Introduction

Research Questions

We aim to answer two research questions with our analysis. Firstly, is predictive performance better with a smaller number of high quality data examples or a larger number of noisy data examples? And secondly, do the 6 additional features improve predictive performance beyond the original 9?

Data

We analyzed three datasets that vary in size and quality. The small dataset consists of 94,710 observations and has the highest quality data; the medium dataset consists of 167,895 observations with medium quality data; and the large dataset consists of 202,335 observations and is the lowest quality data with the most noise. Each dataset contains 15 features, and the outcome is a binary variable which indicates whether a pair of RNA strands interacted or not.

Methods

Data Preparation

The data was first split into a train and validation set, with 80% of our data being used to train our models and the latter 20% being used for model selection and hyperparameter tuning. Since our 15 features use different measurement scales, we standardized all the features in our data.

One challenge we ran into with our data was the severe imbalance across classes. In the small dataset 111 out of 94,710 observations were in class 1 (0.12%); in the medium dataset 347 out of 167,895 observations were in class 1 (0.21%); and in the large dataset 3530 out of 202,335 observations were in class 1 (1.74%). To address this imbalance, we used majority class downsampling. We systematically tested different downsampling values for the different models and recorded performance metrics in Table 1. We concluded that the best downsampling value for the small dataset is 500, for the medium dataset it is 1000, and for the large dataset it is 11,000.

We also have the option of upsampling the minority class using Synthetic Minority Over-sampling Technique (SMOTE) in our code, however this is currently unused. We did not explore the full range of possible oversampling values.

Model Selection

We explored multiple machine learning models, including Logistic Regression, Gradient Boosting, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Neural Networks (MLP). We focused on optimizing True Positive Rate (TPR) while maintaining a False Positive rate (FPR) below 0.05. We concluded that the best model for the small dataset is

Gradient Boost, and the best model for the medium and large datasets is Random Forest (Table 1).

We also explored using different classification thresholds for each model. Using the optimal downsampling values, we adjusted the thresholds such that we got the highest TPR for each model where the FPR was still below 0.05. We conclude that the best models are still Gradient Boost for the small dataset and Random Forest for the medium and large dataset (Table 2).

Results

Model Performance

Our analysis revealed that the small dataset consistently achieved better performance than the medium and large datasets, indicating that predictive performance is better with a smaller number of high quality data examples. The optimal performance on the testing data was achieved on the small dataset using Gradient Boosting with a downsampling value of 500, which attained a mean TPR of 0.5478 and FPR of 0.0604 for the mean threshold of 0.40 (Table 3).

Among the evaluated models, Gradient Boosting and Random Forest consistently had better performance. Looking at the ROC curves and the barplots with the optimal TPR's for FPR < 0.05, we observe that the Gradient Boost and Random Forest models performed similarly across datasets (Graphs 1 and 2). In the small dataset Gradient Boost was better, whereas for the medium and large datasets Random Forest models were marginally better.

Feature Importance

We analyzed whether the additional 6 predictors improved model performance by running the final models on the testing data with 9 vs. 15 predictors, using the optimal threshold obtained earlier (Table 3). Our analysis found that the small dataset performed better with all 15 predictors, while the medium and large dataset performed better when using only the original 9 predictors.

Analysis using SHAP (SHapley Additive exPlanations) revealed the order of feature importance for each dataset (Graph 4). The SHAP graphs indicate which predictors contribute the most to the model's performance. For all three datasets, the features "Distance (either)", "Number of homologs", and "Correlation Expression" were among the top 5 most influential features. And for all three datasets, the features "Is same component", "Are there homologs?", "Upstream (either)", "Upstream (same)", and "Component score" were the least influential features.

Appendix

Table 1
Comparing six models with different downsampling values.

Small Dataset:

	Log Reg	Gradient Boost	Random Forest	SVM	KNN	NN
300	TPR: 0.4497 FPR: 0.0747	TPR: 0.4902 FPR: 0.1223	TPR: 0.4806 FPR: 0.0955	TPR: 0.4568 FPR: 0.0509	TPR: 0.4765 FPR: 0.1120	TPR: 0.5154 FPR: 0.0816
400	TPR: 0.2888 FPR: 0.0406	TPR: 0.3684 FPR: 0.0778	TPR: 0.4083 FPR: 0.0623	TPR: 0.2911 FPR: 0.0350	TPR: 0.3131 FPR: 0.0606	TPR: 0.4199 FPR: 0.0657
500	TPR: 0.3170 FPR: 0.0396	TPR: 0.4244 FPR: 0.0256	TPR: 0.4021 FPR: 0.0416	TPR: 0.2313 FPR: 0.0174	TPR: 0.3468 FPR: 0.0512	TPR: 0.4436 FPR: 0.0626
600	TPR: 0.2169 FPR: 0.0149	TPR: 0.3637 FPR: 0.0563	TPR: 0.3285 FPR: 0.0184	TPR: 0.1951 FPR: 0.0050	TPR: 0.3393 FPR: 0.0399	TPR: 0.4281 FPR: 0.0366
750	TPR: 0.2326 FPR: 0.0189	TPR: 0.3556 FPR: 0.0270	TPR: 0.3707 FPR: 0.0135	TPR: 0.1783 FPR: 0.0026	TPR: 0.2625 FPR: 0.0337	TPR: 0.3732 FPR: 0.0253
1000	TPR: 0.1739 FPR: 0.0146	TPR: 0.2475 FPR: 0.0204	TPR: 0.2778 FPR: 0.0195	TPR: 0.0760 FPR: 0.0009	TPR: 0.1629 FPR: 0.0293	TPR: 0.2887 FPR: 0.0186

