Wray Lab Rotation Presentation

GABRIEL KENNEDY

12/16/2022

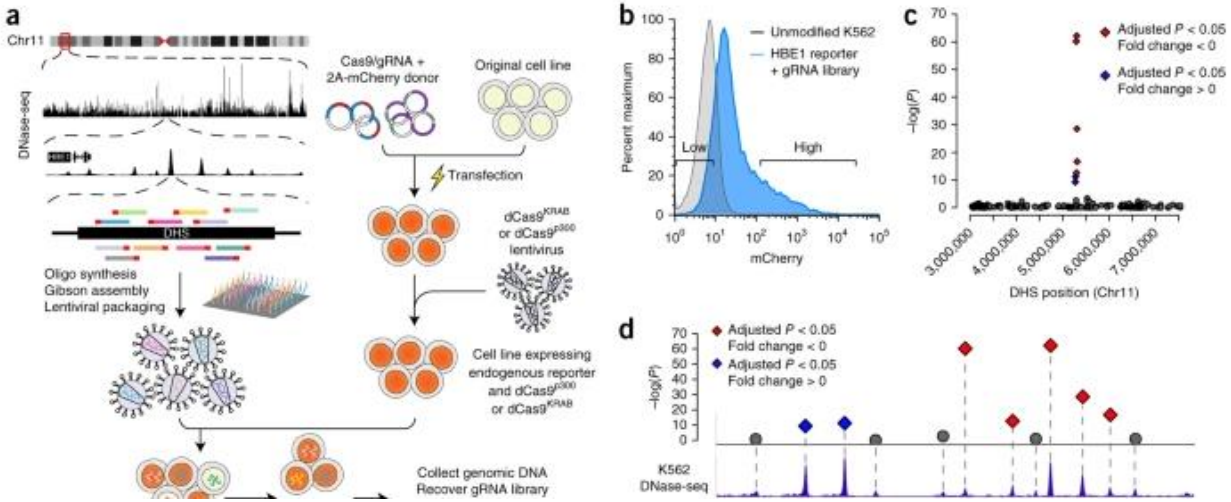
shed: 03 April 2017

ISPR–Cas9 epigenome editing enables hig roughput screening for functional regulat ments in the human genome

S Klann, Joshua B Black, Malathi Chellappan, Alexias Safi, Lingyun Song, Isaac ford ✉, Timothy E Reddy ✉ & Charles A Gersbach ✉

e Biotechnology 35, 561–568 (2017) | Cite this articl

Figure 1: CRISPR–Cas9-based epigenetic regulatory element screening (CERES) identifies regulatory elements of the β-globin locus in a loss-of-function screen.

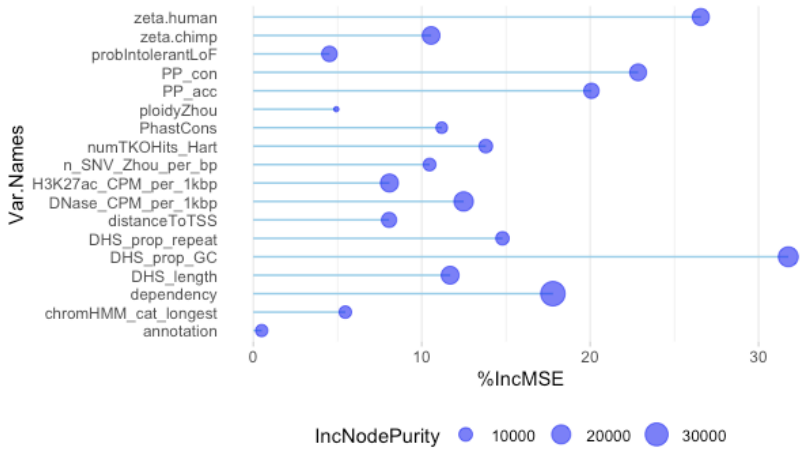


Original Project

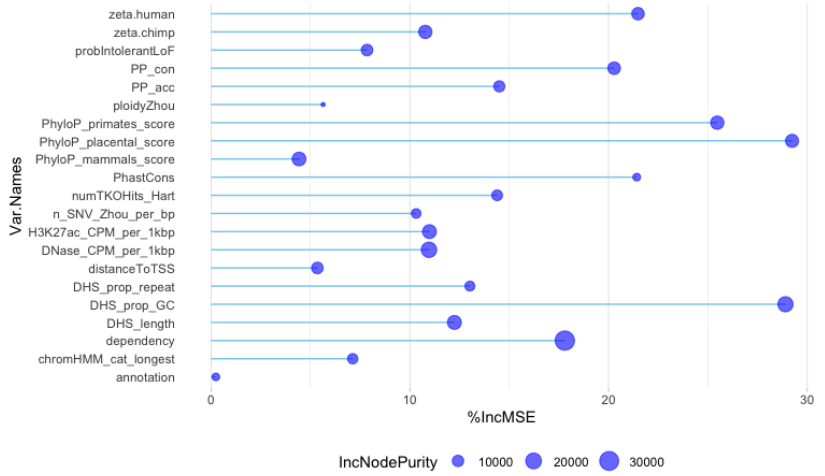
Determine importance of a variety of factors on scoring associated with the Ceres project

Whole-genome CERES score as the response vector

Excluding phyloP scores



Including phyloP scores



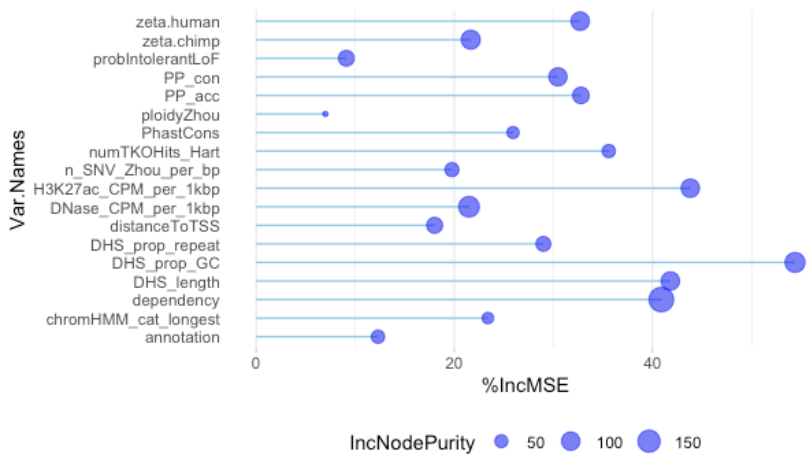
*From Jameson's work

Original Project

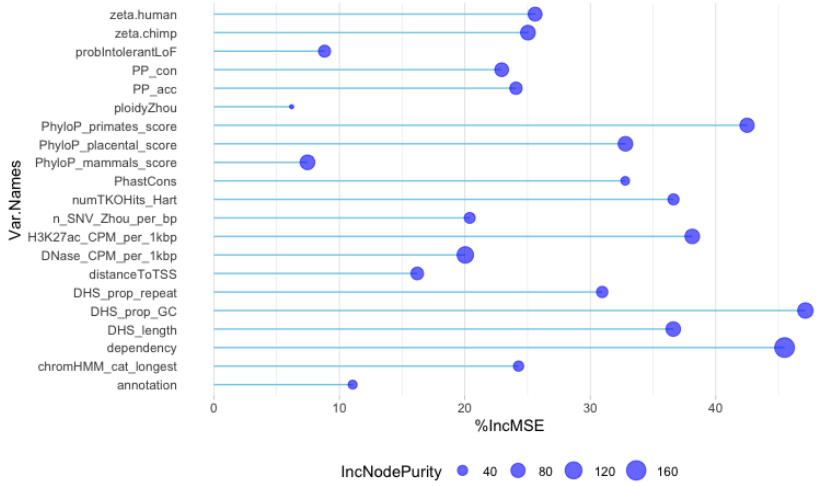
Determine importance of a variety of factors on scoring associated with the Ceres project

Was DHS (all gRNAs grouped) significant across all DHSs?

