Bioinformatics, 37, 2021, i308–i316
doi: 10.1093/bioinformatics/btab300
ISMB/ECCB 2021

Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins

Salma Sohrabi-Jahromi[1] and Johannes Söding[1,2,*]

[1]Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen 37077, Germany and [2]Campus-Institut Data Science (CIDAS), Göttingen 37077, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Understanding how proteins recognize their RNA targets is essential to elucidate regulatory processes in the cell. Many RNA-binding proteins (RBPs) form complexes or have multiple domains that allow them to bind to RNA in a multivalent, cooperative manner. They can thereby achieve higher specificity and affinity than proteins with a single RNA-binding domain. However, current approaches to *de novo* discovery of RNA binding motifs do not take multivalent binding into account.

Results: We present Bipartite Motif Finder (BMF), which is based on a thermodynamic model of RBPs with two cooperatively binding RNA-binding domains. We show that bivalent binding is a common strategy among RBPs, yielding higher affinity and sequence specificity. We furthermore illustrate that the spatial geometry between the binding sites can be learned from bound RNA sequences. These discovered bipartite motifs are consistent with previously known motifs and binding behaviors. Our results demonstrate the importance of multivalent binding for RNA-binding proteins and highlight the value of bipartite motif models in representing the multivalency of protein-RNA interactions.

Availability and implementation: BMF source code is available at https://github.com/soedinglab/bipartite_motif_finder under a GPL license. The BMF web server is accessible at https://bmf.soedinglab.org.

Contact: soeding@mpibpc.mpg.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA in the cell is rarely naked but rather covered with numerous RNA-binding proteins (RBPs) (Singh *et al.*, 2015). These RBPs play crucial roles in regulating the various steps of RNA biochemistry, from RNA maturation and transport to cellular localization, translation and degradation (Gerstberger *et al.*, 2014). RNAs can in turn regulate RBP function by altering their stability, interaction partners and localization (Hentze *et al.*, 2018). These processes require specific binding of RBPs to their target RNAs. RBPs mostly achieve this specificity through RNA-binding domains (RBDs) that engage with specific RNA sequences or structures (Li *et al.*, 2014). Unraveling the target preferences of RBPs is therefore key to understanding cellular regulation.

Many experimental techniques have emerged to generate systematic maps of protein-RNA interactions. To find *in vivo* binding sites, many variants of RNA immunoprecipitation (RIP-seq) (Gilbert and Svejstrup, 2006) and cross-linking immunoprecipitation (CLIP-seq), such as PAR-CLIP (Hafner *et al.*, 2010), iCLIP (König *et al.*, 2010) and eCLIP (Van Nostrand *et al.*, 2016), have been proposed. In both approaches, RNAs bound to the immunoprecipitated protein of interest are sequenced and mapped to the genome. Deriving accurate models of binding affinities from *in vivo* data is problematic because RBP-RNA interactions are influenced by cooperativity and competition with other RBPs, by RNA localization, expression and folding (Änkö and Neugebauer, 2012). Therefore, techniques have been developed to measure binding affinities *in vitro*, in isolation from other RBPs, using random libraries of RNA substrates: RNA Bind-n-Seq (RBNS) (Lambert *et al.*, 2014), RNAcompete (Cook *et al.*, 2017; Ray *et al.*, 2013) and high-throughput RNA-SELEX (HTR-SELEX) (Jolma *et al.*, 2020).

A wide range of motif discovery tools have been developed to learn models of sequence- and secondary structure-dependent binding affinties of RBPs based on datasets of sequences bound *in vitro* or *in vivo* by an RBP of interest (Kazan *et al.*, 2010; Maticzka *et al.*, 2014; Munteanu *et al.*, 2018; Stražar *et al.*, 2016). More recently, a new wave of algorithms have been introduced that use deep neural networks to predict RBP binding sites (Alipanahi *et al.*, 2015; Ghanbari and Ohler, 2020; Grønning *et al.*, 2020; Pan and Shen, 2018; Yan and Zhu, 2020). One challenge is to explain what these complex models have learned, although recently a multitude of methods for interpreting the learned models have been developed, for instance, based on *in silico* mutagenesis, predictions on synthetic sequences, gradient tracing and analyzing the convolutional filters (Alipanahi *et al.*, 2015; Ghanbari and Ohler, 2020; Koo *et al.*, 2020; Pan and Shen, 2018). However, with the increasing number of model parameters and network complexity, the risk grows that such models could also learn experimental biases in the datasets. This is particularly problematic for RBPs, since many of them show short and degenerate sequence preferences. Moreover, RBPs often bind low-complexity untranslated regions in the RNA (Dominguez *et al.*, 2018), unlike transcription factors, which usually bind to more complex sequence motifs and have higher binding specificities.

Around half of eukaryotic RBPs have multiple domains and a majority of the remaining are estimated to have oligomerization tendencies (Stitzinger *et al.*, 2021). In line with this, RBPs have been shown by spaced *k*-mer counting approaches to often bind with multiple RBDs two separated cores with usually similar or identical motifs (Dominguez *et al.*, 2018; Jolma *et al.*, 2020). Transcription factor motif discovery tools have been developed to learn co-occurrence of motif pairs in genomic sequences (Toivonen *et al.*, 2018, 2020), and more recently, it was shown that distance dependent RNA motif pairs can be inferred from neural networks (Koo *et al.*, 2020; Quinn *et al.*, 2020). However, transcription factor binding fundamentally differs from RBP binding as DNA can mediate cooperativity by propagating structural deformations induced by binding of proteins along the helix.

In this work, we present Bipartite Motif Finder (BMF), a tool for learning bipartite RNA motifs in RNA–protein interaction datasets. BMF sums up the contribution of all alternative binding conformations, and not just the best binding configuration. This is critical to accurately model the binding affinity of RBPs, which often have low information content to their many potential binding sites, because combinatorially many binding configurations can have similar total binding energies and thus contribute to the binding probability (Forties and Bundschuh, 2010). It is particularly critical for modeling bipartite binding, and to the best of our knowledge, BMF is the first thermodynamic approach to *de novo* motif discovery for bipartite RNA binding. We demonstrate that BMF is able to detect short and degenerate motifs and to learn the spatial relationship between them. We furthermore show that around half of RBPs manifest multivalent binding with a preferential linker distance between the two binding sites.

Benchmarking the performance of learned binding site models by cross-validation can be problematic when testing methods that train highly parameterized models such as deep neural networks, as these methods can learn biologically irrelevant sequence biases inherent to the experimental method. To compare BMF to existing tools and assess their capacity for learning relevant motif sequences that predict binding events in the cell, we built a cross-platform validation benchmark, training models on HTR-SELEX data and testing on *in vivo* CLIP data. Despite the many complicating effects *in vivo*, we find that the motif and distance preferences learned by BMF can predict RBP binding in the cellular context and that high-quality motifs learned *in vitro* are often very similar to the motifs learned on *in vivo* data. Moreover, BMF can predict binding sites on par with or even better than existing tools.