Medium Dataset:

	Log Reg	Gradient Boost	Random Forest	SVM	KNN	NN
1000	TPR: 0.3245 FPR: 0.0748	TPR: 0.4342 FPR: 0.0779	TPR: 0.4033 FPR: 0.0499	TPR: 0.2398 FPR: 0.0222	TPR: 0.3673 FPR: 0.1181	TPR: 0.4731 FPR: 0.1176
1100	TPR: 0.3114 FPR: 0.0649	TPR: 0.4054 FPR: 0.0685	TPR: 0.3493 FPR: 0.0429	TPR: 0.1968 FPR: 0.0200	TPR: 0.3003 FPR: 0.1066	TPR: 0.4955 FPR: 0.1603
1200	TPR: 0.3100 FPR: 0.0604	TPR: 0.4411 FPR: 0.0612	TPR: 0.3632 FPR: 0.0411	TPR: 0.1532 FPR: 0.0167	TPR: 0.3284 FPR: 0.0744	TPR: 0.4617 FPR: 0.1329
1300	TPR: 0.2606 FPR: 0.0397	TPR: 0.3870 FPR: 0.0517	TPR: 0.3196 FPR: 0.0358	TPR: 0.1323 FPR: 0.0076	TPR: 0.2663 FPR: 0.0738	TPR: 0.4693 FPR: 0.1383
1400	TPR: 0.2040 FPR: 0.0379	TPR: 0.3553 FPR: 0.0436	TPR: 0.2657 FPR: 0.0228	TPR: 0.0988 FPR: 0.0078	TPR: 0.2457 FPR: 0.0673	TPR: 0.4596 FPR: 0.1030

1500	TPR: 0.2036 FPR: 0.0345	TPR: 0.3175 FPR: 0.0359	TPR: 0.3012 FPR: 0.0253	TPR: 0.1118 FPR: 0.0045	TPR: 0.2404 FPR: 0.0586	TPR: 0.3884 FPR: 0.1046

Large Dataset:

	Log Reg	Gradient Boost	Random Forest	SVM	KNN	NN
10000	TPR: 0.1385 FPR: 0.0553	TPR: 0.3961 FPR: 0.0733	TPR: 0.3860 FPR: 0.0577	TPR: 0.2149 FPR: 0.0402	TPR: 0.4213 FPR: 0.1224	TPR: 0.4642 FPR: 0.1393
11000	TPR: 0.0884 FPR: 0.0328	TPR: 0.3536 FPR: 0.0586	TPR: 0.3304 FPR: 0.0477	TPR: 0.1395 FPR: 0.0243	TPR: 0.4074 FPR: 0.1070	TPR: 0.4116 FPR: 0.0875
12000	TPR: 0.0560 FPR: 0.0247	TPR: 0.3056 FPR: 0.0483	TPR: 0.3007 FPR: 0.0435	TPR: 0.0840 FPR: 0.0145	TPR: 0.3589 FPR: 0.0935	TPR: 0.4103 FPR: 0.1180
13000	TPR: 0.0358 FPR: 0.0170	TPR: 0.2890 FPR: 0.0391	TPR: 0.2898 FPR: 0.0346	TPR: 0.0546 FPR: 0.0076	TPR: 0.3379 FPR: 0.0868	TPR: 0.4204 FPR: 0.1124
14000	TPR: 0.0234 FPR: 0.0107	TPR: 0.2677 FPR: 0.0377	TPR: 0.2771 FPR: 0.0333	TPR: 0.0360 FPR: 0.0047	TPR: 0.3409 FPR: 0.0829	TPR: 0.3945 FPR: 0.1044
15000	TPR: 0.0191 FPR: 0.0089	TPR: 0.2432 FPR: 0.0276	TPR: 0.2491 FPR: 0.0257	TPR: 0.0182 FPR: 0.0019	TPR: 0.3009 FPR: 0.0738	TPR: 0.4066 FPR: 0.1028

Note: TPR and FPR values are the mean values across 5 different random seeds.

Table 2

Comparing four best models using the optimal threshold for each.