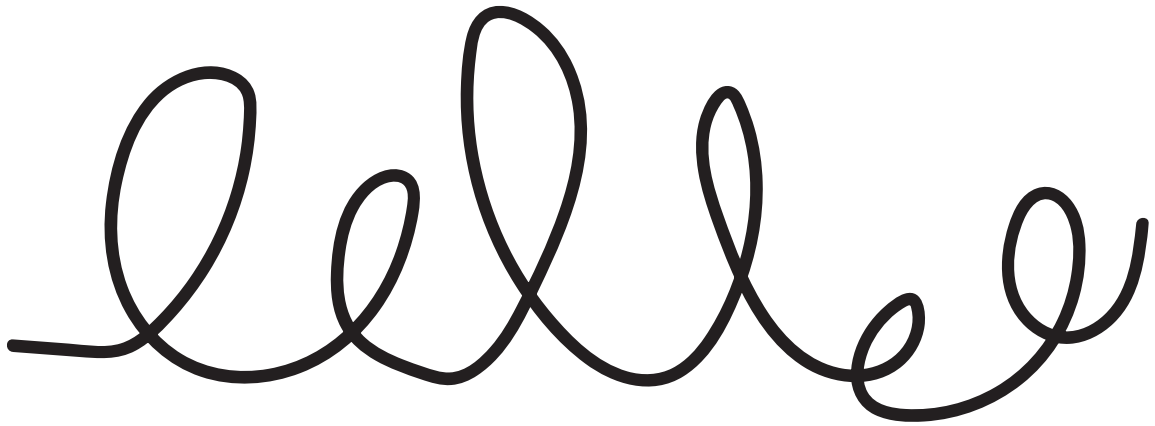
Excluding phyloP scores



Including phyloP scores



*From Jameson's work



Additional Considerations

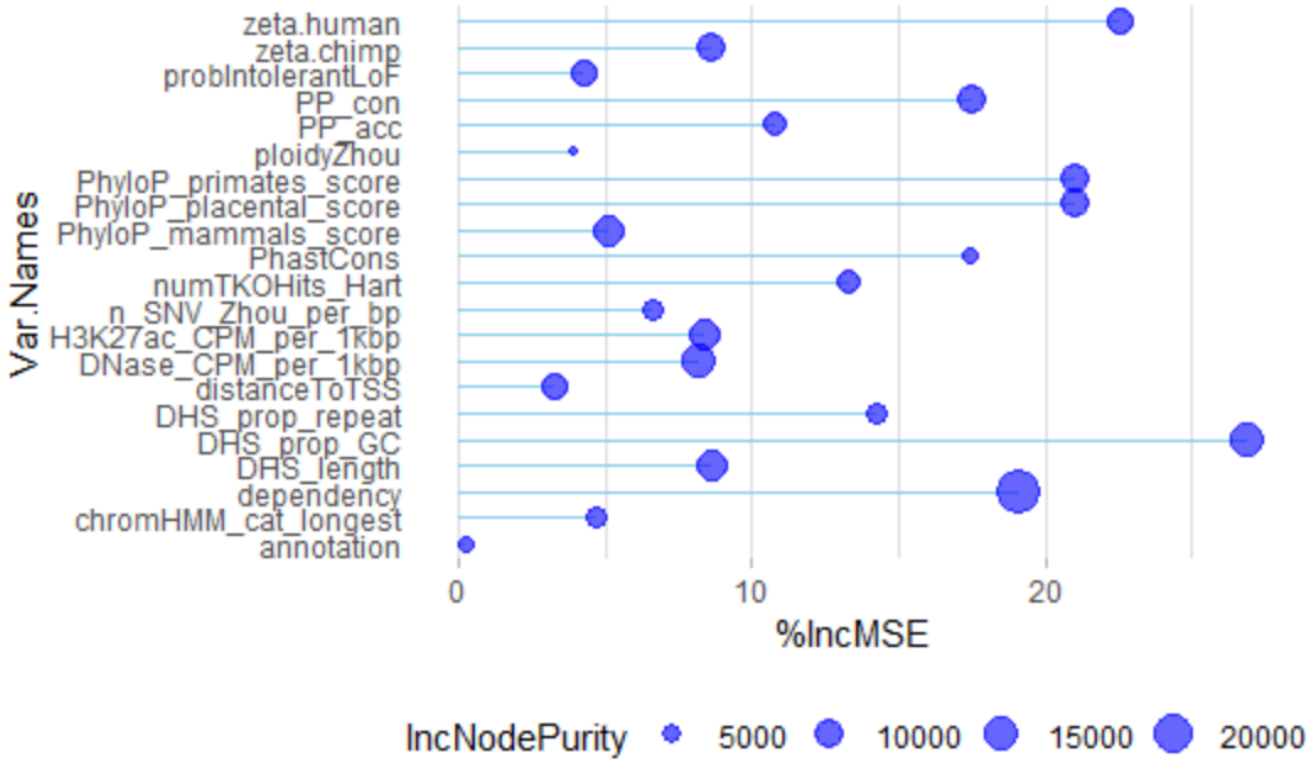
Three-dimensional structure

- TADs found using Hi-C
- How well do we trust the assignment of gene-element pairs?

Does random forest modeling fit this data best?

- What other ML models could be compared against?

How might additional data change these findings?



Updated with elements on TADs with both Enhancers and Promoters

Using Machine Learning Harnesses

Machine Learning Harness techniques compare multiple algorithms against each other by comparing output metrics between each looking for:

- Mean Absolute Error (MAE) and Mean Squared Error (RMSE)

Jameson compared the performances of ML techniques using original DHSs and removing DHSs that aren't contacting any core promoter, but did not find significant differences

Dataset	Machine Learning Technique	Error Metric	Significant DHS as response
Unfiltered	Logistic Regression	MAE	0.03
		RMSE	0.16
	Support Vector Machine	MAE	0.03
		RMSE	0.16
TAD-confirmed	Logistic Regression	MAE	0.03
		RMSE	0.16
	Support Vector Machine	MAE	0.03
		RMSE	0.16

Machine Learning Technique	Error Metric	CERES score as response
Linear Model	MAE	1.28
	RMSE	2.00
Gradient Boosted Machine	MAE	1.29
	RMSE	1.99
Linear Model	MAE	1.30
	RMSE	2.06
Gradient Boosted Machine	MAE	1.30
	RMSE	2.05

