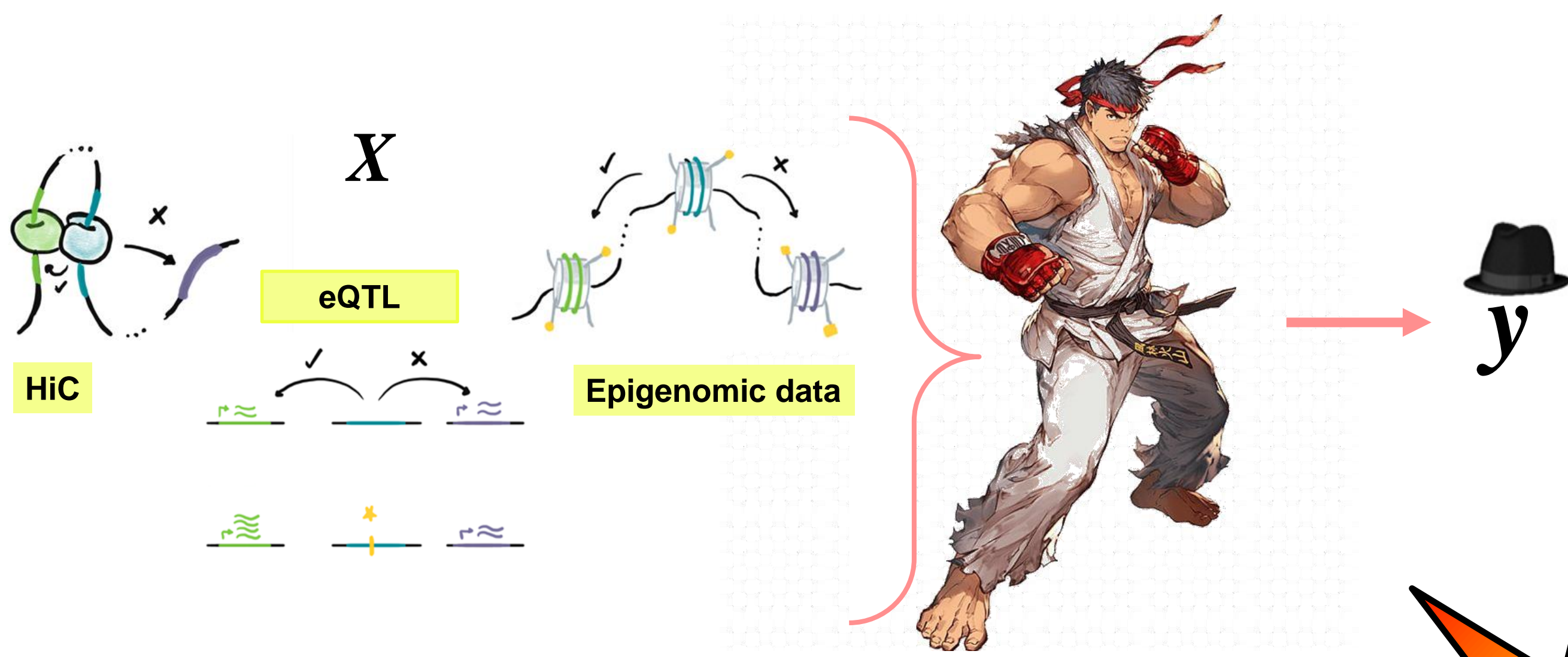


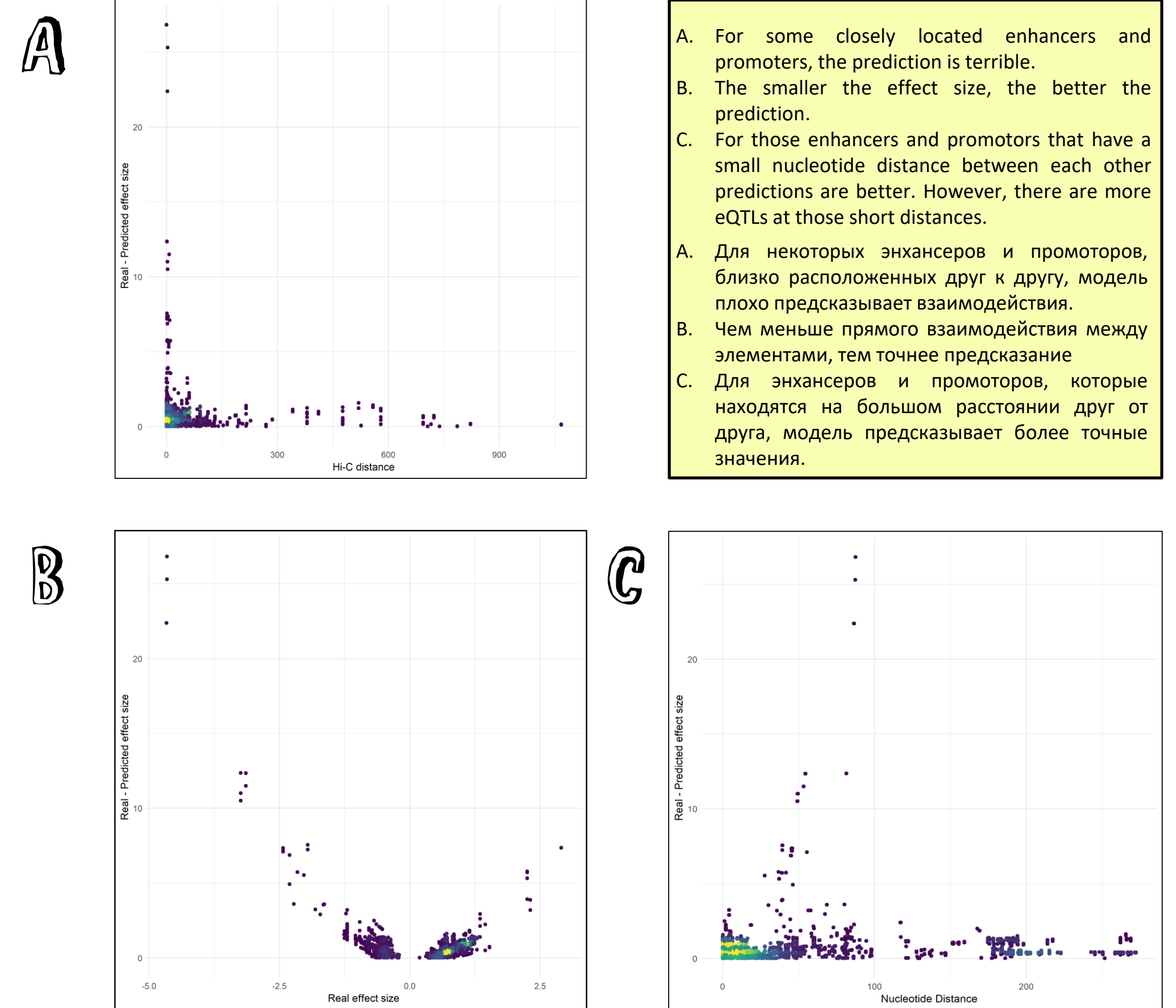
# >> MODEL BATTLE

## INTRODUCTION

- Enhancer - promoter interactions are unknown!
- One way to predict such interactions is using models that would be trained using already know information...
- Нам неизвестны взаимодействия между энхансерами и промоторами!
- Для того чтобы предсказать эти интеракции (взаимодействия), необходимо использовать самообучающиеся модели, которые будут тренироваться за счет уже известной информации...



