

 Open access • Posted Content • DOI:10.1101/2021.06.07.447370

RBPSpot: Learning on Appropriate Contextual Information for RBP Binding Sites Discovery — [Source link](#)

[Nitesh Kumar Sharma](#), [Nitesh Kumar Sharma](#), [Sagar Gupta](#), [Prakash Kumar](#) ...+8 more authors

Institutions: [Council of Scientific and Industrial Research](#), [Academy of Scientific and Innovative Research](#), [Indian Agricultural Statistics Research Institute](#)

Published on: 07 Jun 2021 - [bioRxiv](#) (Cold Spring Harbor Laboratory)

Topics: [Context \(language use\)](#)

Related papers:

- [RBPSpot: Learning on appropriate contextual information for RBP binding sites discovery.](#)
- [Deep neural networks for interpreting RNA binding protein target preferences](#)
- [RNAProt: an efficient and feature-rich RNA binding protein binding site predictor.](#)
- [beRBP: binding estimation for human RNA-binding proteins](#)
- [GraphProt2: A graph neural network-based method for predicting binding sites of RNA-binding proteins](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/rbpspot-learning-on-appropriate-contextual-information-for-14y5smxb0k>

1 **RBPSpot: Learning on Appropriate Contextual Information for RBP Binding**

2 **Sites Discovery**

3 **Nitesh Kumar Sharma^{1,2}, Sagar Gupta¹, Prakash Kumar^{1,2,3}, Ashwani Kumar¹, Upendra**
4 **Kumar Pradhan^{1,2,3}, Ravi Shankar*^{1,2}**

5

6

7

8 ¹ Studio of Computational Biology & Bioinformatics,

9 CSIR-Institute of Himalayan Bioresource Technology (CSIR-IHBT),

10 Palampur (HP), 176061, India.

11

12 ²Academy of Scientific and Innovative Research (AcSIR),

13 Ghaziabad, Uttar Pradesh- 201 002

14

15 ³ICAR-Indian Agricultural Statistics Research Institute

16 Library Avenue, Pusa, New Delhi, Delhi 110012

17

18

19

20 ***Corresponding Author: ravish@ihbt.res.in**

21

22

23

24

25

26

27 **Abstract**

28 Identifying RBP binding sites and mechanistic factors determining the interactions remain a big
29 challenge. Besides the sparse binding motifs across the RNAs, it also requires a suitable sequence
30 context for binding. The present work describes an approach to detect RBP binding sites while
31 using an ultra-fast BWT/FM-indexing coupled inexact k-mer spectrum search for statistically
32 significant seeds. The seed works as an anchor to evaluate the context and binding potential using
33 flanking region information while leveraging from Deep Feed-forward Neural Network (DNN).
34 Contextual features based on pentamers/dinucleotides which also capture shape and structure
35 properties appeared critical. Contextual CG distribution pattern appeared important. The developed
36 models also got support from MD-simulation studies and the implemented software, RBPSpot,
37 scored consistently high for the considered performance metrics including average accuracy of
38 ~90% across a large number of validated datasets while maintaining consistency. It clearly
39 outperformed some recently developed tools, including some with much complex deep-learning
40 models, during a highly comprehensive bench-marking process involving three different data-sets
41 and more than 50 RBPs. RBPSpot, has been made freely available, covering most of the human
42 RBPs for which sufficient CLIP-seq data is available (131 RBPs). Besides identifying RBP binding
43 spots across RNAs in human system, it can also be used to build new models by user provided data
44 for any species and any RBP, making it a valuable resource in the area of regulatory system studies.

45

46

47

48

49

50

51

52 **Introduction**

53

54 It has been reported that at any given time, compared to just 2-3% transcription factors expression
55 share, ~10 times higher volume of RNA binding proteins are expressed (1). Advances with high-
56 throughput techniques like CLIP-seq and Interactome Capture have drastically revised our
57 understanding about RBPs which suggest that human systems are expected to have at least 1,500-
58 2,000 genes coding for RBPs (1,2). Unfortunately, we are still far behind in terms of information for
59 these regulators where hardly ~150 RBPs have been studied so far for their interactions with RNAs.
60 Despite of their critical functional roles in cell systems, very few RBPs have been explored with
61 precise identification of their mechanism of action (1).

62

63 There are certain limitations with these high-throughput experiments. These experiments are costly.
64 They too don't give the entire RBP-RNA interactome spectrum and at a time work for one RBP only
65 in condition specific manner. The CLIP-seq reads provide narrowed down regions to look for
66 interactions but don't provide the mechanistic details and explanations for the interactions. Using
67 general motif discovery tools to identify the interaction spots have got limited success in case of
68 RBPs as they either report too short motifs which have high chances of occurrences across the
69 random data or they don't cover large spectrum of instances. Unlike transcription factors, RBPs
70 binding sites display sparse motif positional conservation. They are usually difficult to detect
71 through such routine motif finding approaches. Besides the binding motifs, contextual sequence
72 environment also guide the RBP-RNA interactions, adding further complexity to the process of
73 discovery of the actual interaction spots. Therefore, this is an area which needs prime focus on
74 deriving the principles of RBP-RNA interactions and their impact of regulation once we have

78 enough CLIP-seq data. One of the most remarkable work, RNAcompete , was done where the
79 authors identified *in-vivo* motifs for 207 different RBPs using pools of 30-41 bases long RNA
80 oligos to which affinity of various RBPs was assessed for binding (3). RNAcompete also
81 highlighted how conventional motif finding tools fail to discover the binding sites motif for RBPs.
82 At computational front some decent progress has been made in dealing with these CLIP-seq data to
83 derive the models for interactions. Initially, to explore the RBPs and their RNA binding sites,
84 databases like RBPDB, CLIPZ, CLIPdb/POSTAR came up (4-7). These databases provided first
85 structured information on RBP-RNA interactions as well as proposed their interaction motifs using
86 traditional motif finding tools while building on publicly available experimental data. As already
87 mentioned above, the motifs being used here are short and occur in abundance even in random data.
88 Also, they don't consider contextual information. Identification of correct RBP:RNA interaction
89 motifs is a critical step which helps in locating the appropriate contextual information to build an
90 accurate model of RBP:RNA interactions.

91
92 RNAcontext is among those first such tools which considered contextual information for RBP-RNA
93 interaction discovery. It applied the structural preferences information for these RNAcompete
94 motifs using *ab-initio* RNA structure prediction tool, sfold (8). However, these *ab-initio* structural
95 prediction methods reliability falls down with the length, making the structural information derived
96 through them not reliable enough (9). The next important stride came with probabilistic tools like
97 RBPmap which extended their previous approach to identify splice sites while applying user
98 provided position specific scoring matrices, supported motif clusters, and phylogenetic conservation
99 to identify RBP RNA interaction spots (10). In the same probabilistic tools category, mCarts was
100 another important addition (11). It works on the similar lines to RBPmap but also applies 6-states
101 Hidden Markov Model (HMM) along with structural information from *ab-initio* secondary structure
102 prediction methods to predicted functional RBP binding sites.

103

104 With Graphprot a new generation of such tools started which applied machine learning as well as
105 leveraged from new data-sets developed from CLIP-seq experiments (12). It also applied the
106 concept of differential RNA secondary structure information in contextual manner to build the
107 interaction models. A recently develop tool, beRBP, carries forward the approach similar to
108 RBPmap while implementing a machine-learning method of Random-Forest (13). It clusters the
109 potential motif sites where it ranks them and uses the highest scoring regions for the matches in the
110 given region while scanning for the user provided motif/PWM. In the followup, they have also
111 applied an approach similar to RNAcontext where RNA structural information is provided for the
112 motif region using *ab-initio* structure prediction tool, RNAfold. Further to this, it added the
113 phylogenetic conservation information similar to RBPmap and mCarts.

114

115 With recent developments in the area of deep learning, many deep-learning based RBP-RNA
116 interaction detection approaches have been implemented recently. DeepBind deserves special
117 mention among them as it pioneered this category where a robust general system was created to
118 model nucleic acids and protein interactions using convolution neural network (CNN) (14).
119 DeepBind has become a sort of prototype for almost all of the recent Deep-learning based tools to
120 identify the RBP-RNA interactions. DeepBind applies 7-mer motif weight matrices are
121 transformation into an image pixel matrix and is scanned for entire sequence while evaluating for 4-
122 stages to derive the binding score: convolution stage, rectification stage which zooms the scanner to
123 most promising regions for the motif, followed by pooling of all such regions and expansion and
124 clustering of motifs, which is finally subjected to a non-linear classifier. However, the authors
125 accepted that compared to transcription factors and their data, running DeepBind with RNAcompete
126 data did not achieve that level of accuracy. They pointed out the importance of accurate RNA
127 secondary structure information and RNA shape readouts in RNA-RBP interactions which most of

128 the approaches have missed so far. Taking the work further on Deep-learning based RBP-RNA
129 interaction detection, another prominent tool system is iDEEP which has come like a series of
130 softwares like iDeep, iDeepS, and iDeepE (15-17). These tools differ from each other for the way
131 they applied various combinations of CNN and RNN layers. iDeepS applied CNN with Long-Short
132 term memory (LSTM) while taking input from sequence and RNAs shape data. iDeepE applies
133 combinations of CNNs which capture local and global sequence properties. A recently developed
134 tool, DeepRiPe, has evolved a CNN and GRU based deep-learning approach while also introducing
135 transcript's regions specific information like splice junctions etc (18). DeepCLIP is another recently
136 developed tool which detects RBP-RNA interaction spots while applying CNN in combination with
137 bidirectional-LSTM and claims to detect sequence position specific importance which could
138 determine the contribution of various nucleotides in RBP binding (19). These very recently
139 developed deep-learning approaches have become much more complex than DeepBind and claim to
140 achieve much higher accuracy. Their complexity comes from adding complex layers above the
141 regular dense hidden layer. These complex layers actually do the job of automatic feature extraction
142 unlike the other machine-learning approaches where expert knowledge is applied to identify the
143 important properties to look into for feature extraction.

144

145 While reviewing these developments and tools, it looked imminent that there is an enormous scope
146 of improvement in the approaches to find and locate RBP-RNA interaction spots. Some of the major
147 points to consider would be: 1) Choice of datasets: A notable issue with all these algorithms is the
148 choice of data-sets, especially the negative data-sets, which have mostly been too relaxed and
149 unrealistic, due to which these tools are prone to over-fitting and imbalance. They are either
150 randomly shuffled sequences or regions randomly selected from those RNAs which did not bind the
151 given RBP. 2) Motif searching approach: most of existing tools, with exception of recent deep-
152 learning based approaches, begin with predefined/user defined motif or PWM derived from

153 traditional motif finding tools with user defined length, which is not a natural approach and one of
154 the prime mistakes. RBP binding sites display sparse conservation which regular motif discovery
155 tools may fail to capture sufficiently. Third, high dependence on *ab-initio* RNA structure prediction
156 tools to derive the structural and accessibility information may be misleading, as already pointed
157 out above, such tools don't provide correct information on actual complete RNA length. A better
158 approach has been consideration of dinucleotide densities for such purpose (20,21). Consideration
159 of RNA-shape appears very much important as pointed out by DeepBind as well as some other
160 recent works (14,22,23). It has been reported that pentamers capture the essence of nucleic acid's
161 shape accurately (24), making them a suitable candidate to be evaluated along with dinucleotide
162 densities to derive RNA structure and shape information. Fourth, though the recent deep-learning
163 approach claim good success through automation of the process of feature extraction at the cost of
164 added complexity, the effectiveness of such automated feature detection needs to be evaluated.
165 Simpler models, if trained with carefully selected properties, are capable to outperform complex
166 models. This is why some of the shallow learning methods have outperformed deep-learning
167 methods on structured data (25, [https://towardsdatascience.com/the-unreasonable-ineffectiveness-](https://towardsdatascience.com/the-unreasonable-ineffectiveness-of-deep-learning-on-tabular-data-fd784ea29c33)
168 [of-deep-learning-on-tabular-data-fd784ea29c33](https://towardsdatascience.com/the-unreasonable-ineffectiveness-of-deep-learning-on-tabular-data-fd784ea29c33)) .

169

170 Considering these all factors, here we present a reliable Deep Neural Net (DNN) based approach to
171 build the mechanistic models of RBP-RNA interactions using high-throughput cross-linking data
172 while considering data from 99 experiments and for 137 RBPs for human system. An ultrafast k-
173 mer spectrum search approach was used to identify the most important seed regions in the sequence
174 for contextual information derivation. Contextual information for 75 bases flanking regions around
175 the identified seed derived motif was extracted in the form of variable windowed position specific
176 dinucleotide, pentamers, and heptamers density based propensities. The combined contextual
177 information was provided to a two hidden layers based dense feed-forward networks to accurately

178 identify the RBP binding spots in RNAs. The developed models were used to identify the
179 interaction spots and scored very high accuracy with remarkable balance between sensitivity and
180 specificity as well as performance consistency when tested across a large number and different
181 types of experimental datasets. Molecular dynamics studies also supported these models. The
182 developed approach has been implemented as a freely available webserver and standalone software,
183 RBPSpot. It was comprehensively bench-marked across three totally unbiased standardized data-
184 sets for performance along with five recently published tools, including more complex deep-
185 learning based tools, where it outperformed all of them consistently across all these datasets for
186 most of the studied RBPs. Unlike most of the existing software which don't provide the option to
187 build new models from data, RBPSpot approach can be applied to detect human system RBP-RNA
188 interactions with its inbuilt models as well as it can be used to develop new models for other species
189 and new RBP data also.

190

191 **Materials & Methods**

192 **Data retrieval and processing**

193 The study has considered human RBP models while using high-throughput sequencing data from
194 cross-linking experiments using various CLIP-seq techniques like CLASH, dCLIP, eCLIP, FLASH-
195 CLIP-seq, HITS-CLIP, iCLIP, PAR-CLIP, sCLIP-seq, uvCLAP-CLIP-seq. This data also includes
196 the two cell lines eCLIP data from ENCODE. Most of them are processed peak data collected for
197 137 RBPs with starBase 2.0 as their primary source (26). A total of 872Mb peak data from 99
198 experiments were covered in this study for RBP-RNA interaction information from CLIP-seq
199 experiments (Supplementary Data 1 Sheet 1). The peak data of RBPs were downloaded in the form
200 of co-ordinates along with their associated RNA information on which they were binding. Peak data
201 were converted into BED file format along with their strand specificity information. Genome
202 sequences of human hg19 builds were obtained from UCSC browser. Peak data were also refined

203 based on the length distribution and peaks laying in extreme range (length >300 bases and <5 bases)
204 were omitted from the study (Supplementary Data 1 Sheet 2).

205

206 **Identification of motif seed candidates: k-mer spectrum search using BWT/FM-Indexing**

207 To search binding sites motifs/seeds for any particular RBP, all the peak regions were transformed
208 into overlapping lists of k-mers of size six to start with. Iteratively and in parallel these generated k-
209 mer spectrum for each such sequence was searched across all the reported cross-linked associated
210 regions in the targets to obtain the enrichment status of the k-mers (seeds) on which motif would be
211 built. These searches were allowed with maximum 30% mismatches. Since normal search would be
212 heavily time consuming step, we implemented an enhanced Burrow-Wheeler transformation with
213 FM-Indexing to search with any number of mismatch which made the search ultra-fast for even in-
214 exact searches. The detailed algorithmic implementation pseudo-code of the implemented algorithm
215 is given in the supplementary methods.

216

217 **Identification of motif seeds candidates: Anchoring with the significant seeds**

218 All the k-mer seeds and their relatives displaying at least 70% similarity were evaluated for
219 existence across at least 70% of peak data. Such motif seed candidates were further evaluated for
220 their statistical significance. Those RBPs where no k-mer and their relatives crossed 70%
221 representation were looked for the highest representation available. The remaining data which did
222 not show the representative k-mer were checked further and recursively with minimum cut-off of
223 20% data representation. Motifs coming from such data were considered as mutually exclusive one.
224 Null model distribution probabilities of occurrence of each k-mer along with its relatives was
225 calculated from the random data set to find their random probabilities. Random data set was
226 generated from unassociated RNAs while randomly carving out the lengths similar to the peak data.

227 Significantly over-represented k-mers were screened using binomial test with p-value cut-off of
228 0.01. These significantly enriched k-mers were used as initial seeds to develop the final motif.
229 These seeds of significantly enriched k-mers were expanded in both the directions by expanding by
230 one nucleotide both sides, followed by search across the peak data with at least 70% occurrence in
231 the peak data while repeating the above mentioned search operation recursively. Expansion of seed
232 region in both directions was allowed till at least 70% match existed. Final motifs were selected on
233 the basis of satisfying both the criteria i.e. the motif displayed least 70% abundance across the
234 CLIP-seq instances at 1% significance level and the maximum k-mer expansion maintained at least
235 70% identity with the associated sequences and relatives. Mutually exclusive motifs were other
236 predominant motifs which existed in the remaining data which were scanned in similar recursive
237 manner as described above. Figure 1. shows the part of the k-mer based motif seed discovery and
238 steps taken afterwards. (Supplementary Data 1 Sheet 3,4)

239

240 **Datasets creation**

241 Once we had prime motifs anchored for each RBP from the given data, their associated peak data
242 sequences were converted into positive datasets. To generate positive datasets for each RBP, start
243 and end co-ordinates from the main motif's both terminals were expanded by +75 and -75 bases
244 into both the directions. In case of multiple motif locations originating from a single peak for the
245 main motif, all the locations were expanded. Different length dataset sequences formed for different
246 RBPs which depended mainly upon the length of the core motif region. However, for any single
247 RBP all the sequences of the dataset were of same length.

248

249 To generate the negative datasets for each RBP, similar condition corresponding RNA-seq data were
250 downloaded from GEO. With minimum three replicates of RNA-seq data expression of each RNA

251 was calculated. Only those transcript sequences were considered which had expression condition
252 available for the same condition but did not bind the RNA or which was not found present in the
253 corresponding condition's CLIP-seq binding data. Associated main motifs for the RBP were
254 searched across these RNA sets also just in the similar manner as was done to the positive dataset
255 instances. Locations of the main motif were reported in the form of start and end co-ordinates from
256 where further expansion of +75 and -75 bases was done on both the sides. This way very strong
257 negative data-sets were built which ensured that learning was in no way influenced by the motif
258 alone as the motif may also occur randomly to some extent and surrounding context is also
259 considered along in a right manner. This approach was carried out for 74 RBPs for which similar
260 condition RNA-seq data were available. Datasets derived this way were called Set A data-sets.

261

262 For 57 RBP similar RNA-seq data were not available for the corresponding conditions. In such
263 scenario the main motifs for negative datasets were searched in those regions which did not appear
264 in the CLIP-seq data but belonged to the same target RNA sequences whose some part appeared in
265 the CLIP-seq, suggesting that though the RNA expressed and even bound to the RBP, these regions
266 despite of having the motif for the RBP did not bind to the RBP and may work as a suitable
267 negative dataset. +75 and -75 flanking bases from both the terminals of the motifs were considered
268 along with the motif region to build the negative datasets. These data-sets were called Set B data-
269 sets.

270

271 **Feature generation for positive and negative datasets**

272 Five different types of properties were considered for input into machine learning: 1) The main
273 motif itself, 2) Di-nucleotide density in the associated region while considering 75 bases flanking
274 regions from both the sides of the motif, 3) Dot bracket representation of the RNA structural triplet

275 for the data-set sequences, covering twenty seven combinations of structure triplets arising from the
276 dot-bracket structural representation from RNAfold predicted RNA structures [.(, .), .(, .)
277 (, .), .), ..(, ..), ..., (((, ((, ((, O(G, O), O), (.G, (.), (.,))((,)O,)(.,))(,))) ,), .), .),], 4) Pentamers
278 density profile for each position which captures the shape information, and, 5) Heptamers densities
279 for the complete region. Dinucleotide densities were evaluated for their discriminatory power for
280 multiple sliding windows starting from 17 to 131. Similarly, the dot brackets structural triplets
281 representation of the data-set sequences were generated using RNAfold (27). They too were
282 evaluated for optimum windows size while testing for window sizes ranging from 29 to full
283 sequence. 1,024 pentamers and 16,384 heptamers densities were evaluated in the similar manner
284 across the data-set sequences.

285

286 To calculate heptamers based feature, all positive datasets were split into k-mers of seven bases.
287 Probability of each k-mer were calculated with maximum of two mismatches for each position and
288 accordingly populated in the tensor. Thus, we had 16,384 X ((sequence length) - 7) tensor of
289 probabilities. 16,384 rows represent the heptamers and 150 columns represent individual positions.
290 In the similar manner pentamer features were calculated. For that we had 1024 X ((sequence
291 length)-5) tensor of probabilities. These both tensors were used to convert the sequence data into
292 vectors of probabilities. All together, based on optimum windows, the combined features sets
293 representation of all the data-set sequences was done. The optimum windows and total features
294 varied for each RBP. Finally, each data-set was broken into training and testing data-sets ensuring
295 that no instances from training ever appeared in the testing data-sets. The breakup for each RBP for
296 their training and testing data-sets is given in Supplementary Data 1 Sheet 5 and 6.

297

298 **Features evaluation on data-sets**

299 After generating all the features from positive and negative data-sets these features were
300 individually checked for their performance using tree based approaches which are expected to
301 perform better on high dimension instances. Random forest and XGBoost were applied. Each
302 property and their associated feature sets were evaluated for the varying window sizes for their
303 discrimination power between the positive and negative sets. Sliding windows of variable sizes
304 were used for dinucleotide and structure based features. These variable sizes windows were
305 evaluated for the performance. Out of these different sized windows the size producing the best
306 performance was kept for final model generation. It was found that the best performing window size
307 varied across the RBPs, resulting into different optimum windows for the RBPs.

308

309 Pentamers and heptamers appeared most informative on the full length window. Equal number of
310 positive and negative instances were chosen for all RBPs considered in the study. From the total
311 chosen instances, 60% were used to create the training set, while remaining 40% instances were
312 used to create the testing set. Python scikit-learn library was used for the same purpose. For feature
313 importance evaluation F-score was used for every considered feature. F-score locates the features
314 which display major difference between their values between negative and positive training sets
315 while comparing the averages for the feature values for positive, negative, and whole set of
316 instances (28). The F-score is represented by the following equation:

317

$$318 \quad F(i) = \frac{\left(\overline{x_i}^+ - \overline{x_i}\right)^2 + \left(\overline{x_i}^- - \overline{x_i}\right)^2}{1/(n_+ - 1) \sum_{k=1}^{n_+} \left(x_{(k,i)}^+ - \overline{x_i}^+\right)^2 + 1/(n_- - 1) \sum_{k=1}^{n_-} \left(x_{(k,i)}^- - \overline{x_i}^-\right)^2}$$

319 Where:

320 $F(i)$ = Feature score for the i th feature,

321 $(\bar{x}_i)^+$ = Average for i -th feature across the positive instances

322 \bar{x}_i = Total average of the i -th feature across the complete data-set

323 $(\bar{x}_i)^-$ = Average for i -th feature across the negative instances

324 $x_{(k,i)}^+$ = Feature value for k -th instance for i -th feature in positive data-set

325 $x_{(k,i)}^-$ = Feature value for k -th instance for i -th feature in negative data-set

326 n_+ = Total number of positive instances

327 n_- = Total number of negative instances

328 Also, for every i -th feature, t-test was conducted between n_+ and n_- to evaluate the significance of i -
329 th feature for its discrimination capability between positive and negative instances.

330

331 **Machine learning implementation**

332 With the optimized windows in the above mentioned section, feature vectors for all the RBPs were
333 used to build models to recognize RBP binding sites using two major machine-learning approaches:
334 XGBoost and Two Hidden Layers based Deep Feed Forward Neural Networks (DNNs). Both were
335 implemented using python scikit-learn, Keras, and Tensorflow libraries. In both the cases 70% and
336 30% of data were retained for train and test sets, respectively.

337

338 The DNNs were built where the input layers had number of nodes equal to the number of features
339 for the RBP considered. Thus, the size of input layer varied from 1,200 to 2,500. The performance
340 of DNN was also evaluated for various numbers of hidden layers where finally total two hidden
341 layers were found performing the best. The connections between the nodes were made dense. For

342 every RBP model the number of nodes across the two hidden layers varied between 700 to 1,300.
343 Different types of activation functions combinations were applied for the layers from a pool of a
344 number of available activation functions. Activation functions define the layers and transform the
345 activation values obtained from previous layer to a non-linear form, creating several hyperplanes to
346 obtain best possible discrimination of instances. In most of the models here, the first hidden layer
347 had RELU and the second hidden layer had ELU (for some cases they interchanged also), while the
348 final output layer had sigmoid function.

349

350 Every learning step provides estimation of error made, measuring the error and accordingly
351 corrections in the learning rate and weights on connections are done. This error estimation is
352 achieved by loss/cost functions. Multiple types of loss functions were tried to optimize the accuracy.
353 The best performance was obtained for Binary Cross Entropy. Since its a feed forward network
354 where the cost function assess the missed targets and accordingly network connection weights are
355 updated though some optimizer. The optimizer parameter which worked the best was 'Adam'
356 optimizer, otherwise SGD with momentum. Usually Adam optimizer works better because of its
357 capability to provide different learning rates per parameter, deals better with sparse gradients, and
358 adapts based on recent learning rates while keeping them in memory. Momentum was applied in the
359 learning which helps to ward-off entrapment under local minima during the minimization steps. The
360 learning rate varied from 0.001 to 0.01 and momentum varied from 0.05 to 0.9. L1 and L2 weight
361 decay regularizers were applied to avoid over-fitting. DNN models were trained using 1000 epochs
362 and batch sizes varying from 50 to 200 instances. All the model from DNN and XGBoost were
363 saved in protobuf format. Since the entire system is implemented here using TensorFlow, the
364 protbuf file provides the graph definition and weights of the model to the TensorFlow structure. The
365 optimum parameter values were fixed using an in-house developed script which tested various
366 combinations of values of the paramters to pick the best ones.

367

368 In XGBoost, grid search was applied for parameter optimization. Following parameters were
369 finalized after the grid search: `params = {"eta/learning rate": 0.2, "max_depth": 4, "objective":`
370 `"binary:logistic", "silent": 1, "base_score": np.mean(yt), 'n_estimators': 1000, "eval_metric":`
371 `"logloss"}`. Gradient boosted decision trees learn very quickly and may overfit. To overcome this
372 shrinkage was used which slows down the the learning rate of gradient boosting models. Size of the
373 decision tree were run on max-depth=9. At the value of 4 stability was gained as the logloss got
374 stabilized and did not change thereafter.

375

376 To evaluate the consistency of performance models developed with the given features, 10-fold cross
377 validation was also performed for each RBP. Everytime, the training dataset was split into 70:30
378 ratio with first used to train and second part used to test, respectively. Each time data was shuffled
379 and random data was selected for building new model from scratch. This process was repeated 10
380 times for each RBP. Accuracy and other performance measure were calculated for each model.
381 (Supplementary data 1 sheet 7)

382

383 The performance on test sets was also evaluated. Confusion matrices containing correctly and
384 incorrectly identified test set instances were built for each RBPs. Frequently used measures for
385 classifier performance evaluation and accuracy of RBPs models were evaluated. Sensitivity
386 (Sn)/Recall/True Positive Rate (TPR) defines the portion of positives which were correctly
387 identified as positives whereas specificity describes the portion of negative instances correctly
388 identified. Precision estimates the proportion of positives with respect to total true and false
389 positives. F1-score was also evaluated which measures the balance between precision and recall.
390 AUC/ROC were also measured for each model. Besides these metrics, Mathew's Correlation

391 Coefficient (MCC) was also considered. MCC is considered among the best metrics to fathom the
392 performance where score equally influenced by all the four confusion matrix classes (true positives,
393 false negatives, true negatives, and false positives) (29). A good MCC score is an indicator of robust
394 and balanced model with high degree of performance consistency.

395 Performance measures were done using the following equations:

396
$$\text{Acc} = \frac{\text{TN} + \text{TP}}{\text{TN} + \text{TP} + \text{FN} + \text{FP}}$$

397
$$\text{Specificity (Sp)} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

398
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

399
$$\text{Recall/Sensitivity (Sn)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

400
$$F1 - \text{Score} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right)$$

401
$$\text{AUC} = \int_0^1 \text{Pr}[\text{TP} | v] dv$$

402
$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

403 Where:

404 TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, Acc =
405 Accuracy, AUC = Area Under Curve

406

407 **Structural analysis of identified binding spots**

408 To assess the stability and dynamics of the RBP-RNA complexes for the identified binding spots,
409 structural analysis was done. The 3D coordinates of RBPs were retrieved from the Protein Data

410 Bank (PDB). X-Ray crystallographic structure for 13 different RBPs were downloaded. Prior to
411 docking, protein structures were prepared by removing water molecules and other hetero-atoms,
412 while adding polar hydrogen atoms. RNA motifs identified through RBSPot algorithm for above
413 mentioned five RBPs were taken as flexible molecules. All docking studies were performed through
414 NPDock (Nucleic Acid–Protein Docking) and PATCHDOCK incorporating more realistic DARS-
415 RNP statistical potential based on reverse Boltzmann statistics to score protein-RNA complexes
416 (30). RNA motifs three dimensional structures were built using RNACOMPOSER web server based
417 on RNA FRABASE database relating the RNA secondary and tertiary structure elements. In order
418 to search for all possible RNA-binding sites and optimize the structural effects of RNA on the
419 construction of complex, short RNA motifs were taken into account. Protein-RNA interface residues
420 were predicted using DR_Bind1 (31) based on evolutionary conservation. Top three representative
421 docking potential-ranked protein-RNA complexes were built for each of the above mentioned RBPs
422 and the best one was considered for further analysis.

423

424 **MD simulations**

425 All molecular dynamics simulations of the RBP alone and the RBP–RNA complex were conducted
426 using GROMACS 5.1 package (32), modeling each system with the AMBER03 force-field of
427 protein and nucleic acids (33) with periodic boundary conditions. The topology files for the selected
428 target RNA motifs were built using pdb2gmx in the framework of AMBER03 force-field. Models
429 were solvated with the TIP3P water model (34). The distance between the biomolecule and the edge
430 of the simulation box was set as minimum 1.0 Å so that they could not directly interact with their
431 own periodic boundary condition and fully immerse with water while rotating freely. Boxes were
432 solvated with TIP3P water. The number of solvated molecules added to each system varied. After
433 the establishment of initial configuration, the systems were minimized. 50,000 steps (steepest
434 descent approach) were used for each system until the maximum force of < 10.0 kJ/mol for energy

435 minimization. For calculation of long range electrostatic interactions, Particle Mesh Ewald (PME)
436 method was used. To establish the systems at constant temperature of 300K, V-rescale thermostat
437 (modified Berendsen thermostat), at a constant pressure of 1 bar, and Parrinello-Rahman thermostat
438 were applied with a 2 ps coupling constant for both parameters. The LINCS algorithm (35) was
439 used to constrain all bond lengths involving hydrogens. During the production run, a time step of 2
440 fs was used and conformations were saved every 10 ps for the analysis of molecular dynamics
441 trajectory of total 20 ns for each RBP and their complexes using leap-frog algorithm (36) to
442 integrate the equation of motion. MD trajectories were further evaluated for considering Root Mean
443 Square Deviation (RMSD). RMSD is suitable to decipher the structural changes in proteins and
444 their complex structures corresponding to initial structure during the course of different time
445 periods of dynamics simulation. RMSD was calculated using the following equation:

$$446 \text{RMSD} = \sqrt{\frac{1}{N} \times \sum_{i=1}^N |u_i - v_i|^2}$$

447 where,

448 u_i =Cartesian coordinates of atom i in the initial structure;

449 v_i =Cartesian coordinates of atom i in the structure during simulation;

450 N =number of atoms;

451 To analyze the structural properties of the individual RBPs and their complexes in the form of root
452 mean square deviation (RMSD), g_rms functions were utilized. Changes in trajectories of molecular
453 dynamics during course of simulation were plotted for evaluation using python plotting library.

