DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Machine Learning for Regulatory Genomics

Report 01 - Multi-tissue mRNA half-life predictions

Author: Yasmine Zakaria Afify, Niklas Bühler, Markus Franke, Pauline Nickel
Supervisor: Prof. Dr. Julien Gagneur
Advisor: Pedro da Silva
Submission Date: 25.07.2022

Contents

1 Introduction

1.1 RNA Halflife

The regulation of genes is a wide field that still contains a lot of unidentified components. Promotor and transcription factors are well-known, but also the mRNA half-life plays an important role. Looking at a transient mRNA, it appears that it is needed short term or only in small amounts. At a long half-life, on the other hand, one can assume that the mRNA is used over a longer period of time or less specifically. In order to understand this topic better, there are two questions that need to be clarified. First, it is important to find out what influences the half-life of mRNA. Second, the exact effects of the mRNA half-life have to be determined. One fact was already discovered. The half-life of an mRNA changes depending on its tissue. That is essential in answering both questions. First of all, we have different conditions that lead to different half-lives. In addition, the effects of different half-lives can be partially seen in the properties of the tissues. However, cause and effect can hardly be separated. Also, the mRNA half-life is not the only gene regulation. First and foremost, the first question should be clarified — what influences the mRNA half-life.

1.2 Machine Learning Approach

In this paper, a first approach is shown as we tried to identify properties of the RNA or RNA binding proteins (RBPs) that have an impact on it. After some data preprocessing (see chapter 3), we developed a linear regression model (see chapter 4), several CNN models to predict half-life (see chapter 5) and based on those models we implemented models for multi-tissue prediction using transfer learning (see chapter 6).

2 Related Work – Saluki

During the preparations for our project, there was a paper that was of particular interest to us. Vikram Agarwal and David Kelley developed the Saluki model to predict half-life based on RNA sequence, coding frame and splice sites.

As a first step, they searched the literature for any RNA half-lives they could find, creating large datasets for the human genome and also for that of mice.

The half-life cannot be measured directly, but must be estimated using markers. This circumstance generally causes large measurement inaccuracies, systematic and methodical errors. These also reflected in the data from the Saluki team and first had to be filtered out using a Principal Component Analysis (PCA).

In a simple genetic model, they first examined the influence of various properties on the half-life and identified RNA length, GC content, exon junction density, codon frequencies and RBP binding sites as the most important factors.

Building on this, the team then developed a hybrid convolutional and recurrent deep neural network – Saluki. This should predict the half-life based on already known sequence data and with the help of the previously cleaned data set. For this purpose, the mouse data were additionally assigned to their respective homologous human sequences. Since it is reasonable to assume that gene-regulating regions tend to be more highly conserved, the mouse data should also provide additional information for the human sequences. In addition to the RNA sequence, the coding frame and the splice sites were also passed as input. A 10-fold cross validation was used for the training. The homologous mouse and human data were processed together. [1]

3 Data Preprocessing

Two data sets were made available to us for our model. First, we used the data prepared by Vikram Agarwal and David Kelley for the Saluki model [1] (see chapter 2). It contains the half-lives of 13.921 transcripts from 54 individual datasets. In addition, 11.363 transcripts from 49 different tissue types were provided to us by the Gagneurlab (based on the Genotype-Tissue Expression (GTEx) Project [2]). This data was also previously cleaned. In contrast to the Saluki dataset, however, this contains relative half-lives. In order to be able to use this data for our models, the coding frame and splice sites still had to be calculated (see chapter 3.1), the RBP factors determined (see chapter 3.2) and the information still had to be coded for CNN models (see chapter 3.3).

3.1 Sequence Features

The RNA sequences were given as 5′ UTR, ORF and 3′ UTR. In some cases the full sequence was needed, which would be a concatenation of those. The length of all sequences was calculated, as well as their logarithmic length and GC content.

In case of the Saluki dataset, an additional GTF file was given containing the coding regions and chromosomes of all transcripts. In order to get these information for the GTex dataset as well, we used the python interface of Ensembl database [3]. With help of this data we were able to calculate coding frame and splice sites for both datasets.

Finally, the datasets were merged with all missing sequence properties.

3.2 RNA Binding Protein Factors

Getting RNAs binding protein (RBPs) factors were one of the data pre-processing tasks for our Linear Mechanistic Models 4 approach. The aim was to get RBPs binding probabilities for each RNA in our dataset. There were different models found in the literature that can predict RBPs probabilities and DeepRiPE [4] was chosen for our task as it has the easiest setup. DeepRiPE predicts the binding probabilities for a list of the most common RBPs listed in figure 3.1. For each of [UTR3′, ORF and UTR5′] regions of the mRNA, predictions on each 50 nt window were generated, padding the sequence with its neighboring 50 nt upstream and downstream sequence (or Ns in the case of sequences at the boundary of a region) to generate a 150 nt input sequence, then getting the max over all predictions per RNA.

RBPs Names List				
DND1	CPSF7	CPSF6	CPSF1	CSTF2
CSTF2T	ZC3H7B	FMR1iso1	RBM10	MOV10
ELAVL1	TARDBP	ELAVL2	ELAVL3	ELAVL4
RBM20	IGF2BP1	IGF2BP2	IGF2BP3	EWSR1
HNRNPD	RBPMS	SRRM4	AGO2	NUDT21
FIP1L1	CAPRIN1	FMR1iso7	FXR2	AGO1
L1RE1	ORF1	MBNL1	P53_NONO	PUM2
QKI	AGO3	FUS	TAF15	ZFP36
DICER1	EIF3A	EIF3D	EIF3G	SSB
PAPD5	CPSF4	CPSF3	RTCB	FXR1
NOP58	NOP56	FBL	LIN28A	LIN28B
UPF1	G35	G45	XPO	-

Table 3.1: DeepRiPE List of RBPs

3.3 CNN-specific Preprocessing

After a general data preparation, the input of our CNN models was one hot encoded. Meaning that the sequence is divided into four channels with binary content. 1 indicates that the corresponding base is present at this position. The coding frame and the splice sites are also encoded in additional channels. The first base in the coding frame is encoded as 1 as well as the first base of the spliced exon. The UTRs are also identified by the coding frame because no codons can be found here.

Figure 3.1: One hot encoding of RNA data including coding frame (CF) and splice sites (SS).

4 Interpretable Mechanistic Models

The term mechanistic model describes a model that is based on fundamental laws of the natural sciences. A benefit of such models is that their variables have an actual meaning, allowing for easier interpretation.

In the domain of mRNA degradation, one fundamental law describes the way an mRNA interacts with its degradation factors. A degradation factor is an RNA-binding protein, also called RBP, that contributes to the degradation of a bound RNA molecule.

A simplified version of this law states that the half-life of an mRNA, considering only a single degradation factor, is inversely proportional to its binding probability with this degradation factor, multiplied by the concentration of both molecules in some medium. In this equation, the two concentrations measure how likely it is that the two molecules collide, while the binding probability defines the conditional probability of an actual interaction upon collision. There is a variety of degradation factors that could possibly interact with a single mRNA molecule and influence its lifespan, so this simple law has to be applied for every single degradation factor and the results have to be aggregated.

Of course, there are other influences on mRNA half-life as well, but in this chapter, we will focus on this simplification.

In the following, we will fit, evaluate and interpret mechanistic models that try to exploit this relationship between RBPs and mRNA half-life.

The desired result of this approach is not so much a model with particularly high prediction capabilities, but rather to gain some insights about biological phenomena by interpreting the models. With respect to the tissue-specific data we're working with, interpretation could lead to the discovery of varying concentrations of different RNA-binding proteins in different tissue types.

4.1 Intuition

The previously described relationship between an RBP and mRNA half-life can be derived as follows.

Let $\{\text{mRNA}_i\}$ be a set of different mRNA's and $\{\text{DF}_j\}$ a set of different degradation factors. The interaction of an mRNA_i with a degradation factor DF_j can then be described as

$$\text{mRNA}_i + \text{DF}_j \xrightarrow{r_{ij}} \emptyset_i + \text{DF}_j,$$

where the reaction rate r_{ij} is given as

$$r_{ij} = \kappa_{ij} \cdot C_{\text{mRNA}_i} \cdot C_{\text{DF}_j}.$$

Here, κ_{ij} is a factor proportional to the binding probability of mRNA_i and DF_j, i.e. a binding-score of those two molecules, C_{mRNA_i} is proportional to the concentration of the mRNA molecule mRNA_i and C_{DF_j} is proportional to the concentration of the degradation factor DF_j in the medium.

