# Merck Challenge
# -- Predicting Molecular Activity

Members: Chen Liang, Liyuan Xie, Tirth Patel, Xinxin Mo, William Xi, Yifei Yan

# Agenda

1 Project Preparation

2 Machine Learning Models and Performance Metrics

3 User Interface

4 Challenges and Bottlenecks

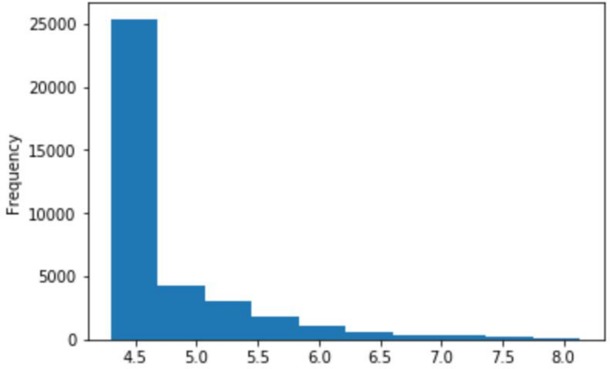
5 Scope for Future Improvement

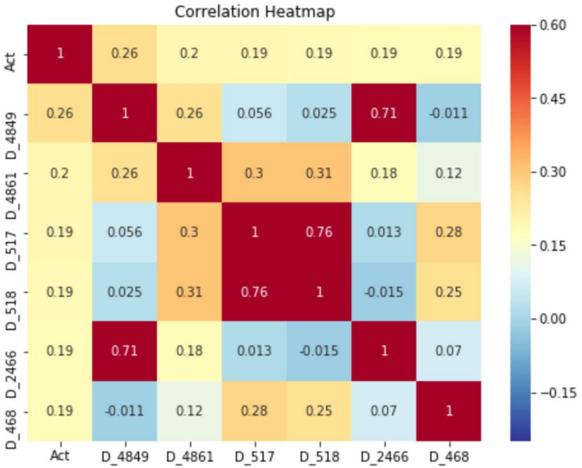
# Project Overview

- Drug discovery is the area of research and development that used the most amount of time and money.
- The time frame can easily range from 3 to 20 years and costs can range between several billion to tens of billions of dollars for the research teams to find the molecules that highly active toward the target through myriad experiments.
- This project aims to utilize machine learning as a cost-effective tool for predicting biological activities of different molecules, both on- and off-target, given numerical descriptors generated from their chemical structures.

# Exploratory Data Analysis

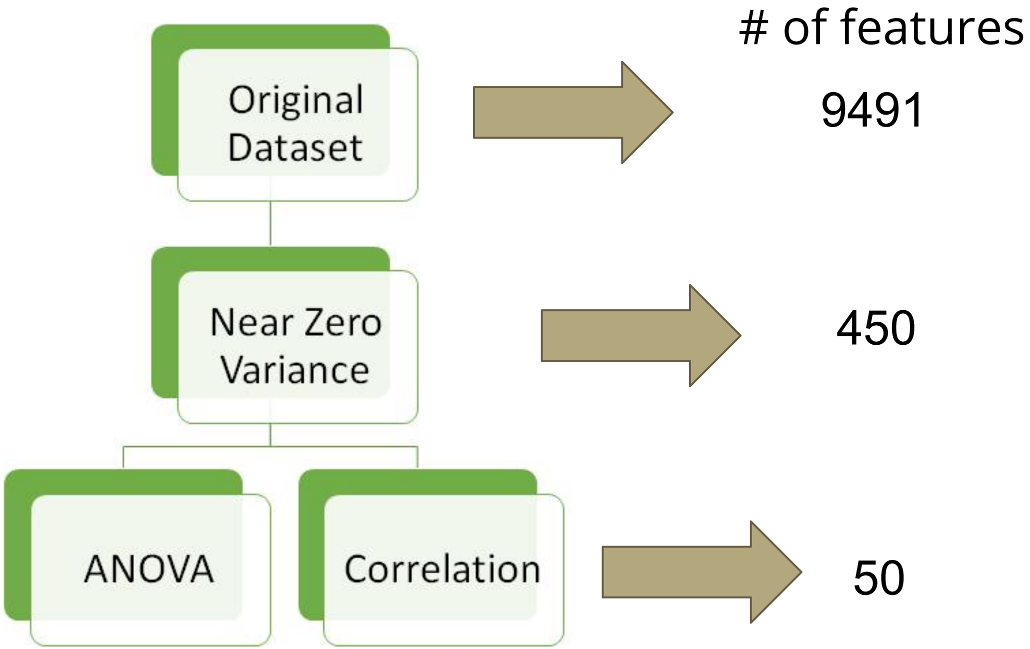
- 9,491 features (fingerprints) and 37,241 rows (molecules)
- No Missing Value
- Imbalanced Distribution of Molecular Activity
  - Mean: 4.69
  - Min: 4.3
  - Max: 8.13
  - Std: 0.65
- Correlation between fingerprints and the molecular activity