454

455

456

457 **Co-occurring RNA motifs group clustering**

458 A two steps statistical approach was employed to identify the co-occurring motif pairs. In this
459 approach, the positive set of RBP was scanned for other most frequent occurring k-mers. Top co-
460 occurring motifs were checked for their statistical significance. KS-test was used to find the
461 significance of distance for two motifs. All the distance between two motifs were calculated from
462 positive and negative data-sets. Distribution plot of random data and positive data were further
463 checked using KS-test. Level of significance were considered $p < 0.05$. They were further checked
464 for frequency ratio (FR). At 5% level of significance, if the hypergeometric test *p-value* was less
465 than 0.05, motif pair of enriched and co-occurring motifs was considered significant. Additionally,
466 frequency ratio (FR) as a measure of co-occurrence of motif pairs was also computed to estimate
467 the tendency of motif pairs to co-occur with each other as proposed previously (37):

$$FR \{ Motif_{M2/M1} \} = \frac{X_{M2/M1} / N_{M1}}{Y_{M2/M1} / M_{M1}}$$

468

469 $X_{M2/M1}$ = Number of sequences containing motif1

470 N_{M1} = Number of sequences containing motif2 co-occurring with motif1

471 $Y_{M2/M1}$ = Number of sequences without motif1

472 M_{M1} = Number of sequences containing motif2 without motif

473

474 **Benchmarking and Performance Evaluation**

475 To evaluate the RBPSpot performance and the importance of dataset constructed in this study, we
476 compared RBPSpot with five different tool: RBPmap, DeepBind, iDeepE, DeepCLIP, and beRBP.
477 Three different datasets were considered separately for the benchmarking process: Datasets used for
478 RBPSpot, beRBP, and Graphprot. Datasets of beRBP and Graphprot are common data source for
479 most of the existing published software built to identify RBP-RNA interactions. As already
480 mentioned above, RBPSpot dataset is based on the positive datasets from ENCORI (the
481 encyclopedia of RNA Interactomes, previously known as StarBase) and the negative datasets based

482 on the protocol mentioned above in the previous section. This dataset contained positive and
483 negative sequences for 131 RBPs in which length of sequence varied from minimum of 156 to
484 maximum of 160 bases. The variation in the length of the sequences for different RBPs was due to
485 the varying length of their major motifs. For benchmarking purpose those RBPs data were
486 considered from this dataset for which at least one tool had model ready for comparison. No such
487 RBP was considered from this dataset for benchmarking for which no other tool had model ready
488 for comparison. This way a total of 52 RBP data were used from RBPSpot dataset for the
489 comparison purpose.

490

491 The beRBP dataset is available for 29 RBPs. This dataset is based on the experimentally validated
492 target sequences (3'-UTRs) for human RBPs (positive datasets) from AURA (38) (v2,
493 8/5/2015;<http://aura.science.unitn.it/>), which is a manually curated and comprehensive catalog of
494 human UTRs bound by regulators including RBPs. Negative instances of this dataset has random
495 sequences chosen from the 3'-UTR pool. The beRBP dataset was obtained from the URL
496 <http://bioinfo.vanderbilt.edu/beRBP/download/TabS1.7z>.

497

498 The third dataset considered in this study was built during the work presenting Graphprot software.
499 Since then, this dataset has been used extensively by many published software to this date. This
500 dataset covers 24 RBPs coming from various CLIP-seq experiments. For each set of CLIP-seq data,
501 they created a set of unbound sites by shuffling the co-ordinates of bound sites within all genes
502 occupied by at least one binding site which worked as the negative dataset. making the
503 corresponding negative dataset instances. This dataset was retrieved from URL
504 http://www.bioinf.uni-freiburg.de/Software/GraphProt/GraphProt_CLIP_sequences.tar.bz2.

505

506 The four out of the compared five tools *viz.* beRBP, RBPmap, DeepCLIP, and DeepBind provide
507 pre-built models. Only iDeepE does not provide any pre-built model. To overcome this, models
508 were generated using iDeepE methodology for the datasets. To make binary decisions with
509 DeepBind, threshold of 0.7 was applied after performing logistic transformation of the raw
510 DeepBind scores (39).

511

512 In the second part of the benchmarking impact of datasets was assessed on model building quality
513 where models were built using different datasets and various combinations of test and train
514 datasets were analysed. Besides RBPSpot, only two tool, iDeepE and DeepCLIP, had provision to
515 build models from user provided datasets. Remaining tools have fixed models with which they
516 work and don't provide the provision to build models from user provided data. Therefore, they
517 could not be included in this part of benchmarking. Thus, for this part, the datasets used by
518 RBPSpot (RBPSpot dataset), iDeepE, and DeepCLIP (Graphprot dataset) were used. Four
519 different combinations of train and test datasets (RBPSpot train and RBPSpot test, RBPSpot train
520 and Graphprot test, Graphprot train and RBPSpot test and Graphprot train and Graphprot test) were
521 used for the benchmarking to evaluate the impact of datasets on the performance of these
522 algorithms.

523

524 **Comparison with experimentally reported motifs**

525 A total of 29 RBPs from RNAcompete study were found overlapping with our set of 131 RBPs.
526 Their IUPAC motifs were downloaded from RNAcompete web portal. For these 29 RBPs a total of
527 44 motifs were reported. Out of these 44 motifs, 35 motifs had a length of 7 bases, eight motifs had
528 a length of 6 bases, and one motif had a length of 5 bases. Four motifs out of 44, were discarded
529 due to more than 3 variable positions in a length of 7 bases. Therefore, in the final analysis a total of
530 40 motifs representing 26 RBPs, were present. These motifs were scanned in the similar manner as

531 was done with the search for motifs identified by RBPSpot approach in order to maintain an
532 unbiased motif search approach. Random data sets to evaluate the random chance observations were
533 generated from the transcriptome data using the length exactly similar to the ones from the cross-
534 linking peak data. The similar above mentioned allowed mismatches based motif searching criteria
535 was used here also to scan the random datasets for motif occurrence in them. Binomial test was
536 applied to find the significance of these motifs in the cross-linking data. Other than RNAcompete
537 motif, experimentally validated motifs were also considered from CISBP-RNA Database. A total of
538 31 RBPs from this dataset were found overlapping with RBPSpot data. Out of these 31 RBPs, 24
539 RBPs were reported from RNACompete study only, two RBPs were reported through SELEX and
540 yeast three-hybrid screening whereas five RBPs were reported from RNAcompete and SELEX/RIP-
541 Chip. These motifs were also searched in the similar manner.

542

543 **Application of RBPSpot across SARS-CoV2 genome**

544 To identify the binding sites of RBPs across SARS-Cov2 genome, we downloaded its genome from
545 NCBI (accession number NC_045512).

546

547

548 **Results and Discussion**

549 **Reads data collection, filtering, and pre-processing**

550 CLIP-seq peak data from various sources were collected for 137 RBPs from starBase 2.0, also
551 know as ENCORI (Supplementary Data 1 Sheet 1). All the data were collected in the form of co-
552 ordinates. These data were from multiple types of CLIP-seq experimental techniques i.e. CLASH,
553 dCLIP, eCLIP, FLASH-CLIP-seq, HITS-CLIP, iCLIP, PAR-CLIP, sCLIP-seq, and uvCLAP. The
554 peak data varied from 234 (PAPD5) to 9,84,503 (U2AF2) peaks. Initially, six RBPs' data were

555 discarded due to insufficient peak data availability. Here we considered only those RBPs which
556 were having >500 unique binding peaks available. These six RBPs viz. PAPD5 (234), EIF3B (298),
557 EIF3A (371), EIF3G (398), EIF3D (399), and PUM1 (473) had lesser number of initial peaks
558 available. Remaining data for 131 RBPs were having a total number of 2,11,23,594 unique peaks.
559 To further filter this data we discarded those sequences which were having a length <5 nucleotides
560 or extreme length sequences (>300 basepairs). With this all, a total of 1,87,14,999 peaks were
561 available for the study, varying from EIF4A1 (1,175) to AGO1-4 (9,41,224). Initial co-ordinate data
562 were extracted into sequences from genome. Initial and final data are given in (Supplementary Data
563 1 Sheet 2).

564

565 **Most of the RBP binding sites display a prime binding motif covering majority and along with**
566 **co-occurring motifs**

567 As discussed in the introduction section, most of the available tools for identifying the RBP
568 bindings sites across the RNAs require either prior information available traditional motif finding
569 approaches like MEME, TOMTOM or HOMER. The application of traditional motif discovery
570 tools may not be much information in case of RBPs which have been reported to be sparse, short,
571 and poorly conserved. Further to this, such motif discovery approaches expect user defined motif
572 length instead of naturally capturing the motif. In general, if such motifs are not considered with
573 proper context they may lead towards false discoveries. Here, we have used the initial deep-
574 sequencing data to find the most frequently occurring k-mers (seeds) to make it an initial step for
575 motif finding. To find a naturally occurring most frequent k-mers, search was started with k=6 with
576 two mismatches. Reason behind this was that some of the previously reported motifs for RBPs
577 were either very sparse or as small as 4 bases long only. Therefore, a k-mer with six bases and with
578 two mismatches would fetch all possible 6-mer spectrum which would agree with each other with
579 two mismatches (relatives to the main k-mer) while also meeting the lowest bound of such motifs.

580 This defined the 6-mer groups within two mismatches. Since a large number of 6-mers spectrum is
581 created whose search with two mismatches across the sequences becomes a computationally
582 intensive and time consuming step, a FM-Indexing and Burrows Wheeler Transformation (BWT)
583 based inexact search step was applied. Since, parallelism through multiprocessing was also
584 implemented, the search becomes more faster with available cores of CPUs.

585

586 4,096 possible combinations of 6-mers were individually searched in the peak data for every RBP.
587 To select the most abundant 6-mers, the first criteria was its occurrence across at least 70% of the
588 CLIP-seq peak region data. All the 6-mers which were occurring in at least 70% data were
589 evaluated for their significance occurrence at $p\text{-value} \leq 0.01$ using binomial test. Many most
590 frequently occurring 6-mers were found whose numbers varied for RBPs (from RBM39 (3) to
591 ELAVL1(17)). The found significant spots for 6-mers for any given type worked as the seed which
592 were subjected to bi-directional expansion. This expansion step every time evaluated the similarity
593 between the expanded region and checked for minimum similarity cut-off of 70% across the
594 considered seed regions which were expanding. The 6-mer seeds were expanded at every found
595 position until they were satisfying both the criteria. The final step resulted into the most frequently
596 occurring elongated k-mers with maximum possible elongation with both criteria met. After
597 elongation, the best scoring expanded k-mer family for each RBP was considered as the primary
598 motif in the RNA sequences interacting with the given RBP. It was found that at least one such
599 primary motif existed for all the RBPs considered in this study, barring four RBPs. These primary
600 motifs were occurring in at least in 70% of the data with high significance. The size of primary
601 motifs varied from 6 bases to 10 bases for different RBPs. The most abundant motif was based on
602 UCUGCAG for ALKBH5 (92.27%), where as the least abundant motif was based on CCUGGAGG
603 for SLBP protein (Supplementary Data 1 Sheet 3).

604

605 There were four different RBPs viz. FXR1, SND1, ILF3, and U2AF1 which did not have any single
606 seed k-mer occurring in at least 70% of the data. This suggested the possibility for multiple motifs
607 working in mutually exclusive manner. It was found that two different motif groups for these four
608 RBPs were working almost in mutually exclusiveness manner with small fractions of overlaps in
609 their instances. The overlap levels between these two motif groups' instances were: FXR1 (7.8%),
610 SND1 (6.8%), ILF3(8.25%) and U2AF1 (9.5%) instances (Supplementary Data 1 Sheet 4).
611 However, for these cases the found 6-mers could not be expanded further as at least 10% of the data
612 was lost due to this.

613

614 This way, the most significant motifs present in the cross-linking data of all these RBPs were
615 discovered which could act as anchor in contextual form. It was interesting to observe that the
616 identified motifs could be clustered into various groups based on their similarity. For every RBP, the
617 motifs obtained from their respective sequences were used to develop their position weight matrices
618 and logos which were compared with each other for similarity based clustering. This resulted into
619 28 clusters of RBPs where RBPs belonging to same cluster shared good level of similarity for their
620 prime motifs (Supplementary Figure 1). Such display of grouping among RBPs is reflection of
621 unity in diversity phenomenon as well as strongly suggest that how much of importance contextual
622 factors could be for RBP-RNA interactions that despite of sharing similarity in their main motifs the
623 binding appeared highly contextual. This also transpires from the study done on the flanking
624 regions of these main motifs for the RBPs belonging to the same cluster. The di-nucleotide,
625 pentamers and heptamers based information content strongly varied among themselves for many
626 cases. The upcoming sections will present some related information on this.

627

628 When these motifs were mapped back to the genome in order to derive the contextual information,
629 several of them hinted for coexistence of secondary supporting motifs for any given RBP. Such
630 cases were studied further for co-occurrence of motifs where the most dominant motif would be
631 supported by some other predominant secondary motif. All those sequences where the dominant
632 motif existed were also searched for the supporting secondary motifs. Obtained co-occurring motif
633 pairs were further evaluated to measure the similarity between them using Jaccard similarity index
634 based approach. The method utilizes the position weight matrices of co-occurring motifs for
635 alignment considering relative shifts to recognize similarity between two motifs (40). All co-
636 occurring motif pairs possessed similarity score < 0.2 ensuring different motif partners being
637 evaluated instead of same motif repeating itself. Sequence regions where the motifs co-occurred
638 displayed high statistical significance of co-occurrence rate for the motif pairs for any given
639 distance ($p\text{-value} \ll 0.05$; KS-test). For all RBP models, big difference was observed for the
640 distribution of co-occurring motif pairs when compared to the random sequence regions, strongly
641 supporting the existence of co-occurrence of motifs in RBP binding models of RNAs. Figure 2
642 illustrates some of these cases. In this way, 178 statistically significant co-occurring motifs pairs out
643 of 297 motif pairs for 127 RBPs were obtained, strongly suggesting again that context holds
644 importance in RBP-RNA interactions. Co-occurring motif details for the RBPs is given in the
645 Supplementary data 1 Sheet 8. Further these motifs were also checked for frequency ratio > 1 as
646 discussed in method section. All 178 statistically significant co-occurring motif pairs were found to
647 have frequency ratio (FR) > 1 . These co-occurring motifs were analyzed for the region flanking 75
648 bases from both sides of the prime motif. The reasons for considering this region becomes more
649 clear in the following next section.

650

651 The motifs reported in the present study were compared with the experimentally reported motifs.
652 Most of the motifs found in this study matched with the experimentally reported motifs. However, it

653 was also observed that several of these experimentally reported motifs were not the prime motif
654 reported here but matched to other lower ranked motifs which either co-occurred with the prime
655 motifs or were exclusively present, covering comparatively lesser amount of CLIP-seq data than the
656 prime motifs reported in the present study. Their occurrence in the cross-linking data varied from
657 34.07% to 81.32% while the prime motifs reported in this study mostly covered at least 70% of
658 CLIP-seq data. Figure 3 provides a snapshot of the comparison between experimentally reported
659 motifs with motifs identified in the present study.

660

661 **Consideration of expression data for targets helps in building more realistic data-sets**

662 The discovered motifs above work as a point to zero upon to consider the potential significant
663 interaction spots in the RNA. However, such motifs alone can't hold much higher stake than that as
664 they may appear in the non-binding regions also, though found statistically significant for the
665 binding regions. Evaluation of their context for their functional role thus becomes essential. In this
666 regard, 75 bases from both the flanking regions were considered where the motif region worked as
667 the anchor. Previously, it has been found that ~75 bases of flanking regions around the potential
668 interaction sites in RNAs capture the local environment for structural and contributory information
669 effectively (20). Also, RBPs which interact with the RNA through multiple domains use multiple
670 interaction sites which are usually concentrated around a local region instead of being long
671 distanced interaction spots. Thus, uniform length sequences with flanking regions were obtained for
672 every individual RBPs which varied for different RBPs depending upon the length of their anchor
673 motifs. This also led to the construction of positive and negative instances datasets, simultaneously.
674 The number of positive instances differed for the RBPs depending upon their available cross-
675 linking sequencing data, ranging from 1,309 (EIF4A1) to 8,48,680 instances (AGO1-4). This
676 covered a total of 19,547 genes experimentally confirmed as targets of these RBPs. Total number of

677 instances was greater than total number of peak data for most of the RBPs due to multiple
678 occurrence of motifs on a single sequence.

679

680 Identifying suitable negative dataset candidates becomes a more crucial task. And it is where most
681 of the previously developed tools have gone too soft and mostly ended up selecting random
682 sequences, which actually does not help to divulge more information. As transpires from above
683 discussions and results, there are many spots across the transcriptomes which possess sequences
684 similar to the interaction motifs but they yet not interact. In usual, chances of finding shorter motif
685 themselves is higher in the random data. In such scenario, considering random sequences really
686 does not add significantly to the purpose of discrimination and does not answer the question raised
687 above. In order to build a better negative dataset, it is better to pick those candidates as negative
688 instances where the region similar to the main motif is present and creates a strong confusion
689 matrices to build a more natural and robust model. Therefore, to create the negative set for RBPs
690 two different kind of strategies were used. In the first strategy we used RNA-seq data for the same
691 condition for which we had the cross-linking data available for the given RBP. Those RNAs were
692 selected which were expressing themselves in the same condition but did not bind to the considered
693 RBP and did not reflect in the CLIP-seq data. They were searched for the prime motifs of the RBP
694 similar to the positive data cases and in similar manner 75 bases flanks were considered along with
695 capturing the contextual information with more discrimination power. For the RBPs for which the
696 negative datasets were created using this strategy are called Set A RBP datasets throughout this
697 study. This way, the negative datasets for 74 RBPs were created (Supplementary Data 1 Sheet 5).

698

699 In the second strategy, the negative datasets were created for those RBPs which did not have similar
700 condition RNA-seq data available for the considered CLIP-seq conditions. In such scenario,

701 therefore, here those RNAs were considered which exhibited binding to their respective RBPs but
702 they also had the motifs on other positions which did reflect in the CLIP-seq data and were also far
703 away from such cross-linking regions. The logic behind is that such RNA sequences whose some
704 regions exhibited binding to RBPs in CLIP-seq data make clear positive instances out of these
705 regions as well as hold a simultaneous evidence that these RNAs were expressed in the given
706 condition. Regions which display the interaction motif in these expressed RNAs but don't bind to
707 the RBPs become an apt case for negative instance consideration with high potential for contextual
708 information unlike the usual random sequences. This particular set of negative dataset instances
709 were called Set B. In this way, the Set B negative dataset were created for the remaining 57 RBPs
710 (Supplementary Data 1 Sheet 6)). Rest of the analysis were same on both the sets of RBPs. This all
711 also reinforces the view that any successful RBP-RNA interaction discovery approach can not be
712 founded solely upon the motifs consideration but needs correctly designed context information
713 extraction approach also which can be provided only after a better a negative instances
714 consideration.

715

716 **Contextual information surrounding the anchored motif is critical for RBP binding sites**
717 **recognition**

718 Motif discovery and anchoring helped in selecting the more appropriate positive and negative
719 instances from which contextual information and features might be derived. The contextual
720 information came in the form of other co-occurring motifs, sequence specific information, position
721 specific information, and structural/shape information which could exhibit sharp discrimination
722 between negative and positive instances. Contextual information were derived from the features
723 based on four major properties: (1) 7-mers frequency probability for each position, (2) 5-mers
724 frequency probability for each position, (3) di-nucleotide densities in the region, and (3) Structural
725 triplet frequency covering 27 combinations of structure triplets arising from the dot-bracket

726 structural representation from RNAfold predicted RNA structures. Consideration of heptamer was
727 for picking up any further sequence specific signals in the flanking region, where similar approach
728 of inexact search was applied with at least 70% similarity match as was done for the prime motifs'
729 6-mer seeds. Pentamers application was motivated from the recent findings which reported that
730 pentamers capture the DNA shape very accurately (24). The nucleic acids shape has been found
731 critical in the interactions with regulatory proteins which scan these shapes for their stationing. So
732 far, this approach has been applied on DNA but hardly on RNAs. The DeepBind work had
733 observed about the importance of using such kind of information which could be beneficial in
734 future developments for the tools reporting RBP-RNA interactions (14). The dinucleotide densities
735 have been found to be highly useful in indirectly evaluating the RNA structure and accessibility
736 (20,21). In fact, it has been found more promising than *ab-initio* RNA structure prediction. *Ab-*
737 *initio* methods' accuracy drastically falls with the length of RNA, and they are suitable for only
738 short RNA sequences (9,20). Pentamer and di-nucleotide frequencies capture better structural and
739 shape information through base stacking and neighborhood contribution. Similarly, RNA structure
740 triplet has been used widely in deriving the structural information of RNA for their propensity
741 towards interaction factors, especially for miRNA:RNA interactions (41).

742

743 Various features generated based on the above mentioned properties were evaluated for their
744 discrimination potential between the positive and negative instances. The most important top 100
745 features are given in supplementary data 1 sheet 9. Among them, the features originating from the
746 dinucleotide densities appeared the most. Some pentamer and heptamer features were also present
747 among these top features. Dinucleotide density reflects the structural and accessibility properties of
748 the nucleic acids, as mentioned above. A very striking observation was also made here. Most of the
749 positive instances flanking regions displayed enrichment of CG. Approximately 69% of RBPs
750 target regions exhibited CG among the most dominant feature for each position. Where as for rest

751 of the RBPs had UU and UA among the most prominent features. Besides this, it was also observed
752 that RBPs which shared high similarity for their binding motifs and were clustered among the same
753 group (Supplementary Figure 1) differed substantially for this contextual information and their
754 flanking regions displayed different distribution patterns. Figure 4 presents an example of one such
755 group, RBPs belonging to AGO4 cluster (Cluster 1). As can be noted in this figure also, CG is
756 remarkably enriched for the binding site regions. Therefore, despite of having binding sites motifs
757 they differ in their binding which is influenced by context. Also, the universal prominence of CG in
758 the RBP binding regions reinforces the theory which suggests their regulatory roles in stationing
759 the binding factors and supporting the binding motifs (42). Also, they may be studied further for
760 RNA modification which are considered critical for RBP binding dynamics.

761

762 For 12 RBPs, pentamers were also found in the top 20 features for different positions whereas
763 heptamers were found for 10 RBPs in the top 20 features. Among top 100 features, almost in 90%
764 cases heptamers and pentamers marked their presence. Significant difference was observed
765 between the positive and negative instances with respect to the F-score for positions which also
766 suggest that substantial amount of information is being held by the flanking regions around the
767 binding motif, which may be one of the determinant for contextual interactions between RBP and
768 RNA. A series of t-tests between the positive and negative instances for various features also
769 supported this. Biologically, heptamers and pentamers were expected to reflect any supporting co-
770 occurring motifs near the prime anchored motif. Pentamers, specifically, were considered to
771 capture the shape properties, which too have been called important in protein and nucleic acids
772 interactions, more so in cases where sequence motifs are not clear or prime (14,24). A closer look
773 with these pentamers and heptamers revealed that for many RBPs binding sites, they were
774 prominent in the flanking regions where the co-occurring secondary motifs existed (Figure 2).
775 Though heptamers were found more reflective to this phenomenon. As could be expected now,

776 these information properties from the flanking regions looked highly promising for identification of
777 a true binding site. The impact of each of these properties on discrimination capacity between true
778 binding sites and negative sites was also clear when evaluated directly on the machine learning
779 models for performance, as transpires in the following section.

780

781 **DNN implementation of the RBP binding site models consistently achieved high accuracy**

782 Before combining the features to build the collective models for RBP-RNA interactions, one more
783 assessment of contributions by the above mentioned properties in discrimination was done.
784 Classification assessment was made for each given properties separately before joining them
785 together while using XGBoosting. This was done to get the preliminary idea about the individual
786 contribution made by each of the contextual properties towards the accurate classification and how
787 important they looked in the process of accurate recognition of the binding spots. For the pentamers
788 based classification the accuracy varied from 60.23% (U2AF2) to 82.01% (FKBP4) for Set A RBPs
789 with an average of 69.8% accuracy. For heptamers it varied from 65.01%(FXR1) to 88.72% (FXR2)
790 with an average accuracy of 76.49% for set A RBPs. Similarly, for set B RBPs pentamer accuracy
791 varied from 55.6% (DHX9) to 86% (EIF4A1) with an average of 66.7% accuracy. For heptamers it
792 varied from 57.23% (MOV10) to 97.47% (EIF4A1) with an average accuracy of 77% for Set B
793 RBPs. For structure triplets we used different window size but none of the windows achieved more
794 than 63.39% accuracy for any RBP, clearly supporting our above made observation that *ab-initio*
795 structure prediction derived features don't add much value due to their innate limitations. Therefore,
796 this feature was not further taken for the final model building. Accuracy for di-nucleotide densities
797 based classification varied from 63.04% (FMR1) at 43 window size to 88.6% (RBFOX2) at 71
798 window size with an average of 75% accuracy at different window sizes which varied from 17 to
799 103 for Set A RBPs. Similarly, for set B RBPs the accuracy of di-nucleotide density based
800 classification varied from 61.46% (DHX9) at 71 window size to 90.57% (EIF4A1) at 91 window

801 size with an average of 75.25% accuracy (Supplementary Data 1 Sheet 5,6). The results here
802 displayed concordance with the observation made in the previous section where importance of
803 contextual dinucleotide density information based features emerged as the most important ones for
804 the binding sites detection. Figure 5(A) presents the violin plots for the accuracy distributions
805 observed for the classifications done by each of these properties for all the RBPs.

806

807 With this all, it was pretty evident that the selected properties and their features had strong
808 discriminatory strength, barring the RNA structural information derived through *ab-initio* structure
809 prediction method. All the features originating from these qualifying properties were combined
810 together to build the final models of RBP-RNA interaction targets.

811

812 After getting optimum window size for di-nucleotide densities, we combined these three features
813 (pentamers probabilities, heptamers probabilities and di-nucleotide densities) together to build the
814 final models. The final models were built using Xgboosting as well as DNN. The reason for
815 considering these two different approaches are that: 1) they reflect two different learning
816 approaches: Shallow and Deep, 2) They complement each other as Xgboost works good for the
817 cases with comparatively lower training data while DNN performance is good where learning data
818 is higher, 3) Both the approaches work very good for conditions where the dimensions are high, as
819 was with this study.

820

821 Combining of the features based on above mentioned properties was done in a gradual manner in
822 order to see the additive effect of them on the classification performance. As it is apparent from
823 Figure 5(B), which showcases the DNN classifier's performance for five RBPs, the performance of
824 the classifiers kept increasing on the addition of more features, where consistency also increased as

825 can be seen through the band width of the plots for the five RBPs. Here also, the dinucleotides
826 based contextual features emerged most critical as the biggest leap in the performance was noted
827 when it joined the heptamers and pentamers based features. Any pair of these three properties
828 features gave almost similar performance, but sharpest rise was observed in the performance when
829 contextual dinucleotide information based features were added to the pentameric and heptameric
830 features.

831

832 After combining the features we had 1,198 (ZNF184) to 2,544 (EIF4A3, EIF4G1,
833 EWSR1,HNRNPD, HNRNPL, KHDRBS3, NOP58 etc.) features for individual RBPs. The feature
834 numbers varied due to different sized best performing dinucleotide densities windows. These
835 features were used in Xgboost machine learning where the average accuracy of 85.07% (Avg. AUC:
836 85.06%, Avg F1-Score: 84.64% Avg MCC:79.26) was obtained and where the values varied from
837 79.19% (FXR2, AUC: 79.19%, F1-Score: 78.58% MCC:66.38) to 90.81% (RBM47, AUC: 90.80%,
838 F1-Score:90.49 % MCC:83.17)) for Set A RBPs. It was found that the average accuracy of 84.08%
839 (Avg. AUC: 84.07%, Avg F1-Score:82.66 % Avg MCC: 69.07) was obtained for Set B RBPs, where
840 accuracy values varied from 66.34% (MOV10, AUC: 66.34%, F1-Score: 64.74% MCC:42.58) to
841 96.78% (EIF4A1,AUC: 96.48%, F1-Score: 96.40% MCC: 92.37). The same set of the combined
842 features was also used in the DNN implementation. DNN works better with higher dimensions and
843 instances to learn from. In the input layer combined features were used where as two hidden layers
844 gave best performance and the number of nodes per hidden layer varied from 700 to 1,300. Details
845 of implementation are already given in the methods section. DNN achieved an average accuracy of
846 92.25% (Avg. AUC: 92.64%, Avg F1-Score: 91.97%, MCC:84.52%) for Set A RBPs which was
847 much higher than XGBoost. Whereas for Set B RBPs an average of 83.47% (Avg. AUC: 89.61%,
848 Avg F1-Score: 83.18%, Avg MCC:67.34%) accuracy was achieved by the DNN models, which was

849 slightly lower than XGBoost. Complete performance details can be found elsewhere
850 (Supplementary Data 1 Sheet 5,6).

851

852 In general, it was apparent that DNN approach was sensitive towards the volume of training
853 instances as it was found performing better where number of instances were higher. But the biggest
854 impact on performance was observed was for the granularity of dataset creation. Performance of
855 DNN was specially more marked here, as can be seen from its performance plot on Set A datasets.
856 On Set A, the DNN models performance hardly touched below 90% accuracy. Even XGBoost's
857 performance was better with Set A when compared to Set B. It needs to be recalled that Set B was
858 made for those RBPs for which the RNA-seq data was not available for the considered CLIP-seq
859 conditions. In such scenario, those RNA were considered to generate the negative instances whose
860 some regions were present in the CLIP-seq data suggesting their expression. From the same RNA,
861 those regions were selected which were having the prime motifs but yet not binding to the RBP and
862 not reflected in the CLIP-seq data and were distant from such binding regions. While Set A negative
863 instances were clearly those regions which were expressed during the CLIP-seq experimental
864 condition and possessed the prime motif but no region of the RNA itself bound to the RBP. Thus,
865 though the over all performance with Set B was still good and better than the datasets used by the
866 compared tools as transpires in the next section, it same time reflects that how important it is to
867 have a refined data-set like Set A. This is possible that some instances covered as negative instances
868 in Set B could be contributing to the RBP-RNA interactions or could not be captured in the CLIP-
869 seq experiments. Yet, as transpires from the various performance metrics plots across various RBPs
870 given in Figure 6 and AUC/ROC plots given in Figure 7, the developed approach in this study,
871 named as RBPSpot, showcases a consistently high and reliable performance for a large number of
872 RBPs. It also provides the largest number of models for RBPs binding developed from CLIP-seq
873 data to this date.

874

875 **Comparative benchmarking: RBPSpot consistently outperforms all the compared tools**

876 A very comprehensive benchmarking study was performed where RBPSpot was compared with five
877 different tools, representing different approaches of RBP RNA interaction detection: RBPmap
878 (probabilistic approach), beRBP (Random Forest machine learning bases claiming highest accuracy
879 in its category), DeepBind (the first deep-learning based approach), iDeepE and DeepCLIP
880 (representing some very recent and more complex deep-learning based tools). Besides this, the
881 benchmarking has also considered three different datasets as this work also presents a new dataset
882 while underlining the importance of better datasets in creating better models as well as to carry out
883 a totally unbiased assessment of performance of these tools on different datasets.