Leaving aside other influences for this simplification, the half-life of a fixed mRNA molecule mRNA_i is determined by all the r_{ij}'s in combination:

$$\text{mRNA}_i + \text{DF}_1 \xrightarrow{r_{i1}} \emptyset_i + \text{DF}_1 \quad \cdots \quad \text{mRNA}_i + \text{DF}_n \xrightarrow{r_{in}} \emptyset_i + \text{DF}_n.$$

Let $\varphi_{ij} = \kappa_{ij} \cdot C_{\text{DF}_j}$ and note that $r_{ij} = \varphi_{ij} \cdot C_{\text{mRNA}_i}$. Then, $\varphi_i = \sum_j \varphi_{ij}$ leads to the simplified chemical equation

$$\text{mRNA}_i \xrightarrow{\varphi_i} \emptyset_i,$$

where $\frac{1}{\varphi_i}$ is proportional to the half-life of mRNA_i.

This equation can now be set up for each individual mRNA_i in different tissue types and we can build models that incorporate this relationship into their structure.

After fitting these models to the available tissue-specific half-life data, we can then in turn extract the fitted coefficients to learn about the predicted concentration of different DF_j's, or more generally speaking of different RNA-binding proteins, across different tissue types.

4.1.1 Model Structure

As our fundamental modeling approach, we choose linear regression, as this type of modeling allows for easy interpretation by comparing the absolute values of the learned coefficients.

We define different model types by the range of features they're fitted on. For example, there will be a baseline model which is fitted on codon frequencies only, and there will be another model which will only incorporate RBP binding-scores as features, thus realizing the mechanistic approach outlined above. For each model type, there will be one linear regression model per type of tissue and we will evaluate several of these model bundles using cross-validation to test out different regularization hyperparameters. In the end, the models from the best-performing cross-validation fold will constitute the final bundle of models per model type.

In this section, we will outline and derive the model structure of the purely mechanistic model, which we will later define as Model 2.

Based on the previously derived equation, the general structure of this linear model will be the following:

$$\kappa_{i1} \cdot C_{\text{DF}_1} = \varphi_{i1} \quad \cdots \quad \kappa_{in} \cdot C_{\text{DF}_n} = \varphi_{in}.$$

Since we don't know the exact split of the summed values φ_i into the separate summands

φ_{ij}, we reformulate these equations into a single matrix equation. For this, note that

$$\varphi_i = \sum_j \varphi_{ij} = \sum_j \kappa_{ij} \cdot C_{\mathrm{DF}_j} \Rightarrow \begin{bmatrix} \kappa_{i1} & \dots & \kappa_{in} \end{bmatrix} \cdot \begin{bmatrix} C_{\mathrm{DF}_1} \\ \vdots \\ C_{\mathrm{DF}_n} \end{bmatrix} = \varphi_i,$$

which leads to

$$\begin{bmatrix} \kappa_{11} & \dots & \kappa_{1n} \\ & \vdots & \\ \kappa_{N1} & \dots & \kappa_{Nn} \end{bmatrix} \cdot \begin{bmatrix} C_{\mathrm{DF}_1} \\ \vdots \\ C_{\mathrm{DF}_n} \end{bmatrix} = \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_N \end{bmatrix}.$$

This matrix equation constitutes a linear regression model, where the κ_{ij} matrix contains the input data (in this case each row contains the binding-scores of one mRNA_i with every DF_j), the C_{DF_j} vector contains the learned coefficients and the φ_i vector constitutes the outcome variables.

The structure of this model will allow us to extract meaningful information, i.e. the learned C_{DF_j} coefficients, from a fitted model. These coefficients should predict a tissue-specific measure of concentration of the degradation factors DF_j.

4.2 Preparations

In order to fit the regression models for different tissue types, we rely on the data set containing relative measures of half-life per mRNA and tissue.

As mentioned above, different model types will be defined by the types of features they're being fit on. But a distinction between feature categories is important not only for fitting different models on different categories of features, but also with regard to interpreting the models coefficients later on. In order to correctly assess the influence of a single feature on the output of a model relative to the other features, one has to consider the magnitude of input values of different features. Therefore, we plot a separate heatmap per feature category to not lose sight of important features which happen to have inputs with lower magnitude.

The feature categories we are considering are: codon frequencies; RBP binding-scores as obtained from DeepRiPE (see [4]); CDS, 5'UTR and 3'UTR sequence length in the log scale and CDS, 5'UTR and 3'UTR GC-content.

4.2.1 Outcome Variables

The outcome variables φ_i are a measure of relative half-life per tissue, compared to half-life in all tissues. There is one such outcome variable for every mRNA_i in every tissue type.

4.2.2 Cross-Validation on Chromosomes

In order to evaluate our models for different regularization hyperparameters, we define cross-validation folds.

It is important to note that genes which lie on the same chromosome are often more similar than genes from different chromosomes. In order to avoid artificially increasing our models prediction accuracy on the test set, we don't split the data into training and test sets randomly, but instead adhere to the policy of only distributing the complete data from a chromosome to either set.

Folds are then chosen to result in a train/test split of 80/20, including a tolerance of 2 percentage points in both directions, in order to make the chromosome distribution policy work. The test sets of all folds are also mutually exclusive.

4.2.3 Ridge Regularization

In some of our models, we're supplying codon frequencies as features. This poses a problem, as codon frequencies are in general highly correlated and thus can lead to model coefficients that are poorly determined and exhibit high variance. This problem can be alleviated by applying ridge regularization, as this regularization technique imposes a penalty on coefficient size (see also [5]). The strength of this regularization is determined by a hyperparameter which we call α.

The tested hyperparameters across all models are $\alpha = 0$, 10^{-5}, $5 \cdot 10^{-5}$, 10^{-4}, $5 \cdot 10^{-4}$, 0.001, 0.005, 0.01 and 0.05.

4.3 Model Types

All model types are fitted with relative measures of half-life as outcome variables and only differ in what they take as input.

The baseline model, called "Model 1" solely considers codon frequencies as features and "Model 2" considers solely RBP binding-scores. These two feature categories are combined in "Model 3", which is fit on codon frequencies as well as RBP binding-scores. In the final model, "Model 4", the sequence lengths of the CDS, 5'UTR and 3'UTR, as well as their GC-content, are supplied as additional features.

Table 4.1 reports on the best performing models of each type. The aggregated R^2 values were calculated across all different tissue types. A more detailed visualization of achieved R^2 values per tissue type can be found in the appendix.

The model types consistently improved by providing more features. Figure 4.1 compares the performance of the final model with the performance achieved in the baseline model. Similar comparisons for Model 2 and 3 can be found in the appendix as well.

4.3.1 Model Interpretation

The goal of interpreting the models is to extract some information about the relative concentration of various RBPs in different types of tissue.

By plotting the coefficients of different features (extracted from the tissue-specific models in the best-performing fold) for every tissue type, we can gain insight into the influence of

Model	Feature Categories	α	Mean R^2	Max R^2
Model 1	C	$5 \cdot 10^{-5}$	0.032555	0.110555
Model 2	R	0.005	0.034458	0.123484
Model 3	C, R	$5 \cdot 10^{-5}$	0.048725	0.141194
Model 4	C, R, E	0.0001	0.055069	0.149735

Table 4.1: Results of fitting the different model types. The feature categories are defined as *C* for codon frequencies, *R* for RBP binding-scores and *E* for extra features.

