Supplementary Information: Deep neural network for interpreting RNA binding protein target preferences

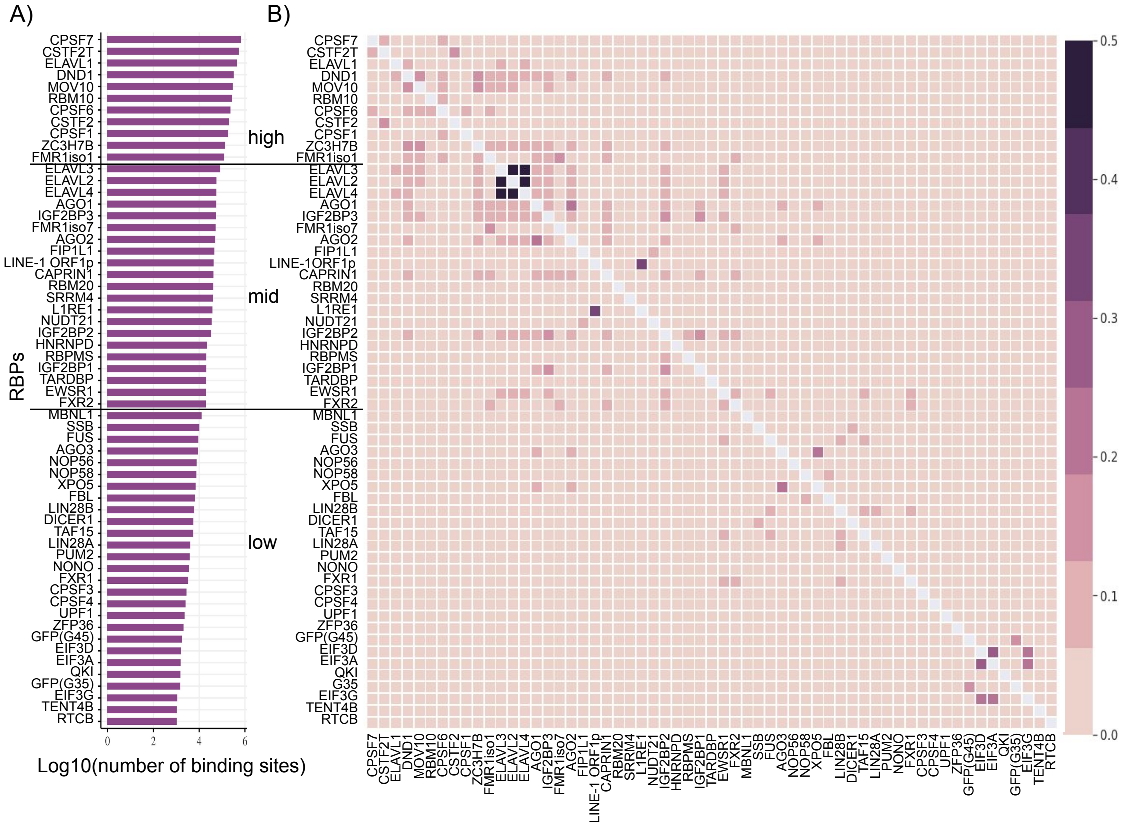
Mahsa Ghanbari[1] and Uwe Ohler[1,2,3]

[1]The Berlin Institute for Medical Systems Biology, Max Delbrück Center for Molecular Medicine, Hannoversche Str. 28, Berlin 10115, Germany
[2]Department of Biology, Humboldt Universität zu Berlin, Unter den Linden 6, Berlin 10117, Germany
[3]Department of Computer Science, Humboldt Universität zu Berlin, Unter den Linden 6, Berlin 10117, Germany

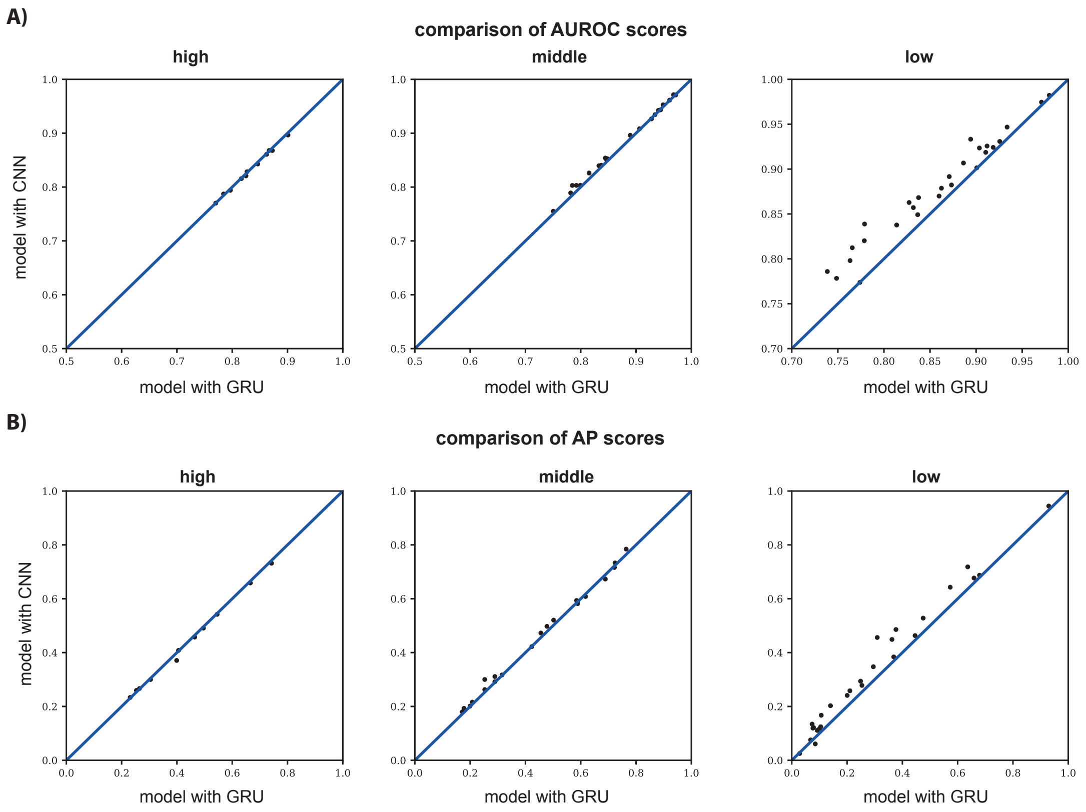
Supplemental information includes Supplemental Table S1 and Supplemental Figures S1-13.

