symmetrical extension. For both the motif and context, we aligned extracted instances and regions, respectively, to visualize the sequence logos and nucleotide binding preferences (**Figure 2A**, bottom right).

To validate the algorithm, we demonstrate our ability to recapitulate known motifs from the literature for a series of diverse RBPs. For this evaluation, we opted against performing a cross-reference analysis using tools that compare user-provided motifs with those from existing databases in order to determine the identity of the proteins to which the motifs belong (Bailey et al. 2015); while such an approach is certainly useful in specific circumstances, we deemed it unsuitable for our purposes because (1) most available databases are transcription factor-centric, (2) the available RBP databases are primarily derived from *in vitro* data only, (3) the available RBP databases are limited in the number of RBPs and motifs they contain with respect to the full scope of literature available, and (4) the outcomes of such analyses may be misleading, as they do not account for RBPs which may recognize different binding motifs depending on which RBD is interacting with its target mRNA at the time of capture. Instead, we thoroughly reviewed the literature to curate a set of ground-truth motifs representing only those RBPs having well-characterized binding patterns, with either specified logos, consensus sequences, or nucleotide preferences that are

concordant across multiple sources. Of the total 14 RBPs in this ground-truth set, we successfully discovered the motifs for 13 RBPs, yielding an accuracy rate of 92.9% (**Supplemental Table S4**); we highlight 4 representative examples in **Figure 2B**. The sole RBP from this set for which our discovered motif did not correspond with that from the literature was PABPN1, a nuclear poly(A) binding protein (Huang et al. 2023) for which we discovered a motif consensus of GCUGG. The poly(A) *k*-mer itself was not enriched in our predictions, signifying that the absence of the poly(A) motif is inherent to the dataset. We hypothesize that our discovered motif represents either (1) a novel motif for PABPN1 that reflects a non-canonical function other than regulating poly(A) tail formation in the nucleus (Huang et al. 2023), or (2) the motif for an as-of-yet unknown, RNA-bound RBP with which PABPN1 physically interacts, in which case the PABPN1-specific antibody would immunoprecipitate mRNAs crosslinked with the other RBP. These hypotheses may be supported, respectively, by the fact that over 60% of PABPN1's discovered motif instances are located in coding sequences and by previous reports of PABPN1 interacting with numerous spliceosomal proteins (Huang et al. 2023), thus pointing overall to a motif with PABPN1-mediated roles in mRNA splicing.

As a more comprehensive proof-of-principle, we further inspected RBFOX2, a well-characterized RBP with established regulatory functions in mRNA splicing and mRNA stability that is implicated in neurological developmental processes and disease (Begg et al. 2020). Our algorithm was able to reproduce RBFOX2's canonical GCAUG motif, as supported by motifs derived from both *in vitro* and *in vivo* assays (**Figure 2B**, bottom row) (Van Nostrand et al. 2020a). Previous studies have also demonstrated that approximately half of RBFOX2's binding sites contain alternative, non-GCAUG motifs, which have since been identified *in vitro* after successive rounds of iterative enrichment calculations from a synthetic RNA library modeled solely after 3'UTRs (Begg et al. 2020). Crucially, our discovered motif contains 6 of the 8 *k*-mers identified in the *in vitro* study: 2 primary (GCAUG and GCACG) and 4 secondary (GCUUG, GUAUG, GUGUG, and GUUUG) (**Figure 2C**). To the best of our knowledge, this marks the first documentation of an *in vivo*-derived motif that successfully captures the majority of RBFOX2's alternative binding *k*-mers.

Our RBFOX2 motif also contained 4 additional *k*-mers that have not been previously identified. GACUG and GCGUG maintain the critical G1, U4, and G5 positions of the canonical motif; although these 2 *k*-mers were not included in the set of GHNUG *k*-mers characterized in (Begg et al. 2020), they can still be denoted by this IUPAC nucleotide representation (Cornish-Bowden 1985). In conjunction with their relatively frequent occurrence in our discovered motif, especially in comparison to some of the established primary and secondary *k*-mers (**Supplemental Table S5**), these properties suggest that GACUG and GCGUG may serve as novel candidates for lower-affinity RBFOX2 secondary *k*-mers; in fact, GACUG has been previously reported in another study (Zhou et al. 2007). The remaining 2 *k*-mers, GGUAG and GUUAG, which do not share the GHNUG representation and which occur less frequently, may also present novel, lower-affinity secondary motifs with a distinct nucleotide content and lower similarity to the GCAUG consensus, or they may be attributed to noise; because they were detected by the algorithm and satisfy the same requirements as the other established motif instances, we posit that the former case is more probable.

Our context logo for RBFOX2 revealed a strong U directly preceding the GCAUG motif, a finding that coincides with its well-profiled canonical *6*-mer motif, (U)GCAUG (Begg et al. 2020). The nucleotide preference of RBFOX2's motif context also corroborated *in vitro* findings (Dominguez et al. 2018), with RBFOX2 demonstrating a clear predisposition for binding in G-rich regions (**Figure 2E, 2F**). For a more widespread comparison, we compared our discovered contexts with *in vitro*-derived contexts for 8 RBPs that were present in our collection and deemed to have significant nucleotide preferences in (Dominguez et al. 2018), showing that we could recapitulate the putative nucleotide preferences for 5 of them (**Supplemental Table S6A**); for the remaining 3, we closely examined the contexts to explain any discrepancies observed in comparison to the *in vitro* study. For HNRNPL, we discovered a symmetrically oscillating CA-rich context, while an average preference for Cs was reported *in vitro*; however, the *in vitro* context also had a relatively strong underlying presence of As – especially at the positions adjacent to the motif – and displayed a depletion of Gs similar to that observed in our context. These differences were observed despite the fact that our discovered HNRNPL motif consensus – ACACA – was concordant with the *in vitro*-derived motif. Second, while our discovered context for IGF2BP1 did not exhibit a strong preference for

any specific nucleotide, it did demonstrate a slight, albeit asymmetric, preference for Cs; on the other hand, the *in vitro* context was strongly C-rich on average. Finally, SFPQ's discovered context was strongly G-rich, while its *in vitro* average nucleotide preference was reported to be C-rich; interestingly, the more detailed *in vitro* context for SFPQ showed that the content was actually C-rich on the left-hand side of the motif but G-rich on the right-hand side of the motif.

In short, we attribute the discordant context preferences for these 3 RBPs to two main reasons. First, an averaged representation of nucleotide preference can be misleading, as it may obscure the underlying position-specific preferences. Second, and perhaps with the most implications, in general *in vitro* experiments are inherently limited by (1) the synthetic, randomized nature of the RNAs used, which may not reflect the intrinsic nucleotide content observed in the transcriptome (e.g., repeat regions), (2) the short length of the utilized RNAs, which may not fully encompass the native context required *in vivo* for RBP binding, and (3) the absence of a natural cellular environment, causing RBPs to be studied in isolation without the presence of endogenous *trans*-factors and other signals that may influence its binding partners or target RNAs (Khoroshkin et al. 2024). The accuracy of our discovered contexts is supported not only by the 5 RBPs that corroborated *in vitro* findings, but additionally and importantly by the facts that (1) we previously established that our discovered motifs are accurate, which in turn suggests that their corresponding contexts must be valid, (2) the accuracy of our motif discovery algorithm demonstrates that the *k*-mer properties used to derive motifs are valid, and (3) two of those properties – namely, *k*-mer enrichment and *k*-mer co-occurrence – use information from the context of the motif instances to construct the motif itself.

A comprehensive characterization of the binding motifs and contexts for 71 RBPs

We report the discovered motifs, contexts, and binding preferences for a total of 71 RBPs in the HepG2 cell line (**Supplemental Table S7**). To gain an improved understanding of our discovered motifs for the full set of RBPs, we classified each into one of 5 categories (**Supplemental Table S8**):

- Consistently reported (N = 14): The RBP has a well-characterized binding motif that has been consistently reported across multiple sources in the literature.
- Multiple reported (N = 22): Multiple different motifs have previously been reported for the RBP; we either present a motif that supports at least one of these reported motifs or a novel motif that does not correspond with any of those previously reported.
- Weakly characterized (N = 25): Limited information is available regarding the binding motif of the RBP (e.g., only 1 source available, only nucleotide preferences [but not a concrete motif] available); we present a more specific motif.
- Non-sequence-specific (N = 6): The RBP is believed to recognize features other than sequence, such as RNA secondary structure.
- No existing characterization (N = 4): To the best of our knowledge, no motif has been previously reported for the RBP, and we present a novel binding motif.

We have addressed the "Consistently reported" classification in the preceding section. In our "Multiple reported" category, our discovered motif for a given RBP often corresponded with at least one of the reported motifs. For example, FUBP3, an RBP for which we discovered a motif with a UGUGU consensus, corroborated the UG-repeat motif discovered in an *in vivo* study (Gau et al. 2011) but was distinct from the UA-rich motifs discovered *in vitro* (Dominguez et al. 2018); this discrepancy could be attributed to technical reasons underlying the environments of the respective assays used to generate the binding data. On the other hand, discrepancies between multiple reported motifs could also be attributed to biological reasons based on the RBP's binding mode. For example, KHSRP, an RBP belonging to the same protein family as FUBP3 (Gau et al. 2011) and for which we also discovered a UGUGU motif, has been shown to bind both GU-rich and AU-rich elements *in vivo* (Briata et al. 2016; Cook et al. 2011). KHSRP has 4 KH RBDs, each having varying degrees of specificity, with the third KH domain in particular demonstrating a strong preference for G-containing sequences (García-Mayoral et al. 2008; Palzer et al. 2022). We thus hypothesize that our UGUGU motif may reflect that KHSRP's utilization of KH-3 for RNA binding was

more prevalent during eCLIP capture. Our KHSRP motif also exhibited a characteristic depletion in Cs that has previously been observed (García-Mayoral et al. 2008) (**Figure 3C**).

For RBPs in this category whose discovered motif did not correspond with at least one of the reported motifs, we hypothesized biological explanations for our novel motif. Interestingly, we discovered a GGGGG motif for CSTF2T, a widely-expressed RBP with roles in 3'end cleavage that is a component of the cleavage stimulation factor (CSTF) complex along with its paralog CSTF2, with which it shares similar RNA binding specificity (Yao et al. 2013). The CSTF complex has been shown to bind U-/GU-rich core downstream elements (CDEs) located following a polyadenylation signal; G-rich auxiliary downstream elements (ADEs) even further downstream are thought to stabilize CSTF's binding to these CDEs (Darmon and Lutz 2012). Although CSTF2T has the highest affinity for poly(GU) and poly(U) sequences, as observed in CDEs, it also has moderate affinity for poly(G) sequences (Monarez et al. 2007) and has been proposed as a binder of RNA G-quadruplex (rG4) structures (Lee et al. 2020). The majority (~78%) of our discovered motif instances were located in either proximal or distal introns; alternative polyadenylation (APA) has been shown to occur frequently in introns (Di Giammartino et al. 2011), and intronic APA is often coordinated with co-transcriptional splicing (Mitschka and Mayr 2022). Taken together, we hypothesize that our discovered G-rich motif may either (1) provide support for CSTF2T as a rG4-binding RBP, (2) suggest that CSTF2T is associated with G-rich ADEs during 3'-end processing to stabilize binding of the CSTF complex, and/or (3) indicate that CSTF2T (and the CSTF complex) has a more prominent role than previously appreciated in the cleavage of transcripts undergoing intronic APA. These hypotheses will require further investigation.

We also examined the additional motifs that were discovered by our algorithm but not ranked first by our scoring strategy, which we refer to collectively as secondary motifs. In the case of LIN28B, we discovered a GGAGA secondary motif that supported the GGAG motif discovered for LIN28B in a previous study of its transcriptomic targets (Graf et al. 2013) as well as a GGAGA motif discovered for LIN28B's paralog, LIN28A, with which it

shares a similar structure (Wilbert et al. 2012; Maklad et al. 2023). This demonstrates that our method still has the power to discover previously reported motifs, even if not ranked first by our scoring method.

For our "Weakly characterized" category, we propose new, specific motifs for 25 RBPs that we considered to be weakly characterized. For 10 of these RBPs, a previous study had reported their candidacy for binding to rG4s (Lee et al. 2020; Busa et al. 2021; He et al. 2021), and our discovered motifs (and contexts) – which were all G-rich – supported this theory. For some RBPs that only had a reported motif from a single source, our discovered motifs were concordant, thus providing additional support; for example, PPIG and PRPF8, which both have GUGAG motifs (**Figure 3C**) (Feng et al. 2019). For other RBPs, we propose novel motifs (e.g., GGGUG for BUD13, CACAC for CDC40, GAAGA for FXR2; **Supplemental Table S8**).

Our "Non-sequence-specific" category contains a number of RBPs that have previously been hypothesized to bind in a non-sequence-specific manner. Many of these RBPs are DEAD-box helicases, which recognize the RNA sugar-phosphate backbone as opposed to its nucleotides (Linder and Jankowsky 2011), while the other two have either been reported to have double-stranded RBDs (ILF3) (Li et al. 2020) or to bind highly structured RNAs (XPO5) (Wang et al. 2020). We nevertheless present novel binding motifs for these RBPs, hypothesizing that the motifs may instead reveal the sequence preferences of other *trans*-factors that interact with the RBP in question, as we elaborate on in the following sections. Finally, to the best of our knowledge, the motifs for 4 of the 71 total RBPs have not been previously characterized, comprising our "No existing characterization" category. This set includes EXOSC5, HLTF, RPS3, and ZNF800.

In the same vein, and most importantly, we sought to comprehensively characterize the binding contexts for the set of RBPs covered in our study. In total, we discovered novel *in vivo* contexts for 71 RBPs. For the contexts of the 5 RBPs for which we corroborated with *in vitro* data, we provide additional detail by reporting a larger context region, but also by delineating nucleotide preferences for each component of the context (below). For other RBPs, such as those known to bind polypyrimidine tracts (e.g., PTBP1, U2AF2) (Sayers et al. 2022), we recapitulated the

nucleotide preferences of their contexts, but also provided a more detailed report of these preferences per position. For the remaining RBPs, which comprise the majority of the set, we present for the first time, to the best of our knowledge, a comprehensive profile of their preferred contexts for binding *in vivo*.

Notably, we observed a strong degree of symmetry for the majority of our discovered contexts (N = 59), where we define symmetry to occur if the left and right contexts flanking the motif share the same nucleotide preference. (**Supplemental Table S6B**; **Materials and Methods**). We noticed that the symmetric preference was for either a single nucleotide or for 2 frequently occurring nucleotides. The remaining 12 RBPs had contexts that were asymmetric in nucleotide preference, which displayed a bit more variance. In the case of ILF3, for example, each side of the context displayed preference for a single, albeit different, nucleotide, where its left context showed preference for C but its right for G (**Supplemental Table S6B**). In other cases, each flank of the context displayed preference for 2 nucleotides of different identity (e.g., IGF2BP1; **Supplemental Table S6B**), or the left and right contexts each displayed preference for a different number of nucleotides (e.g., PABPN1; **Supplemental Table S6B**). By visualizing the PFMs, we noted that some RBPs had stronger, more pronounced nucleotide preferences than other RBPs.

Furthermore, we observed that the majority of RBPs (N = 46) had preference for the same nucleotide in both their motif and context (**Materials and Methods**). For example, G3BP1 is AG-rich in both its motif and contexts (**Supplemental Table S6B**). While the remaining 25 RBPs had different nucleotide preferences in their motif and contexts, the difference in preference was generally not very striking. For example, while HLTF's motif is AG-rich, its contexts are overall A-rich, indicating that there is still some degree of similarity between the nucleotide preference of each element (**Supplemental Table S6B**). One exception to this phenomenon is QKI, which demonstrates a marked difference in the nucleotide preferences of its elements, having an AC-rich motif but a U-rich context (**Figure 3C**; **Supplemental Table S6B**).

Investigating the relationships of binding patterns between RBPs

Equipped with our characterization of RBP binding preferences, we next sought to understand the similarities and differences in these preferences between RBPs. First, we used HDBSCAN (McInnes et al. 2017) to cluster the flattened position frequency matrices (PFMs) of the motifs and contexts, respectively, of all RBPs, then visualized the dimension-reduced clusters using a two-dimensional UMAP embedding (McInnes et al. 2018) (**Figure 3A, 4B**). Using HDBSCAN and UMAP with the Euclidean metric to compute distances between points – with the former technique seeking to assign clusters based on point proximity and the latter technique aiming to preserve the relative distances observed in high-dimensional space in its lower-dimensional mapping – we can use the results as a method of motif comparison, where a small Euclidean distance and same cluster membership suggest that two RBPs in question may possess similar motif and context patterns, as well as similar nucleotide binding preferences.

Indeed, we observed that RBPs clustered together by nucleotide preference. Clustering by motif yielded 9 clusters in total (**Figure 3A**), with distinct groups forming for RBPs with generally UG-rich (Cluster 1), U-rich (Cluster 2), C-rich (Cluster 3), CG-rich (Cluster 4), and A-rich (Cluster 9) motifs, each with a moderate degree of variation (**Figure 3C**). Cluster 6 contained 3 RBPs with relatively different motifs from one another but are characterized by their strong G1 and U5 positions. The last three clusters consisted of RBPs with largely G-rich motifs, but their division into subgroups can be attributed to the presence of a distinct C5 position (Cluster 8), strong G's at the odd-indexed or solely peripheral positions and U's at either or both of the even-indexed positions (Cluster 5), and the remaining, predominantly G-rich motifs (Cluster 7). The latter cluster is fairly large (N = 17), with a considerable degree of variation present, but with less well-defined patterns, which may help to explain why it was not further divided into subgroups.