Correlation Heatmap

| | Act | D_4849 | D_4861 | D_517 | D_518 | D_2466 | D_468 |
|---|---|---|---|---|---|---|---|
| Act | 1 | 0.26 | 0.2 | 0.19 | 0.19 | 0.19 | 0.19 |
| D_4849 | 0.26 | 1 | 0.26 | 0.056 | 0.025 | 0.71 | -0.011 |
| D_4861 | 0.2 | 0.26 | 1 | 0.3 | 0.31 | 0.18 | 0.12 |
| D_517 | 0.19 | 0.056 | 0.3 | 1 | 0.76 | 0.013 | 0.28 |
| D_518 | 0.19 | 0.025 | 0.31 | 0.76 | 1 | -0.015 | 0.25 |
| D_2466 | 0.19 | 0.71 | 0.18 | 0.013 | -0.015 | 1 | 0.07 |
| D_468 | 0.19 | -0.011 | 0.12 | 0.28 | 0.25 | 0.07 | 1 |

# Dimensionality Reduction



Original Dataset

Near Zero Variance

ANOVA

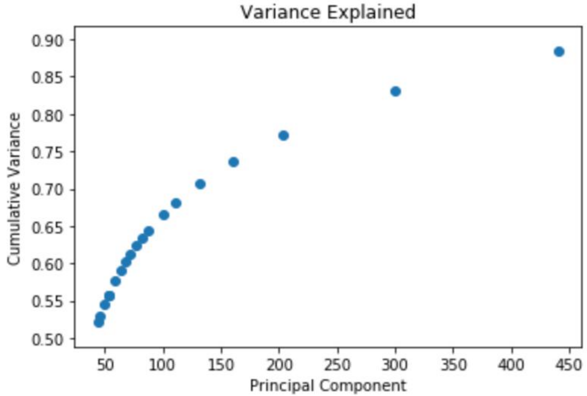
Correlation

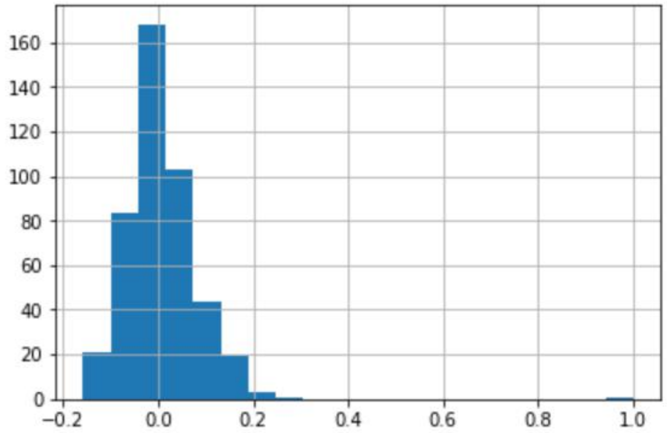
# of features

9491

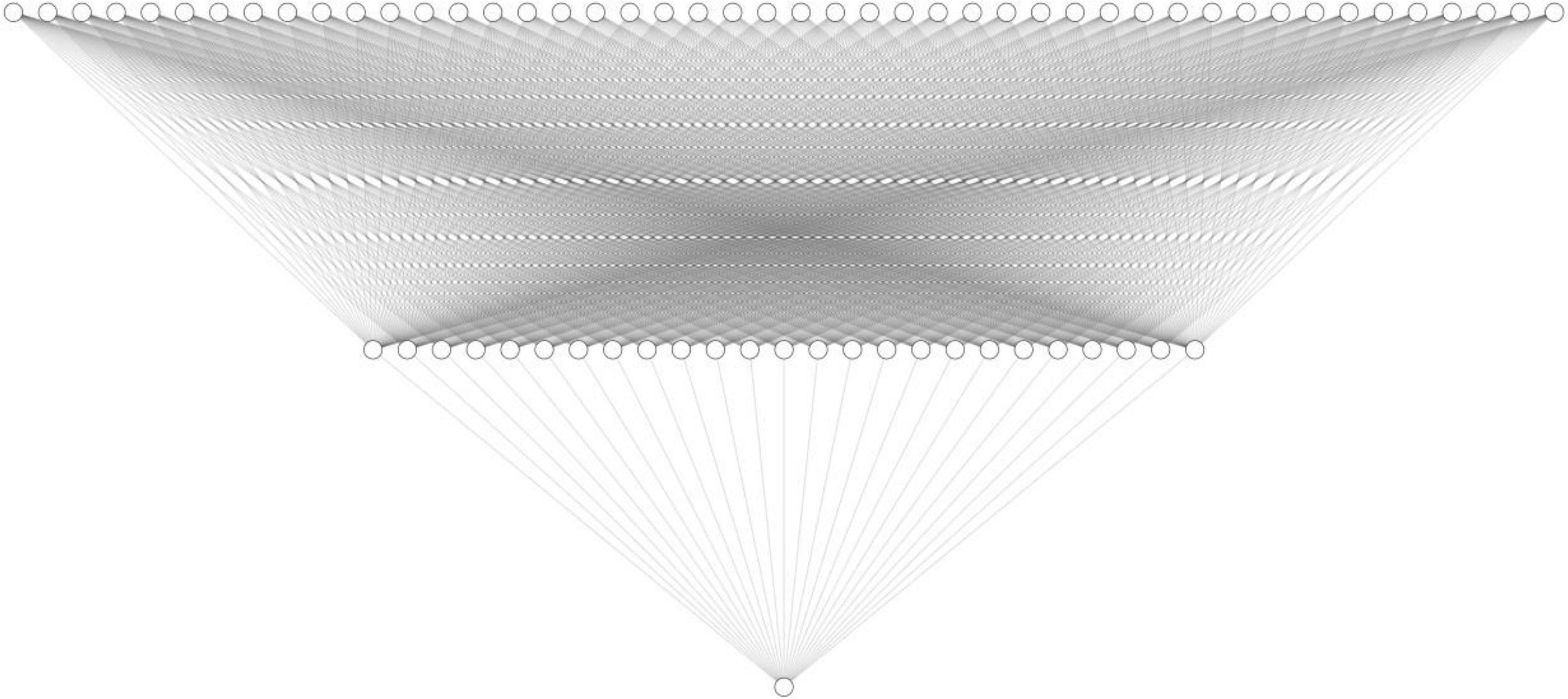
450

50

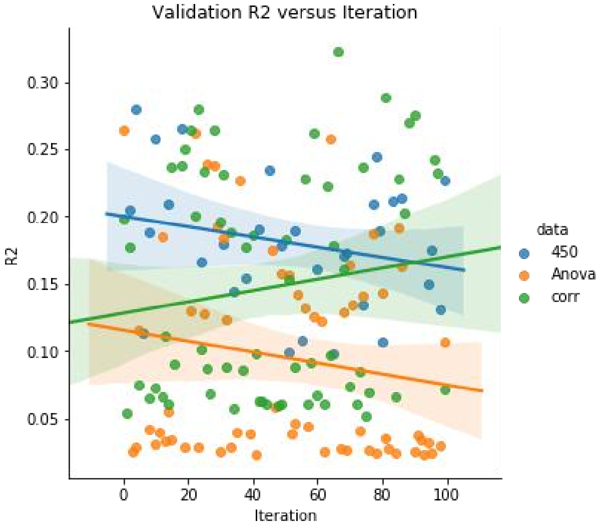


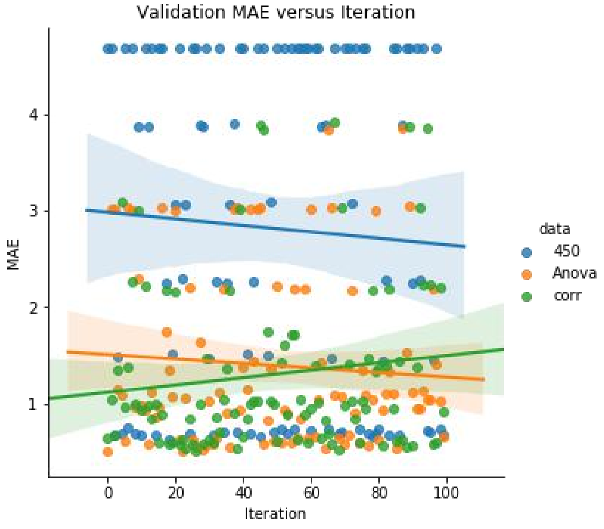


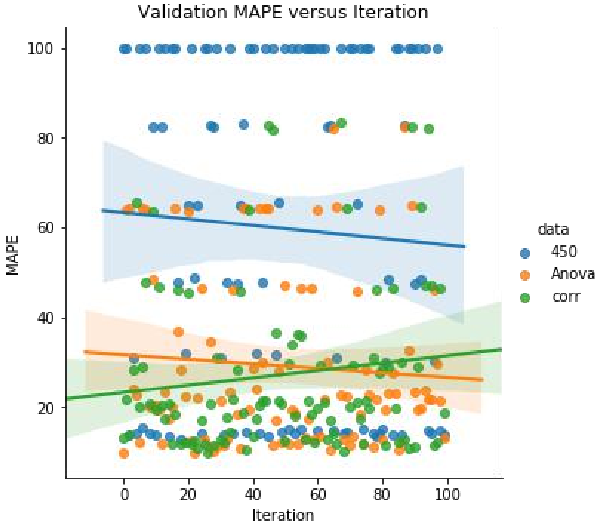
# Tuned Models - Artificial Neural Networks