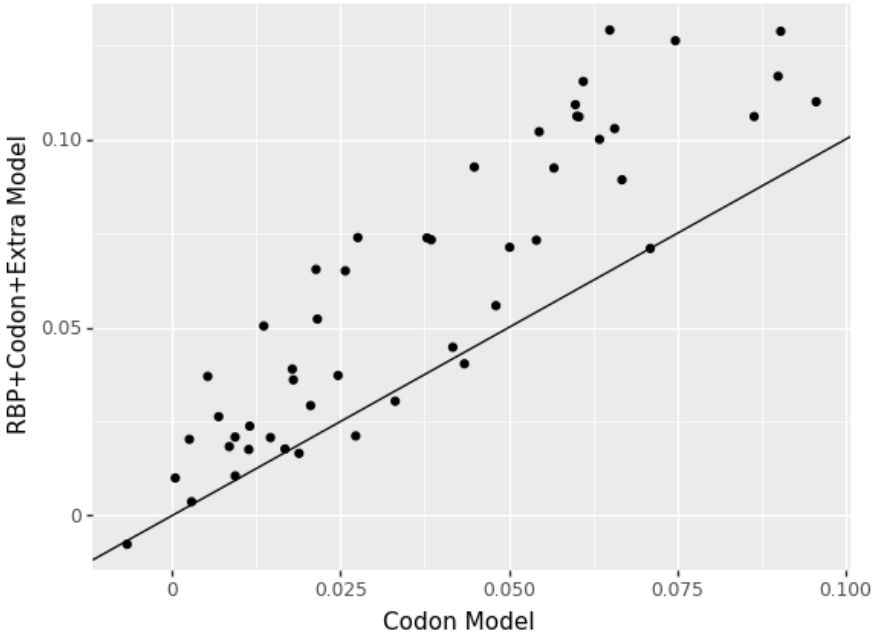

Figure 4.1: Comparing the achieved R^2 scores per tissue type in the final model with the respective scores in the baseline model.

these features on mRNA half-life in different tissues. According to the mechanistic model we introduced, especially the learned parameters of RBP binding-scores are of interest, as they might correlate with a mixture of RBP concentration and influence of the specific RBP on mRNA half-life per tissue. By examining the coefficients of these features with the highest absolute value, and especially comparing their influence across tissues, we can try to predict the concentration and influence of the RBPs on mRNA half-life in different tissues.

We visualize the models coefficients in several heatmaps, which can be found in the appendix. The heatmap of Model 2, containing only the RBP binding-scores as features, is shown in figure 4.2. Every horizontal line which is either of bright or dark color highlights a consistent positive or negative influence of a certain feature across tissue types.

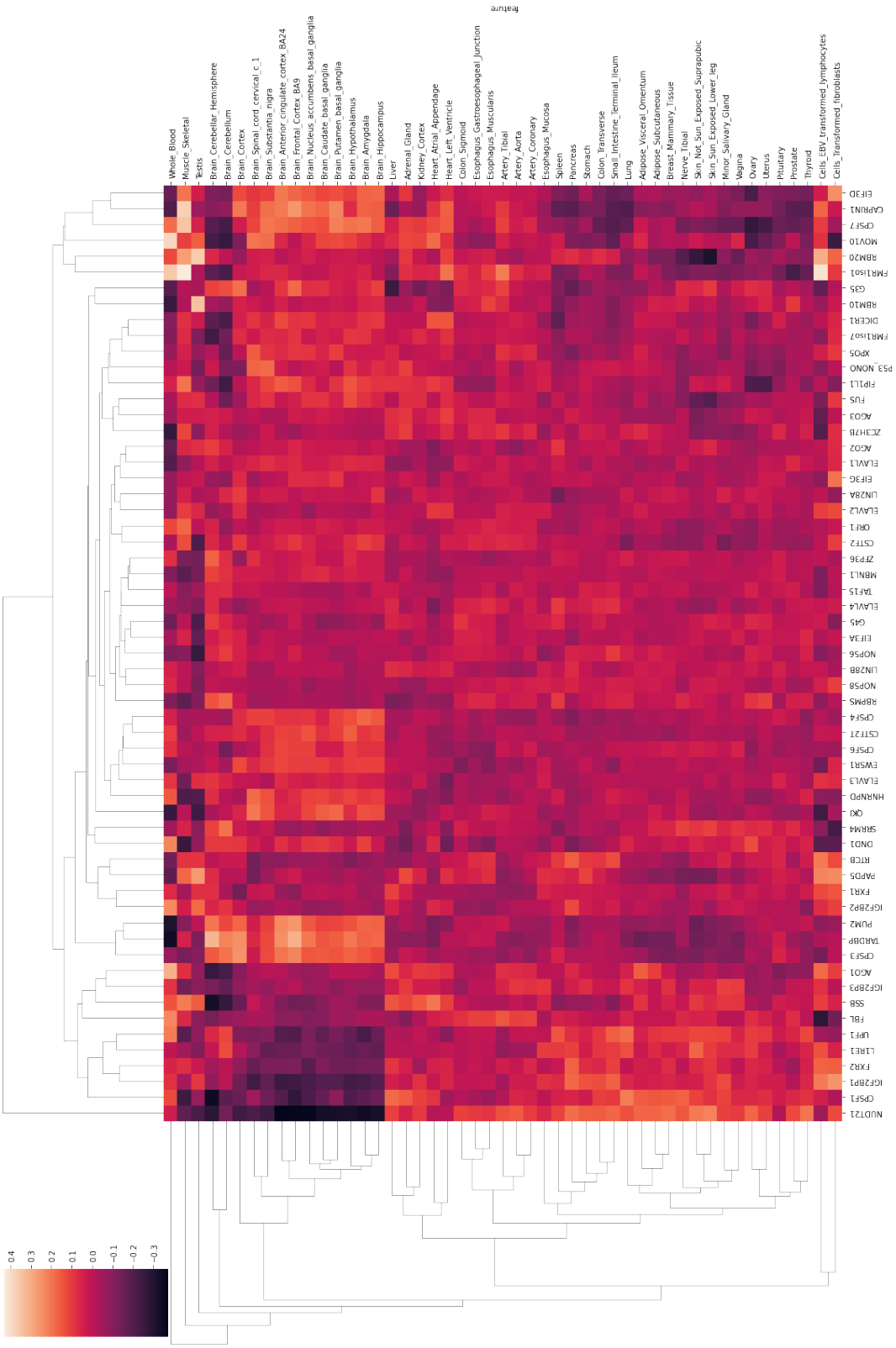

Figure 4.2: Clustered heatmap visualizing coefficients of RBP binding-scores in Model 2.

4.4 Conclusion

Based on the plotted heatmaps, we made several predictions of which RBPs could be more prevalent in specific tissue types than in others. By browsing the *TISSUES Tissue Expression Database*[1], we can now check these predictions against the biological literature and experimental data. The TISSUES database offers confidence scores ranging from 1 to 5 for the expression of genes in different tissue types. It employs different sources for tissue associations of genes, namely *Knowledge*, which is manually curated knowledge from UniProtKB (see [7]), *Experiments* and *Text mining*. In the process of validating our predictions, only the *Knowledge* and *Experiments* scores were considered and listed. The summary of this validation is listed in table 4.2.

RBP	Tissue	Score (K)	Score (E)
NUDT21	Brain	4/5	5/5
FXR2	Brain	4/5	5/5
CPSF3	Brain	4/5	5/5
EWSR1	Brain	4/5	5/5
MOV10	Brain	4/5	4/5
FMR1iso1	Brain	4/5	3/5
TARDBP	Brain	4/5	3/5
ELAVL2	Brain	4/5	2/5
CAPRIN1	Brain	–	5/5
CPSF6	Brain	–	4/5
CPSF1	Brain	–	3/5
IGF2BP1	Brain	–	–
CAPRIN1	Liver	4/5	3/5
CPSF6	Heart	4/5	2/5
CPSF7	Skin	–	2/5
RBM20	Skin	–	–
ELAVL4	Skin	–	–

Table 4.2: Table of all predicted RBP concentrations with their respectively predicted tissue types. *Score (K)* represents the *TISSUES* confidence score in the category *Knowledge* and *Score (E)* represents the experimental confidence score.

Most predictions of increased concentration of RBPs in specific tissue types are supported by the literature. These results lead to the conclusion that by expanding the applied technique on other RBPs and by obtaining more predictions from the heatmaps or via other approaches, e.g. extracting predictions using absolute coefficient values per tissue or considering p-values, might result in the discovery of previously unknown expression patterns of RBPs in different tissue types.

[1] https://tissues.jensenlab.org/About, see also [6].

5 Deep Learning Approaches to Predict Half-life

Given that the tissue-specific data was directly measured from RNA-seq data as the ratio of split to unsplit reads, we decided that a transfer learning approach on general half-life would be valuable by leveraging data that is less noisy due to being measured by transcriptional inhibitors or pulse-chase based methods.
V. Agarwal and D. Kelley performed a meta-analysis of studies compiling half-life values obtained with one of these two methods and found that the values were clustered closely by laboratory of origin.[1] Thus they "standardized the samples in each matrix, used iterative PCA to impute missing gene measurements, and performed quantile-normalization to align the samples into similar distributions"[1] in order to reduce the bias induced by the different laboratories.
Whereas their Deep Learning model uses both mouse and human data we concentrated on a human RNA only approach, and used their corrected values for RNA half-life on 13230 human Genes.