In our context-based clustering results, we observed that the majority of RBPs remained clustered with the same members as in the motif-based clustering, with 16 of the 71 (22.5%) migrating to a new cluster (**Figure 3B**, labeled in black). We sought to understand the underlying causes that fostered these migrations, and we did so by carefully inspecting the motif and context logos of the associated RBPs (**Figure 3C**). For example, due to its strong CUUUC

consensus, PTBP1's motif was distinct enough from the set of RBPs having U-rich motifs (Motif Cluster 2, orange) and instead clustered with the set of RBPs having C-rich motifs (Motif Cluster 3, green). However, upon the addition of context, PTBP1's overwhelmingly strong preference for U's in its binding context led to its subsequent clustering with the set of RBPs having U-rich contexts (Context Cluster 3, orange), as opposed to remaining in the same cluster that it previously belonged to, whose members ultimately have C-rich contexts (Context Cluster 5, green). A similar phenomenon was observed for CDC40, HNRNPL, and QKI which, despite demonstrating a binding preference for C's and A's, have motifs with strong enough C's that they were also clustered with the set of RBPs having C-rich motifs (Motif Cluster 3, green). When contexts were incorporated, these 3 RBPs detached to form their own cluster (Context Cluster 4, blue). CDC40 and HNRNPL both had CA-rich contexts, although the latter had strong U undertones. Interestingly, in spite of its context preference for Us, QKI was still included in this cluster, likely due to its nucleotide preference similarities in motif with CDC40 and HNRNPL. This demonstrates a case in which the motif is more influential than the context in the clustering outcome. In short, the migration of certain RBPs between the motif and context clustering demonstrates that context plays an important role in differentiating between binding outcomes, particularly for RBPs sharing similar target motifs.

We sought to examine the similarities and differences between RBP binding patterns using a probabilistic approach, reasoning that the PFM representation of motifs and contexts fundamentally comprises a series of position-specific probability distributions across the 4 canonical nucleotides. To this end, we used the Kullback-Leibler Divergence (KLD) to quantify the degree to which one probability distribution Q (i.e., the approximating distribution) can approximate another probability distribution P (i.e., the reference distribution), where smaller values of KLD represent a closer approximation (Bishop and Nasrabadi 2006). Unlike the Euclidean distance metric, the KLD does not qualify as a *bona fide* metric in the formal, mathematical sense due to its asymmetric nature, and thus we instead refer to KLD as a measure of comparison. By computing the cumulative KLD across each position of a given pair of PFMs, we can obtain a quantitative method to compare PFMs between RBPs. For this reason, we used the KLD as a quantitative measure of the similarities and differences between RBP motifs and contexts for all future analyses.

We also computed the cumulative KLD between RBPs in a pairwise manner, performed hierarchical clustering on the global KLD profiles across all RBPs, and annotated each branch in the resulting dendrogram based on the corresponding nucleotide preferences of the motif (**Supplemental Fig. S1A**) and context (**Supplemental Fig. S1B**). Groups of RBPs with mutually low KLDs are visible as patches of blue in the heatmap. Notably, the hierarchy of the dendrogram allowed for ease of visualization and interpretation of the comparison between RBP binding patterns. For example, for motif KLDs (**Supplemental Fig. S1A**), the first main branch of the dendrogram splits based on whether the motif is rich in G's or rich in C's, U's or A's; because this is a major divide, and based on the number of RBPs in each of the two sets, we observe that (1) the majority of RBPs in this collection have predominantly G-rich motifs (albeit with some variation) and (2) the preference for G's is a major differentiator between the two main groups. On the other hand, when we performed the same analyses for the context KLDs, we observed deviations from the patterns observed in the motif KLD clusters, in that the nucleotide preferences at the context level did not always heavily influence the clustering of RBPs; some clusters exhibited a strong, consistent nucleotide preference, while other clusters did not. Such behavior may be attributed to (1) the increased length and complexity upon addition of the contexts, (2) the combined influences of both the motif and the contexts and/or (3) the overall nucleotide preference of each RBP not being the decisive factor in clustering, as RBPs may have contexts with sufficient similarity to be assigned to same cluster while also having different nucleotide preferences. (**Supplemental Fig. S1B**).

Cross-prediction: a novel approach to investigate RBP binding specificity

The determinants governing RBP binding specificity, particularly for those RBPs sharing similar motifs and contexts, remains an important outstanding question. RBPs recognizing similar sequence patterns must be able to differentiate between candidate sites *in vivo* in order to maintain the degree of specificity necessary for efficient and accurate mRNA regulation. We asked whether we could simulate the biological process of binding site selection *in silico*, especially to determine whether our models are capable of capturing the binding specificity exhibited by RBPs that recognize similar motifs. To this end, we developed and implemented a novel strategy that we refer to as cross-prediction, whereby the model trained on one RBP's contexts is presented with a distinct set of contexts belonging

to another RBP and predicts the binding outcome for each context in that set (**Figure 4A**). The cross-prediction performance, an aggregated measure across all contexts in the dataset, serves as a proxy for how successfully an RBP, by way of its model, can distinguish between its cognate binding sites versus those belonging to another RBP. Because this approach implicitly compares the salient features learned by each RBPs' model, it provides a measure of comparison between the binding patterns of different RBPs.

We hypothesized that a strong cross-prediction performance between a given pair of RBPs may suggest that the RBPs share certain salient features that are significant to their respective binding activities, including similarity in motifs, contexts, nucleotide preferences, and targets; it may also signify that the motif of one RBP is strongly enriched in the contexts of another RBP's motif. At the global level, a high cross-prediction performance may indicate structural, behavioral, and functional similarity between the two involved RBPs. In cases where the cross-prediction between two RBPs is strongly asymmetric, it suggests that the RBP possessing the higher cross-prediction performance recognizes a larger percentage of the salient features present in the second RBP's contexts than the other way around. We performed cross-prediction in a pairwise manner to comprehensively compare the binding patterns and activities of all RBPs. Only 316 pairs, or 6.36% of the total 4970 possible pairs (i.e., 71 choose 2), had a cross-prediction AUPR greater than 86.5% (i.e., the median cognate performance from above), indicating that the vast majority of RBP pairs demonstrate binding patterns that are distinct from one another (**Supplemental Table S9**). We subsequently generated a heatmap of cross-prediction performances, which provides an effective medium for visualizing the pairs of RBPs with strong cross-predictions (**Figure 4B**). We also conducted hierarchical clustering on the RBPs based on their respective cross-prediction profiles across the entire set of RBPs (**Figure 4C**), where assignment to the same cluster is suggestive of RBPs exhibiting similar global binding patterns to one another.

Of note, certain collections of RBPs demonstrated striking cross-prediction patterns (**Figure 4D-H**), with the largest group (**Figure 4B**, top left; **Figure 4D**, inset) being most prominent: more than 50% (40 of 71 RBPs, excluding HNRNPM and DDX55) predicted relatively well on a set of 14 RBPs. This set of 14 comprised RBPs

with strongly G-rich motifs and contexts, and different combinations of these RBPs clustered together by motif and context similarity measures (**Figure 3, Supplemental Fig. S1**). Upon further investigation, we noticed that each of the 14 RBPs have either exhibited binding of RNA G-quadruplexes (rG4s) (CSTF2T, DDX3X, FAM120A, GRSF1, GTF2F1, PRPF4), or showed binding enrichment at putative rG4s (EFTUD2, FASTKD2, FKBP4, NCBP2, NKRF, RBFOX2, SLTM, XRN2) in previous studies (Herdy et al. 2018; Lee et al. 2020; Busa et al. 2021; Dumas et al. 2021; Bourdon et al. 2023). The considerable degree of universality exhibited by cross-predictions on these 14 RBPs may indicate that their contexts contain highly similar and prevalent nucleotide patterns and/or structures; furthermore, the commonality between target genes for certain subsets of these RBPs (Lee et al. 2020) may suggest that these regions undergo coordinated regulation by a large number of factors, which might in turn reveal the importance of their target genes.

MATR3 and SUGP2 demonstrated an interesting asymmetry, in which MATR3 showed a much weaker performance when predicting on SUGP2's contexts (AUPR = 78.34%) compared to SUGP2's prediction on MATR3's contexts (AUPR = 91.58%) (**Figure 4E**, **Supplemental Table S9**). Interestingly, SUGP2 performed even better on MATR3's contexts than it did on its own cognate contexts (AUPR = 82.27%) (**Supplemental Table S9**). These RBPs, both with U-rich motifs and contexts, consistently clustered together in all motif and context similarity analyses performed (**Figure 3; Supplemental Fig. S1**), and they are furthermore the sole members of their own cross-prediction cluster, indicating a strong similarity in global binding profiles.

A group of RBPs (DDX59, HNRNPK, PCBP1, and PCBP2) with predominantly C-rich motifs and contexts (**Figure 3**) showed reciprocated, strong cross-prediction (**Figure 4F**, **Supplemental Table S9**). The lower cross-prediction by IGF2BP1/3 could be attributed to the more variable content of their motifs and contexts that did not exhibit a very strong preference for any given nucleotide. Importantly, PTBP1 also performed relatively well on the 4 strongly C-preferring RBPs (and vice versa) (**Supplemental Table S9**), suggesting that they together possess similar global binding patterns, despite PTBP1's distinct, highly dissimilar U-rich motifs and contexts (**Figure 3**); PTBP1 did not cluster with any of the 4 C-preferring RBPs by motif and context KLD. This exemplifies that our

cross-prediction results are informative regarding the identification of potential relationships between RBPs, as they can reveal what other analyses cannot.

In the final two cases, we observe that 5 RBPs predict very highly on TRA2A (**Figure 4G**), and they are also part of the same cross-prediction cluster (**Figure 4C**), indicating that the 6 RBPs together share global binding patterns (**Supplemental Table S9**). BCLAF1 is the most highly similar to TRA2A, as the 2 RBPs consistently cluster together by motif and context similarities, and often as the sole members of the cluster in question. FXR2, GRWD1, HLTF, and ZNF800 also have AG-rich motifs and contexts, which could help to explain their strong performance on TRA2A. However, this cross-prediction pattern is strongly unilateral by way of TRA2A, as the other pairs within this set exhibit considerably low cross-prediction performance; the underlying reason remains to be seen. Lastly, FUBP3 and KHSRP, two members of the same protein family (Gau et al. 2011), share high cross-prediction with SUB1 and ZC3H11A, suggesting a potentially novel relationship with RBPs outside of their family.

Taken together, the novel cross-prediction strategy we introduce proves informative of the degree of similarity between the global recognition and binding patterns of RBPs. Importantly, we posit that a strong cross-prediction performance and/or membership to the same cross-prediction cluster may be suggestive of potential RBP-RBP interactions. While the clustering assignments observed by cross-prediction may differ with those produced by HDBSCAN (**Figure 3**) or KLD (**Supplemental Fig. S1**), each of the three measures provides an alternative viewpoint of the underlying data (i.e., pattern-based, Euclidean, and probabilistic, respectively). As we demonstrate below, integration of these features and the information they carry not only helps to provide a more complete depiction of the relationship between RBPs, but also helps to increase our confidence in interaction discovery – especially if the corresponding RBPs consistently cluster together across methods.

A novel method to discover proximal interactions between RBP-RBP pairs

Given that RBPs frequently interact cooperatively or competitively to jointly regulate common targets (Dassi 2017), we developed a method to investigate RBP-RBP interactions. We first devised a conceptual delineation of the

potential RBP-RBP interaction modes into concrete categories based on properties of the involved partners, including whether the RBPs bound proximally or distally to one another, whether one or both RBPs bound RNA, and whether the RBPs physically interact (**Figure 5A**). We used this classification to outline a scope within which we performed our RBP-RBP interaction discovery. Because our analyses are context-based, do not yet incorporate RNA secondary structure data, and do not yet account for long-range interactions across genomic features, we limited our method to consider only proximal interactions between pairs of RBPs (**Figure 5A**, left-hand major branch). We define the following 4 proximal interaction modes – 3 cooperative and 1 competitive – for consideration by our method:

1. Interaction Mode 1 – **Cooperative** (proximal, dual-binding, physical): both RBPs bind the same target mRNA at distinct, non-overlapping sites to promote a shared regulatory goal. The binding sites of both RBPs are located near enough to each other such that the RBPs physically interact (**Figure 5A**, first leaf node).
2. Interaction Mode 2 – **Cooperative** (proximal, dual-binding, non-physical): both RBPs bind the same target mRNA at distinct, non-overlapping sites to promote a shared regulatory goal. The binding sites of both RBPs are located far enough apart from each other such that the RBPs do not physically interact (**Figure 5A**, second leaf node).
3. Interaction Mode 3 – **Cooperative** (proximal, single-binding, physical): the two RBPs physically interact with one another to regulate their target mRNA synergistically, but only one partner binds the target mRNA (**Figure 5A**, third leaf node).
4. Interaction Mode 4 – **Competitive** (proximal, single-binding, non-physical): the two RBPs vie for the same or overlapping binding sites on a common mRNA target to exert opposing regulatory effects (**Figure 5A**, fourth leaf node).

Note that in the competitive mode, the binding of one partner to its motif prevents the binding of the second, suggesting that their respective motifs have some degree of overlap with one another. On the other hand, in the cooperative modes, the involved RBPs either bind non-overlapping motifs in close proximity to one another, or

they form a complex in which only one partner binds its motif. Taken together, these characteristics necessitate the overlap of regions (i.e., contexts) containing motif instances in order for a proximal competitive or cooperative interaction to occur, and further implicate three main features as important for examination in interaction discovery: motif similarity, context similarity, and the relative degree of motif co-occurrence to motif overlap.

We used this information to formulate an RBP-RBP interaction discovery method based on empirical observations and guided by reported interactions in the literature, which allowed us to define a set of characteristic features that we represent as a set of conditions in the form of a decision tree (**Figure 5B**). Each branch of the tree in turn represents a set of conditions that must be satisfied for a given candidate pair to qualify as belonging to the corresponding interaction category (**Figure 5B**; **Materials and Methods**). We delineate the following 4 paths (**Figure 5B**):

- Path I: RBPs demonstrate similar motifs and contexts, and their motif co-occurrence exceeds their motif overlap, suggesting **cooperative** mechanisms corresponding to Interaction Modes 1 and 2; whether or not the interaction is physical depends on the distance between their respective binding sites.
- Path II: RBPs demonstrate similar motifs and contexts, and their motif overlap exceeds their motif co-occurrence, suggesting either **cooperation** corresponding to Interaction Mode 3, in which the RBPs interact physically but only 1 partner binds RNA, or **competition** corresponding to Interaction Mode 4, in which the RBPs contend for the same or overlapping binding sites.
- Path III: RBPs demonstrate dissimilar motifs but similar contexts, and their motif co-occurrence exceeds their motif overlap, **cooperative** mechanisms corresponding to Interaction Modes 1 and 2; whether or not the interaction is physical depends on the distance between their respective binding sites. Path III differs from Path I in its condition for motif dissimilarity, while Path I requires motif similarity.
- Path IV: RBPs demonstrate dissimilar motifs but similar contexts, and their motif overlap exceeds their motif co-occurrence, suggesting either **cooperation** corresponding to Interaction Mode 3 or **competition** corresponding to Interaction Mode 4. Path IV differs from Path II in its condition for motif dissimilarity,

while Path II requires motif similarity. Because dissimilar motifs are not likely to overlap very often, if at all, this path is expected to occur at very low frequencies.

If a candidate RBP-RBP pair does not meet the set of conditions required for any of the 4 paths, then our method considers it a non-interacting pair. Note that each of these 4 paths classifies an RBP-RBP pair to the interaction mode(s) for which it has the highest propensity, but does not preclude the possibility that the pair may interact via other modes. For example, for an RBP-RBP pair classified as cooperative according to Path I, a higher degree of motif co-occurrence than motif overlap suggests cooperation via Interaction Modes 1 and/or 2, but a non-zero motif overlap may still suggest the occurrence of interactions via Modes 3 and/or 4, albeit at lower frequencies. As such, we hypothesize that it is possible for a candidate RBP-RBP pair to exhibit both cooperative and competitive modes of interaction, but it may have stronger inclinations towards one or the other.

As a strategy to reduce the search space of candidate interacting pairs, we began our investigation by referencing the clustering results from KLD and cross-prediction analyses, reasoning that RBPs assigned to the same cluster(s) may share similar motifs, contexts, and/or global binding profiles, and thus have a stronger likelihood of being potential interacting pairs, especially if the RBPs cluster together repeatedly across multiple metrics.