884

885 Thus, the first dataset considered in the benchmarking study was derived from the RBPSpot dataset.
886 Only those RBPs were considered for comparison for which at least one tool had model built,
887 besides RBPSpot itself. This way comparison was done for 52 RBPs. The second dataset considered
888 was the one evolved during development of Graphprot software which has been used largely by
889 various other datasets for model building and performance benchmarking purposes. The third
890 dataset used in this benchmarking study was the one used by beRBP software which too has been
891 used by many other tools for the same purpose. Details about these datasets have already been
892 discussed above and in the methods sections.

893

894 All these six software were tested across all these three datasets and RBPSpot outperformed all of
895 them across all the datasets, and for all the performance metrics considered (Figure 8). Figure 8
896 gives a detailed view of the data analysis of this benchmarking across the three datasets studied for
897 all these software. RBPSpot scored the average accuracy of 88.43% and the average MCC value of
898 0.77 on RBPSpot dataset, the average accuracy of 91.63% and the average MCC value of 0.83 on

899 Graphprot dataset, and the average accuracy of 88.9% and the average MCC value of 0.74 on
900 beRBP dataset. Among all the considered performance metrics, MCC stands as the most important
901 one as it gives high score only when a software scores high on all the four performance parameters
902 (true positive, false positive, true negative, false negative). A good MCC score signifies the
903 robustness of the model and its performance consistency. RBPSpot emerged as the most robust
904 algorithm among all these compared software with very high consistency of performance. As it is
905 visible from the score distribution for all the metrics, RBPSpot also exhibited least dispersion of
906 scores for all the studies RBPs and for all the three datasets, confirming the precise performance
907 achieved by RBPSpot compared to other tools.

908

909 After RBPSpot, the best performance was observed for the complex deep-learning based software
910 iDeepE and DeepCLIP. On RBPSpot dataset, iDeepE performed better than DeepCLIP, but for other
911 two datasets they attained almost similar metrics scores for performance. Undeniably, they emerged
912 far superior than their deep-learning predecessor, DeepBind, and other compared tools. They even
913 displayed much smaller dispersion of their scores than other compared tools. However, RBPSpot's
914 performance points out that more appropriate features may be learned through training on
915 biologically relevant properties to derive better discrimination power using machine learning
916 approach, which can be amalgamated with Deep Neural Nets with much lesser complexity and
917 superior performance than applying complex deep-learning layers to automate feature extraction.
918 The observations made in the introduction part of this work appeared true in this study that such
919 complex deep-learning approaches score good on unstructured data where clear features
920 identification and extraction is difficult to be done by expert and automation is required for feature
921 extraction. The problems where features are identifiable and can be structured, simpler machine
922 learning models may outperform the complex deep-learning approaches.

923

924 The above mentioned benchmarking was done for all the tools while keeping their original training
925 dataset and models for RBPs. Most of the existing tools don't provide the option to build user
926 specified models of RBPs using their algorithms but come with their own pre-built models. This
927 limits the scope to test the algorithms with different combinations of datasets. Fortunately, the two
928 best performing tools after RBPSpot, iDeepE and DeepCLIP, provided this scope where the users
929 may build their new models with their own datasets. Also, since these two tools performance were
930 next to RBPSpot, they stood as a natural choice to study the performance impact with datasets
931 variations. Both iDeepE and DeepCLIP have implemented Graphprot dataset for their original
932 model building. For this part of benchmarking study the training and testing datasets of RBPSpot,
933 iDeepE, and DeepCLIP were swapped and studied for four different combinations of training and
934 testing datasets: RBPSpot training and testing datasets, RBPSpot training and Graphprot testing
935 dataset, Graphprot training and testing datasets, Graphprot training and RBPSpot testing dataset.

936

937 Figure 9 presents the results for this part of benchmarking where RBP models were rebuilt and
938 tested using the four different combinations of training and testing datasets. RBPSpot outperformed
939 the remaining two software, iDeepE and DeepCLIP for all the combinations of datasets, for all the
940 considered performance metrics. Like the previous benchmarking study, here also RBPSpot scored
941 the highest among all the software for all the combinations of datasets with a remarkable
942 consistency. As transpires from the kernel density plots in Figure 9, RBPSpot maintained its least
943 variability and dispersion of performance scores and continued to display its strong balance in
944 detecting the positive and negative instances with high and similar level of precision. This was
945 reflected by high scoring on all the four parameters of performance resulting into consistently
946 highest MCC values, which confirmed the robustness of the algorithm. Also, it was observed that
947 performance of all the compared software was better when RBPSpot dataset was used for training.
948 The original implementation of iDeepE and DeepCLIP have used Graphprot dataset. Both these

949 software performed better when their original dataset for model building was replaced by RBPSpot
950 training dataset, underscoring better and more realistic composition of RBPSpot dataset.

951

952 The benchmarking done here stands among one of the most comprehensive ones. It looked into
953 various aspects of performances and has involved a large number of RBPs for comparison as well
954 as evaluated the role of datasets in performance. RBPSpot consistently scored high across all the
955 comparative tests and clearly outperformed the compared tools. The full details and data for the
956 benchmarking studies are given in supplementary Data 1 Sheet 10-15.

957

958 **Structural and molecular dynamics analysis supports the RBP binding site models**

959 Depending upon the availability of complete experimentally validated 3D structures in PDB
960 database, structures for 13 RBPs (IGF2BP1, DIS3L2, CNBP, SRSF3, FKBP4, KHDRBS1,
961 LIN28A, CAPRIN2, DICER1, GTF2F1, HNRNPC, CPSF6 and AGO2) were selected for the
962 structural interaction analysis for the identified binding sites (43). In order to examine
963 conformational variations of the RBPs within the hydrated controlled environment, the root-mean-
964 square deviation (RMSD) of the atomic positions of RNAs containing motif with respect to RBP
965 backbone were calculated and compared with the RNA complexes without the prime motif. In
966 comparative analysis of RMSD measures these RBPs complexes were considered with three
967 different RNA sequences for each RBP. These sequences were randomly selected from positive
968 datasets having 75 bases flanking regions. To analyze the structural behavior of RBPs and their
969 complexes, 20 ns simulation job was performed. For this purpose, selected RBPs and complexes
970 were immersed in the cubic boxes of varying dimensions based on the system size. Prior to the
971 energy minimization process, different charged molecules like NA^+ or Cl^- were added to neutralize
972 the system (44).

973

974 Once the simulation was finished, the last step was to analyze the simulation result in term of
975 RMSD plot during the course of simulation for 20ns. RMS module in GROMACS was executed
976 while choosing "Backbone" for least-squares fitting and "RNA_Heavy" for the RMSD calculation.
977 By doing so, the overall rotation and translation of the protein was removed via fitting and the
978 RMSD reported about how much the RNA position varied relative to the protein. This is considered
979 as a good indicator of how well the binding pose was preserved during the simulation. Comparative
980 analysis of RMSD trajectories of 13 different RBPs-RNA complexes with three replicates each for
981 the two conditions clearly suggested that the presence of the identified prime motifs was giving
982 stability to the RBP-RNA complexes (Figure 10).

983

984 For example, in case of AGO2, on comparative analysis of RMSD value of the AGO2-RNA first
985 sequence complex with the prime motif, the value ranged from 0.1 nm to 0.7 nm and got stabilized
986 at 0.5 nm whereas RMSD values for the complex without the motif ranged from 0.1 nm to 1.7 nm
987 and got stabilized at 1.5 nm, which was less stable. Similarly the second pair with motif had RMSD
988 ranging from 0.1 nm to 1.7 nm which got stabilized at 0.6 nm, whereas the same pair without the
989 prime motif ranged had RMSD ranging from 0.3 nm to 1.4 nm and got stabilized at 1.4 nm. For the
990 third pair, the AGO2-RNA complex of the third sequence with the prime motif showed deviation
991 from 0.1 nm to 1.0 nm and got settled at 0.7 nm whereas the same sequence without the prime motif
992 showed deviation from 0.0 nm to 2.0 nm and settled at 1.4 nm. In all the three cases of AGO-RNA
993 complexes, the sequence with the prime motif was found to be more stable when compared to the
994 one without the motif in the dynamic environment. Similar pattern was observed for all the 13 RBP
995 and their triplicate pairs. Details can be found in Table 1.

996

997 In the nutshell, the structural molecular dynamics study supported the identified binding spots for
998 the RBP where it was clearly evident that the identified binding motif provided structural stability to
999 the considered RBP-RNA complexes.

1000

1001 **Application: SARS-Cov2 genome was found to host RBP binding sites**

1002 Most of the deadly viruses are RNA viruses which exploit the host proteins to replicate, spread and
1003 survive. The best living example is nSARS-CoV-2. The emergence of the novel human corona-virus
1004 SARS-CoV-2 in Wuhan, China has caused a pandemic of respiratory disease (Covid19). The big
1005 scientific concern is that to this date very scarce and uncertain molecular information is available
1006 about the Covid19 patient's molecular system as not much high-throughput studies have been
1007 carried out so far. There is almost absolutely no information on the host RBPs response during
1008 Covid19 infection despite of the fact that all such virus essentially require host RBPs to survive and
1009 replicate And RBP-RNA interaction studies hold prime importance in this regard also.

1010

1011 Therefore, we scanned the SARS-CoV-2 genome through RBPSpot to find the binding sites for
1012 RBPs which could have therapeutic value. Interestingly, out of 131 different model we found 22
1013 different binding sites for 7 different RBPs (AIFM1 (2), BUD13 (3), CELF2 (4), RBM6 (3), UPF1
1014 (2), TARBP2 (4) and KHSRP (4)) (Figure 11). Among these, AIFM1 interaction with viral
1015 polymerases in influenza virus infected cells is well studied (45). These all binding sites were found
1016 on anti-sense strand of the genome whose importance is for viral replication. During the infection,
1017 majority of immunoprecipitated RNA of Coronavirus were found originating from the anti-sense
1018 strand (46). Therefore, there is a possibility that these RBPs are helping in it's transcription by
1019 binding to it's negative strand. To check the stability of these RBPs with their binding site we also
1020 performed MD simulations study on two different sequence forms for each identified binding site

1021 (One with the binding site and another without it). Prior to this, we obtained complete 3D structures
1022 for AIFM1 and UPF1 from PDB and modeled the remaining five RBPs through homology
1023 modeling due to lack of complete defined structures for them. After modeling we evaluated the built
1024 3D structure models using SAVES v6.0 (structure Activity validation server). Five RBPs PDB
1025 structures namely AIFM1, BUD13, CELF2, TARBP2 and UPF1 passed through verification filter
1026 like PROCHECK and WHATCHECK except KHSRP and RBM6. When we analyzed the model
1027 structure for KHSRP and RBM6 with both program it gives 80.9% and 83.5% residues in allowed
1028 regions in the Ramachandran plot but for good quality model, over 90% residues are expected in the
1029 most favored region and lack of loop filtering causing side-chain packing inaccuracies.
1030 Subsequently, on analyzing the RMSD graph (Supplementary Figure 2) for all the seven RBPs it
1031 was found that that five out of seven RBP-RNA complexes were stable with prime motif compared
1032 to the RBP-RNA complexes counterpart without the main motif. This part of the study was done
1033 just to showcase the application of the developed approach. The finding made in this section may be
1034 used for further study for Covid research groups.

1035

1036

1037 **Conclusion**

1038 A living system is a continuous outcome of the regulatory setup working for that system in the
1039 background. RNA binding proteins define one such critical regulatory component of the system
1040 which is present at almost every post-transcriptional regulatory event but about which our
1041 understanding is still nascent and evolving. How they select their targets and carry out interactions
1042 in functional manner is largely ambiguous. With the advent of high-throughput techniques like
1043 CLIP-seq and interactome capture, the information on genes recognized as RBPs and their
1044 interactions are growing continuously. Such high-throughput data on interactions are very valuable

1045 resources to construct the interaction models. The present study used the same from CLIP-seq
1046 experiments. However, there are several other critical factors involved which are required to be
1047 build these interactions models with high accuracy. This involves proper negative data-sets
1048 screening, appropriate motif discovery strategy, and contextual information derivation. All of them
1049 are interconnected with each other and success of any such RBP binding site discovery tool depends
1050 highly on this. Without proper datasets, correct binding specific motif candidates are hard to be
1051 found. The motif finding step itself needs to consider the sparse nature of RBP binding sites and
1052 need to anchor correctly so that correct surrounding could be recognized to provide the contextual
1053 information. Otherwise, such motifs occur frequently even in the non-binding regions, and wrong
1054 context may easily compromise the accuracy. When all these information are applied through
1055 effective machine learning algorithms, consistently high level accuracy is achievable. It was
1056 comprehensively and comparatively benchmarked against some recent tools where it outperformed
1057 them consistently across a wide number of datasets and RBPs. It also showcased that when a DNN
1058 is trained properly on suitable properties with appropriate biological insights, the developed system
1059 could easily outperform much complex deep-learning based approaches where such learning is done
1060 through automated feature extraction process using complex layers like CNN and LSTM etc. Such
1061 complex deep learning approach may be suitable for unstructure data where features could not be
1062 identified easily. However, when features are identifiable and structured, simpler machine learning
1063 approaches can outperform them easily. The developed approach in this study, RBPSpot, can
1064 identify the binding sites of existing RBPs in human system as well as it becomes one of few tools
1065 where users can put their own data and raise their own models for any species and any RBP. The
1066 software is freely available as a webserver as well as as an standalone program.

1067

1068 From here, we visualize that incorporation of spatio-temporal and other interactome network
1069 information for RBPs as the another dimension to explore to further improve our understanding on

1070 RBP-RNA interactions. This is something which still remains largely unaddressed. Some
1071 encouraging recent developments have happened (47,48) which promise that incorporation of back-
1072 end network and interaction information on RBP RNA interactions could add more value towards
1073 recognition of functional and dynamic nature of RBP RNA interactions which could further boost
1074 interaction spot identification process. Also, the findings made here from the contextual information
1075 like CG enrichment in the flanking regions must be explored further for their functional roles
1076 associated with such binding sites. RNA modifications on CG and likewise other important
1077 contextually important factors found in this study may further provide reasoning for spatio-temporal
1078 nature of these interactions which would mark another level of development in our understanding
1079 towards RBP RNA interactions and regulation.

1080

1081 **Declarations**

1082 **Availability of data and materials**

1083 All the secondary data used in this study were publicly available and their due references and
1084 sources have been provided. All data and information generated/used, methodology related details
1085 etc have also been made available in the supplementary data files provided along with and also
1086 made available through the related open access server at <https://scbb.ihbt.res.in/RBPSpot/>. The
1087 software has also been made available at Github at: <https://github.com/SCBB-LAB/RBPSpot>

1088

1089 **Competing interests**

1090 The authors declare that they have no competing interests.

1091

1092 **Funding**

1093 RS is thankful to Department of Biotechnology, Govt. of India for supporting this study through
1094 grant in Big Data analysis[Grant number: BT/PR16331/BID 17/589/2016 (GAP-0228)] to RS.

1095

1096 **Authors' contributions**

1097 NKS and SG carried out the computational part and benchmarking of the study. PK developed the
1098 FM-Index and BWT based inexact k-mer search script. AK carried out the structural analysis and
1099 molecular dynamics simulation. UKP helped in statistical analysis. RS conceptualized, designed,
1100 analyzed and supervised the entire study. NKS, SG, AK and RS wrote the MS.

1101

1102 **Acknowledgments**

1103 We are thankful to the Director, CSIR-IHBT, for his kind support. We are thankful to Dr. Indu
1104 Gangwar for her inputs for the study. We are thankful to DBT for the funding support they gave for
1105 this project. NKS is thankful to CSIR for financial support as project associateship. UKP and PK
1106 are thankful to ICAR, New Delhi for providing support in Ph.D. UKP, NKS, and PK are thankful to
1107 Academy of Scientific and Innovative Research (AcSIR).

1108

1109 **Ethics approval and consent to participate**

1110 Not applicable.

1111

1112 **Consent for publication**

1113 Not applicable.

1114

1115 **References**

- 1116 1. Gerstberger S., Hafner M., Tuschl T. 2014. A census of human RNA-binding proteins.
1117 *Nature Reviews Genetics* 15:829–845.
- 1118 2. Castello,A., Hentze,M.W. and Preiss,T. (2015) Metabolic Enzymes Enjoying New
1119 Partnerships as RNA-Binding Proteins. *Trends Endocrinol Metab*, **26**, 746–757.
- 1120 3. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S.,
1121 Albu,M., Zheng,H., Yang,A., *et al.* (2013) A compendium of RNA-binding motifs for
1122 decoding gene regulation. *Nature*, **499**, 172–177.
- 1123 4. Cook KB., Kazan H., Zuberi K., Morris Q., Hughes TR. 2011. RBPDB: a database of RNA-
1124 binding specificities. *Nucleic Acids Research* 39:D301–D308.
- 1125 5. Khorshid M., Rodak C., Zavolan M. 2011. CLIPZ: a database and analysis environment for
1126 experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Research*
1127 39:D245-252.
- 1128 6. Hu,B., Yang,Y.-C.T., Huang,Y., Zhu,Y. and Lu,Z.J. (2017) POSTAR: a platform for
1129 exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic*
1130 *Acids Res*, **45**, D104–D114.
- 1131 7. Yang,Y.-C.T., Di,C., Hu,B., Zhou,M., Liu,Y., Song,N., Li,Y., Umetsu,J. and Lu,Z.J. (2015)
1132 CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, **16**, 51.
- 1133 8. Kazan,H., Ray,D., Chan,E.T., Hughes,T.R. and Morris,Q. (2010) RNAcontext: A New
1134 Method for Learning the Sequence and Structure Binding Preferences of RNA-Binding
1135 Proteins. *PLOS Computational Biology*, **6**, e1000832.
- 1136 9. Gardner,P.P. and Giegerich,R. (2004) A comprehensive comparison of comparative RNA
1137 structure prediction approaches. *BMC Bioinformatics*, **5**, 140.

- 1138 10. Paz,I., Kosti,I., Ares,M., Cline,M. and Mandel-Gutfreund,Y. (2014) RBPmap: a web server
1139 for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res*, **42**, W361–W367.
- 1140 11. Weyn-Vanhentenryck,S.M. and Zhang,C. (2016) mCarts: Genome-Wide Prediction of
1141 Clustered Sequence Motifs as Binding Sites for RNA-Binding Proteins. *Methods Mol Biol*,
1142 **1421**, 215–226.
- 1143 12. Maticzka, D., Lange, S.J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding
1144 preferences of RNA-binding proteins. *Genome Biol* *15*, R17.
- 1145 13. Yu,H., Wang,J., Sheng,Q., Liu,Q. and Shyr,Y. (2019) beRBP: binding estimation for human
1146 RNA-binding proteins. *Nucleic Acids Res*, **47**, e26.
- 1147 14. Alipanahi,B., DeLong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence
1148 specificities of DNA- and RNA-binding proteins by deep learning. *Nature Biotechnology*,
1149 **33**, 831–838.
- 1150 15. Pan,X. and Shen,H.-B. (2017) RNA-protein binding motifs mining with a new hybrid deep
1151 learning based cross-domain knowledge integration approach. *BMC Bioinformatics*, **18**, 136.
- 1152 16. Pan, X., and Shen, H.-B. (2018). Predicting RNA-protein binding sites and motifs through
1153 combining local and global deep convolutional neural networks. *Bioinformatics* *34*, 3427–
1154 3436.
- 1155 17. Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of RNA-protein sequence
1156 and structure binding preferences using deep convolutional and recurrent neural networks.
1157 *BMC Genomics* *19*, 511.
- 1158 18. Ghanbari, M., and Ohler, U. (2020). Deep neural networks for interpreting RNA-binding
1159 protein target preferences. *Genome Res* *30*, 214–226.
- 1160 19. Grønning, A.G.B., Doktor, T.K., Larsen, S.J., Petersen, U.S.S., Holm, L.L., Bruun, G.H.,
1161 Hansen, M.B., Hartung, A.-M., Baumbach, J., and Andresen, B.S. (2020). DeepCLIP:

- 1162 predicting the effect of mutations on protein-RNA binding with deep learning. *Nucleic*
1163 *Acids Res* **48**, 7099–7118.
- 1164 20. Heikham,R. and Shankar,R. (2010) Flanking region sequence information to refine
1165 microRNA target predictions. *J Biosci*, **35**, 105–118.
- 1166 21. Černý,J., Božíková,P., Svoboda,J. and Schneider,B. (2020) A unified dinucleotide alphabet
1167 describing both RNA and DNA structures. *Nucleic Acids Research*, **48**, 6367–6381.
- 1168 22. Jankowsky,E. and Harris,M.E. (2015) Specificity and nonspecificity in RNA–protein
1169 interactions. *Nature Reviews Molecular Cell Biology*, **16**, 533–544.
- 1170 23. A,G. and M,G. (2011) The role of RNA sequence and structure in RNA--protein
1171 interactions. *Journal of molecular biology*, **409**.
- 1172 24. Zhou,T., Yang,L., Lu,Y., Dror,I., Dantas Machado,A.C., Ghane,T., Di Felice,R. and Rohs,R.
1173 (2013) DNASHape: a method for the high-throughput prediction of DNA structural features
1174 on a genomic scale. *Nucleic Acids Research*, **41**, W56–W62.
- 1175 25. Ryan M. (2021). *Deep Learning with Structured Data* [Book] . Manning Publications.
- 1176 26. Li J-H., Liu S., Zhou H., Qu L-H., Yang J-H. 2014. starBase v2.0: decoding miRNA-
1177 ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq
1178 data. *Nucleic Acids Research* **42**:D92-97.
- 1179 27. Lorenz,R., Bernhart,S.H., Höner zu Siederdisen,C., Tafer,H., Flamm,C., Stadler,P.F. and
1180 Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, **6**, 26.
- 1181 28. Chen,Y.-W. and Lin,C.-J. (2006) Combining SVMs with Various Feature Selection
1182 Strategies. In Guyon,I., Nikravesh,M., Gunn,S., Zadeh,L.A. (eds), *Feature Extraction:*
1183 *Foundations and Applications*, Studies in Fuzziness and Soft Computing. Springer, Berlin,
1184 Heidelberg, pp. 315–324.
- 1185 29. Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient
1186 (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**,

- 1187 6.
- 1188 30. Tuszynska I., Bujnicki JM. 2011. DARS-RNP and QUASI-RNP: new statistical potentials
1189 for protein-RNA docking. *BMC bioinformatics* 12:348.
- 1190 31. Chen A-J., Paik J-H., Zhang H., Shukla SA., Mortensen R., Hu J., Ying H., Hu B., Hurt J.,
1191 Farny N., Dong C., Xiao Y., Wang YA., Silver PA., Chin L., Vasudevan S., Depinho RA.
1192 2012. STAR RNA-binding protein Quaking suppresses cancer via stabilization of specific
1193 miRNA. *Genes & Development* 26:1459–1472.
- 1194 32. Berendsen HJC., van der Spoel D., van Drunen R. 1995. GROMACS: A message-passing
1195 parallel molecular dynamics implementation. *Computer Physics Communications* 91:43–56.
- 1196 33. Duan Y., Wu C., Chowdhury S., Lee MC., Xiong G., Zhang W., Yang R., Cieplak P., Luo
1197 R., Lee T., Caldwell J., Wang J., Kollman P. 2003. A point-charge force field for molecular
1198 mechanics simulations of proteins based on condensed-phase quantum mechanical
1199 calculations. *Journal of Computational Chemistry* 24:1999–2012.
- 1200 34. Jorgensen WL., Chandrasekhar J., Madura JD., Impey RW., Klein ML. 1983. Comparison of
1201 simple potential functions for simulating liquid water. *The Journal of Chemical Physics*
1202 79:926–935.
- 1203 35. Hess B., Bekker H., Berendsen HJC., Fraaije JGEM. 1997. LINCS: A linear constraint
1204 solver for molecular simulations. *Journal of Computational Chemistry* 18:1463–1472.
- 1205 36. Gunsteren WFV., Berendsen HJC. 1988. A Leap-frog Algorithm for Stochastic Dynamics.
1206 *Molecular Simulation* 1:173–185.
- 1207 37. Vandenbon,A., Kumagai,Y., Akira,S. and Standley,D.M. (2012) A novel unbiased measure
1208 for motif co-occurrence predicts combinatorial regulation of transcription. *BMC Genomics*,
1209 **13 Suppl 7**, S11.

- 1210 38. Dassi,E., Re,A., Leo,S., Tebaldi,T., Pasini,L., Peroni,D. and Quattrone,A. (2014) AURA 2:
1211 Empowering discovery of post-transcriptional networks. *Translation (Austin)*, **2**, e27738.
- 1212 39. Yuan,H., Kshirsagar,M., Zamparo,L., Lu,Y. and Leslie,C.S. (2019) BindSpace decodes
1213 transcription factor binding signals by large-scale sequence embedding. *Nature Methods*, **16**,
1214 858–861.
- 1215 40. Vorontsov IE., Kulakovskiy IV., Makeev VJ. 2013. Jaccard index based similarity measure
1216 to compare transcription factor binding site models. *Algorithms for molecular biology: AMB*
1217 **8**:23.
- 1218 41. Xue,C., Li,F., He,T., Liu,G.-P., Li,Y. and Zhang,X. (2005) Classification of real and pseudo
1219 microRNA precursors using local structure-sequence features and support vector machine.
1220 *BMC Bioinformatics*, **6**, 310.
- 1221 42. Hartl,D., Krebs,A.R., Grand,R.S., Baubec,T., Isbel,L., Wirbelauer,C., Burger,L. and
1222 Schübeler,D. (2019) CG dinucleotides enhance promoter activity independent of DNA
1223 methylation. *Genome Res*, **29**, 554–563.
- 1224 43. Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., ... &
1225 Bourne, P. E. (2010). The RCSB Protein Data Bank: redesigned web site and web services.
1226 *Nucleic acids research*, 39(suppl_1), D392-D401.
- 1227 44. Pfeiffer, S., Fushman, D., & Cowburn, D. (1999). Impact of Cl⁻ and Na⁺ ions on simulated
1228 structure and dynamics of β ARK1 PH domain. *Proteins: Structure, Function, and*
1229 *Bioinformatics*, 35(2), 206-217.
- 1230 45. Bradel-Tretheway,B., Mattiacio,J., Krasnoselsky,A., Stevenson,C., Purdy,D., Dewhurst,S.
1231 and Katze,M. (2011) Comprehensive Proteomic Analysis of Influenza Virus Polymerase
1232 Complex Reveals a Novel Association with Mitochondrial Proteins and RNA Polymerase
1233 Accessory Factors. *Journal of virology*, **85**, 8569–81.

- 1234 46. Hackbart,M., Deng,X. and Baker,S.C. (2020) Coronavirus endoribonuclease targets viral
 1235 polyuridine sequences to evade activating host sensors. *PNAS*, **117**, 8094–8103.
- 1236 47. Pradhan,U.K., Anand,P., Sharma,N.K., Kumar,P., Kumar,A., Pandey,R., Padwad,Y. and
 1237 Shankar,R. (2020) Various RNA-binding proteins and their conditional networks explain
 1238 miRNA biogenesis.(Under review).
- 1239 48. Mukherjee,N., Wessels,H.-H., Lebedeva,S., Sajek,M., Ghanbari,M., Garzia,A.,
 1240 Munteanu,A., Yusuf,D., Farazi,T., Hoell,J.I., et al. (2019) Deciphering human
 1241 ribonucleoprotein regulatory networks. *Nucleic Acids Research*, **47**, 570–581.

1242

1243 Tables

1244 **Table 1:** Table for RMSD value for selected 13 RBPs complexes with and without the prime motif.
 1245 The identified prime motifs were found statistically enriched in the target sequences when
 1246 compared to random regions. Molecular dynamics studies with and without these motifs clearly
 1247 suggested their important role in binding where they were found responsible for stable complex
 1248 formation between RBP and RNA.

| RBPs name | Sequence name | RMS Deviation range value with motif (nm) | Stablized_RMSD value with motif (nm) | RMS Deviation range value without motif (nm) | Stabilized RMSD value without motif (nm) |
|----------------|---------------|---|--------------------------------------|--|--|
| AGO2 | RNA_seq1 | 0.1-0.7 | 0.5 | 0.1-1.7 | 1.5 |
| | RNA_seq2 | 0.1-1.7 | 0.6 | 0.3-1.4 | 1.4 |
| | RNA_seq3 | 0.1-1.0 | 0.7 | 0.1-2.0 | 1.4 |
| CAPRIN2 | RNA_seq1 | 0.1-0.3 | 0.2 | 0.6-2.1 | 1.9 |
| | RNA_seq2 | 0.3-1.8 | 1.3 | 0.7-2.2 | 1.5 |
| | RNA_seq3 | 0.2-2.1 | 1.8 | 0.1-2.9 | 2.7 |
| CNBP | RNA_seq1 | 0.1-1.3 | 1.1 | 0.3-4.8 | 4.3 |
| | RNA_seq2 | 0.3-1.8 | 0.5 | 0.3-2.5 | 2.0 |
| | RNA_seq3 | 0.1-2.7 | 2.5 | 0.1-4.7 | 3.4 |
| CPSF6 | RNA_seq1 | 0.3-1.2 | 1.0 | 0.3-3.2 | 3.0 |
| | RNA_seq2 | 0.3-2.0 | 1.8 | 0.3-2.8 | 2.4 |
| | RNA_seq3 | 0.1-2.9 | 2.7 | 0.1-4.8 | 4.1 |

| | | | | | |
|----------------|----------|---------|-----|----------|-----|
| | | | | | |
| DICER1 | RNA_seq1 | 0.3-1.3 | 1.0 | 0.3-3.5 | 2.9 |
| | RNA_seq2 | 0.3-1.5 | 1.5 | 0.3-2.7 | 2.0 |
| | RNA_seq3 | 0.1-2.5 | 2.3 | 0.1-4.8 | 4.2 |
| | | | | | |
| DIS3L2 | RNA_seq1 | 0.1-0.3 | 0.3 | 0.3-2.5 | 2.5 |
| | RNA_seq2 | 0.3-1.8 | 1.2 | 0.3-2.4 | 2.0 |
| | RNA_seq3 | 0.1-2.3 | 1.8 | 0.1-4.2 | 4.0 |
| | | | | | |
| FKBP4 | RNA_seq1 | 0.3-1.5 | 1.3 | 0.3-1.2 | 1.3 |
| | RNA_seq2 | 0.1-1.7 | 1.1 | 0.1-2.5 | 2.2 |
| | RNA_seq3 | 0.1-1.9 | 0.5 | 0.1-2.3 | 2.2 |
| | | | | | |
| GTF2F1 | RNA_seq1 | 0.3-1.3 | 1.3 | 0.3-3.7 | 2.5 |
| | RNA_seq2 | 0.3-1.9 | 1.7 | 0.5-2.5 | 2.0 |
| | RNA_seq3 | 0.1-2.6 | 2.1 | 0.1-4.7 | 4.0 |
| | | | | | |
| HNRNPC | RNA_seq1 | 0.1-0.5 | 0.3 | 0.1-2.7 | 2.3 |
| | RNA_seq2 | 0.1-2.3 | 1.8 | 0.1-2.3 | 2.0 |
| | RNA_seq3 | 0.1-2.5 | 2.2 | 0.3-4.3 | 4.1 |
| | | | | | |
| IGF2BP1 | RNA_seq1 | 0.3-1.7 | 0.9 | 0.3-2.6 | 2.1 |
| | RNA_seq2 | 0.3-1.5 | 0.6 | 0.07-1.7 | 1.3 |
| | RNA_seq3 | 0.1-1.7 | 1.5 | 0.1-2.6 | 1.6 |
| | | | | | |
| KHDRBS1 | RNA_seq1 | 0.3-1.9 | 0.5 | 1.0-4.8 | 3.5 |
| | RNA_seq2 | 0.3-1.2 | 1.2 | 0.3-2.1 | 1.2 |
| | RNA_seq3 | 0.3-2.5 | 2.5 | 0.1-4.7 | 4.2 |
| | | | | | |
| LIN28A | RNA_seq1 | 0.3-1.4 | 0.7 | 0.3-3.0 | 2.2 |
| | RNA_seq2 | 0.5-2.1 | 1.5 | 0.5-3.0 | 1.6 |
| | RNA_seq3 | 0.1-2.2 | 2.2 | 0.1-4.9 | 3.1 |
| | | | | | |
| SRSF3 | RNA_seq1 | 0.2-0.5 | 0.8 | 0.5-3.7 | 3.0 |
| | RNA_seq2 | 0.3-1.2 | 1.2 | 0.5-2.5 | 2.1 |
| | RNA_seq3 | 0.1-3.0 | 2.5 | 1.0-4.8 | 3.2 |

1249

1250 **Figure legends**

1251 Figure 1: Detailed pipeline of the workflow. The image provides the brief outline of entire
1252 computation protocol implemented to develop accurate RBP RNA interaction model and identify
1253 the correct RBP binding sites across given RNA sequences. The process of model building starts
1254 from identifying prime motifs through k-mer spectrum search from the CLIP-seq regions where
1255 BWT/FM indexing based inexact search algorithm was implemented. The statistically enriched k-
1256 mers were expanded across all reporting sequences region till at least 70% similarity between them
1257 was present. The final prime motifs were established as the anchors. The flanking regions around

1258 such anchored prime motifs were used to derive the contextual information, together which worked
1259 as feature vector elements for discrimination using XGBoost and Deep-Learning.

