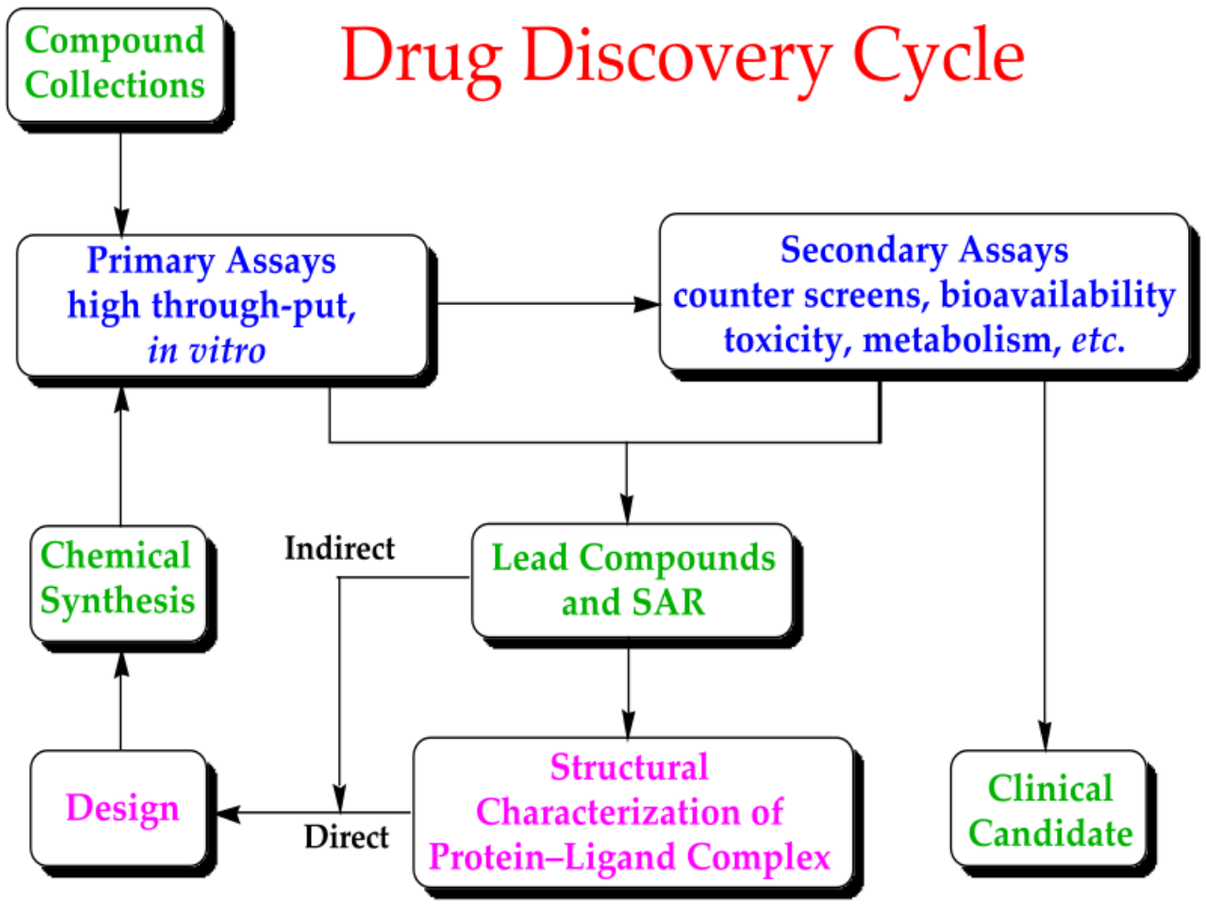
DEEP LEARNING FOR QUANTUM MECHANICS

MITCH MURPHY

MEDICINE IS A SCIENCE OF UNCERTAINTY AND AN ART OF PROBABILITY.

