Predicting interactions of m^{6}A-regulated RNA-binding proteins

Gebele, Jonas[1], Wilming, Frederic[1]

Abstract
The rising success of mRNA vaccines has had a major role in showing the importance of chemical modifications and how they can influence interactions drastically. We put our focus on a specific modification and how it can affect interactions between RNA and proteins. It has been shown for the often-occurring m^{6}A modification that it seems to play a part in regulating the interaction between RNA and proteins, specifically RNA-binding proteins (RBPs). As this is a well-researched topic, it provides us with many crosslinking immunoprecipitation (CLIP) datasets for RBPs binding sites and the methylation of those, which we used to train a convolutional neural network (CNN). We then assessed different encodings of methylation, models and dataset distribution regarding their predictions on m^{6}A regulated RBPs.

Keywords
Machine Learning — RNA binding proteins — Methylation

[1] *Contributed equally to this work*
*__Corresponding authors__: jonas.gebele@in.tum.de, frederic.wilming@tum.de

Contents

Introduction

Deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) are the code that all life we as humans have encountered runs on [1]. Ever sinking sequencing costs lead to them being a hot topic in the machine learning community to use as data to learn part of the complexity of life from [2]. In DNA is all the genomic information stored, which acts as a blueprint for life, but RNA and its myriad of sub forms form the second most common group of molecules in an average E. coli cell, that actually forms the network around it [3]. This can be seen by the fact, that the most effective vaccines against Covid-19 are all mRNA-based [4]. What is less known is the fact that RNA modifications play a huge role in that process as they are able to reduce the recognition of the immune system and therefore the immune reaction [5]. A critical player in

[1] Hood, L., Galas, D. The digital code of DNA. Nature 421, 444–448 (2003). https://doi.org/10.1038/nature01410
[2] https://www.illumina.com/techniques/sequencing.html
[3] www.bionumbers.org
[4] Szabó GT, Mahiny AJ, Vlatkovic I. COVID-19 mRNA vaccines: Platforms and current developments. Mol Ther. 2022;30(5):1850-1868. doi:10.1016/j.ymthe.2022.02.016
[5] Karikó K, Buckstein M, Ni H, Weissman D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modifica-

regulating different mechanisms of mRNA processing are RBPs [6].

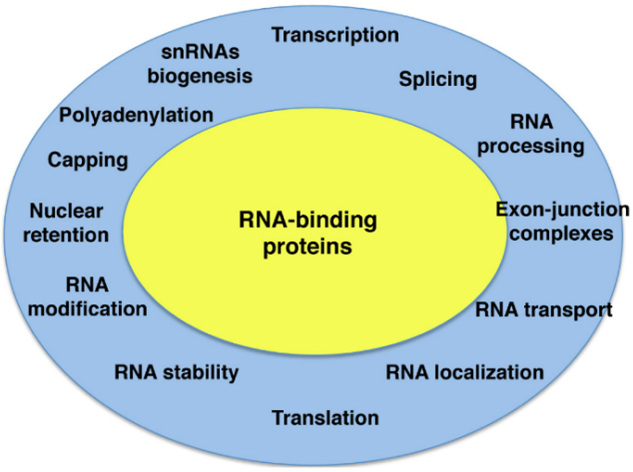

Figure 1. RBPs and their different roles[7]

This is done by directly binding to specific sequence regions; therefore, the investigation of these interactions is of high relevance. Machine Learning has been used trying to predict these interactions based on the RNA sequence but fails at capturing high-fidelity interaction footprints in vivo [8]. A reason for this could be that they do not take into account highly dynamic factors of the cell environment, that can impact binding. One of these dynamic factors could be chemical modifications of the RNA. Modifications generally are just deviations of the basic structure of the 4 nucleotides, often in form of additions like methylation and acetylation or atom substitution [9]. Depending on where and what modification is happening they are labeled. In our case we will focus on m^{6}A or N^6-methyladenosine, a relatively often occurring RNA modification [10] that affects adenosines, one of the 4 bases, and has proven to be able to disrupt or facilitate access to binding sites by proteins through modifying RNA structure [11].