2 Materials and methods

Most RBPs can bind RNA using several structured RBDs and often also using disordered regions, some of which contain typical RGG/RG and RS motifs, which can modulate RNA-binding activity (Calabretta and Richard, 2015; Lunde *et al.*, 2007; Ozdilek *et al.*, 2017). Furthermore, many RNA-binding proteins dimerize or homo- and hetero-oligormerize. This effectively leads to two and more RBDs binding cooperatively to RNA molecules. Here, we present Bipartite Motif Finder (BMF), a motif search tool and algorithm to describe the sequence specificity of monovalently and multivalently binding proteins or protein complexes.

2.1 Thermodynamic model for bivalent RNA binding

We consider the simple case in which the RBP consists of two RBDs, *A* and *B* (Fig. 1A). We describe the binding of proteins at concentration c_{AB} to a single, specific RNA sequence $\mathbf{x} = (x_0 \dots x_{L-1}) = x_{0:L-1}$ composed of nucleotides x_i. We consider not only the most likely binding configuration but rather all possible binding configurations, involving zero, one or more proteins bound to the RNA (Fig. 1B). According to Boltzmann's law, each binding configuration $\mathbf{c}$ has a probability $p(\mathbf{c})$ proportional to its so-called *statistical weight* $e^{(-E(\mathbf{c})-T\Delta S(\mathbf{c}))/k_{\mathrm{B}}T}$, where $F(\mathbf{c}) = -E(\mathbf{c}) - T\Delta S(\mathbf{c})$ is the free energy composed of the binding enthalpy $-E(\mathbf{c})$ and a part related to the change in entropy $\Delta S(\mathbf{c})$ between the completely unbound and bound states. To obtain probabilities, the statistical weights need to be normalized at the end by dividing by their total sum, the partition sum $Z(\mathbf{x})$.

The change in entropy due to the binding of a single protein that is present at concentration c_{AB} is equal to its chemical potential, which is $\Delta S = k_{\mathrm{B}} \log c_{\mathrm{AB}}$. In the following, we compute all energies in units of $k_{\mathrm{B}}T$, so we set $k_{\mathrm{B}}T = 1$. In our model, the concentration $c_{\mathrm{B}}(d)$ of the downstream domain *B* at the RNA depends on the distance *d* to the binding site of the upstream domain *A* (see next section).

We compute the statistical weights of all binding configurations iteratively using dynamic programming. We split the configurations into two sets, A and B, and define $Z_{\mathrm{A}}(i)$ to be the sum of statistical weights of all binding configurations on the RNA up to position i, $x_{0:i}$, for which domain A is bound at position $i-k+1$ to i, where k is the length of RNA bound by the domains. Similarly, we define $Z_{\mathrm{B}}(i)$ to be the sum of statistical weights of all binding configurations on the RNA sequence $x_{0:i}$ for which no domain is bound or domain B is bound with its right edge upstream of or at position i. With the knowledge of $Z_{\mathrm{A}}(i')$ and $Z_{\mathrm{B}}(i')$ for $0 \leq i' < i$, we can compute $Z_{\mathrm{A}}(i)$ and $Z_{\mathrm{B}}(i)$ (Fig. 1B):

$$Z_{\mathrm{A}}(i) = \left(Z_{\mathrm{B}}(i-l) + \sum_{j=0}^{i-l} Z_{\mathrm{A}}(j) \right) c_{\mathrm{AB}} e^{-E_A(x_{i-k+1:i})}, \tag{1}$$

$$\begin{aligned} Z_{\mathrm{B}}(i) = Z_{\mathrm{B}}(i-1) + \sum_{j=0}^{i-l} Z_{\mathrm{A}}(j)\, c_{\mathrm{B}}(i-k-j)\, e^{-E_B(x_{i-k+1:i}))} \\ + Z_{\mathrm{B}}(i-l) c_{\mathrm{AB}} e^{-E_B(x_{i-k+1:i})}, \end{aligned} \tag{2}$$

where $E_A(x_{i-k+1:i})$ and $E_B(x_{i-k+1:i})$ represent the binding energies of domains A and B to the RNA sequence $x_{i-k+1:i}$. The concentration of the single B domain, defined as expected number of B per volume, is simply its probability density. The dynamic programming is initialized using

$$Z_{\mathrm{A}}(i) = 0 \mathrm{forall} i \in \{0, \dots, k-2\}, \tag{3}$$

$$Z_{\mathrm{B}}(i) = 1 \mathrm{forall} i \in \{0, \dots, k-2\}. \tag{4}$$

The first equation follows from requiring all k positions in the binding motif to be part of sequence $x_{0:L-1}$. The second equation follows from the fact that $Z_{\mathrm{B}}(i)$ for $i < k-1$ sums up only the statistical weight of the unbound configuration.

The partition sum $Z(\mathbf{x})$ for RNA sequence $\mathbf{x}$ is the sum of statistical weights of all configurations,

$$Z(\mathbf{x}) = Z_{\mathrm{B}}(L-1) + \sum_{i=0}^{L-1} Z_{\mathrm{A}}(i)\,. \tag{5}$$

The probability for an RNA to not be bound by any protein (neither *A* nor *B* domains) is just the statistical weight of the unbound configuration, set to 1, times the normalization factor $1/Z(\mathbf{x})$, so the probability for a RNA $\mathbf{x}$ to be bound by a protein is $p(\mathrm{bound}|\mathbf{x}) = 1 - 1/Z(\mathbf{x})$.

By taking the partial derivatives of equations (1) and (2) with respect to the model parameters (Supplementary Methods), we obtain update equations for the partial derivatives with which we can in turn compute the partial derivatives of $Z(\mathbf{x})$, $p(\mathrm{bound}|\mathbf{x})$, and the log likelihood in equation (8). These allow us to find optimum model parameters by gradient-based maximization of the log-likelihood.

2.2 Motif model of a single RNA-binding region

Position weight matrices (PWMs) and Bayesian Markov models (BaMMs) have been used to represent RBP binding preferences through positional or conditional probabilities of observing each nucleotide at a given position (Hartmann *et al.*, 2013; Siebert and Söding, 2016). Since RBPs are known to bind shorter and more repetitive sequences, we learn binding energies for all 4^k *k*-mers at