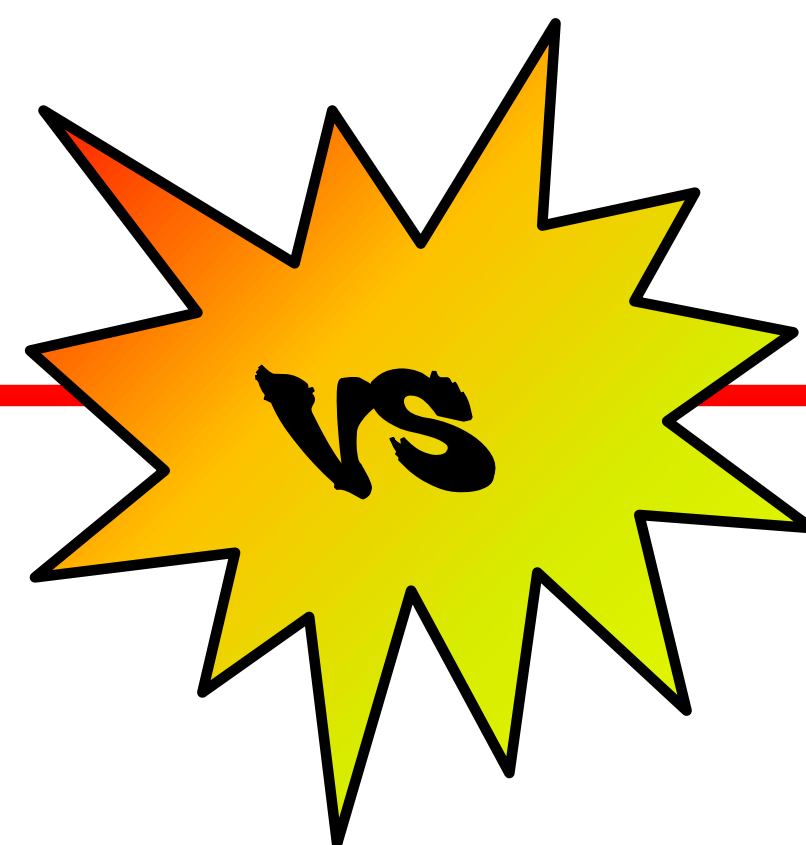
## WHAT CAN THE PARAMETERS SAY ABOUT ERROR?