1260

1261 Figure 2: The co-occurring motifs positional preference. The plots are showing the position specific
1262 existence of the co-occurring motifs with respect to the prime motif (coordinated at “0”). F-score
1263 values of other contextual features like position specific pentamers and heptamers distribution
1264 reflect this to some extent. Most of these RBPs exhibited some secondary motif which co-occured
1265 with the prime motif in a position specific manner.

1266

1267 Figure 3: Comparison between experimentally reported motifs and motif identified in the present
1268 study. Most of the previously reported motifs for the RBPs were detected by the approach
1269 presented in the current study. However, it also observed that several of previously reported motifs
1270 are not the prime motifs but comparatively cover lesser CLIP-seq data than the prime motifs
1271 identified in the present study. The last three columns show the matching motifs similar to the
1272 previously reported motifs, their status in CLIP-seq data coverage, and the corresponding motif
1273 rank.

1274

1275 Figure 4: F-score distribution of dinucleotide densities at different positional windows for the target
1276 regions and their flanks for Cluster#1 members. Context specific dinucleotide density distribution
1277 emerged among the most important features for all RBPs taken in this study. Their densities worked
1278 as important features at variable windows and distances for different RBPs. Here, Cluster# 1
1279 members data is shown. They shared high similarity among themselves for their prime binding
1280 motifs, yet their contextual information and density profiles differed a lot. Enriched contextual
1281 “CG” distribution of these regions was found consistently distinguished property for the regions

1282 binding the RBPs.

1283

1284 Figure 5: Assessment for three main properties in discriminating between the negative and positive
1285 instance,(A) Violin plot distribution of accuracy when dinucleotide, pentamer and heptamer were
1286 used alone for set A and set B RBP. (B) Impact of combination of the dinculetodie, pentamers, and
1287 heptamers properties based features. These features appeared highly additive, complementary to
1288 each other as the performance in accurately identifying the binding regions increases substantially
1289 as these are combined.

1290

1291 Figure 6: Performance metrics for RBPSpot. (A) First plot showing the accuracy, AUC, sensitivity,
1292 specificity, F1 score for the DNN model for set A RBPs. The second plot is showing the same
1293 metrics for the gradient boosting method. The third plot is showing the corresponding instances in
1294 the test, train and in total data for set A RBPs, (B) The first plot is showing the accuracy, AUC,
1295 sensitivity,specificity, F1 score for the deep learning models for Set B RBPs, where the second plot
1296 is showing the same metrics values for the gradient boosting method with Set B RBPs. The third
1297 plot is showing the number of instances in the test, train and in total data for Set B RBPs. RBPSpot
1298 scored highly on all the performance metrics where the most remarkable thing about it was its
1299 consistent performance across a large number of RBPs and dataset.

1300

1301 Figure 7:AUC/ROC plots for Set A RBPs. The AUC/ROC plots for the deep-learning models for
1302 some of the RBPs clearly showcase the robustness and highly reliable performance of the
1303 implemented DNN models.

1304

1305 Figure 8: Comparative bench-marking results of RBPSpot when compared to beRBP, DeepBind,

1306 RBPmap, iDeepE, and DeepCLIP for three different datasets. (A) Benchmarking result on RBPSpot
1307 dataset, (B) Graphprot dataset, and (C) beRBP dataset. Each these datasets performances was
1308 evaluated for various performance metrics where the heatmaps are for accuracy, F-1 score, and
1309 MCC values for each dataset for some of the evaluated RBPs. The rightmost plots are radar charts
1310 view of the average Accuracy, F1 score, and MCC attained by each software for the corresponding
1311 dataset. The last plot is the box plot which provides the average distribution of these metrics scores.
1312 From the plots it is clearly visible that for all these datasets and for almost all of the RBPs, RBPSpot
1313 consistently outperformed the compared tools for all the metrics. More all the radar plots it scored
1314 the highest and nearest to the isosceles triangle suggesting consistent and better average
1315 performance also. The box plot suggests that RBPSpot not only performed best in overall but also
1316 the dispersion of its various metric scores were much lesser than other compared tools. Some of
1317 these tools exhibited enormous variation in the distribution of their metric score suggesting unstable
1318 performance by them.

1319

1320 Figure 9: Performance benchmarking with different combinations of train and test datasets. In this
1321 part of performance benchmarking the impact of datasets was also evaluated. Since this part
1322 required rebuilding of RBP-RNA interaction models from the scratch and from the provided user
1323 defined data, only two other tools other than RBPSpot qualified this criteria (DeepCLIP and
1324 iDEEPE). These tools provide the capability to build new models from user given datasets. These
1325 tools were originally developed on Graphport dataset. Therefore, in this part of benchmarking
1326 RBPSpot and Graphprot datasets were considered and 4 different train-test datasets combinations
1327 were studied. Distributions for various performance metrics for the compared tools and the
1328 corresponding datasets have been given as Kernel density plots: (A) RBPSpot train and test, (B)
1329 RBPSpot train and graphprot test, (C) Graphprot train and test, and (D) Graphprot train and
1330 RBPSpot test. For every such combinations, the average performance metrics scores are given in the

1331 form of heatmap (E). The plots clearly underline that RBPSpot consistently outperforms the two
1332 tools for all the metrics on all these different combinations of train and test datasets, where again the
1333 consistent and precise performance of RBPSpot was an important observation, Consistently high
1334 MCC scoring by RBPSpot underlined it as a robust and balanced algorithm where dispersion in
1335 performance metrics was least. Also, the performance of all the compared tools increased when
1336 RBPSpot dataset was used in training, clearly suggesting the importance of having a right dataset.
1337 RBPSpot dataset presented here emerged as a better dataset for such studies.

1338

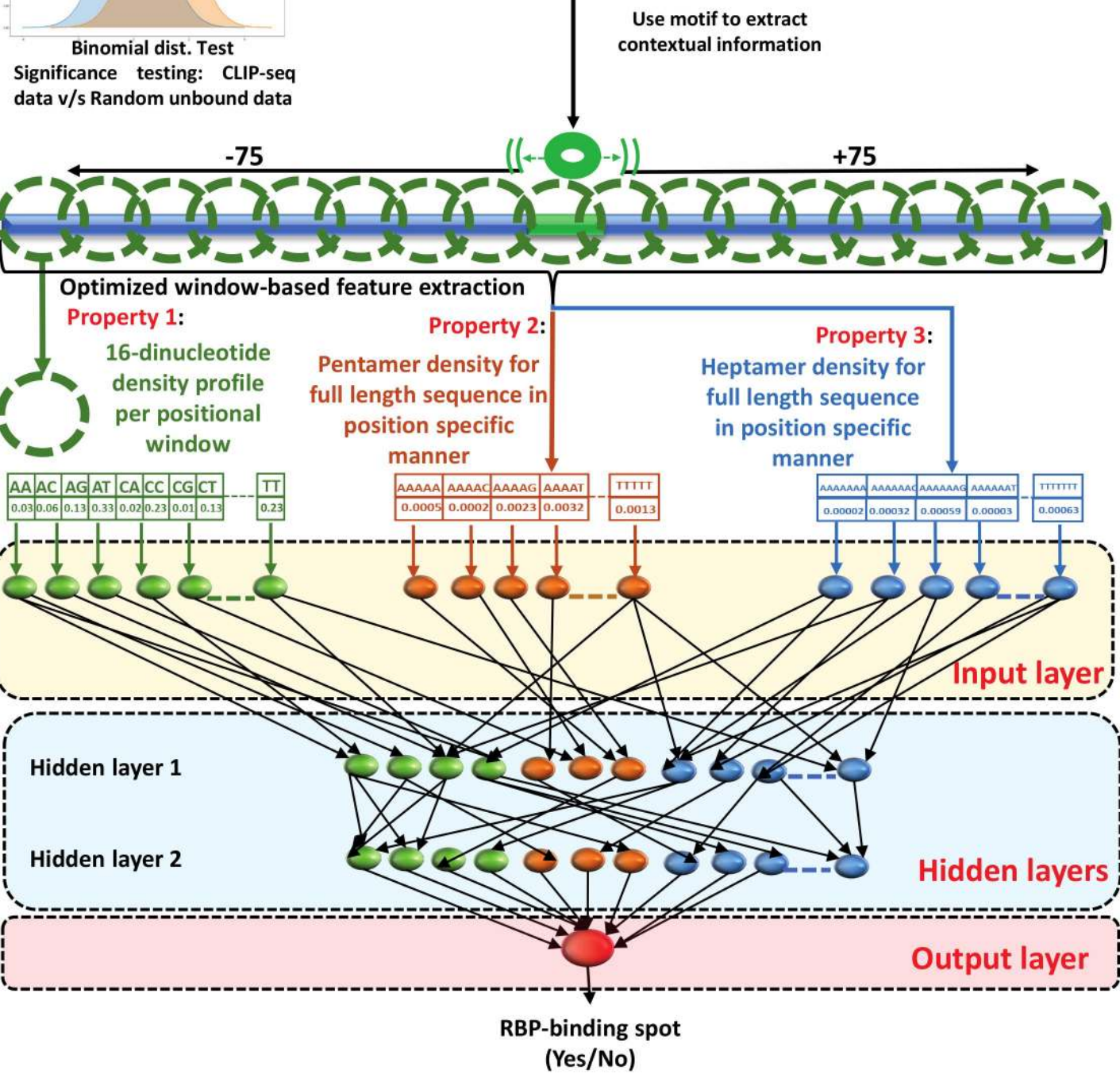
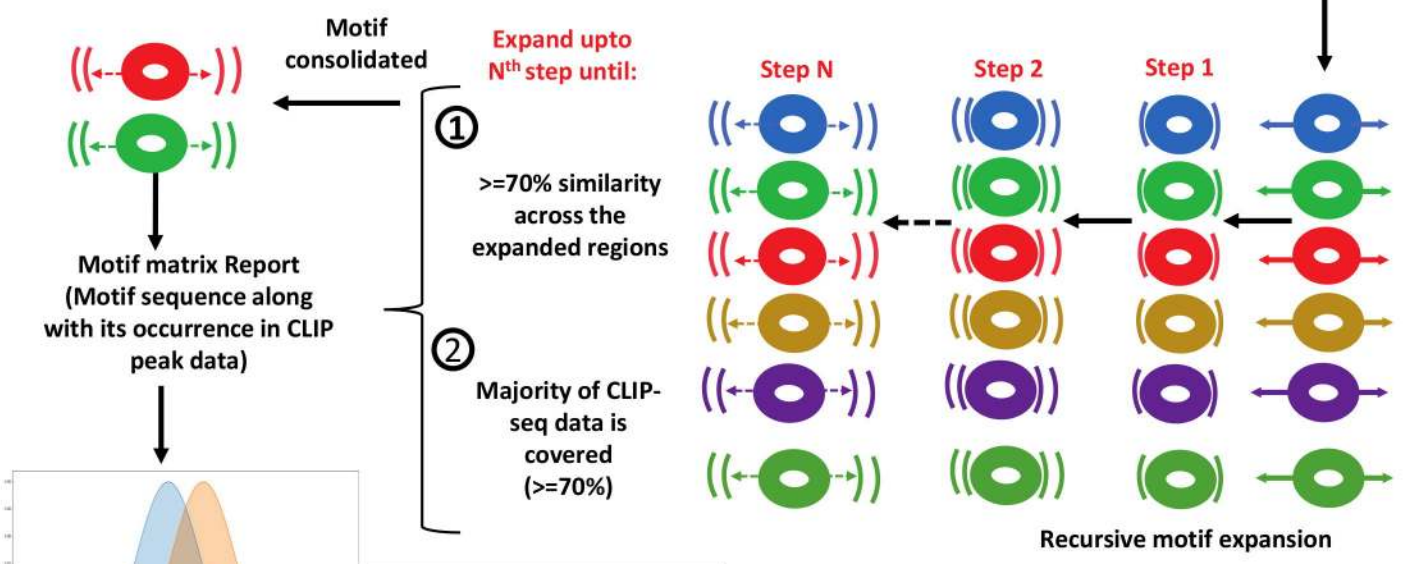
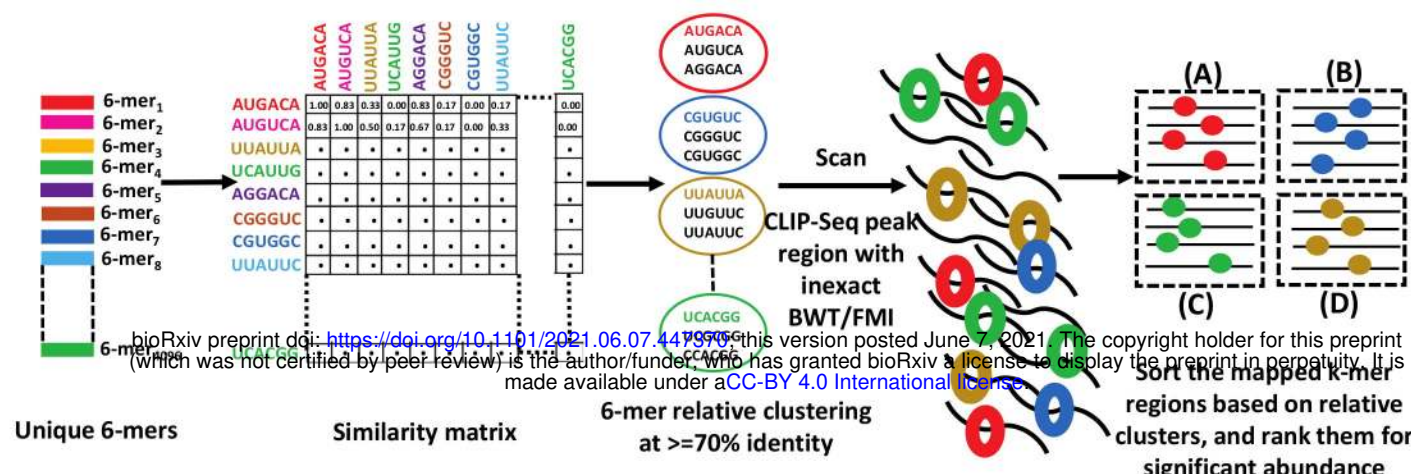
1339 Figure 10: Comparative time dependent root mean square deviations (RMSD) plots for 12 different
1340 RBP-RNA complexes of with and without the prime motif. The trajectory was measured at 300 K
1341 for the 20-ns. Trajectory arcs for RBP-complex of three randomly selected RNA sequences with
1342 motifs are shown in blue, green and violet spike arcs whereas trajectory spike-arcs for RBP-
1343 complex without motif were shown in orange, red and brown color. The complexes with the prime
1344 motifs were found much stable than their counterparts without the prime motif.

1345

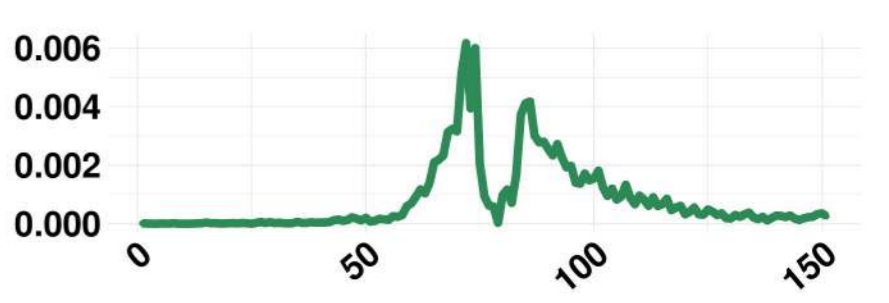
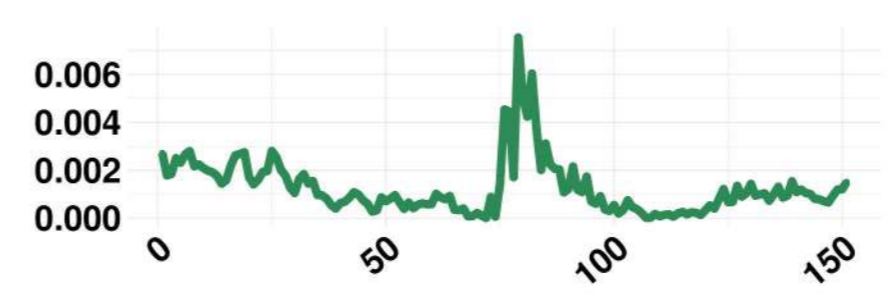
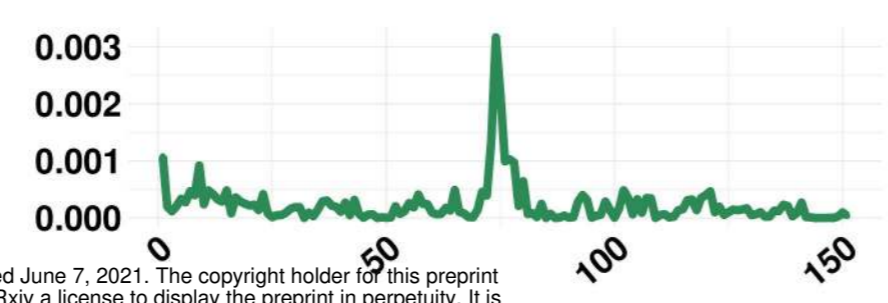
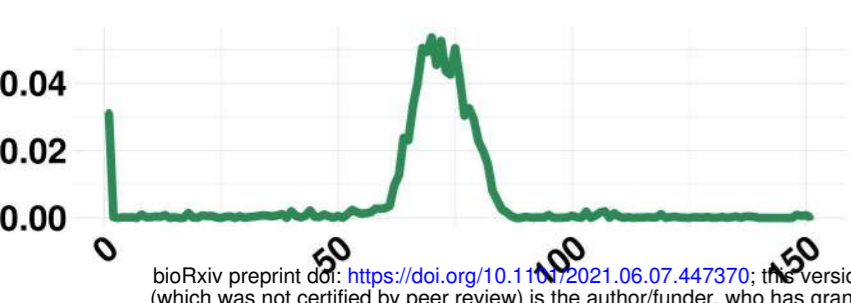
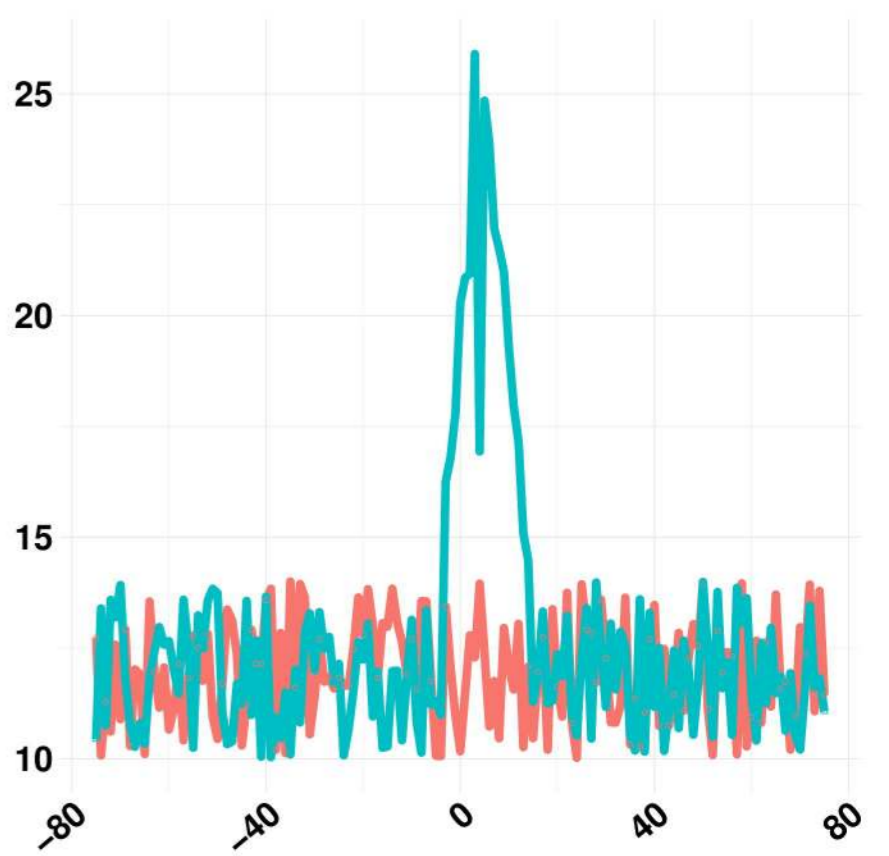
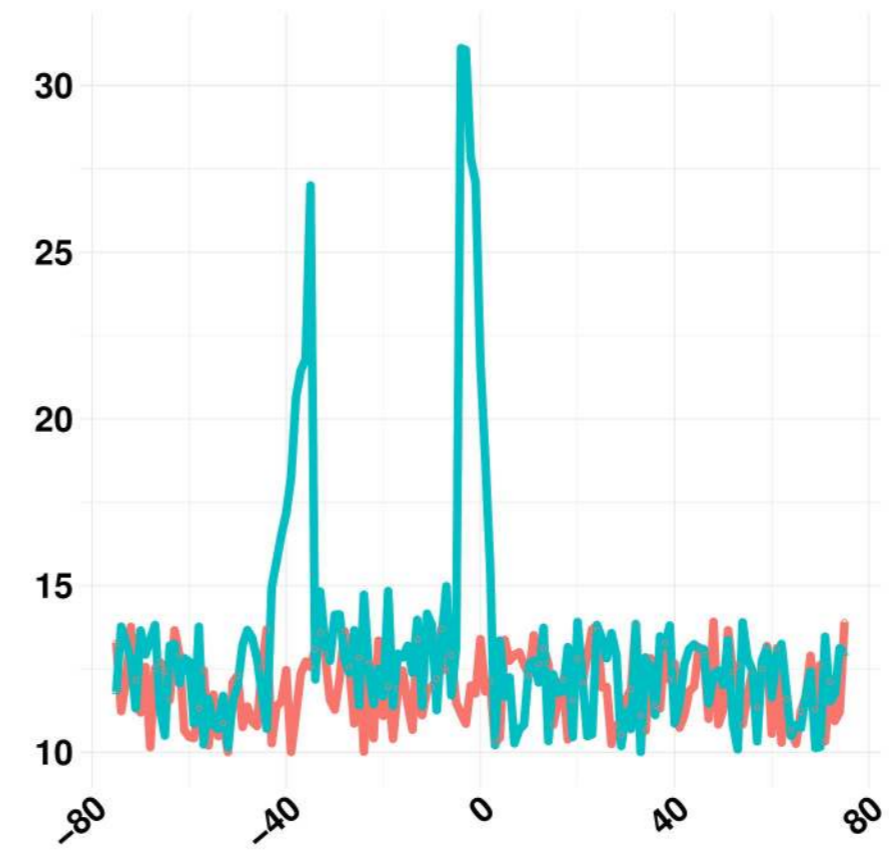
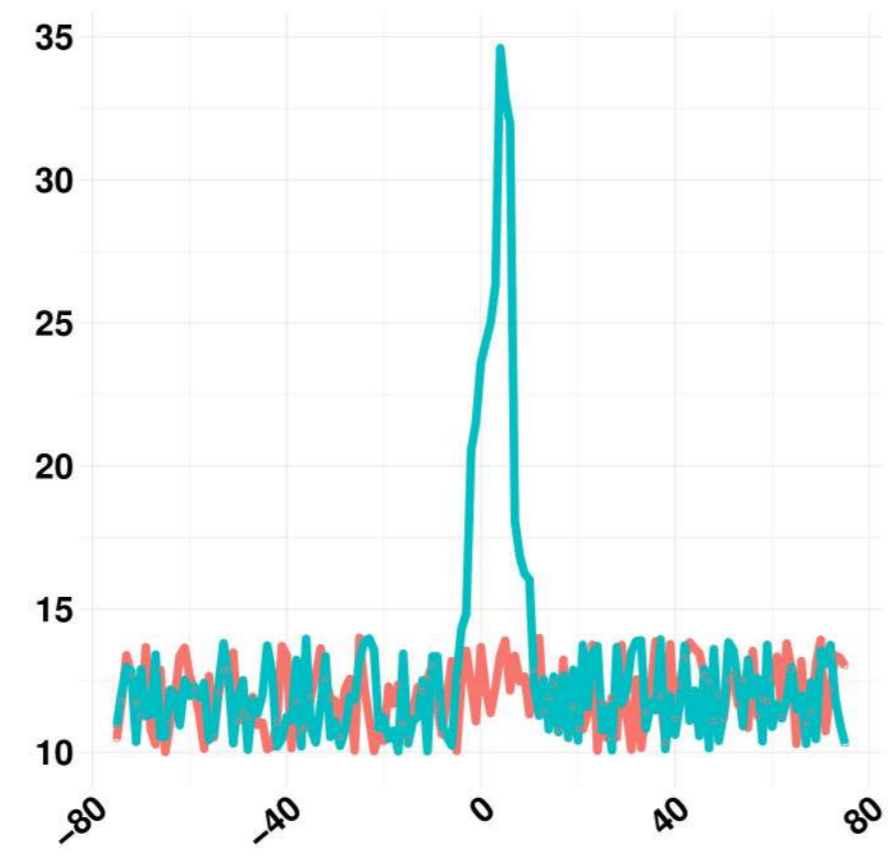
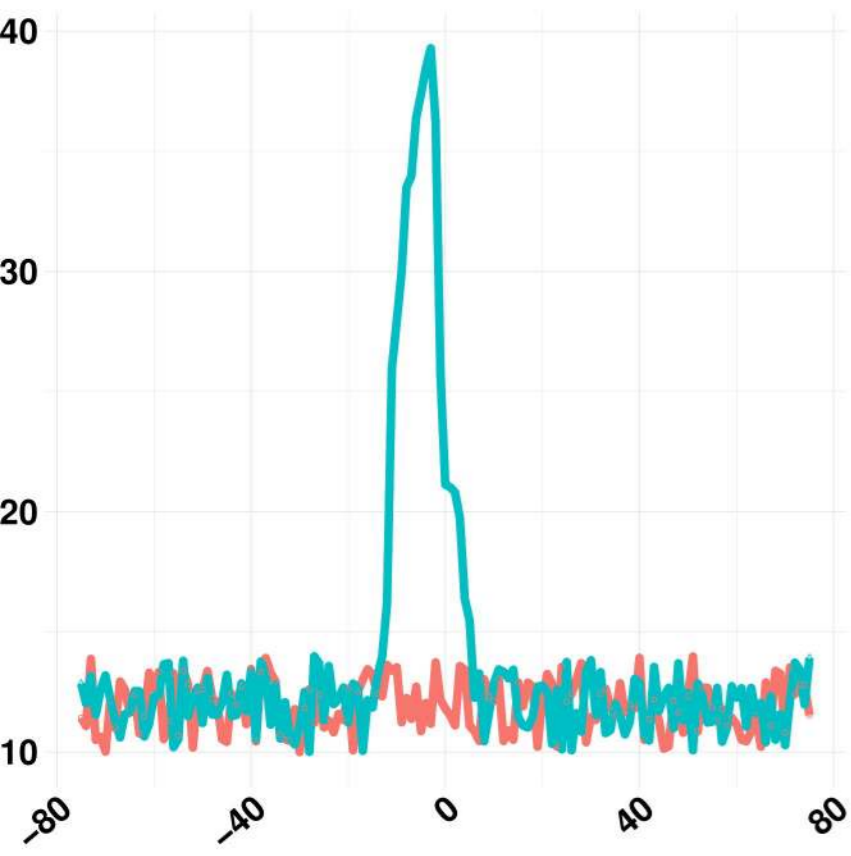
1346 Figure 11: Application of RBPSpot reports the binding sites for seven different RBP on nSARS-
1347 CoV2 genome. A total for 22 such binding sites were discovered across the nSARS-CoV2 genome,
1348 all which existed across the negative strand of the virus genome.