RBP	AUROC	AP
AGO1	0.789076	0.317035
AGO2	0.853832	0.49748
AGO3	0.868226	0.485672
CAPRIN1	0.755036	0.216009
CPSF1	0.770088	0.233365
CPSF3	0.798064	0.118253
CPSF4	0.778281	0.0757451
CPSF6	0.787158	0.259414
CPSF7	0.793916	0.542165
CSTF2	0.815647	0.3
CSTF2T	0.842871	0.658515
DICER1	0.857085	0.241041
DND1	0.820734	0.457522
EIF3A	0.882287	0.202734
EIF3D	0.869932	0.124661
EIF3G	0.891675	0.134418
ELAVL1	0.896556	0.731773
ELAVL2	0.926606	0.60828
ELAVL3	0.943415	0.716039
ELAVL4	0.934444	0.581832
EWSR1	0.852801	0.201107
FBL	0.906787	0.347475
FIP1L1	0.803026	0.300094
FMR1iso1	0.867917	0.26645
FMR1iso7	0.896127	0.520597
FUS	0.901412	0.463117
FXR1	0.862783	0.2582
FXR2	0.803092	0.180022
GFP(G35)	0.820182	0.0609622
GFP(G45)	0.838807	0.111121
HNRNPD	0.942418	0.47266
IGF2BP1	0.826114	0.192926
IGF2BP2	0.839561	0.291862
IGF2BP3	0.840454	0.422296
L1RE1	0.961325	0.589827
LIN28A	0.785889	0.167328
LIN28B	0.923507	0.448803
MBNL1	0.982158	0.944225
MOV10	0.828301	0.408187
NOP56	0.924164	0.687188
NOP58	0.930788	0.676819
NUDT21	0.850143	0.26264
LINE-1 ORF1p	0.971041	0.673126
NONO	0.925687	0.383958
TENT4B(PAPD5)	0.8492	0.121876
PUM2	0.946767	0.718361
QKI	0.97455	0.642795
RBM10	0.860757	0.490855
RBM20	0.908388	0.5935
RBPMS	0.971549	0.784266
RTCB	0.773793	0.0252931
SRRM4	0.803274	0.311076
SSB	0.918654	0.52801
TAF15	0.878692	0.278177
TARDBP	0.952737	0.733116
UPF1	0.812371	0.119511
XPO5	0.837698	0.293879
ZC3H7B	0.867818	0.370814
ZFP36	0.933268	0.456097

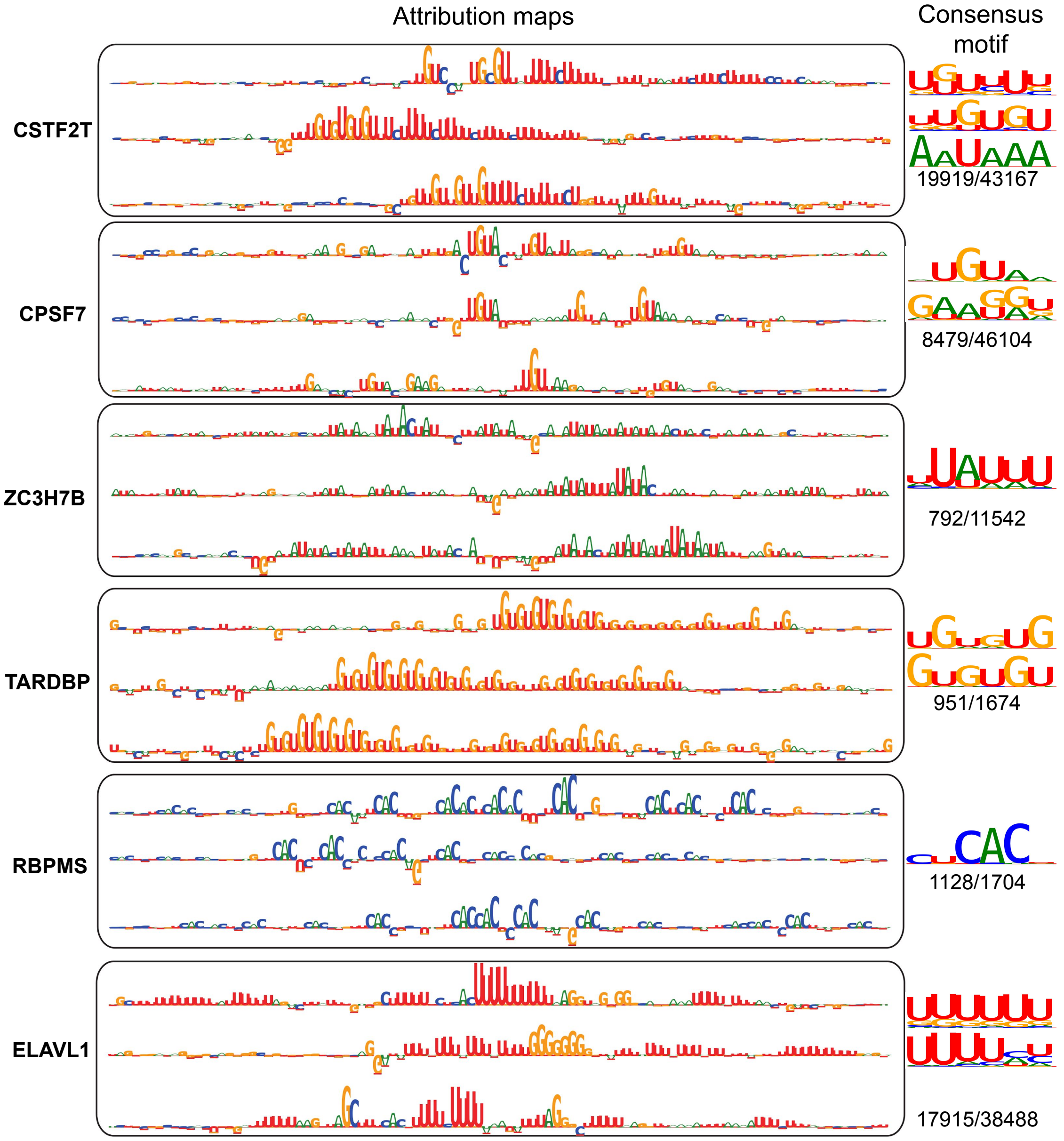
Supplemental Table S1: Classification performance of DeepRiPe: AUROC as well as AP scores for all 59 PAR-CLIP datasets.