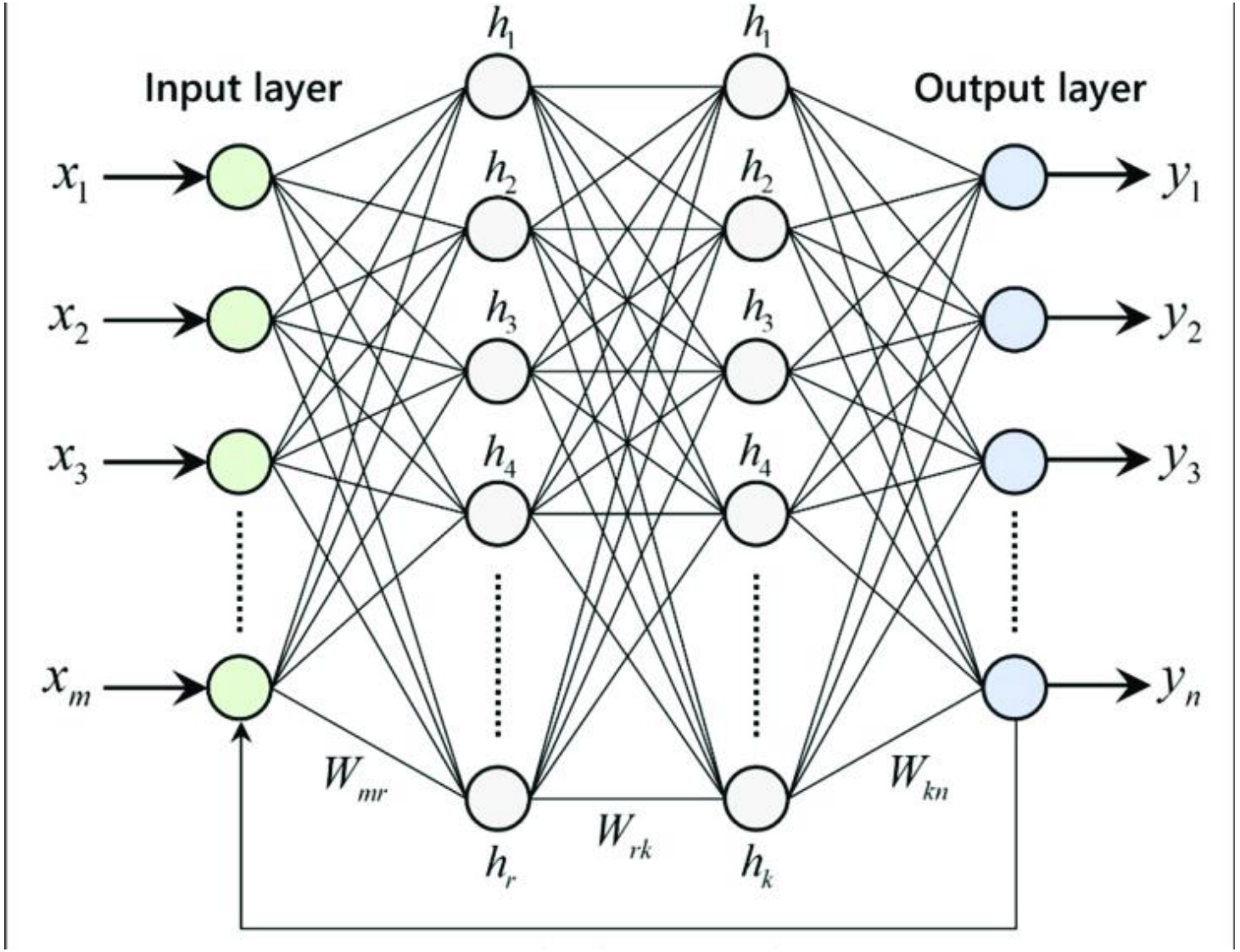
Adding to the ML Harness Models

Unfiltered

Linear Regression	MAE	1.28
-	RMSE	1.99
Gradient Boosting Machine	MAE	1.28
-	RMSE	1.97
Support Vector Regression	MAE	1.28
-	RMSE	2.02
RANSAC Regression	MAE	1.26
-	RMSE	2
Random Forest Regression	MAE	1.32
-	RMSE	2.01

TAD-confirmed

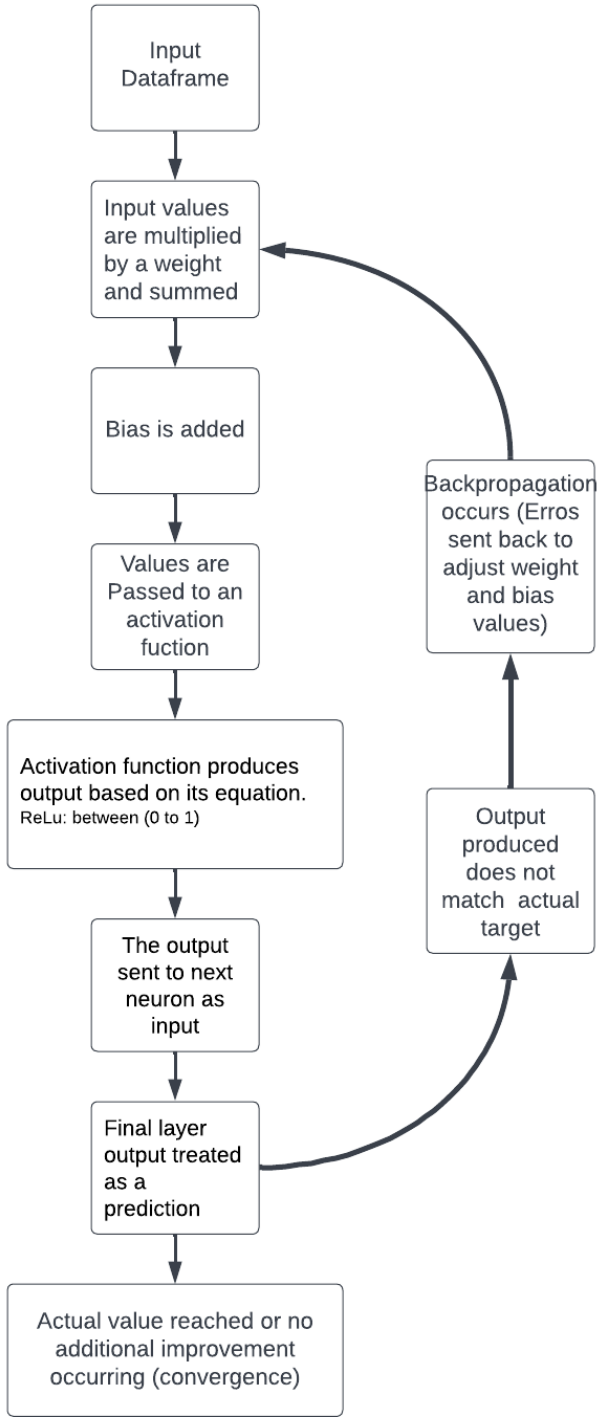
Linear Regression	MAE	1.3
-	RMSE	2.06
Gradient Boosting Machine	MAE	1.3
-	RMSE	2.03
Support Vector Regression	MAE	1.3
-	RMSE	2.09
RANSAC Regression	MAE	1.29
-	RMSE	2.07
Random Forest Regression	MAE	1.11
-	RMSE	1.84



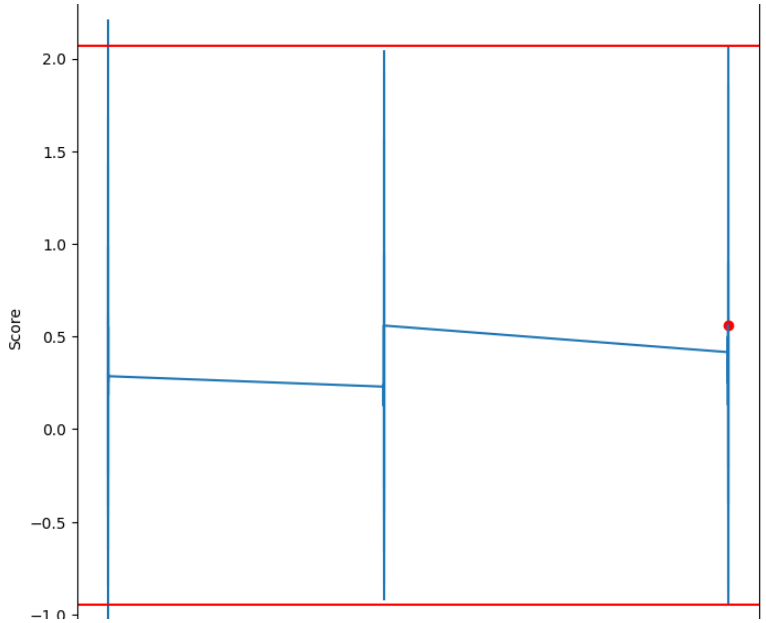
ML models – Neural Network

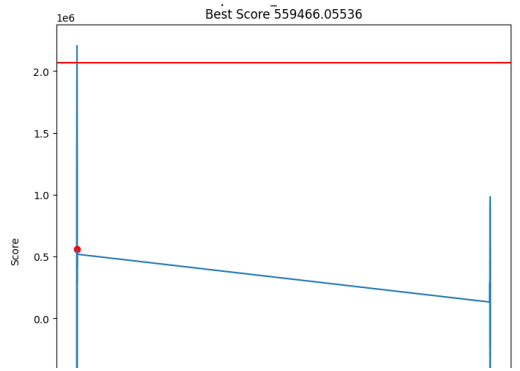
Unsupervised Models: Full data as input, no target, let the algorithm determine the targets

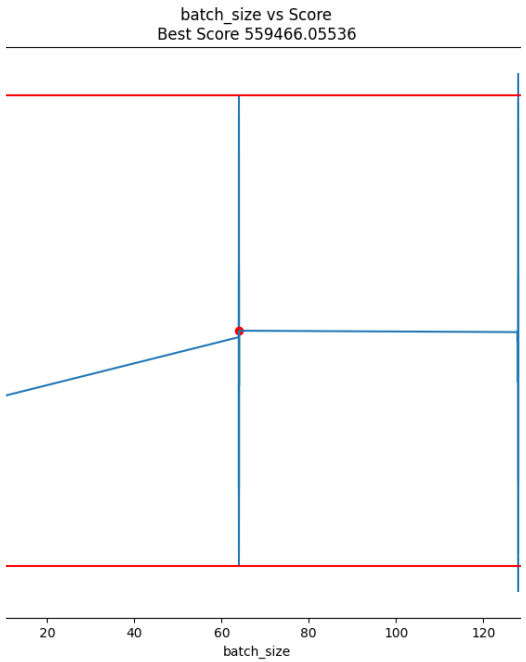
Supervised Models: Trained with specific input and output examples, let the algorithm figure out the pattern that works within the restrictions