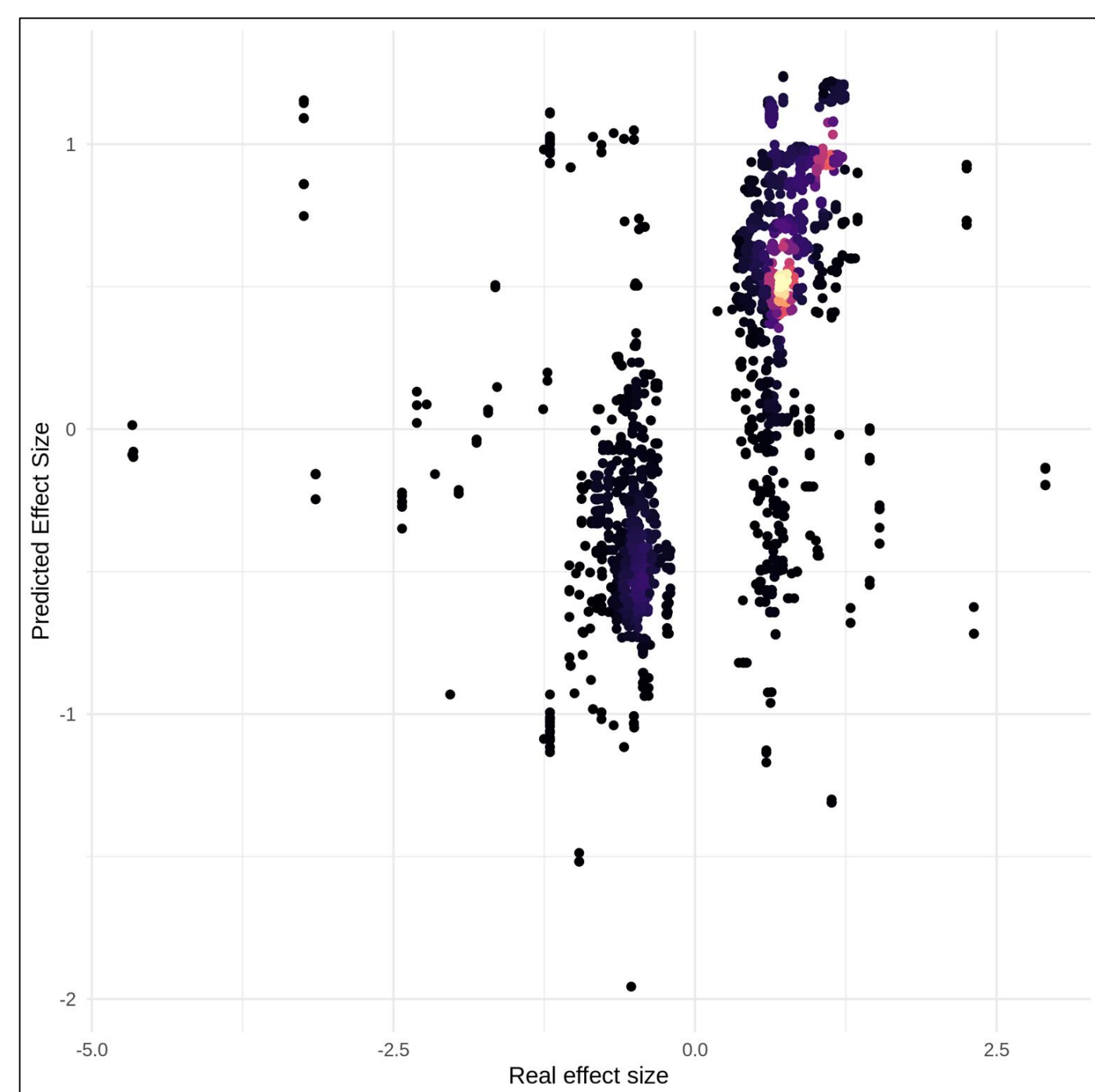
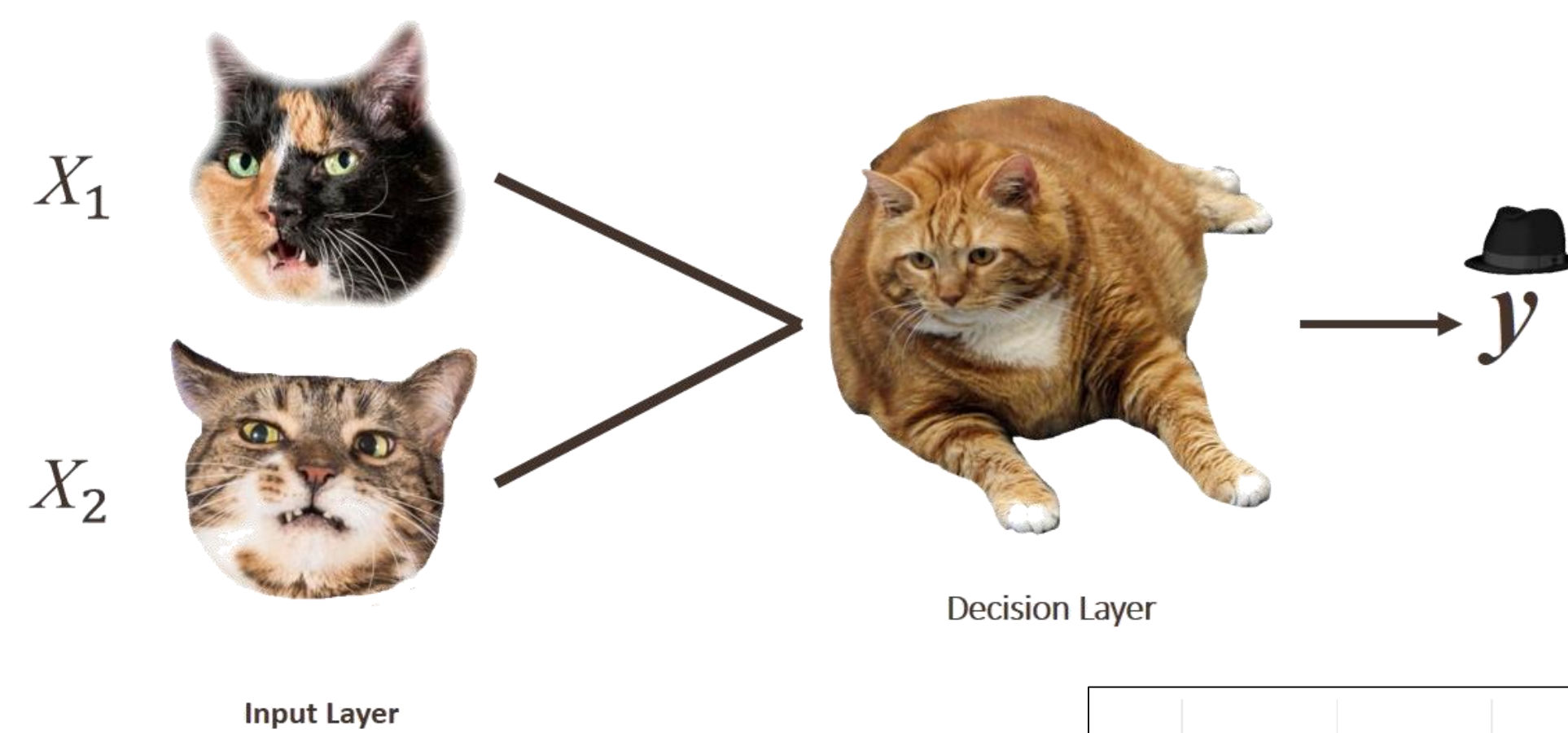
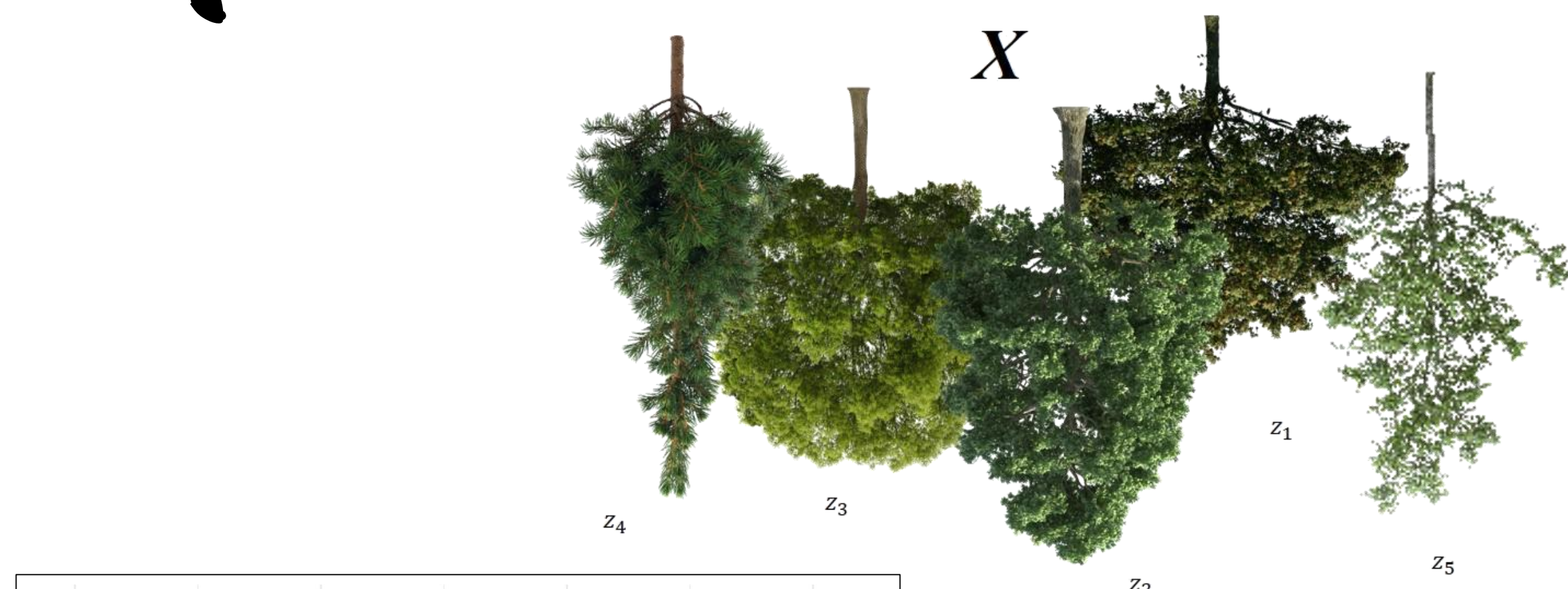
A. For some closely located enhancers and promoters, the prediction is terrible.  
 B. The smaller the effect size, the better the prediction.  
 C. For those enhancers and promoters that have a small nucleotide distance between each other predictions are better. However, there are more eQTLs at those short distances.