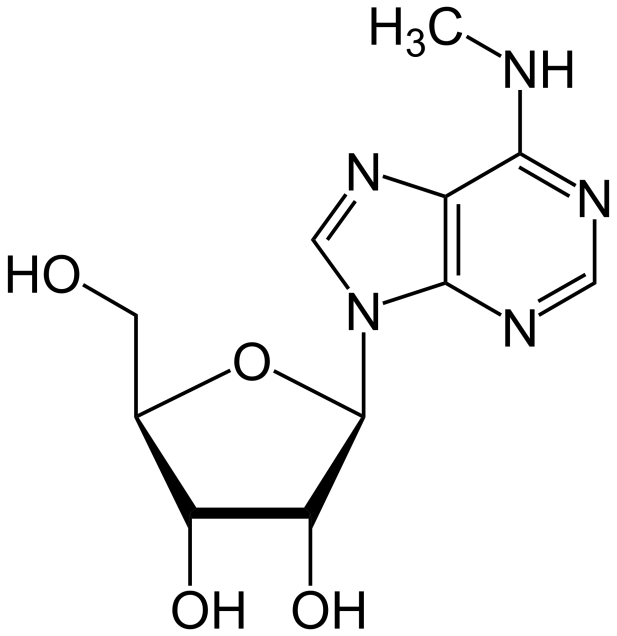

Figure 2. N^6-Methyladenosine (m^{6}A)

The data that we want to use is gathered with CLIP. It is a very common method and it is used to find the binding sites of RBPs as well as the position of m^{6}A [13]. The principle stays mostly the same, anything interaction with DNA or RNA is getting crosslinked, in our case, RBPs are covalently bound to RNA using UV light [15]. After that, the cell is lysed and the RBP of interest is isolated via immunoprecipitation using the power of highly specific binding like the antibodies of our immune system. For m6A it is similar but with an adapted procedure to capture the methylation information.

While this is a fairly good method, it cannot catch all transcript as some might not be expressed in the cell line of the CLIP-seq experiment. This opens the door for computational prediction, often CNN architectures that use the nucleotides of an RNA sequence as input similar to pixels in computer vision.

tion and the evolutionary origin of RNA. Immunity. 2005;23(2):165-175. doi:10.1016/j.immuni.2005.06.008

[6] Zhao et al, POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Research 50, Issue D1, D287-D294 (2022).

[8] Sun, L., Xu, K., Huang, W. et al. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. Cell Res 31, 495–516 (2021). https://doi.org/10.1038/s41422-021-00476-y

[9] Frye, M., Jaffrey, S., Pan, T. et al. RNA modifications: what have we learned and where are we headed?. Nat Rev Genet 17, 365–372 (2016). https://doi.org/10.1038/nrg.2016.47

[10] Jiang X, Liu B, Nie Z, Duan L, Xiong Q, Jin Z, Yang C, Chen Y. The role of m6A modification in the biological functions and diseases. Signal Transduct Target Ther. 2021 Feb 21;6(1):74. doi: 10.1038/s41392-020-00450-x. PMID: 33611339; PMCID: PMC7897327.

[11] Arguello et al, RNA Chemical Proteomics Reveals the N6-Methyladenosine (m6A)-Regulated Protein–RNA Interactome. Journal of the American Chemical Society 139, Issue 48, 17249-17252 (2017).

[13] Hafner et al, CLIP and complementary methods. Nature Reviews Methods Primers 1, 1, 1-23 (2021).

[15] Huppertz I.; et al. iCLIP: protein-RNA interactions at nucleotide resolution. Methods. 2014, 65(3): 274-287.

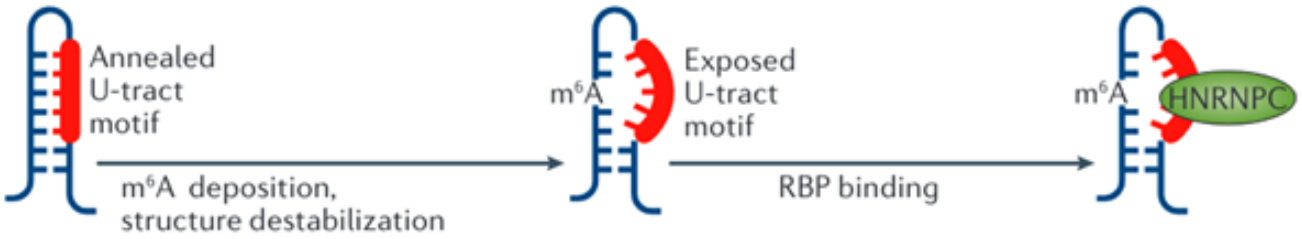

Figure 3. Mechanism of how methylation could affect protein-RNA interactions [12]