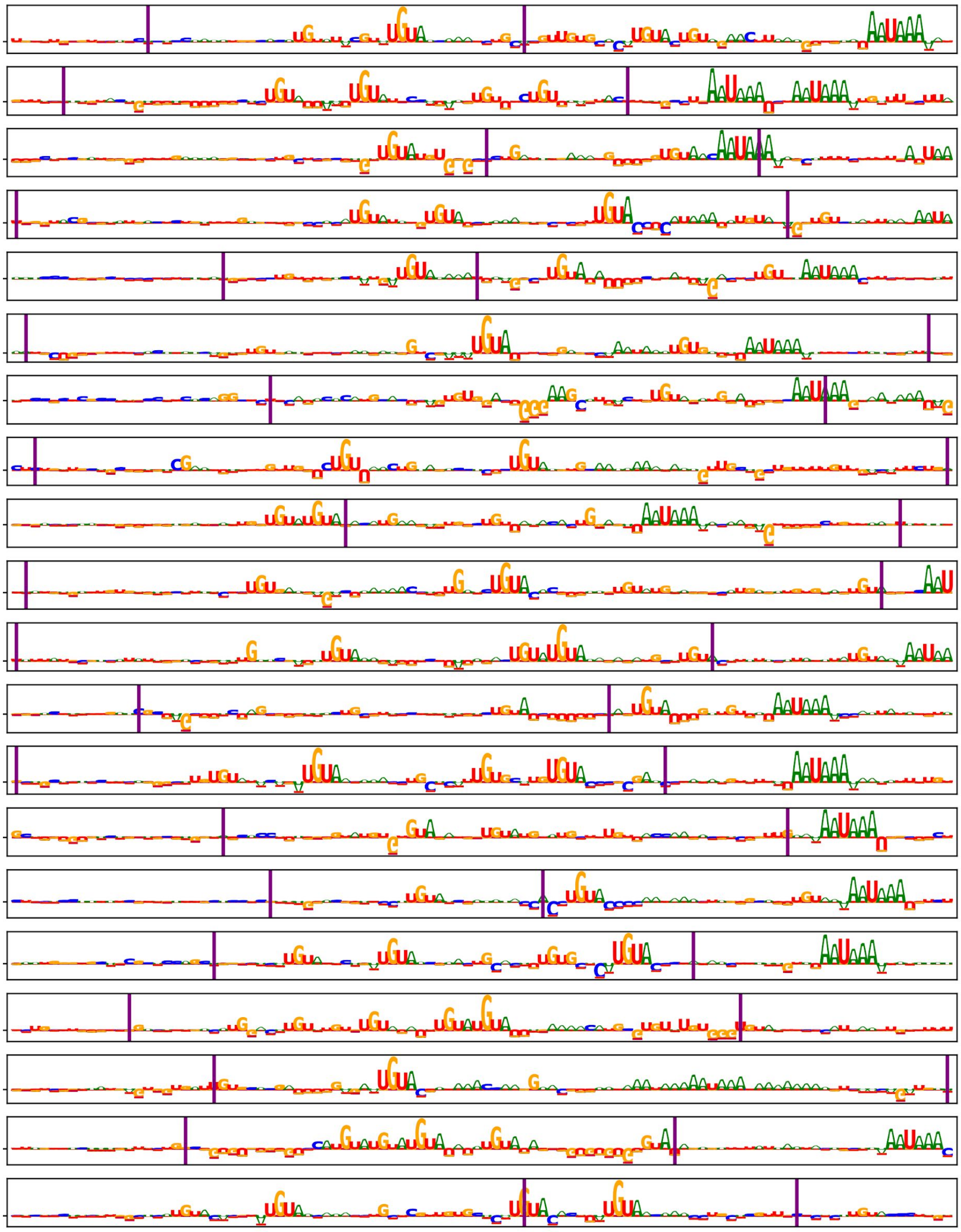
Supplemental Figure S1: Overview of datasets: A) The number of PAR-CLIP peaks for RBPs used in this study. We divided RBPs into 3 categories: RBPs with more than 10^5 peaks (high), RBPs that have between 15000 and 10^5 peaks (mid) and RBPs with less than 15000 peaks (low). B) Overlap between peaks of RBPs in terms of Jaccard index. For each pairs of RBPs, we calculate the Jaccard index = number of genome bins that both RBPs bind / number of genome bins that at least RBP binds.



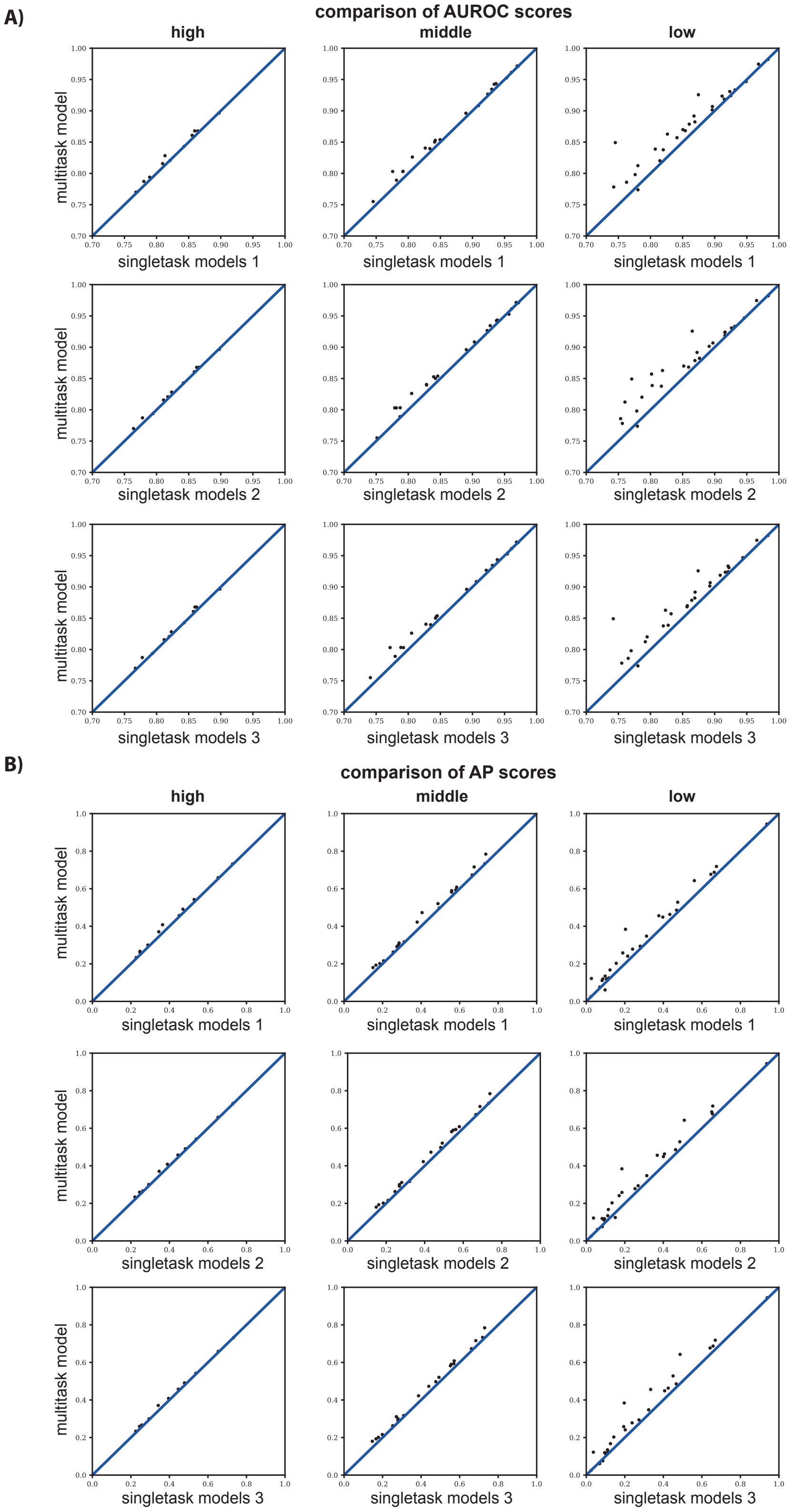
Supplemental Figure S2: Assessing the performance of the DeepRiPe when using GRU instead of CNN in the multitask module. Scatter plots comparing the AUROC (A) and AP scores (B) of DeepRiPe with CNN vs DeepRiPe with GRU model. Each data point represents an RBP and it falls above the diagonal when model with CNN outperforms the one with GRU. The results show that GRU does not help the model specially for low-model, most likely due to the lack of data for training GRU with more parameters compare to CNN.