Dataset	Gradient Boost	Random Forest	KNN	NN
Small, 500	TPR: 0.5058 FPR: 0.0396 Thresh: 0.40	TPR: 0.4737 FPR: 0.0299 Thresh: 0.40	TPR: 0.2442 FPR: 0.0157 Thresh: 0.51	TPR: 0.3583 FPR: 0.0434 Thresh: 0.62
Med, 1000	TPR: 0.3405 FPR: 0.0385 Thresh: 0.62	TPR: 0.3648 FPR: 0.0395 Thresh: 0.54	TPR: 0.1837 FPR: 0.0273 Thresh: 0.60	TPR: 0.2913 FPR: 0.0424 Thresh: 0.80
Large, 11000	TPR: 0.2844 FPR: 0.0405	TPR: 0.3014 FPR: 0.0399	TPR: 0.1892 FPR: 0.0365	TPR: 0.2625 FPR: 0.0432

	Thresh: 0.54	Thresh: 0.53	Thresh: 0.60	Thresh: 0.63

Note: TPR, FPR, and Threshold values are the mean values across 5 different random seeds.

Table 3

Comparing the optimal models with the first 9 features to the same models with all 15 features on test data (using optimal thresholds from Table 2).

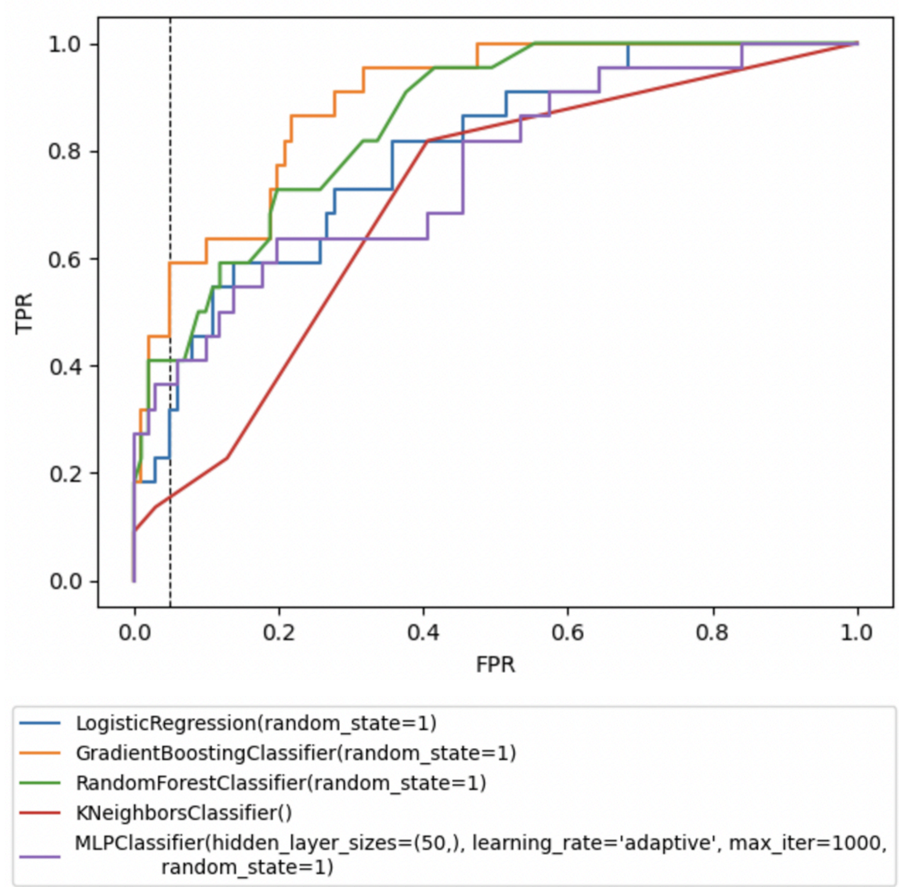
	9 features	All features
Small (Gradient Boost, 500, 0) Threshold: 0.40	TPR: 0.4695 FPR: 0.0604	TPR: 0.5478 FPR: 0.0604
Medium (Random Forest, 1000, 0) Threshold: 0.54	TPR: 0.4829 FPR: 0.0687	TPR: 0.3804 FPR: 0.0505
Large (Random Forest, 11000, 0) Threshold: 0.53	TPR: 0.2178 FPR: 0.0678	TPR: 0.1270 FPR: 0.0317

Note: TPR and FPR values are the mean values across 5 different random seeds.