William Osler

- Motivation
- Data
- Feature engineering
- Visualizations
- Models
- Implications/ applications

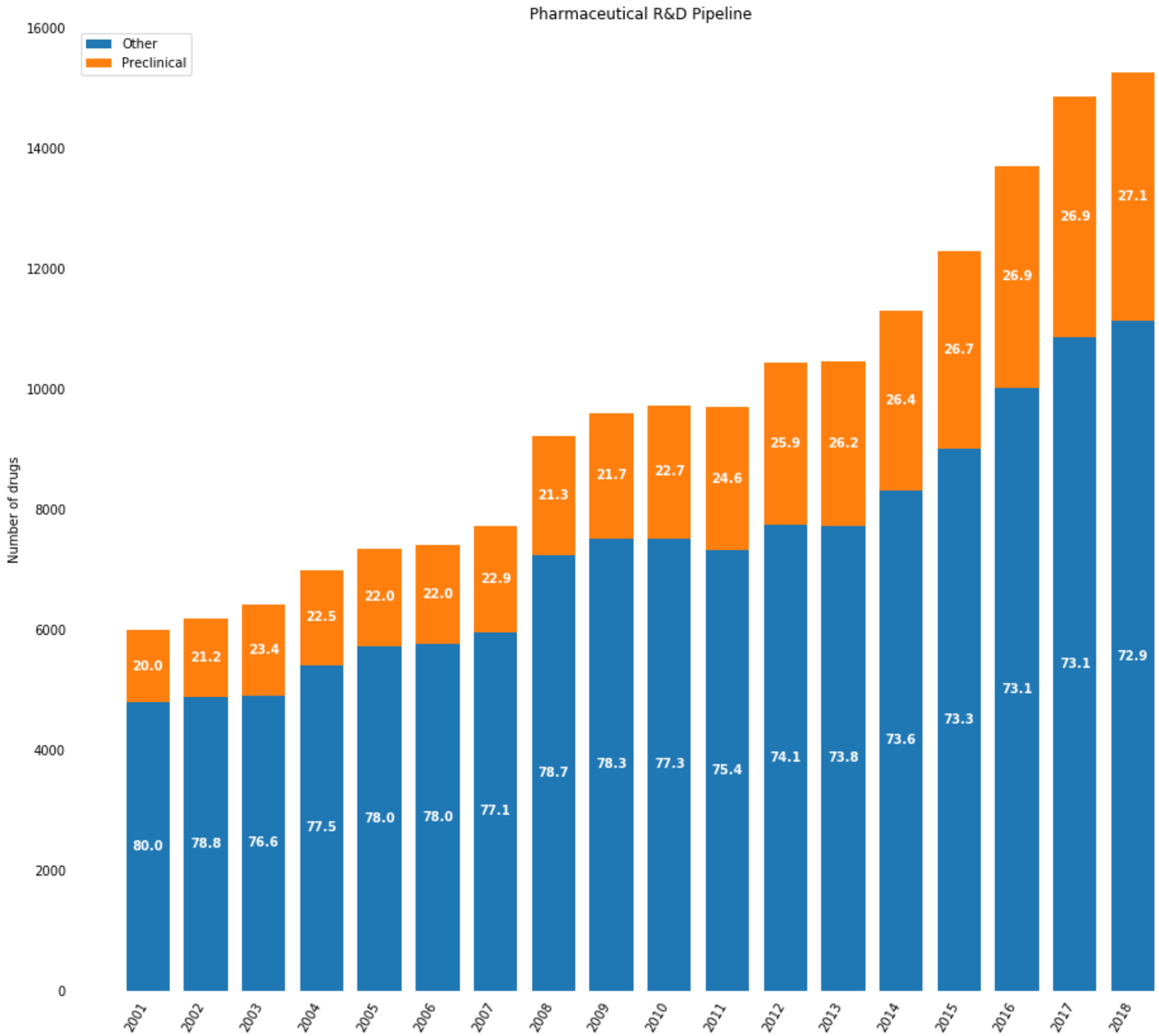


$71.4 BILLION

DRUGMAKERS SAY R&D SPENDING HIT RECORD IN 2017

DEEP LEARNING FOR QUANTUM MECHANICS

2018 VS. 2017



- **2.65%** more drugs
- **3.27%** more preclinical drugs
- **3.27%** more pharmaceutical companies doing R&D
- **$2.48 million** per preclinical drug