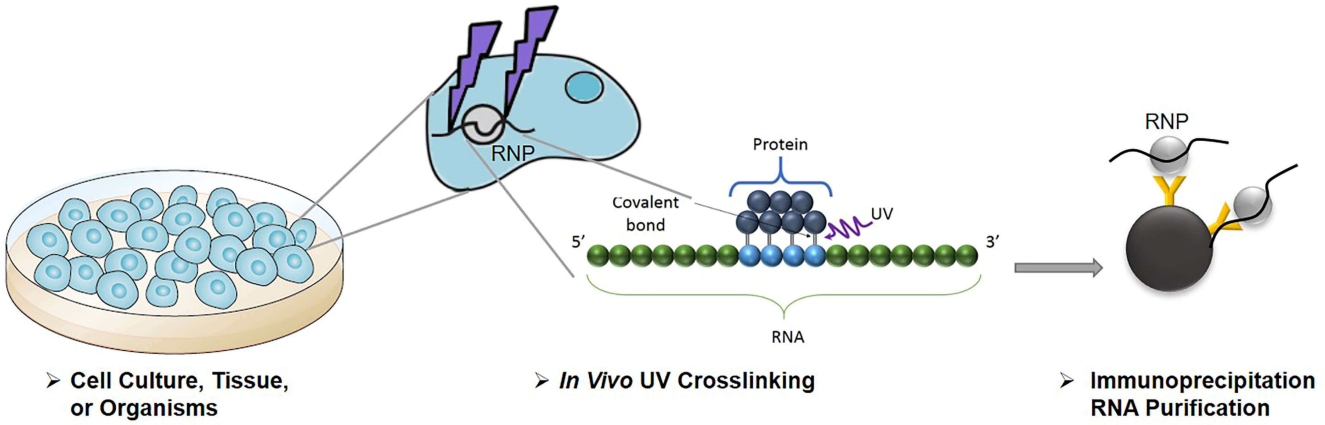

Figure 4. Schematic illustration of the principle behind CLIP[14]

Using this we want to build a deep learning model that can incorporate m6A and RNA sequence data to better grasp the binding preferences of RBPs in vivo.

1. Methods

1.1 Data Pre-Processing

For the protein-RNA interactions, we used the PAR-CLIP dataset from Mukherjee et al. 2019 [16], which analyzed 66 proteins in HEK293 cells. The methylation data came from the miCLIP dataset that has been also done on HEK293 cells [17]. To train a model we had to combine and process the data into a representation that the model can learn from. As a basis, we used Bedtools v2.30.0 on the Windows Subsystem for Linux (Ubuntu 20.04), a practical toolset for genome arithmetic to work with sequence data and the common .bed file format. Additionally, we developed and made use of an additional python script in the same execution environment [18] in order to parse and modify the bed files which could not be done with the standard Bedtool commands.

In the first step, we generated positive and negative data instances, which we used as classified data instances for our model to train and learn on. The positive samples were available in the PAR-CLIP dataset, but we had to extend them to a uniform length of 200 nucleotides as the sequence length of the original sequences in the bed files would vary. The Bedtools slop function could not be utilized so we developed a bash script [19] that would dynamically determine the center of each sequence and extend from that position the sequence in both directions with a length of 100. For the negative samples, we separated them into two different sub-types. As a first negative type, we sampled randomly in a 200 nucleotide

[16] Neelanjan Mukherjee, Hans-Hermann Wessels, Svetlana Lebedeva, Marcin Sajek, Mahsa Ghanbari, Aitor Garzia, Alina Munteanu, Dilmurat Yusuf, Thalia Farazi, Jessica I Hoell, Kemal M Akat, Altuna Akalin, Thomas Tuschl, Uwe Ohler, Deciphering human ribonucleoprotein regulatory networks, Nucleic Acids Research, Volume 47, Issue 2, 25 January 2019, Pages 570–581, https://doi.org/10.1093/nar/gky1185

[17] Linder, B., Grozhik, A., Olarerin-George, A. et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods 12, 767–772 (2015). https://doi.org/10.1038/nmeth.3453

[18] https://github.com/jonasgebele/Predicting-Interactions-of-m6A-regulated-RNA-Binding-Proteins/blob/main/data/EncodingPreprocessing.py

[19] https://github.com/jonasgebele/Predicting-Interactions-of-m6A-regulated-RNA-Binding-Proteins/blob/main/README.md

window around the positive sequences from both sides of the binding sites as these are safe to be areas where the protein of choice is not binding. The second negative type of sequence is positive samples of other RBPs randomly chosen from the PAR-CLIP dataset.
After we extracted and parsed the methylation data from the miCLIP dataset, we dynamically mapped the positions of the m6As to their binding sites of the respective proteins of the PAR-CLIP data instances.
To build and train our model with Keras we combined our raw sequence data with methylation-rate encodings. This has been done in multiple ways and we compared three different encoding styles, which we later compared and evaluated on their effect on the performance of our models. The original methylation data is saved as a methylation rate of how many sequenced reads contained methylation at this position of all sequenced reads. We encoded it first with one-hot encoding into a binary encoding with every position that contained any methylation as a 1 (see encoding 2). Additionally, we used for the second encoding the rounded-down integer values of the original methylation rates (see encoding 3). As the third option, we used rounded integer values of the methylation rates with an included cutoff value of 0.4 (see encoding 4).
Each instance of our generated datasets consisted of three lines, the label indicating positives or negatives, the sequence data and the methylation encoding in various variances.

```
>1
ACACACAAAAAGCTAATTTGGGATATATATATACAGAATATA...
>0
CAGTAAAATCCAGGGCACAGTGGATTTCAGGCCTTGCTTTTC...
>1
ATATGGGGGATGCCCAATATCTATGCAGGCAGGTGGGGGGAT...
```