Neural Network Process







Neural Network – Hyperparameter tuning

Determine which parameters lead to the best fit:

Hyperparameter tuning neural network: compared batch sizes of 8, 64, 128; epochs of 10, 50, 100; and optimizers of Adam and rmsprop

Best fit:

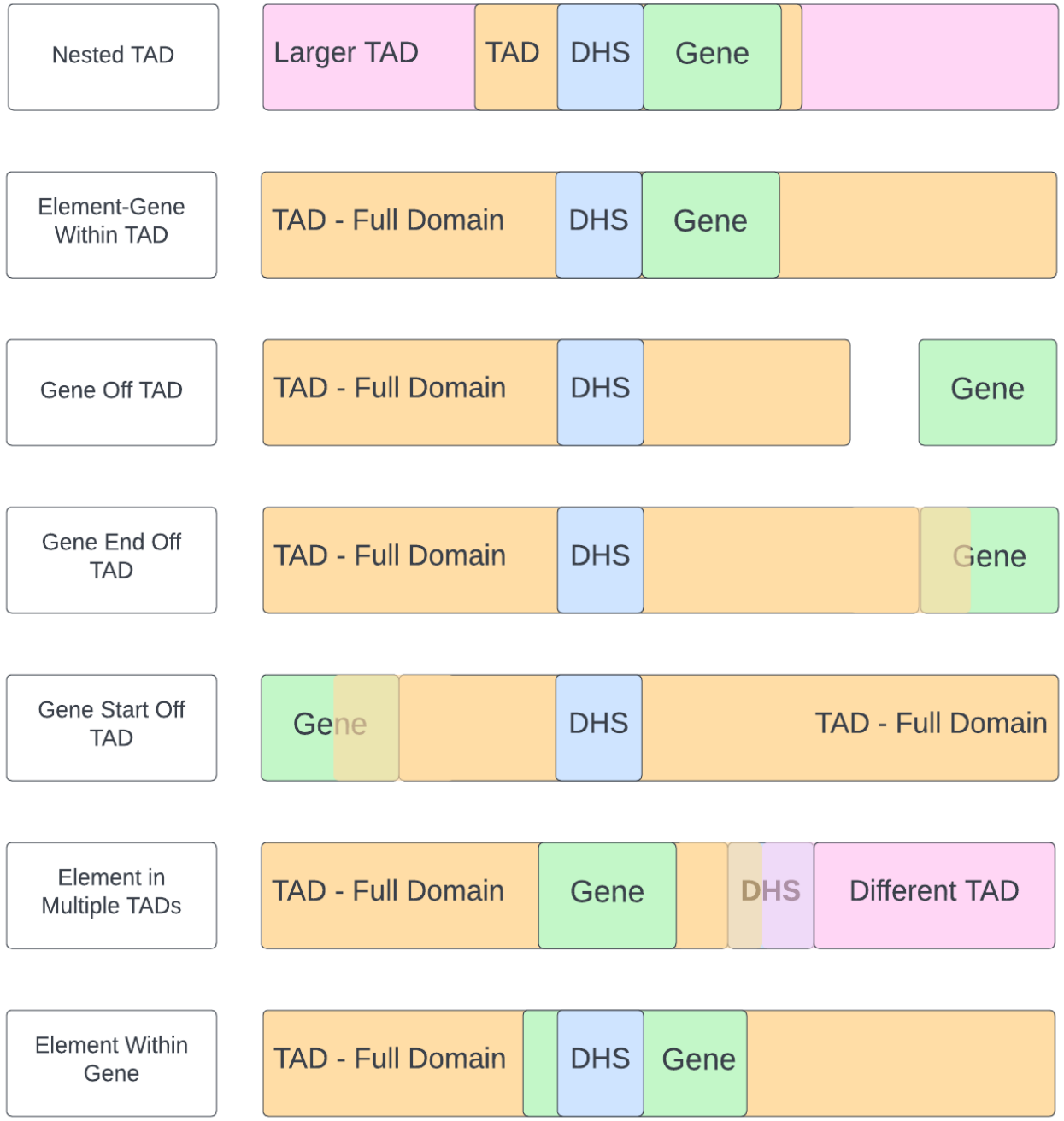
{Optimizer_trial': 'adam', 'batch_size': 64, 'epochs': 100}

MAE Output: 1.3

RMSE Output: 2.01

The models may not be that different.

SO WHAT OTHER IMPORTANT FACTORS ARE THERE?

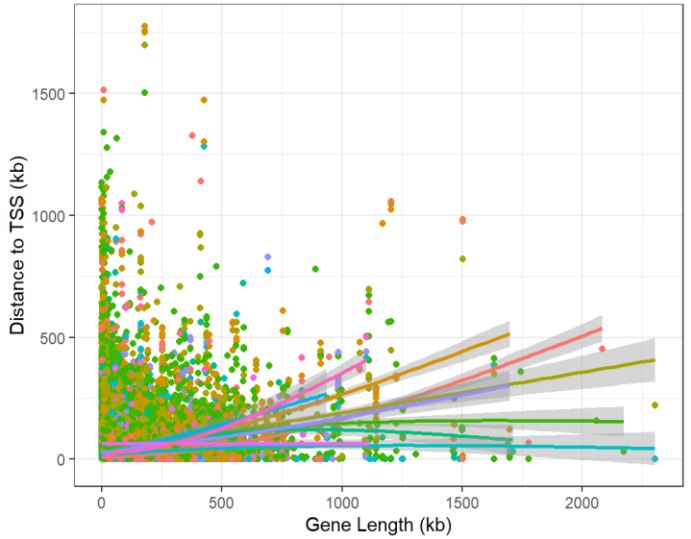


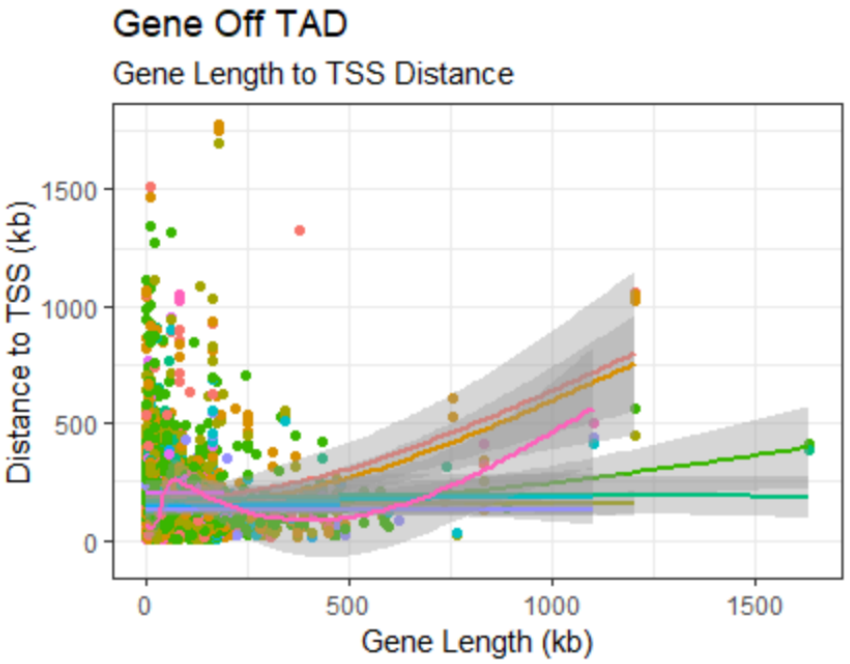
Where is the gene or element in relation to the TAD?