MOTIVATION

- Demand for new medications
- Lower health costs
- B. Himmetoglu (2016) Tree based machine learning framework for predicting ground state energies of molecules

MoleculeNet

A Benchmark for Molecular Machine Learning

DATA

- QM7 dataset, Pande Group @ Stanford
  - Coulomb Matrix
  - atomization energy
  - atomic charges
  - cartesian coordinates

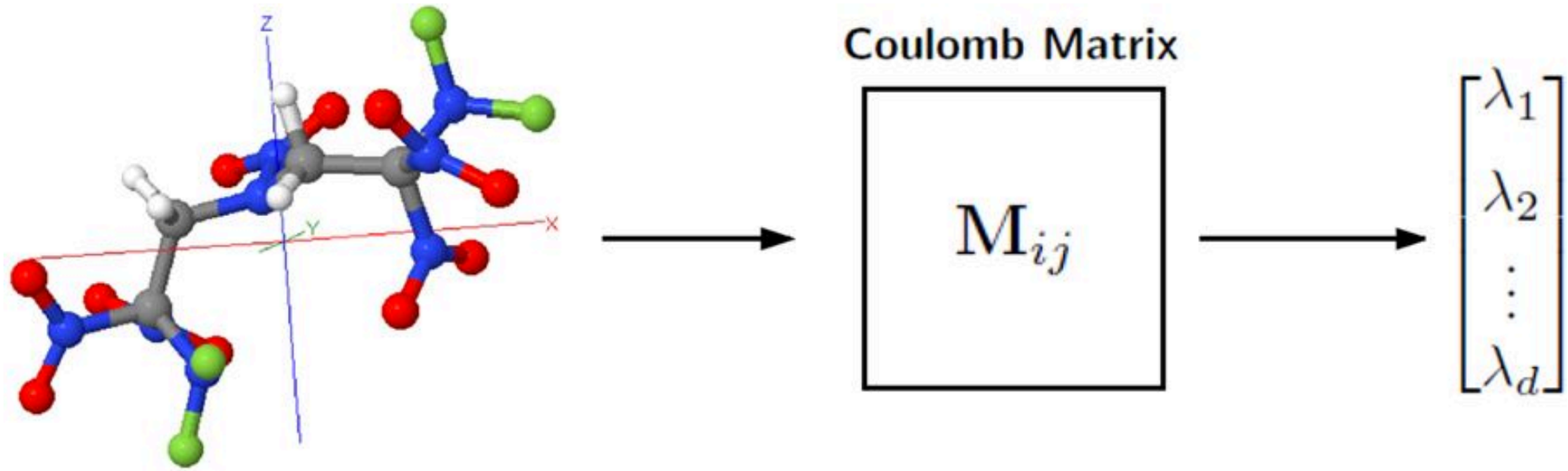
$$C_{IJ} = \begin{cases} 0.5\, Z_I^{2.4} & I = J \\ \frac{Z_I\, Z_J}{|\mathbf{R}_I - \mathbf{R}_J|} & I \neq J \end{cases}$$

