DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Machine Learning for Regulatory Genomics

Interpretable Mechanistic Models for Predicting Tissue-specific RBP Expression

Author:	Niklas Bühler
Supervisor:	Prof. Dr. Julien Gagneur
Advisor:	Pedro da Silva
Submission Date:	25.07.2022

Contents

1 Introduction

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.

2 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.

2.1 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:

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

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.

3 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 [1]); CDS, 5'UTR and 3'UTR sequence length in the log scale and CDS, 5'UTR and 3'UTR GC-content.

3.1 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.

3.2 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.

3.3 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 [2]). 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 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.

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.

4.1 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 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.

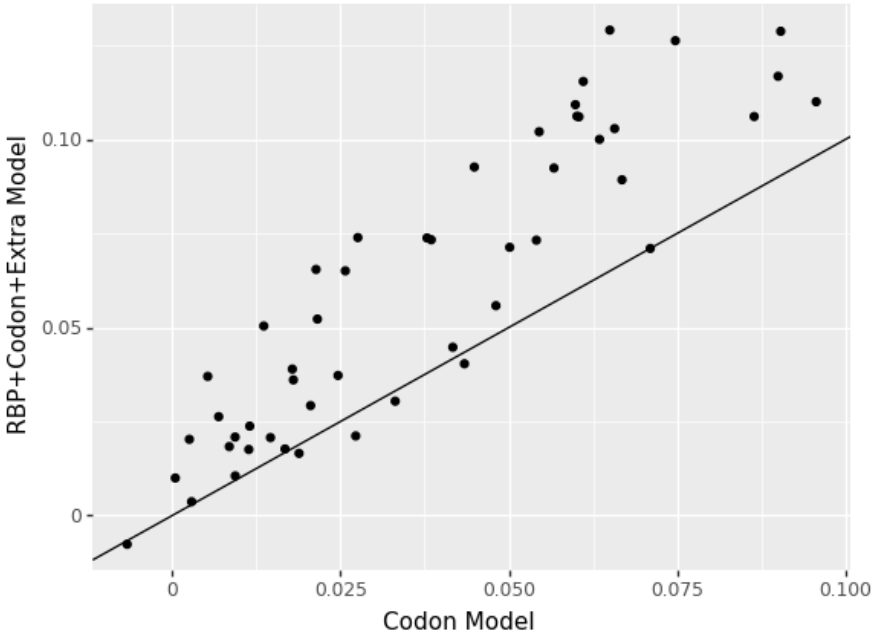

Figure 4.1: Comparing the achieved R^2 scores per tissue type in the final model with the respective scores in the baseline model.

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.