Encoding 1. Baseline data instances without encoding of methylation-sites.

```
>1
ACACACAAAAAGCTAATTTGGGATATATATATACAGAATATA...
00000000000001000000000001000000000000000...
>0
CAGTAAAATCCAGGGCACAGTGGATTTCAGGCCTTGCTTTTC...
00000000000000000010000000000000000000000...
```

Encoding 2. Data instances with one-hot encoding of methylation-sites.

```
>1
ACACACAAAAAGCTAATTTGGGATATATATATACAGAATATA...
00000000000005000000000003000000000000000...
>0
CAGTAAAATCCAGGGCACAGTGGATTTCAGGCCTTGCTTTTC...
00000000000000000040000000000000000000000...
```

Encoding 3. Data instances with methylation-rate encoding of methylation-sites.

```
>1
ACACACAAAAAGCTAATTTGGGATATATATATACAGAATATA...
00000000000005000000000000000000000000000...
>0
CAGTAAAATCCAGGGCACAGTGGATTTCAGGCCTTGCTTTTC...
00000000000000000040000000000000000000000...
```

Encoding 4. Data instances with methylation-rate encoding with cut-off of 0.4 of methylation-sites.

As part of the pre-processing, we also filtered out duplicates of the sequences for each dataset as well as empty sequences caused by Bedtools generating the sequences from the bed files that were *out-of-range* with *getfasta*.
To later analyze how methylation partakes in protein binding and compare different encodings, we generated multiple datasets for each of the examined proteins. This includes one baseline dataset with only the sequence data and labels (see encoding 1), the datasets with the different encodings of the methylation rate and last but not least a dataset with only sequences that contained methylation site encodings. We ascertained that across all examined proteins only around 5 percent of all sequences have matching methylation sites. Therefore, these datasets were significantly reduced in size by a factor of around 20.
The datasets were balanced by label type and contained around 50 percent positives, each 12.5 percent negatives of type one from the left and from the right as well as 25 percent negatives of type two.
We build these datasets for the CAPRIN1, EIF3A, HNRNPs and IGF2BPs m6A reader proteins and then trained our models on them.

1.2 Network Architecture

The convolutional neural network architecture we used utilizes 3 sequential convolutional blocks followed by flattening the output and two fully connected layers in the end (see figure 6).
Our convolutional neural network took a 200x5 matrix as input where 200 was the sequence length representing the number of nucleotides examined combined with the methylation rate encoding at each position. The

Table 1. Analyzed proteins that have been reported to regulate or interact with m^{6}A.

RBP	SRA	Reads	Name
CAPRIN1	SRR500485	733138	$CAPRIN1_SRR500485$
EIF3A	SRR1761289	150438	$EIF3A_SRR1761289$
HNRNPD	SRR1042842	870518	$HNRNPD_SRR1042842$
IGF2BP1	SRR048951	571507	$IGF2BP1_SRR048951$
IGF2BP2	SRR048957	891603	$IGF2BP2_SRR048957$
IGF2BP3	SRR048963	984629	$IGF2BP3_SRR048963$

four base pairs that the sequence is made of are encoded in a four-dimensional identity vector representing a one-hot encoding with a fifth additional dimension for the methylation site encoding: [1,0,0,0,M], [0,1,0,0,M], [0,0,1,0,M], [0,0,0,1,M]. The first layer can be thought of as a feature scanner of the sequences, which used a kernel size of 8 for the length of the 1D convolution window and 16 filters for the output space. As the second layer, a max-pooling layer was used for each convolution with a pool size of 2, which only outputs the maximum value of all its respective convolutional layer outputs to filter the relevant sequence features. Different parameters, variance or changing pool sizes after each layer did not show remarkable improvements in the performance of the model. This architecture of convolution and max pooling was repeated in three blocks. Regularization with a dropout layer after each convolution layer showed material improvement in performance on the validation set and was set to 0.4 after extensive hyperparameter tuning. Adding more layers of convolution followed by max-pooling and dropout did not improve the performance of the model materially, which could be due to the limitation of the size of the data. After the repeated convolution we flattened the convolution of the kernel tensors into a flat array. This array was then used as input for one fully connected layer of size 64 which condensed the input of the whole sequence feature tensor into 64 outputs. After this procedure, an additional dropout layer followed in order to generalize from our training data in order not to overfit.