FEATURE ENGINEERING

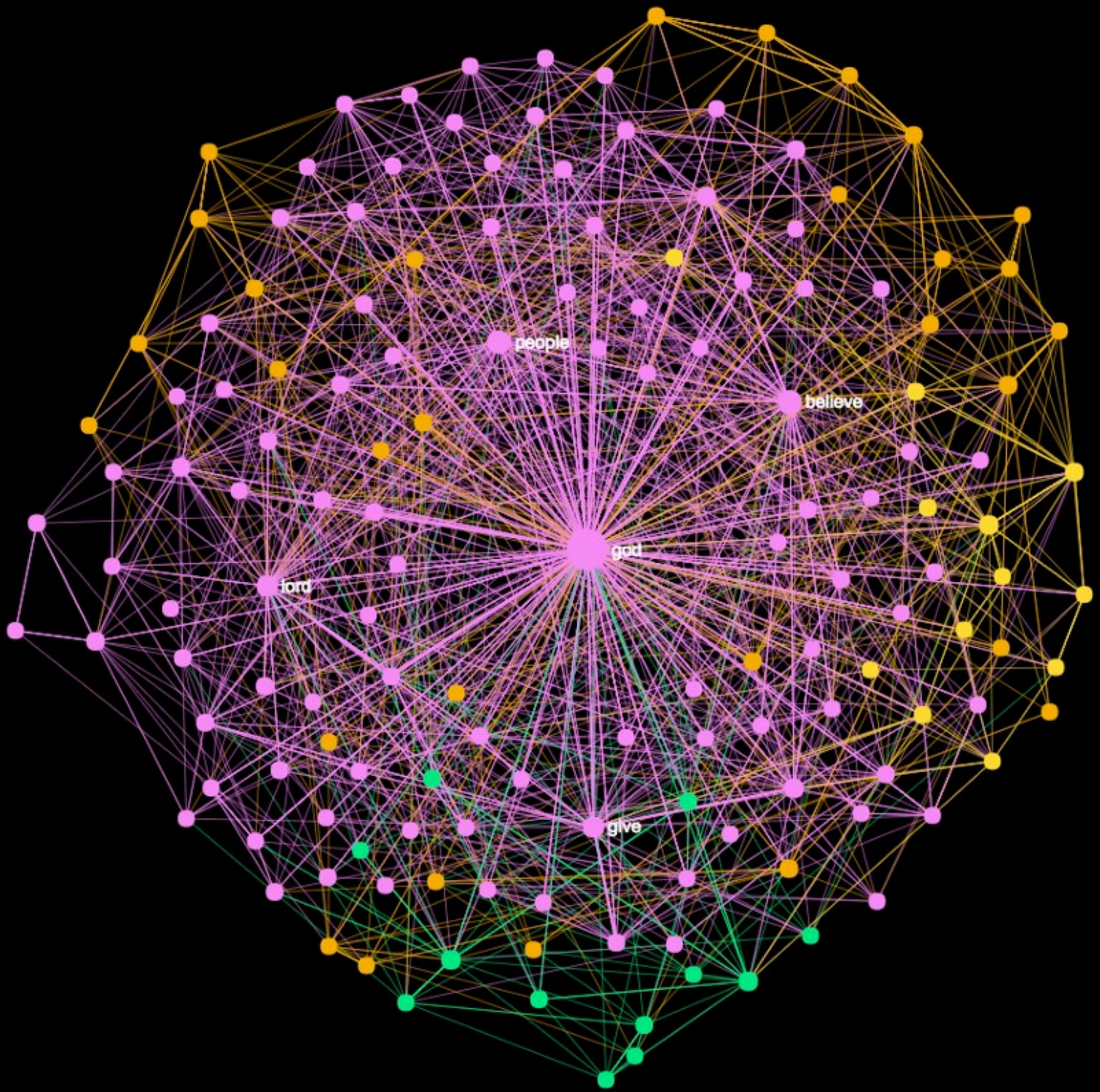
- Coulomb Matrix
  - upper triangle, unrolled and sorted
  - Eigenvectors