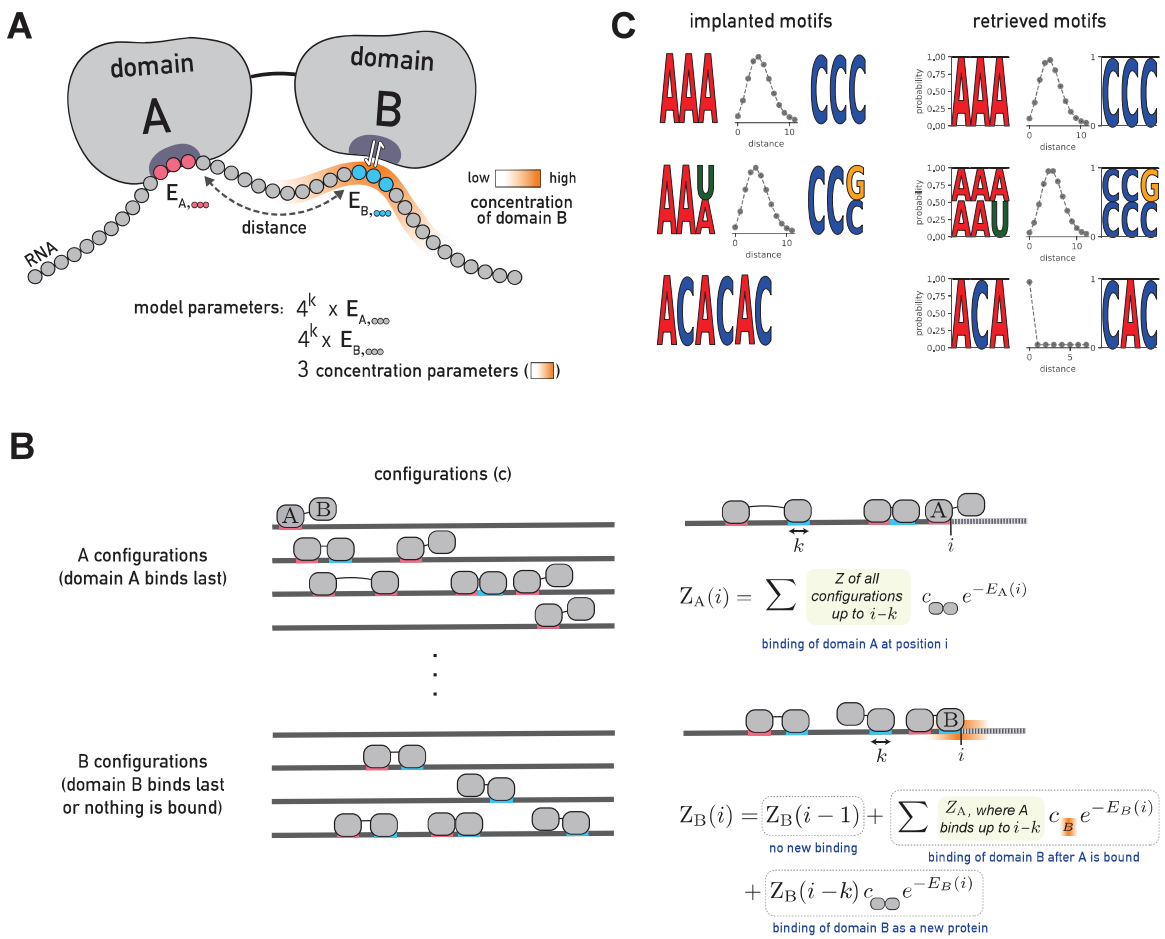

Fig. 1. BMF can learn multivalent binding preferences for RBPs. (**A**) RBP-RNA interaction model for a protein with two RBDs. BMF optimizes the binding energies of each domain to all possible RNA k-mers ($k = 3$ here) and learns the distance distribution between the motif cores. BMF models the high RNA local concentration at the second binding site, when the first domain is bound to the RNA. (**B**) BMF calculates binding probabilities for all binding configurations of one or several proteins to the RNA sequence. $Z_A(i)$ is the sum of statistical weights of all binding configurations on the RNA up to position i, for which domain A is bound at position i. Similarly, $Z_B(i)$ is the sum of statistical weights of all binding configurations on the RNA subsequence for which no domain is bound or domain B is bound with its right edge upstream of or at position i. Z_A and Z_B are calculated iteratively (right panel). The first term in the second equation accounts for configurations for which position i is not bound by anything, the second term accounts for configurations for which domain A of the same protein is bound at j (as seen in the example illustration) and the last term accounts for configurations for which domain B binds whose A domain is not bound upstream of i. (**C**) BMF recovers the correct RNA motifs implanted in synthetic datasets for all tested cases. Here and in the following figures, the two learned core motifs are visualized by plotting the energies of the top five k-mers, converted to k-mer probabilities according to Boltzmann's law and normalized to 1

each motif core, $E_A(k\text{-mer})$ and $E_B(k\text{-mer})$. The length k of the motif can be set by the user.

2.3 Model for the effective concentration $c_B(d)$

Spaced k-mer analyses on high-throughput RNA-binding datasets pointed to a length preference of the RNA linker connecting two motif cores (Dominguez *et al.*, 2018; Jolma *et al.*, 2020; Schneider *et al.*, 2019). The concentration of domain B after domain A binds the RNA molecule is equal to its probability distribution. While according to the flexible chain model of the RNA fragment the concentration should be a Gaussian distribution centered on domain A (Rubinstein *et al.*, 2003), for short RNA linkers the concentration can peak some distance away from domain A. To describe multivalent binding for both short-range and long-range co-occurrence of motif sequences, we model the effective concentration at the second binding site with a negative binomial (NB) distribution,

$$c_B(d) = c_{AB} + S \cdot \binom{d+r-1}{d} \cdot p^r (1-p)^d, \qquad (6)$$

where d represents the the number of nucleotides between the binding sites of A and B on the RNA, and r and p are parameters of the negative binomial distribution. The total concentration of B is the cellular concentration (c_{AB}) plus $c_B(d)$, the local concentration of B linked to a bound A. We scale the negative binomial with the factor S as a conversion to protein concentration values. Since only the ratio between S and c_{AB} determine the binding dynamics, we fix c_{AB} to one and optimize our bipartite model for S, r and p.

2.4 Parameter initialization

The absolute values of the energy parameters in our model do not reflect the physical binding energies, however their relative values determine the probability of binding to a given sequence. We therefore draw initial energy parameters randomly (in units of k_BT) from a normal distribution with the average of 12 and standard deviation of one. The initial value of 12 k_BT was chosen based on experimentally determined binding energies (Yang *et al.*, 2013) and additionally ensures that the algorithm does not overflow. The scaling factor S is initialized as 10^4. The spacer parameter r is drawn from a uniform distribution from one to five and p is randomly drawn between zero and 0.5.

2.5 Likelihood estimation for HTR-SELEX measurements

In HTR-SELEX experiments (and similarly for bind-n-Seq), we have input (background) library sequences $\mathbf{x} \in \mathcal{X}^{bg}$ and sequences

enriched after competitive binding, $\mathbf{x} \in \mathcal{X}^+$. We denote with $p_b(\mathbf{x})$ the fraction of sequence $\mathbf{x}$ in the input library. To find a sequence in $\mathbf{x} \in \mathcal{X}^+$, it must have first been present in the input library (probability $p_b(\mathbf{x})$) and then have been bound to the RNA (probability $p(\text{bound}|\mathbf{x})$). The probability to find a sequence $\mathbf{x} \in \mathcal{X}^+$ after the selection is therefore, according to Bayes' theorem,

$$p(\mathbf{x}|\text{bound}) = \frac{p(\text{bound}|\mathbf{x})p_{bg}(\mathbf{x})}{\sum_{\hat{\mathbf{x}}\in\mathcal{X}^{bg}} p(\text{bound}|\hat{\mathbf{x}})p_{bg}(\hat{\mathbf{x}})}, \tag{7}$$

and, using $p(\text{bound}|\mathbf{x}) = 1 - 1/Z(\mathbf{x})$, the log-likelihood is

$$LL = \ln \prod_{x\in\mathcal{X}^+} p(\mathbf{x}|\text{bound}) = \sum_{\mathbf{x}\in\mathcal{X}^+} \left(\ln p_{bg}(\mathbf{x}) + \ln\left(1 - \frac{1}{Z(\mathbf{x})}\right)\right) - N^+ \ln \sum_{\hat{\mathbf{x}}\in\mathcal{X}^{bg}} p_{bg}(\hat{\mathbf{x}})\left(1 - \frac{1}{Z(\mathbf{x})}\right). \tag{8}$$