Trying to add more fully connected layers in order to reduce the gap from the 1344 long flattened input to the 64 output size of the fully-connected layers showed no improvement in the performance and leads to further overfitting of the model. The final layer consists of one neuron with a sigmoid activation function doing the classification of the prediction.

Adam was the optimizer of our choice. Manual hyperparameter tuning, as well as multiple runs with Keras Tuner, helped us determine the best optimizer as well as our optimal learning rate of 0.0005.

Different convolutional neural network architectures with more layers or different kernel sizes were empirically discarded through hyperparameter tuning.

Additionally, we tried using a recurrent neural network architecture, but likewise, like other papers on similar topics [20], we couldn't materially improve the accuracy of our model.

1.2.1 Baseline Architecture

The architecture of the baseline model for comparing methylation encoded data with sequences without encoding is similar to the main model (see figure 5).

As input size, the model utilizes 200x4 tensors instead of 200x5 as the baseline data lacks the additional methylation site encoding. The reasoning of the hyperparameters and the architecture is similar as described in the section above.

Then we define a model that can integrate the m6A signal in different ways.

1.3 Training of the model

We split our randomly sampled data into 80 percent as the training set, 10 percent as the validation set and 10 percent as the test set.

For the training and validation, we chose a batch size of 128 as the number of samples that are propagated through the network at each iteration. Choosing such a small batch size was necessary due to storage limitations in Google Colab and reduced the training time as well.

From that point on we trained a model for each protein, which would converge at an optimum after around 15 to 20 epochs. At this point, we would stop the training as the model began to overfit the training data. We also

[20] Zeng, Haoyang et al. "Convolutional Neural Network Architectures for Predicting DNA–protein Binding." Bioinformatics 32, 12 (June 2016): i121–i127 © 2016 The Authors

Model: "ML4RG_SS22_Model_BASELINE"

Layer (type)	Output Shape	Param #
conv1d_3 (Conv1D)	(None, 192, 64)	2368
dropout_3 (Dropout)	(None, 192, 64)	0
max_pooling1d_3 (MaxPooling 1D)	(None, 96, 64)	0
conv1d_4 (Conv1D)	(None, 93, 64)	16448
dropout_4 (Dropout)	(None, 93, 64)	0
max_pooling1d_4 (MaxPooling 1D)	(None, 46, 64)	0
conv1d_5 (Conv1D)	(None, 43, 64)	16448
dropout_5 (Dropout)	(None, 43, 64)	0
max_pooling1d_5 (MaxPooling 1D)	(None, 21, 64)	0
flatten_1 (Flatten)	(None, 1344)	0
dense_2 (Dense)	(None, 128)	172160
dropout_6 (Dropout)	(None, 128)	0
dense_3 (Dense)	(None, 1)	129

Total params: 207,553
Trainable params: 207,553
Non-trainable params: 0

Figure 5. Architecture of the baseline model

Model: "ML4RG_SS22_Model"

Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 192, 64)	2944
dropout (Dropout)	(None, 192, 64)	0
max_pooling1d (MaxPooling1D)	(None, 96, 64)	0
conv1d_1 (Conv1D)	(None, 93, 64)	16448
dropout_1 (Dropout)	(None, 93, 64)	0
max_pooling1d_1 (MaxPooling1	(None, 46, 64)	0
conv1d_2 (Conv1D)	(None, 43, 64)	16448
dropout_2 (Dropout)	(None, 43, 64)	0
max_pooling1d_2 (MaxPooling1	(None, 21, 64)	0
flatten (Flatten)	(None, 1344)	0
dense (Dense)	(None, 128)	172160
dropout_3 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 1)	129

Total params: 208,129
Trainable params: 208,129
Non-trainable params: 0

Figure 6. Architecture of the model that incorporates methylation

trained a baseline model that we compared with a one-hot encoded model, each for all three proteins. Then we evaluated the models of the different methylation site encodings for all three proteins. Finally, we tested the performance on the subset of the data filtered for sequences with methylation sites in the encoding against the normal distribution of methylation sites which only occurred in approximately five percent of all sequences for the researched proteins.

2. Results

We trained the model and tuned the hyperparameters on each of the datasets for each protein. For all proteins, we generated the test accuracy for each of the different methylation site encodings with 5 runs and stored the results in a five-element vector. EIF3A was removed from the test measurements as its small size of only 2950 sequences was too limited for a concise analysis.