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

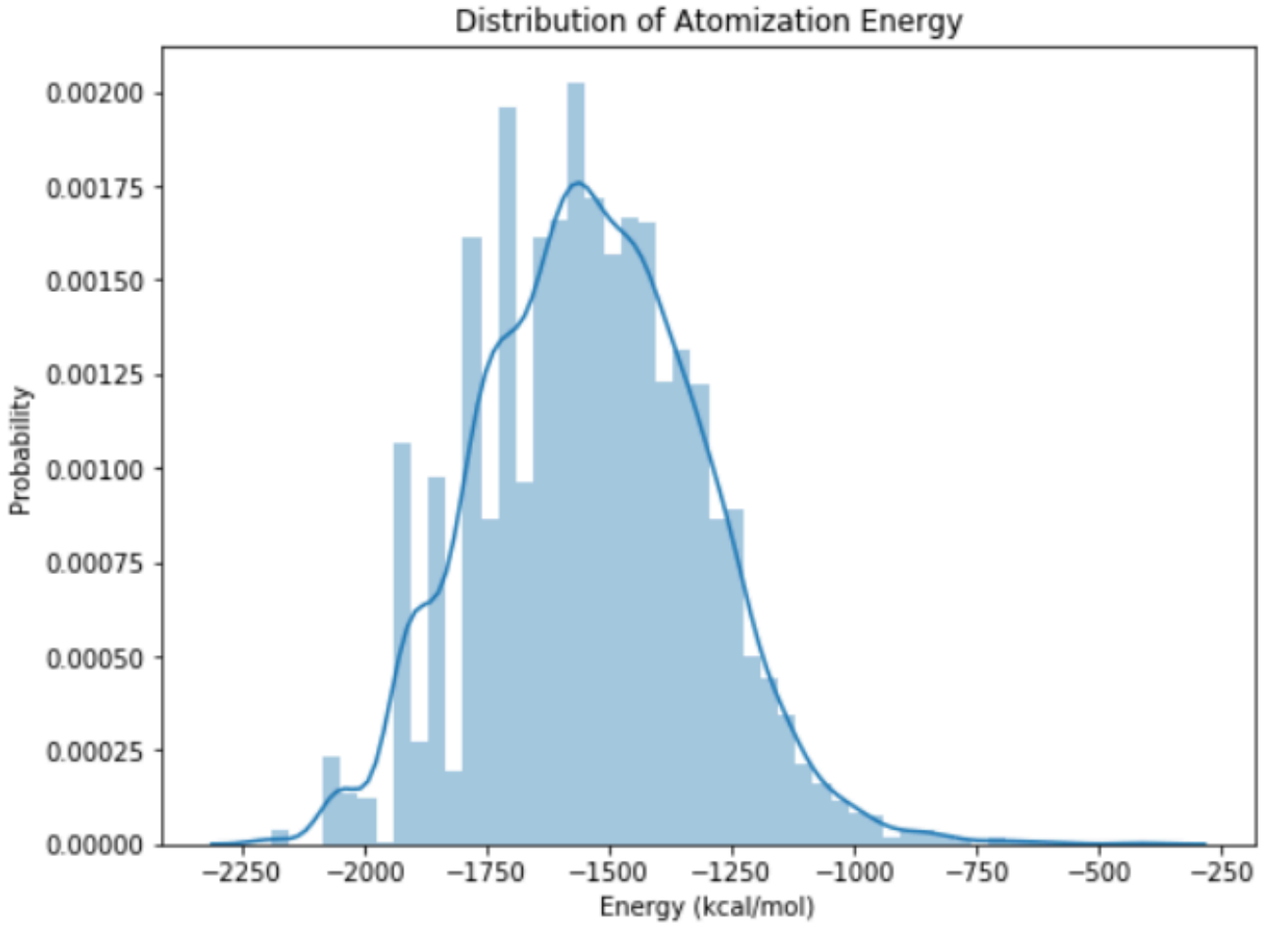
- Interatomic distance matrix
  - Eigenvector centrality