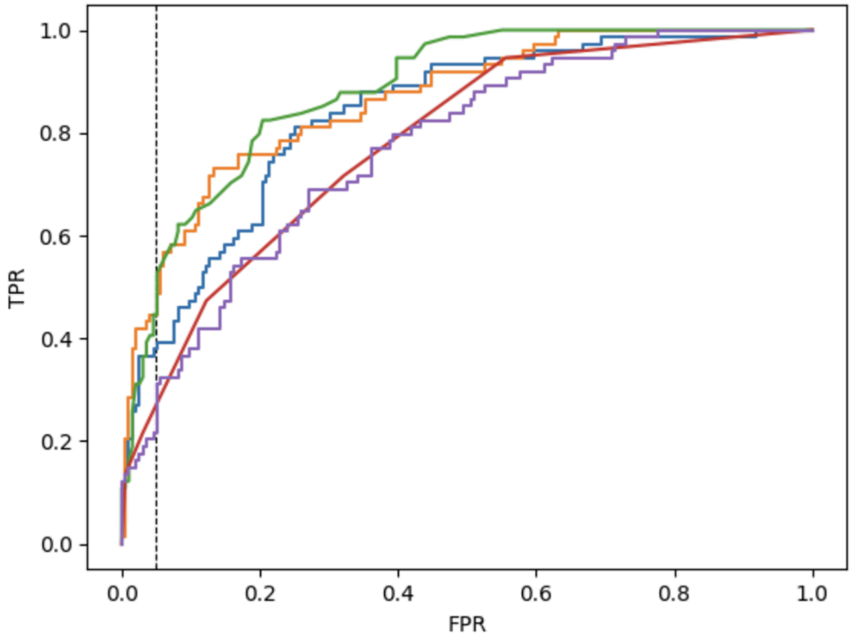
Graphs 1

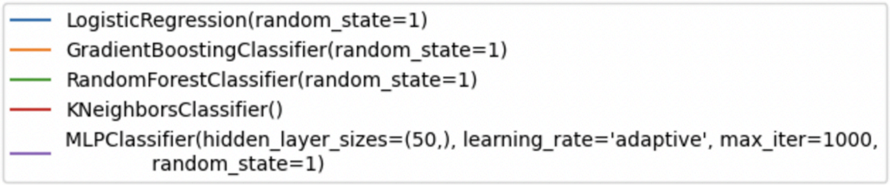
ROC Curves on validation data.

Small dataset:

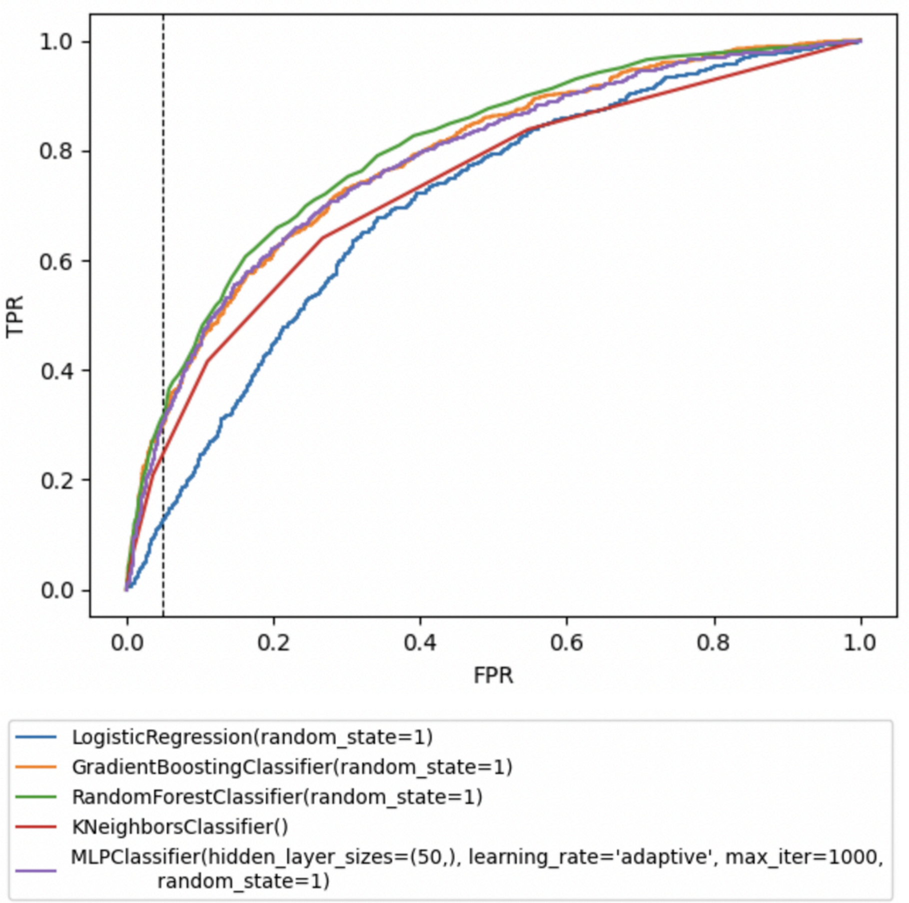


Medium dataset:



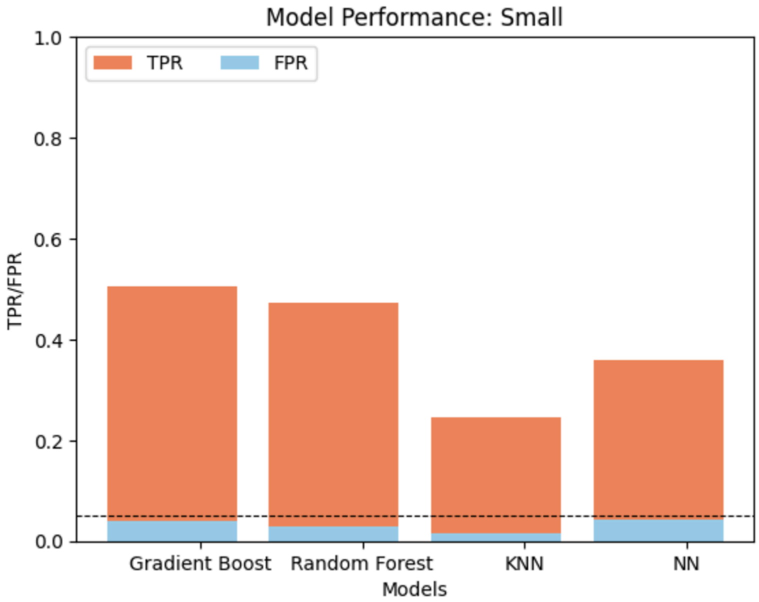


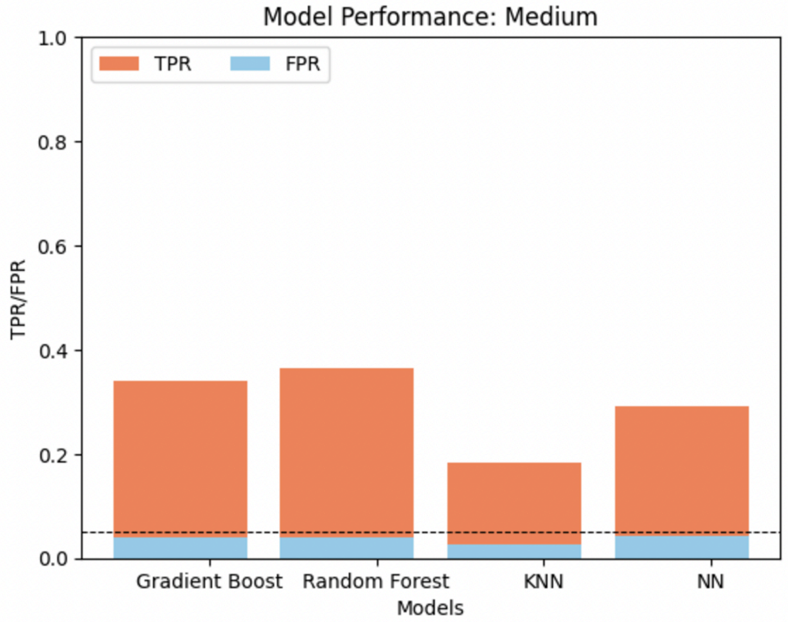
Large dataset:

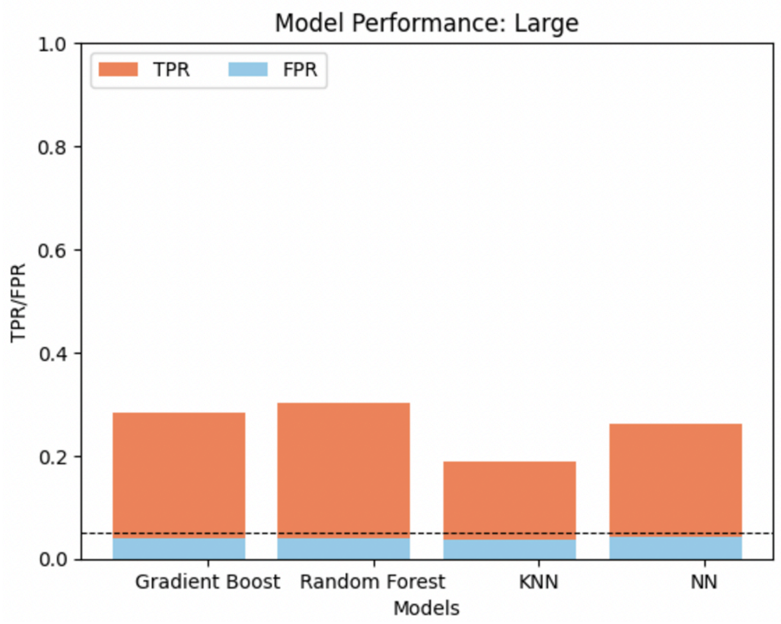


Graphs 2

Bar plots of model performance on validation data. TPR overlaid with FPR.