2.1 Encoding m^{6}A data potentially increases performances over a baseline model

While we can see in the boxplot in figure 7 that the model with the one-hot encoded methylation rate performs better on average across all examined proteins in terms of accuracy, it is just slightly more accurate and the overall performance is relatively low.

2.2 Evaluation of different encodings for m^{6}A

In the boxplot in the figure 8 we compared the different encodings of the methylation sites, namely the one-hot encoding of the methylation rate, the rounded-down methylation rate and the rounded-down methylation as a cut-off value with 0.4.
Using a rounded integer version of the original methylation rate seems to be resulting in the highest accuracy across all tested proteins. The one-hot encoded model follows shortly leading to the model with the abbreviated methylation rate performing worst. Overall, the accuracies were considerably low and lay all close to the baseline model.

2.3 Improving performance by modifying distribution of positive and negative samples

In the boxplot in the figure 9 we compared the filtered dataset contained only sequences that encoded any methylation in them, which were around 5 percent of all sequences with the normal distribution of sequences with methylation rate encoding. We can see that the test accuracy of the curated dataset is higher than the normal distribution and that the difference is considerably higher compared to the differences between previous models.

3. Discussion

While we do see an improvement in accuracy by including data regarding methylation sites, it remains unclear

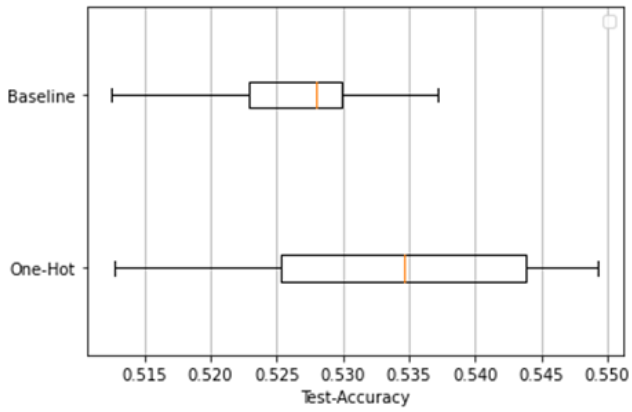

Figure 7. Averaged test accuracies of the baseline model that only takes the sequence as the input with the One-hot model, that additionally uses one hot encoded methylation data.

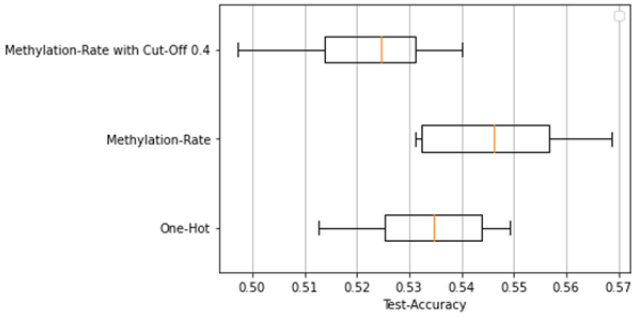

Figure 8. Averaged test accuracies of different methylation encodings. The averaged test accuracies of each protein have been plotted of the different methods used to encode the methylation rate of the sequence.

how strong its actual effect is on naturally distributed datasets, as the overall accuracy of our model was relatively low for all encodings. Compared to other models from the literature the performance of our model was considerably lower [21].

It is unlikely that this is a result of a bad or a too small architecture as we tested different sizes and depths for it. We orientated ourselves for both of our models at other common models from scientific literature with an accuracy of around 80 percent and a paper that tested different CNN architectures for predicting DNA-protein binding [22]. A partial reason for the low accuracy could

[21] Zeng, Haoyang et al. "Convolutional Neural Network Architectures for Predicting DNA–protein Binding." Bioinformatics 32, 12 (June 2016): i121–i127 © 2016 The Authors

[22] Zeng, Haoyang et al. "Convolutional Neural Network Architectures for Predicting DNA–protein Binding." Bioinformatics 32, 12 (June 2016): i121–i127 © 2016 The Authors

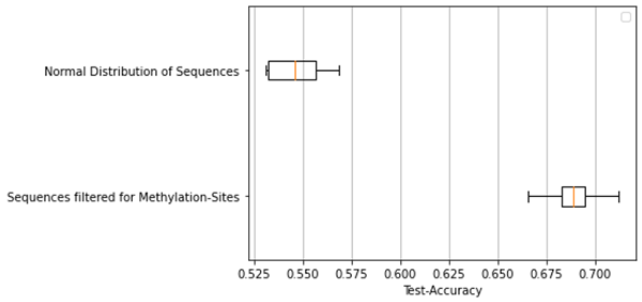

Figure 9. Accuracy scores for a dataset using a normal distribution of methylation and a filtered dataset using only methylated sequences.

due to the fact that some of our proteins scored relatively low in other papers' predictions 10 [23]. A major issue

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

RBP	AUROC	AP
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

Figure 10. Classification performance of DeepRiPe: AUROC as well as AP scores for all 59 PAR-CLIP datasets

was probably due to the size and the quality of the data. The amount of data that was available for some of the proteins was partially so small, that we removed them (EIF3A) from the analysis as we could not reasonably include them.