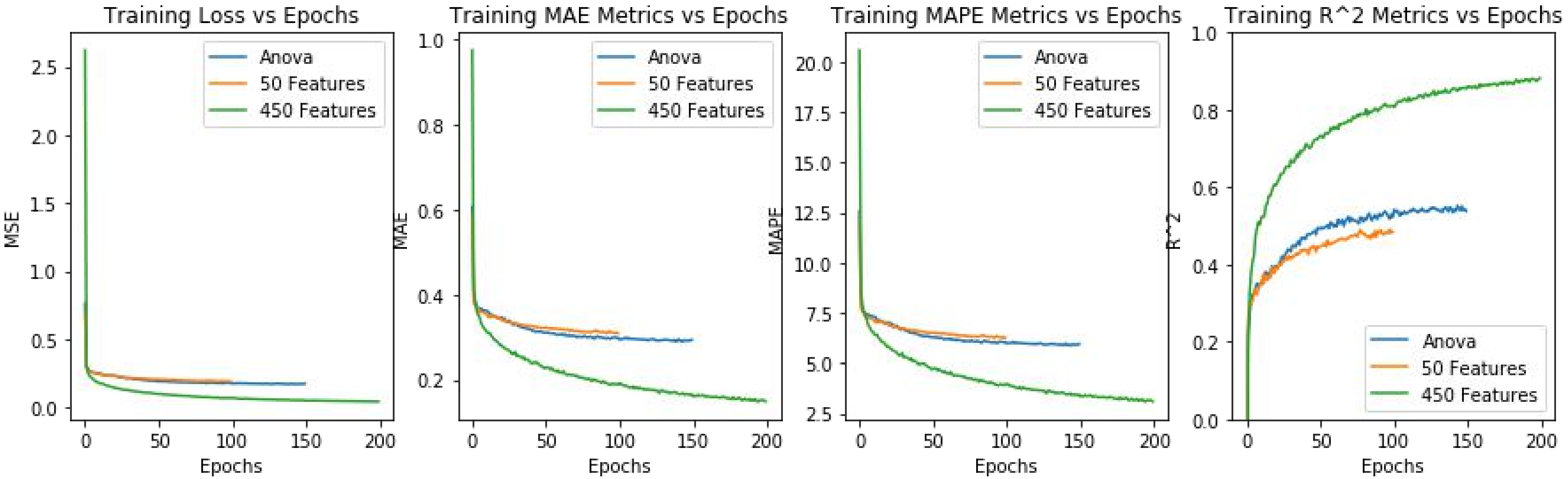
# Tuned Models - Artificial Neural Networks