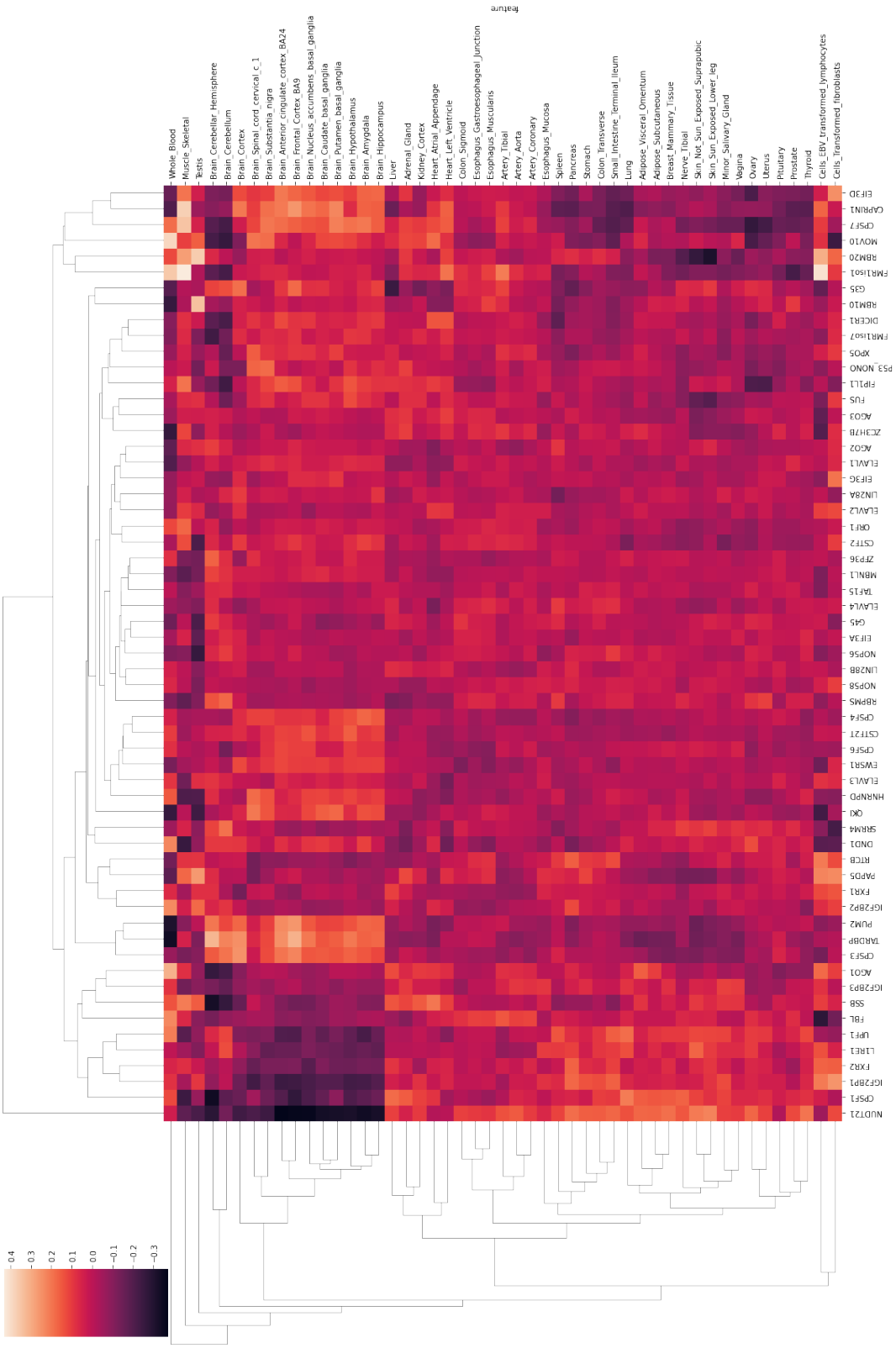

Figure 4.2: Clustered heatmap visualizing coefficients of RBP binding-scores in Model 2.

5 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 [4]), *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 5.1.

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 5.1: 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

[1] https://tissues.jensenlab.org/About, see also [3].

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.