To validate the conditions of this method, we first examined known interacting RBP pairs established in the literature as positive controls (**Supplemental Table S11**). We started by examining pairs of RBPs with strong, reciprocated cross-prediction scores from **Figure 4F**. We noticed that HNRNPK and PCBP2 consistently demonstrated high similarity to one another, as witnessed by their respectably low motif and context KLDs (**Supplemental Table S11A**, "Motif KLD" and "Context KLD", respectively), membership to the same cluster in global KLD comparisons (**Supplemental Fig. S1**), and remarkably strong and symmetric cross-prediction performances (**Figure 4F**; **Supplemental Table S11A**, "Cross-prediction AUPR"). We discovered C-rich motifs and contexts for both of these RBPs (**Figure 3C**), which is supported by their belonging to the poly(rC) binding protein family (Choi et al. 2009). As heterogeneous ribonucleoproteins, their functional range is extensive and

diverse, shuttling between the nucleus and cytoplasm to regulate pre-mRNA splicing, mRNA transport, mRNA stability, and mRNA translation (Kim et al. 2000; Lang et al. 2021). Experimental evidence has revealed an interaction between HNRNPK and PCBP2, with hypotheses that they may form an hnRNP complex to subsequently facilitate secondary interactions with regulatory effectors (Kim et al. 2000), but it remains unclear whether this interaction is cooperative or competitive, and whether one or both partners bind target mRNAs. Based on our finding that their motif co-occurrence (**Supplemental Table S11A**, "Percent of Motif Instance Co-occurrence") substantially exceeds their motif overlap (**Supplemental Table S11A**, "Percent of Motif Instance Overlap"), we believe that HNRNPK and PCBP2 have a higher propensity to form cooperative interactions in which both partners bind RNA (**Figure 5B**, Path I). For the less frequent cases in which their motifs overlap, this could be attributed to a cooperative mechanism in which the RBPs physically interact but only one partner binds RNA, or a competitive mechanism in which the RBPs compete for a binding site (**Figure 5B**, Path II). This case study serves as a validation of our method's ability to detect an established RBP-RBP interaction with existing experimental evidence, and we additionally propose a novel hypothesis for their potential mode(s) of interaction.

For PCBP1 and PCBP2, paralogs with previously reported interactions (Zhao et al. 2022), we noticed that the percentage of motif instances overlapping and co-occurring were strongly asymmetric, with 18.78% and 24.29% of PCBP1's motif instances overlapping and co-occurring, respectively, with PCBP2's, while only 2.11% and 3.68% of PCBP2's motif instances overlapped and co-occurred, respectively, with PCBP1's (**Supplemental Table S11B**). We inspected the size of their respective datasets and observed that PCBP1's dataset was much smaller than PCBP2's, with 1,061 motif instances being discovered for PCBP1 compared to PCBP2's 9,240 (**Supplemental Table S11B**). We hypothesize that this asymmetry, in which PCBP1 has fewer motif instances but a larger proportion overlapping or co-occurring with PCBP2's, may indicate a skewed relationship in which PCBP1 relies on PCBP2 to a greater extent than the other way around. This line of reasoning can be extended to other similar cases in which the degrees of overlap and co-occurrence demonstrate asymmetry. We also noticed that PCBP1 had a higher cross-prediction performance on PCBP2's contexts (AUPR = 94.79%) than on its own (AUPR = 90.92%), suggesting that the salient features within PCBP1's contexts that were learned by its cognate model are very strong, as well as highly

similar to the salient features present in PCBP2's contexts (**Supplemental Table S11B**). In addition, while our interaction discovery method classifies this pair as a cooperative interaction corresponding to Path I (**Materials and Methods**; **Supplemental Table S10**), we note that the degree of motif overlap is relatively close to the degree of motif co-occurrence for each RBP, which may suggest some proclivity for either a cooperative or competitive binding mode represented by Path II. Using other RBPs from **Figure 4F** as a reference, we additionally verified PCBP2-PTBP1 (**Supplemental Table S11C**; Path III, **Supplemental Table S10**), whose interactions have also been previously reported (Kim et al. 2000), and we hypothesize that this pair interacts cooperatively along Path III.

We used the KLD and cross-prediction results to guide our identification of additional positive controls. We noticed that, although their membership to the same cluster depended on the analysis, MATR3 predicted very strongly on PTBP1's contexts (AUPR = 95.75%; **Supplemental Table S11D**), and that the cross-prediction performance was even greater than MATR3's performance on its own contexts (AUPR = 91.08%), suggesting that there were significant patterns within PTBP1's contexts that are recognized by MATR3. Indeed, an interaction between MATR3 and PTBP1 has been reported to co-regulate splicing repression for a subset of MATR3 targets (Coelho et al. 2015). Based on our discovered motifs and contexts, we found that the co-occurrence between MATR3 and PTBP1 motif instances exceeds their overlap considerably (**Supplemental Table S11D**), thus supporting this established cooperative interaction; we add that the interaction may be more likely to occur in cooperative binding modes corresponding to Path I (**Figure 5B**; **Supplemental Table S10**).

As a final example, MATR3 and SUGP2 caught our attention due to their consistent clustering in all analyses performed (**Figure 3**; **Supplemental Fig. S1**; **Figure 4C**), even being the sole constituents of their cross-prediction cluster (**Figure 4C**, **E**), which suggests that their respective global binding profiles are significantly similar to one another while simultaneously being distinct from those of all other RBPs. These two RBPs both have binding sites enriched in antisense L1 elements (Van Nostrand et al. 2020b) and sense L2 elements (Attig et al. 2018), and multiple experimental studies have established SUGP2 as a strong interactor of MATR3 (Chi et al. 2018; Iradi et al. 2018), although the precise role of their interaction has yet to be determined. The asymmetry in their cross-

prediction scores (**Supplemental Table S11E**), with SUGP2 performing noticeably stronger on MATR3's contexts (AUPR = 91.58%) than vice versa (MATR3 on SUGP2's contexts, AUPR = 78.34%), and even higher than on its own contexts (AUPR = 82.27%), suggests that MATR3's contexts contain a subset of SUGP2's contexts that have stronger saliency, while SUGP2's contexts as a whole contain more variability. This in turn may suggest a relationship of dependence in which at least some of MATR3's binding and functions are not only enhanced but perhaps even contingent on the presence of SUGP2, while SUGP2's regulatory activities may be more widespread and independent of MATR3. Interestingly, a similar degree of co-occurrence and overlap was observed for their respective motifs (**Supplemental Table S11E**), suggesting that this interaction has an equal propensity to occur both in cooperative modes corresponding to Path I and in cooperative and/or competitive modes corresponding to Path II (**Figure 5B**; **Supplemental Table S10**).

In short, we demonstrate with these positive controls that our proposed method successfully identifies known RBP-RBP interactions. These examples also demonstrate the combinatorics of RBP-RBP interactions (Khoroshkin et al. 2024), with certain RBPs (e.g., PCBP2, PTBP1, MATR3) being involved in multiple different interaction pairs, thus pointing to the exponential scale at which the full range of human RBPs interact. In total, we discovered 796 interactions with our method, which accounts for approximately 32% of the total 2,485 possible RBP-RBP pairs in our collection (**Supplemental Table S10**; **Materials and Methods**). As further validation, we cross-referenced our method with known, documented interactions in the BioGrid database (Oughtred et al. 2021) (**Materials and Methods**). Of the 436 documented interacting RBP-RBP pairs that were present both in our collection and in BioGrid, 145 of these interactions (33.36%) were detected by our method, demonstrating our method's ability and efficacy. We next sought to further investigate novel interacting pairs detected by our method and to hypothesize their potential functions and mechanisms.

NMD factor UPF1's strong association with AKAP1 suggests its surveillant role in local translation at organelle membranes

We noted that AKAP1 and UPF1, despite visually appearing to have dissimilar motifs and nucleotide preferences, occurred in the same cross-prediction cluster (**Figure 4C**), which may be suggestive of a potential interaction. The relatively large divergence between their target motif logos (**Figure 5C**, "Motif KLD"), with AKAP1's being predominantly C-rich (consensus: CUCCC) and UPF1's being predominantly G-rich (consensus: GGGGG), indicated dissimilarity between the binding motifs of the 2 RBPs. Conversely, the binding contexts of the 2 RBPs exhibited some degree of similarity: AKAP1's context logo appears C-rich but has G-rich undertones, while UPF1's context logo is GC-rich. This context similarity is especially reflected in the fact that the context KLD value increased only slightly compared to the motif KLD value (from 31.36 for motif to 33.32 for context for AKAP1 vs. UPF1, and from 33.64 for motif to 35.59 for context for UPF1 vs. AKAP1; **Figure 5C**, "Context KLD"), despite being a cumulative sum across a considerably larger region of 55 nt compared to the target length of 5 nt. Their relatively high and reciprocated cross-prediction scores (**Figure 5C**, "Cross-prediction AUPR"), as well as their membership in the same cross-prediction cluster (**Figure 4C**), indicate that the 2 RBPs likely share similar global binding patterns. Finally, as expected from their highly divergent motif PFMs, AKAP1's (N = 3775) and UPF1's (N = 6225) motif instances shared very little overlap (N = 14; **Figure 5C**, "Percent of Motif Instance Overlaps"); however, their respective co-occurrence rates were relatively high, with 30.65% of AKAP1's overlapping UPF1's and 18.57% of UPF1's overlapping AKAP1's (**Figure 5C**, "Percent of Context Overlaps"). Taken together, these features correspond with our conditions for cooperative interactions along Path III, and we therefore hypothesized that AKAP1 and UPF1 may work together in a synergistic manner to regulate common targets.

Although an interaction between AKAP1 and UPF1 has briefly been reported (Gabrovsek et al. 2020), the underlying mechanism has not yet been resolved. We sought to understand how this RBP pair may cooperate. AKAP1 – a member of the A-kinase anchoring protein (AKAP) family best known for its localization of protein kinase A (PKA) to different cellular compartments to facilitate signaling pathway propagation (Turnham and Scott 2016) – spatially regulates (with the help of LARP4 and PABPC1) the local translation of essential mitochondrial

proteins at the mitochondrial outer membrane (Gabrovsek et al. 2020; Sharma and Fazal 2024). AKAP1 has also been shown to bind ribonucleoprotein (RNP) complexes with roles in translation and mRNA decay through its KH and Tudor domains (Gabrovsek et al. 2020). On the other hand, Up-frameshift 1 (UPF1) is an RNA helicase with numerous functions in mRNA decay pathways, the most prominent of which is nonsense-mediated decay (NMD), for which it is an essential RBP (Kim and Maquat 2019). When a premature termination codon (PTC) in an erroneously transcribed RNA is detected by the ribosome, the mRNA is targeted for degradation when UPF1 outcompetes PABPC1 (a translation-promoting factor) for binding to the eRF1-eRF3 complex (Qi et al. 2022); this in turn prevents efficient translation termination from occurring, and an activated UPF1 triggers the NMD pathway upon its phosphorylation by the kinase SMG1 (Kim and Maquat 2019).

We investigated the genomic regions containing overlapping AKAP1-UPF1 contexts, noting that the vast majority (93.3%) occurred in 3'UTRs, which is known to have roles in mRNA translation regulation (Mayr 2019). As expected, a number of these overlapping contexts occurred in genes encoding important mitochondrial proteins, such as *SDHAF1* (a component of succinate dehydrogenase, which is involved in mitochondrial metabolism), *SOD2* (oxidative stress reducer), *MRPL12/49* (two mitochondrial ribosomal proteins), *SLC25A1/10/42* (three solute carriers regulating small molecule transport across mitochondrial membranes), *TIMM13/50* and *TOMM34* (three translocases regulating protein import into the mitochondria), and numerous other genes encoding proteins that are either components of or localized to the inner or outer mitochondrial membrane (e.g., *ATAD3B*, *BCL2L1*, *MFN2*, *MUL1*, *RHOT2*, and *TMBIM6*) (Ginsberg et al. 2003; Gabrovsek et al. 2020; Rath et al. 2021; Sayers et al. 2022). Finally, we performed gene ontology (GO) analysis (Kolberg et al. 2023) on the gene set containing AKAP1-UPF1 context overlaps (**Figure 5D**; **Materials and Methods**). Despite the lack of enrichment for mitochondria-specific GO terms, the recurrence of significantly enriched localization (e.g., protein localization), membrane (e.g., endomembrane system organization), and transport (e.g., vesicle-mediated transport) terms led us to hypothesize that AKAP1's functions of local protein synthesis may also occur at components of the endomembrane. Indeed, we found many genes in the set coded for proteins with roles in the endoplasmic reticulum (ER) membrane (e.g., *ABHD4*, *DHCR7*, *ESYT1*, *LPCAT3*, *MPDU1*, *NCLN*, *POR*, *REEP6*, *RRBP1*, *TRAM2*), Golgi membrane (e.g.,

CDIPT, *CHPF*, *FUT6*, *MYO18A*, *RFNG*), trans-Golgi network (e.g., *AP1M1*, *ATG9A*, *TGOLN2*), and ER-to-Golgi transport (e.g., *ACTR1A*, *GET4*, *RAB1B*, *TEX261*, *YKT6*) (Sayers et al. 2022).

In light of these findings, we propose a cooperative mechanism between AKAP1 and UPF1 in which they work jointly, by way of Path III (**Figure 5B**), to ensure accurate local translation of specialized mRNA transcripts into full-length protein products at certain organelles. Our results suggest that UPF1 maintains a sustained interaction with AKAP1, whereby UPF1 is on vigilant standby to intervene in circumstances of local translation where AKAP1 has anchored PTC-containing mRNA transcripts to organelle membranes. This could present a quality control mechanism for the accurate translation and efficient use of proteins that are essential to organelle function, especially at the mitochondria, ER, and Golgi apparatus, such that the resultant proteins become immediately available for insertion into the membrane or for import into the organelle. In summary, we infer novel subcellular localizations for AKAP1, and we propose a novel mechanism for the cooperative interaction between AKAP1 and UPF1.

The DEAD-box RNA helicase DDX59 facilitates the mRNA regulatory roles of HNRNPK and PCBP2

Interestingly, DDX59 demonstrated highly similar features with both HNRNPK and PCBP2 individually (**Figure 5E**), despite being a distinct RBP belonging to a large family of highly-conserved DEAD-box RNA helicases (Shamseldin et al. 2013; Linder and Jankowsky 2011). Of note, DDX59's model demonstrated superior performance on the contexts of HNRNPK (AUPR = 97.44%) and PCBP2 (AUPR = 94.75%) compared to its own contexts (AUPR = 92.44%), suggesting a very strong degree of similarity in the underlying patterns of their respective binding regions. DDX59 also exhibited a higher degree of motif co-occurrence and motif overlap with HNRNPK and PCBP2, compared to those observed between the known interactors HNRNPK and PCBP2 (**Figure 5E**), which provides additional support for a potential interaction with DDX59. Although DEAD-box helicases interface with the RNA backbone of their targets in a non-sequence-specific manner (Linder and Jankowsky 2011; Jarmoskaite and Russell 2011), we discovered a C-rich nucleotide binding preference (**Figure 3C**) for DDX59, which may be a byproduct of frequent associations with poly(rC) binding proteins like HNRNPK and

PCBP2. While the exact functions of DDX59 have not yet been deduced, its family members have annotated roles in ATP-dependent local RNA duplex unwinding, RNA clamping, and RNP assembly, stabilization, remodeling, and disassembly (Linder and Jankowsky 2011; Bohnsack et al. 2023). Based on our empirical evidence, we hypothesize that DDX59 interacts cooperatively with both HNRNPK and PCBP2 by any or all of the following possible mechanisms: (1) unwinding short regions of double-stranded RNA to enable binding by the single-stranded RNA-preferring KH domains of HNRNPK and PCBP2, (2) serving as an anchor point at which HNRNPK and PCBP2 can interact to form an hnRNP complex, (3) remodeling hnRNPs, and/or (4) disassembling hnRNPs. This hypothesis is further supported by the fact that HNRNPK has previously been shown to interact with a number of DEAD-box helicases to enact various regulatory processes (Chen et al. 2002; Li et al. 2019; Good et al. 2019; Chen et al. 2021).

To investigate the functional implications of this interaction, we first studied each pair of RBPs in this trio independently. We performed intersections between the respective contexts of the RBPs in a pairwise manner, making sure to exclude any resultant intersections that also overlapped the third RBP (e.g., context overlaps between HNRNPK and PCBP2 that do not overlap with DDX59 contexts; **Materials and Methods**). Next, we performed GO analysis on each set of intersections to determine the gene functions that are uniquely regulated by each pair (**Materials and Methods**). As expected, HNRNPK and PCBP2, as hnRNPs with a large breadth of mRNA regulatory functions (Kim et al. 2000; Lang et al. 2021), were enriched for the largest number of unique GO terms (N = 88), which were also very diverse in nature; while some of these terms pertained to more specialized processes (e.g., axonogenesis, miRNA metabolic process, postsynapse organization, tricuspid valve development) or certain signaling pathways (e.g., serine/threonine kinase, Rho GTP-ase, SMAD, semaphorin-plexin), others were more general terms related to cell, tissue, organ, and organism development. HNRNPK and DDX59 were enriched for 12 unique GO terms, which included dendrite morphogenesis, dendritic spine development, and dendritic spine morphogenesis terms, suggesting that these 2 RBPs may have specific, coordinated roles in dendrite development. On the other hand, PCBP2 and DDX59 were enriched for only 3 unique GO terms, 2 of which were related to RNA Pol II transcription and the third of which was "regulation of phosphatidylinositol 3-kinase/protein kinase B

signal transduction", which may suggest that they have coordinated roles in the PI3K/AKT pathway. Previous studies have demonstrated associations for HNRNPK in axon (Liu et al. 2008; Liu and Szaro 2011) and dendrite (Proepper et al. 2011; Folci et al. 2014; Leal et al. 2017) development and for PCBP2 in the PI3K/AKT signaling pathway (Qin et al. 2022; Yang et al. 2023) in various organisms. Our pairwise findings implicate novel roles for DDX59 in regulating axon and dendrite development in cooperation with HNRNPK and PI3K/AKT signaling in cooperation with PCBP2 in human.