Additionally, a problem with most methylation data is the high natural imbalance as methylation overall rarely

[23] Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. 2020 Feb;30(2):214-226. doi: 10.1101/gr.247494.118. Epub 2020 Jan 28. PMID: 31992613; PMCID: PMC7050519.

occurs [24]. Therefore for our dataset, only five percent of all sequences contain matching m^{6}A methylation sites, which is high compared to other modifications, which often have much lower rates [25]. While this fact generally can't influence the performance of a model for the worse as the model can simply learn to ignore the methylation site feature, the positive effect is limited as the feature can only be applied to five percent of all data.
For this reason, we additionally tested the performance of our model on a filtered dataset containing only sequences with matching methylation site encodings against an unfiltered and normally distributed dataset.
This led us to our most significant finding from this paper, which is that our model indeed learns from the methylation sites and can therefore improve the accuracy materially 9 for predicting RBP binding sites. This finding can be underscored by the indication that the accuracy of our models was reduced by the cut-off encoding of the methylation rate.
This observation might be due to the fact that fewer methylation site encodings were available through the cut-off and therefore fewer features of our model detecting the methylation data were applicable to the test data. Also from all the methylation rate encodings, the methylation rates produced the best accuracy followed by the one-hot encoding. The methylation rate encodings with cut-off produced the worst performance which is probably due to the fact that we materially reduced the methylation rates. We originally intended to use a more classic threshold of 0.5 as is commonly used in the literature to split methylation rates, but due to the imbalance of methylation rates smaller than 0.5 and higher than 0.5 we decided to lower it to not exclude too much information.

4. Conclusions

Overall, our results seem to suggest that methylation plays a role in the binding of RBPs and we were able to evaluate different encodings and models of including it in a machine learning approach. Nevertheless, more experiments need to be done for further investigation and to perhaps confirm this hypothesis. Our models show an overall low accuracy, but also small differences between the encodings of the methylation sites.
While our model seemed not to be the main issue, we hope that the used architecture could perform considerably better with improved data and show an increase in the clearness of the results. This could for example be achieved by generally using more data.
An interesting approach might be also to combine the data sets of the different proteins, which would result in a huge increase in data size. While we would lose the negative samples from the binding sites of other proteins, negative samples can be ideally generated from the transcriptome, unlike positive samples. Combined it would be less likely to learn specific binding motifs of a protein and tend more to a general prediction of binding sites.
Altogether we could show that our models learned from the methylation site encodings for the prediction of RNA-binding proteins by filtering the datasets for sequences with methylation rate encodings which showed material improvements in accuracy. This indicates that the model learned features from the methylation rate encodings and utilized them for the predictions.

Data Availability Statement

The data that support the findings of this study are openly available and reproducible on GitHub at https://github.com/jonasgebele/Predicting-Interactions-of-m6A-regulated-RNA-Binding-Proteins.

References

[1] Hood, L., Galas, D. The digital code of DNA. Nature 421, 444–448 (2003). https://doi.org/10.1038/nature01410.

[2] https://www.illumina.com/techniques/sequencing.html

[3] www.bionumbers.org

[4] Szabó GT, Mahiny AJ, Vlatkovic I. COVID-19 mRNA vaccines: Platforms and current developments. Mol Ther. 2022;30(5):1850-1868. doi:10.1016/j.ymthe.2022.02.016

[5] Karikó K, Buckstein M, Ni H, Weissman D. Suppression of RNA recognition by Toll-like receptors: the impact of nucleoside modification and the evolutionary origin of RNA. Immunity.

[24] Stephanie L. Battle, R. David Hawkins, Chapter 2 - The epigenetics of pluripotent stem cells: Epigenome dynamics in vitro and in vivo mammalian embryonic stem cell systems, Editor(s): Eran Meshorer, Giuseppe Testa, In Translational Epigenetics, Stem Cell Epigenetics, Academic Press, Volume 17, 2020, Pages 25-74, ISBN 9780128140857, https://doi.org/10.1016/B978-0-12-814085-7.00002-7.