A. Для некоторых энхансеров и промоторов, близко расположенных друг к другу, модель плохо предсказывает взаимодействия.  
 B. Чем меньше прямого взаимодействия между элементами, тем точнее предсказание.  
 C. Для энхансеров и промоторов, которые находятся на большом расстоянии друг от друга, модель предсказывает более точные значения.

## Random Forest

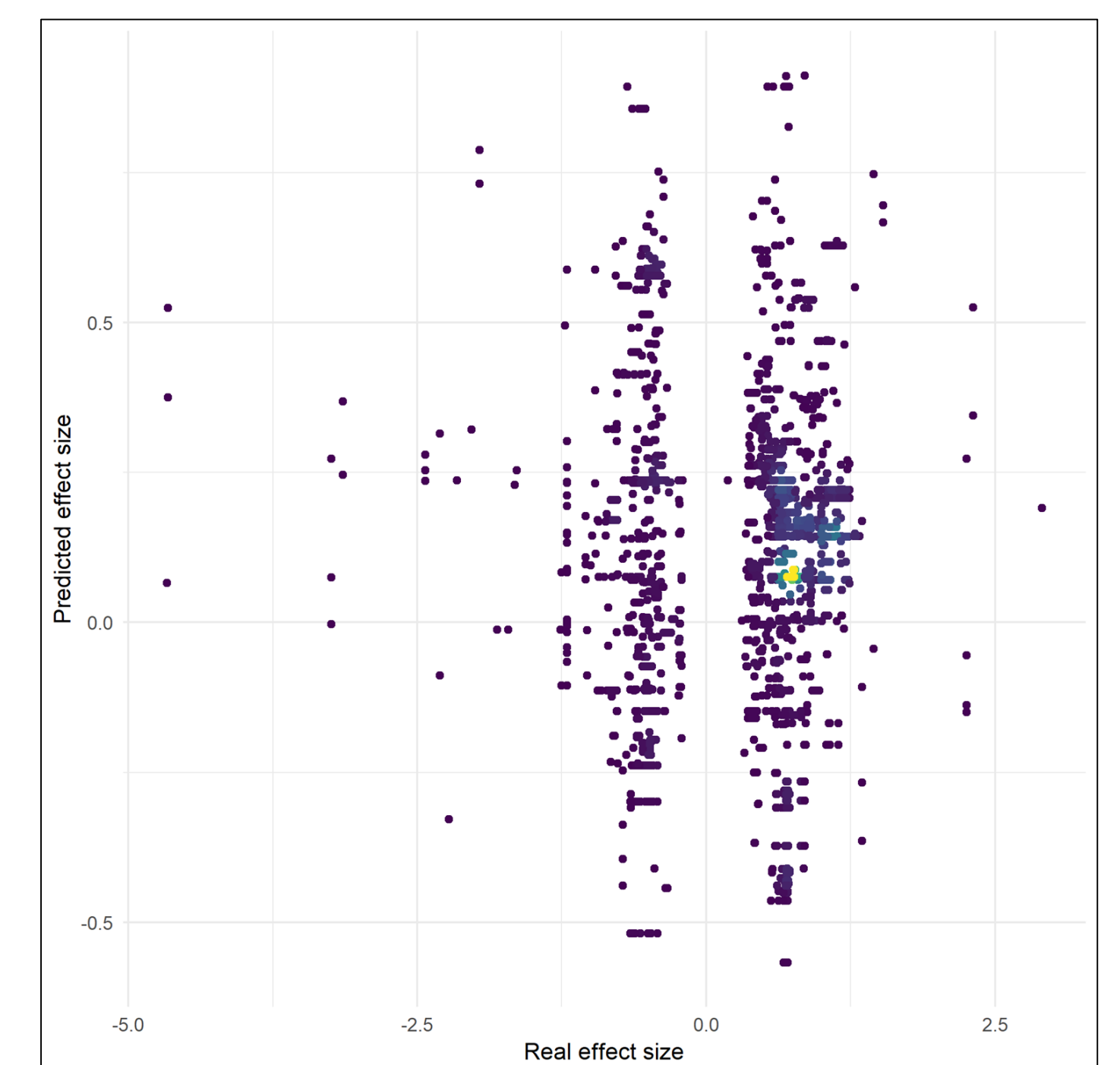


## Neural Net

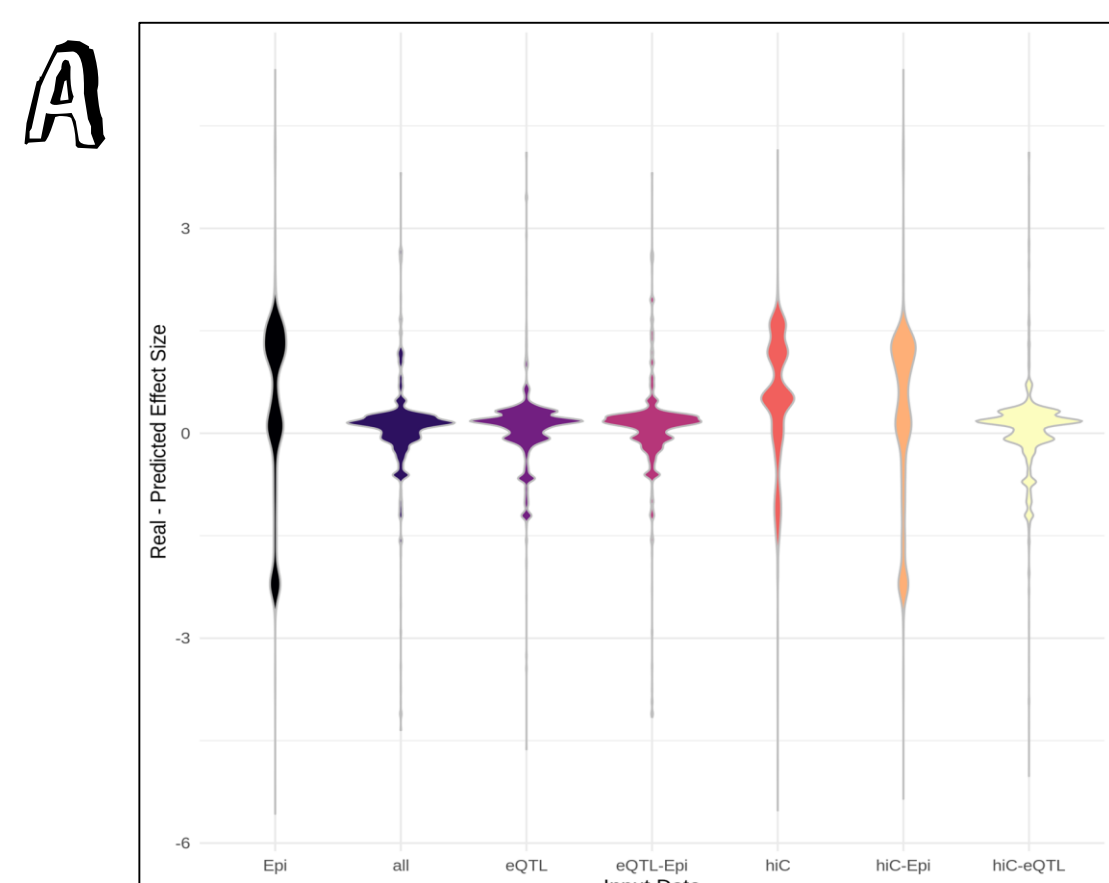


Relation between real eQTL value and predicted value. Most of the points are in the diagonal. Moreover, seems that both models predict better positive effects.