1349

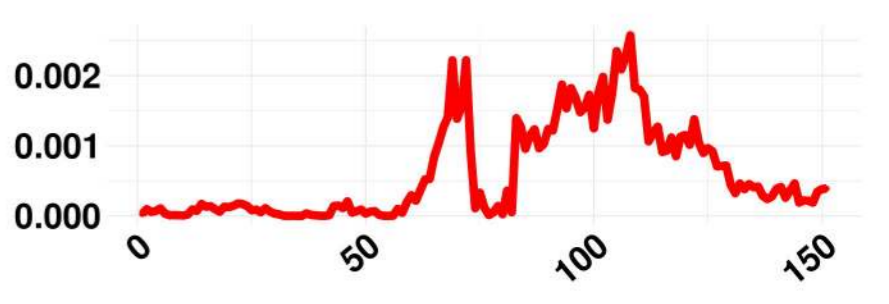
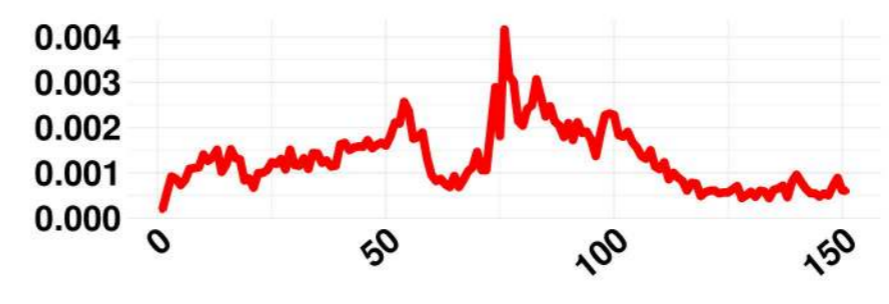
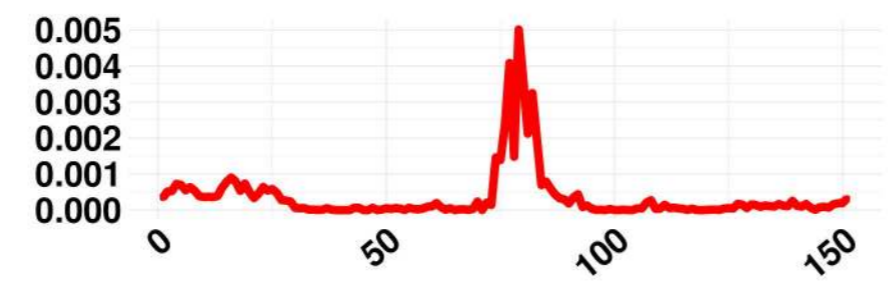
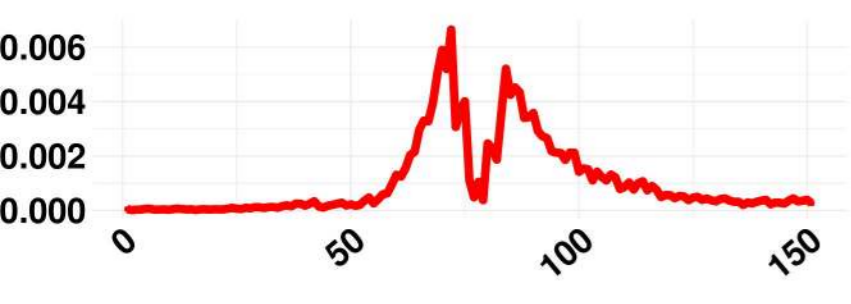
bioRxiv preprint doi: <https://doi.org/10.1101/2021.06.07.447970>; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Random Data Cross-linking Data Pentamer Heptamer



bioRxiv preprint doi: <https://doi.org/10.1101/2021.06.07.447370>; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

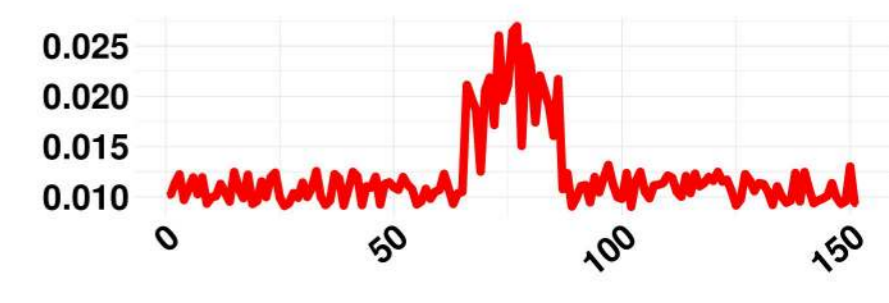
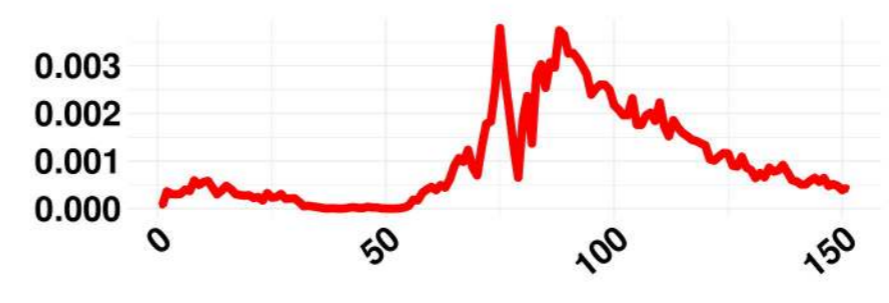
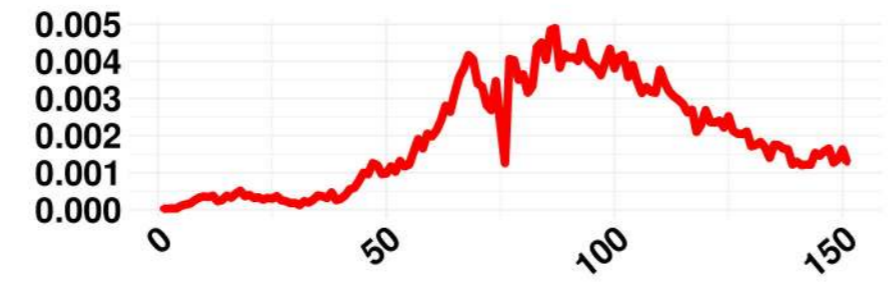
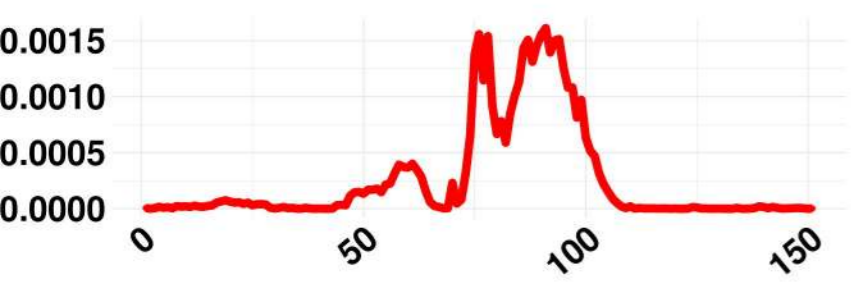
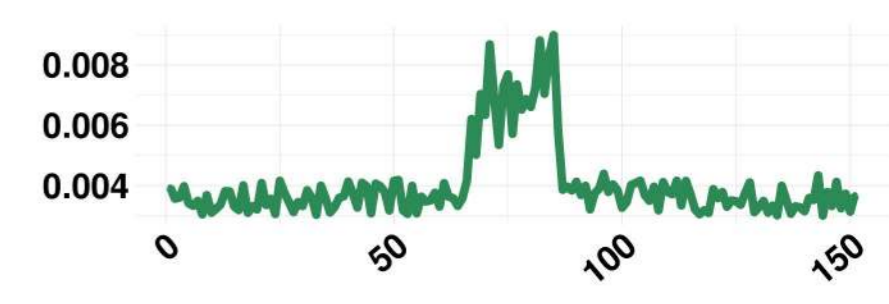
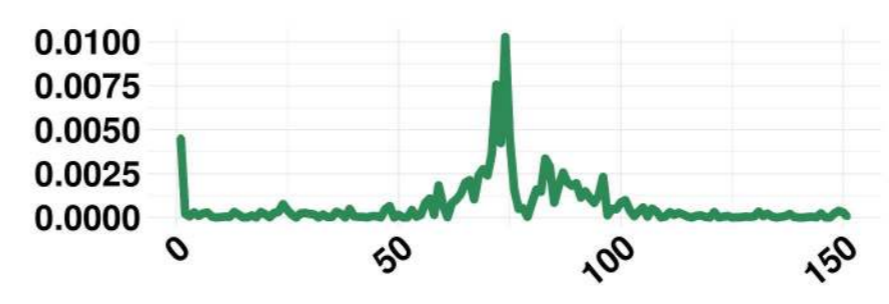
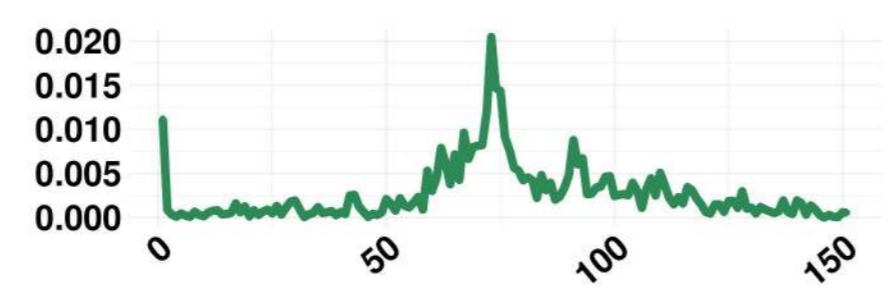
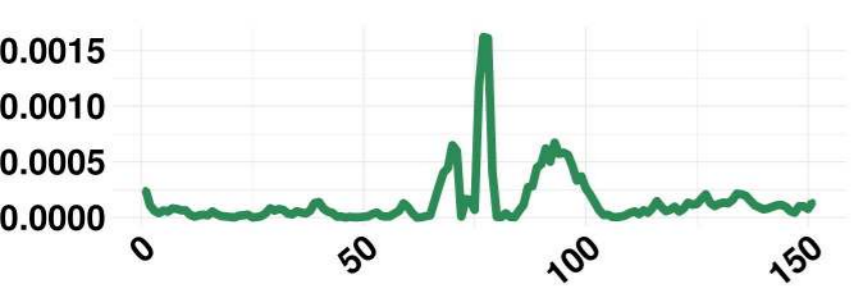
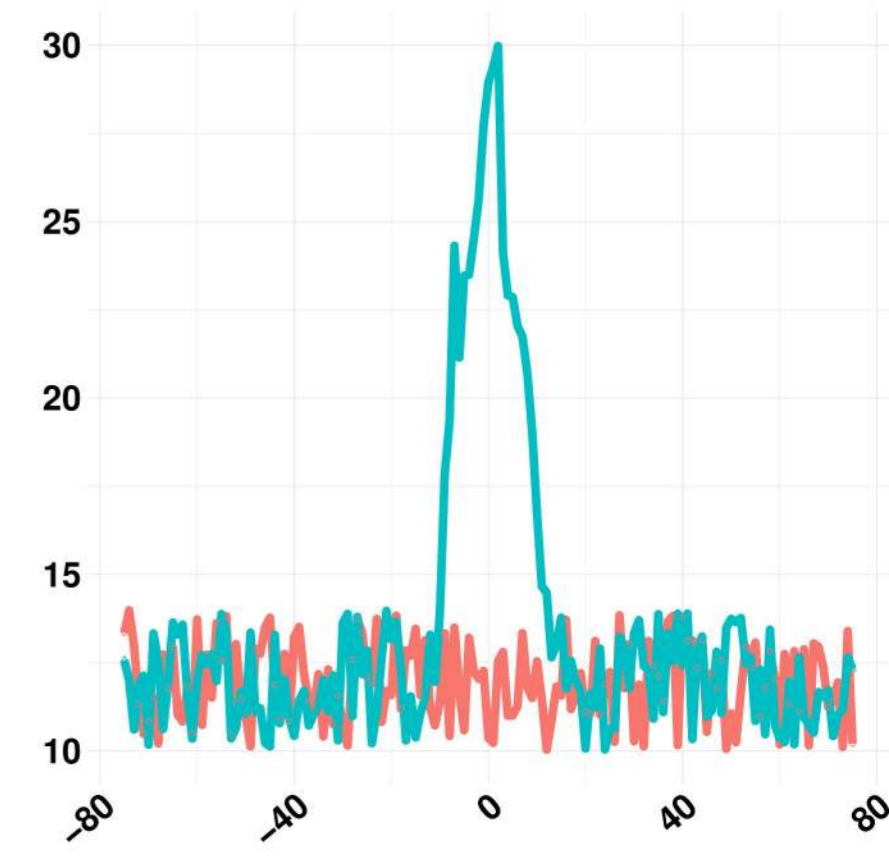
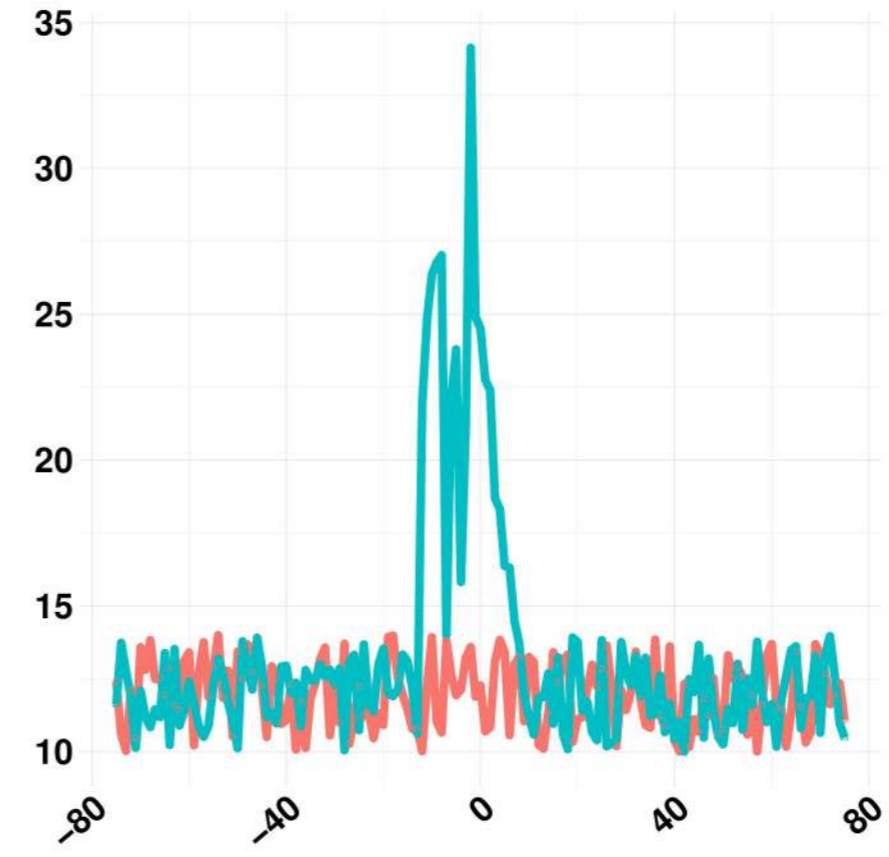
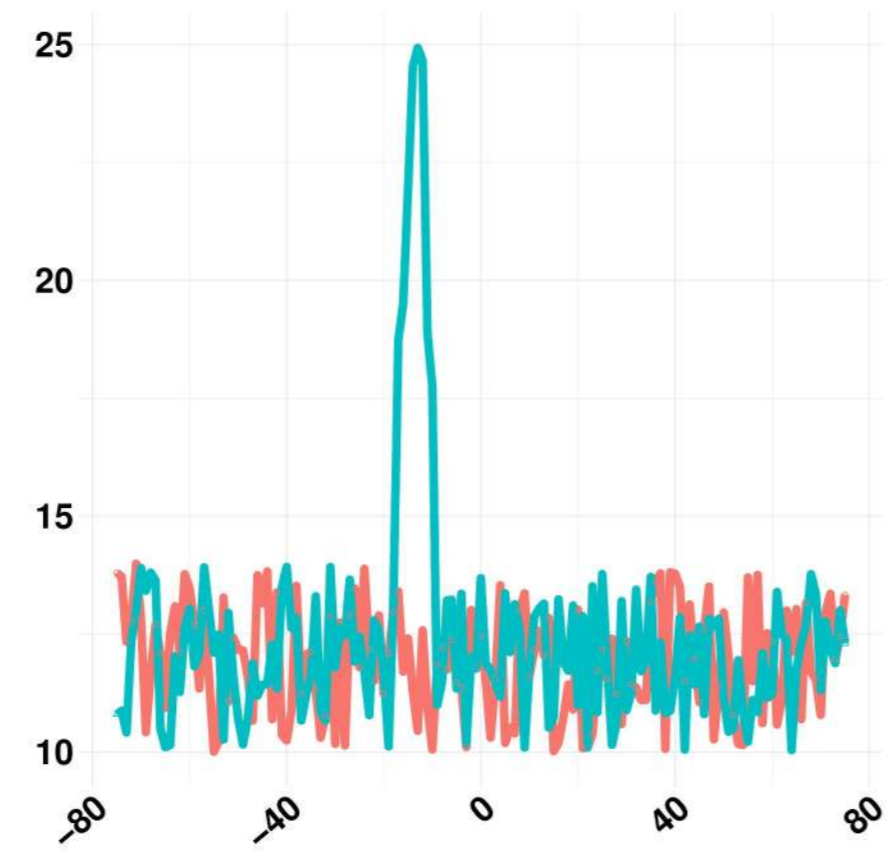
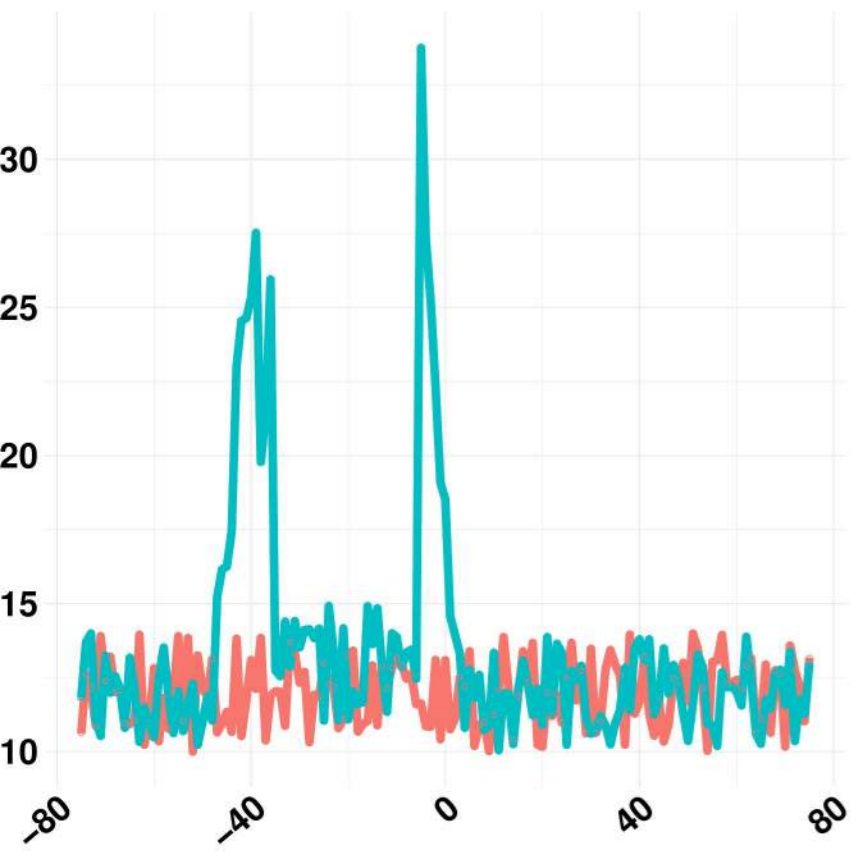


FMR1

IGF2BP2

MSI2

DDX54








































RBM39

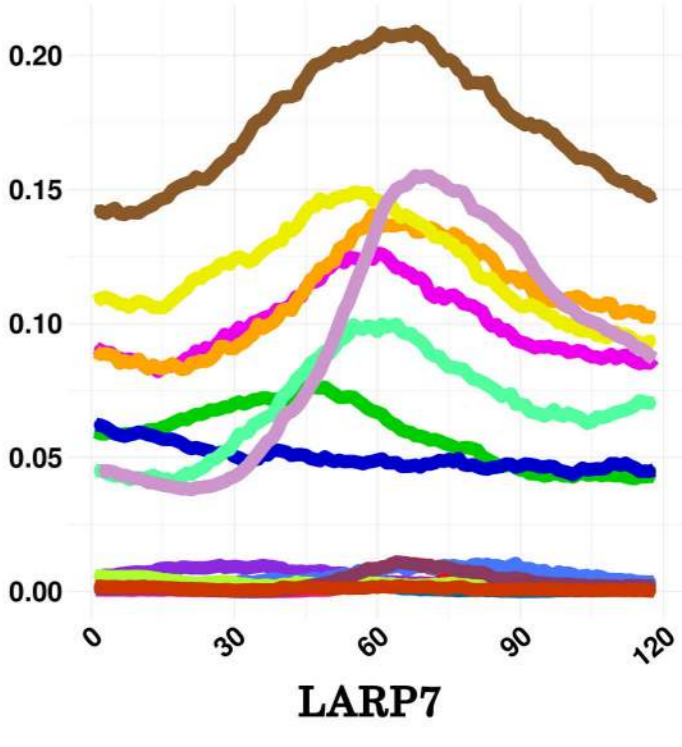
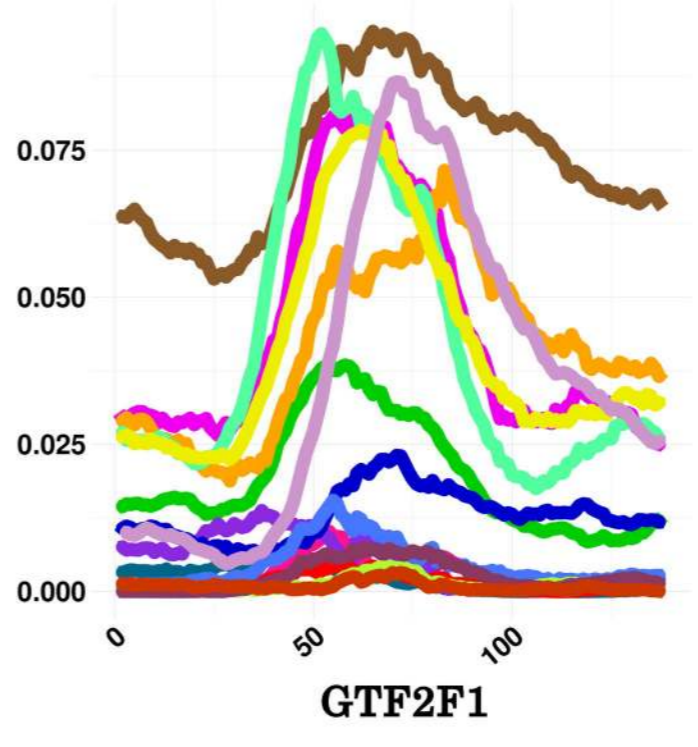
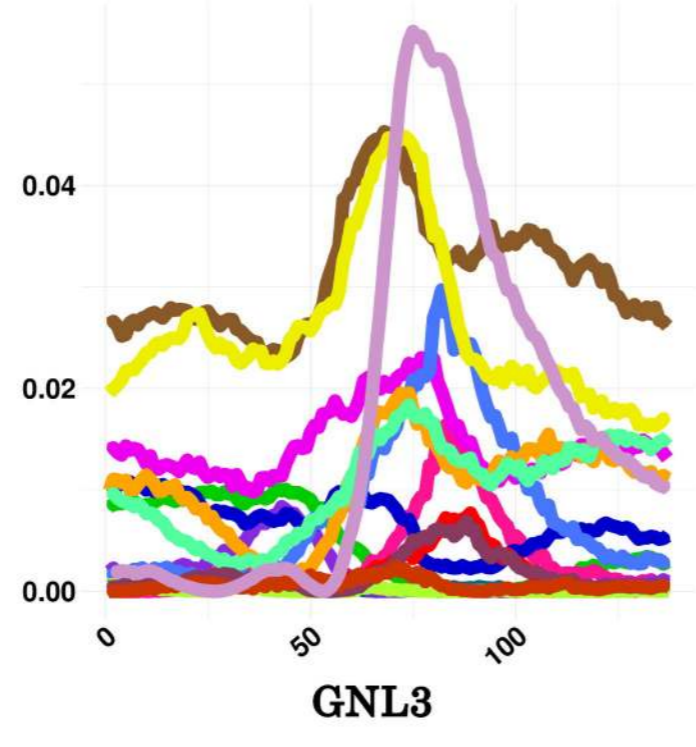
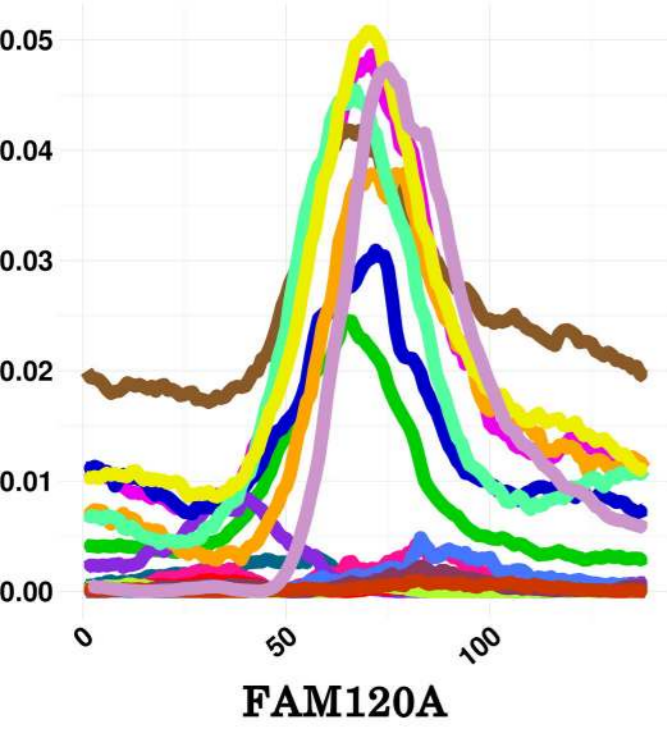
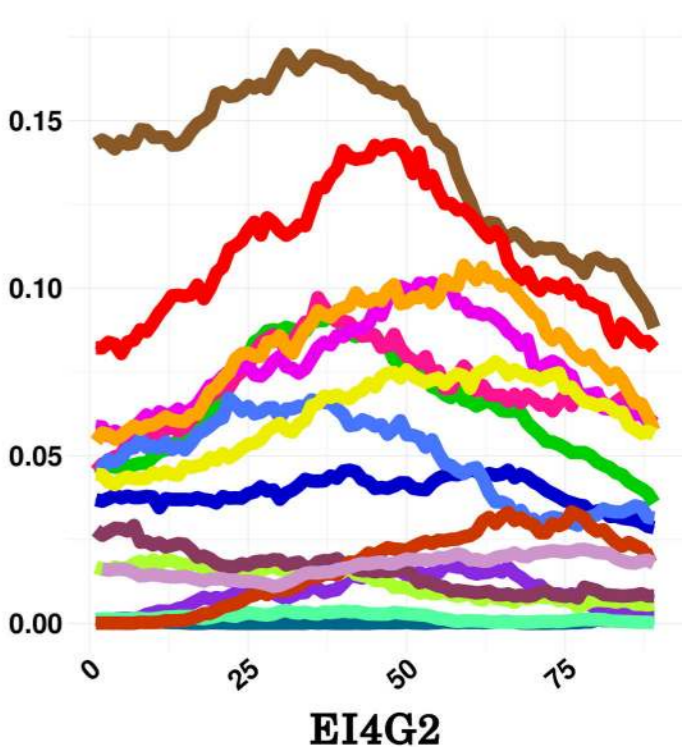
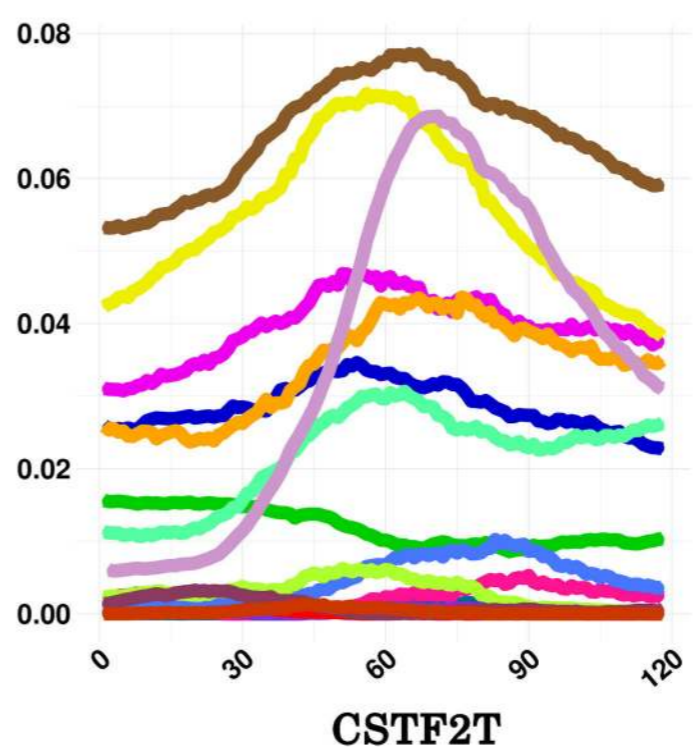
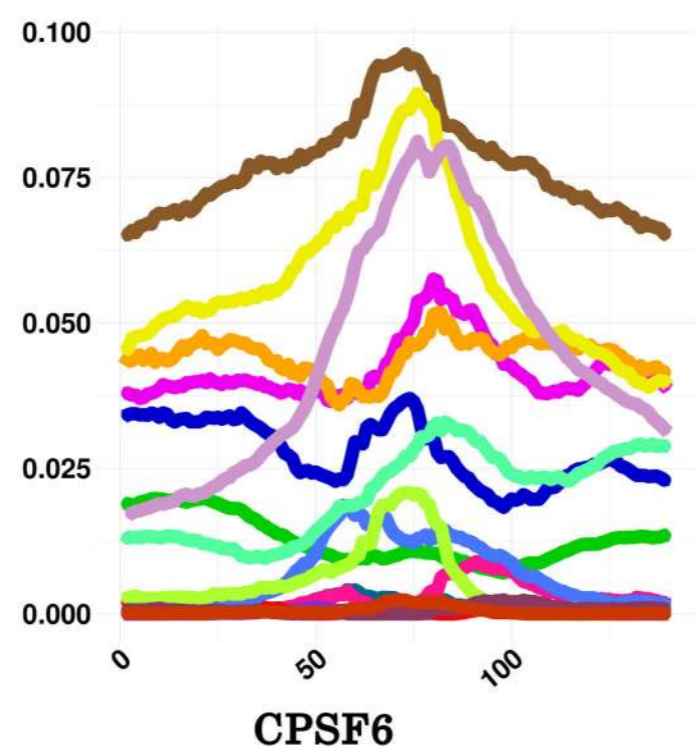
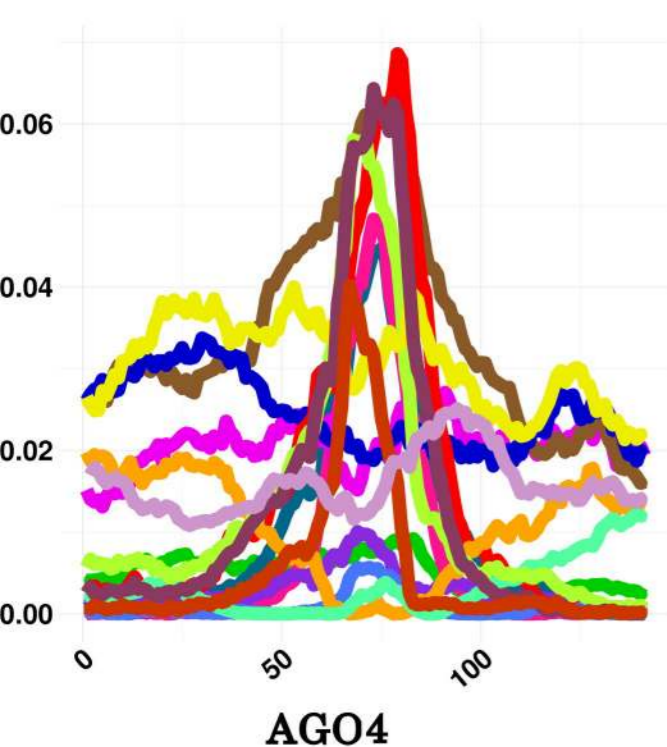
SAFB2