2.6 Parameter optimization

We learn the model parameters by maximizing the likelihood function (eq. 8). For an efficient optimization using stochastic gradient descent, we computed the partial derivative of the likelihood function with respect to all of the model parameters (Supplementary Methods). For parameter optimization, we used ADAM (Kingma and Ba, 2014) with its default hyperparameters $\alpha = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-8}$, and a minibatches size of 512. We parameterized $r = e^{\rho}$ and $p = 1/(1 + e^{-\pi})$ to ensure that r and p stay within bounds. Optimization was terminated when 1000 iterations were reached or when the variation ν_θ for the best bound k-mer of each domain as well as for p and r fall under a threshold of 0.03. The variation for the parameter θ up to iteration t was defined as $\nu_\theta = (\max\{\theta_{t-4:t}\} - \min\{\theta_{t-4:t}\})/\theta_t$. We stop BMF after 1000 iterations since continuing for another 500 iterations did not change the prediction performance of MBF models in a cross-validated HTR-SELEX dataset (Supplementary Fig. S1).

2.7 Evaluating the performance of BMF on synthetic data

In order to evaluate BMF's ability to learn bipartite motifs, we generated two sets of 2000 RNA sequences, an artificial input set and an enriched set. For the enriched set, we inserted the first core of the simulated bipartite motif at random positions. The second core was inserted with a linker length drawn from a binomial distribution with a specific p and r. We ran BMF 10 times with random parameter initializations to assess its robustness.

2.8 HTR-SELEX datasets

We obtained 177 HTR-SELEX datasets of 86 distinct factors from (Jolma *et al.*, 2020, Supplementary Table S1). The longer length of oligomers used in this dataset (40 nucleotides) enables the search for co-occurrence of motif pairs with longer spacers. We used sequences of the input library and the last cycle to train BMF. Even though our model describes one cycle of selection, the retrieved motifs were more prominent in the later cycles. Moreover, the cross-platform validation discussed below resulted in slightly better performance for all the tools when choosing the input and last cycles for motif detection in comparison to second and third cycles. Whenever several experimental or technical replicates were available, we built a separate model for each replicate and averaged the corresponding metric over all replicates of an RBP at the end. We used BMF's default hyper-parameters throughout the manuscript.

2.9 Cross-platform validation of *in vitro* motifs

Each experimental technique for measuring RNA binding has its own biases. When measuring the quality of predictions of motif models by cross-validation, methods can learn these biases to distinguish bound from background sequences. Highly parameterized models could learn such subtle, complex biases. These platform-dependent biases can be a result of library preparation, amplification, or can depend on the type and concentration of RNase that is used (Kishore *et al.*, 2011; Orenstein and Shamir, 2014). There have been efforts to reduce the effect of such biases when training motif models, e.g. by learning binding models for many RBPs at the same time (Ghanbari and Ohler, 2020). In order to ensure that BMF does not over-train on the *in vitro* HTR-SELEX data, we performed cross-platform validation: We trained BMF on HTR-SELEX datasets and used the resulting models to predict binding sites in *in vivo* CLIP data.

We collected eCLIP datasets of 15 RBPs (Van Nostrand *et al.*, 2020) and PAR-CLIP datasets of 10 RBPs (Mukherjee *et al.*, 2019) for which we also have HTR-SELEX data. We used the pre-processed CLIP peaks as enriched sequences. Since the PAR-CLIP dataset contained larger numbers of peaks, we restricted our analysis to the top 2000 reported binding sites per RBP. For each eCLIP and PAR-CLIP dataset, we created a background set of the same size by drawing random PAR-CLIP or eCLIP peaks of other factors measured with the same technique. We applied a sliding window with length of 50 and a stride of 20 to generate same-size fragments that fully cover each peak. The prediction scores were averaged over these fragments when the region was longer than 50 bases. We compared our simple model with deep learning approaches, the popular RBP binding predictors iDeepE (Pan and Shen, 2018), DeepCLIP (Grønning *et al.*, 2020) and GraphProt (Maticzka *et al.*, 2014). iDeepE and DeepCLIP use deep learning to predict RBP binding, while GraphProt's model is based on Support Vector Machines (SVMs). We had to use 40 nucleotide fragments for DeepCLIP as it required the same length for training and testing.

2.10 BMF software and web server

The BMF command-line tool offers three commands: (i) learning a BMF model given enriched and background sequences. Output is a BMF model file. (ii) Bipartite motif visualization, given the BMF file learned in step 1. (ii) Predicting binding scores for new sequences with the BMF model trained in the first step. The first two functionalities (*de novo* motif discovery) are also available on the BMF web server.

3 Results

We present BMF, a method for *de novo* discovery of RNA-binding motifs that uses a bipartite motif model capable of learning multivalent binding specificities among RBPs. BMF models the protein binding with up to two domains to its RNA substrate. We assume that due to the structure of the RBP (or RBP complex), the distance between the two binding sites is spatially constrained. BMF therefore consists of two short sequence motif models and a distance probability distribution (Fig. 1A). Binding with just one domain is modeled using a distance distribution peaked at 0 base pairs. In the following sections we demonstrate that this model can reliably detect bipartite motifs in synthetic and real sequences, and we evaluate its performance at identifying binding sites compared to other models of RBP binding in HTR-SELEX, PAR-CLIP and eCLIP datasets.

3.1 BMF accurately discovers implanted synthetic motifs

To test BMF's ability to learn bipartite motifs, we generated 2000 artificial sequences containing first an AAA and then a CCC with a distance distribution of around 3–5 bases between them (Fig. 1B, top). BMF retrieved the implanted motifs and spacer distribution accurately. The results were similarly accurate when sequence degeneracy was introduced by flipping the last base (Fig. 1B, middle), or when implanting the repeat sequence ACACAC (Fig. 1B, bottom). This demonstrates that BMF can not only reveal multivalent specificities but can also recover longer sequence motifs by placing the two cores adjacent to one another.

The log-likelihood increases during stochastic gradient descent and the optimization terminates when the log-likelihood has reached a plateau (Supplementary Fig. S2A). The distance parameters and

the binding energies of *k*-mers in the motif cores all reach a plateau before termination (Supplementary Fig. S2B, C). To test robustness to parameter initialization, we ran BMF ten times with random initial parameter values and verified that the *k*-mer energies and distance parameters match across all runs (Supplementary Fig. S2D, E).

3.2 Most RBPs show multivalent binding, often to multiple occurrences of the same motif

We applied BMF to 177 HTR-SELEX datasets consisting of 86 distinct RBPs to investigate the importance of multivalent binding in the formation of RBP target specificity. BMF detected bipartite binding for many RBPs including ELAVL1, KHDRBS3 and RBPMS (Fig. 2A). Interestingly, BMF restricted the distance of the motif cores strictly to zero when the RBP binds repeat sequences (e.g. CELF1 binding GU repeats) or when the RBP binds a longer RNA sequence that requires a longer motif core (e.g. RBFOX3 binding UGCAUG, and PUM2 binindg UGUANA). The sequence and spacing preferences were also reproducible across experimental replicates (Supplementary Fig. S3), and match for proteins that belong in the same family (Supplementary Fig. S4, Supplementary Table S1). All 177 BMF models with core lengths of 3-5 can be found at BMF's GitHub repository. These results show that BMF can identify bipartite motifs in HTR-SELEX data.