6 Appendix

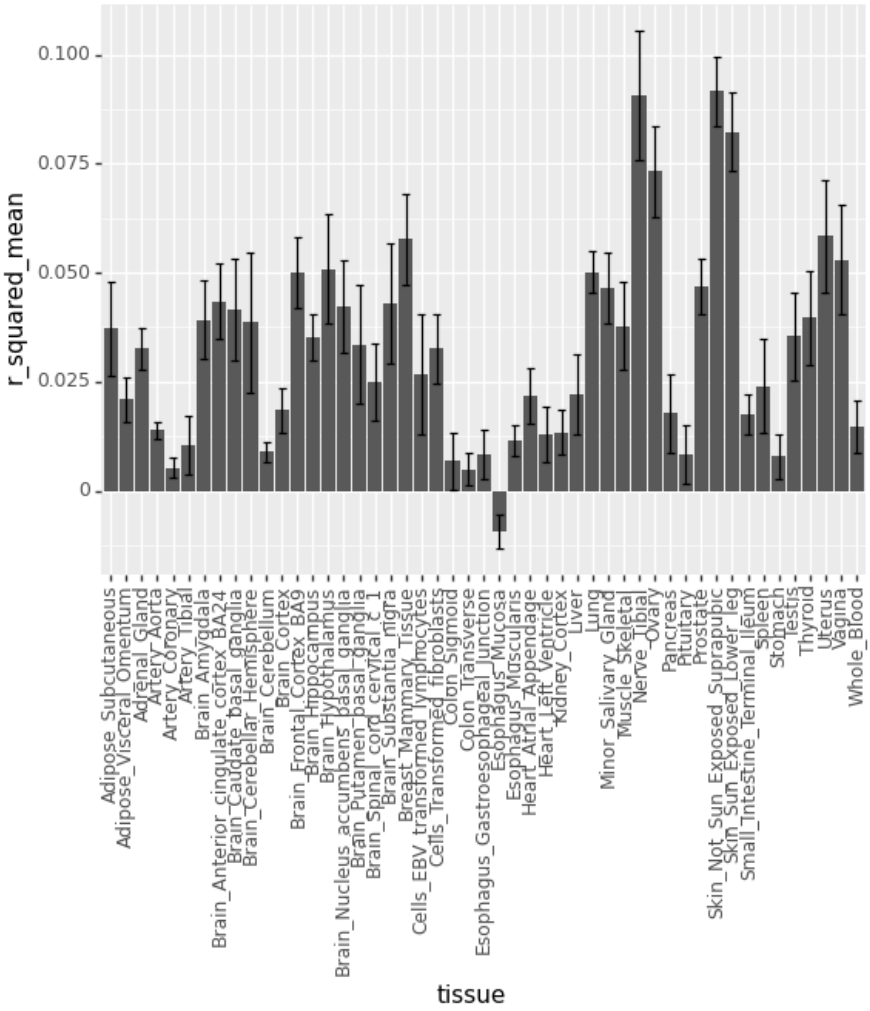

Figure 6.1: Achieved R^2 values on test set for different tissues in Model 1.

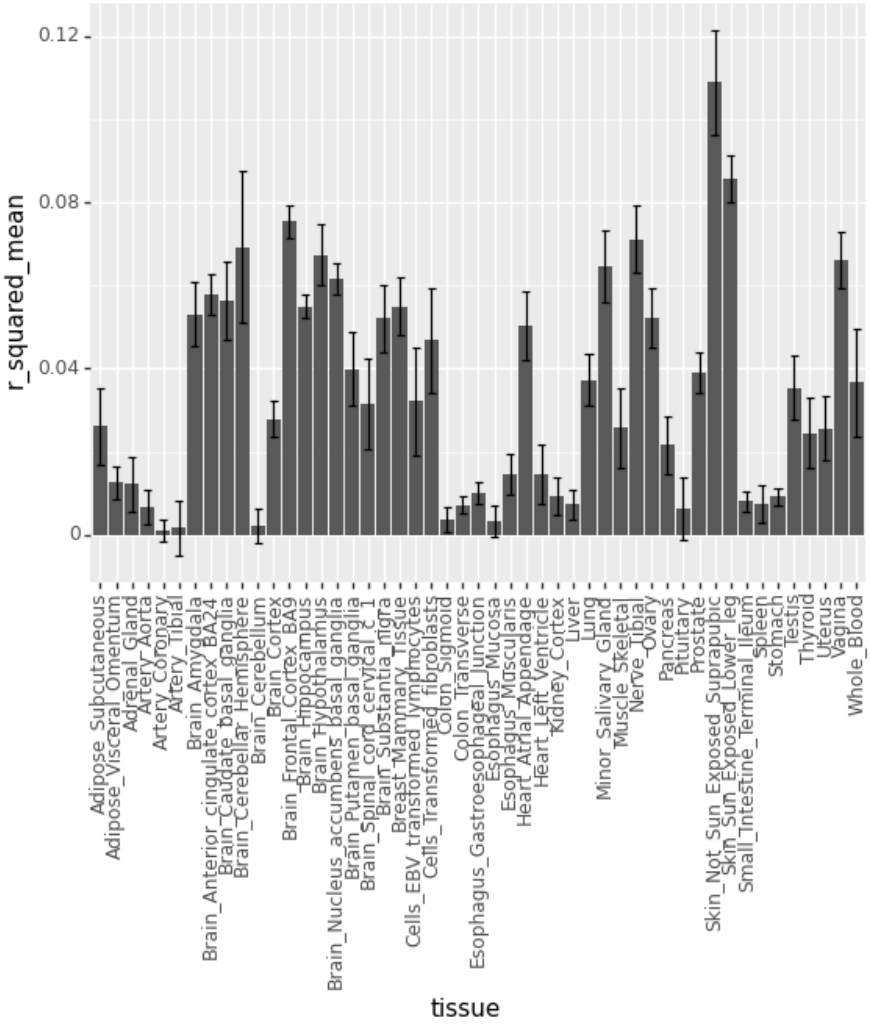

Figure 6.2: Achieved R^2 values on test set for different tissues in Model 2.