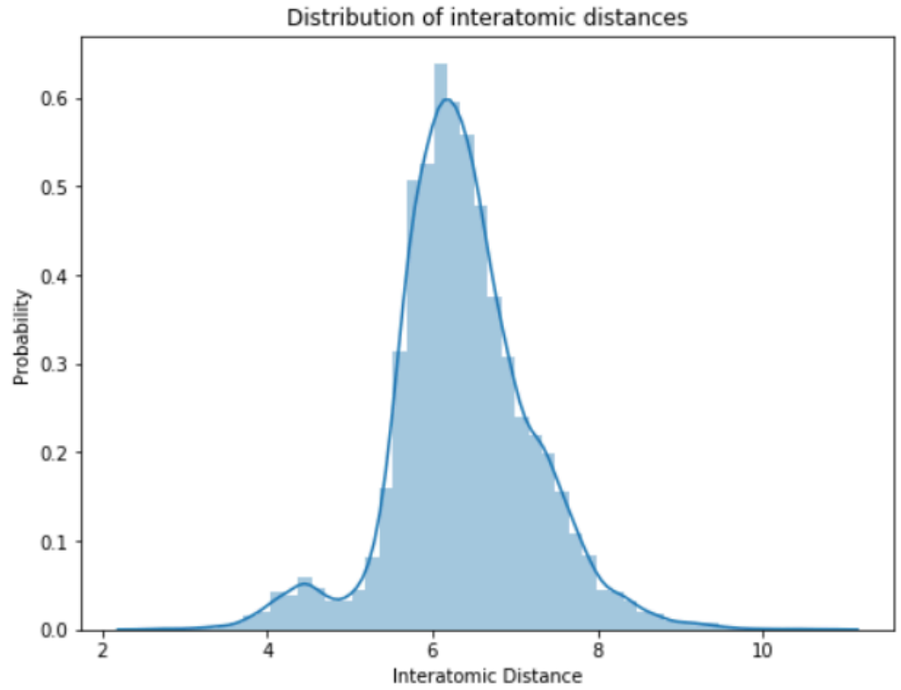


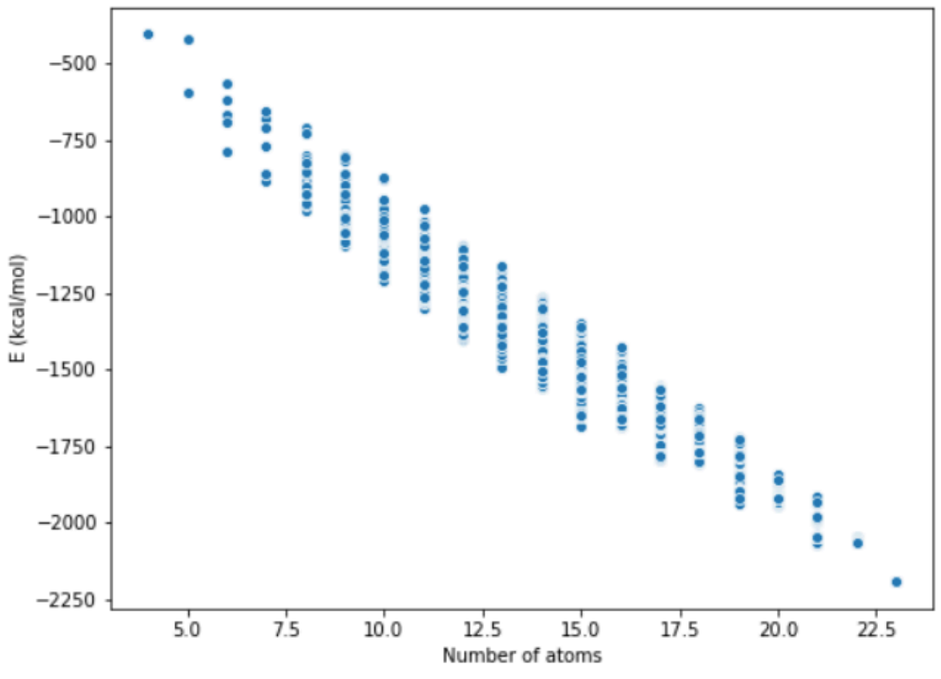