5.1 Benchmarking pure CNN-based models

To benchmark models on sequences with CNNs, we prepared the sequences separated into 5'UTR, ORF and 3'UTR, and padded as well as one-hot encoded these separately. Similarily to the data processing in the Saluki model we added a track for exon junctions and a track to mark the beginning of each codon in the ORF.
The approach to split the sequence thus was chosen in order to exploit the different structures and utilisation between those blocks. Additionally providing the UTRs separately brings the benefit of not having to add zero-valued tracks for the codon starts to those, reducing dimensionality and improving the stability of the normalization.
These blocks were later concatenated into a big Dense Layer, which output was fed into one medium sized additional hidden layer and then into the final prediction layer.
The models we tried usually achieved an R^2 between 0.2 and 0.3, and attempts to add additional basic features such as codon frequencies either only improved those models negligibly or didn't increase validation error at all, suggesting that most of the models were able to pick up most of the basic sequence based features.

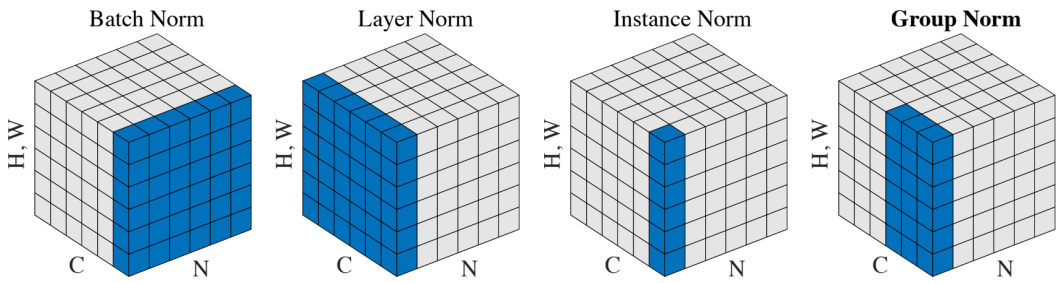

Figure 5.1: Normalization methods, N is the batch axis, C the channels and (H,W) the spacial axis. Source: Yuxin Wu, Kaiming He: Group Normalization[8]

5.2 Benchmarking hybrid CNN and RNN models

While we kept 5'UTR, ORF and 3'UTR separate for pure CNNs, we decided to feed the model the whole sequence for hybrid CNN and RNN models to exploit sequential characteristics, using the same data processing approach as the Saluki model (see Chapter 3)[1].
Firstly we benchmarked a model almost identical to Saluki, with the key differences that we split training, validation and test set by chromosome for better generalization and only used human data, which reproduced an R^2 of 0.39 on the validation set.
Tuning most hyper-parameters had little to adverse effect, and we found that the hyper-parameter with the greatest effect was the choice of normalization. In Saluki Kelley et al "chose Layer Normalization over Batch Normalization because most of the 3' positions are zero padded and would confuse the batch statistics."[1]
Yet this does not come without cost, as Layer Normalization artificially increases the weights in short sequences, due to the higher prevalence of zero values in the padded sequence.
As a matter of fact we found that both our reproduction of Saluki and our final model architecture both performed better with Batch Normalization, achieving an R^2 of 0.42 on the validation set for our Saluki reproduction with Batch Normalization.
It is worth noting that Yuxin Wu and Kaiming He have experimentally demonstrated in a recent paper that Group Normalization usually outperforms Layer Normalization [8], and we theorize that using Group Normalization could improve this model further.

5.3 Final Model Architecture

One of the most more famous network architectures in Deep Learning history is GoogleLeNet.[9] It is known for it's Inception Layers based on the intuition of multi-scale processing. Our best performing model (on the validation set) is based on the same intuition, adapted to our needs by adding MaxPooling Layers to reduce dimensionality (see Figure 5.2). We used early stopping with a patience of 25 and restored the best weights based on the mean squared validation error, furthermore we use a dropout parameter of 0.33 throughout and

we found GRU to be faster to train with similar performance in comparison to LSTM which coincides with the findings of the paper on GRUs by Chung et al.[10]
Testing the the performance of this model on the test set provided an explained variance and R^2 of 0.47.

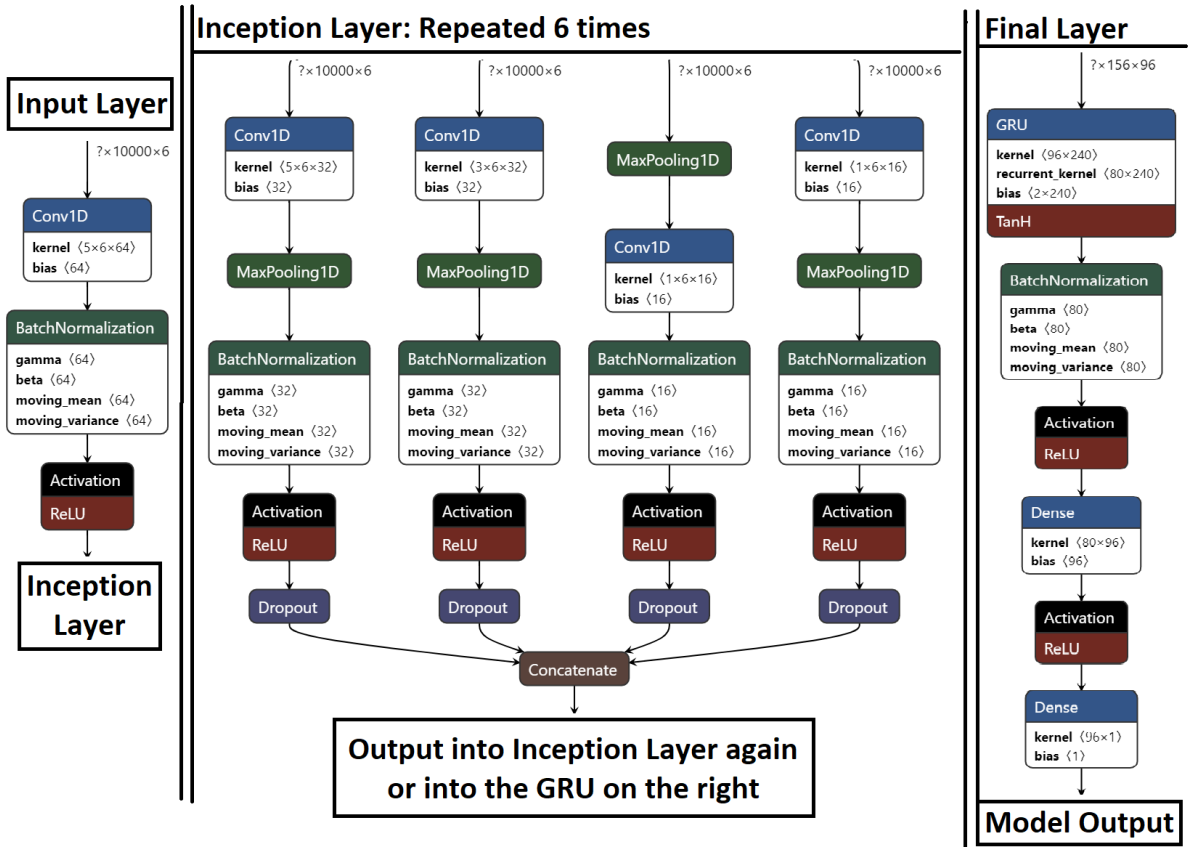

Figure 5.2: Model Architecture of the hybrid CNN-Inception and RNN model, figure produced with Netron[11]

5.4 Checking the influence of known motifs related to general RNA half-life

The 3'UTR motif UAASUUAU is known to destabilize mRNAs.[12]
Following up on the results of that paper about this motif, we selected 20 sequences for both 3'UTR and 5'UTR respectively, such that the lengths of the UTRs we were inserting the motif into were between the median and mean length calculated accross the sequences for each UTR respectively.
Then we created 150 variations of each sequence where we inserted the motif into randomly selected positions (without duplicates, in the UTR which we choose for its average length),

and predicted the half-life scores for these altered sequences using the hybrid CNN and RNN model we developed in this chapter.
The following plot shows the shift in prediction, which we attain by subtracting the original half-life prediction of the unaltered sequence from the new predictions for each of the 20 sequences and UTR's respectively. We used different colors for the sequences to visualize the changes in prediction for each sequence and the influence of the position where we inserted the motif UAASUUAU into the UTRs.