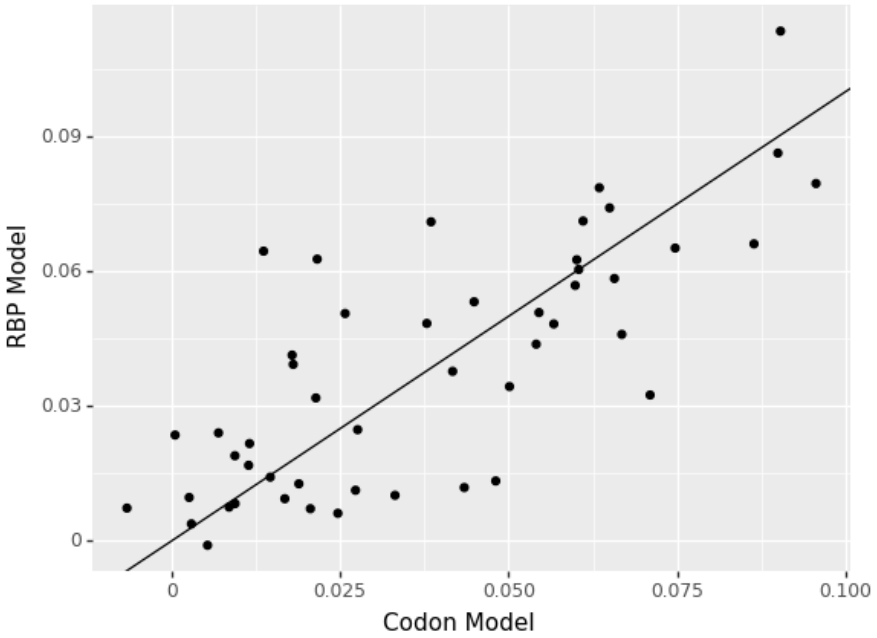

Figure 6.3: Comparing the achieved R^2 scores per tissue type in Model 2 with the respective scores in the baseline model.