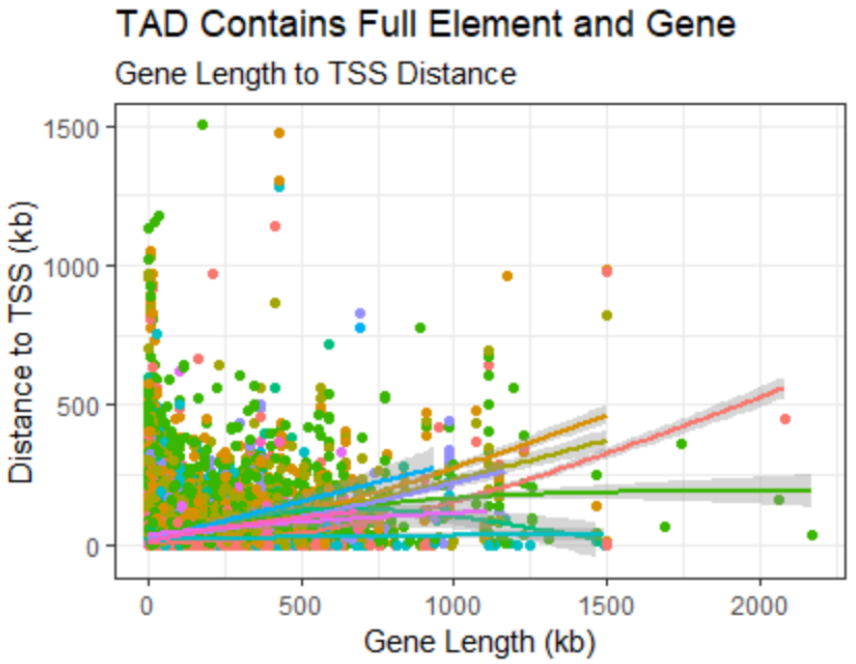
Relationship	Active Promoter	Candidate Strong Enhancer	Heterochromatin/Repetitive/Copy Number Variation	Candidate Weak Enhancer	Distal CTCF/Candidate Insulator	Transcription asociated	Inactive Promoter	Polycomb repressed	Low activity proximal to active states
Element-Gene Within TAD	2623 (18.2%)	6685 (49.7%)	1057 (45%)	8307 (51.8%)	8020 (54.3%)	1923 (45.3%)	1321 (35.9%)	4215 (50.2%)	342 (56.2%)
Element in Multiple TADs	54 (0.4%)	15 (0.1%)	10 (0.4%)	23 (0.1%)	15 (0.1%)	4 (0.1%)	9 (0.2%)	12 (0.1%)	1 (0.2%)
Element within Gene	11090 (77.1%)	4683 (34.8%)	1035 (44.1%)	5269 (32.8%)	4054 (27.5%)	2076 (48.9%)	1984 (53.9%)	3133 (37.3%)	177 (29.1%)
Gene End off TAD	205 (1.4%)	573 (4.3%)	54 (2.3%)	621 (3.9%)	647 (4.4%)	74 (1.7%)	103 (2.8%)	283 (3.4%)	23 (3.8%)
Gene off TAD	254 (1.8%)	1033 (7.7%)	131 (5.6%)	1237 (7.7%)	1378 (9.3%)	106 (2.5%)	168 (4.6%)	530 (6.3%)	47 (7.7%)
Gene Start off TAD	152 (1.1%)	459 (3.4%)	62 (2.6%)	586 (3.7%)	652 (4.4%)	66 (1.6%)	98 (2.7%)	223 (2.7%)	18 (3%)