Because each pair of RBPs within this trio exhibited strong features indicative of cooperative interactions, we leveraged transitive relations to hypothesize that DDX59, HNRNPK, and PCBP2 may also interact in unison to form a complex. To investigate this possibility, we performed a three-way intersection between their respective contexts (**Materials and Methods**) and found that 21.79% of DDX59's contexts overlapped with both HNRNPK's and PCBP2's, with most of these events located in introns or coding sequences. From GO analysis of the genes containing context overlaps between all three RBPs, we discovered enrichment for anatomical structure development and its child term anatomical structure morphogenesis (**Figure 5F**); further inspection of additional child terms for both revealed many related to the development of the face (face, head, & craniofacial suture morphogenesis), oral cavity (hard & soft palate morphogenesis, roof of mouth development, tooth eruption), and digits (limb joint & appendage morphogenesis) (Binns et al. 2009). Mutations in DDX59 have been associated with orofaciodigital syndrome, a disease which leads to varied disfigurements in features of the mouth, face, and digits (Shamseldin et al. 2013; Faily et al. 2017), which may help to explain the discovered GO terms. We additionally looked at terms that were uniquely enriched for genes in the three-way intersection but not enriched in any of the three pairwise GO analyses (N = 12). Along with adherens junction organization, cell-matrix/cell-substrate adhesion, and intracellular signal transduction terms, we found multiple terms related to signal transduction (intracellular signal transduction; chemical synaptic transmission, postsynaptic; regulation of postsynaptic membrane potential; excitatory postsynaptic potential). These latter GO terms reinforce our proposed role of DDX59 in the nervous system, and may suggest that DDX59 cooperates with both HNRNPK and PCBP2 to regulate signal transmission across the synapse.

In conclusion, we posit several hypotheses regarding cooperative interactions, by way of Path I (**Figure 5B**; **Supplemental Table S10**), involving DDX59, HNRNPK, and PCBP2. At the molecular level, we hypothesize novel and specific roles for DDX59 in regulating pre-mRNA splicing, mRNA transport, mRNA stability, and mRNA translation, via its strong association with either one or both of the hnRNPs possessing such functions. At the developmental level, we propose that DDX59, HNRNPK, and PCBP2 work together cooperatively – either in pairs or as a trio – to regulate various aspects of nervous system development and synaptic signaling. Finally, at the phenotypic level, we uncover a potential implication for HNRNPK and PCBP2's involvement in orofaciodigital syndromes, via their strong association with DDX59. It may be that mutations in DDX59 impair its protein-protein interactions with HNRNPK and/or PCBP2, thereby disrupting the joint regulation of shared target genes or signaling pathways that are critical to proper development of the face, oral cavity, and digits. In conclusion, utilizing combinatorics and transitive inference, we propose an interaction between DDX59, HNRNPK, and PCBP2 – either as individual pairs, a complex of three proteins, or both – that holds mechanistic, functional, and disease significance, thereby demonstrating the utility of our approach.

DISCUSSION

In this paper, we presented a novel, comprehensive computational pipeline to study the binding preferences, patterns, activities, functions, and interactions of RBPs. Inspired by state-of-the-art concepts from NLP, we developed a sequence representation strategy that enabled us to clearly delineate RBP binding motifs and their contexts, and we subsequently leveraged Multiple Instance Learning to predict RBP binding of these contexts in a weakly supervised fashion. We also developed an accurate, consensus-based motif discovery algorithm to identify the binding and context motifs for 71 RBPs in the HepG2 cell line. Importantly, we discovered novel *in vivo* contexts for these 71 RBPs, which is an important contribution to understanding the inherent nucleotide binding preferences exhibited by RBPs *in vivo* as well as their role in facilitating RBP binding specificity. With feature integration and transitive inference, we were able to leverage these findings and introduce a method for identifying known and novel RBP-RBP interactions, illustrating examples of each.

Our approach to RBP-RBP interaction discovery allowed us to identify many novel interacting pairs, and we were additionally able to hypothesize functional mechanisms pertaining to mRNA regulation for a subset of those pairs. In total, we characterized 2,485 RBP-RBP pairs (**Supplemental Table S10**; **Materials and Methods**). Based on our findings, the majority of detected interactions are likely cooperative, and this appears to align with previous work that has remarked on the biological and technical challenges involved in detecting competitive interactions (Lang et al. 2021). In addition, the mode of interaction between RBPs (Street et al. 2024) greatly influences its ability to be detected by our method, which for the time being addresses proximal, RNA-mediated interactions between pairs of RBPs. Long-range cooperative interactions in which RBPs bind distally from one another and are subsequently brought into closer proximity by virtue of RNA secondary structures may be more difficult to detect (Dassi 2017), as are long-range competitive interactions in which RBPs do not directly contend for the same binding site but instead bind distally from one another to promote divergent outcomes, such as in the proposed regulatory mechanism of alternative splicing in muscle differentiation between PTBP1 and QKI (Hall et al. 2013) (**Figure 5A**). For other binding modes, such as the cooperative and competitive modes in Path II (**Figure 5B**), we do not yet have the resolution to distinguish whether or not the feature set indicates a cooperative or competitive interaction, for which other types of data will be required. Though not explored in detail here, BCLAF1 and TRA2A present a strong example for the cooperative and/or competitive modes of Path II (**Supplemental Table S10**), and it would be of great interest to understand by which mode and for which functions they interact. We also do not consider homomeric interactions here. We additionally note that, due to its reliance on the percentage of motif instance and context overlaps, our approach is more attuned to detect RBP-RBP interactions that occur more often, which in turn implies that these interactions have more widespread regulatory roles on a larger set of mRNA targets. Indeed, the case studies we highlighted here (**Figure 5**) are those with very prominent features. On the other hand, less frequently occurring interactions that may govern the regulation of a small, specialized subset of mRNAs – which are still detected but have much lower overlap percentages – will require more detailed investigation in order to elucidate their precise functions.

By using the theory of transitive relation, we were also able to propose a novel *n*-way interaction ($n > 2$), as exemplified by our DDX59 case study. Importantly, this potential cooperativity between DDX59, HNRNPK, and PCBP2 has significant physiological and disease implications, thereby demonstrating the utility of our discovery. As additional support for this approach, we were also able to detect a putative interaction between RBPs recognizing G-rich motifs (namely, CSTF2T, FAM120A, GTF2F1, and PRPF4; **Supplemental Table S12A**), which has previously been identified (Lee et al. 2020) and therefore serves as a form of validation for our use of transitivity. We further propose that interactions may occur between other members of the 14 rG4-binding candidates isolated from our cross-prediction results (**Figure 4D**; **Supplemental Table S12B**); for example, the highly similar G-rich motif, preceded by a U and followed by a C, within a strongly G-rich context, belonging to ETUD2, FAM120A, GRSF1, and NKRF may be indicative of a potential interaction between the 4 RBPs. It is logical, therefore, to also extend this method in the future to infer larger macromolecular complexes. In fact, we were able to obtain preliminary results in this direction, identifying pairwise interactions between multiple components of the (pre-)catalytic spliceosome (**Supplemental Table S13**), with some RBPs being associated with either the U2, U5, or U4/U6 snRNPs, or the U4/U6 x U5 tri-snRNP complex (Sayers et al. 2022), and many having been captured together in existing Cryo-EM structures of the spliceosome (Szklarczyk et al. 2023). Two RBPs in this set, AQR and BUD13, have not, to the best of our knowledge, been reported in the literature, and we present them as a potential novel interacting pair within the spliceosome that is also supported by their discovered interactions with other spliceosomal proteins (**Supplemental Tables S10**, **S13**). We may even be able to obtain a finer resolution by using our feature integration and transitive inference to build up such complexes piecewise based on progressive degrees of similarity. To facilitate detections of these large-scale interactions, the cross-prediction and KLD dendrograms may prove helpful, as their inherent hierarchical nature may provide clues regarding the combinatorics, hierarchy, and order in which RBP-RBP interactions are formed. We also note that, while we used the KLD and cross-prediction clusters to guide our identification of candidate RBP pairs, this does not preclude the possibility that interactions also occur between RBPs from different clusters; such cases, which will also contribute to the discovery of larger complexes, require further examination. Overall, our results provide a promising advancement and outlook for charting the vast landscape of RBP-RBP interactions.

Finally, future avenues could entail analyses from numerous angles. Firstly, additional studies will be needed to investigate the proposed mechanisms of novel interactions discussed in this paper. The many interactions we identified but did not elaborate on here (**Supplemental Table S10**) will also require additional exploration to understand their underlying mechanisms, functions, and scale. We are additionally interested in extending our analyses to encompass mutual interactions, which we were able to preview in our AKAP1-UPF1 case study: the gene set containing context overlaps between the 2 RBPs contains the UPF1 gene itself, indicating that (1) UPF1 regulates the translation of its own cognate mRNA, a process referred to as an autogenous interaction, and (2) AKAP1 is involved in the regulation of UPF1's mRNA, a process referred to as a heterogeneous interaction (Dassi 2017). Solely within our collection of RBPs, our preliminary results reveal a strong prevalence of mutual interactions: ~97% (69 of the 71) are involved in any type of mutual interaction, while 100% of those detected mutual interactions demonstrated heterogeneous interactions and ~30% (N = 21) also demonstrating autogenous interactions. The full extent of the mutual interaction landscape will become clearer as these analyses are expanded to encompass the entire set of human RBPs, as will their functional consequences and implications. Other biological questions of interest include (1) to what extent RBP binding patterns and RBP-RBP interactions exhibit cell type specificity, (2) how RNA secondary structures contribute to enable specific RBP binding and distal RBP-RBP interactions, (3) how RBP binding preferences are influenced by cooperative and competitive interactions with microRNAs (Ciafrè and Galardi 2013; Kim et al. 2021), and (4) how RBP activities, functions, and interactions can be further elucidated via RBP regulatory network modeling. In conclusion, the method we have presented herein serves as a valuable contribution to the field's current knowledge of RBP binding patterns, context preferences, and interactions, all enabled by the advent of state-of-the-art machine learning, deep learning, and modeling techniques; we believe our work sets an important foundation for future studies in these areas.

MATERIALS AND METHODS

Sequence definition

We used the ENCODE REST API to download merged replicate eCLIP peaks in HepG2 from the ENCODE portal (Luo et al. 2020) (**Supplemental Table S1**). For each RBP, we assigned each of its peaks to a genomic region based on the classification scheme and priority order from (Van Nostrand et al. 2020a). We anchored each peak at its center and extended in both directions by 50 bp to achieve a uniform sequence length of 101 bp (**Figure 1C**). We also filtered out any sequences that remained unassigned by genomic region, contained non-canonical nucleotides, or originated from chromosomes other than chr1-22 and chrX. We defined the resulting sequences as our positive observations. To derive an analogous set of negative observations, we randomly sampled 101 bp sequences from matched genomic regions (to maintain representative nucleotide content) and matched chromosomes (to ensure balanced cross-validation groups), resampling as needed to ensure a disjoint set with respect to the positive observations (**Figure 1C**). We divided our entire dataset into 12 cross-validation groups, assigning each sequence to a group based on its chromosome of origin, as in the chromCV method (Moore et al. 2020). For each RBP, we designated the largest of the 12 groups as the test dataset, with the remaining 11 groups serving as the training dataset.

Sequence decomposition

We implemented a sliding window-based decomposition of all sequences in two stages. In the first stage, a sliding window of size k_{target} and step p_{target} was used to deconstruct a given sequence into a series of consecutive k-mers, each of which represents a potential target (**Figure 1A**); these series of k-mers, representing a full sequence, were used to train the baseline model (see below). In the second stage, another sliding window of size $k_{contexts}$ and step $p_{contexts}$ was used to isolate a specified number of k-mers $n_{contexts}$ in the left and right flanking regions of a given potential target to generate the left and right context, respectively, for that target (**Figure 1B**). The derived structure, comprising the left context, target, and right context, is collectively referred to as a context; these contexts were used to train the context model (see below). To assign each context an initial label, we implemented a label inheritance strategy that assigns the sequence label to all contexts derived from that sequence; (i.e., all contexts

derived from positive sequences were labeled positive and all contexts derived from negative sequences were labeled negative).

Model pre-processing, construction, training, and iterative relabeling pipeline (Figure 1D)

Data encoding and embedding

We trained a keras (Chollet and Others 2015) Tokenizer on the training dataset to compile a dictionary of *k*-mers present in our corpus. We then used the trained Tokenizer to encode the data as either a Bag-of-Words m x n matrix (such that m = number of observations and n = vocabulary size) or with an index-based encoding using the dictionary's word index (such that each constituent *k*-mer in the context would be represented by its index in the learned dictionary); in the latter case, the encoded form is fed as input to the Embedding layer of the neural networks. We indicate below, as needed, the encoding method used in each step of model training. We used the trained Tokenizer to apply the same operations to the test dataset.

Constructing the model architecture

We tested multiple different neural network architectures: namely, a shallow neural network containing a sole hidden layer with either a Bag-of-Words encoding (for the baseline model; see below) or an embedding layer (for the context model; see below); a convolutional neural network (CNN) with max pooling; and three forms of a recurrent neural network (RNN), each consisting of a convolutional layer, max pooling layer, and a different RNN architecture (long short-term memory [LSTM], bidirectional LSTM, and gated recurrent unit [GRU]). In preliminary testing stages of the project, we found that the more complex model architectures did not confer a significant advantage in overall performance compared to a simple neural network model. In line with our goals to devise an architecture-independent method for context discovery and analysis, we thus settled on using a simple neural network architecture for this study. Our neural networks contain an embedding layer (in the case of our context model, whose input is our encoded context data), a hidden layer with a ReLu activation function (whose size, or number of neurons, is determined by values from hyperparameter tuning [see below]), a dropout layer with a rate of

0.5 (to counter overfitting) and an output layer with a sigmoid activation function (to enable binary classification). We used the Adam algorithm for stochastic gradient descent and the binary cross-entropy loss function, which is suitable for binary classification tasks (Chollet and Others 2015).

Model training and hyperparameter tuning

For RBP binding prediction, we first trained a baseline model on the *k*-mer based sequence representation from the first stage of sequence decomposition (encoded as a Bag-of-Words matrix) to obtain a reference model. Next, we trained an initial contexts model on the contexts from the second stage of sequence decomposition (encoded as a series of dictionary indices). We performed hyperparameter tuning to attain baseline and context models for each RBP with the strongest performance. To minimize runtime, we opted to use the RandomizedSearchCV approach from scikit-learn (Pedregosa et al. 2011) as opposed to the GridSearchCV approach, which exhaustively tests each combination of hyperparameter values. We considered 4 hyperparameters influential to the neural network's performance, and we specified the following ranges for each hyperparameter based on preliminary exploratory analyses (not shown):

- Embedding dimension: [8, 16, 32, 64, 128]
- Number of neurons: [16, 32, 64, 128, 256, 512, 1024, 2048]
- Batch size: [32, 64, 128, 256, 512, 1024, 2048]
- Number of epochs: [2, 3, 4, 5, 6, 7, 8, 9, 10]

We performed 25 iterations of random hyperparameter value sampling, training, cross-validation, and testing. To select the optimal set of hyperparameters from the tuning results, we defined criteria that balanced a model's generalization gap (measured by the difference in training and test performance) with its overall performance (measured by the test performance). We set a small threshold (starting at 0.005) and identified the tuning iterations whose difference in training and test performance was less than this threshold; if no tuning iterations met this requirement, we progressively increased the threshold at small increments until at least one tuning iteration demonstrated a generalization gap below the threshold. Next, for the subset of iterations that met the generalization gap requirement, we selected the iteration that had the maximum test performance for the baseline and context

models, respectively. Using the optimized hyperparameter values from the selected iterations, we trained the final baseline and contexts models on the full training dataset.

Iterative relabeling

In our goal to identify the precise regions within a positive sequence that contribute to its binding event, we sought to remove weak contexts within these sequences that do not possess salient features that were identified by the model. To this end, we performed the following procedure. First, using the baseline model as our gold standard, we ran the set of contexts (encoded as a Bag-of-Words matrix) through the trained baseline model to get the baseline's predictions on the contexts, then used the resultant predictions to assign new labels to the contexts. Contexts originating from negative sequences were not relabeled, as they are not expected to contain any salient features and are not associated with a binding event. Contexts originating from positive sequences (i.e., having an initial label of 1) were relabeled based on whether their respective predicted probabilities were greater than 0.5 (relabel as 1) or below 0.5 (relabel as 0), a process which we refer to as relabeling.