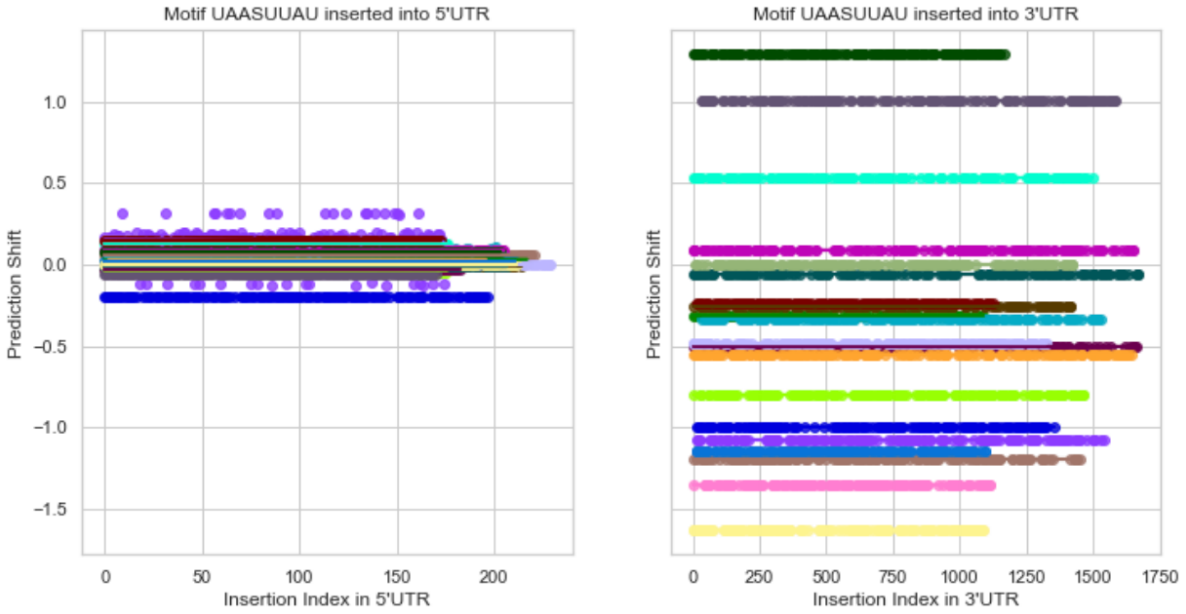

Figure 5.3: caption

As we can see in Figure 5.3 inserting the motif into the 5'UTR did not overly change the prediction, with shifts ranging from -0.2 to 0.3, while inserting the motif into the 3'UTR resulted in bigger shifts ranging from -1.6 to 1.3 in standard deviation of the log-transformed half-life prediction.
All of the 20 tested 3'UTR's were shifted independently of insertion position, and 15 out of the 20 had their position shifted to shorter half-life values.
This indicates that the model does recognize that the 3'UTR motif UAASUUAU, which is known to destabilize RNAs[12], has an effect on half-life when present in the 3'UTR.

6 Multi-tissue prediction of mRNA half-life from sequence

Our approach here was multi-task learning. We addressed the problem as the following, using some sequence features of the mRNA, we trained a model that can predict a vector of values. Each value in our output vector is a predicted half-life of one of the tissues. After reaching a good R2 score in our approach of predicting half-life in section 5, we thought that using that model for transfer learning is a worth trying approach. Thus, we extracted the same sequence features from our multi-tisssue dataset so that our input to the model is a 6-track of one-hot encoding.

6.1 Data Pre-processing

NaNs where handled by keeping the mRNAs with annotated 3′ and 5′ UTRs only. Same 6-track one-hot encoding computation implemented in section 3.1 was done on the dataset. Dataset was splitted into test, validation and training based on chromosomes. Namely ['chr2', 'chr3', 'chr4'] were used for our validation set, ['chr1', 'chr8', 'chr9'] were used for our test set and the rest were used for our training set. Tissues with nans half-life values were masked by -1000 value in order to handle them in our loss function.

6.2 Approaches

Our base model architectures here were the implemented ones in section 5 which are RNN-CNN with Batch Normalization and RNN-CNN with Layer Normalization models, we tried out training them from scratch on our multi-tissue dataset and tried their pre-trained weights for transfer learning. Explained-variance score per tissue was our watched metric to compare between different models. We dropped the last 2 layers (dense and activation) and added 2 dense (units = 128) layers and 1 activation layer of type GELU [13] to adapt the model on our dataset.

6.2.1 Transfer Learning

Pre-trained weights were used as starting weights for our fine-tuning. Different layers combinations of the transferred model were frozen and a re-training experiment was done to check which would perform better. A low learning rate of 0.0001 was used. The best performance was gotten by unfreezing all the layers and training them 6.1. This supported

the idea of trying to train the models from scratch without transfer learning (next section 6.2.2)

6.2.2 Training from Scratch

RNN-CNN models showed a good/promising training performance on our dataset when all layers were not frozen (trainable). That's why we tried training them from scratch on our multi-tissue dataset. To compare between the 2 models, minimum, maximum and mean explained variance across all tissues were calculated (figure 6.1)

	Batch Normalization Model	Layer Normalization model
Minimum	0.08	0.04
Maximum	0.17	0.11
Mean	0.12	0.08

Table 6.1: Explained variance scores stats -training from scratch-

6.3 Model Interpretation

We used the insight concluded in chapter 4 that RBP CPSF7 is prevalent in skin-type tissues with a high concentration to test how the half-life will get affected by appending a motif that CPSF7 RBP bind with into some RNAs sequences. The motif chosen was 'UGUA' [7]. Sequences were computed the same way as mentioned in section 5.4. The model used for prediction was the output model of the transfer learning approach 6.2.1. As shown in figure 6.1, some sequences got about double the half-life when adding the motif into some specific places in the sequence.

6.4 Conclusion

RNN-CNN with Batch Normalization showed the best performance across all tissues in both trials (training from scratch and transfer learning). Transfer learning and training from scratch showed almost the same performance across tissues (figure 6.2). Finally, tissues with top 4 explained variance score were almost the same between both models which are listed in table 6.3.

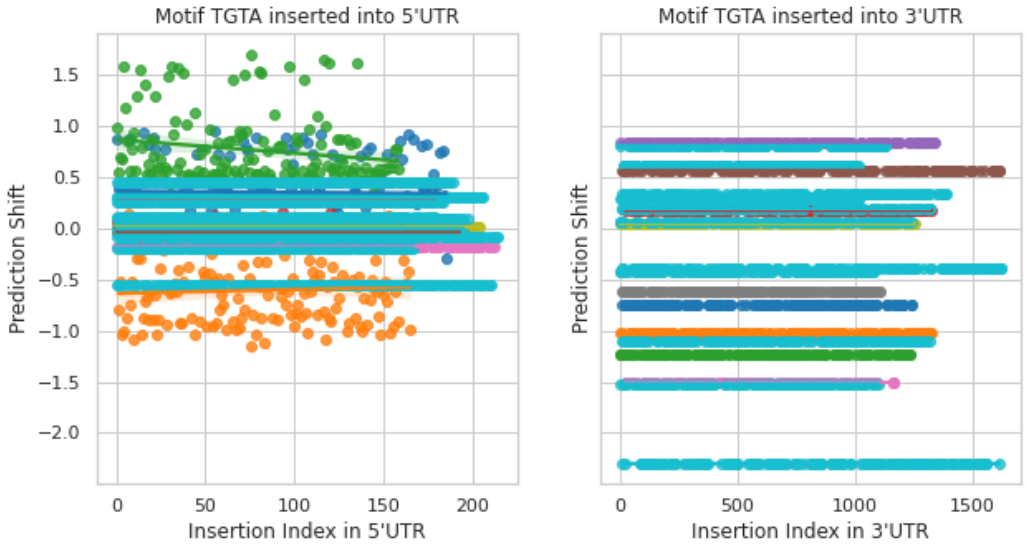

Figure 6.1: Explained variance scores stats -training from scratch-

	Training from scratch Model	Transfer learning model
Minimum	0.08	0.09
Maximum	0.17	0.16
Mean	0.12	0.12
Median	0.12	0.12

Table 6.2: Explained variance score stats comparison between transfer learning and scratch models