SLTM

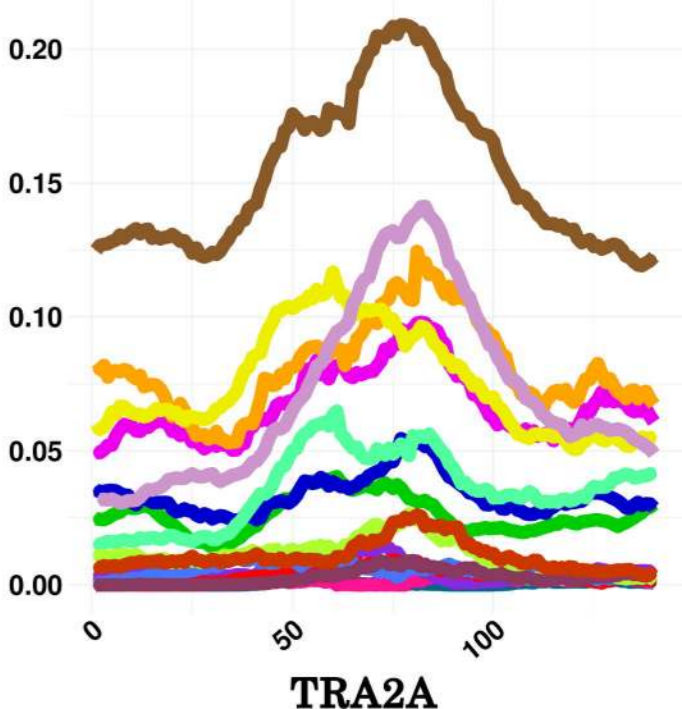
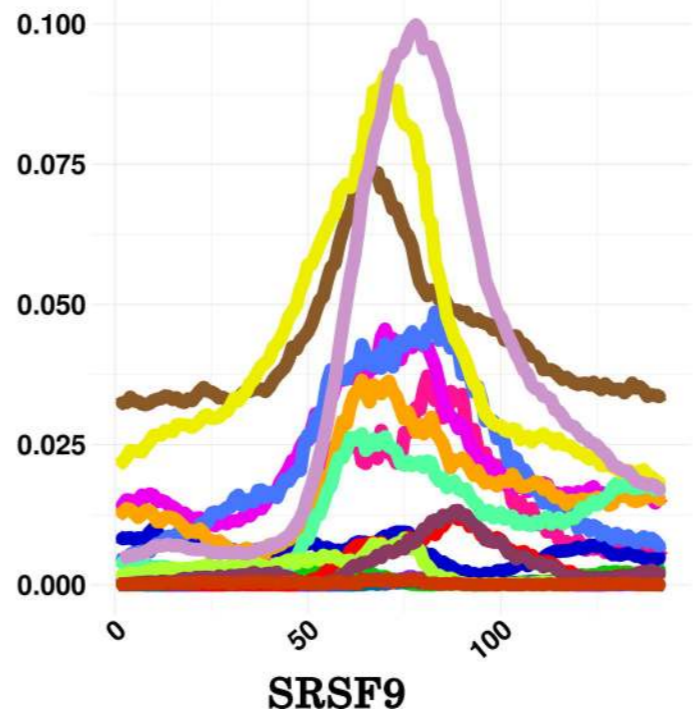
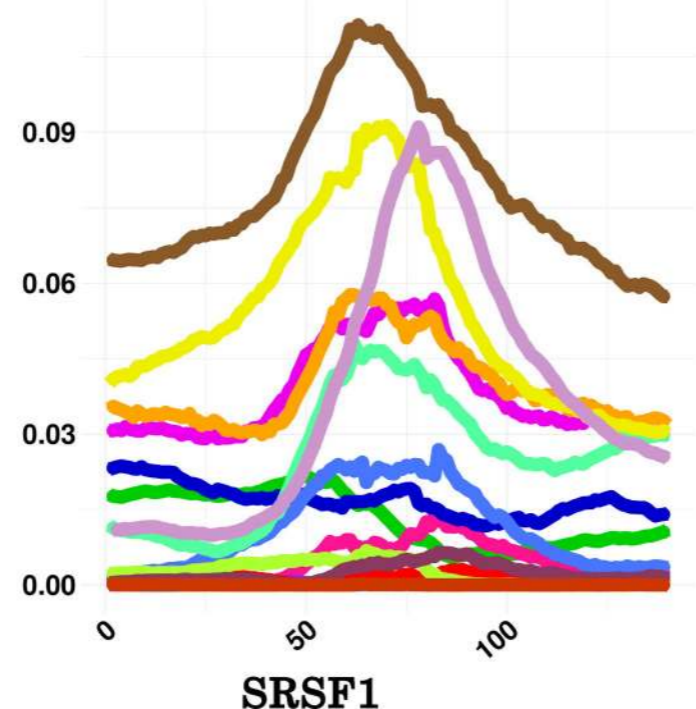
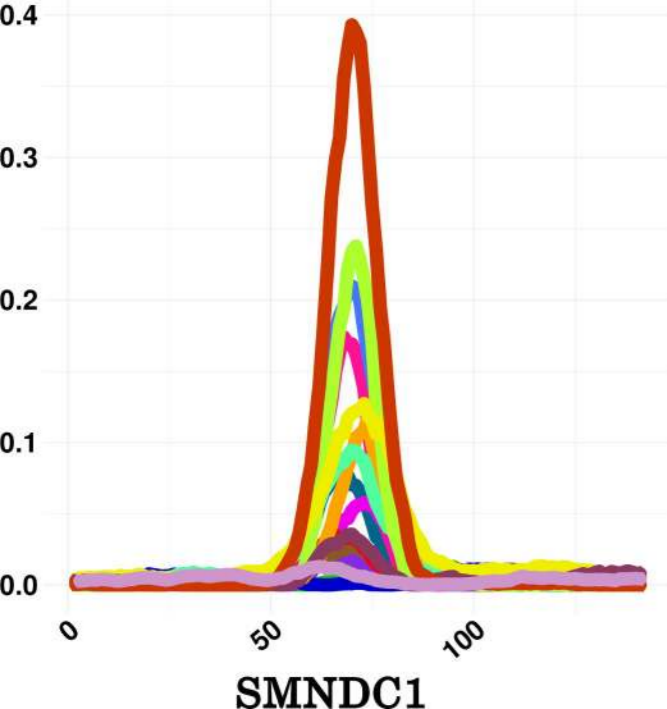
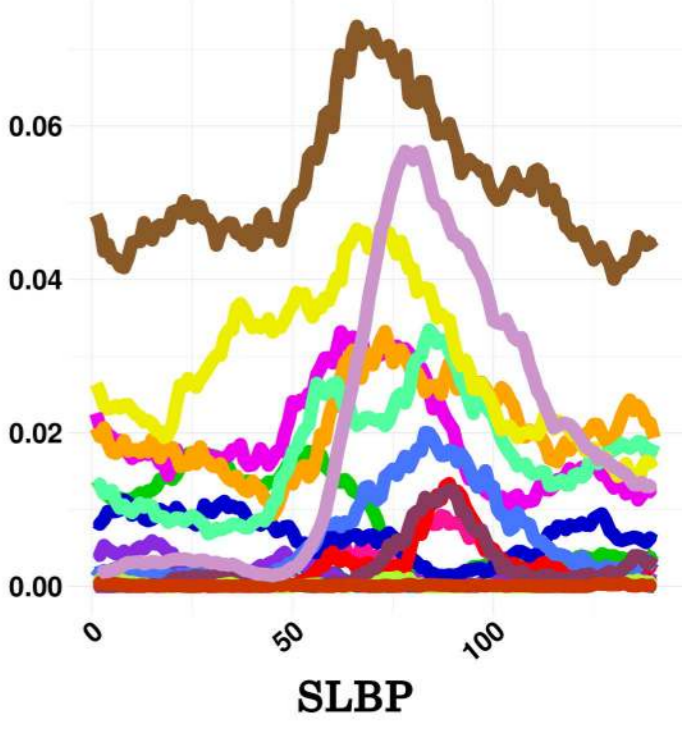
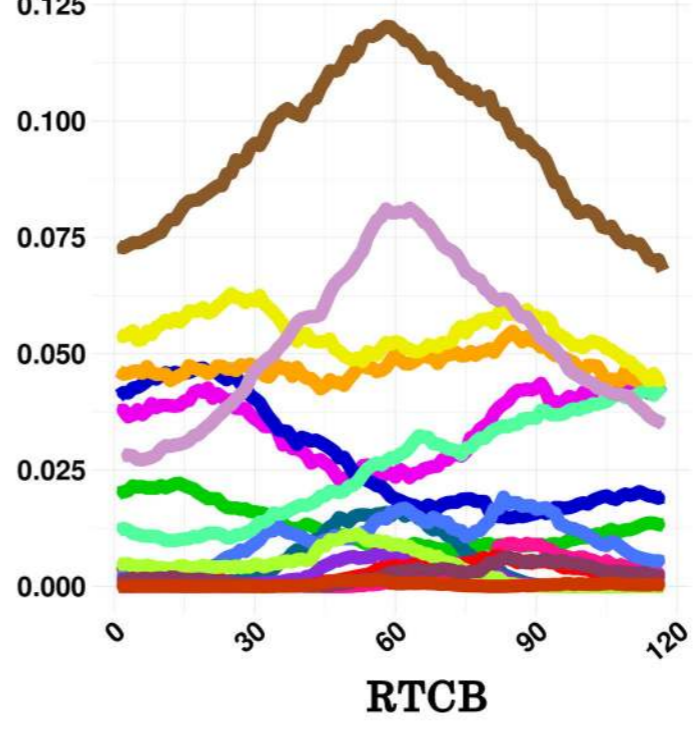
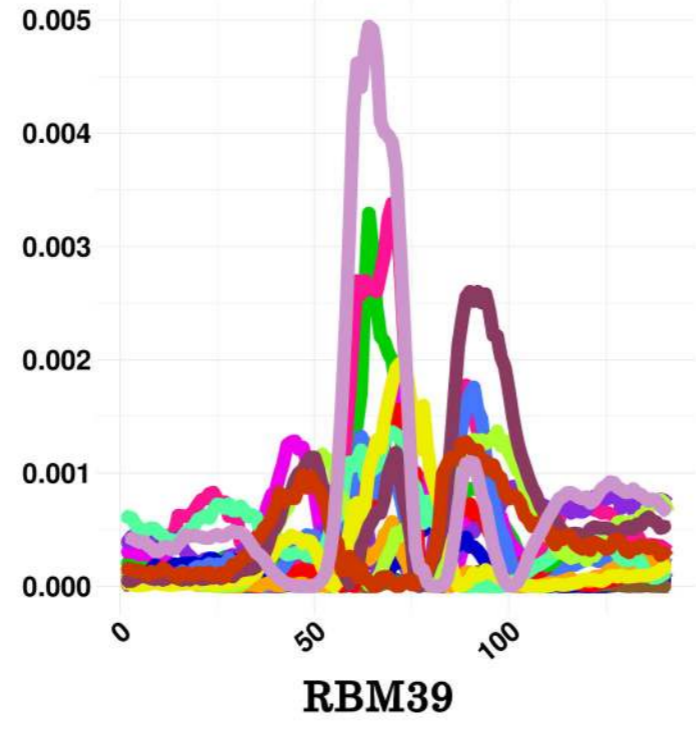
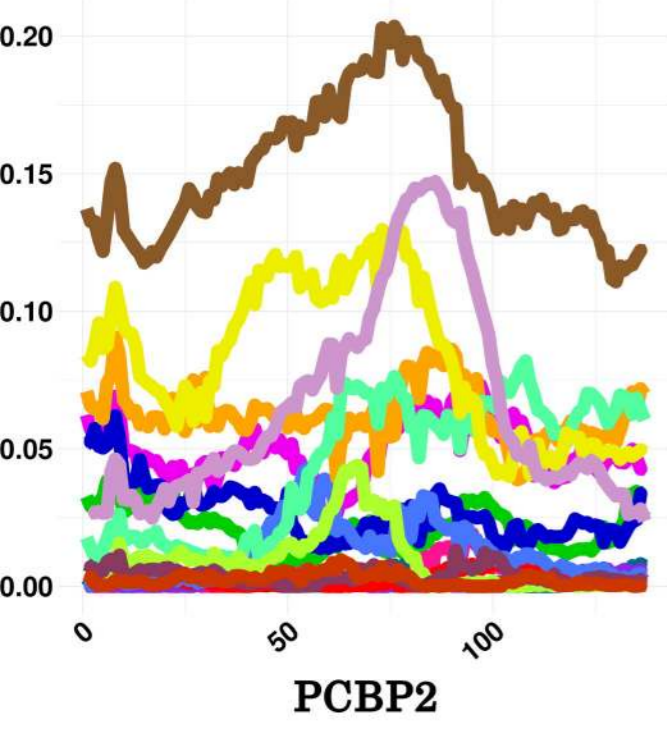
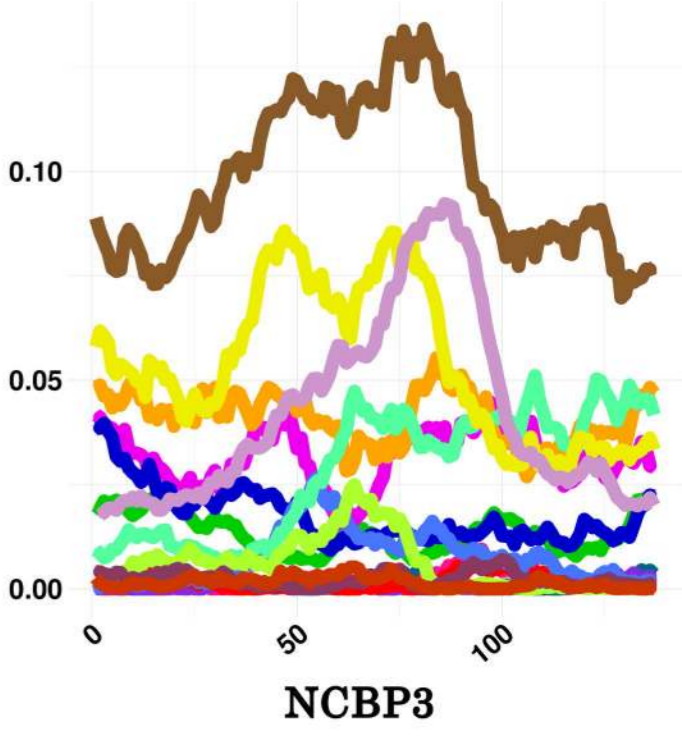
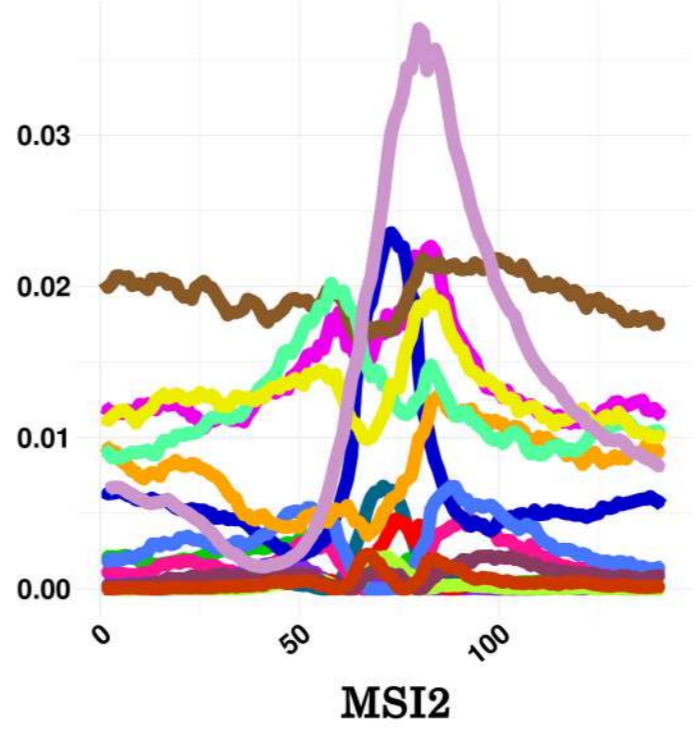
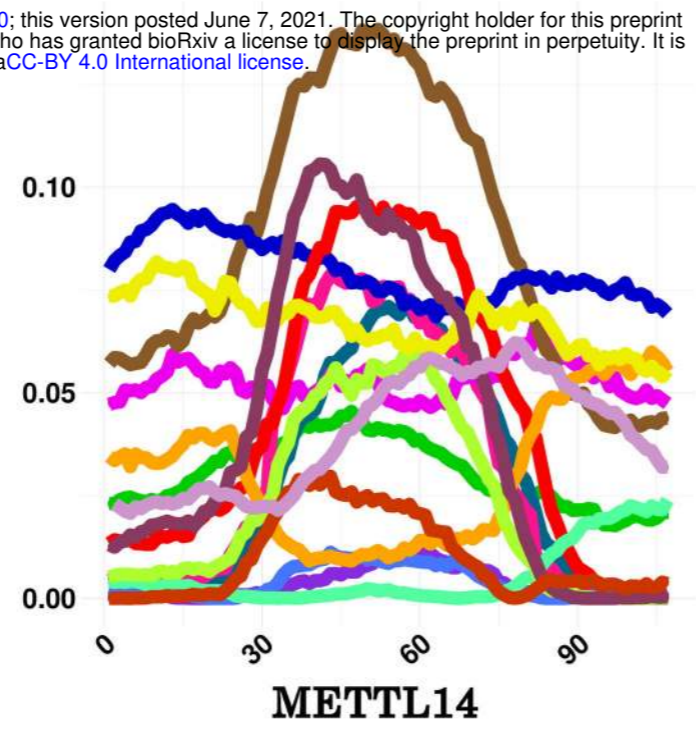
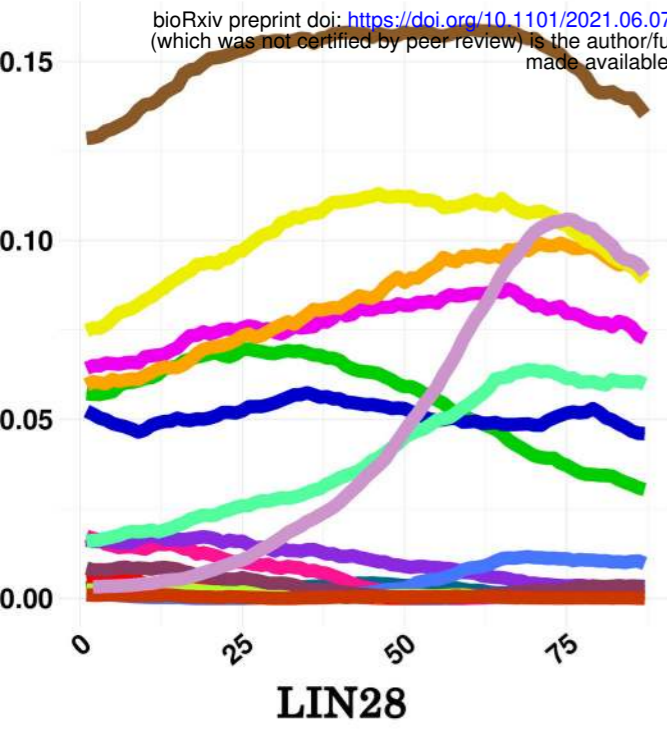
NOP58

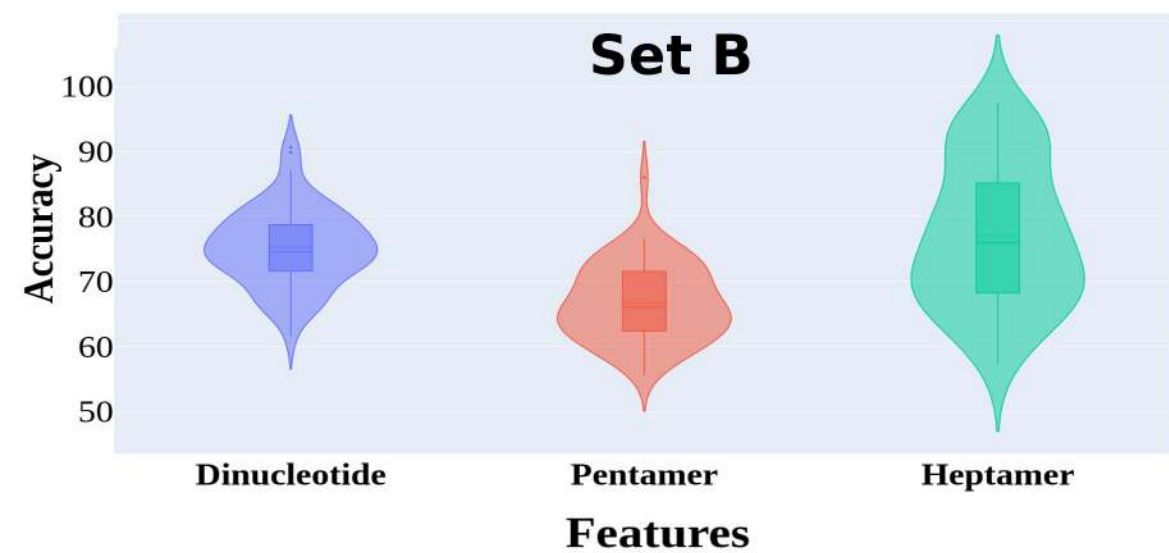
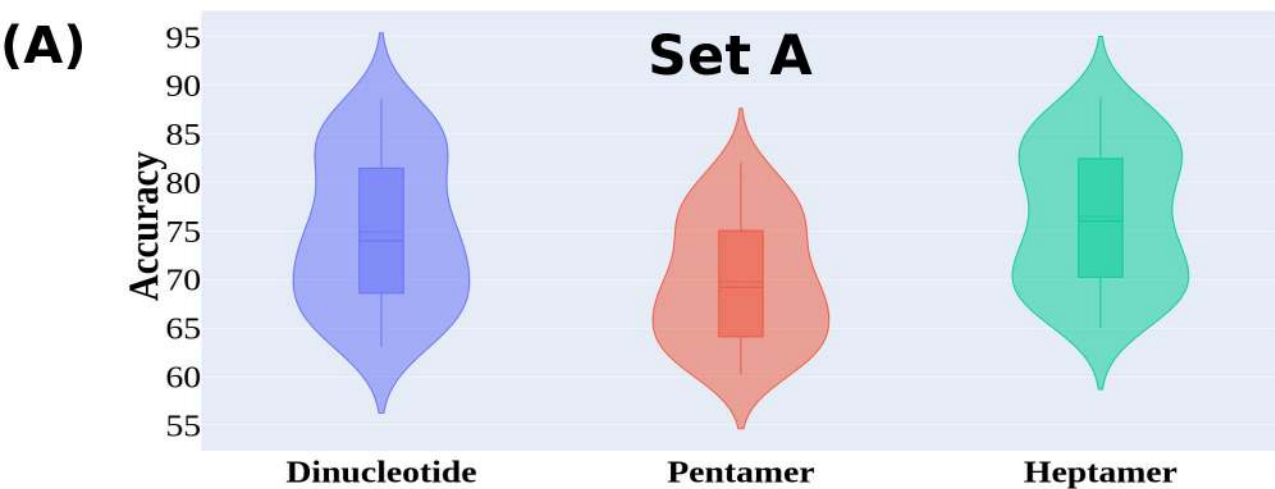
| RBP | RNACompete motif | RBPSpot Prime motif | SELEX/Y3H /RIP-ChiP support | Alignment with prime motif | RBPSpot motif matching to the experimentally reported motif | % of occurrence of the matching motif from RBPSpot in CLIP-Seq instances | Rank of the matching RBPSpot motif |
|---------|---|---|-----------------------------|---------------------------------|---|--|------------------------------------|
| HNRNPK |  |  | | CCC CCC |  | 47.66% | 11 th |
| QKI |  |  | ACUAACA | ACU ACU |  | 62.74% | 5 th |
| KHDRBS3 |  |  | | AUAAA · UAAAA |  | 70.48% | 1 st |
| TIA1 |  |  | | UUUUG · UUCUG |  | 59.68% | 9 th |
| LIN28A |  |  | | CGGA · CUGA |  | 34.07% | 9 th |
| HNRNPC |  |  | UUUUU | UUUUUU · UUUUGU |  | 66.19% | 1 st |
| SRSF1 |  |  | AGGACA | GGACA · GGAGA |  | 62.97% | 7 th |
| KHDRBS1 |  |  | | AUAAA · AUGAAA |  | 71.85% | 1 st |
| HNRNPA1 |  |  | | GGGA · GAGA |  | 55.97% | 7 th |
| HNRNPL |  |  | ACACACA | CACAC CACAC |  | 63.45% | 1 st |
| KHDRBS2 |  |  | AAUAAAA | UAAAA · UAAGA |  | 68.18% | 5 th |
| TIAL1 | |  | UUUUUU | UUUUUU · UUUUAU |  | 81.32% | 1 st |
| PUM2 | |  | UGUAUAUA | UGUAUAUA · UUUAUAUA |  | 78.39% | 1 st |

AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT



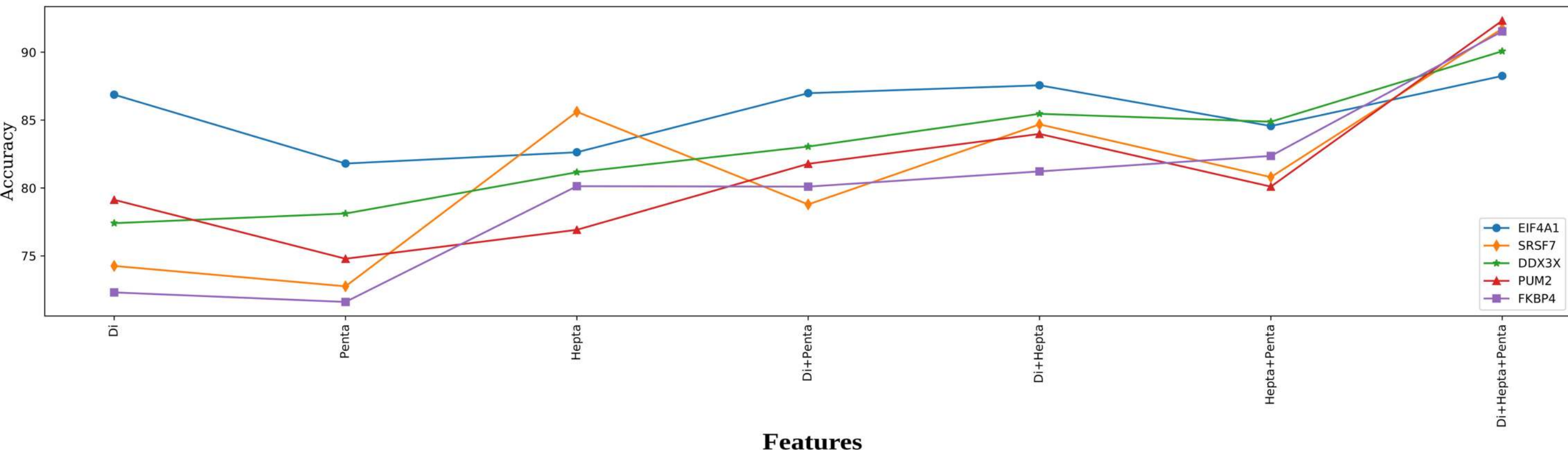
bioRxiv preprint doi: <https://doi.org/10.1101/2021.06.07.447370>; this version posted June 7, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



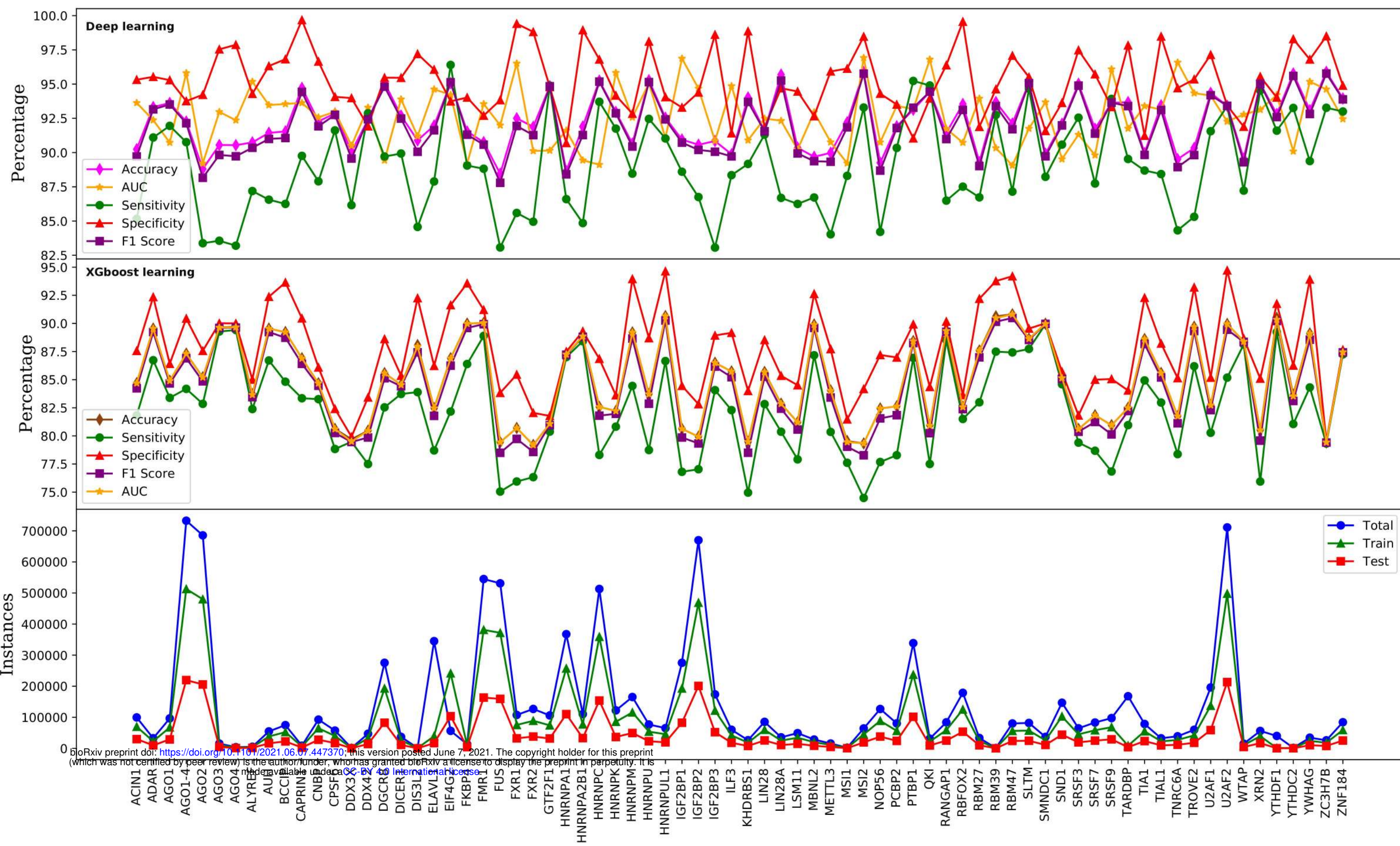


(B)