We then performed, in an iterative manner, additional rounds of both retraining the contexts model and relabeling the contexts using the resultant predictions; we refer to this process as iterative relabeling. Note that iterative relabeling should be run until a given termination condition is met; for example, until convergence, until a pre-defined number of iterations is completed, or until the first occurrence of overfitting is detected. To minimize runtime, we ran iterative relabeling for a total of 3 iterations. To select the best-performing iteration of relabeling, we implemented a procedure akin to that for the selection of the best-performing iteration of hyperparameter tuning (see above), incrementally applying a generalization gap threshold to identify iterations with the smallest difference in training and test performance, then selecting from these the iteration with the highest test performance.

Evaluation of model performance

Because the data was imbalanced after relabeling (with respect to the relative number of positive and negative observations), we used the area under the precision-recall (AUPR) curve to evaluate the performance of our context predictions (Grau et al. 2015). We computed AUPRs at the instance- and bag-levels before and after relabeling (**Figure 1E**).

Computation of MIL witness rates

We computed the witness rates (i.e., number of positive contexts) (Carbonneau et al. 2018) per sequence before and after relabeling. To survey the global witness rate across all contexts belonging to a given RBP, we computed the proportion of all contexts that were assigned a positive label by our predictive model for that RBP. We then computed summary statistics, such as the median and average, to report global witness rates across all RBPs (**Supplemental Table S2**). Because our initial dataset was balanced (i.e., equal number of positive and negative observations), by definition the starting global witness rate for every RBP is 0.5. To inspect the witness rates at the local level, we computed the proportion of positively-predicted contexts per true positive sequence, performed an average across all sequences for a given RBP, and again calculated summary statistics. Because each positive sequence is initially composed of solely positive contexts due to inherited labeling, by definition the starting local witness rate for each such sequence is 1.0. We computed and compared the global and local witness rates at successive stages of training to examine the effects of iterative relabeling.

Motif discovery algorithm (Figure 2A)

The motif discovery algorithm consists of 6 phases, described below.

Phase I: Identification of candidate consensuses

From our post-relabeling context predictions, we designated a context as positive if its prediction was greater than 0.5, and negative otherwise. For each unique target *k*-mer, we computed its enrichment based on its frequency in the positively-predicted contexts relative to its frequency in the negatively-predicted contexts. We also used these

quantities to perform chi-square testing and obtain a *p*-value for each *k*-mer indicating the significance of its enrichment. We only considered all enriched *k*-mers (i.e., enrichment > 1) as candidate consensuses, disregarding all other non-enriched *k*-mers in the motif discovery process to reduce the motif search space.

Phase II: Identification of position-dependent similar k-mers

To create a preliminary partition for each candidate consensus, we generated one list for each nucleotide position of the consensus, where each list contained all other enriched *k*-mers that had the same nucleotide as the consensus at that position. We then merged these lists by taking the intersection of the inner and outer lists, respectively, then taking the union. This operation yielded a preliminary partition consisting of *k*-mers with at most 3 mismatches (i.e., a Hamming distance of 3) relative to the corresponding candidate consensus, thus modeling the degeneracy of RBP motifs. Note that a preliminary partition is constructed on the basis of similarity to the corresponding consensus. While we use the Hamming distance as a measure of similarity between partition *k*-mers and their consensus, we do not use it to construct the partition itself.

Phase III: Identification of motif instances with k-mer co-occurrence

To remove *k*-mers from the preliminary partition that are not likely to be true motif instances, we relied on each *k*-mer's co-occurrence with the corresponding candidate consensus. We defined the *k*-mer co-occurrence frequency as the frequency at which a particular partition *k*-mer occurred in the same sequence as its corresponding candidate consensus, then used a co-occurrence threshold to filter out any *k*-mers whose co-occurrence frequency was above this threshold. To determine the value of the co-occurrence threshold, we developed a minimization-based tuning procedure. We initialized the tuning procedure at a co-occurrence threshold of 0.5 and progressively increasing the threshold by small increments until a ceiling of 1; at each iteration, we computed the co-occurrence frequency of each *k*-mer in the partition, filtered out any *k*-mers whose frequency did not satisfy the threshold, constructed the resultant motif (see below) based on the remaining *k*-mers in the partition, and computed the Kullback-Leibler Divergence (KLD) between the motif PFMs from the successive iterations. We continued incrementing the threshold as long as the successive KLDs continued to decrease, and stopped tuning if the KLD began to increase

(i.e., we minimized the difference between probability distributions of motifs at successive co-occurrence thresholds). We also performed tuning in the opposite direction by decreasing the threshold from 0.5 in small increments (until a floor of 0) and repeating the minimization procedure. To arrive at the final co-occurrence threshold, we selected the motif corresponding to the smallest difference in KLD between the two tuning directions.

Phase IV: Motif construction

For all *k*-mers in a candidate consensus's final partition, we collated all of the *k*-mers' instances where they occurred as targets of positively-predicted contexts. If more than one *k*-mer occurred as the targets of contexts originating from the same sequence, we retained only one *k*-mer using the following conditions: (1) if the identities of the *k*-mers were different, we retained that with the higher enrichment; (2) if the identities of the *k*-mers were the same, we retained that with the higher prediction score. We then aligned the compiled instances to generate the corresponding motif for that consensus. We visualized the motif logos using WebLogo (Crooks et al. 2004) and Logomaker (Tareen and Kinney 2020).

Phase V: Motif scoring and primary motif selection

For each RBP, the algorithm discovers all possible motifs within the dataset that satisfy the algorithm's conditions; the number of discovered motifs is different for each RBP. To select a primary motif for each RBP, we developed an iterative sorting strategy relying on *k*-mer and motif metrics to rank all discovered motifs. Briefly, we first sorted motifs by the enrichment of the consensus *k*-mer, selecting the top 10 for further consideration. We then sorted these 10 motifs by the significance of their consensus's enrichment, selecting the top 5. We next sorted these 5 motifs by their weighted relative entropy (Tompa 1999), selecting the top 2. From these 2, we designated the motif whose candidate consensus had the highest enrichment as the primary motif for the RBP, which we used to perform all subsequent analyses.

Phase VI: Context discovery

To extract the contexts of our discovered motifs, we extracted the genomic coordinates of the left and right flanking regions surrounding each motif instance, then aligned the resulting 55-nt long regions. We visualized the logos for the contexts using WebLogo (Crooks et al. 2004) and Logomaker (Tareen and Kinney 2020).

Determination of motif and context nucleotide preference (Supplemental Table S6)

We determined the nucleotide preference of each RBP's motif and context by examining its PFM representation. Starting with the left context, we identified the most prevalently occurring nucleotide for every position, then ranked each of the four nucleotides in descending order based on the number of positions at which each occurred most prevalently. We also computed the ratio of the number of positions for each nucleotide relative to the maximum of these values. To account for cases in which RBPs demonstrate a clear preference for more than one nucleotide (e.g., HNRNPL), we deemed any nucleotide for which its position count was at least half that of the top-ranking nucleotide as also demonstrating a strong preference. We performed these operations for the left context, motif, and right context individually to obtain nucleotide preferences at a finer resolution, and finally performed it for the full context (excluding the motif) to determine the overall nucleotide preference of the context. We deemed a context symmetric if its preferred nucleotide(s) in the left and right contexts were the same. Similarly, we compared the preferred nucleotide(s) in the motif and the full context to determine whether they shared the same preferences.

Data preparation for discovery of RBP-RBP interactions

Because the largest chromCV group from each RBP was used as its test dataset, and because the distribution of binding sites across chromosomes was different for each RBP, the contexts in the test predictions belonged to different chromosomes, making it difficult to conduct comparisons across RBPs. This particularly posed a challenge when proceeding to the RBP-RBP interaction analyses, as two important features for discovery are the relative degrees of motif instance overlap and motif instance co-occurrence (see below) which, by definition, must be detected in the same loci. To resolve this challenge, for each RBP we fit 11 additional new and independent models,

such that for each model, one of the other 11 chromCV groups served as the test dataset and the remaining data served as the training dataset. For each of these additional models, we used the same hyperparameters that were used to train the original model for the RBP. Upon completion, we pooled the predictions across all 12 total models to arrive at a set of RBP binding predictions with coverage across all chromosomes, which enabled our RBP-RBP interaction discovery.

Calculation of motif instance overlap and motif instance co-occurrence (Figure 5B)

We used Bedtools (Quinlan and Hall 2010) to perform a strand-specific intersection of motif instances by genomic coordinate. We considered an intersection between any two motif instances – each belonging to an RBP within a given pair – by at least one nucleotide as a motif instance overlap. We performed the same operation for the contexts. We note that the resulting set of overlapping contexts contains two distinct subsets: one subset in which the corresponding motif instances themselves also overlap and the other subset in which the corresponding motif instances do not overlap. To obtain the motif co-occurrences, we performed a set difference between the full set of overlapping contexts and the set of overlapping motif instances to arrive at the set of overlapping contexts whose constituent motif instances do not overlap but instead co-occur. We divided each quantity by the total number of motif instances to obtain the corresponding percentages. We applied this procedure in a pairwise fashion to account for all possible RBP-RBP pairs.

Application of the RBP-RBP interaction discovery method

We implemented the decision tree in **Figure 5B** programmatically. To achieve this, we defined measures for each of the 4 characteristics in the decision tree: motif similarity, context similarity, motif instance co-occurrence, and motif instance overlap. Note that the order of operations of the conditions displayed in the decision tree in **Figure 5B** is not fixed but simply represents one possible implementation.

As a measure for motif similarity, we considered any two RBPs to have similar motifs if they belong to the same motif KLD cluster (**Supplemental Fig. S1**).

To define a measure of context similarity, we first considered two RBPs to have similar contexts if at least one partner had a cross-prediction score greater than 0.8, as an AUPR of this magnitude would suggest that the first RBP recognizes the majority of salient features in the contexts of the second RBP. We discovered 507 interactions, which accounted for 20.40% of the total 2,485 RBP-RBP pairs. In addition, we cross-referenced our discovered interactions with those in the BioGrid database (Oughtred et al. 2021) by downloading the BIOGRID-ALL-4.4.240.tab3.zip release and filtering out (1) non-human interaction data and (2) any interaction data for RBP-RBP pairs not present in our collection. Of the 2,485 possible pairs in our collection, 436 were documented as interacting in the BioGrid database, and we discovered 91 (20.87%) of these documented pairs. We note that while this database serves as a useful reference, it is not a comprehensive compendium of RBP-RBP interactions for two reasons: first, upon our survey of the RBP-RBP interactions within the database, most of these interactions were identified by experimental assays that detect physical interactions, and second, upon our survey of the literature, we identified reports of many RBP-RBP interactions that are not contained in BioGrid. However, cross-referencing our results with BioGrid still provides a method to approximate true positive discovery rates.

Upon examining our results, however, we noticed that a number of known interactions, including AKAP1-LARP4 (Gabrovsek et al. 2020) and EFTUD2-PRPF8 (Malinová et al. 2017), were not detected, along with other RBP-RBP pairs that we believe demonstrated strong signs of interaction but whose cross-prediction performance did not meet the threshold requirements (e.g., LARP4-UPF1, AQR-BUD13). Upon closer inspection, we observed that many of these pairs demonstrated a notable asymmetry, in which the first partner predicted relatively strongly (but below 0.8) on the second partner, but the second partner had a considerably lower score on the first partner. Despite this asymmetry, the presence of a higher score in one direction still indicates the presence of salient features in the second partner's contexts that are significant and similar to the salient features learned by the model of the first partner, and this pair could still be considered a potential interaction. We also took into consideration the importance of dataset size – in some cases, a smaller dataset for an RBP can affect model training such that the most significant salient features are not captured or learned by the model at a high resolution, which can lead to a

lower score not only on its own contexts, but also on the contexts of other RBPs. Based on these facts, we concluded that a cross-prediction threshold of 0.8 was too stringent for interaction discovery, and we thus decided to lower the threshold in order to capture such cases in which there is a disparity in cross-prediction values. To address this, we tested different thresholds, cross-referencing the detected interactions with BioGrid for each test.

After experimentation, we defined two RBPs to have context similarity if at least one partner in the pair has a cross-prediction score on the second partner's contexts that is greater than 0.65. To ensure that the second partner's cross-prediction score on the first partner's contexts is not extremely low, we include an additional condition that its cross-prediction score must not be less than 5% of the 0.65 threshold. This definition resulted in the discovery of 796 interactions (32.03% of the total 2,485 RBP-RBP pairs in this study), with 145 of these interactions (33.36% of the 436) corresponding with the physical interactions in the BioGrid database, an increase of 85 additional interactions compared to the previous threshold at 0.8. Since our method considers only proximal interactions, we can infer that the 145 true positive interactions from BioGrid that were detected are proximal interactions. At these conditions, we also detected 289 more interactions overall than at the previous threshold of 0.8; although the cross-prediction scores for these 289 are lower, they still meet the other interaction conditions, which supports their discovery. Overall, these findings serve as a good validation for our interaction discovery method. Additionally, the increase in validated interactions from 91 at a threshold of 0.8 to 145 at a threshold of 0.65 suggests that higher cross-prediction thresholds can be too stringent; while the use of the lower 0.65 threshold may seem too relaxed, the detection of more reported physical interactions supports its validity. Furthermore, these findings demonstrate that an RBP-RBP pair does not have to exhibit very strong context similarity (close to 0.8 or higher) in order for the RBPs to participate in interactions with one another, provided that the other conditions of our method are also satisfied.

To determine the mode of interaction for the 796 pairs, we used inequality relationships between motif instance co-occurrence and motif instance overlap to facilitate classification of pairs into different paths (**Figure 5B**). For RBP pairs with a similar motif and similar context, we assigned them to Path I (cooperative) if their motif instance co-

occurrence was greater than their motif instance overlap, and to Path II (cooperative and/or competitive) if their motif instance overlap was greater than their motif instance co-occurrence. Similarly, for RBP pairs with a dissimilar motif but a similar context, we assigned them to Path III (cooperative) if their motif instance co-occurrence was greater than their motif instance overlap, and to Path IV (cooperative and/or competitive) if their motif instance overlap was greater than their motif instance co-occurrence. Note that these classifications indicate the path (and interaction binding mode) that an RBP-RBP is most likely to follow; in other words, the interaction mode(s) that the RBP-RBP pair demonstrates the highest propensity for; however, it does not preclude the possibility that the RBP-RBP pair may interact via another mode if both their motif instance overlap and co-occurrence are non-zero, regardless of which is greater. Classifications for all 2,485 RBP-RBP pairs are available in **Supplemental Table S10**.

Context intersections for DDX59, HNRNPK, and PCBP2

To identify distinct interactions between contexts belonging to any pair of RBPs within this trio, we first performed Bedtools intersect (Quinlan and Hall 2010) between the contexts of two RBPs (e.g., HNRNPK and PCBP2), then performed Bedtools intersect again, this time with the contexts of the third RBP (e.g., DDX59) using the -v option; this produced the set of contexts from HNRNPK and PCBP2 that overlapped with each other but that simultaneously did not overlap with DDX59's contexts. We repeated this procedure for the other two pairs (HNRNPK and DDX59; PCBP2 and DDX59).

To identify all overlaps between contexts belonging to each of the three RBPs, we sought to account for all possible combinations of orderings (with respect to genomic coordinates) in which these overlaps could potentially occur. For example, three contexts – one from each RBP – may all overlap one another simultaneously. In other cases, the three contexts may sequentially (with respect to genomic coordinates) overlap each other in a staggered manner, such that there is continuity between all three, but they do not all overlap with one other; the order in which the three can overlap in such a manner is combinatorial. We first performed intersections with Bedtools (Quinlan and Hall 2010) between two RBPs (e.g., HNRNPK and PCBP2) to identify their set of overlapping contexts. Then, we intersected the third RBP's contexts (e.g., DDX59) independently with the subset of HNRNPK's contexts that

overlapped PCBP2's, and with the subset of PCBP2's contexts that overlapped HNRNPK's; note that both of these operations are required to capture all possible interactions with DDX59. We merged the output that resulted from these last two intersections and filtered out duplicates to arrive at the subset of DDX59's contexts that overlap both HNRNPK's and PCBP2's contexts. Finally, to account for the combinatorics of the possible order of overlaps, we repeated this multi-step intersection procedure two more times: once where we first intersected HNRNPK's and DDX59's contexts, then with PCBP2's contexts; and once where we first intersected PCBP2's and DDX59's contexts, then with HNRNPK's contexts. Note that all intersections were performed in a strand-specific manner.

Gene Ontology (GO) analysis

In accordance with the pipeline used to process eCLIP data at the ENCODE portal (Luo et al. 2020), we used the GENCODE v29 annotation (Harrow et al. 2012) for all gene-level analyses. For a given RBP-RBP pair, we used Bedtools (Quinlan and Hall 2010) to intersect the genomic coordinates of their motif instances and co-occurrences with the GENCODE annotation to determine in which genes they were located; the resulting unique set of genes served as our test set for GO analysis. To derive a background set, we used all genes that were expressed in an ENCODE poly(A) RNA-seq data set from the same HepG2 cell line (experiment accession: ENCSR000CPE; file accession: ENCFF168QKW; TPM threshold = 1). We used the gProfiler package (Kolberg et al. 2023) to perform GO analysis of biological processes, using FDR < 0.05 as a cutoff to identify significantly enriched GO terms.