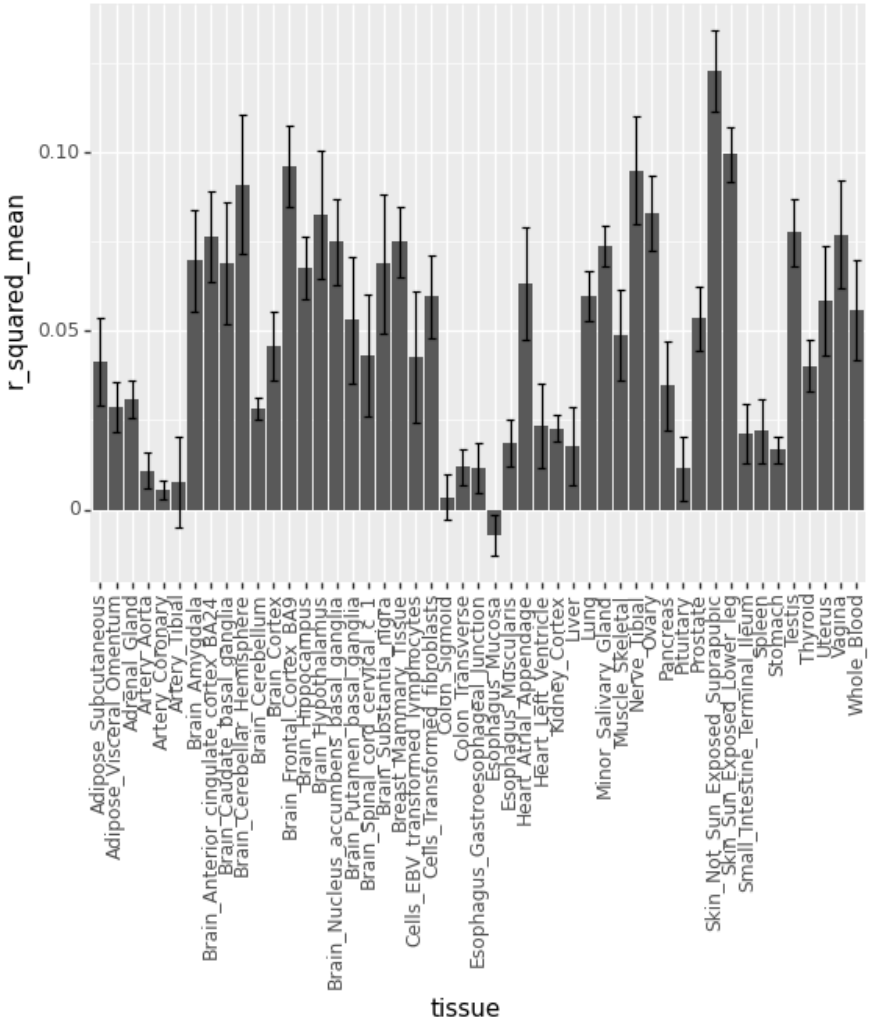

Figure 6.4: Achieved R^2 values on test set for different tissues in Model 3.

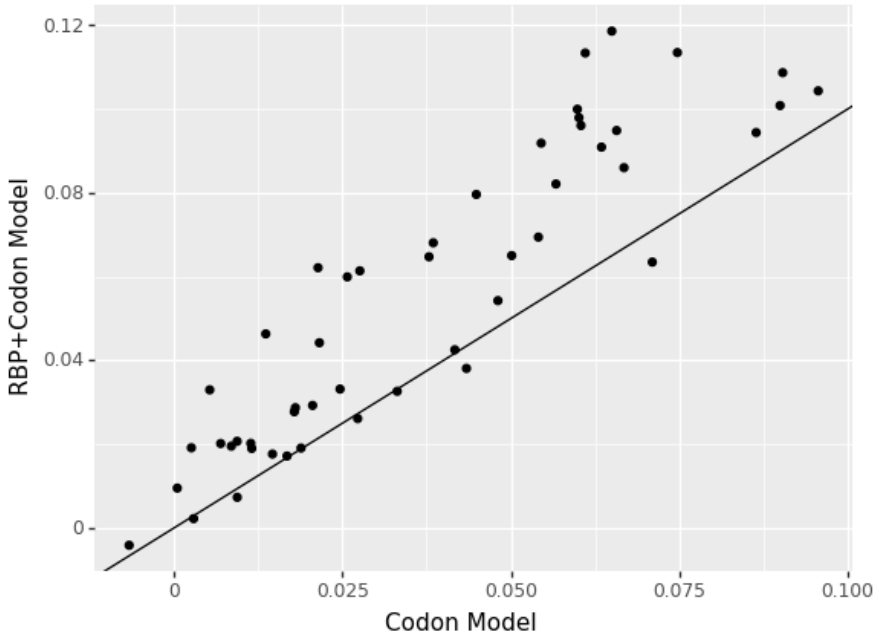

Figure 6.5: Comparing the achieved R^2 scores per tissue type in Model 3 with the respective scores in the baseline model.