We then looked for the frequency of such multivalent, bipartite motifs and calculated the probability of observing the two core motifs at distances beyond zero for each motif model (Fig. 2B). At two extremes, this probability would be zero for RBPs like RBFOX3, which consist of a larger binding sequence, and one for RBPs like KHDRBS3, which prefer a larger spacer between the motif cores. Interestingly, the majority of RBPs lie at the two extremes, and about half of them show a bipartite binding behavior. This ratio is higher than estimated in previous studies, which were based on *k*-mer counting approaches (Dominguez *et al.*, 2018; Jolma *et al.*, 2020). The number of bipartite motifs could be furthermore underestimated as some RBPs show bipartite binding only when BMF's core size is increased to four or five nucleotides (Supplementary Fig. S5). Overall, these results highlight the importance of multivalent binding as a common strategy to achieve high specificity despite having individually small and weak binding sites.

We noted that many motif models (like ELAVL1 and KHDRBS3) have similar sequence preferences on both cores. We

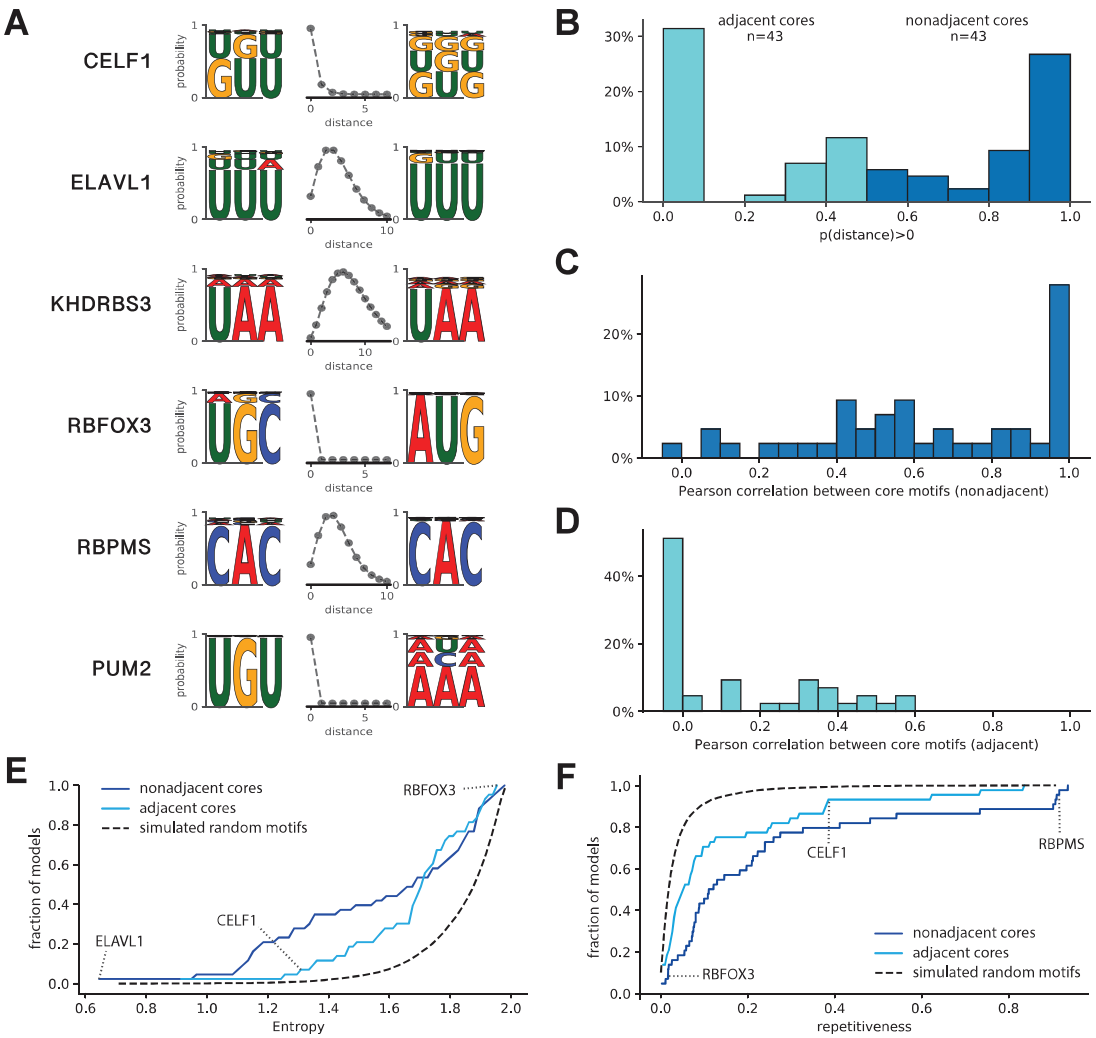

Fig. 2. Many RBPs are multivalent, bind low-complexity sequences and often bind two similar motif cores. (**A**) Examples of motifs that represent a wide range of binding modes, learned by BMF on HTR-SELEX data. When the RBP has a larger motif than allowed by the core size (3 here), the distance between cores is learned to be zero to accommodate a longer binding sequence (e.g. CELF1, RBFOX3 and PUM2). (**B**) Distribution of the probability of the spacer length between the two motif cores to be above 0. As seen in the examples in A, most RBPs either clearly bind adjacent cores (distance= 0, turquoise) or have a multivalent binding mode with two non-adjacent cores (dark blue). (**C**) and (**D**) Similarities between binding preferences of the two cores for RBPs with adjacent cores (turquoise) or multivalent non-adjacent cores (dark blue), according to panel B. (**E**) Cumulative distribution of the entropy of BMF models for all RBPs in the HTR-SELEX dataset. In general the optimized bipartite motif models have much lower complexity than randomly generated bipartite models (dashed black line). (**F**) Cumulative distribution of '*sequence repetitiveness*' of BMF models for all RBPs in the HTR-SELEX dataset. Overall, BMF models are more often repetitive that those of randomly generated bipartite models (dashed black line)

quantified their similarity by the Pearson correlation between the probabilities of observing each of the 4^k k-mers. As expected from the individual examples, the core motifs are mostly similar for RBPs that exhibit bipartite binding (Fig. 2C) as opposed to adjacent motif cores (Fig. 2D). This demonstrates that RBPs have often evolved to bind multiple occurrences of the same or similar short sequence motifs, either using multiple same-chain RBDs or by homodimerization and oligomerization.