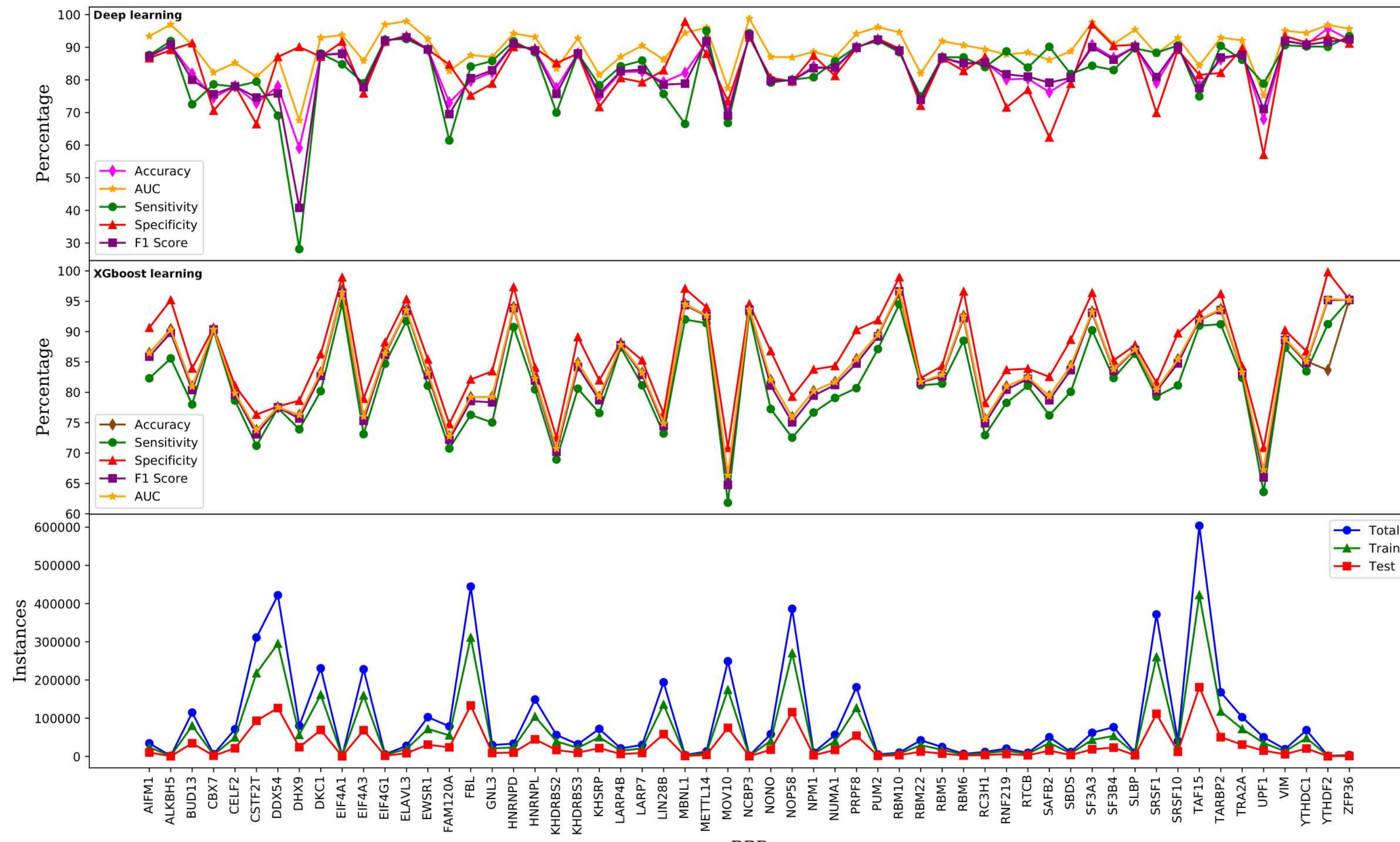
Features

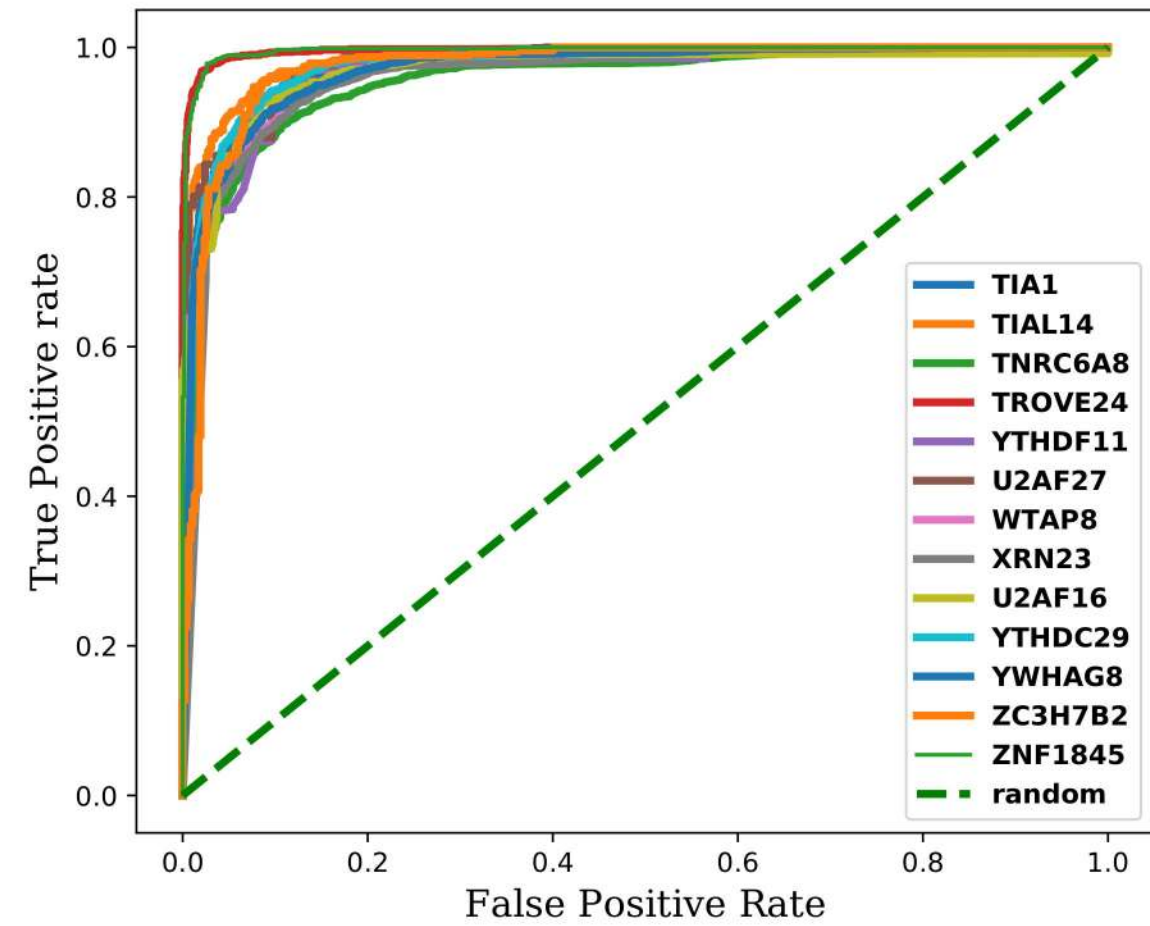
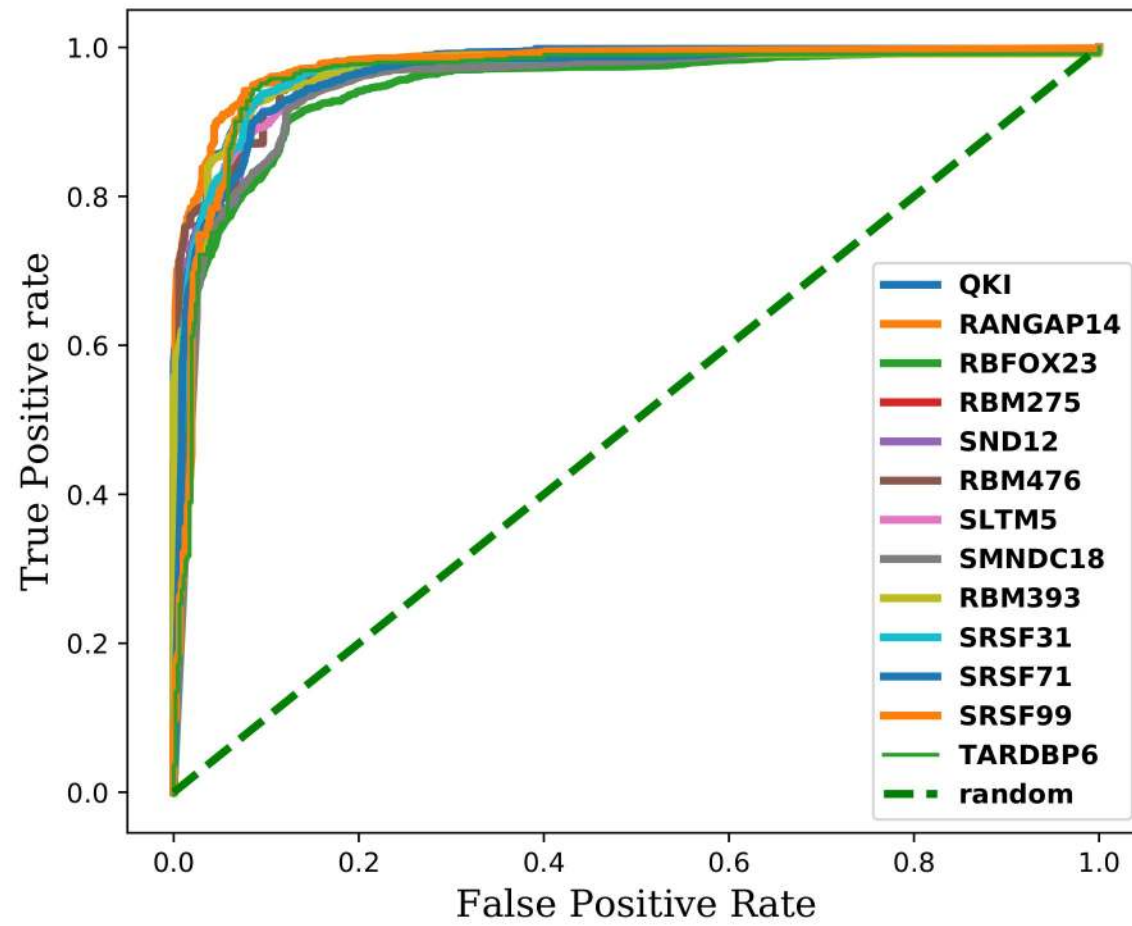
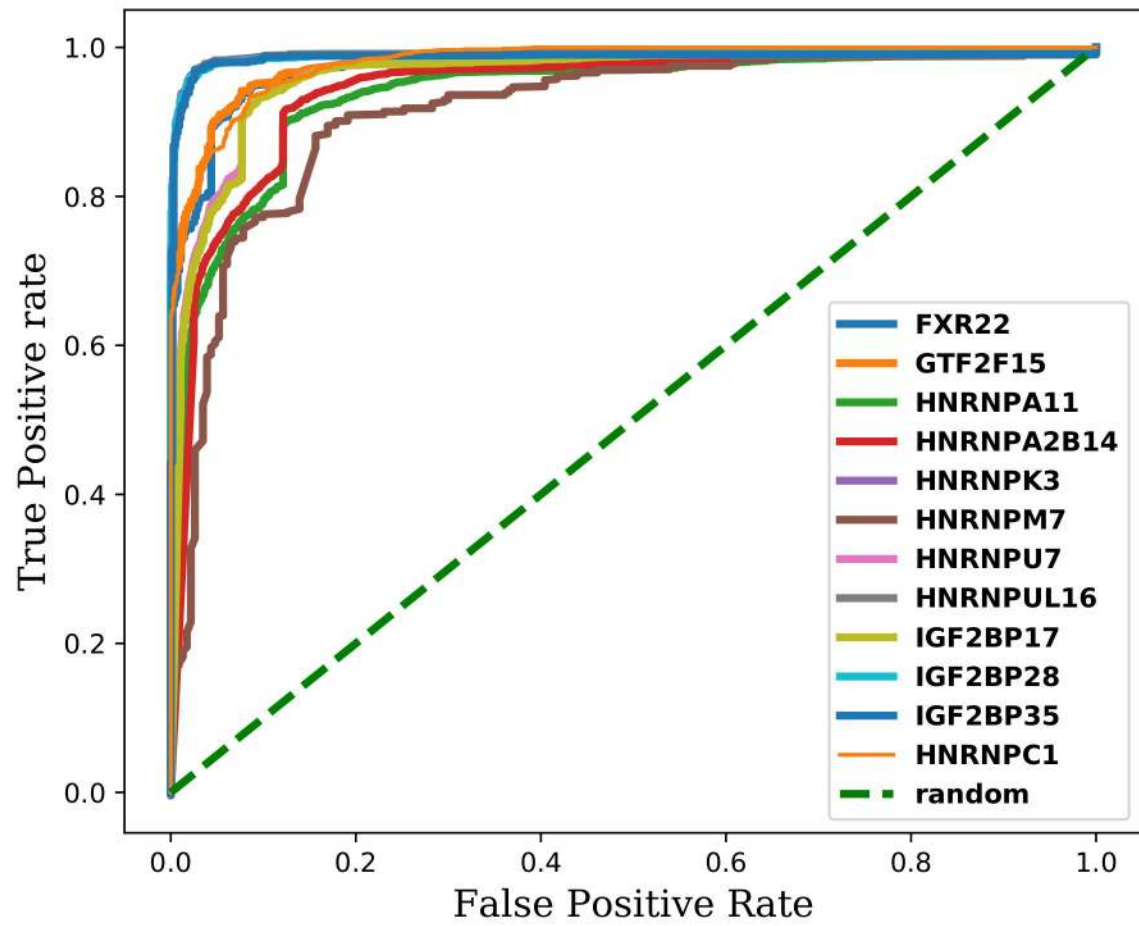
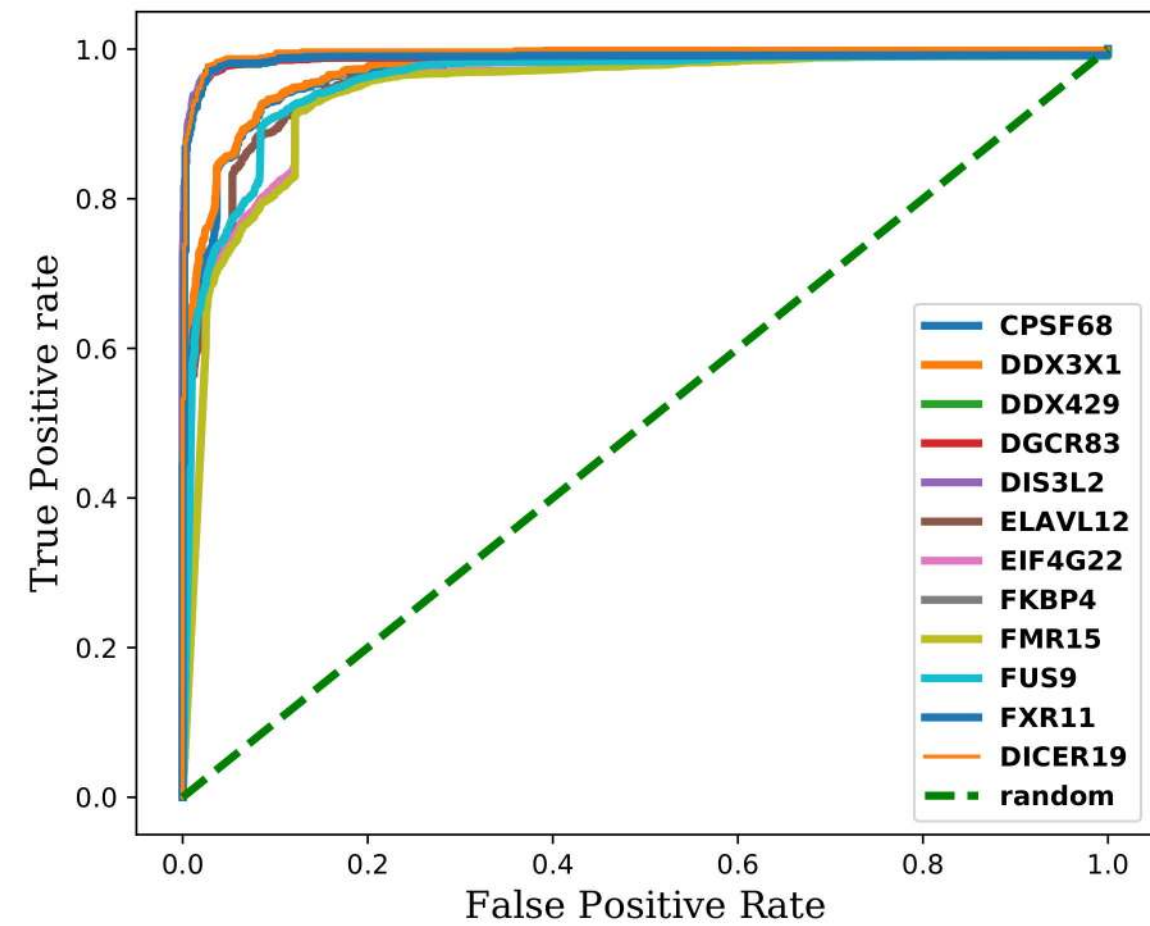
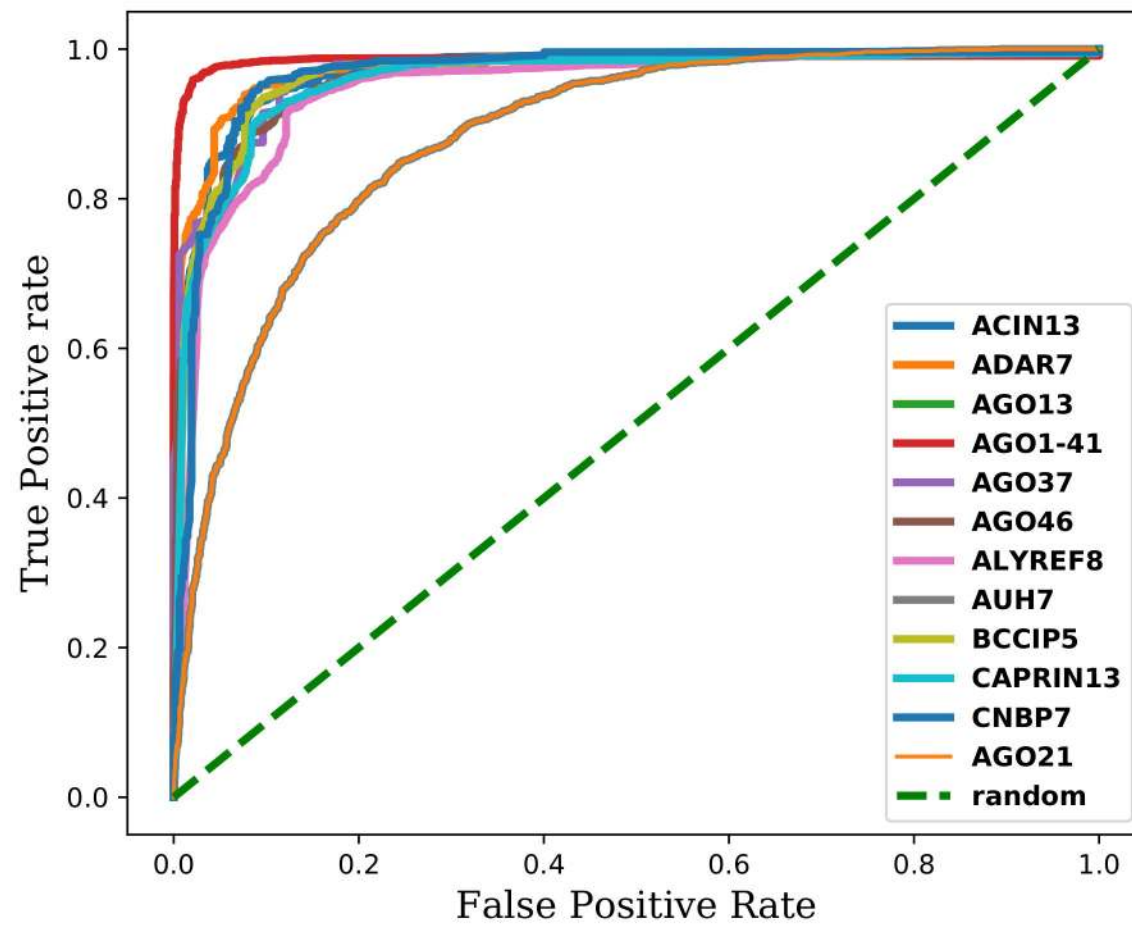
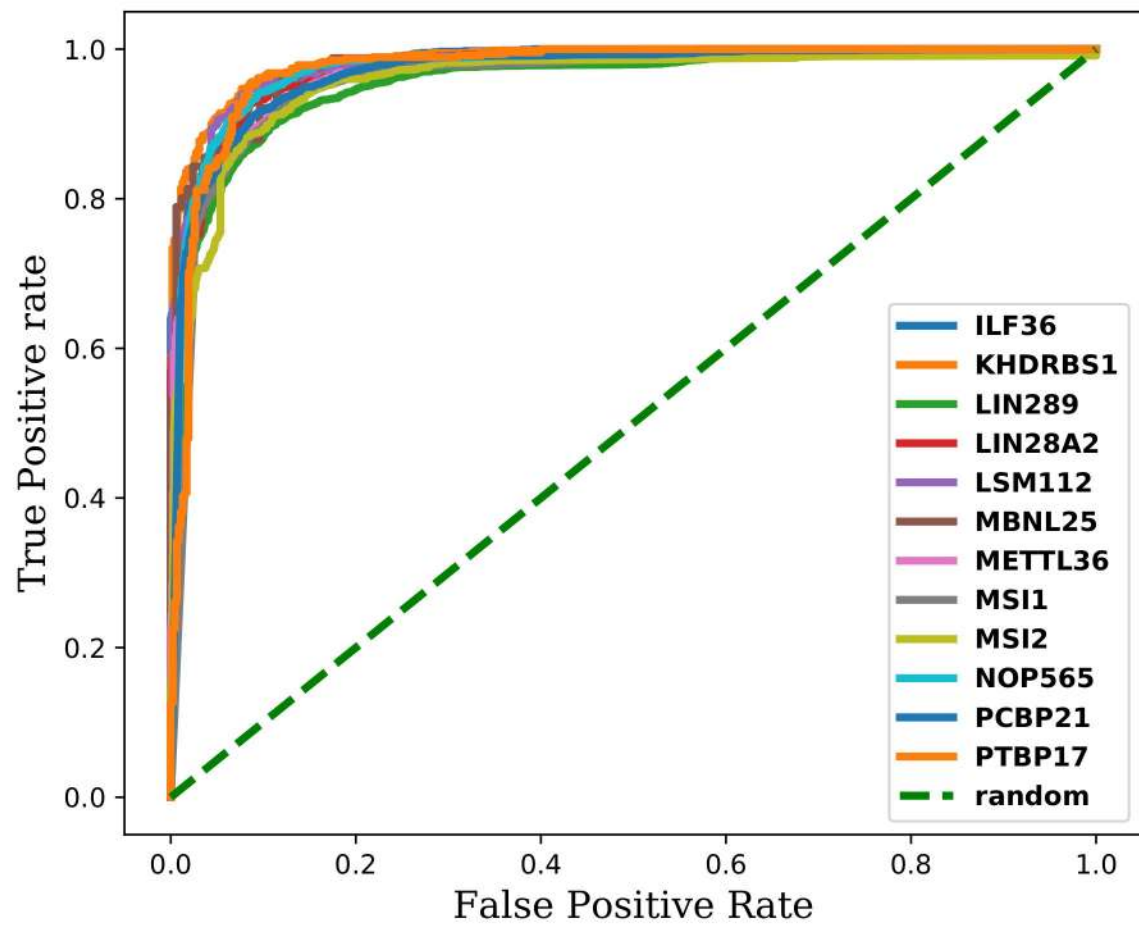