DATA AVAILABILITY

Data is publicly available at the ENCODE portal. Accession for eCLIP data used in this study are detailed in **Supplemental Table S1**.

CODE AVAILABILITY

Code is publicly available at https://github.com/orbitalse/novel_RBP_discovery_methods.

AUTHOR CONTRIBUTIONS

S.I.E.: Problem Formulation, Conceptualization, Data Curation, Data Processing, Investigation, Formal Analysis, Methodology, Visualization, Writing of Manuscript (original draft), Review of Manuscript, Editing of Manuscript. Z.W.: Research Area, Discussions, Review of Manuscript. All authors reviewed and approved the final version of the manuscript.

ACKNOWLEDGEMENTS

We thank Dr. Athma A. Pai for critical reading of and feedback on the manuscript.

FUNDING

National Institutes of Health grant U24HG012343 to Z.W.

CONFLICT OF INTEREST

Zhiping Weng co-founded Rgenta Therapeutics, and she serves as a scientific advisor for the company and is a member of its board.

REFERENCES

Alipanahi B, Delong A, Weirauch MT, Frey BJ. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* **33**: 831–838.

Attig J, Agostini F, Gooding C, Chakrabarti AM, Singh A, Haberman N, Zagalak JA, Emmett W, Smith CWJ, Luscombe NM, et al. 2018. Heteromeric RNP assembly at LINEs controls lineage-specific RNA processing. *Cell* **174**: 1067–1081.e17.

Bailey TL, Johnson J, Grant CE, Noble WS. 2015. The MEME suite. *Nucleic Acids Res* **43**: W39–49.

Begg BE, Jens M, Wang PY, Minor CM, Burge CB. 2020. Concentration-dependent splicing is enabled by Rbfox motifs of intermediate affinity. *Nat Struct Mol Biol* **27**: 901–912.

Bentley DL. 2014. Coupling mRNA processing with transcription in time and space. *Nat Rev Genet* **15**: 163–175.

Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. 2009. QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**: 3045–3046.

Bishop CM, Nasrabadi NM. 2006. *Pattern recognition and machine learning*. Springer.

Bohnsack KE, Yi S, Venus S, Jankowsky E, Bohnsack MT. 2023. Cellular functions of eukaryotic RNA helicases and their links to human diseases. *Nat Rev Mol Cell Biol* **24**: 749–769.

Bourdon S, Herviou P, Dumas L, Destefanis E, Zen A, Cammas A, Millevoi S, Dassi E. 2023. QUADRatlas: the RNA G-quadruplex and RG4-binding proteins database. *Nucleic Acids Res* **51**: D240–D247.

Briata P, Bordo D, Puppo M, Gorlero F, Rossi M, Perrone-Bizzozero N, Gherzi R. 2016. Diverse roles of the nucleic acid-binding protein KHSRP in cell differentiation and disease. *Wiley Interdiscip Rev RNA* **7**: 227–240.

Busa VF, Favorov AV, Fertig EJ, Leung AKL. 2021. Spatial correlation statistics enable transcriptome-wide characterization of RNA structure binding. *Cell Rep Methods* **1**: 100088.

Carbonneau M-A, Cheplygina V, Granger E, Gagnon G. 2018. Multiple instance learning: A survey of problem characteristics and applications. *Pattern Recognit* **77**: 329–353.

Castello A, Fischer B, Hentze MW, Preiss T. 2013. RNA-binding proteins in Mendelian disease. *Trends Genet* **29**: 318–327.

Chen C-C, Yang J-H, Fu S-L, Lin W-J, Lin C-H. 2021. Arginine methylation of hnRNPK inhibits the DDX3-hnRNPK interaction to play an anti-apoptosis role in osteosarcoma cells. *Int J Mol Sci* **22**: 9764.

Chen H-C, Lin W-C, Tsay Y-G, Lee S-C, Chang C-J. 2002. An RNA helicase, DDX1, interacting with poly(A) RNA and heterogeneous nuclear ribonucleoprotein K. *J Biol Chem* **277**: 40403–40409.

Chi B, O'Connell JD, Yamazaki T, Gangopadhyay J, Gygi SP, Reed R. 2018. Interactome analyses revealed that the U1 snRNP machinery overlaps extensively with the RNAP II machinery and contains multiple ALS/SMA-

causative proteins. *Sci Rep* **8**. http://dx.doi.org/10.1038/s41598-018-27136-3.

Choi HS, Hwang CK, Song KY, Law P-Y, Wei L-N, Loh HH. 2009. Poly(C)-binding proteins as transcriptional regulators of gene expression. *Biochem Biophys Res Commun* **380**: 431–436.

Chollet F, Others. 2015. Keras. https://keras.io.

Ciafrè SA, Galardi S. 2013. microRNAs and RNA-binding proteins: a complex network of interactions and reciprocal regulations in cancer. *RNA Biol* **10**: 935–942.

Coelho MB, Attig J, Bellora N, König J, Hallegger M, Kayikci M, Eyras E, Ule J, Smith CWJ. 2015. Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. *EMBO J* **34**: 653–668.

Conlon EG, Manley JL. 2017. RNA-binding proteins in neurodegeneration: mechanisms in aggregate. *Genes Dev* **31**: 1509–1528.

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res* **39**: D301–8.

Cornish-Bowden A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**: 3021–3030.

Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res* **14**: 1188–1190.

Darmon SK, Lutz CS. 2012. Novel upstream and downstream sequence elements contribute to polyadenylation efficiency. *RNA Biol* **9**: 1255–1265.

Dassi E. 2017. Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins. *Front Mol Biosci* **4**: 67.

D'haeseleer P. 2006a. How does DNA sequence motif discovery work? *Nat Biotechnol* **24**: 959–961.

D'haeseleer P. 2006b. What are DNA sequence motifs? *Nat Biotechnol* **24**: 423–425.

Dietterich TG, Lathrop RH, Lozano-Pérez T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell* **89**: 31–71.

Di Giammartino DC, Nishida K, Manley JL. 2011. Mechanisms and consequences of alternative polyadenylation. *Mol Cell* **43**: 853–866.

Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA, et al. 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Mol Cell* **70**: 854–867.e9.

Dumas L, Herviou P, Dassi E, Cammas A, Millevoi S. 2021. G-quadruplexes in RNA biology: Recent advances and future directions. *Trends Biochem Sci* **46**: 270–283.

Durbin R, Eddy SR, Krogh A, Mitchison G. 1998. *Biological sequence analysis*. Cambridge University Press, Cambridge, England.

Faily S, Perveen R, Urquhart J, Chandler K, Clayton-Smith J. 2017. Confirmation that mutations in DDX59 cause an autosomal recessive form of oral-facial-digital syndrome: Further delineation of the DDX59 phenotype in two new families. *Eur J Med Genet* **60**: 527–532.

Feng H, Bao S, Rahman MA, Weyn-Vanhentenryck SM, Khan A, Wong J, Shah A, Flynn ED, Krainer AR, Zhang C. 2019. Modeling RNA-binding protein specificity in vivo by precisely registering protein-RNA crosslink sites. *Mol Cell* **74**: 1189–1204.e6.

Folci A, Mapelli L, Sassone J, Prestori F, D'Angelo E, Bassani S, Passafaro M. 2014. Loss of hnRNP K impairs synaptic plasticity in hippocampal neurons. *J Neurosci* **34**: 9088–9095.

Foulds J, Frank E. 2010. A review of multi-instance learning assumptions. *Knowl Eng Rev* **25**: 1–25.

Gabrovsek L, Collins KB, Aggarwal S, Saunders LM, Lau H-T, Suh D, Sancak Y, Trapnell C, Ong S-E, Smith FD,

et al. 2020. A-kinase-anchoring protein 1 (dAKAP1)-based signaling complexes coordinate local protein synthesis at the mitochondrial surface. *J Biol Chem* **295**: 10749–10765.

García-Mayoral MF, Díaz-Moreno I, Hollingworth D, Ramos A. 2008. The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. *Nucleic Acids Res* **36**: 5290–5296.

Gau B-H, Chen T-M, Shih Y-HJ, Sun HS. 2011. FUBP3 interacts with FGF9 3' microsatellite and positively regulates FGF9 translation. *Nucleic Acids Res* **39**: 3582–3593.

Gebauer F, Schwarzl T, Valcárcel J, Hentze MW. 2021. RNA-binding proteins in human genetic disease. *Nat Rev Genet* **22**: 185–198.

Gehring NH, Wahle E, Fischer U. 2017. Deciphering the mRNP code: RNA-bound determinants of post-transcriptional gene regulation. *Trends Biochem Sci* **42**: 369–382.

Gerstberger S, Hafner M, Tuschl T. 2014. A census of human RNA-binding proteins. *Nat Rev Genet* **15**: 829–845.

Ghanbari M, Ohler U. 2020. Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res* **30**: 214–226.

Ginsberg MD, Feliciello A, Jones JK, Avvedimento EV, Gottesman ME. 2003. PKA-dependent binding of mRNA to the mitochondrial AKAP121 protein. *J Mol Biol* **327**: 885–897.

Good AL, Haemmerle MW, Oguh AU, Doliba NM, Stoffers DA. 2019. Metabolic stress activates an ERK/hnRNPK/DDX3X pathway in pancreatic β cells. *Mol Metab* **26**: 45–56.

Graf R, Munschauer M, Mastrobuoni G, Mayr F, Heinemann U, Kempa S, Rajewsky N, Landthaler M. 2013. Identification of LIN28B-bound mRNAs reveals features of target recognition and regulation. *RNA Biol* **10**: 1146–1159.

Grau J, Grosse I, Keilwagen J. 2015. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**: 2595–2597.

Hall MP, Nagel RJ, Fagg WS, Shiue L, Cline MS, Perriman RJ, Donohue JP, Ares M Jr. 2013. Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation. *RNA* **19**: 627–638.

Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, et al. 2012. GENCODE: The reference human genome annotation for The ENCODE Project. *Genome Res* **22**: 1760–1774.

Herdy B, Mayer C, Varshney D, Marsico G, Murat P, Taylor C, D'Santos C, Tannahill D, Balasubramanian S. 2018. Analysis of NRAS RNA G-quadruplex binding proteins reveals DDX3X as a novel interactor of cellular G-quadruplex containing transcripts. *Nucleic Acids Res* **46**: 11592–11604.

Herrera F, Ventura S, Bello R, Cornelis C, Zafra A, Sanchez-Tarrago D, Vluymans S. 2016. *Multiple instance learning*. 1st ed. Springer International Publishing, Cham, Switzerland.

He X, Yuan J, Wang Y. 2021. G3BP1 binds to guanine quadruplexes in mRNAs to modulate their stabilities. *Nucleic Acids Res* **49**: 11323–11336.

Huang L, Li G, Du C, Jia Y, Yang J, Fan W, Xu Y-Z, Cheng H, Zhou Y. 2023. The polyA tail facilitates splicing of last introns with weak 3' splice sites via PABPN1. *EMBO Rep* **24**: e57128.

Hwang H, Jeon H, Yeo N, Baek D. 2024. Big data and deep learning for RNA biology. *Exp Mol Med* **56**: 1293–1321.

Iradi MCG, Triplett JC, Thomas JD, Davila R, Crown AM, Brown H, Lewis J, Swanson MS, Xu G, Rodriguez-Lebron E, et al. 2018. Characterization of gene regulation and protein interaction networks for Matrin 3 encoding mutations linked to amyotrophic lateral sclerosis and myopathy. *Sci Rep* **8**. http://dx.doi.org/10.1038/s41598-018-21371-4.

Jarmoskaite I, Russell R. 2011. DEAD-box proteins as RNA helicases and chaperones. *Wiley Interdiscip Rev RNA* **2**: 135–152.

Khoroshkin M, Buyan A, Dodel M, Navickas A, Yu J, Trejo F, Doty A, Baratam R, Zhou S, Lee SB, et al. 2024.

Systematic identification of post-transcriptional regulatory modules. *Nat Commun* **15**: 7872.

Kim JH, Hahm B, Kim YK, Choi M, Jang SK. 2000. Protein-protein interaction among hnRNPs shuttling between nucleus and cytoplasm. *J Mol Biol* **298**: 395–405.

Kim S, Kim S, Chang HR, Kim D, Park J, Son N, Park J, Yoon M, Chae G, Kim Y-K, et al. 2021. The regulatory impact of RNA-binding proteins on microRNA targeting. *Nat Commun* **12**: 5057.

Kim YK, Maquat LE. 2019. UPFront and center in RNA decay: UPF1 in nonsense-mediated mRNA decay and beyond. *RNA* **25**: 407–422.

Kolberg L, Raudvere U, Kuzmin I, Adler P, Vilo J, Peterson H. 2023. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic acids research* **51**: W207–W212.

Lang B, Yang J-S, Garriga-Canut M, Speroni S, Aschern M, Gili M, Hoffmann T, Tartaglia GG, Maurer SP. 2021. Matrix-screening reveals a vast potential for direct protein-protein interactions among RNA binding proteins. *Nucleic Acids Res* **49**: 6702–6721.

Leal G, Comprido D, de Luca P, Morais E, Rodrigues L, Mele M, Santos AR, Costa RO, Pinto MJ, Patil S, et al. 2017. The RNA-binding protein hnRNP K mediates the effect of BDNF on dendritic mRNA metabolism and regulates synaptic NMDA receptors in hippocampal neurons. *eNeuro* **4**. http://dx.doi.org/10.1523/ENEURO.0268-17.2017.

Lee DSM, Ghanem LR, Barash Y. 2020. Integrative analysis reveals RNA G-quadruplexes in UTRs are selectively constrained and enriched for functional associations. *Nat Commun* **11**: 527.

Li J, Chen Y, Xu X, Jones J, Tiwari M, Ling J, Wang Y, Harismendy O, Sen GL. 2019. HNRNPK maintains epidermal progenitor function through transcription of proliferation genes and degrading differentiation promoting mRNAs. *Nat Commun* **10**: 4198.

Li K, Wu J-L, Qin B, Fan Z, Tang Q, Lu W, Zhang H, Xing F, Meng M, Zou S, et al. 2020. ILF3 is a substrate of SPOP for regulating serine biosynthesis in colorectal cancer. *Cell Res* **30**: 163–178.

Linder P, Jankowsky E. 2011. From unwinding to clamping - the DEAD box RNA helicase family. *Nat Rev Mol Cell Biol* **12**: 505–516.

Liu Y, Gervasi C, Szaro BG. 2008. A crucial role for hnRNP K in axon development in Xenopus laevis. *Development* **135**: 3125–3135.

Liu Y, Szaro BG. 2011. hnRNP K post-transcriptionally co-regulates multiple cytoskeletal genes needed for axonogenesis. *Development* **138**: 3079–3090.

Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, Myers Z, Sud P, Jou J, Lin K, et al. 2020. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**: D882–D889.

Maklad A, Sedeeq M, Chan KM, Gueven N, Azimi I. 2023. Exploring Lin28 proteins: Unravelling structure and functions with emphasis on nervous system malignancies. *Life Sci* **335**: 122275.

Malinová A, Cvačková Z, Matějů D, Hořejší Z, Abéza C, Vandermoere F, Bertrand E, Staněk D, Verheggen C. 2017. Assembly of the U5 snRNP component PRPF8 is controlled by the HSP90/R2TP chaperones. *J Cell Biol* **216**: 1579–1596.

Mayr C. 2019. What are 3' UTRs doing? *Cold Spring Harb Perspect Biol* **11**: a034728.

McInnes L, Healy J, Astels S. 2017. hdbscan: Hierarchical density based clustering. *J Open Source Softw* **2**: 205.

McInnes L, Healy J, Melville J. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [statML]*. http://arxiv.org/abs/1802.03426.

Mitschka S, Mayr C. 2022. Context-specific regulation and function of mRNA alternative polyadenylation. *Nat Rev Mol Cell Biol* **23**: 779–796.

Monarez RR, Macdonald CC, Dass B. 2007. Polyadenylation proteins CstF-64 and τCstF-64 exhibit differential binding affinities for RNA polymers. *Biochem J* **401**: 651–658.

Moore JE, Pratt HE, Purcaro MJ, Weng Z. 2020. A curated benchmark of enhancer-gene interactions for evaluating enhancer-target gene prediction methods. *Genome Biol* **21**: 17.

Müller-McNicoll M, Neugebauer KM. 2013. How cells get the message: dynamic assembly and function of mRNA-protein complexes. *Nat Rev Genet* **14**: 275–287.

Nussbacher JK, Batra R, Lagier-Tourenne C, Yeo GW. 2015. RNA-binding proteins in neurodegeneration: Seq and you shall receive. *Trends Neurosci* **38**: 226–236.

Oughtred R, Rust J, Chang C, Breitkreutz B-J, Stark C, Willems A, Boucher L, Leung G, Kolas N, Zhang F, et al. 2021. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci* **30**: 187–200.

Palzer K-A, Bolduan V, Käfer R, Kleinert H, Bros M, Pautz A. 2022. The role of KH-type splicing regulatory protein (KSRP) for immune functions and tumorigenesis. *Cells* **11**: 1482.