Tissue Name
Skin_Sun_Exposed_Lower_leg
Skin_Not_Sun_Exposed_Suprapubic
Minor_Salivary_Gland
Vagina

Table 6.3: Tissues with top 5 explained variance

7 Appendix

7.1 Chapter 4: Interpretable Mechanistic Models

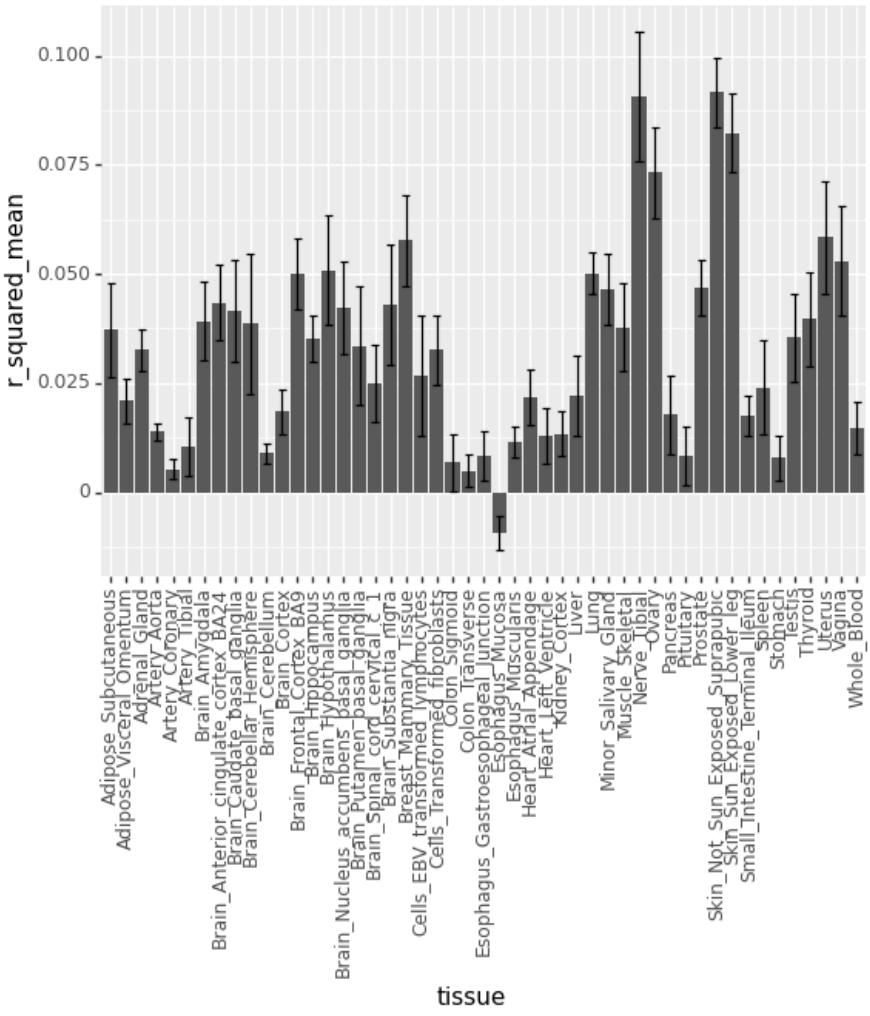

Figure 7.1: Achieved R^2 values on test set for different tissues in Model 1.

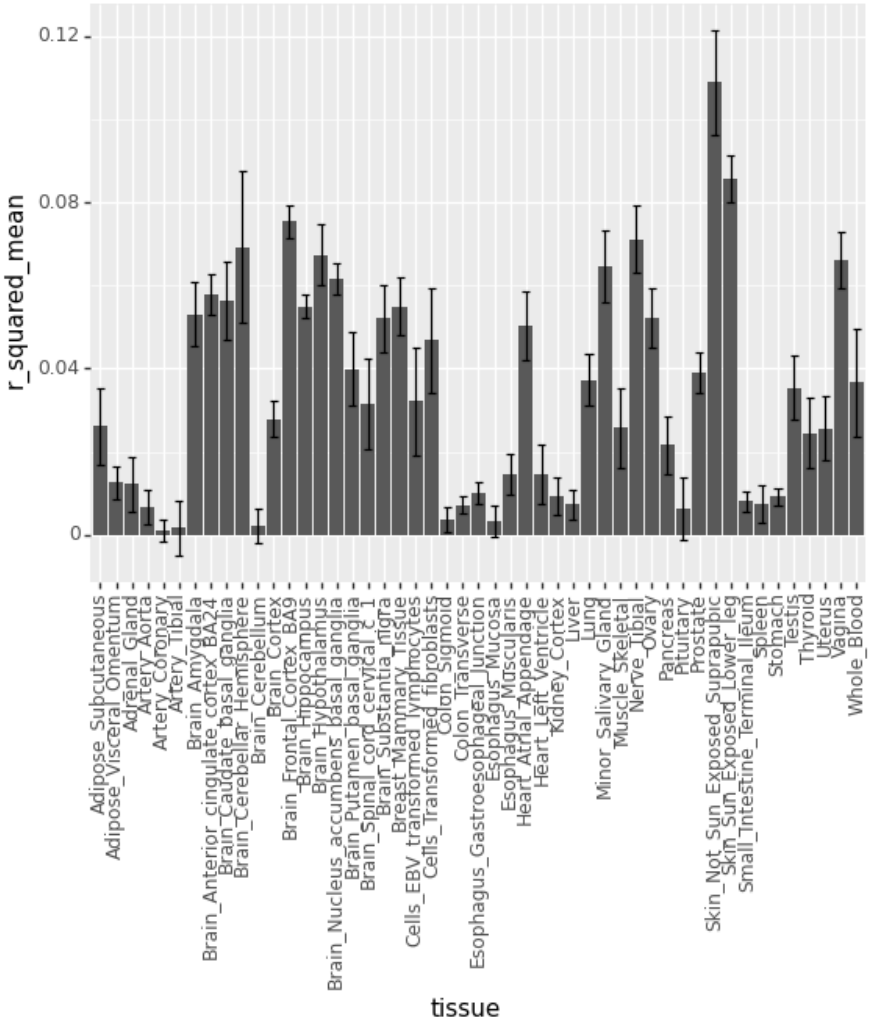

Figure 7.2: Achieved R^2 values on test set for different tissues in Model 2.

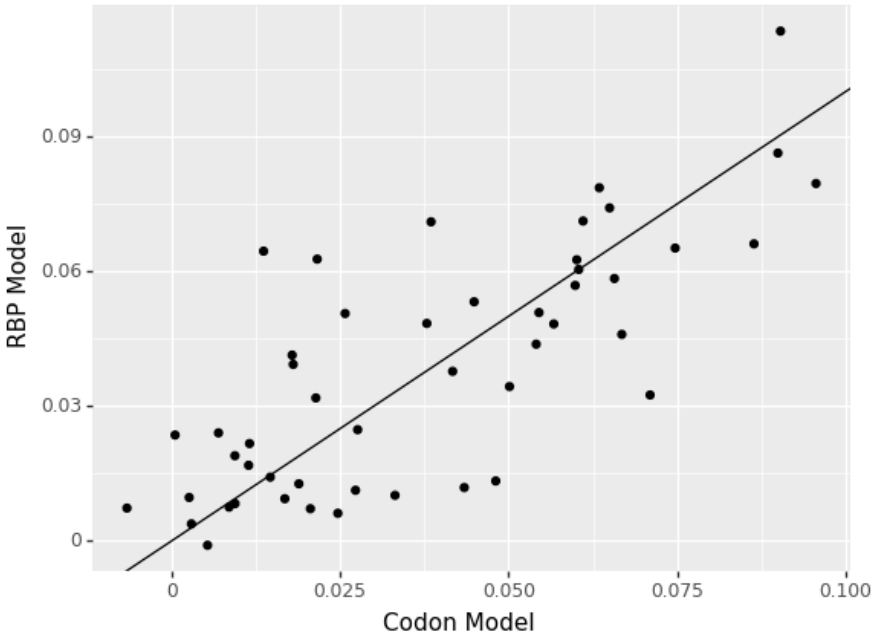

Figure 7.3: Comparing the achieved R^2 scores per tissue type in Model 2 with the respective scores in the baseline model.