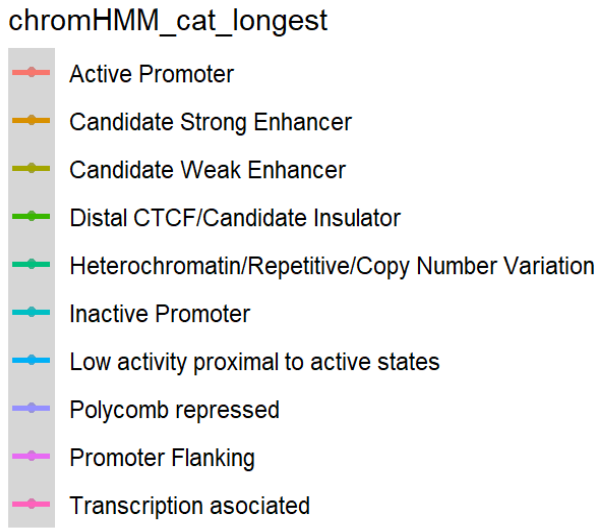
All Regulatory Elements

All Regulatory Elements

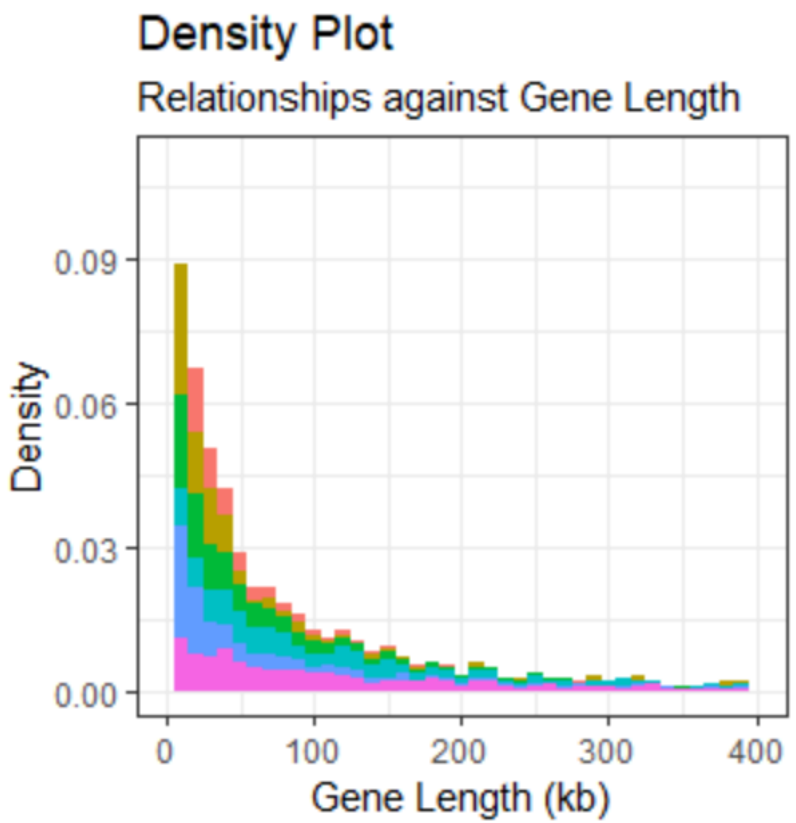


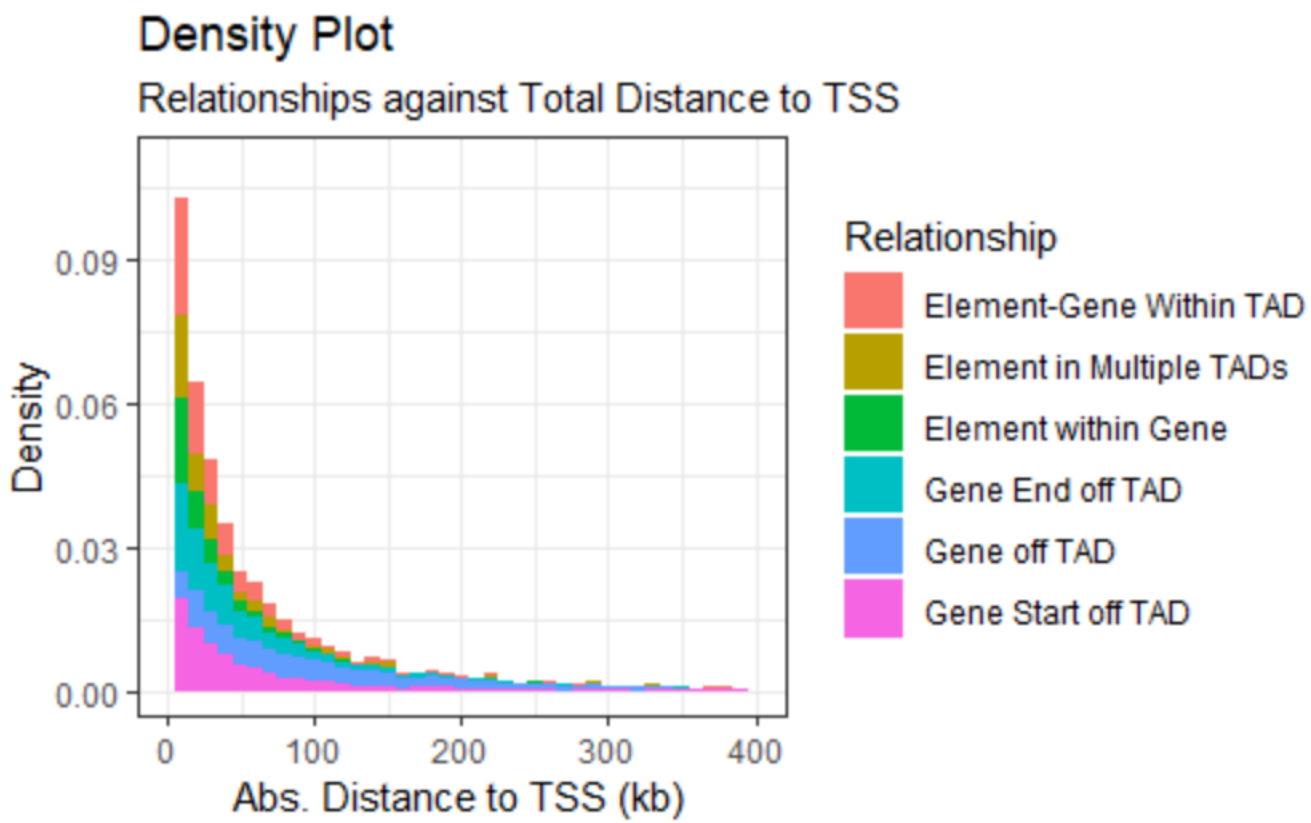






All Regulatory Elements





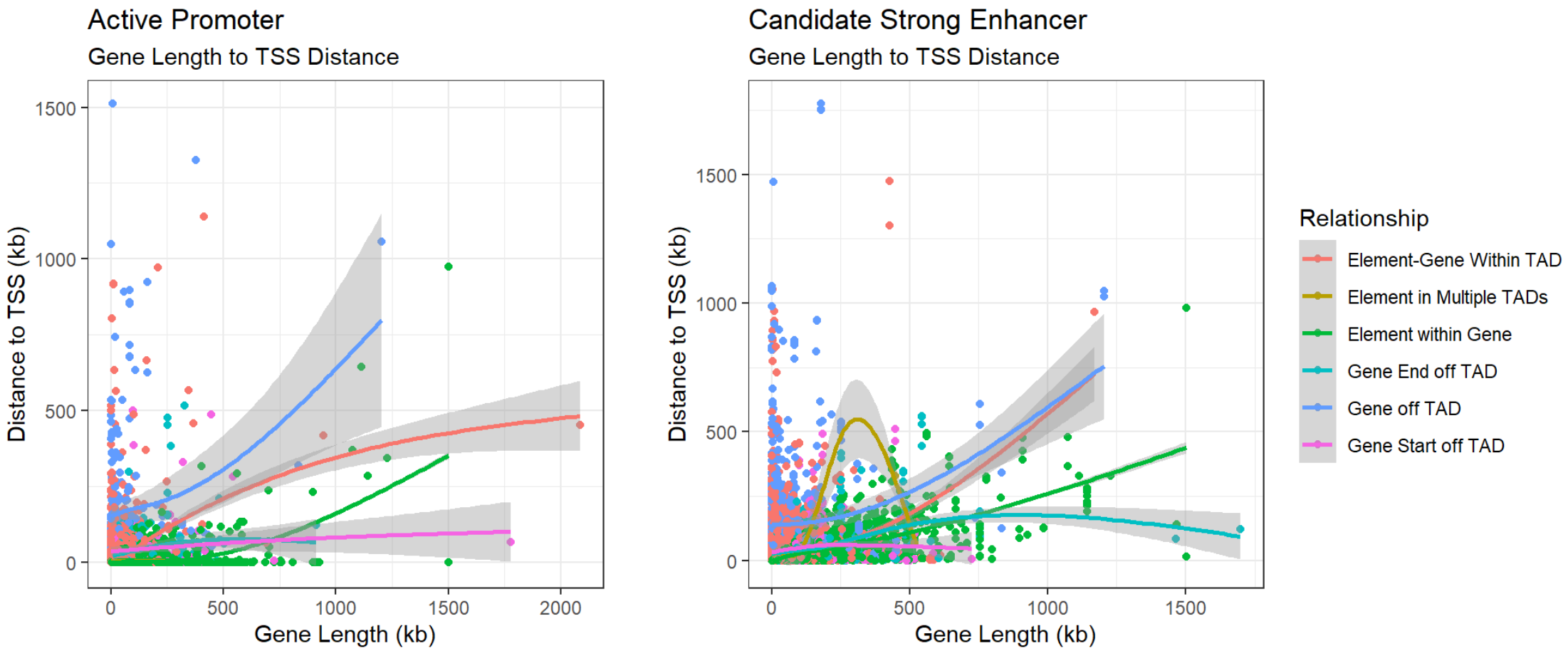
Outliers were removed to make this clearer (1415 Outliers)

Outliers were removed to make this clearer (581 Outliers)

Element-Gene Relationships

Relationship	Cand. Enhancer (Weak)	Cand. Enhancer (Strong)	Active Promoter	Inactive Promoter
Proximity Assignement Likely Correct				
Element within Gene	5269 (32.8%)	4683 (34.8%)	11090 (77.1%)	1984 (53.9%)
Element-Gene Within TAD	8307 (51.8%)	6685 (49.7%)	2623 (18.2%)	1321 (35.9%)
Proximity Assignment Less Likely				
Gene off TAD	1237 (7.7%)	1033 (7.7%)	254 (1.8%)	168 (4.6%)
Gene Start off TAD	586 (3.7%)	459 (3.4%)	152 (1.1%)	98 (2.7%)
Gene End off TAD	621 (3.9%)	573 (4.3%)	205 (1.4%)	103 (2.8%)
Proximity Assignment Uncertain				
Element in Multiple TADs	23 (0.1%)	15 (0.1%)	54 (0.4%)	9 (0.2%)

Promoter and Enhancer Comparisons

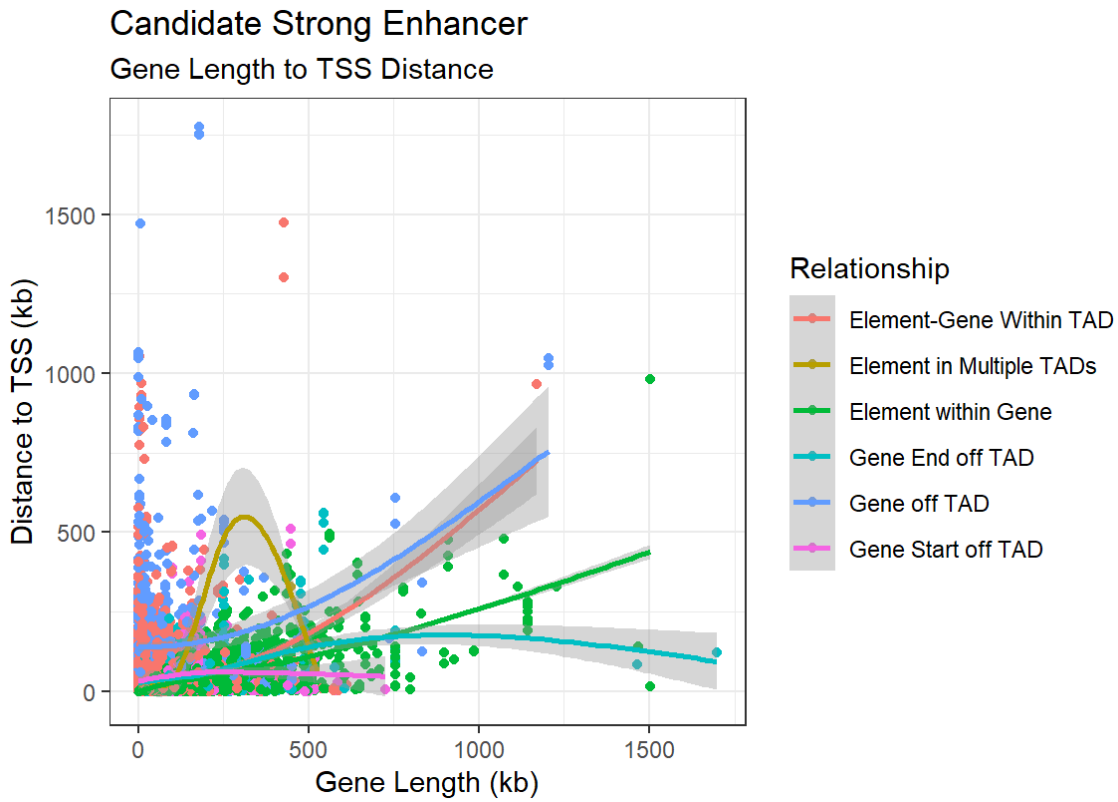


Candidate Strong Enhancer

Gene off TAD

TSS - Decile	N	Percent
1	6	0.6
2	8	0.8
3	8	0.8
4	23	2.2
5	37	3.6
6	68	6.6
7	79	7.6
8	129	12.5
9	199	19.3
10	476	46.1