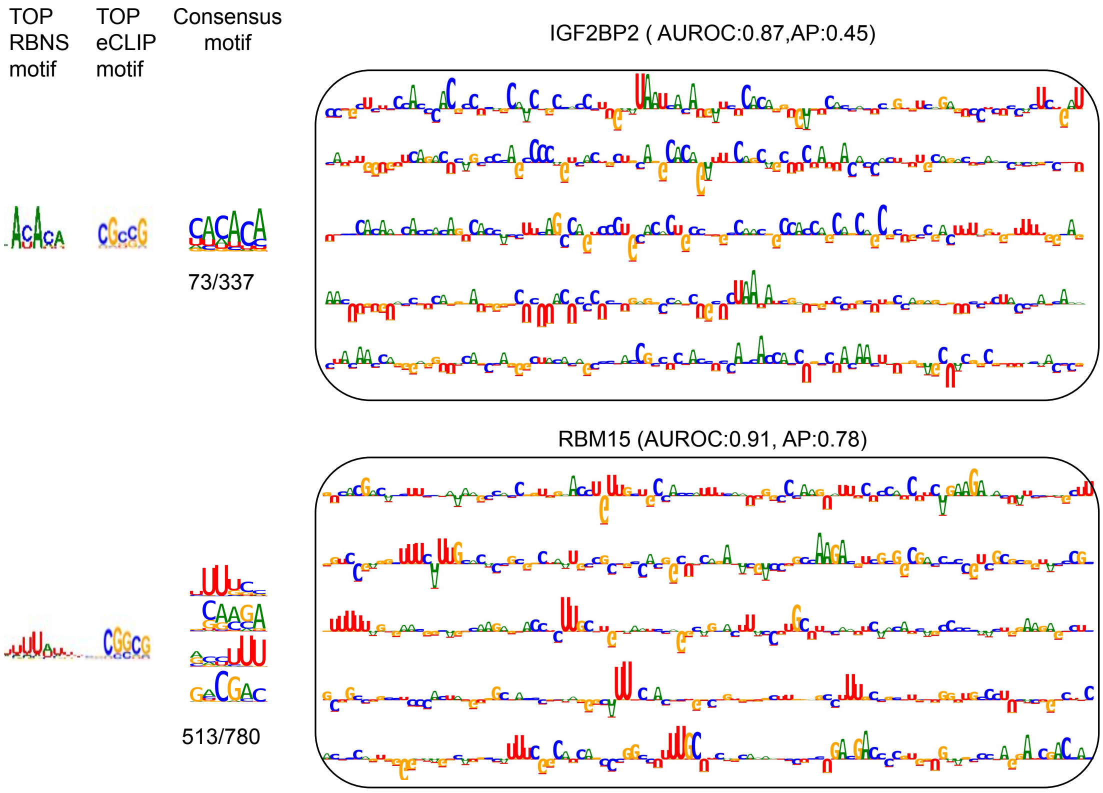
Supplemental Figure S3: Attribution maps of several RBPs.For each RBP,the sequence logos corresponding to attribution maps of three true binding sites with the highest DeepRiPe prediction scores are shown. Consensus motifs, obtained from attribution maps of all true positive binding sites of the RBP in the test set with prediction scores larger than 0.5, are shown next to attribution maps. The ratio of the number of binding sites larger than 0.5 to the total number of CLIP binding sites in the test set is listed below the consensus motif



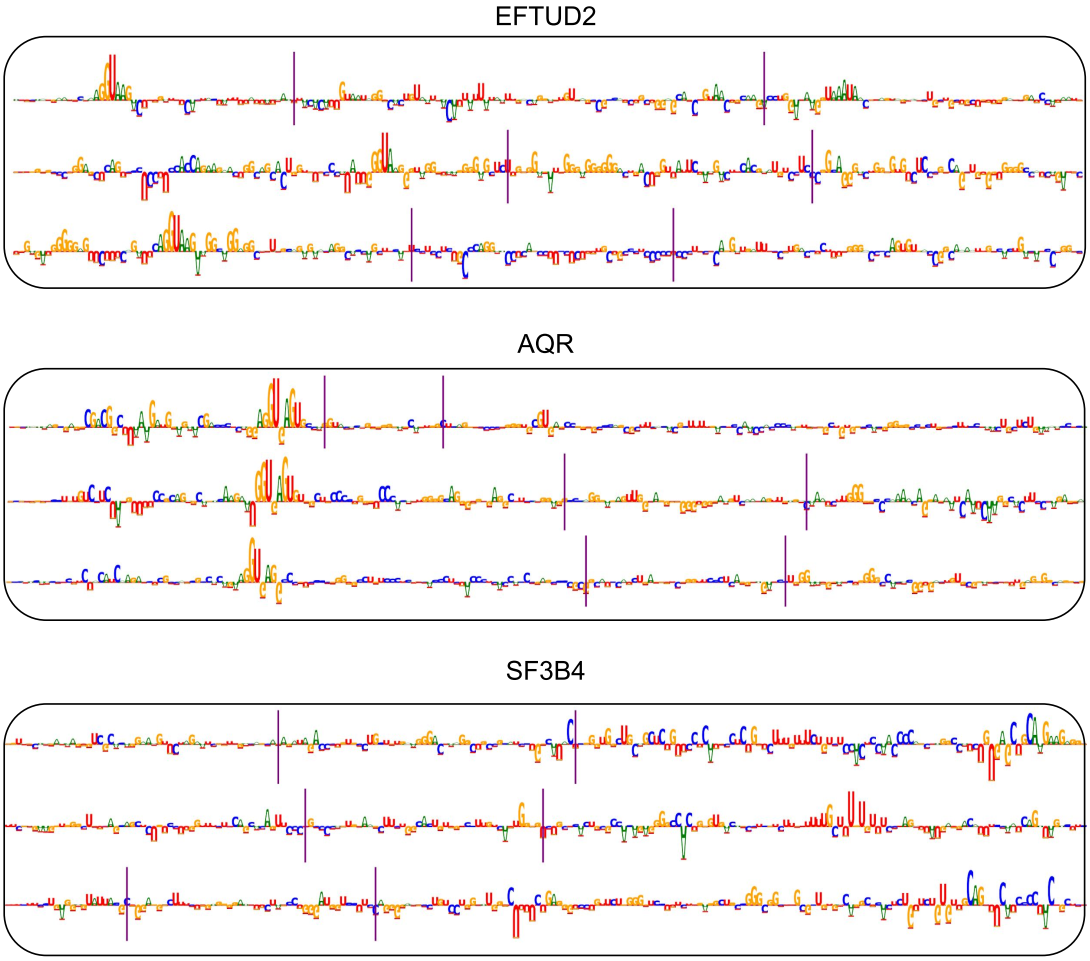
Supplemental Figure S4: Attribution maps of CPSF6. The sequence logos corresponding to attribution maps for 20 true binding sites of CPSF6 with the highest DeepRiPe prediction scores. The lines indicate the position of actual peaks along input sequences. UGUA motif is always located inside the peak, while this is not the case for AAUAAA motif.