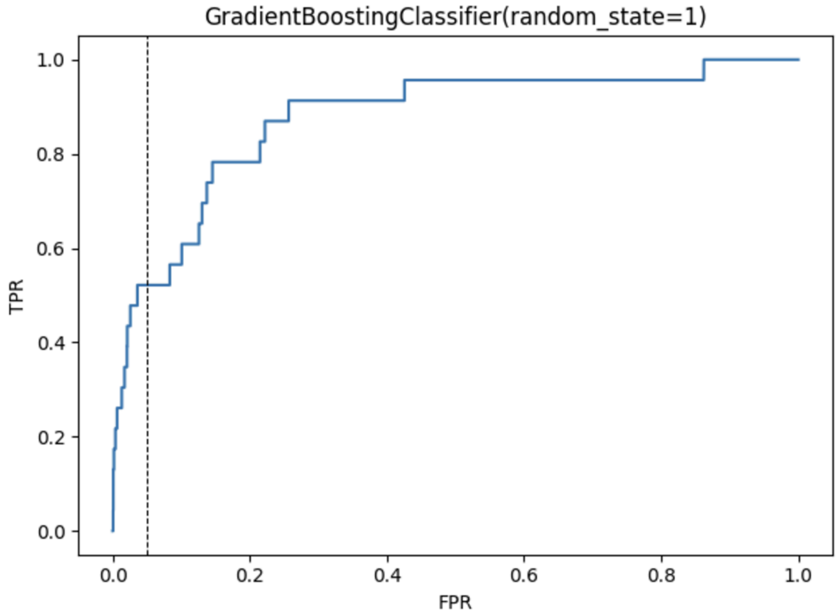

Model Performance: Small
TPR
FPR
1.0
0.8
0.6
0.4
0.2
0.0
TPR/FPR
Gradient Boost
Random Forest
KNN
NN
Models


Model Performance: Medium
TPR
FPR
1.0
0.8
0.6
0.4
0.2
0.0
TPR/FPR
Gradient Boost
Random Forest
KNN
NN
Models

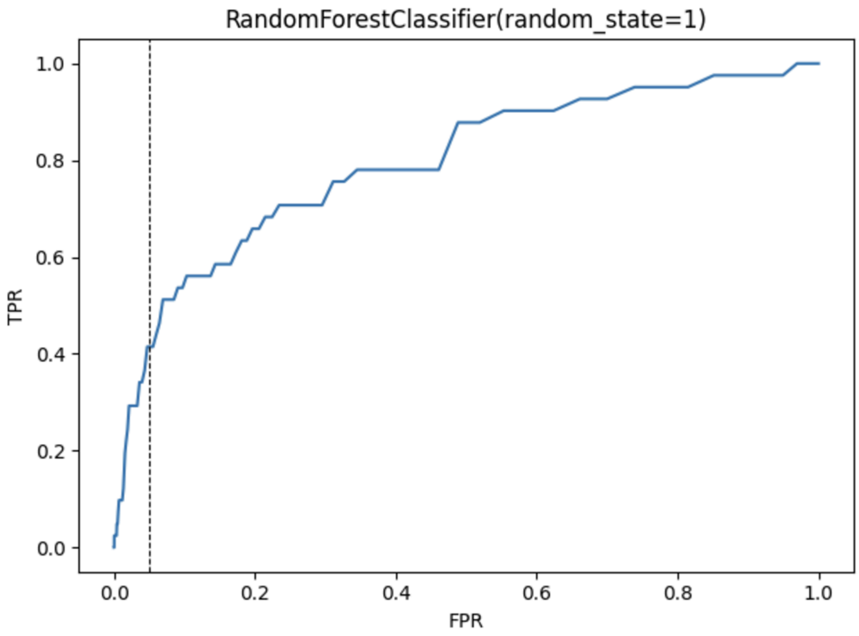


Graphs 3

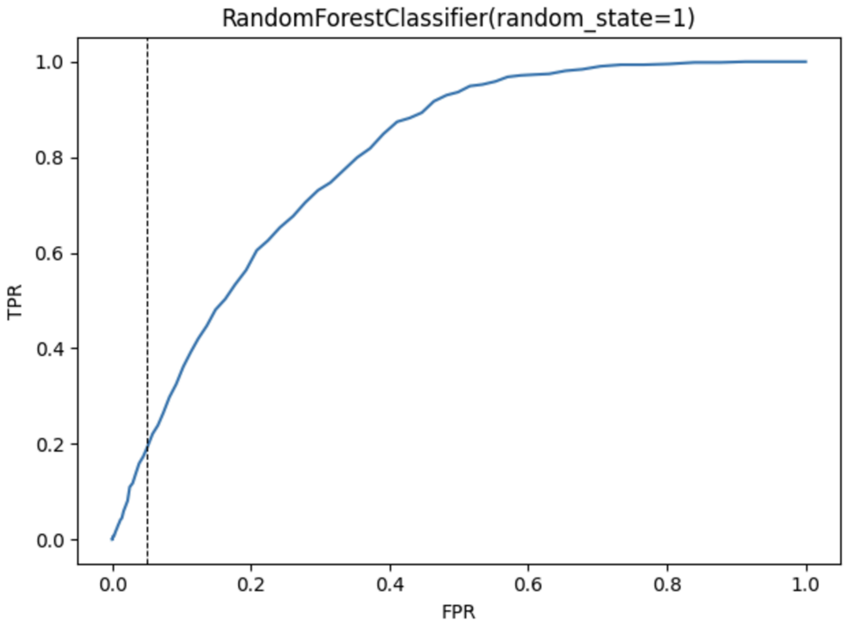
ROC curves of optimal models on testing data with all features.
Small testing data:



Medium testing data:

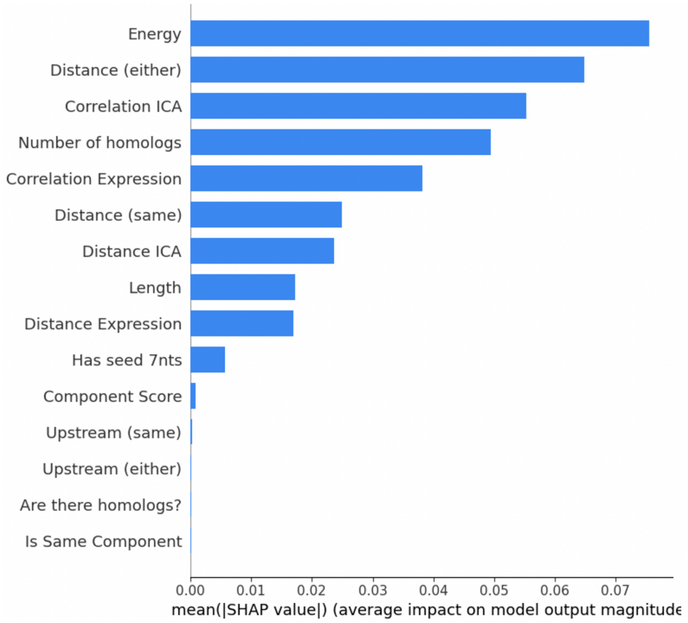


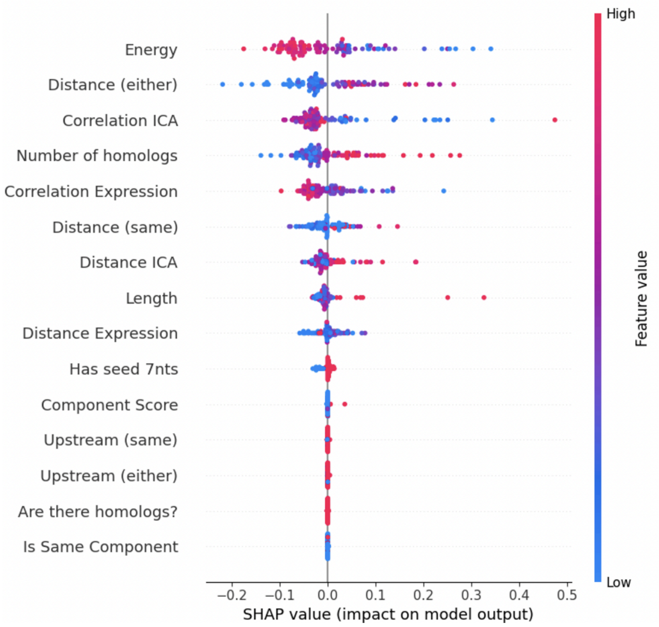
Large testing data:



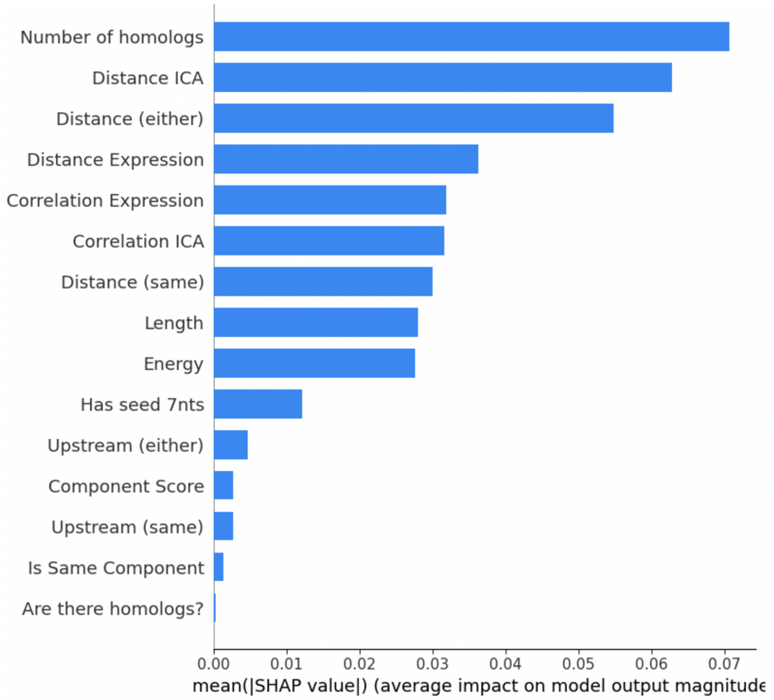
Graphs 4

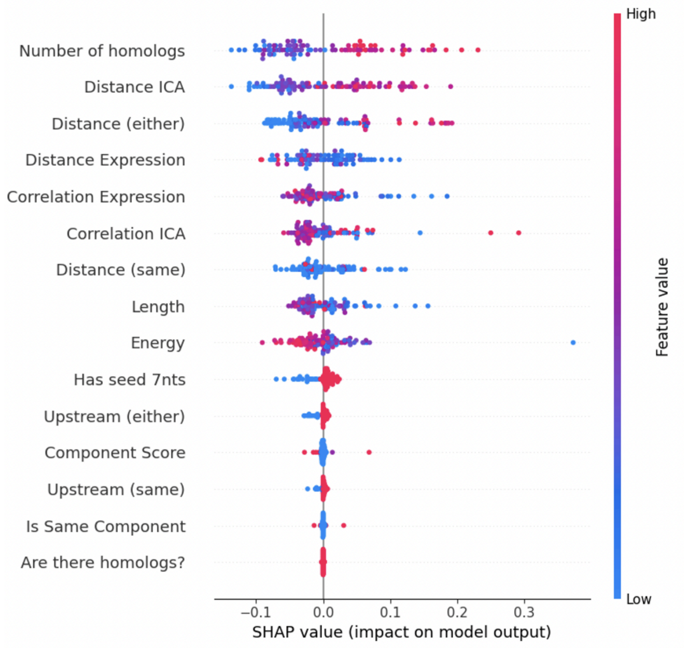
Small Dataset:



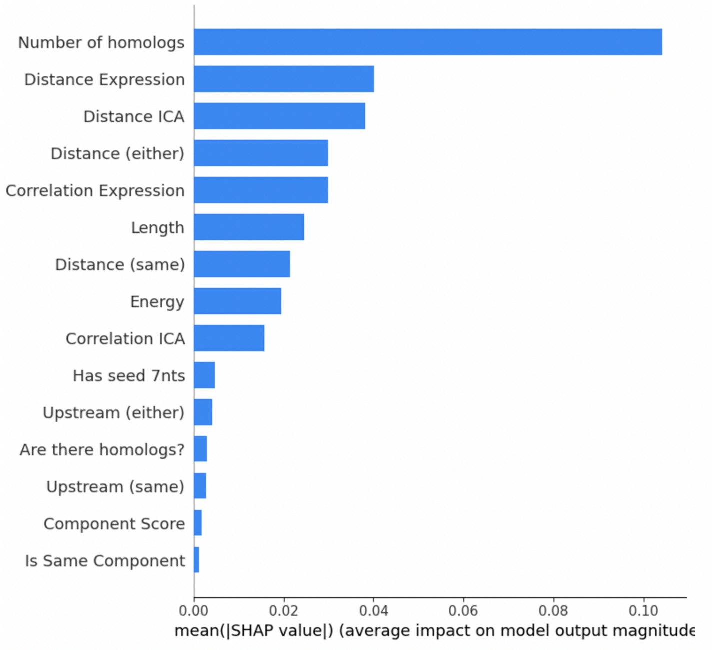


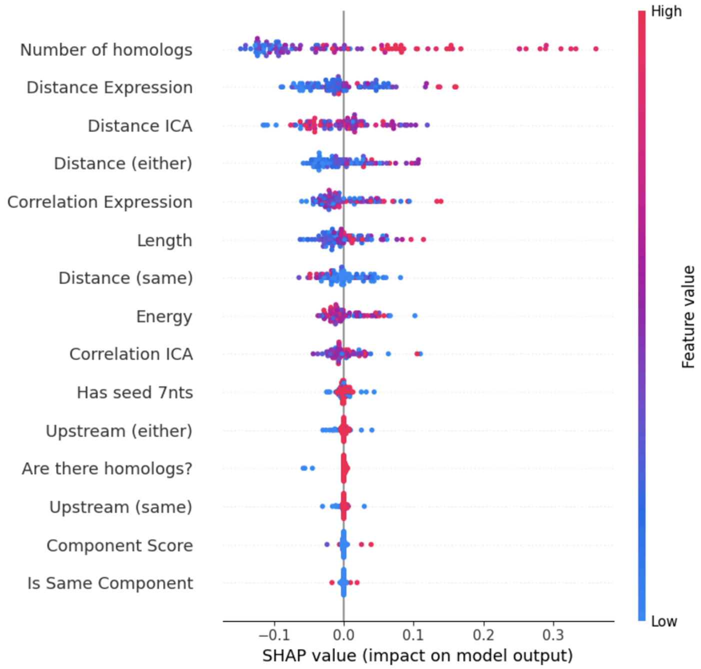
Medium dataset:





Large dataset:


Number of homologs
Distance Expression
Distance ICA
Distance (either)
Correlation Expression
Length
Distance (same)
Energy
Correlation ICA
Has seed 7nts
Upstream (either)
Are there homologs?
Upstream (same)
Component Score
Is Same Component
0.00
0.02
0.04
0.06
0.08
0.10
mean(|SHAP value|) (average impact on model output magnitude


Number of homologs
Distance Expression
Distance ICA
Distance (either)
Correlation Expression
Length
Distance (same)
Energy
Correlation ICA
Has seed 7nts
Upstream (either)
Are there homologs?
Upstream (same)
Component Score
Is Same Component
−0.1
0.0
0.1
0.2
0.3
SHAP value (impact on model output)
High
Low
Feature value