Gene Length - Decile	N	Percent
1	142	13.7
2	135	13.1
3	105	10.2
4	83	8.0
5	111	10.7
6	97	9.4
7	69	6.7
8	111	10.7
9	96	9.3
10	84	8.1

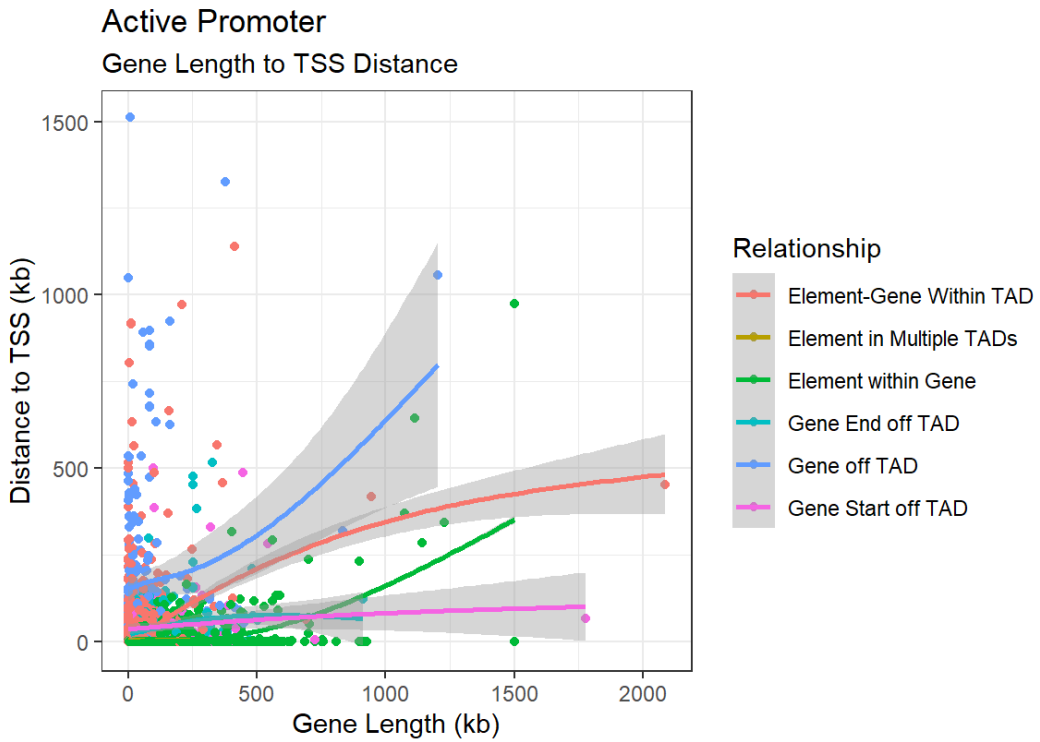


Active Promoter

Gene off TAD

TSS - Decile	N	Percent
7	1	0.4
8	2	0.8
9	35	13.8
10	216	85.0

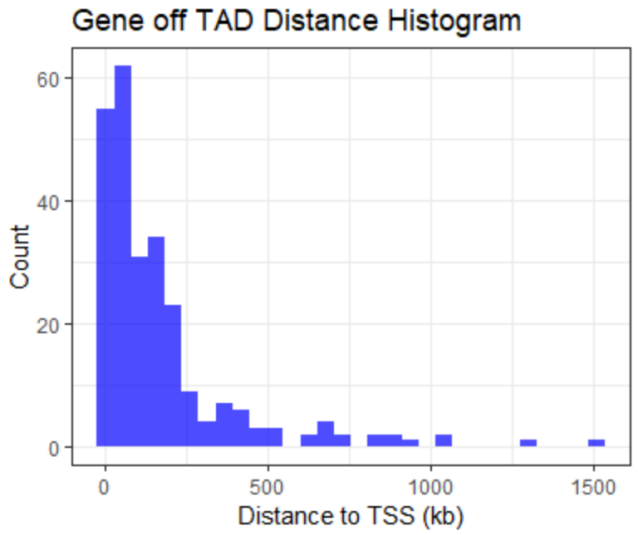
Gene Length - Decile	N	Percent
1	31	12.2
2	33	13.0
3	23	9.1
4	20	7.9
5	35	13.8
6	14	5.5
7	14	5.5
8	19	7.5
9	29	11.4
10	36	14.2



Active Promoter

Gene off TAD

TSS - Decile	N	Percent
7	1	0.4
8	2	0.8
9	35	13.8
10	216	85.0



In the 10th decile, the active promoters that are off TAD from the assigned gene go from being uncommon to being ~9% of the total elements seen, demonstrating uncertainty at that level of distance from the nearest TSS.

Note: This decile includes everything from 20kb to beyond 1500kb.

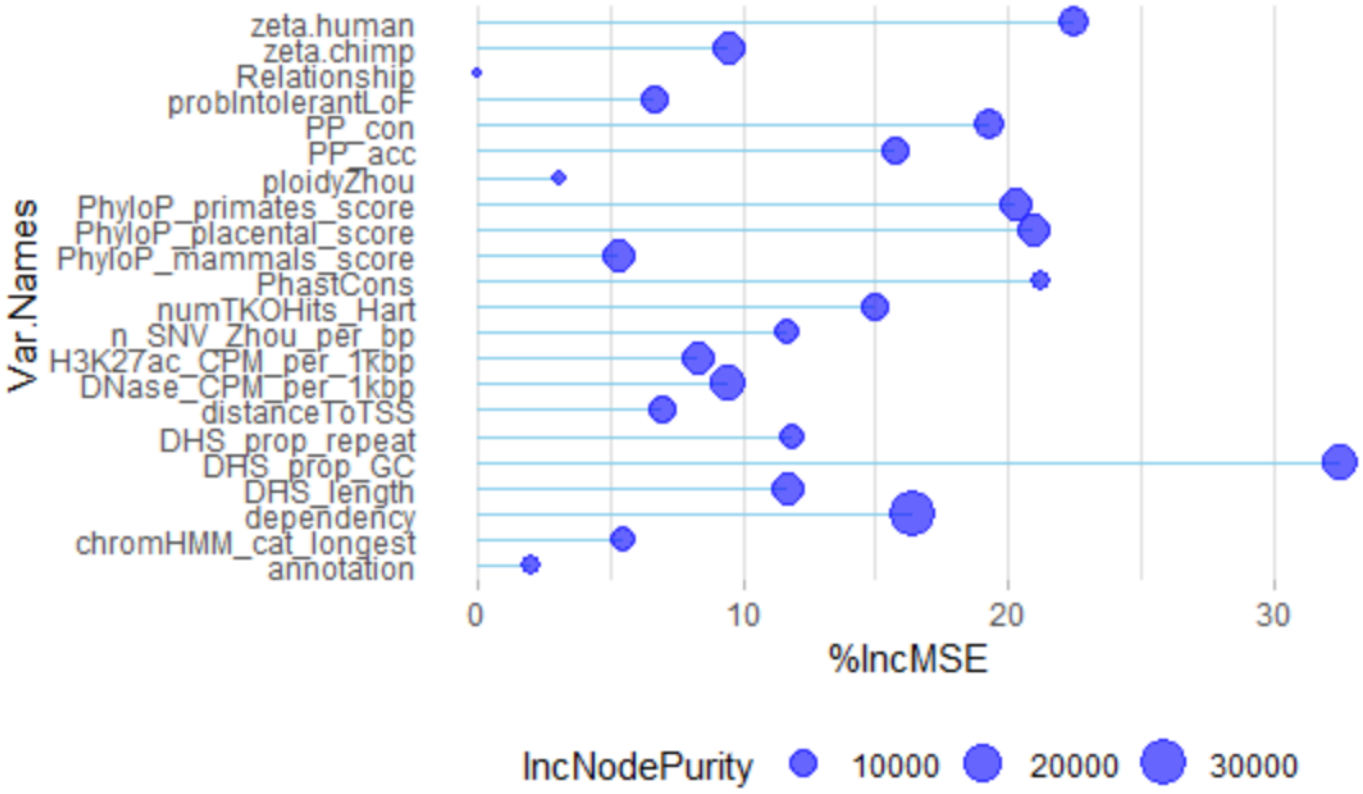
All Others Relationships

TSS - Decile	N	Percent
1	1438	10.2
2	1438	10.2
3	1438	10.2
4	1438	10.2
5	1438	10.2
6	1438	10.2
7	1437	10.2
8	1436	10.2
9	1402	9.9
10	1221	8.6

Redone without “Gene off TAD”

By removing elements that had an assigned gene in a separate TAD, I wanted to see how the models could change.