Supplemental Figure S5: Assessing the performance of multitask model vs singletask models. We subsample from negative samples of the training and validation datasets to ensure an equal number of negative samples as positive samples in these datasets (single models 3). Using all negative samples for training singletask models that have less positives samples leads to a bad performance due to imbalanced data. , A) Scatter plots comparing AUROC scores of DeepRiPe and singletask models. , B) Scatter plots comparing AP scores of DeepRiPe and singletask models.



Supplemental Figure S6: Comparison of motifs obtained from in vitro (RBNS) and in vivo (eCLIP) experiments with patterns observed in attribution maps. For each RBP, the motifs obtained from RBNS, eCLIP and attribution maps along with attribution maps for top five inputs with highest prediction scores are shown. The consensus motifs obtained from attribution maps correspond to all true binding sites with prediction scores larger than 0.5. The ratio of the number of binding sites used to obtain consensus motif to the number of all true binding sites is mentioned along with corresponding consensus motif.

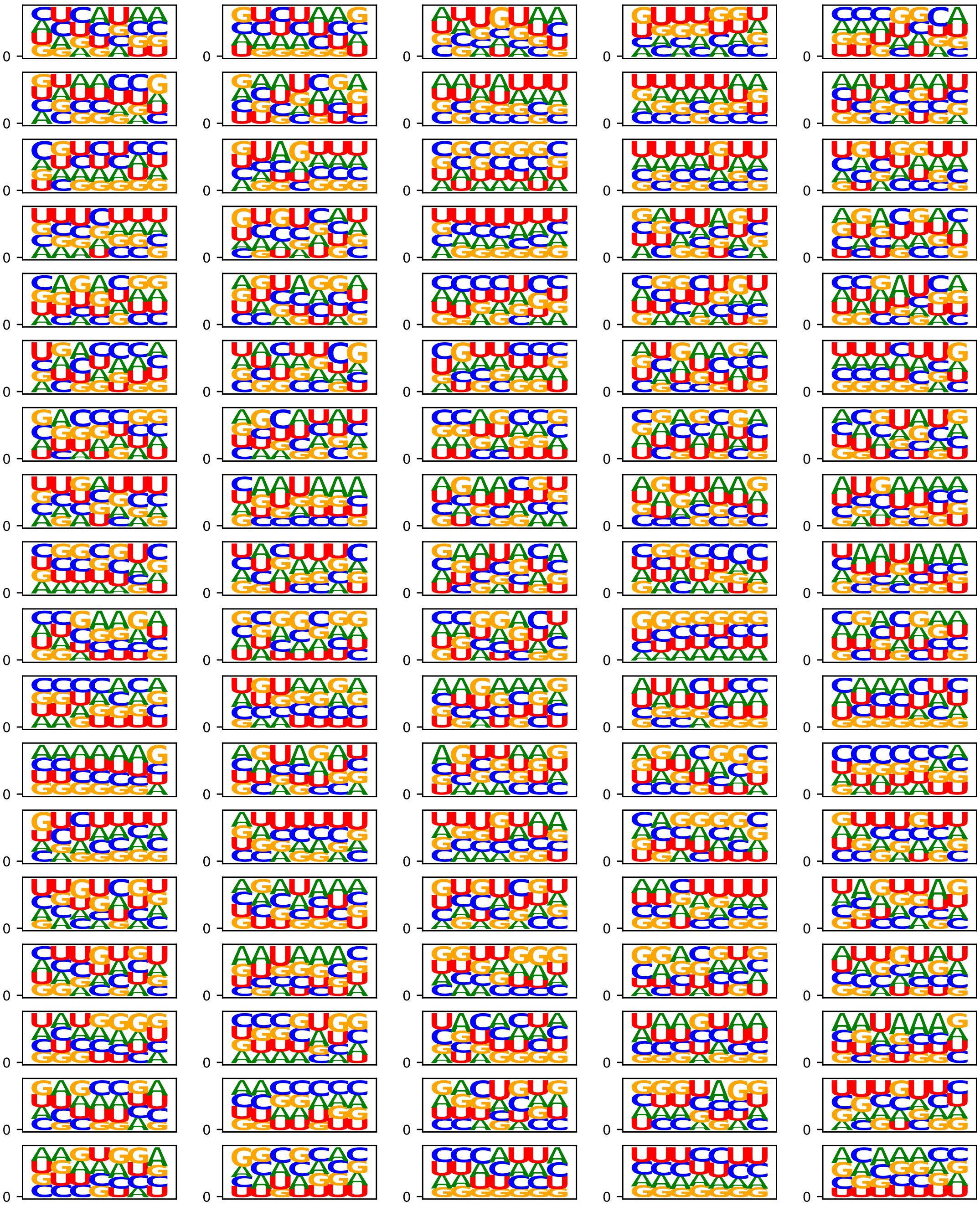


EFTUD2

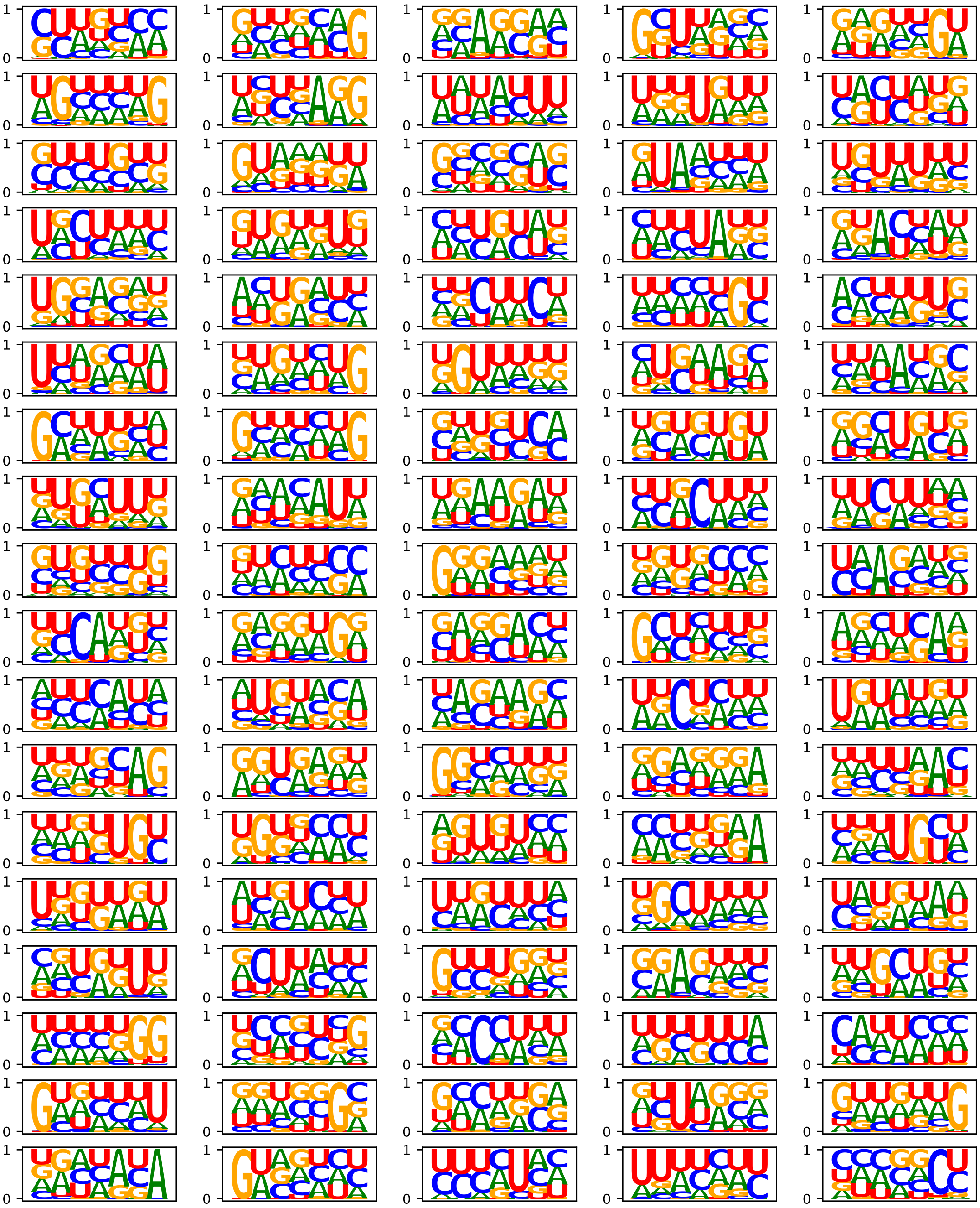
AQR

SF3B4

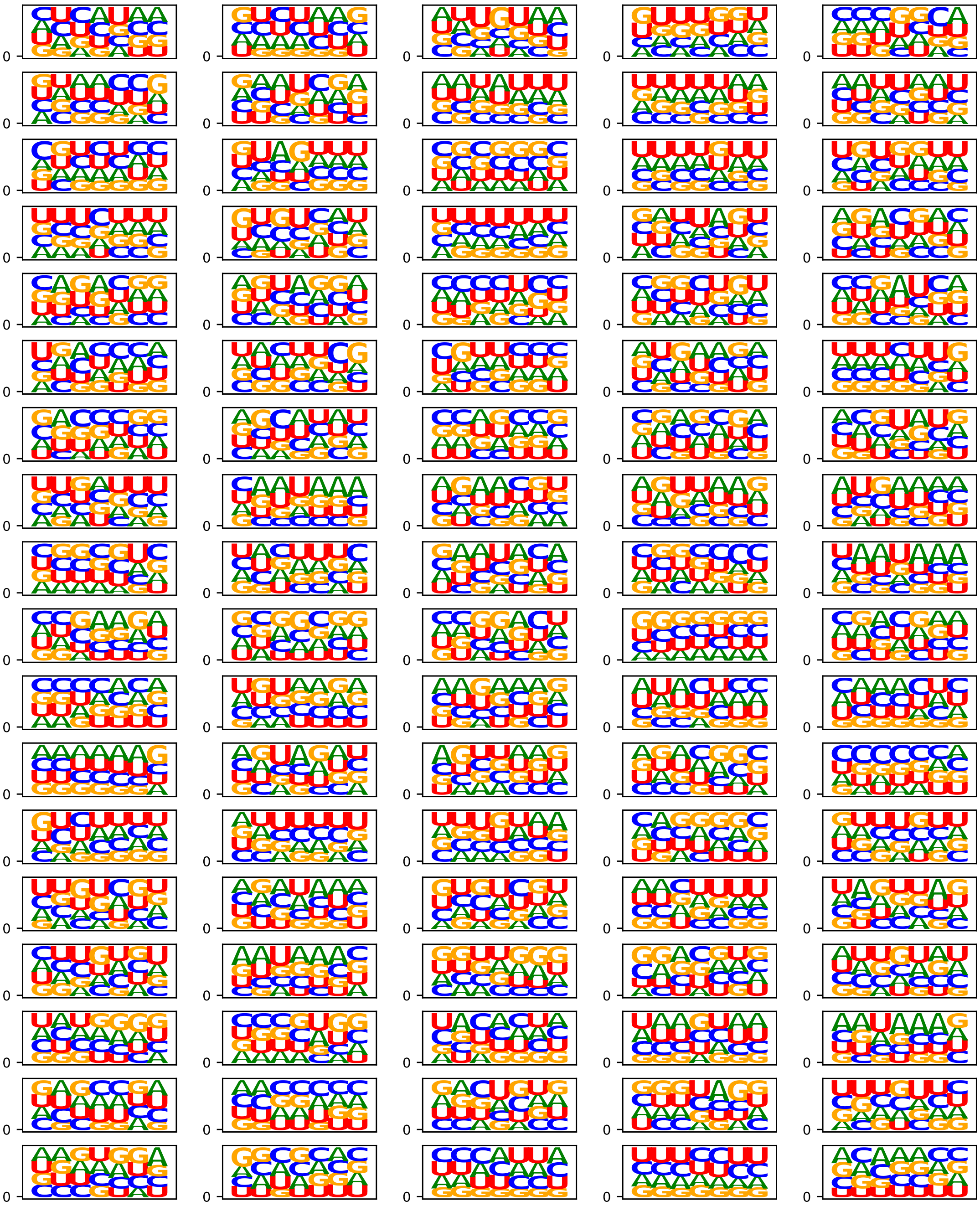
Supplemental Figure S7: Attribution maps of some splicing factors. 3' and 5' splice site motifs (CAG, GGUAAG) are observed in the attribution maps of EFTUD2, AQR and SF3B4. The lines indicate the position of actual peaks along input sequences. The observed motifs are not always located inside the peaks and they are not involved in direct interactions.