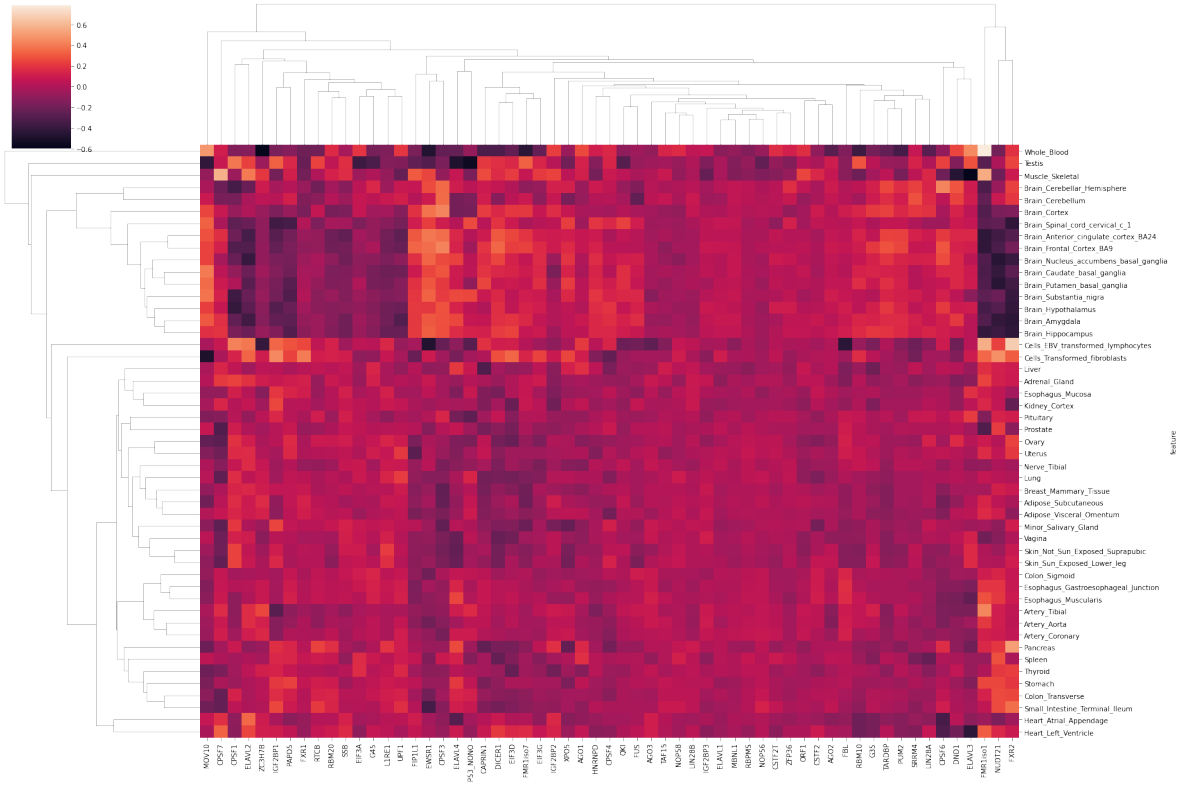

Figure 6.6: Clustered heatmap visualizing coefficients of RBP binding-scores in Model 3.