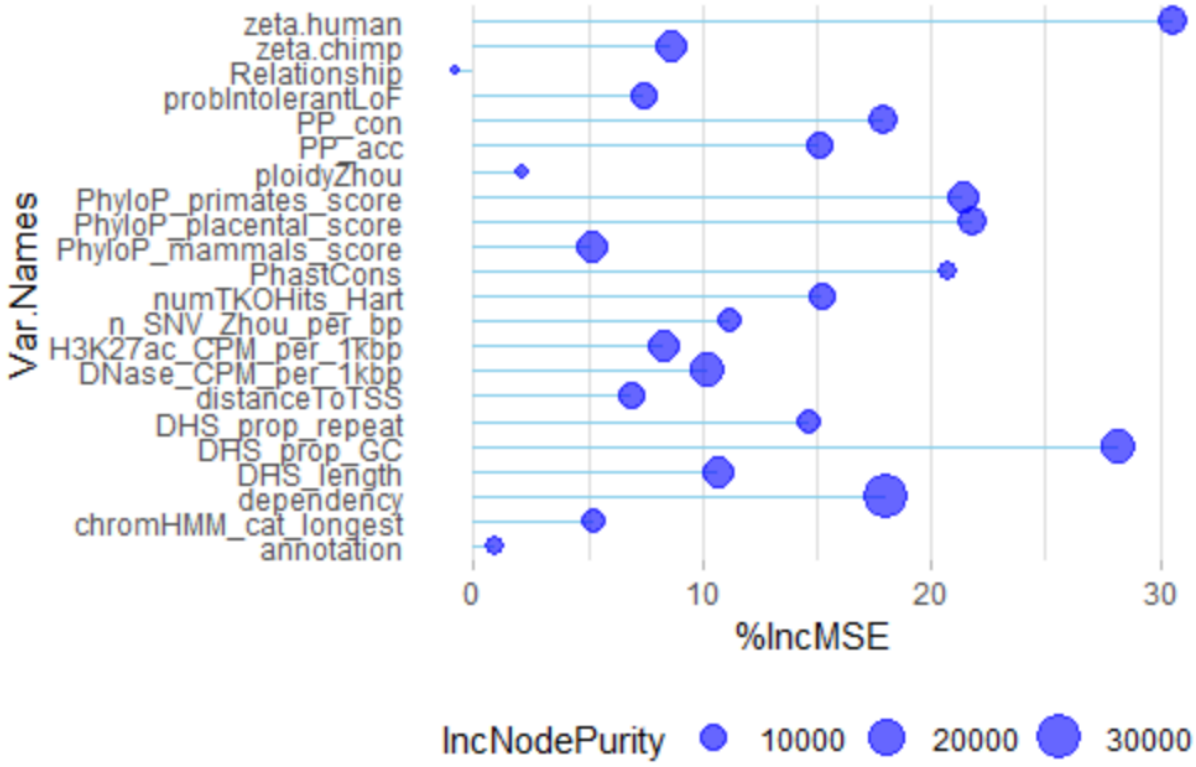
Linear Regression	MAE	1.27
-	RMSE	1.89
Gradient Boosting Machine	MAE	1.27
-	RMSE	1.88
Support Vector Regression	MAE	1.26
-	RMSE	1.92
RANSAC Regression	MAE	1.29
-	RMSE	1.9
Random Forest Regression	MAE	1.31
-	RMSE	1.93



Redone with the “Relationship” variable

By including how TADs were interacting with the element-gene pairs, I wanted to see how the models could change.

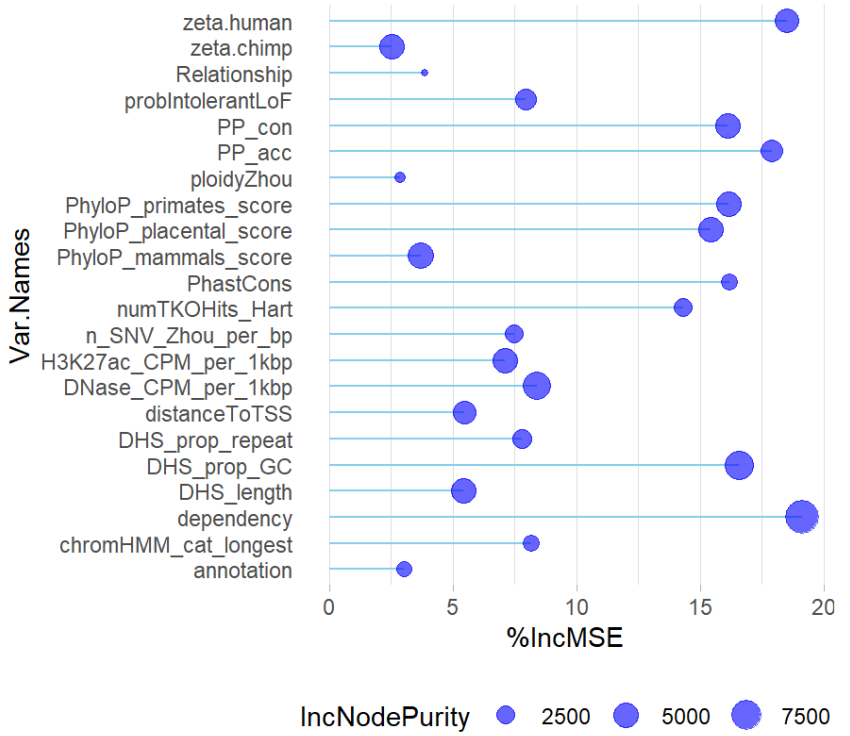
Linear Regression	MAE	1.28
-	RMSE	2.05
Gradient Boosting Machine	MAE	1.28
-	RMSE	2.03
Support Vector Regression	MAE	1.28
-	RMSE	2.09
RANSAC Regression	MAE	1.29
-	RMSE	2.06
Random Forest Regression	MAE	1.32
-	RMSE	2.07



Only Promoter-Enhancer Gene Interactions seen

I removed all genes that did not have both a strong candidate enhancer and an active promoter on the same TAD

Linear Regression	MAE	1.26
-	RMSE	1.91
Gradient Boosting Machine	MAE	1.27
-	RMSE	1.91
Support Vector Regression	MAE	1.27
-	RMSE	1.94
RANSAC Regression	MAE	1.29
-	RMSE	1.92
Random Forest Regression	MAE	1.31
-	RMSE	1.97



Only Promoter-Enhancer Gene Interactions

Relationship	Cand. Enhancer (Weak)	Cand. Enhancer (Strong)	Active Promoter	Inactive Promoter
Element within Gene	1472 (30%)	3315 (39.4%)	4550 (76%)	231 (36.7%)
Element-Gene Within TAD	2524 (51.4%)	3956 (47%)	1125 (18.8%)	292 (46.4%)
Gene off TAD	364 (7.4%)	364 (4.3%)	54 (0.9%)	41 (6.5%)
Gene Start off TAD	245 (5%)	335 (4%)	112 (1.9%)	26 (4.1%)
Gene End off TAD	302 (6.1%)	445 (5.3%)	131 (2.2%)	36 (5.7%)
Element in Multiple TADs	4 (0.1%)	6 (0.1%)	17 (0.3%)	3 (0.5%)

Takeaways

Hi-C information does seem to be important and highlights potentially misaligned elements, however, such effects are relatively uncommon across the dataset.

Beyond 20kb from the TSS, annotated active promoters increase in the likelihood of being out of TAD from their proximity assigned gene.

Our models do not seem to improve as a function of the increasing complexity of model type nor removal of potentially problematic elements.

Additional data would go a long way in confirming these conclusions and retesting our models.

Future Steps

Use the next iteration of CERES data to redevelop and test ML models, confirm general findings, and replicate the output

Complete more hyperparameter search mechanisms (e.g. GridSearch, Bayesian Optimization, Random Search) on each model type will each dataset. (probably on DCC as total time will be massive)

Scale with nested TADs to get at different nested effects