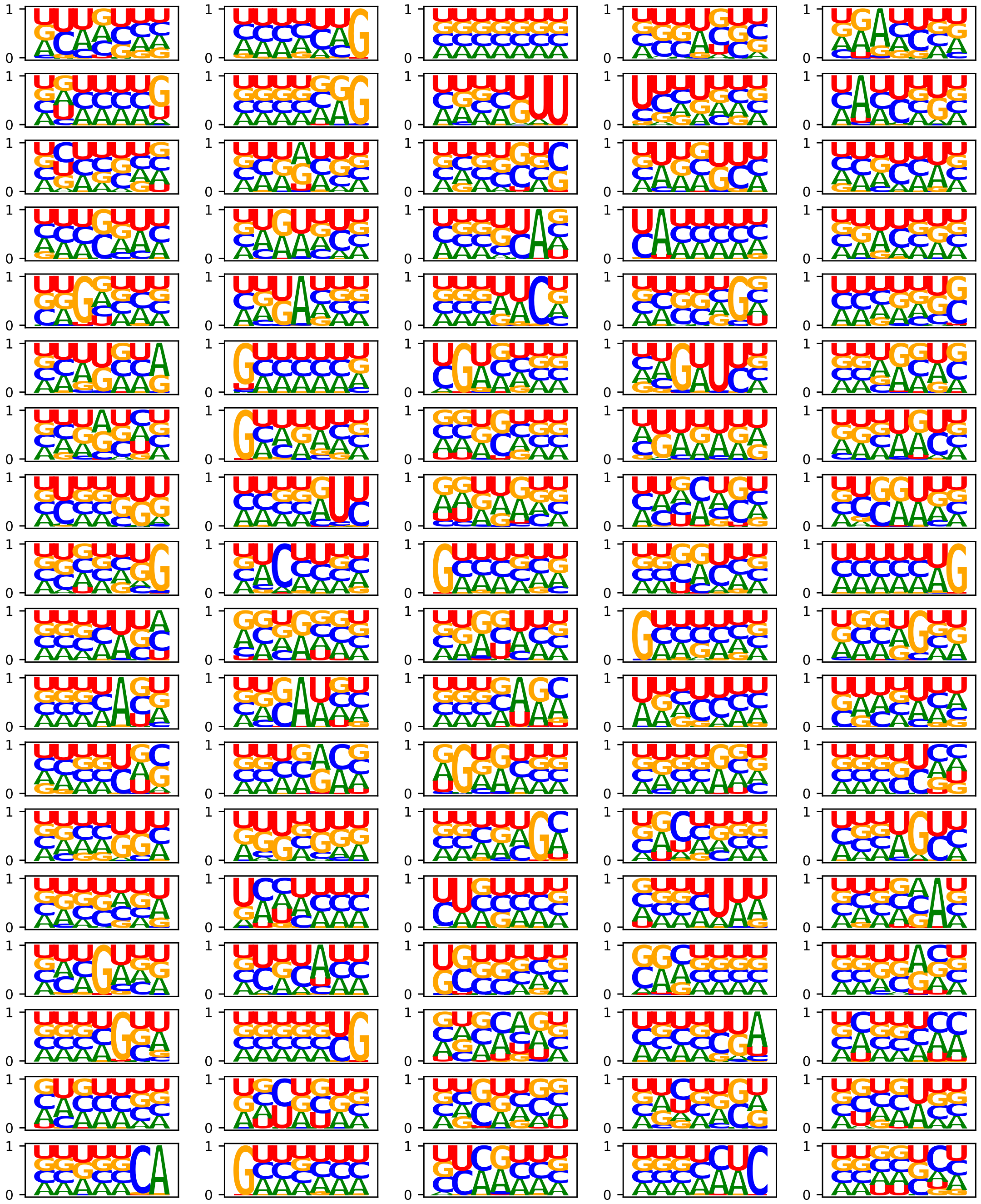
Supplemental Figure S8: Visualization of the filters' weights in the first convolutional layer of low-model in the form of PWM.