3.3 RBPs often bind low-complexity and repetitive sequences

It has been shown that RBPs bind sequences of lower complexity than DNA-binding transcription factors (Dominguez *et al.*, 2018; Singh and Valcárcel, 2005). This can be seen at its extreme for some of our binding models, which are composed of only one to two types of nucleotides (Fig. 2A). Looking at all 78 RBP binding models, we observed that many proteins bind repetitive sequences or have the same simple k-mer affinities for each of their valencies. In order to quantify this, we calculated the entropy of the motif sequences as a measure of sequence complexity (Fig. 2E, Supplementary Methods) (Dominguez *et al.*, 2018). For highly complex sequence affinities (e.g. RBFOX3), the entropy gets close to two, while this value is closer to zero for degenerate and repetitive sequences (e.g. ELAVL1). A similar trend is visible when quantifying the repetitiveness of BMF models, resulting in high scores when both cores consist of mono- or di-nucleotide repeats (Fig. 2F, Supplementary Methods). Overall, more than half of RBP motifs show levels of degeneracy that are highly unlikely in artificially generated random motif models. This binding preference toward low complexity sequences fits to the previous observation that bipartite motifs tend to bind multiple occurrences of the same sequence.

3.4 Including all binding configurations and cooperativity enhances the accuracy of RBP binding predictions

To assess the value of cooperativity and multivalency, we compared BMF to a spaced k-mer motif model which scores the sequences by finding the best binding site (Supplementary Fig. S6, Supplementary Methods). Interestingly, for all RBPs but particularly for those that show bipartite binding, BMF's performance is superior to that of the k-mer enrichment model. This highlights the value of two distinct BMF features: considering all binding configurations, and including the cooperative effect of multi-domain binding.

3.5 *In vitro* bipartite models learned by BMF can predict *in vivo* binding

Experimental techniques for measuring RNA binding have individual biases that can be learned by motif discovery tools. This is particularly problematic when evaluating computational methods with many model parameters that can capture complex structures in their input datasets (Ghanbari and Ohler, 2020). Cross-platform validation, i.e. using binding models trained on an experimental dataset to predicting binding sites in another experimental platform ensures a fair assessment of the quality of motif models. We therefore trained models on HTR-SELEX data to predict binding sites on sequences derived from PAR-CLIP and eCLIP experiments (Mukherjee *et al.*, 2019; Van Nostrand *et al.*, 2020). We compared the performance of BMF to iDeepE (Pan and Shen, 2018), DeepCLIP (Grønning *et al.*, 2020) and GraphProt (Maticzka *et al.*, 2014) (Fig. 3A–E). iDeepE and DeepCLIP are deep learning tools and GraphProt is based on support vector machines. Thanks to their more complex architecture and higher number of parameters, these models are able to learn more complex aspects of the training data,

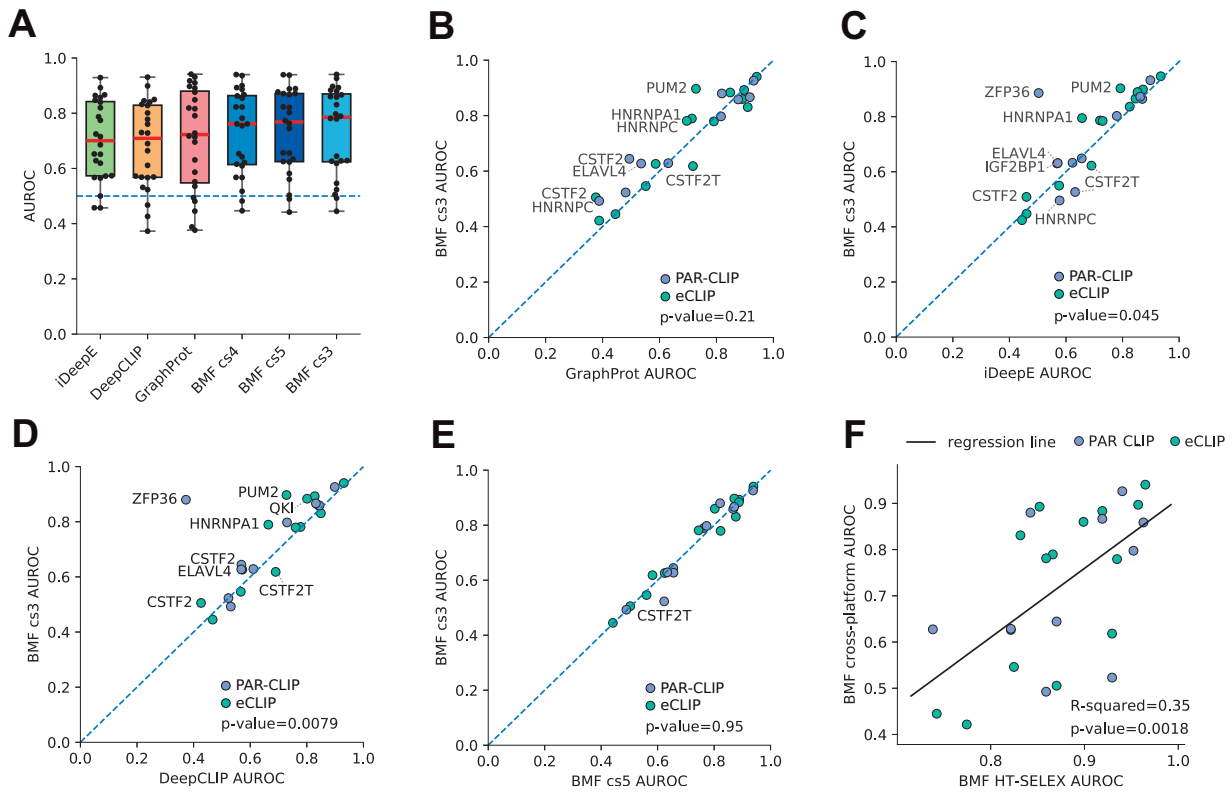

Fig. 3. Cross-platform validation shows *in vitro* BMF motifs can predict *in vivo* binding sites in transcriptomes. We used BMF, iDeepE, DeepCLIP and GraphProt to identify eCLIP and PAR-CLIP RBP binding sites after training their motif models on HTR-SELEX datasets. (**A**) AUROC distribution for iDeepE, DeepCLIP, GraphProt and BMF with motif sizes ranging from 3 to 5. The tools are sorted based on their median AUROC performance. The values for each RBP dataset is shown with a black dot. (**B–E**) AUROC from BMF (core size 3) compared to GraphProt, iDeepE, DeepCLIP and BMF with core size 5. Statistical significance was assessed through Wilcoxon signed-rank tests. (**F**) BMF AUROC values from cross-validated HTR-SELEX analysis correlate with cross-platform benchmark performance. Both BMF models are built with core size 3. Linear regression line is shown in black. In all plots AUROC values are averaged over all replicate combinations wherever replicates were available

while GraphProt additionally takes the RNA structure as an input. Interestingly, despite these advantages, BMF showed a competitive prediction quality as measured by the area under the receiver operating characteristic curve (AUROC), with a better median AUROC than iDeepE, DeepCLIP and GraphProt. Simlar results are obtained when replacing AUROC with the area under the precision recall curve (AURPC, Supplementary Fig. S7).