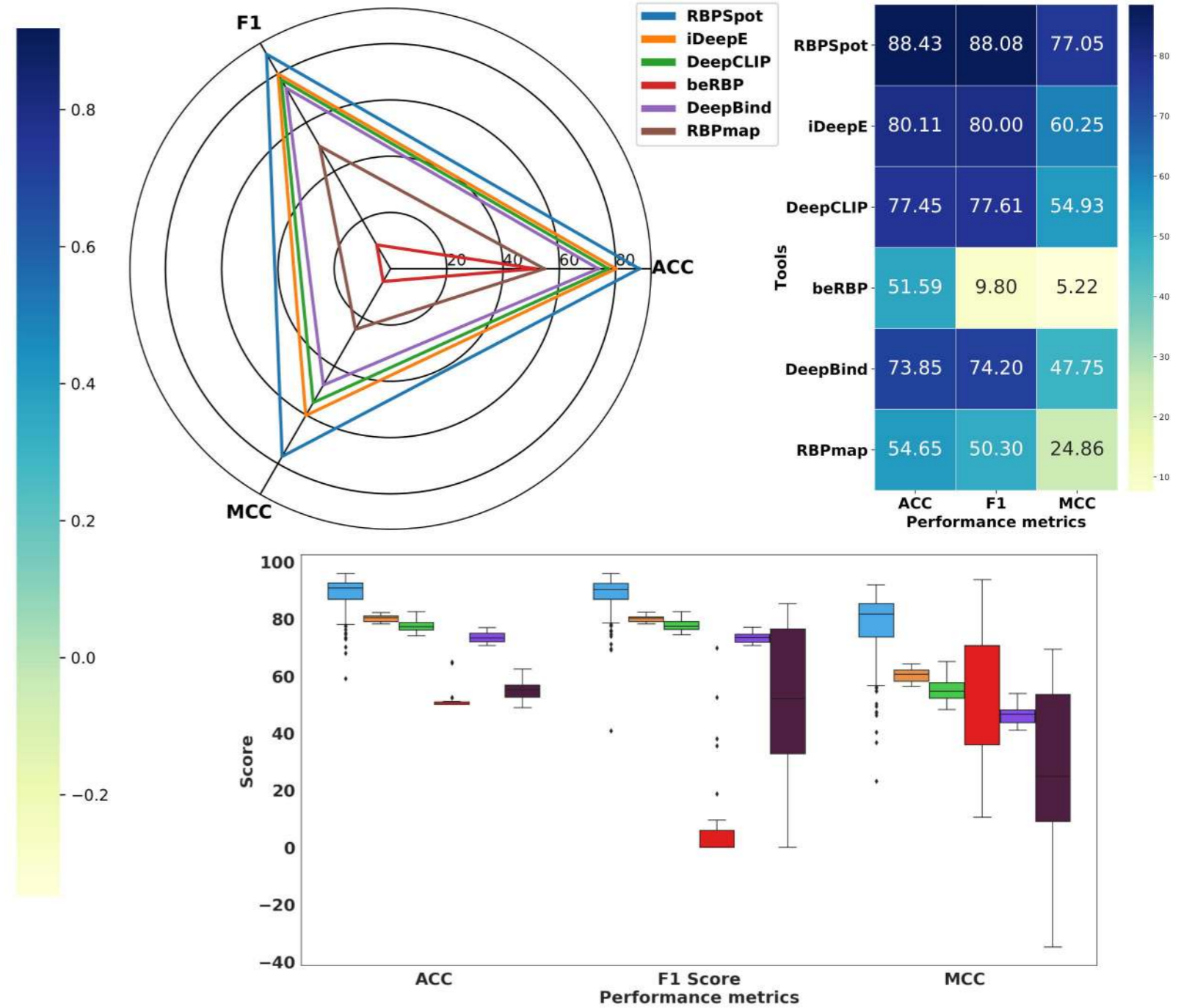
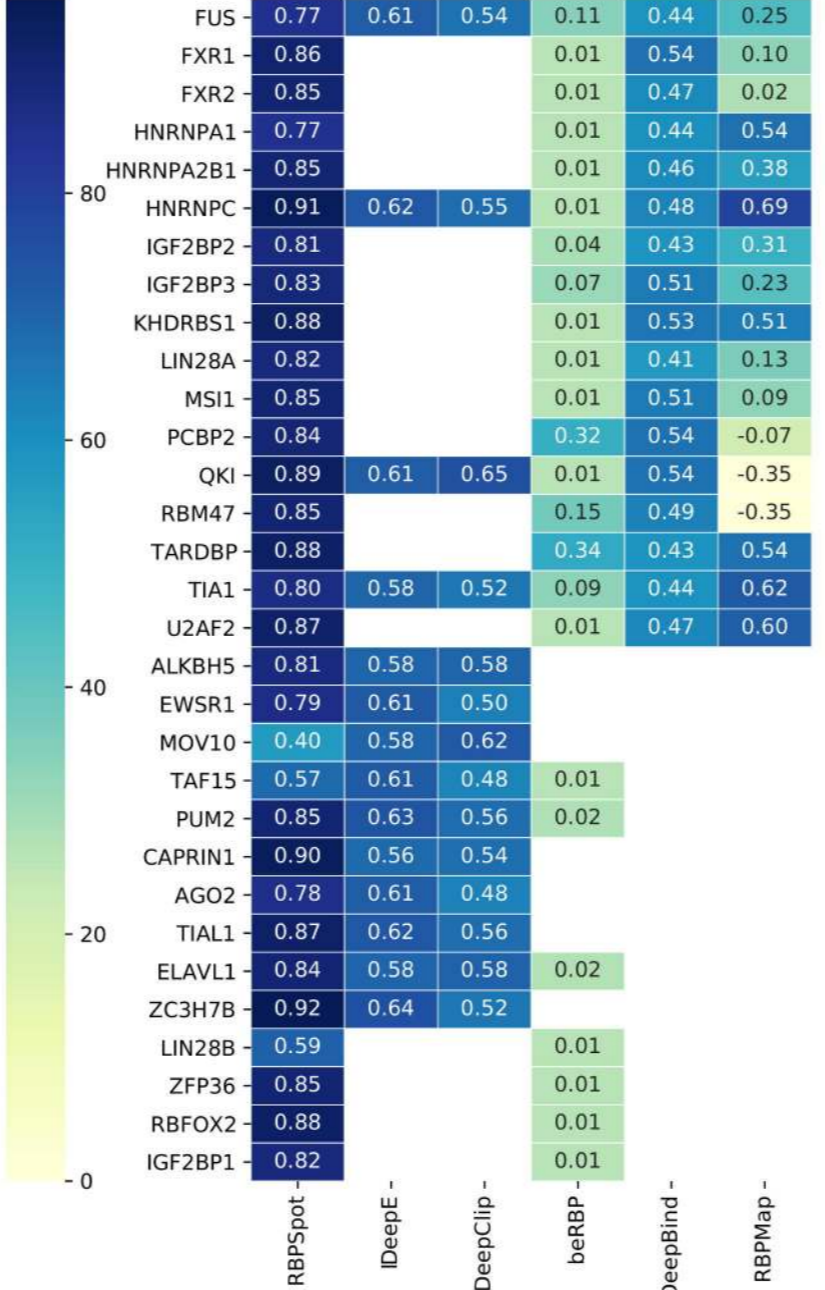
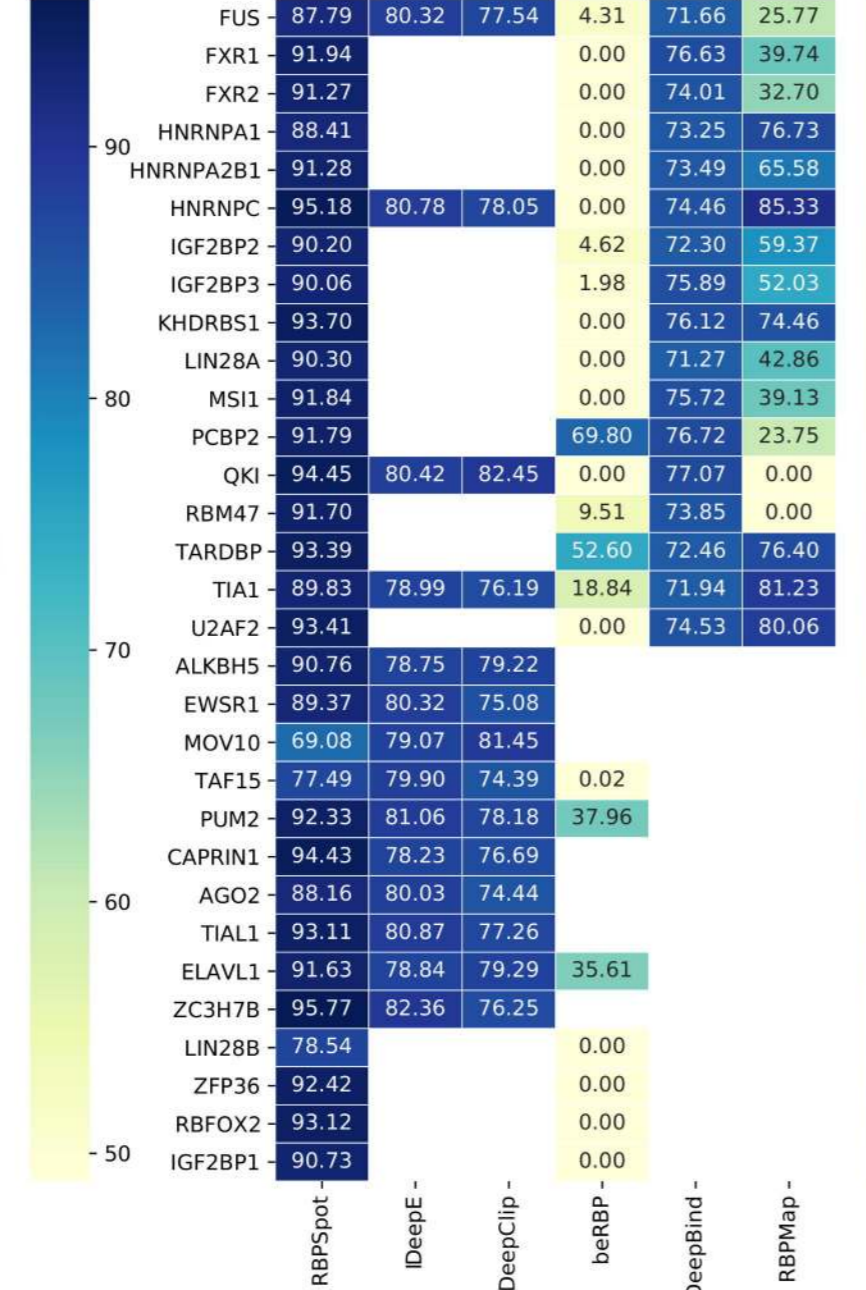
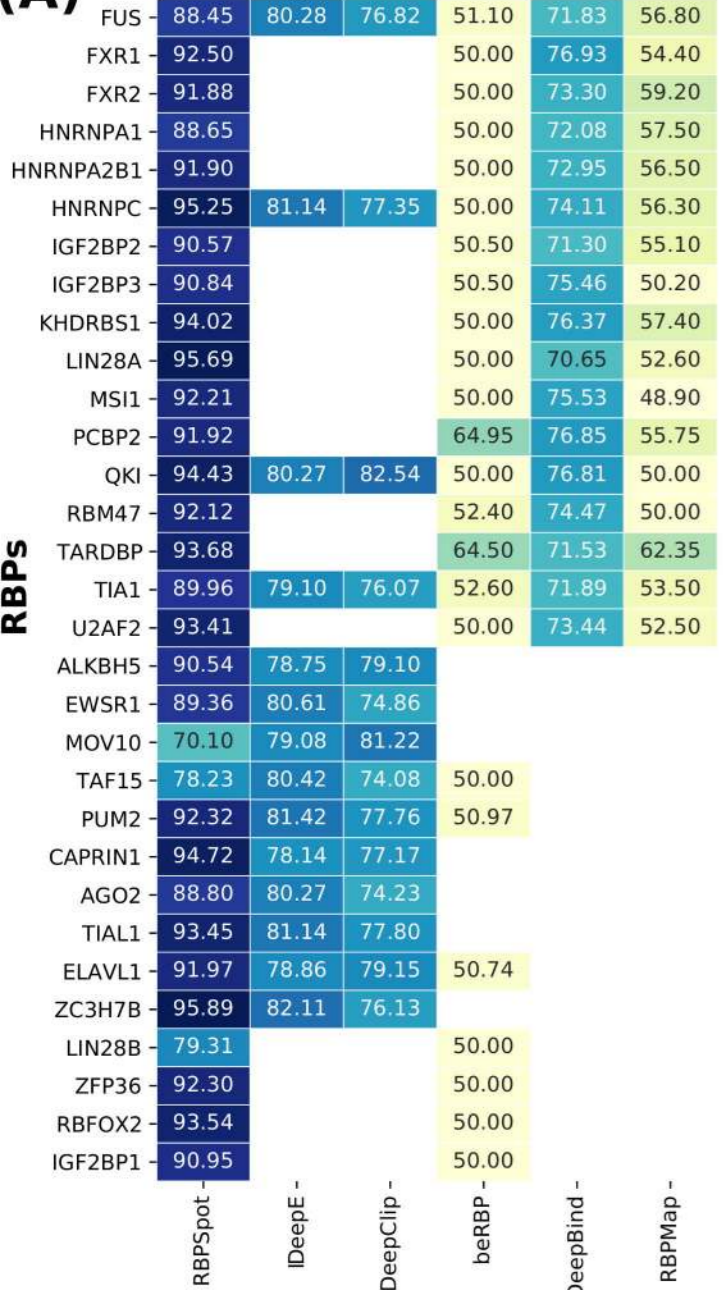
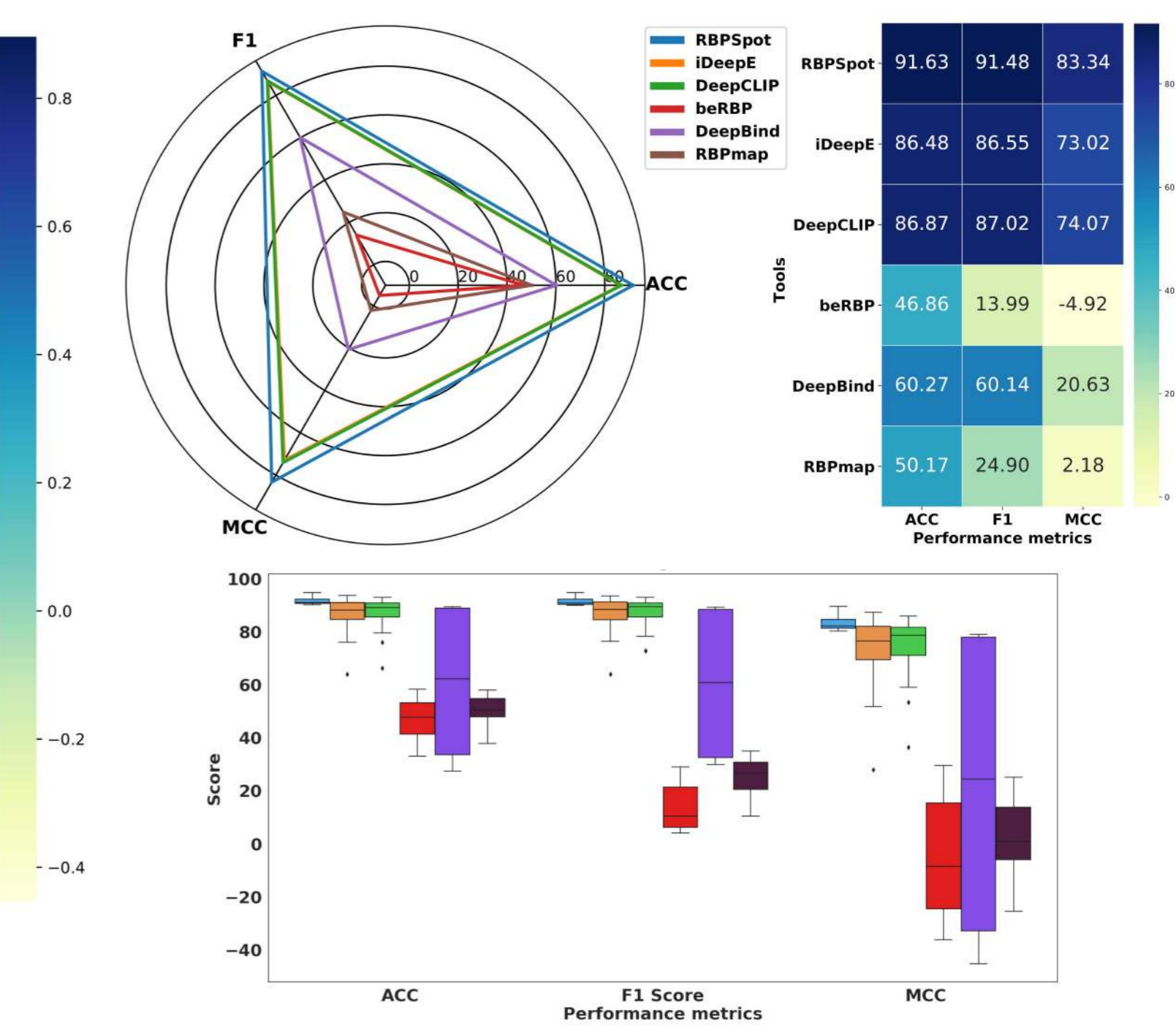
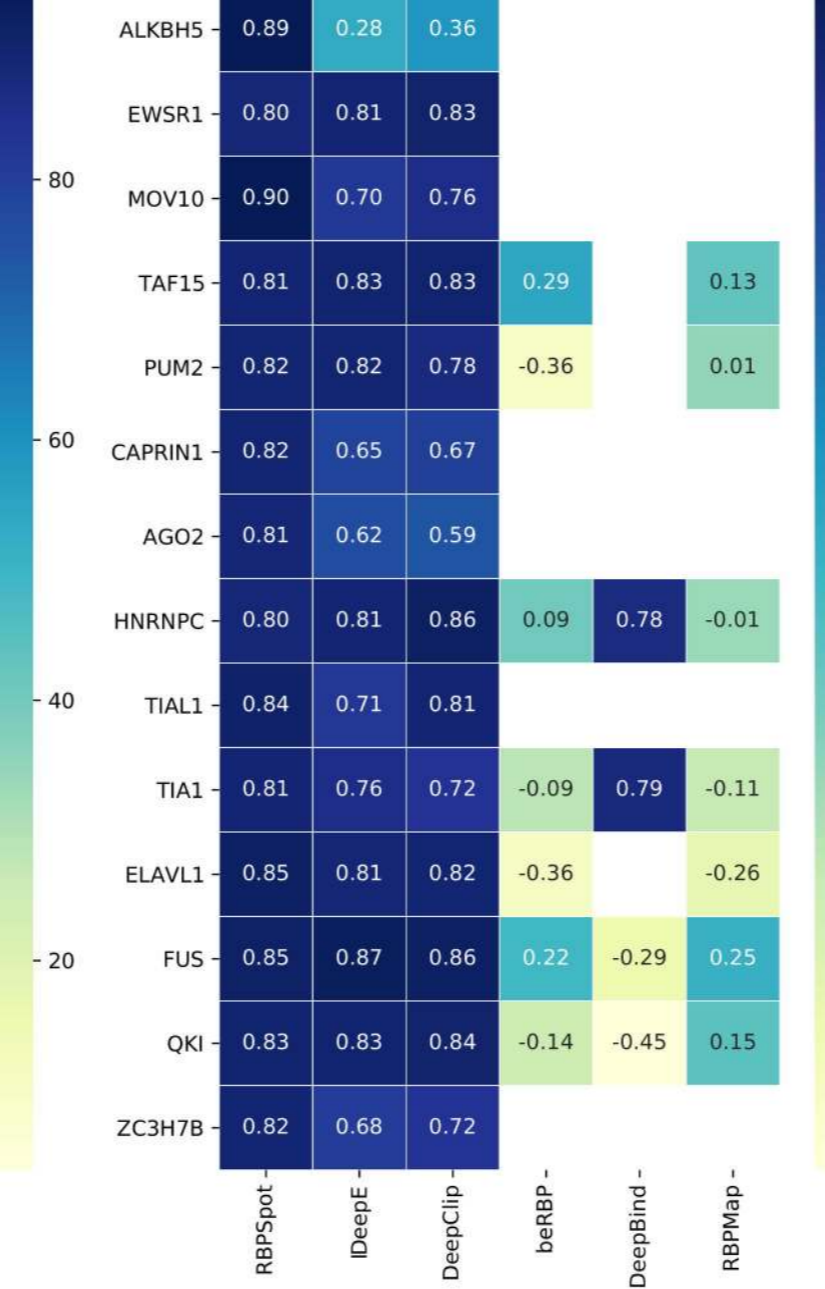
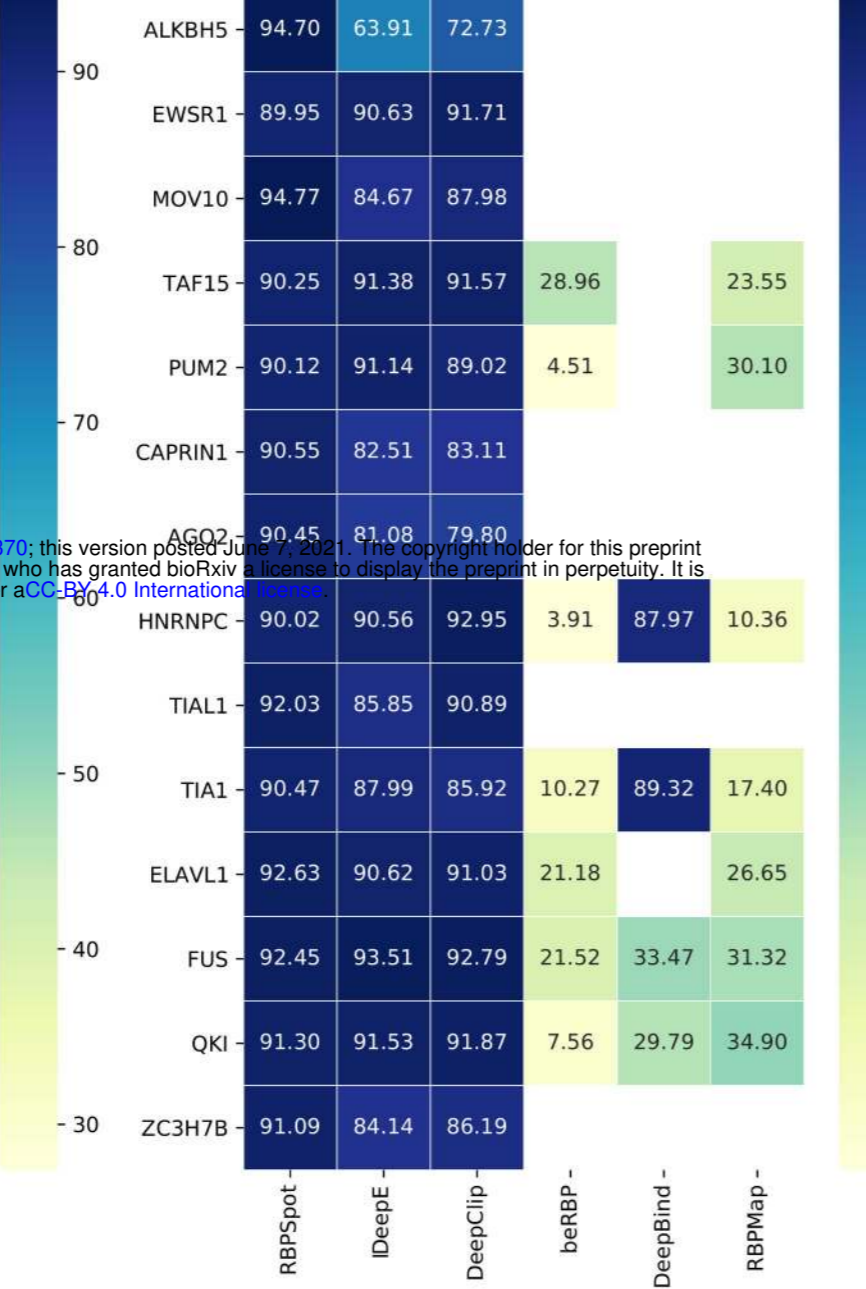
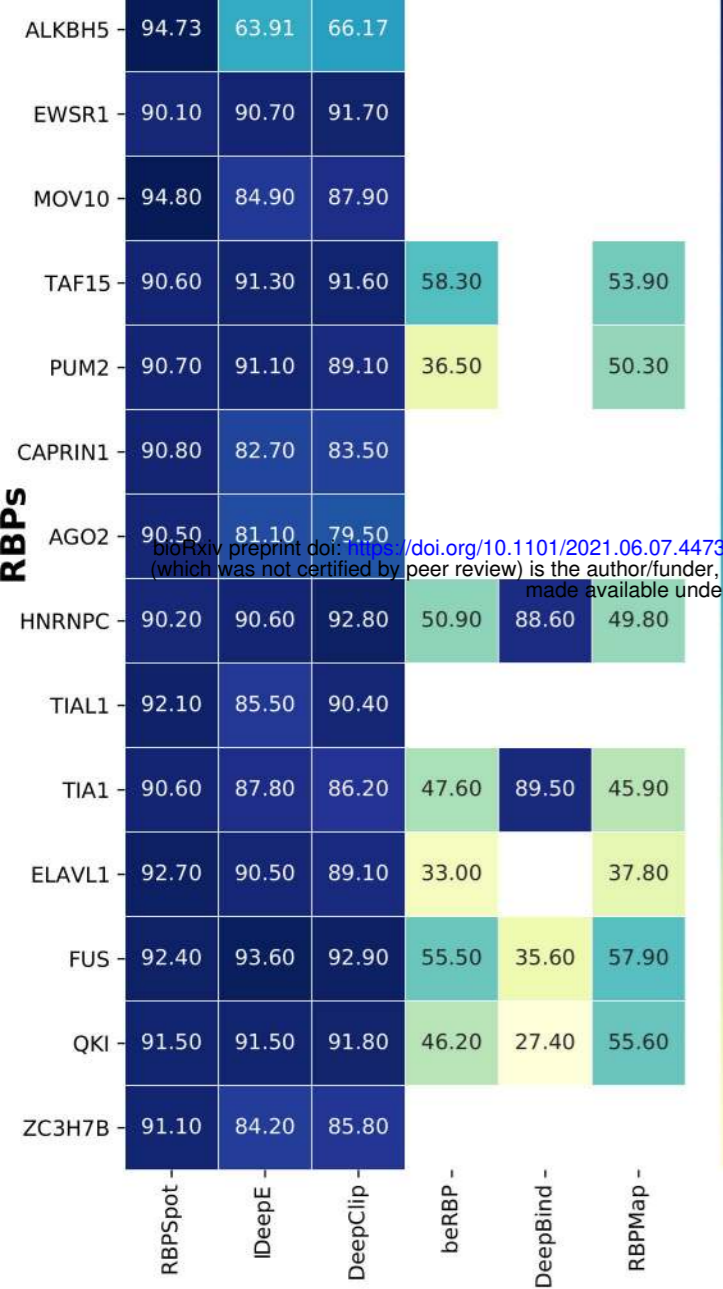
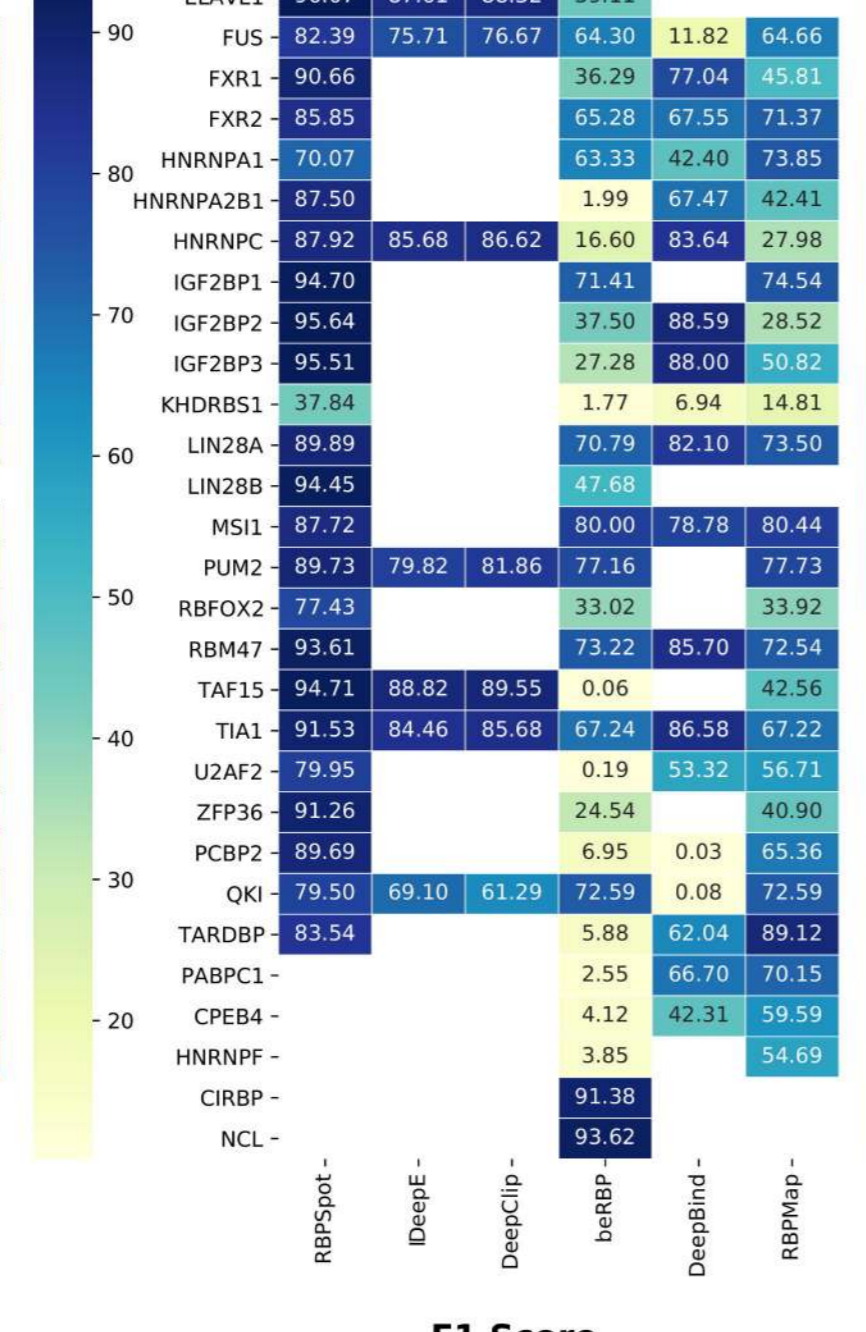
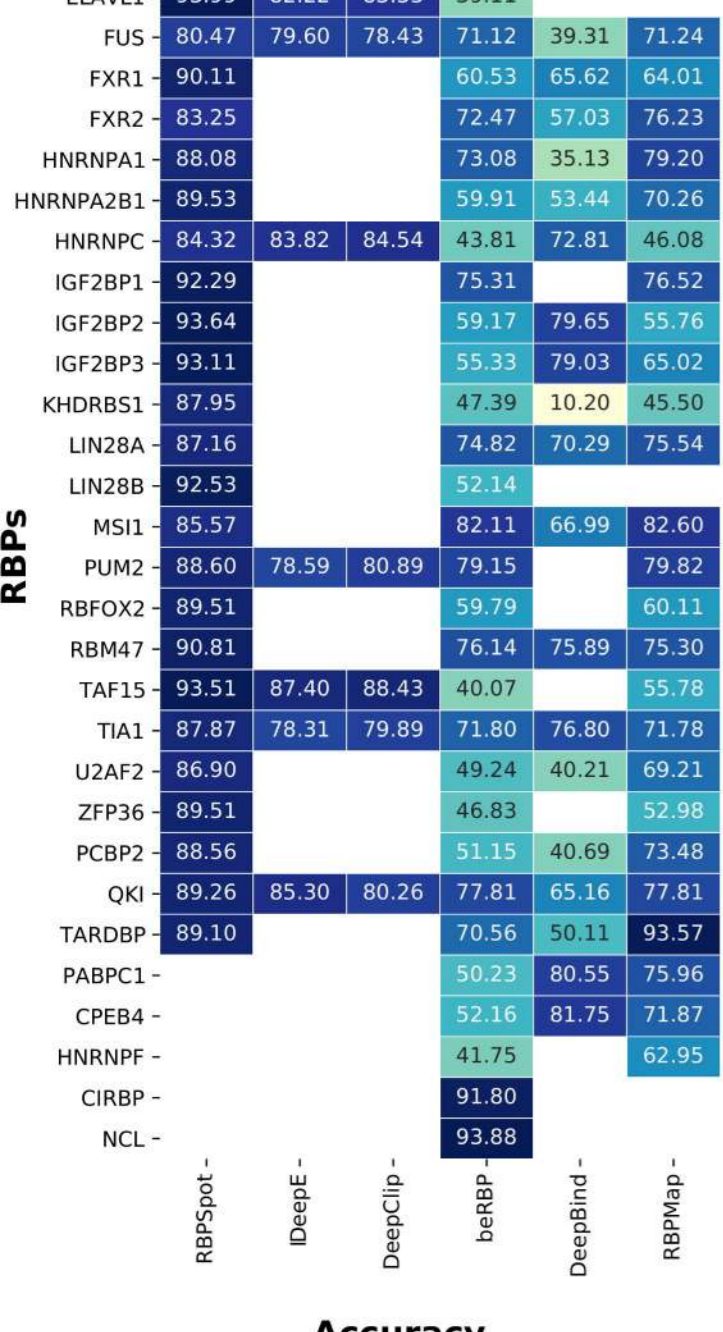
Statistics for set A RBPs



Statistics for set B RBPs

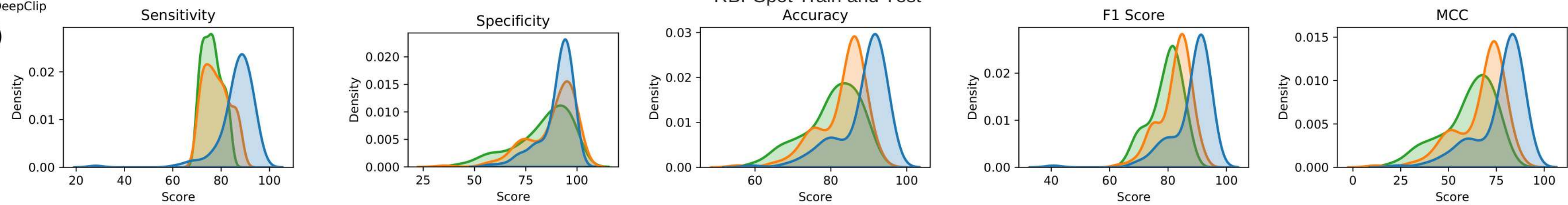




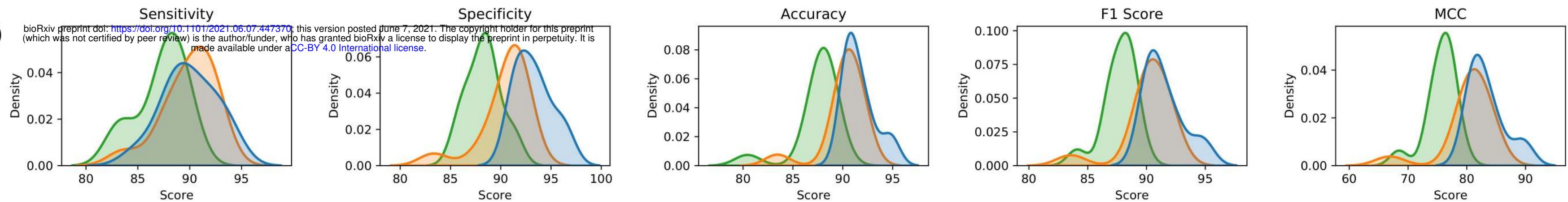
(A) RBPSpot Dataset**(B)****Graphprot Dataset****(C)****beRBP Dataset**

RBPSpot
iDeepE
DeepClip

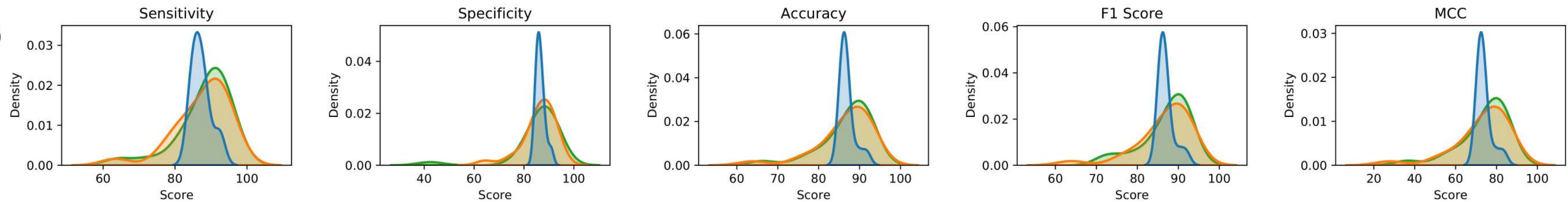
(A)



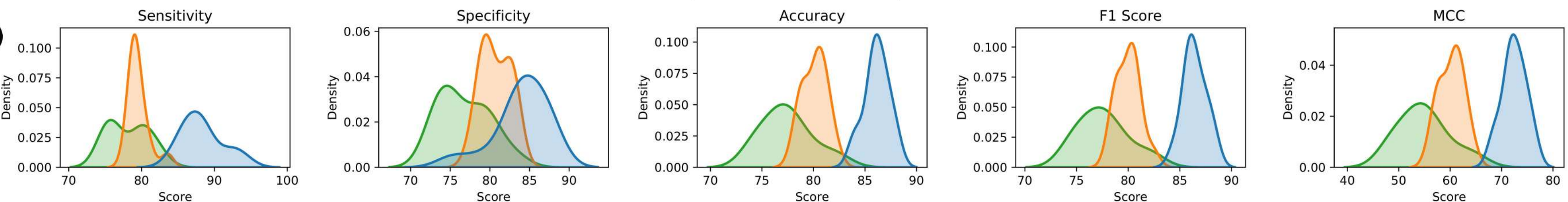
(B)



(C)



(D)



(E)

| | RBPSpot Train and Test | | | | | RBPSpot Train and Graph Prot Test | | | | | Graph Prot Train and Test | | | | | Graph Prot Train and RBPSpot Test | | | | |
|----------|------------------------|-------------|-------------|----------|----------|-----------------------------------|-------------|-------------|----------|----------|---------------------------|-------------|-------------|----------|----------|-----------------------------------|-------------|-------------|----------|----------|
| Tools | MCC | SENSITIVITY | SPECIFICITY | ACCURACY | F1 SCORE | MCC | SENSITIVITY | SPECIFICITY | ACCURACY | F1 SCORE | MCC | SENSITIVITY | SPECIFICITY | ACCURACY | F1 SCORE | MCC | SENSITIVITY | SPECIFICITY | ACCURACY | F1 SCORE |
| RBPSpot | 77.04 | 86.43 | 90.32 | 88.43 | 88.08 | 83.34 | 89.99 | 93.27 | 91.60 | 91.48 | 73.78 | 87.22 | 86.53 | 86.88 | 86.91 | 72.47 | 88.34 | 83.95 | 86.14 | 86.44 |
| DeepCLIP | 60.39 | 76.24 | 83.31 | 80.38 | 79.35 | 75.61 | 87.27 | 88.30 | 87.59 | 87.72 | 74.06 | 88.15 | 85.81 | 86.86 | 87.01 | 54.93 | 78.15 | 76.76 | 77.44 | 77.60 |
| iDeepE | 67.02 | 78.24 | 87.89 | 83.06 | 82.40 | 80.39 | 89.89 | 90.46 | 90.18 | 90.15 | 73.01 | 87.56 | 85.48 | 86.48 | 86.55 | 60.24 | 79.53 | 80.72 | 80.11 | 79.99 |