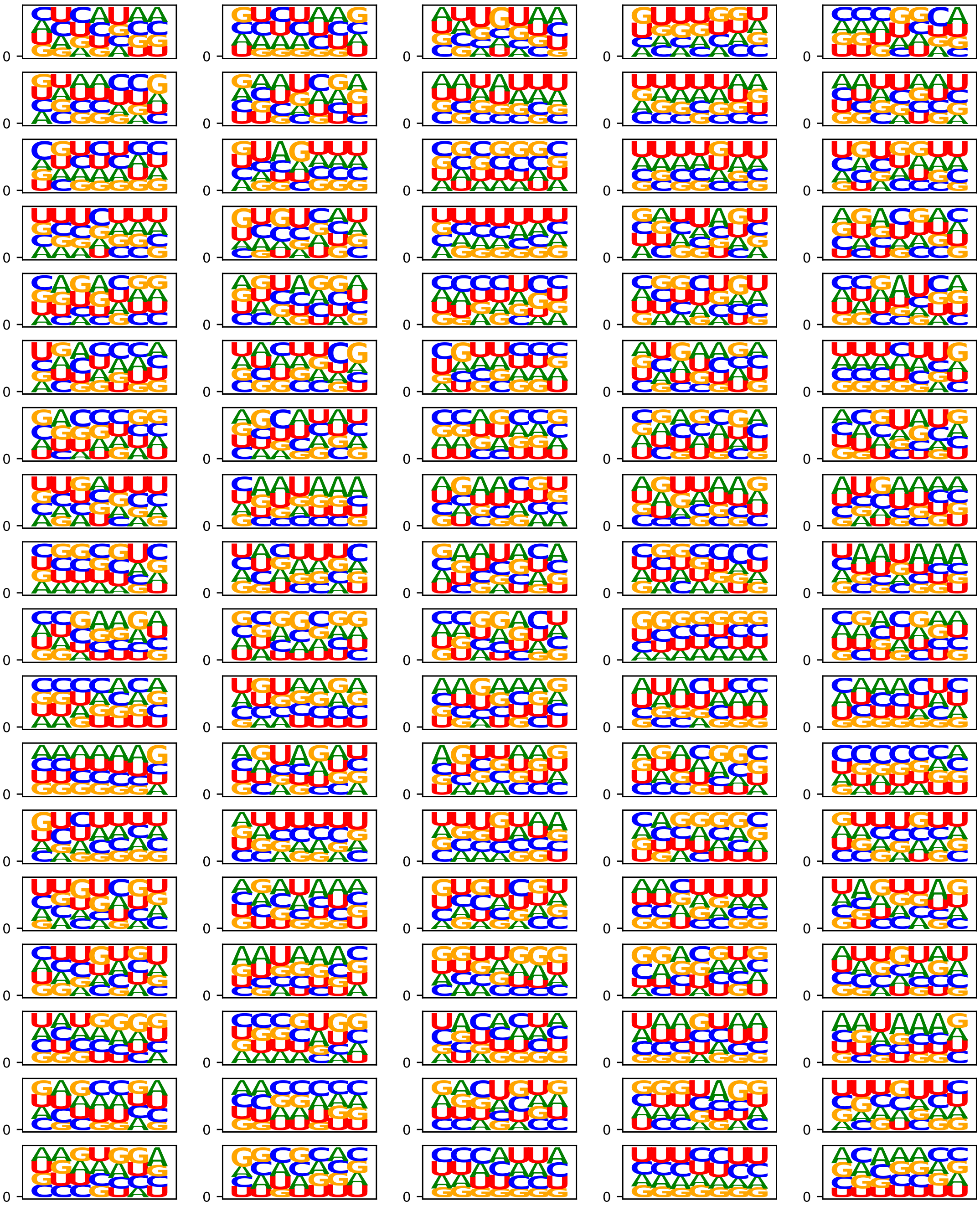
Supplemental Figure S9: Visualization of the filters in the first convolutional layer of low-model by averaging over inputs that activates the filter. For each filter, we averaged over inputs subsequences that activate the neuron corresponding to the filter.



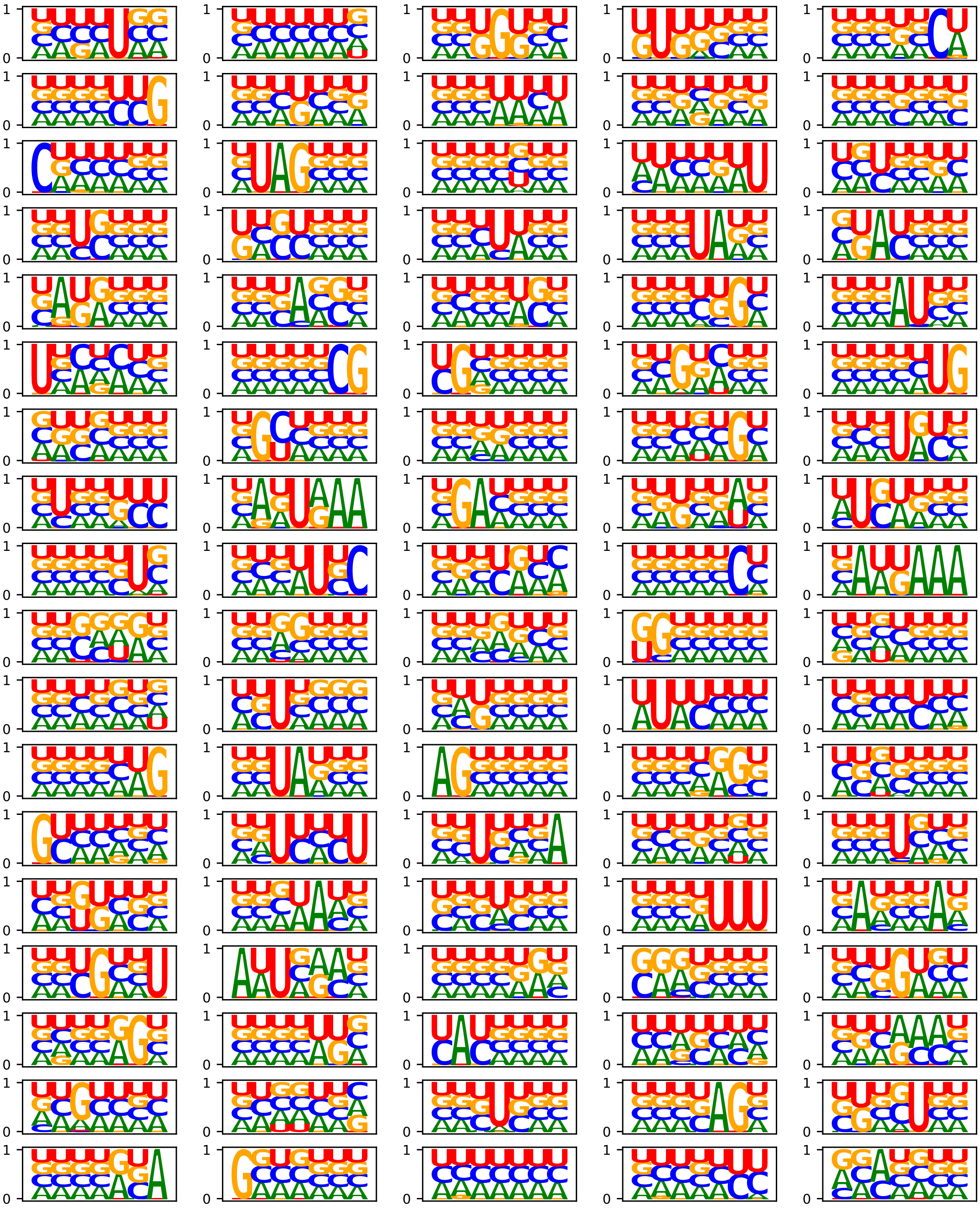
Supplemental Figure S10: Visualization of the filters' weights in the first convolutional layer of mid-model in the form of PWM.



Supplemental Figure S11: Visualization of the filters in the first convolutional layer of mid-model by averaging over inputs that activates the filter. For each filter, we averaged over inputs subsequences that activate the neuron corresponding to the filter.



Supplemental Figure S12: Visualization of the filters' weights in the first convolutional layer of high-model in the form of PWM..



Supplemental Figure S13: Visualization of the filters in the first convolutional layer of high-model by averaging over inputs that activates the filter. For each filter, we averaged over inputs subsequences that activate the neuron corresponding to the filter.