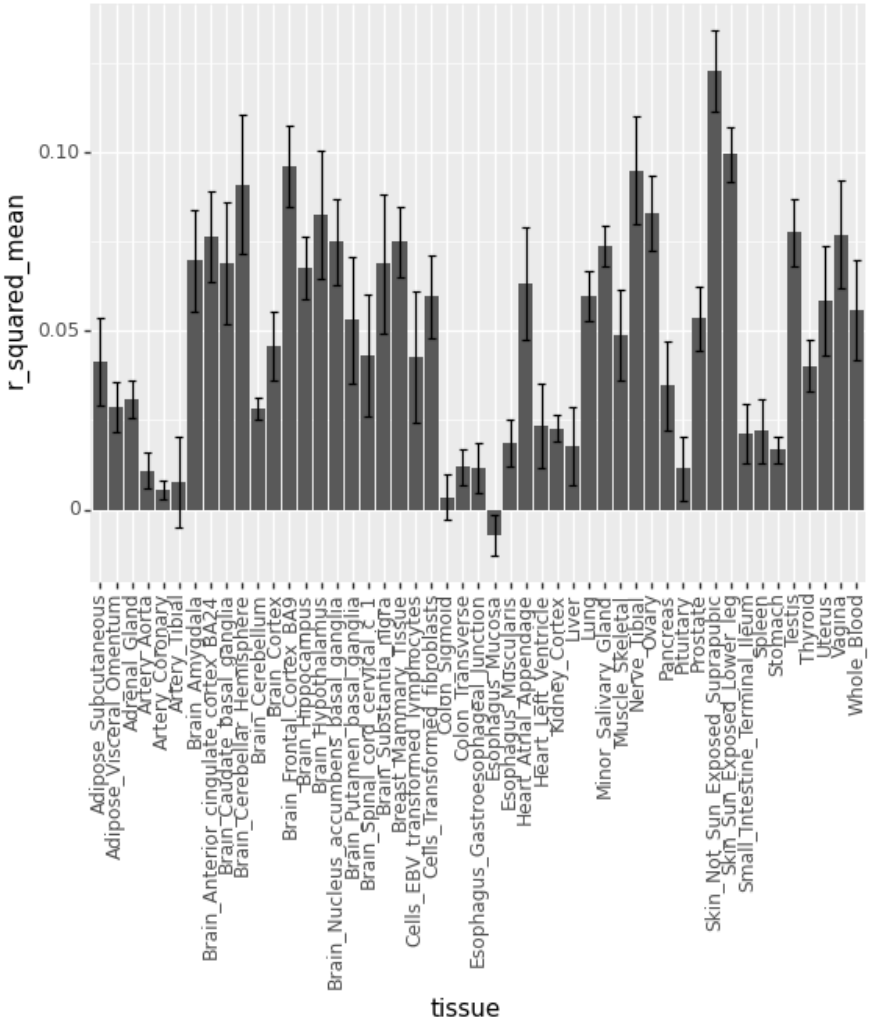

Figure 7.4: Achieved R^2 values on test set for different tissues in Model 3.

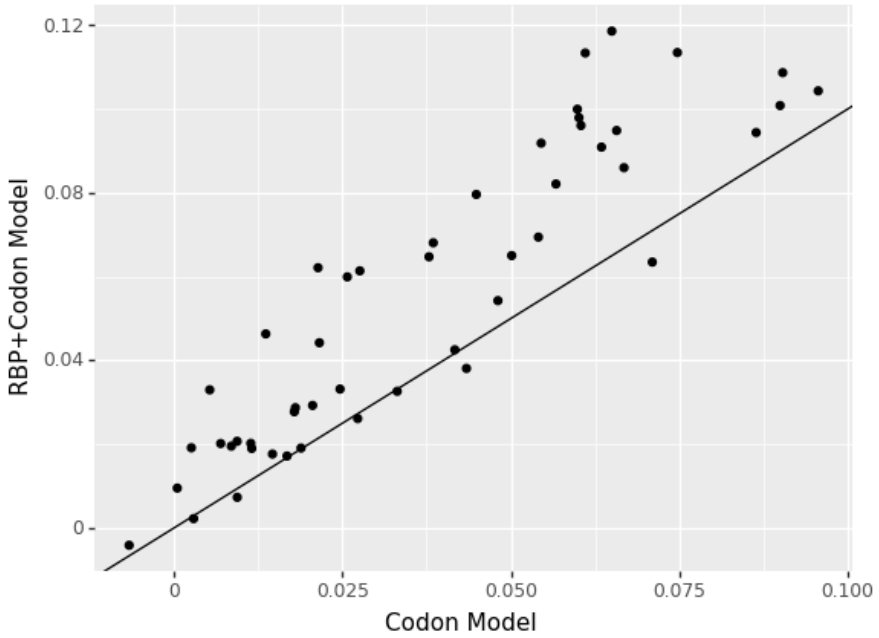

Figure 7.5: Comparing the achieved R^2 scores per tissue type in Model 3 with the respective scores in the baseline model.

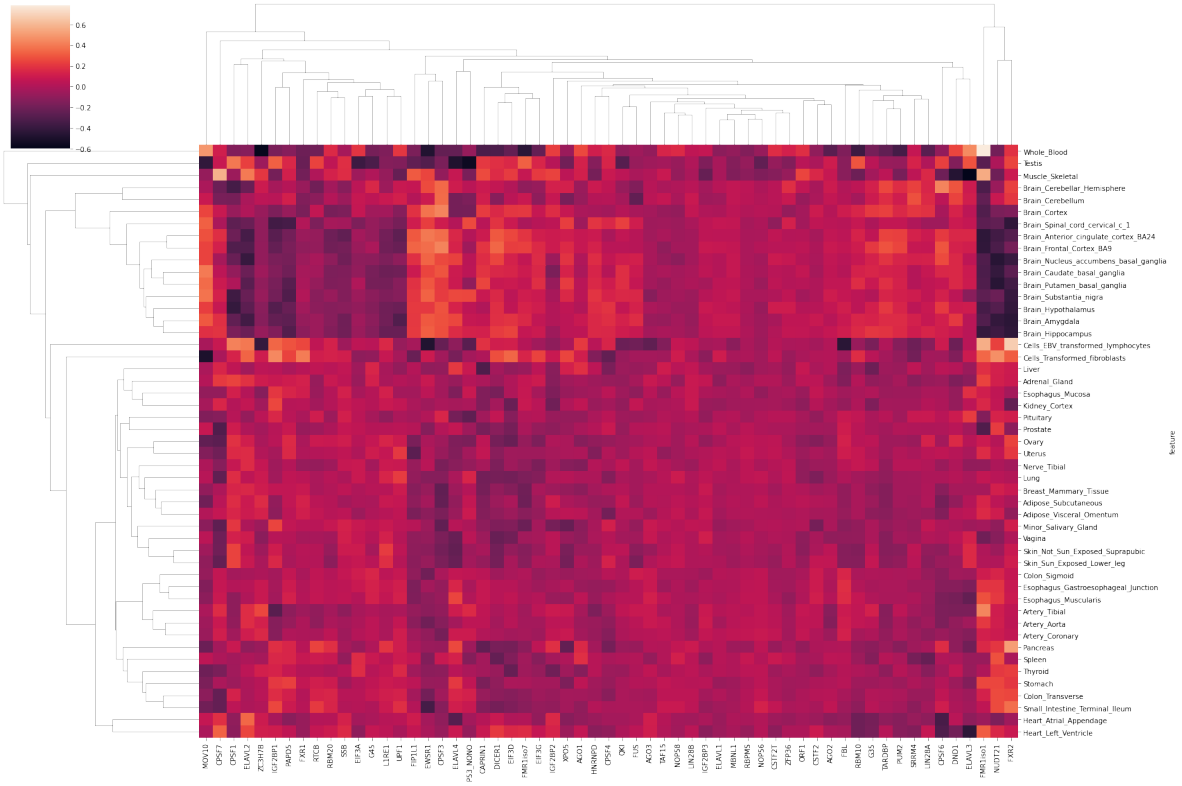

Figure 7.6: Clustered heatmap visualizing coefficients of RBP binding-scores in Model 3.