# Tuned Models - Artificial Neural Networks



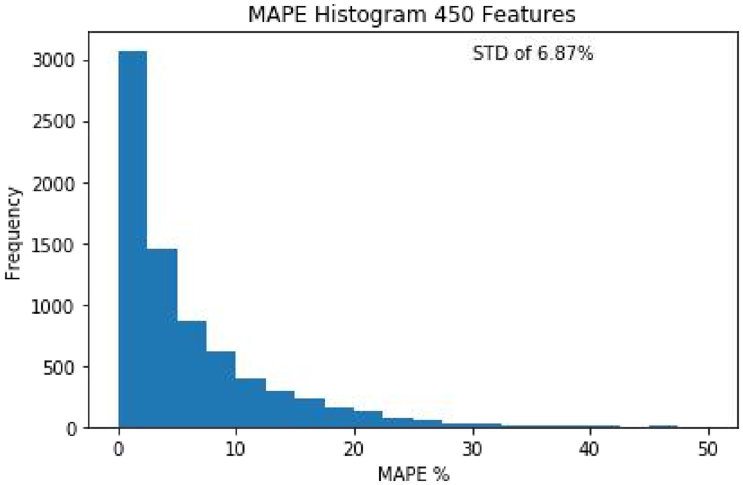
Training Loss vs Epochs

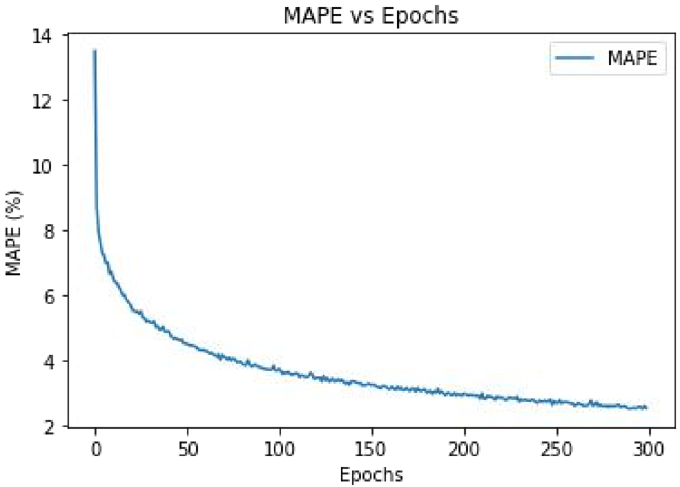
Training MAE Metrics vs Epochs

Training MAPE Metrics vs Epochs

Training R^2 Metrics vs Epochs

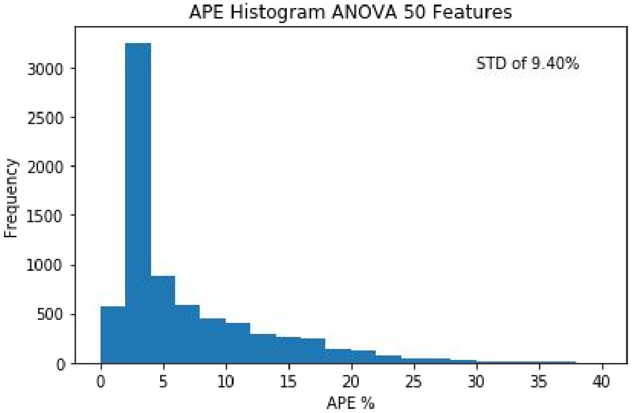
# Untuned Model - Artificial Neural Networks

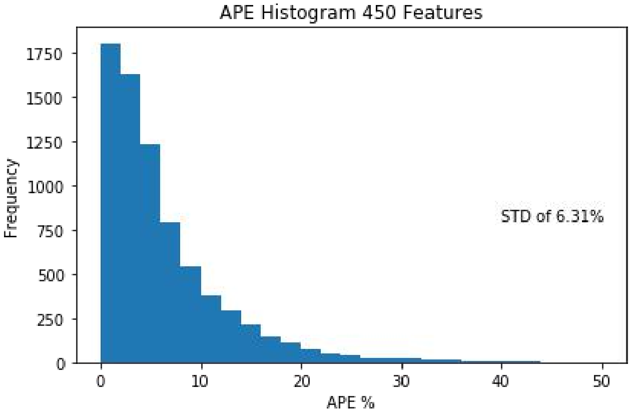


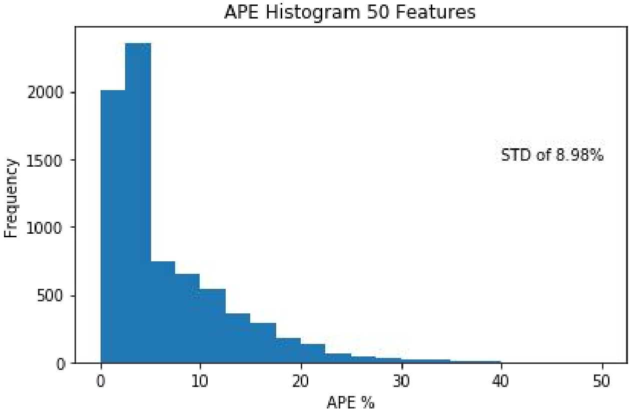


| Data | $R^2$ | MAPE | MAE | Batch_size | No_epochs | Learning Rate |
|---|---|---|---|---|---|---|
| 450 Features | 0.313 | 6.88% | 0.332 | 300 | 300 | 0.001 |