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. *Journal of Machine Learning Research* **12**: 2825–2830.

Pereira B, Billaud M, Almeida R. 2017. RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends Cancer Res* **3**: 506–528.

Proepper C, Steinestel K, Schmeisser MJ, Heinrich J, Steinestel J, Bockmann J, Liebau S, Boeckers TM. 2011. Heterogeneous nuclear ribonucleoprotein k interacts with Abi-1 at postsynaptic sites and modulates dendritic spine morphology. *PLoS One* **6**: e27045.

Qin Y, Hou Y, Liu S, Zhu P, Wan X, Zhao M, Peng M, Zeng H, Li Q, Jin T, et al. 2022. A novel long non-coding RNA lnc030 maintains breast cancer stem cell stemness by stabilizing SQLE mRNA and increasing cholesterol synthesis. *Adv Sci (Weinh)* **9**: e2204046.

Qi Y, Wang M, Jiang Q. 2022. PABPC1--mRNA stability, protein translation and tumorigenesis. *Front Oncol* **12**:

1025291.

Quattrone A, Dassi E. 2019. The architecture of the human RNA-binding protein regulatory network. *iScience* **21**: 706–719.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

Rath S, Sharma R, Gupta R, Ast T, Chan C, Durham TJ, Goodman RP, Grabarek Z, Haas ME, Hung WHW, et al. 2021. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Res* **49**: D1541–D1547.

Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* **499**: 172–177.

Ray D, Laverty KU, Jolma A, Nie K, Samson R, Pour SE, Tam CL, von Krosigk N, Nabeel-Shah S, Albu M, et al. 2023. RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Sci Rep* **13**: 5238.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, et al. 2022. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**: D20–D26.

Shamseldin HE, Rajab A, Alhashem A, Shaheen R, Al-Shidi T, Alamro R, Al Harassi S, Alkuraya FS. 2013. Mutations in DDX59 implicate RNA helicase in the pathogenesis of orofaciodigital syndrome. *Am J Hum Genet* **93**: 555–560.

Sharma NK, Gupta S, Kumar P, Kumar A, Pradhan UK, Shankar R. 2021. RBPSpot: Learning on appropriate contextual information for RBP binding sites discovery. *bioRxiv*. http://dx.doi.org/10.1101/2021.06.07.447370.

Sharma S, Fazal FM. 2024. Localization of RNAs to the mitochondria-mechanisms and functions. *RNA* **30**: 597–608.

Street LA, Rothamel KL, Brannan KW, Jin W, Bokor BJ, Dong K, Rhine K, Madrigal A, Al-Azzam N, Kim JK, et al. 2024. Large-scale map of RNA-binding protein interactomes across the mRNA life cycle. *Mol Cell* **84**: 3790–3809.e8.

Sun L, Xu K, Huang W, Yang YT, Li P, Tang L, Xiong T, Zhang QC. 2021. Predicting dynamic cellular protein-RNA interactions by deep learning using in vivo RNA structures. *Cell Res* **31**: 495–516.

Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, et al. 2023. The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res* **51**: D638–D646.

Tareen A, Kinney JB. 2020. Logomaker: beautiful sequence logos in Python. *Bioinformatics* **36**: 2272–2274.

Tompa M. 1999. An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proc Int Conf Intell Syst Mol Biol* 262–271.

Turnham RE, Scott JD. 2016. Protein kinase A catalytic subunit isoform PRKACA; History, function and physiology. *Gene* **577**: 101–108.

Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen J-Y, Cody NAL, Dominguez D, et al. 2020a. A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**: 711–719.

Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, Blue SM, Nguyen TB, 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.

Van Nostrand EL, Pratt GA, Yee BA, Wheeler EC, Blue SM, Mueller J, Park SS, Garcia KE, Gelboin-Burkhart C, Nguyen TB, et al. 2020b. Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol* **21**: 90.

Wang J, Lee JE, Riemondy K, Yu Y, Marquez SM, Lai EC, Yi R. 2020. XPO5 promotes primary miRNA processing independently of RanGTP. *Nat Commun* **11**: 1845.

Wilbert ML, Huelga SC, Kapeli K, Stark TJ, Liang TY, Chen SX, Yan BY, Nathanson JL, Hutt KR, Lovci MT, et al. 2012. LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Mol Cell* **48**: 195–206.

Yang B, Wu Y, Chen Y, Li Y, Wang J, Cha X, Liu J. 2023. MiR-5195-3p targets the PCBP2/PI3K/AKT pathway to inhibit melanoma cell proliferation and migration. *Heliyon* **9**: e19227.

Yao C, Choi E-A, Weng L, Xie X, Wan J, Xing Y, Moresco JJ, Tu PG, Yates JR 3rd, Shi Y. 2013. Overlapping and distinct functions of CstF64 and CstF64τ in mammalian mRNA 3' processing. *RNA* **19**: 1781–1790.

Zhao H, Wei Z, Shen G, Chen Y, Hao X, Li S, Wang R. 2022. Poly(rC)-binding proteins as pleiotropic regulators in hematopoiesis and hematological malignancy. *Front Oncol* **12**: 1045797.

Zhou H-L, Baraniak AP, Lou H. 2007. Role for Fox-1/Fox-2 in mediating the neuronal pathway of calcitonin/calcitonin gene-related peptide alternative RNA processing. *Mol Cell Biol* **27**: 830–841.

FIGURE CAPTIONS

Figure 1: Overview of the NLP representation, MIL formulation, and implementation of the RBP binding prediction task. (A) The first phase of the sliding window-based sequence decomposition strategy deconstructs a sequence into candidate target *k*-mers. (B) The second phase of sequence decomposition constructs the context elements for each candidate target generated in panel A. (C) Sequence definition and Multiple Instance Learning formulation. Positive and negative observations, derived from the same genomic region, are deconstructed into their constituent contexts with the sliding window approach, then organized into bags. The precise RBP binding region within the positive sequence is highlighted in green and represents the main prediction objective, indicating the specific contexts that confer upon the parent sequence its positive label. (D) Process diagram of the model training and iterative relabeling pipeline. (E) Distribution of instance-level and bag-level model performances across 71 RBPs in HepG2, before and after iterative relabeling.

Figure 2: A novel algorithm for motif and context discovery. (A) Schematic overview of the motif discovery algorithm. (B) Exemplification of motif discovery validation, with respect to the literature, for 4 representative RBPs with motifs of diverse nucleotide content. Literature motifs obtained from (Ray et al. 2013; Dominguez et al. 2018; Van Nostrand et al. 2020a). For each RBP, one *in vitro* and one *in vivo* motif logo is displayed, with the exception of QKI, for which only *in vitro* logos are shown; however, QKI's motif has been well-characterized *in vivo* by multiple studies, such as in (Hall et al. 2013). (C) Unique instances comprising the RBFOX2 canonical GCAUG motif; 6 motif instances recapitulate known primary and secondary *k*-mers bound by RBFOX2 *in vitro* (Begg et al. 2020). (D) Context motif for RBFOX2; the target motif GCAUG is preceded by a U and enriched for in a G-rich sequence context, as previously demonstrated in the literature. Nucleotide preference of RBFOX2 binding, visualized as (E) a line plot and (F) a heatmap.

Figure 3: RBPs cluster based on similarities between their respective motifs and contexts. Clustering of motif (A) and context (B) PFMs with HDBSCAN and visualization with UMAP. Clusters are assigned number labels, with cluster colors roughly mapping between the two embeddings. (C) Motif and context logos for representative RBPs from selected clusters; cluster labels and nucleotide preferences underlying the clustering assignments from panels A and B for target (left) and context (right) motifs are labeled in the periphery. RBPs whose names are labeled in black are those that migrated clusters between the target and context clustering, with the cause for migration evident from the corresponding logos.

Figure 4: Cross-prediction as a novel approach to reveal similarities in global binding patterns between RBPs. (A) Schematic demonstrating the operations involved in the cross-prediction approach. (B) Clustered heatmap of the pairwise cross-prediction AUPRs between all RBPs. Regions highlighted in panels D through H are labeled in the heatmap. (C) Enlarged dendrogram from the clustered heatmap in panel B, revealing hierarchical clustering of RBPs based on similarity of their respective cross-prediction profiles across all other RBPs. (D-H) Highlighted regions of high cross-prediction performance from the heatmap in panel B; the arrow symbol in the

bottom left corner of panels D and G indicates that the inset was rotated 90° clockwise relative to the original heatmap in B.

Figure 5: A method to infer RBP-RBP interactions. (A) Conceptual representation for the classification of RBP-RBP interaction categories, which depend on properties of the involved components. Distal interactions, while demonstrated here, are not considered in this analysis. (B) Interaction discovery method, represented as a decision tree based on features of the candidate RBP-RBP pair. Each path (I through IV) along the tree represents a set of possible interaction modes based upon the fulfilled conditions along the path. A candidate pair may be involved in a cooperative interaction, competitive interaction, or, potentially, both. (C) Feature summary and (D) GO analysis (top 50 terms with FDR < 0.05 shown) for the proposed interaction between AKAP1 and UPF1. (E) Feature summary and (F) GO analysis (top 50 terms with FDR < 0.05 shown) for the proposed interaction between DDX59, HNRNPK, and PCBP2.

Supplemental Fig. S1: Pairwise Kullback-Leibler (KLD) clustering. (A) Heatmap and clustering of motif KLDs between RBP pairs. (B) Heatmap and clustering of context KLDs between RBP pairs.

SUPPLEMENTAL TABLES

Supplemental_Table_S1.xlsx: ENCODE eCLIP data experiment and file accessions

Supplemental_Table_S2.xlsx: MIL performance and statistics

Supplemental_Table_S3.xlsx: Reduction in witness rate

Supplemental_Table_S4.xlsx: Motif discovery ground-truth comparison

Supplemental_Table_S5.xlsx: RBFOX2 motif instance itemization

Supplemental_Table_S6.xlsx: Context discovery comparison and characterization

Supplemental_Table_S7.xlsx: Motifs, contexts, and nucleotide preferences for all RBPs

Supplemental_Table_S8.xlsx: Motif characterization

Supplemental_Table_S9.xlsx: Cross-prediction performance

Supplemental_Table_S10.xlsx: Interaction discovery for all RBP-RBP pairs

Supplemental_Table_S11.xlsx: Features supporting known interactions

Supplemental_Table_S12.xlsx: Features of candidate interacting rG4-binding proteins