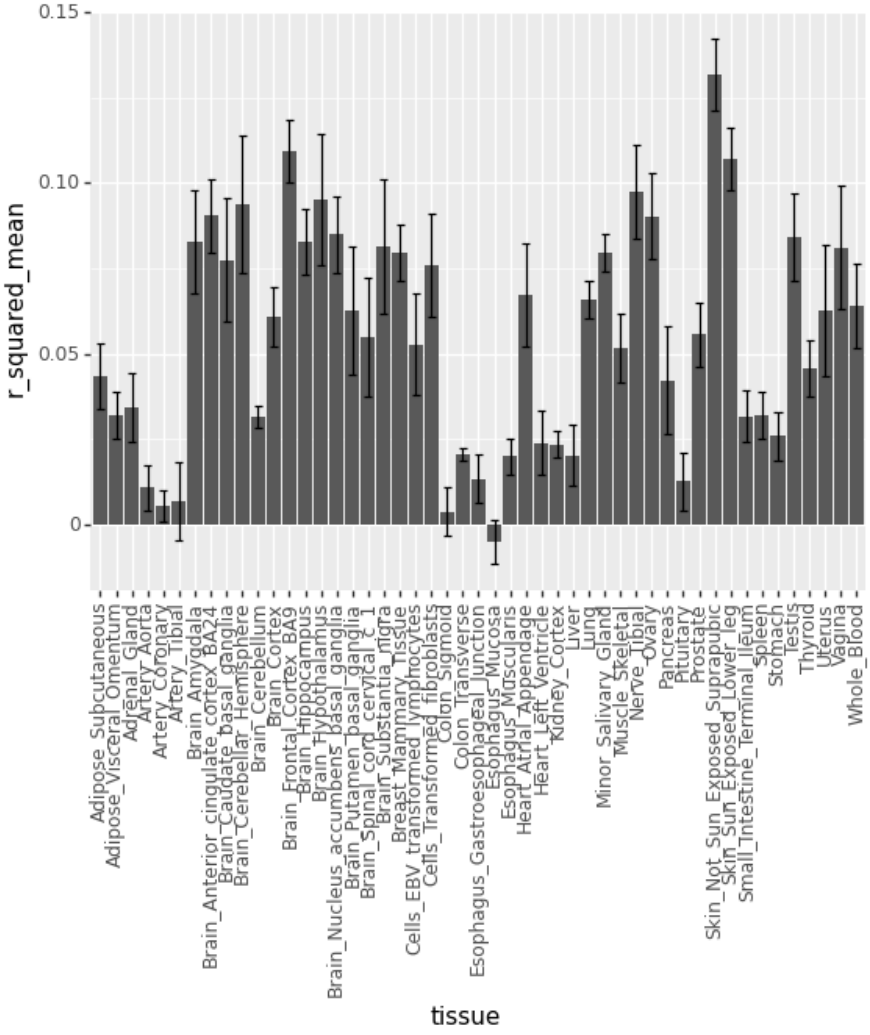

Figure 7.7: Achieved R^2 values on test set for different tissues in Model 4.

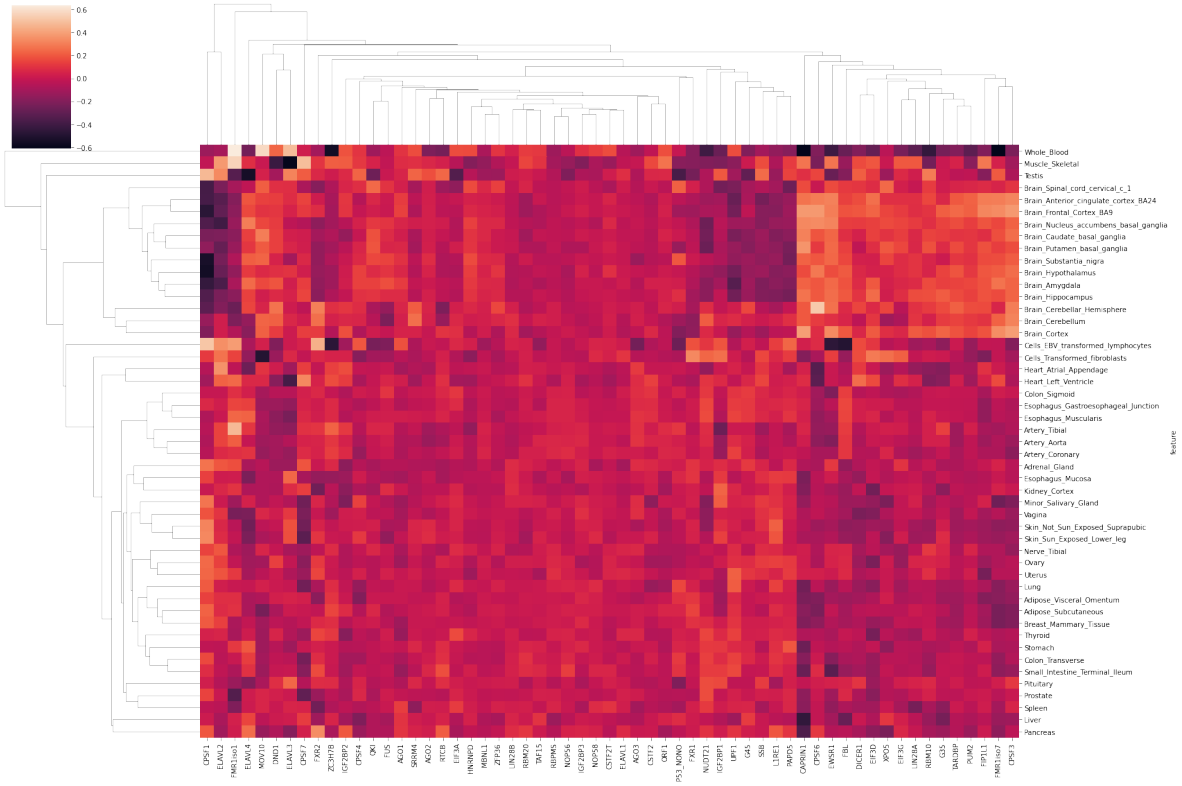

Figure 7.8: Clustered heatmap visualizing coefficients of RBP binding-scores in Model 4.

List of Figures

List of Tables

Bibliography

[1] V. Agarwal and D. Kelley. "The genetic and biochemical determinants of mRNA degradation rates in mammals". In: *bioRxiv* (2022). DOI: 10.1101/2022.03.18.484474. eprint: https://www.biorxiv.org/content/early/2022/03/19/2022.03.18.484474.full.pdf. URL: https://www.biorxiv.org/content/early/2022/03/19/2022.03.18.484474.

[2] *Genotype-Tissue Expression (GTEx) Project*. 2021. URL: https://gtexportal.org/home/datasets (visited on 07/25/2022).

[3] *Ensembl*. 2022. URL: https://www.ensembl.org/index.html (visited on 07/25/2022).

[4] M. Ghanbari and U. Ohler. "Deep neural networks for interpreting RNA-binding protein target preferences". In: *Genome research* 30.2 (2020), pp. 214–226.

[5] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[6] O. Palasca, A. Santos, C. Stolte, J. Gorodkin, and L. J. Jensen. "TISSUES 2.0: an integrative web resource on mammalian tissue expression". In: *Database* 2018 (2018).

[7] T. U. Consortium. "UniProt: the universal protein knowledgebase in 2021". In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. D480–D489. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1100. eprint: https://academic.oup.com/nar/article-pdf/49/D1/D480/35364103/gkaa1100.pdf. URL: https://doi.org/10.1093/nar/gkaa1100.

[8] Y. Wu and K. He. *Group Normalization*. 2018. arXiv: 1803.08494. URL: http://arxiv.org/abs/1803.08494.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions". In: *CoRR* abs/1409.4842 (2014). arXiv: 1409.4842. URL: http://arxiv.org/abs/1409.4842.

[10] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *CoRR* abs/1412.3555 (2014). arXiv: 1412.3555. URL: http://arxiv.org/abs/1412.3555.

[11] R. Lutz. *Netron, Visualizer for neural network, deep learning, and machine learning models*. Dec. 2017. URL: https://github.com/lutzroeder/netron.

[12] R. Geissler, A. Simkin, D. Floss, R. Patel, E. Fogarty, J. Scheller, and A. Grimson. "A widespread sequence-specific mRNA decay pathway mediated by hnRNPs A1 and A2/B1". In: *Genes Development* 30 (May 2016), pp. 1070–1085. DOI: 10.1101/gad.277392.116.

[13] D. Hendrycks and K. Gimpel. *Gaussian Error Linear Units (GELUs)*. 2016. DOI: 10.48550/ARXIV.1606.08415. URL: https://arxiv.org/abs/1606.08415.