Interestingly, generally performance of BMF is best for $k = 3$, although it changes little between core size of $k = 3$, 4 or 5 (Fig. 3A, Supplementary Fig. S8). For some RBPs increasing the core size reduced the predictive power for the resulting models. This could be due to over-fitting on biases of the HTR-SELEX data and might be a reason for why the more highly parameterized RNA motif models of GraphProt, DeepCLIP and iDeepE often do not perform as well as the simpler ones of BMF. On the other hand, longer BMF models, as well as other tools in the benchmark, could better learn binding preferences for factors such as CSTF2T that bind more complex RNA sequences. To summarize, BMF can capture RBP specificities with reduced risk of overfitting.

A comparable trend emerges when predicting bound RNAcompete sequences from Ray *et al.* (2013) using HTR-SELEX models. Again, the median performance is best for Graphprot and BMF, followed by DeepCLIP and iCLIP (Supplementary Fig. S9). Larger BMF models perform slightly better than smaller ones, perhaps because these *in vitro* assays enrich for the best RNA binding sequences and therefore might yield motifs of higher information content than those relevant in the cell.

To see whether the core spacing of HTR-SELEX motif models exist in *in vivo* data, we trained BMF models on the CLIP data and compared them to their *in vitro* counterparts. Interestingly for the models that were learned well on the HTR-SELEX data (tool-averaged AUROC ≥ 0.75), both the motif core sequences and their distance distribution match between the two experimental platforms (Fig. 4). The sequence and/or spacer length preferences do vary for other factors with lower AUROC values (Supplementary Figs S10 and S11).

A comparison of the AUROC values from the cross-validated HTR-SELEX data (Supplementary Fig. S6) and those from the cross-platform validation shows a correlation between BMF motif quality and its performance in the cross-platform benchmark (P-value $= 0.0018$, Fig. 3F). It could help explain why some HTR-SELEX models fail at predicting binding to new sequences, possibly as they have little sequence preference for their target RNA or due to the absence of this information in the HTR-SELEX data. Overall, this shows that BMF can be used to learn RNA motifs from *in vitro* data to predict binding sites of the protein in the cell despite numerous factors confounding binding *in vivo*.

4 Discussion

We present BMF, the first bipartite motif model to describe multivalent binding preferences in RBPs. The motif models learned on *in vivo* and *in vitro* datasets imply the following multipartite binding strategy is common—adapted by about half of RBPs in our datasets—to bind their target RNA molecules: First, these RBPs bind multiple short (3–4 nucleotides) RNA segments simultaneously and cooperatively with their multiple RBDs, which can be either on a single chain or part of dimer or oligomer complexes (Lunde *et al.*, 2007; Wang *et al.*, 2002). Second, the recognition motifs of their single RBDs are usually similar (Fig. 2). These two aspects make it simple to evolve the sequence features in the target RNAs required for highly specific cooperative binding: a sufficient density of the simple core recognition motifs. We have recently shown that the RBP binding affinity through cooperative binding of multiple RBDs depends on the motif density on the target RNA with a Hill-like coefficient that is similar in size to the number of binding domains (Stitzinger *et al.*, 2021, Fig. 4D). Via di- and oligomerization of RBPs the number of cooperatively binding domains and thereby the Hill-like coefficient can be further increased, by which it is possible to distinguish between targets with, say, a core binding motif every 20 versus every 30 nucleotides (e.g Schulz *et al.*, 2013, Fig. 3E, F).

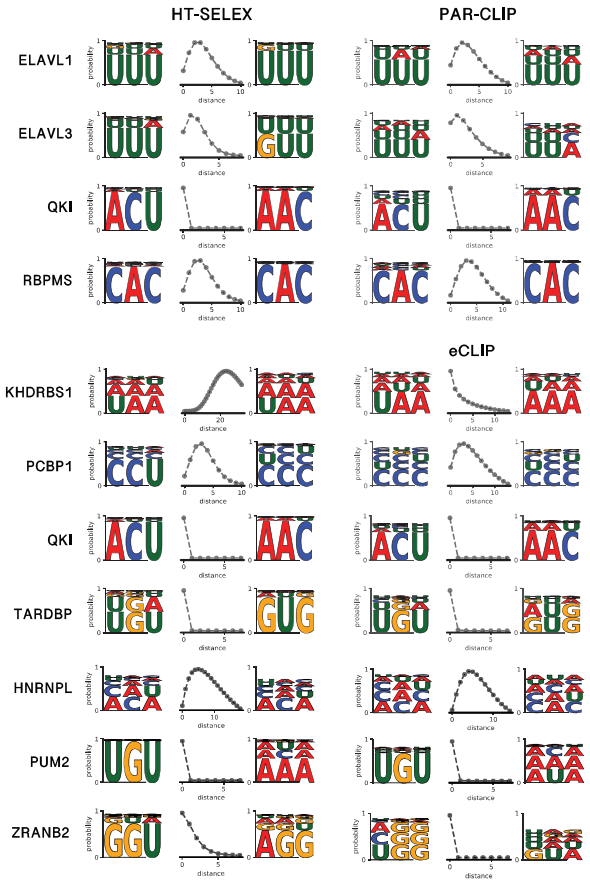

Fig. 4. Bipartite motif models learned on *in vitro* data match their *in vivo* counterparts. Bipartite motifs are shown for those RBPs in Figure 3 whose best replicate has a tool-averaged AUROC of at least 0.75. The models learned *in vitro* and *in vivo* match not only in the sequence preference but also the relative positioning of the two motif cores, with the exception of KHDRBS1, which shows a bipartite motif only in the HTR-SELEX data

An encoding of binding affinity via the density of motifs makes sense for the many RNA-binding proteins for which the precise binding sites on their target RNAs is not important to perform their function.

Mono- and dinucleotide repeats are particularly attractive as target motifs because they possess one binding site per position and per two positions, respectively. The high density of motifs gives rise to high affinities through the combinatorially many possible binding configurations of two or more RBDs. BMF takes full account of this combinatorial complexity.

A limitation of the evolutionary strategy to bind low-complexity sequences using multiple domains with near-identical motifs is the much smaller number of motifs than can be distinguished, only 64 for length-3 cores. This low number might be sufficient, however, for targeting such RBPs to their RNA targets because specificity is enhanced by compartmentalization—an RBP occurring only in the nucleus cannot bind to cytosolic mRNAs, for example. Furthermore, only a fraction of RBPs is expressed in any one cell type at any one time, in a similar way as the many transcription factors having the same binding affinities are usually expressed in different cells or at different times.

Our results agree with previous studies that reported bipartite motifs in HTR-SELEX and RBNS datasets by counting spaced k-mers of various linker lengths (Dominguez *et al.*, 2018; Jolma *et al.*, 2020). The motifs we report are congruent with those reported before and additionally provide a distance distribution to describe the best binding geometry. The observation that motifs are repetitive

and degenerate is also consistent with previous high-throughput studies (Dominguez *et al.*, 2018).

Interestingly, BMF motifs were shorter and less complex than those reported by Jolma *et al.*, 2020. For RBPs for which Jolma et al obtained long motifs (i.e. PCBP1, PUM1 and TARDBP), longer motif cores than 3 nucleotides in BMF could not improve prediction performance in the cross-platform benchmark. This indicates that 3-6 base long motifs would suffice in explaining the sequence specificities for the majority of RBPs.