# Tuned Models - Artificial Neural Networks

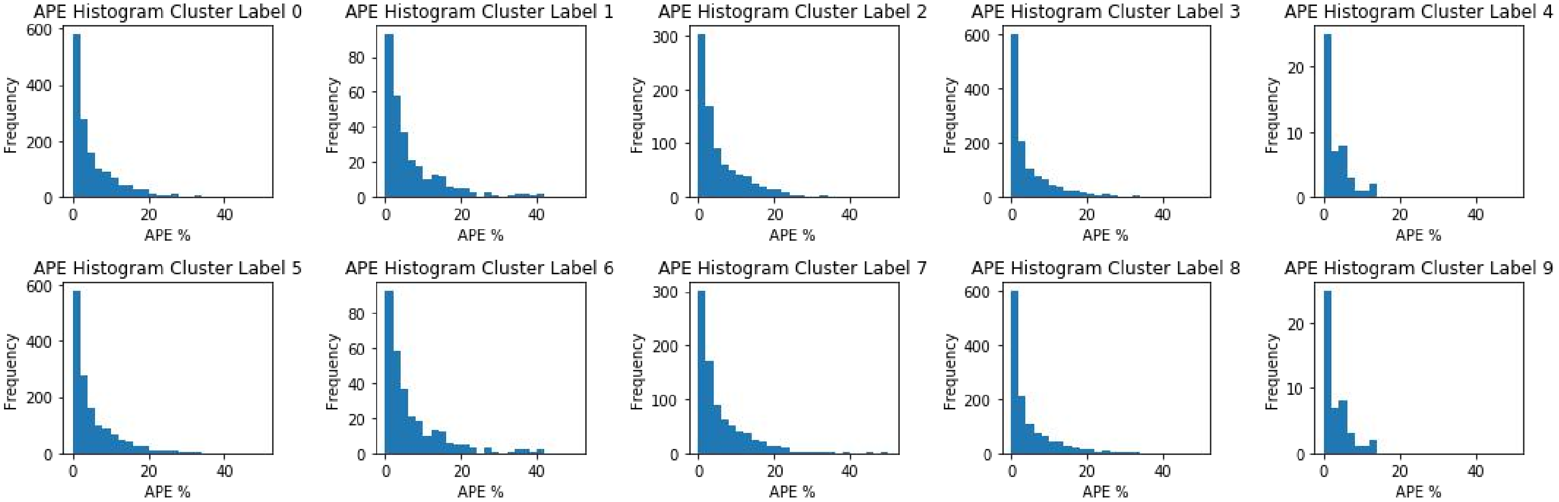

APE Histogram ANOVA 50 Features

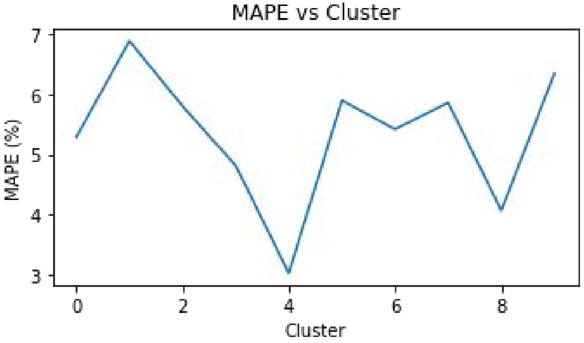

APE Histogram 450 Features

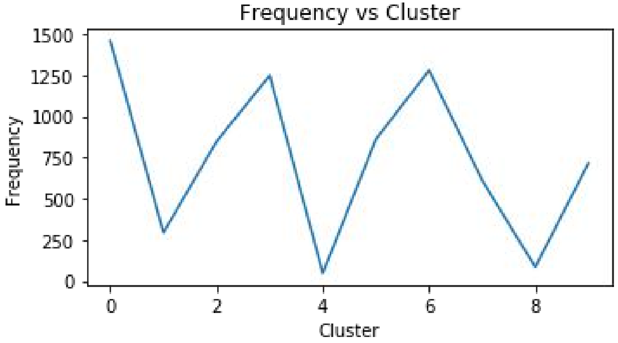

APE Histogram 50 Features

| Data | $R^2$ | MAPE | MAE | Batch_size | No_epochs | Learning Rate |
|---|---|---|---|---|---|---|
| 450 Features | 0.432 | 6.31% | 0.307 | 60 | 200 | 0.001 |
| Anova 50 Features | 0.332 | 6.82% | 0.336 | 60 | 150 | 0.01 |
| 50 Features | 0.394 | 6.50% | 0.323 | 50 | 100 | 0.01 |

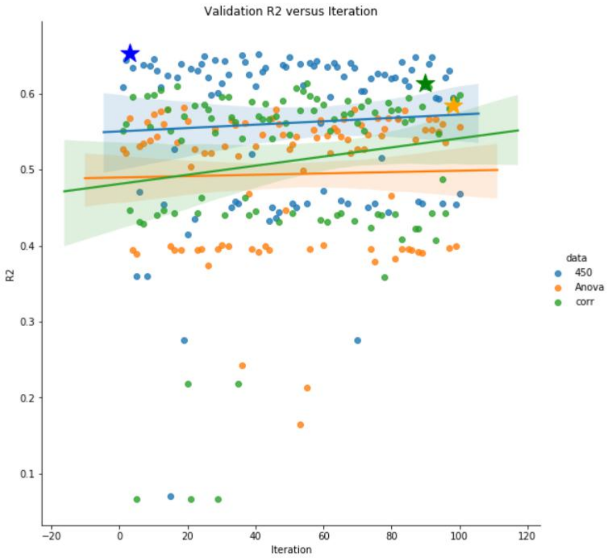
# Tuned Models - Artificial Neural Networks