Supplemental_Table_S13.xlsx: Features of spliceosomal proteins

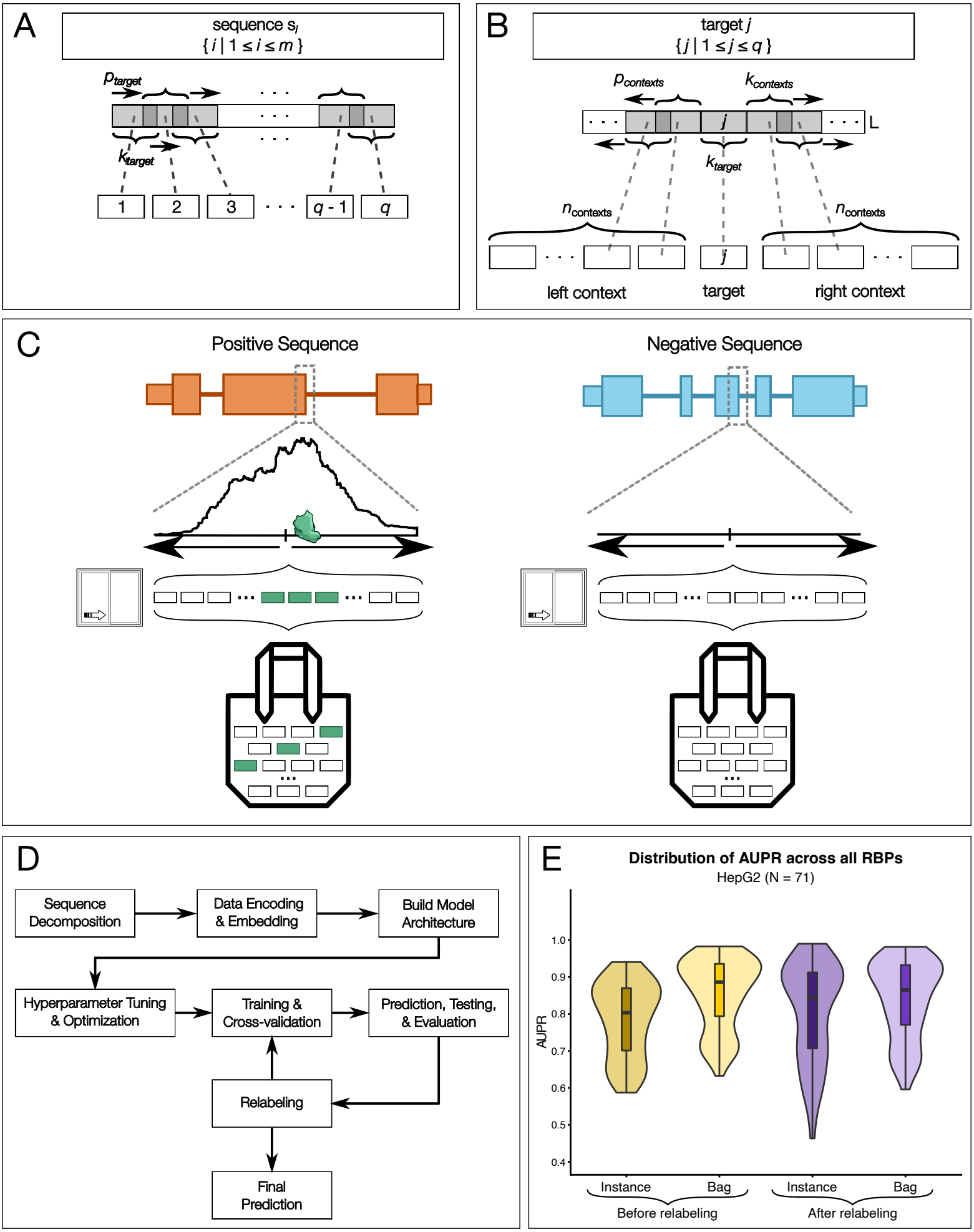


Figure 1

Elhajjajy et al. 67

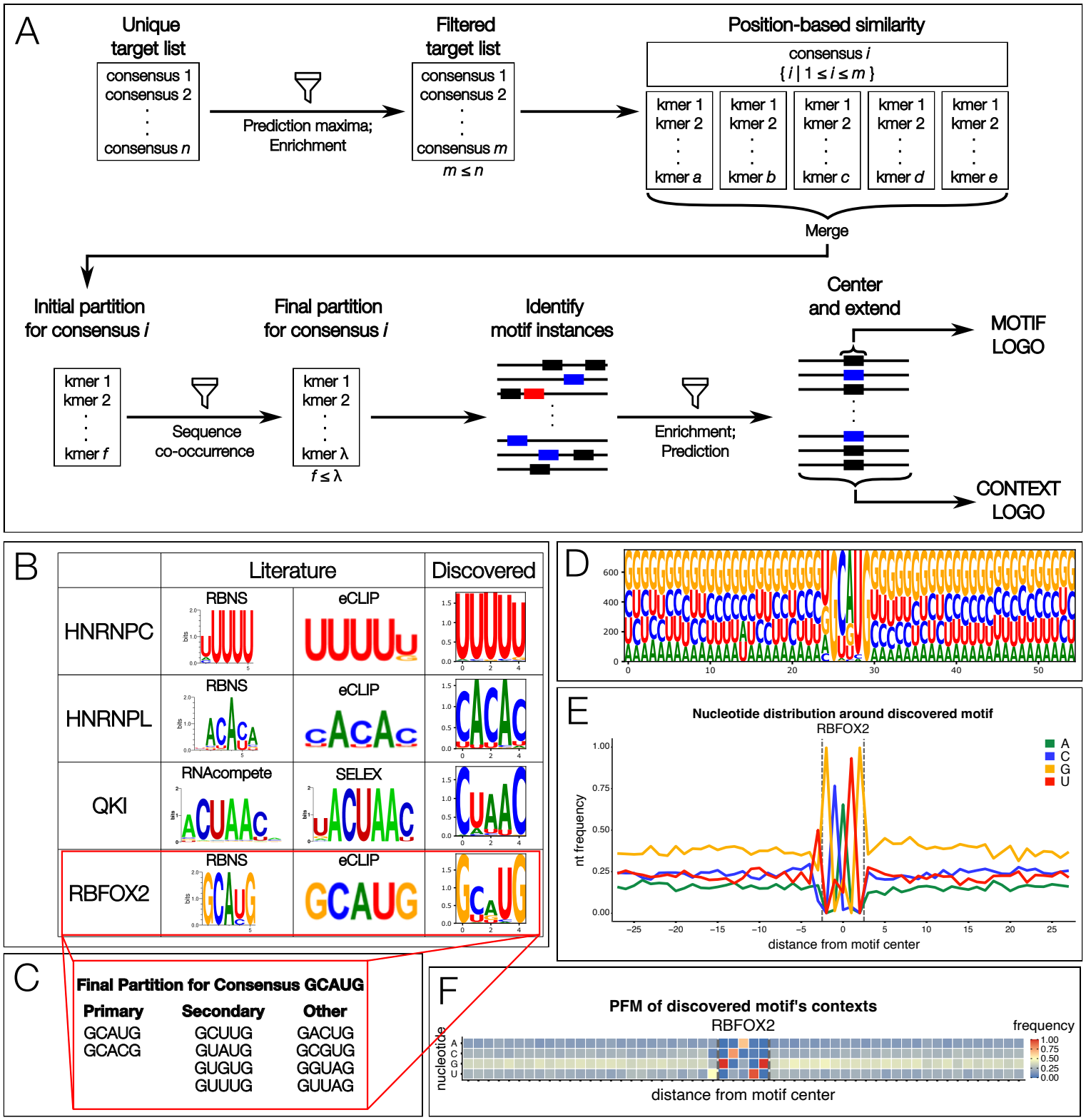
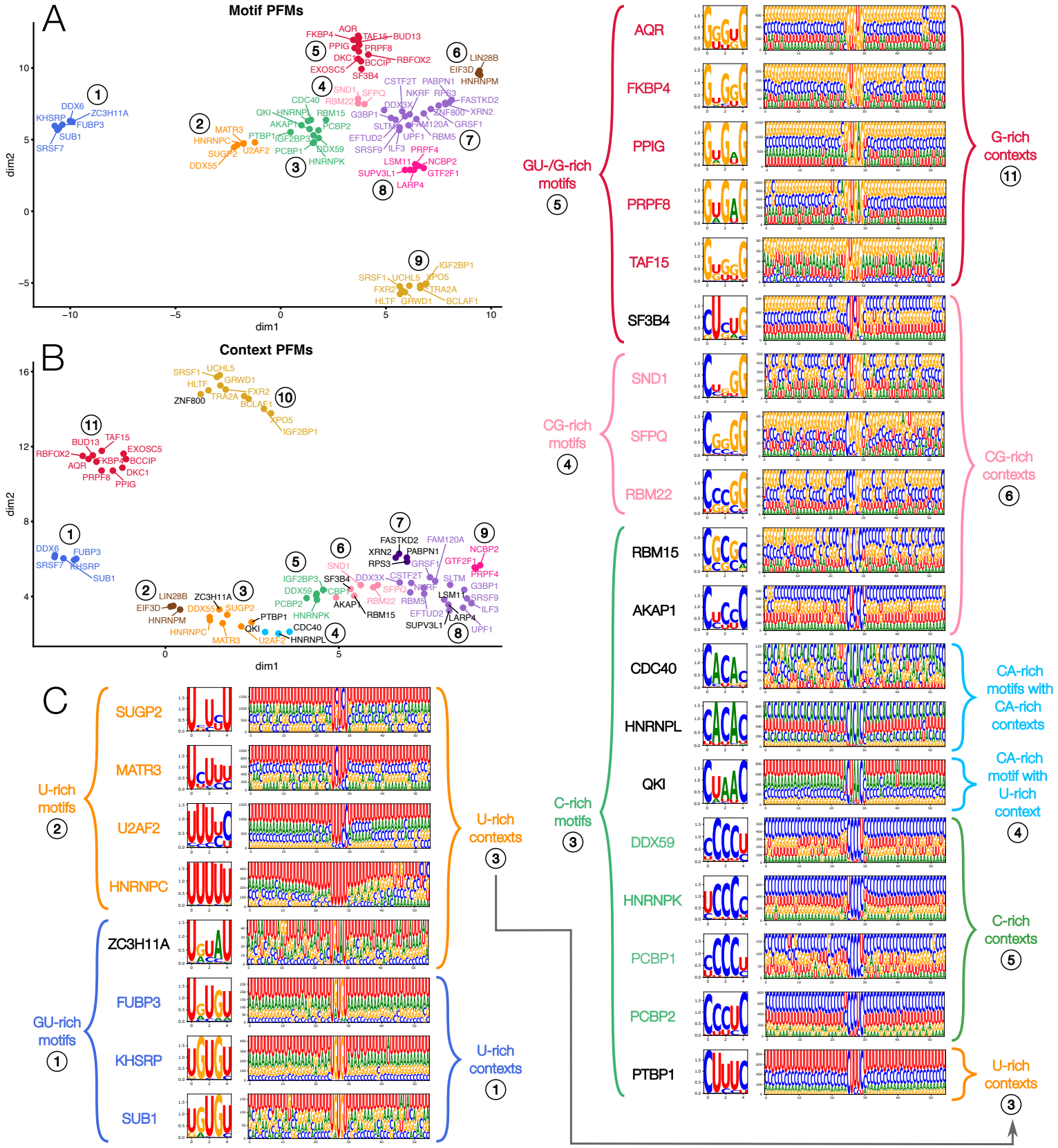


Figure 2

Elhajjajy et al. 68


A
Motif PFMs
dim1
dim2
B
Context PFMs
dim1
dim2
C
U-rich motifs
2
SUGP2
MATR3
U2AF2
HNRNPC
GU-rich motifs
1
ZC3H11A
FUBP3
KHSRP
SUB1
U-rich contexts
3
U-rich contexts
1
GU-/G-rich motifs
5
AQR
FKBP4
PPIG
PRPF8
TAF15
SF3B4
G-rich contexts
11
CG-rich motifs
4
SND1
SFPQ
RBM22
CG-rich contexts
6
C-rich motifs
3
RBM15
AKAP1
CDC40
HNRNPL
QKI
DDX59
HNRNPK
PCBP1
PCBP2
PTBP1
CA-rich motifs with CA-rich contexts
CA-rich motif with U-rich context
4
C-rich contexts
5
U-rich contexts
3

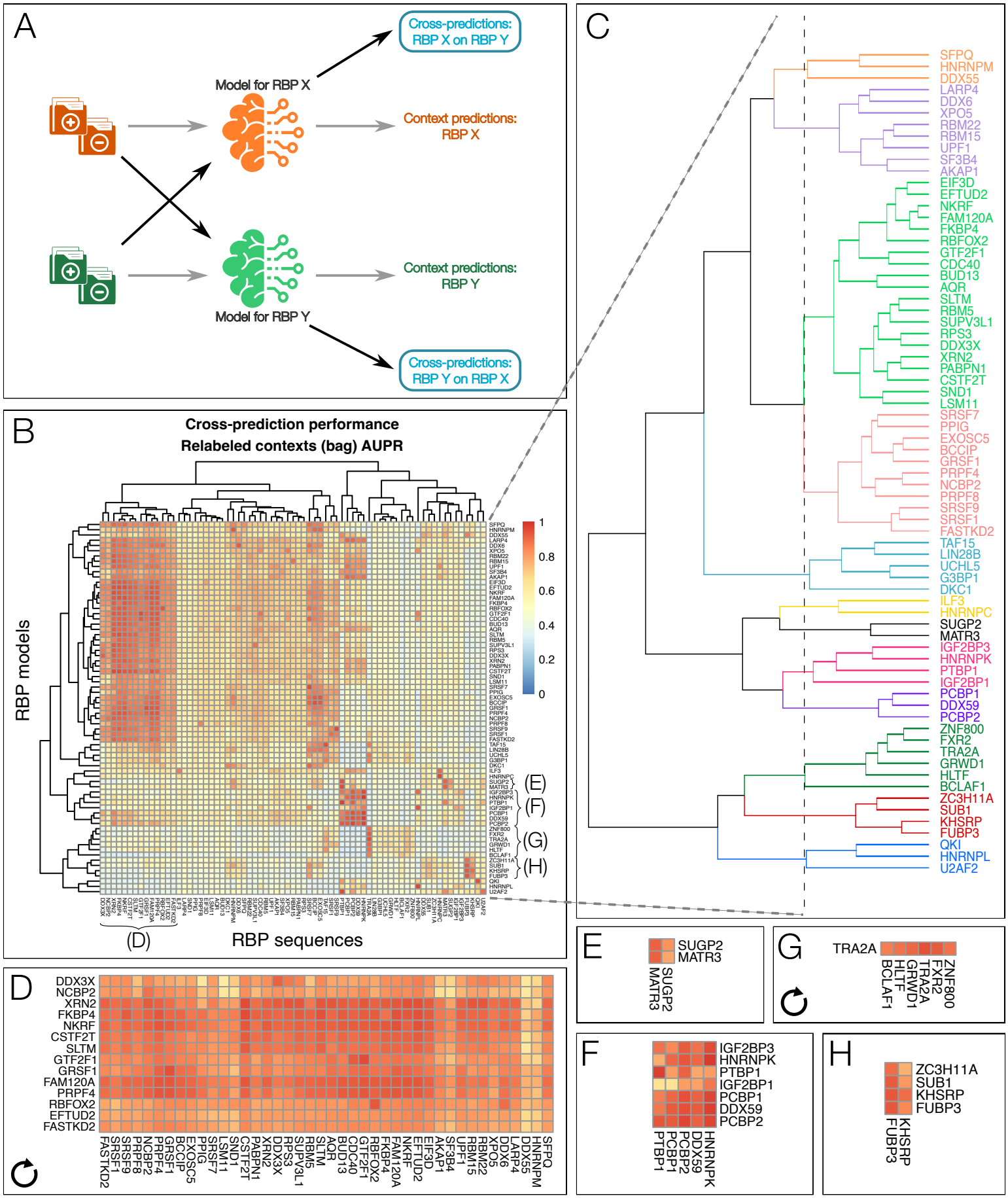
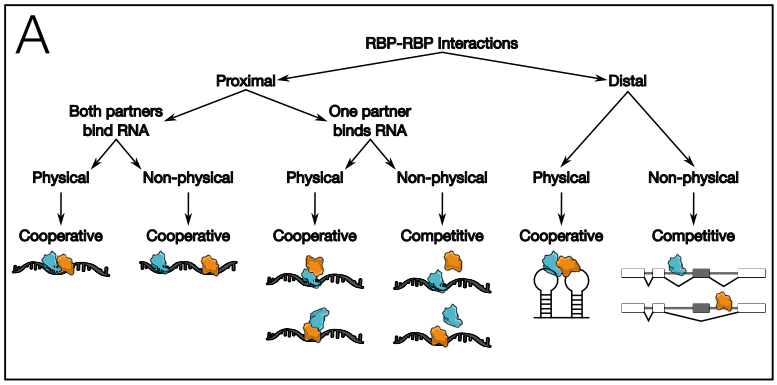
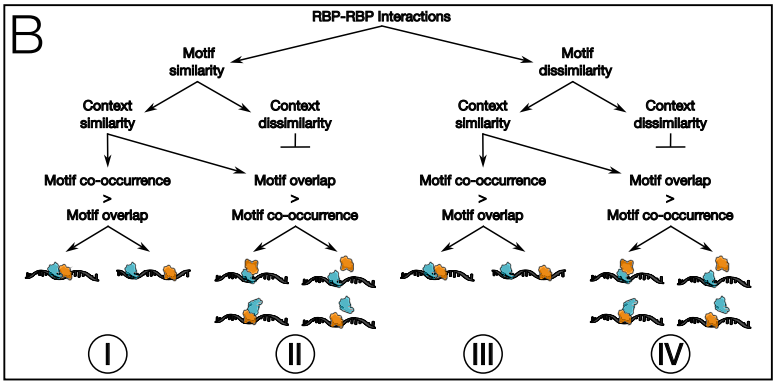

A
Cross-predictions: RBP X on RBP Y
Model for RBP X
Context predictions: RBP X
Context predictions: RBP Y
Model for RBP Y
Cross-predictions: RBP Y on RBP X
B
Cross-prediction performance
Relabeled contexts (bag) AUPR
RBP models
RBP sequences
(D)
(E)
(F)
(G)
(H)
C
D
E
SUGP2
MATR3
F
IGF2BP3
HNRNPK
PTBP1
IGF2BP1
PCBP1
DDX59
PCBP2
G
TRA2A
H
ZC3H11A
SUB1
KHSRP
FUBP3

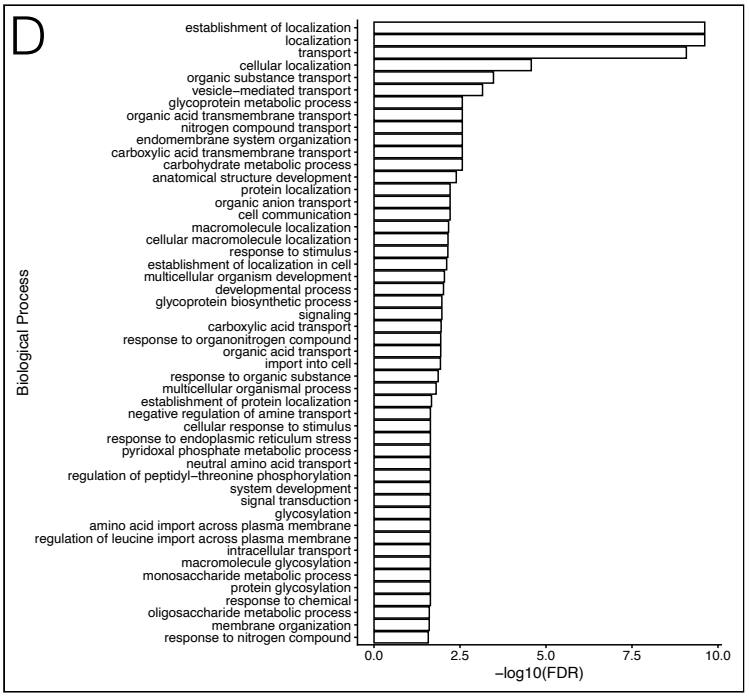
Figure 4

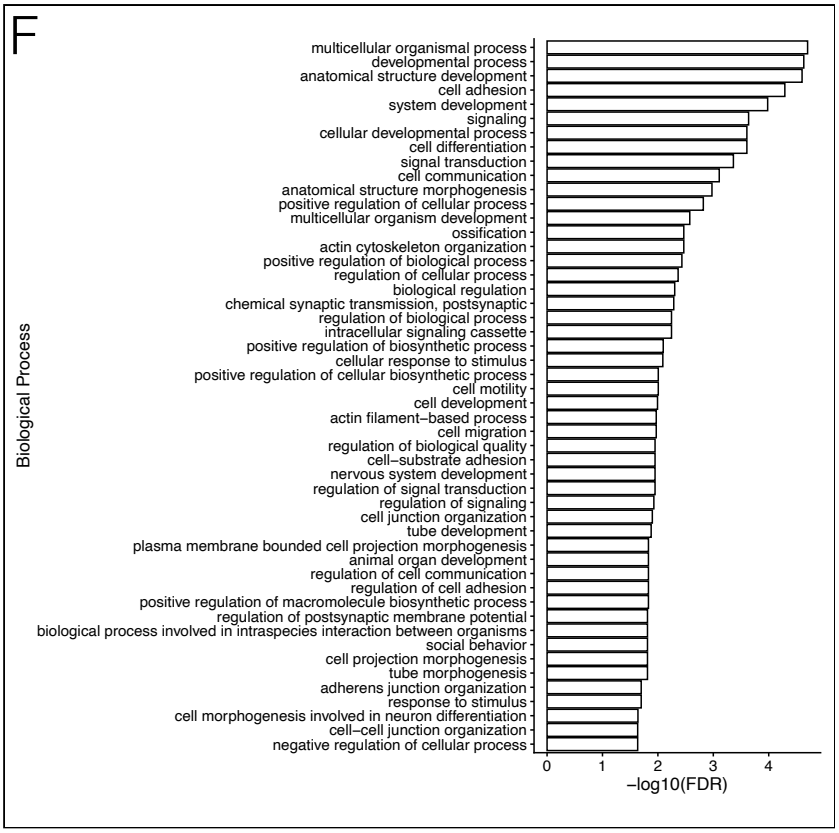




C

			Reference	
			AKAP1	UPF1
Motif KLD	Comparison	AKAP1	0	33.64
		UPF1	31.36	0
Context KLD		AKAP1	0	35.59
		UPF1	33.32	0
Cross-prediction AUPR		AKAP1	80.84	81.51
		UPF1	82.25	81.32
Percent of Motif Instance Overlaps		AKAP1	100	0.22
		UPF1	0.37	100
Percent of Motif Instance Co-occurrence		AKAP1	0	18.57
		UPF1	30.65	0
Total Number of Motif Instances			4064	6603



E

			Reference		
			DDX59	HNRNPK	PCBP2
Motif KLD	Comparison	DDX59	0	3.04	4.81
		HNRNPK	3.07	0	7.22
		PCBP2	6.34	6.70	0
Context KLD		DDX59	0	3.90	7.75
		HNRNPK	3.93	0	9.36
		PCBP2	9.56	8.84	0
Cross-prediction AUPR		DDX59	92.44	90.98	91.77
		HNRNPK	97.44	98.14	96.98
		PCBP2	94.75	93.87	96.64
Percent of Motif Instance Overlaps		DDX59	100	8.40	3.51
		HNRNPK	16.35	100	4.17
		PCBP2	9.72	5.89	100
Percent of Motif Instance Co-occurrence		DDX59	0	12.53	9.12
		HNRNPK	24.30	0	13.77
		PCBP2	23.35	18.25	0
Total Number of Motif Instances			3299	6323	9240