BMF does not take RNA secondary structure into account, both the change of total energy upon binding by modifying or breaking secondary and tertiary structure interactions and their associated entropy changes. It has been shown that some RBPs at least partially identify their target RNA molecules through binding specific structural elements (Jones *et al.*, 2001; Mackereth and Sattler, 2012). This could further narrow the search space of proteins to fewer potential binding partners and open new ways for cellular regulation. Despite ignoring structure, BMF's performance is comparable if not better than GraphProt, a tool that includes detailed modeling of secondary structure. We expect that expanding our bipartite motif model to include RNA structure could further improve its predictive power. We also simplify BMF by assuming that proteins bind as constitutive complexes. This does not describe well proteins that interact more weakly and, for example, oligomerize only upon binding to nearby RNA segments.

Overall, BMF's performance is promising in the following regards: Owing to its multi-domain binding model BMF can (i) find pairs of sequence motifs over-represented in a sequence set, and can (ii) learn the distance between the motif pairs, reflecting the best binding configurations. This information can be further used to (iii) asses whether or not an RBP displays bipartite binding. We believe that looking at RNA motifs as combinations of individual low-affinity interactions can improve our understanding of RNA regulation in the cell and shed a new light on how some RBPs can find their targets despite the weak sequence and structural preferences of individual domains.

Financial Support: The authors acknowledge support by the focus program SPP2191 of the Deutsche Forschungsgemeinschaft.

Conflict of Interest: none declared.

Code and data availability

The HTR-SELEX data of Jolma *et al.* (2020) were downloaded from the European Nucleotide Archive under accession PRJEB25907 (https://www.ebi.ac.uk/ena/browser). The preprocssed eCLIP datasets were collected from the ENCODE at https://www.encodeproject.org (Van Nostrand *et al.*, 2020). PAR-CLIP peaks were obtained from https://github.com/BIMSBbioinfo/RCAS_meta-analysis (Mukherjee *et al.*, 2019). BMF source code, documentation and motif models can be found at https://github.com/soedinglab/bipartite_motif_finder.

Acknowledgements

The authors thank Christian Roth for his help on AVX-optimization, web server implementation and for discussions on tool development and benchmark design. They thank Jilin Zhang for feedback on the manuscripts. SSJ acknowledges support from the International Research School for Molecular Biology (IMPRS-MolBio).

References

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Änkö,M.-L. and Neugebauer,K.M. (2012) RNA–protein interactions in vivo: global gets specific. *Trends Biochem. Sci.*, **37**, 255–262.

Calabretta,S. and Richard,S. (2015) Emerging roles of disordered sequences in RNA-binding proteins. *Trends Biochem Sci.*, **40**, 662–672.

Cook,K.B. *et al.* (2017) RNAcompete-S: combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. *Methods*, **126**, 18–28.

Dominguez,D. *et al.* (2018) Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell*, **70**, 854–867.

Forties,R.A. and Bundschuh,R. (2010) Modeling the interplay of single--stranded binding proteins and nucleic acid secondary structure. *Bioinformatics*, **26**, 61–67.

Gerstberger,S. *et al.* (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.

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

Gilbert,C. and Svejstrup,J.Q. (2006) RNA immunoprecipitation for determining RNA–protein associations in vivo. *Curr. Prot. Mol. Biol.*, **75**, 24.

Grønning,A.G.B. *et al.* (2020) DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. *Nucleic Acids Res.*, **48**, 7099–7118.

Hafner,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.

Hartmann,H. *et al.* (2013) P-value-based regulatory motif discovery using positional weight matrices. *Genome Res.*, **23**, 181–194.

Hentze,M.W. *et al.* (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, **19**, 327–341.

Jolma,A. *et al.* (2020) Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res.*, **30**, 962–973.

Jones,S. *et al.* (2001) Protein–RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.

Kazan,H. *et al.* (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.

Kingma,D.P. and Ba,J. (2014) Adam: a method for stochastic optimization. *arXiv*, 6980.

Kishore,S. *et al.* (2011) A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nat. Methods*, **8**, 559–564.

König,J. *et al.* (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.

Koo,P.K. *et al.* (2020) Global importance analysis: a method to quantify importance of genomic features in deep neural networks. *bioRxiv*, 288068.

Lambert,N. *et al.* (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell*, **54**, 887–900.

Li,X. *et al.* (2014) Finding the target sites of RNA-binding proteins. *Wiley Interdisc. Rev. RNA*, **5**, 111–130.

Lunde,B.M. *et al.* (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.

Mackereth,C.D. and Sattler,M. (2012) Dynamics in multi-domain protein recognition of RNA. *Curr. Opin. Struct. Biol.*, **22**, 287–296.

Maticzka,D. *et al.* (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, **15**, R17–18.

Mukherjee,N. *et al.* (2019) Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res.*, **47**, 570–581.

Munteanu,A. *et al.* (2018) SSMART: sequence-structure motif identification for RNA-binding proteins. *Bioinformatics*, **34**, 3990–3998.

Orenstein,Y. and Shamir,R. (2014) A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res.*, **42**, e63.

Ozdilek,B.A. *et al.* (2017) Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Res.*, **45**, 7984–7996.

Pan,X. and Shen,H.-B. (2018) Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, **34**, 3427–3436.

Quinn,T.P. *et al.* (2020) Learning distance-dependent motif interactions: an interpretable CNN model of genomic events. *bioRxiv*, 270967.

Ray,D. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.

Rubinstein,M. *et al.* (2003) *Polymer Physics*, **Vol. 23**. Oxford University Press, New York.

Schneider,T. *et al.* (2019) Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein imp3. *Nat. Commun.*, **10**, 1–18.

Schulz,D. *et al.* (2013) Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell*, **155**, 1075–1087.

Siebert,M. and Söding,J. (2016) Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, **44**, 6055–6069.

Singh,G. *et al.* (2015) The clothes make the mRNA: past and present trends in mrnp fashion. *Ann. Rev. Biochem.*, **84**, 325–354.
Singh,R. and Valcárcel,J. (2005) Building specificity with nonspecific RNA-binding proteins. *Nat. Struct. Mol. Biol.*, **12**, 645–653.
Stitzinger,S.H. *et al.* (2021) Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains. *bioRxiv*, 428308.
Stražar,M. *et al.* (2016) Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, **32**, 1527–1535.
Toivonen,J. *et al.* (2018) Modular discovery of monomeric and dimeric transcription factor binding motifs for large data sets. *Nucleic Acids Res.*, **46**, e44.
Toivonen,J. *et al.* (2020) Moder2: first-order Markov modeling and discovery of monomeric and dimeric binding motifs. *Bioinformatics*, **36**, 2690–2696.
Van Nostrand,E.L. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
Van Nostrand,E.L. *et al.* (2020) Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol.*, **21**, 1–26.
Wang,X. *et al.* (2002) Modular recognition of RNA by a human pumilio-homology domain. *Cell*, **110**, 501–512.
Yan,J. and Zhu,M. (2020) A review about RNA–protein-binding sites prediction based on deep learning. *IEEE Access*, **8**, 150929–150944.
Yang,X. *et al.* (2013) The dataset for protein–RNA binding affinity. *Protein Sci.*, **22**, 1808–1811.