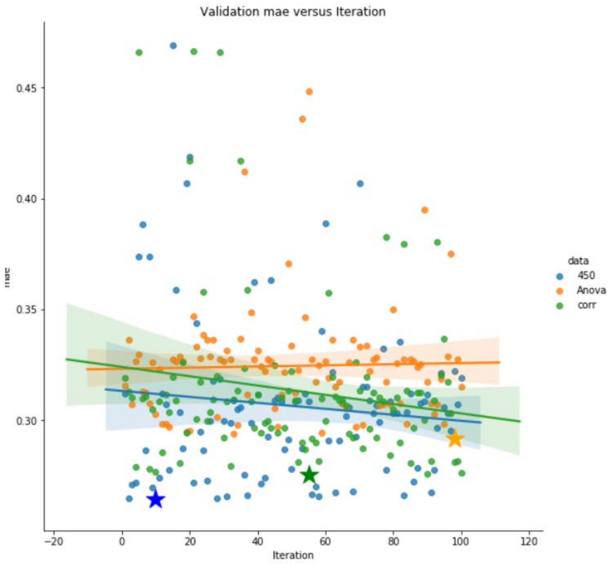




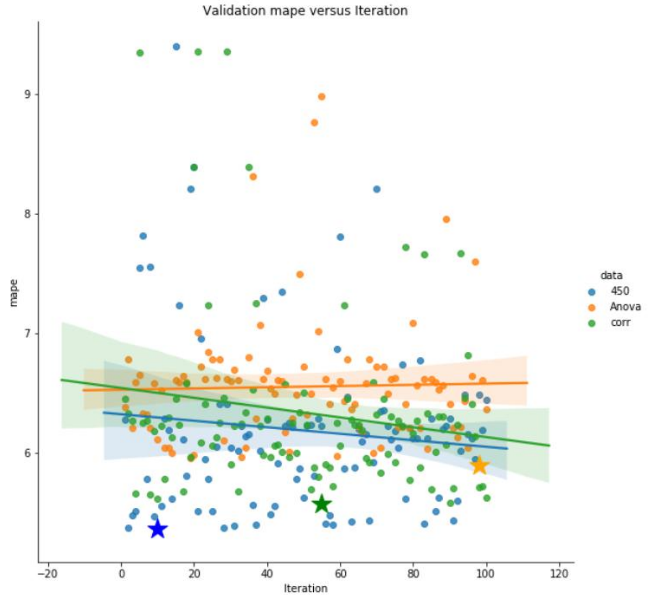
# Random Forest Model



Validation R2 versus Iteration



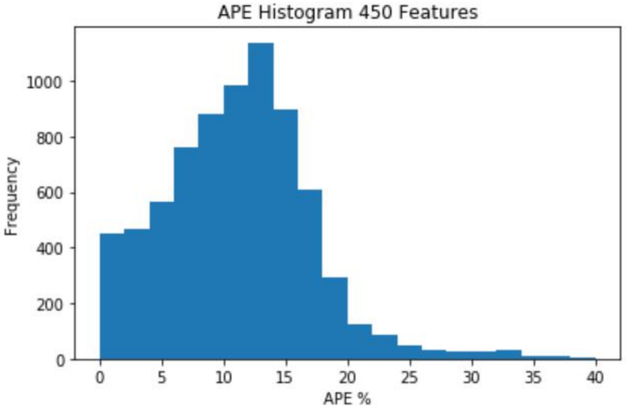
Validation mae versus Iteration

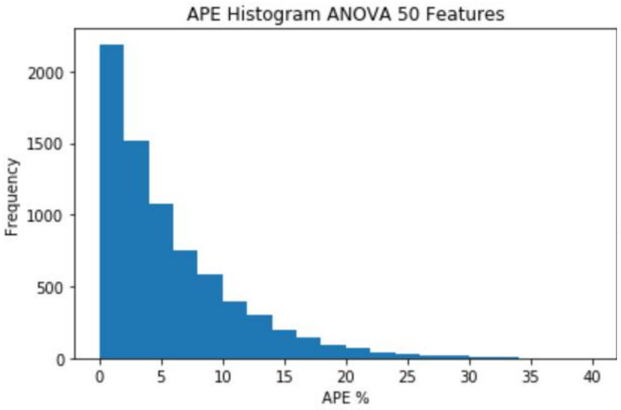


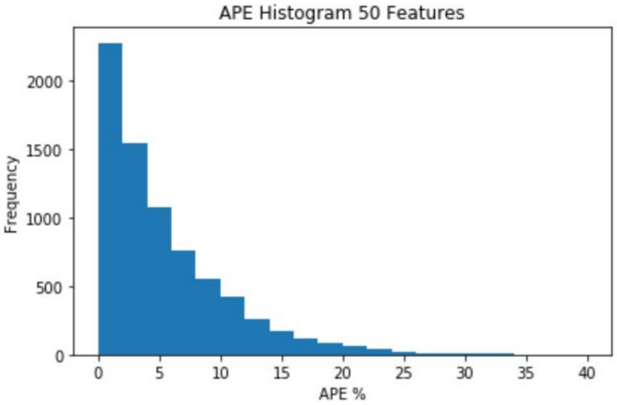
Validation mape versus Iteration

| | Dataset | MAPE | MAE | R² |
|---|---|---|---|---|
| Baseline Model | 450 features | 5.50% → 5.48% | 0.27 → 0.27 | 0.65 → 0.65 |
| | 50 features (corr) | 5.55% → 5.58% | 0.54 → 0.28 | 0.61 → 0.61 |
| | 50 features (Anova) | 5.77% → 5.89% | 0.29 → 0.29 | 0.58 → 0.58 |