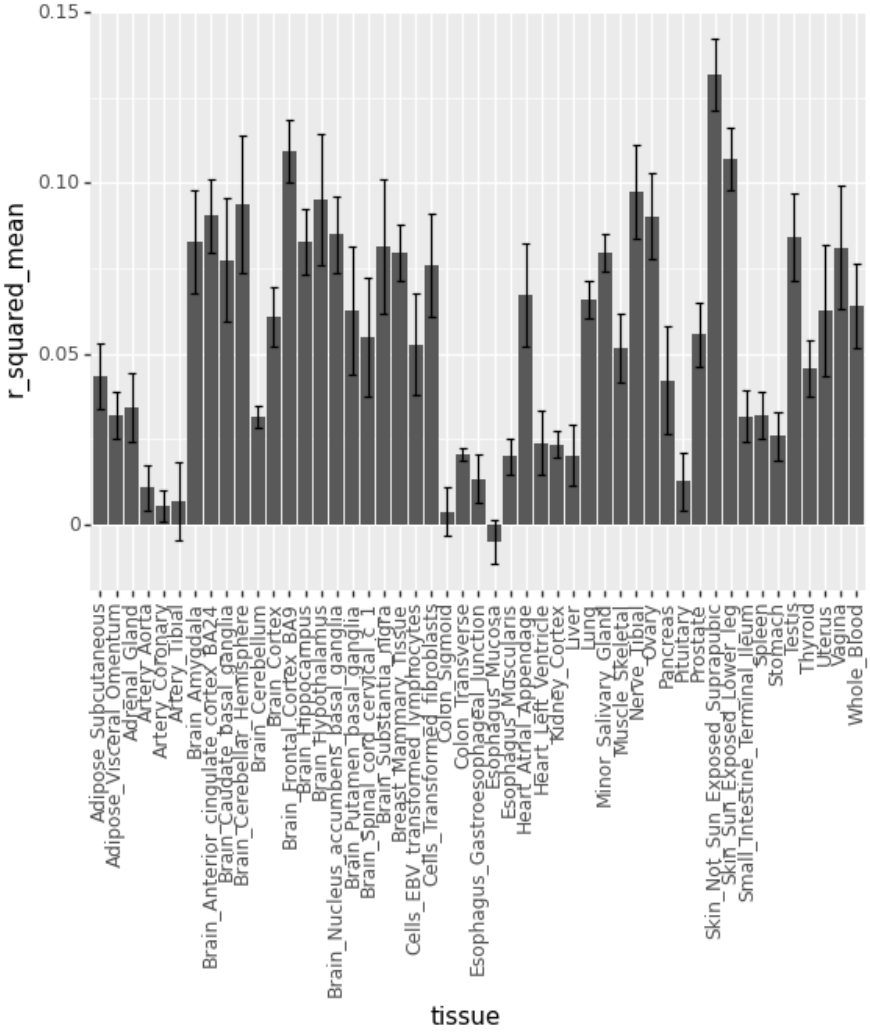

Figure 6.7: Achieved R^2 values on test set for different tissues in Model 4.

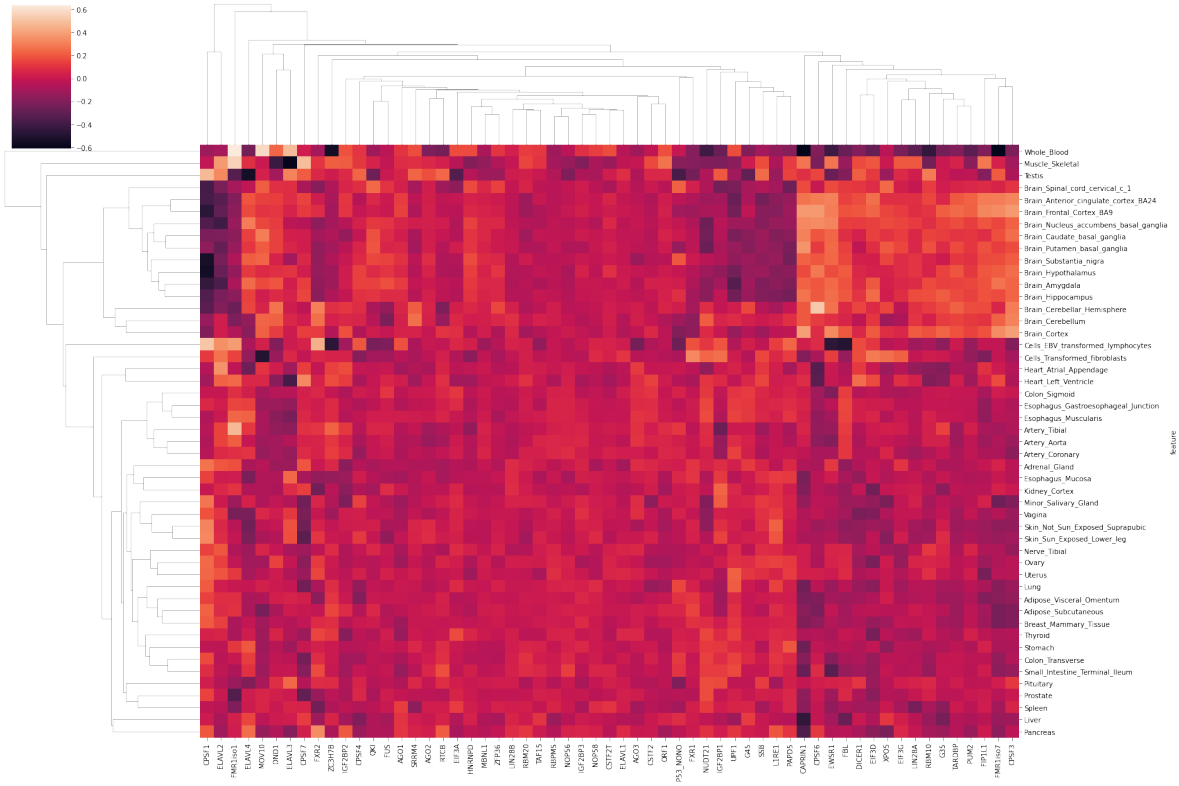

Figure 6.8: Clustered heatmap visualizing coefficients of RBP binding-scores in Model 4.

Bibliography

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

[2] 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.

[3] 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).

[4] 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.