[25] Debnath TK, Xhemalçe B. Deciphering RNA modifications at base resolution: from chemistry to biology. Brief Funct Genomics. 2021 Mar 27;20(2):77-85. doi: 10.1093/bfgp/elaa024. PMID: 33454749; PMCID: PMC8008165.

2005;23(2):165-175. doi:10.1016/j.immuni.2005.06.008

[6] Zhao et al, POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Research 50, Issue D1, D287-D294 (2022).

[7] Oliveira C, Faoro H, Alves LR, Goldenberg S. RNA-binding proteins and their role in the regulation of gene expression in Trypanosoma cruzi and Saccharomyces cerevisiae. Genet Mol Biol. 2017;40(1):22-30. doi:10.1590/1678-4685-GMB-2016-0258

[8] Sun, L., Xu, K., Huang, W. et al. Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures. Cell Res 31, 495–516 (2021). https://doi.org/10.1038/s41422-021-00476-y

[9] Frye, M., Jaffrey, S., Pan, T. et al. RNA modifications: what have we learned and where are we headed?. Nat Rev Genet 17, 365–372 (2016). https://doi.org/10.1038/nrg.2016.47

[10] Jiang X, Liu B, Nie Z, Duan L, Xiong Q, Jin Z, Yang C, Chen Y. The role of m6A modification in the biological functions and diseases. Signal Transduct Target Ther. 2021 Feb 21;6(1):74. doi: 10.1038/s41392-020-00450-x. PMID: 33611339; PMCID: PMC7897327.

[11] Arguello et al, RNA Chemical Proteomics Reveals the N6-Methyladenosine (m6A)-Regulated Protein–RNA Interactome. Journal of the American Chemical Society 139, Issue 48, 17249-17252 (2017).

[12] Lewis et al, RNA modifications and structures cooperate to guide RNA-protein interactions. Nature Reviews Molecular Cell Biology 18, Issue 3, 202-210 (2017)

[13] Hafner et al, CLIP and complementary methods. Nature Reviews Methods Primers 1, 1, 1-23 (2021).

[14] https://www.creativebiomart.net/epigenetics/services/rna-protein-interaction-analysis/crosslinking-immunoprecipitation-clip-service/

[15] Huppertz I.; et al. iCLIP: protein-RNA interactions at nucleotide resolution. Methods. 2014, 65(3): 274-287.

[16] Neelanjan Mukherjee, Hans-Hermann Wessels, Svetlana Lebedeva, Marcin Sajek, Mahsa Ghanbari, Aitor Garzia, Alina Munteanu, Dilmurat Yusuf, Thalia Farazi, Jessica I Hoell, Kemal M Akat, Altuna Akalin, Thomas Tuschl, Uwe Ohler, Deciphering human ribonucleoprotein regulatory networks, Nucleic Acids Research, Volume 47, Issue 2, 25 January 2019, Pages 570–581, https://doi.org/10.1093/nar/gky1185

[17] Linder, B., Grozhik, A., Olarerin-George, A. et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. Nat Methods 12, 767–772 (2015). https://doi.org/10.1038/nmeth.3453

[18] https://github.com/jonasgebele/Predicting-Interactions-of-m6A-regulated-RNA-Binding-Proteins/blob/main/data/EncodingPreprocessing.py

[19] https://github.com/jonasgebele/Predicting-Interactions-of-m6A-regulated-RNA-Binding-Proteins/blob/main/README.md

[20] Zeng, Haoyang et al. “Convolutional Neural Network Architectures for Predicting DNA–protein Binding.” Bioinformatics 32, 12 (June 2016): i121–i127 © 2016 The Authors

[21] Ghanbari M, Ohler U. Deep neural networks for interpreting RNA-binding protein target preferences. Genome Res. 2020 Feb;30(2):214-226. doi: 10.1101/gr.247494.118. Epub 2020 Jan 28. PMID: 31992613; PMCID: PMC7050519.

[22] Stephanie L. Battle, R. David Hawkins, Chapter 2 - The epigenetics of pluripotent stem cells: Epigenome dynamics in vitro and in vivo mammalian embryonic stem cell systems, Editor(s): Eran Meshorer, Giuseppe Testa, In Translational Epigenetics, Stem Cell Epigenetics, Academic Press, Volume 17, 2020, Pages 25-74, ISBN 9780128140857, https://doi.org/10.1016/B978-0-12-814085-7.00002-7.

[23] Debnath TK, Xhemalçe B. Deciphering RNA modifications at base resolution: from chemistry to biology. Brief Funct Genomics. 2021 Mar 27;20(2):77-85. doi: 10.1093/bfgp/elaa024. PMID: 33454749; PMCID: PMC8008165.