# Random Forest

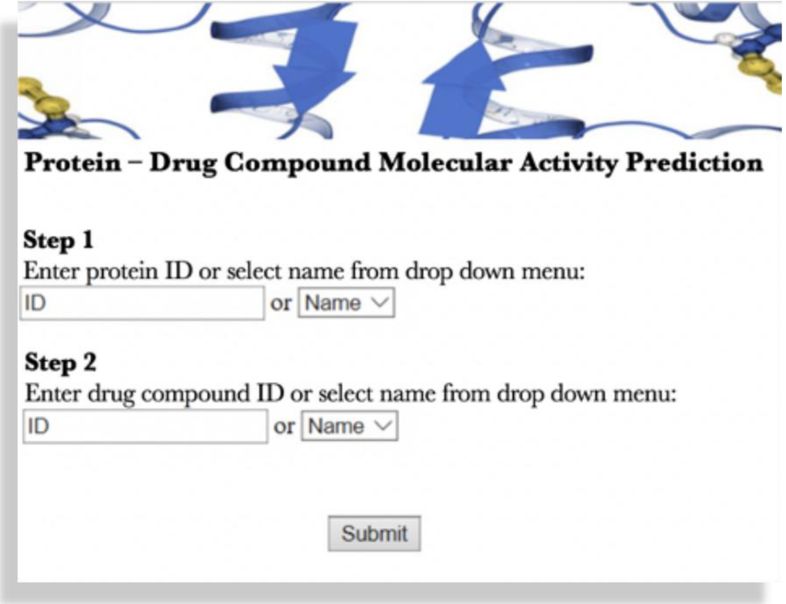






| Data | $R^2$ | MAPE | MAE | n_estimators | min_samples_leaf | min_samples_split |
|---|---|---|---|---|---|---|
| 450 Features | 0.101 | 11.1% | 0.530 | 500 | 1 | 2 |
| Anova 50 Features | 0.595 | 5.70% | 0.280 | 500 | 1 | 2 |
| 50 Features | 0.631 | 5.70% | 0.270 | 360 | 2 | 5 |

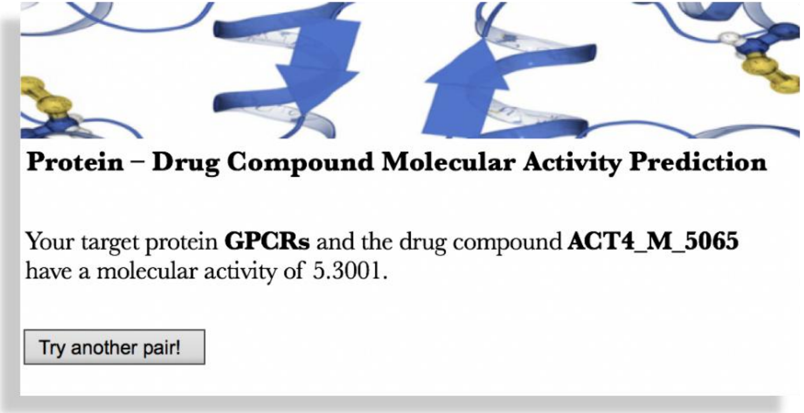
# User Interface



**Protein – Drug Compound Molecular Activity Prediction**

**Step 1**
Enter protein ID or select name from drop down menu:
ID or Name

**Step 2**
Enter drug compound ID or select name from drop down menu:
ID or Name



**Protein – Drug Compound Molecular Activity Prediction**

Your target protein **GPCRs** and the drug compound **ACT4_M_5065** have a molecular activity of 5.3001.

# Bottlenecks and Challenges

- Data Sampling
  - Imbalance dataset
- Data Dimensionality
  - High dimensionality while need to preserve information
- Data Sparsity
  - Overfitting while using more features
  - A lack of generalization
- Difficult to improve performance of ANN past a bottleneck of 6-7% MAPE
  - Tuning sees marginal improvement
  - Overfitting isn’t an issue at lower dimensionality -> dropout/batch normalization/regularization has minimal impact

# Scope for Further Work

- Minimal or no preprocessing for the dataset
  - Requires deeper nets and a dedicated GPU
- Try Ensemble Models
- More iterations of the neural network architecture
- Identify data clusters that are difficult to predict
  - Reduce overall variance in prediction accuracy
- Test the model on the other Merck datasets to evaluate generalizability
- The compound name could be linked to the compound ID to determine biosimilars

# References

[1]https://github.com/CathyQian/Data_Science_Projects/tree/master/Predicting_Merck_Molecular_Activity

[2]https://arxiv.org/pdf/1406.1231.pdf

https://github.com/CathyQian/Data_Science_Projects/tree/master/Predicting_Merck_Molecular_Activity
https://arxiv.org/pdf/1406.1231.pdf

# Questions?