Соотношение между настоящими значениями взаимодействия энхансеров и промоторов и теми, что были предсказаны моделью. Большинство точек лежат на диагонали, однако, обе модели предсказывают больше положительных значений, чем это есть на самом деле

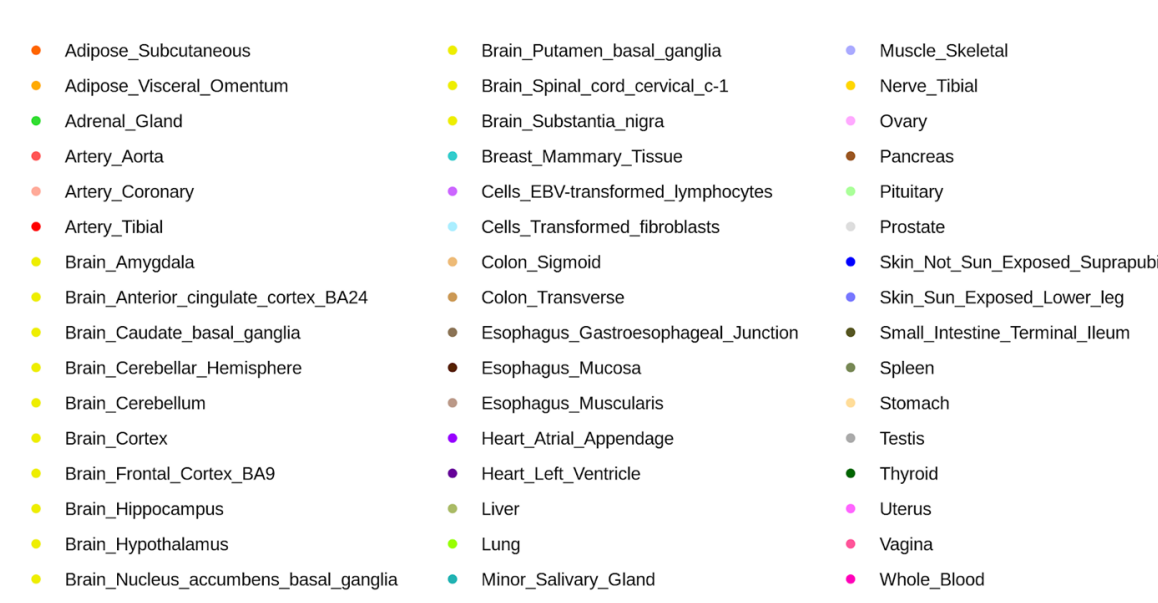


## WHICH FEATURES ARE IMPORTANT?



A. Using all types of data (eQTL+ Epigenomic data+ HIC) for prediction, we get the error similar to that produced using only eQTL, eQTL+Epigenomic data, eQTL + HIC. That means the epigenomic data and HIC are not adding a lot of predicting power.  
 B. The tissues that are similar to our tissue of interest are more important for making predictions.

A. Предсказания, для которых используются все типы информации (взаимодействия между элементами (eQTL), эпигенетические значения (Epi) и физические интеракции (HIC), имеют те же отклонения от реальных данных, что и предсказания, сделанные на основании только eQTL, eQTL+ Epi, eQTL + HIC. Из этого следует, что эпигенетическая информация и физическое взаимодействие в меньшей степени участвуют в предсказании интеракций между энхансерами и промоторами.  
 B. Данные тканей, аналогичных нашей, наиболее значимы для создания точных предсказаний.



## CONCLUSIONS AND PERSPECTIVES

1. Predictions made by Random Forest are better than the predictions made by Neural Net.
  2. There is a lot of variation across the chromosomes, that's why to improve the prediction we might apply individual approaches to each chromosome.
  3. To make the predictions more accurate, we might try other neural networks that have different architectures. We could also use other input data, such as the sequences of the elements.
1. Предсказания, сделанные RandomForest, точнее, чем предсказания, выполненные нейронной сетью.
  2. Для каждой хромосомы значимость различных параметров сильно варьируется, поэтому для качественных предсказаний необходим индивидуальный подход для всех хромосом.
  3. Для того чтобы сделать предсказания более точными, следует экспериментировать с различными архитектурами нейронных сетей, а также добавить сиквенсы элементов для входной информации.