NETWORK CENTRALITY



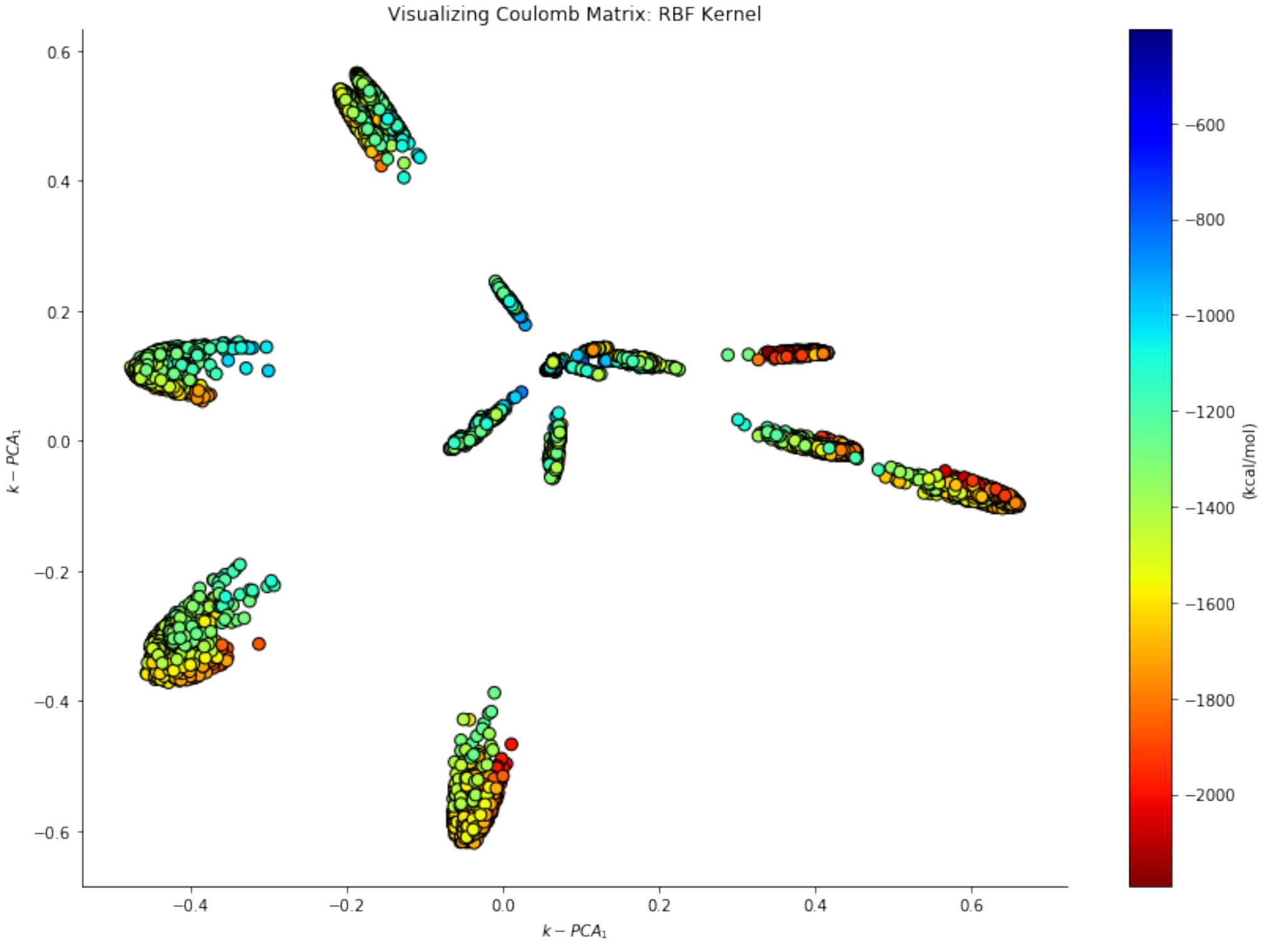
VISUALIZATION



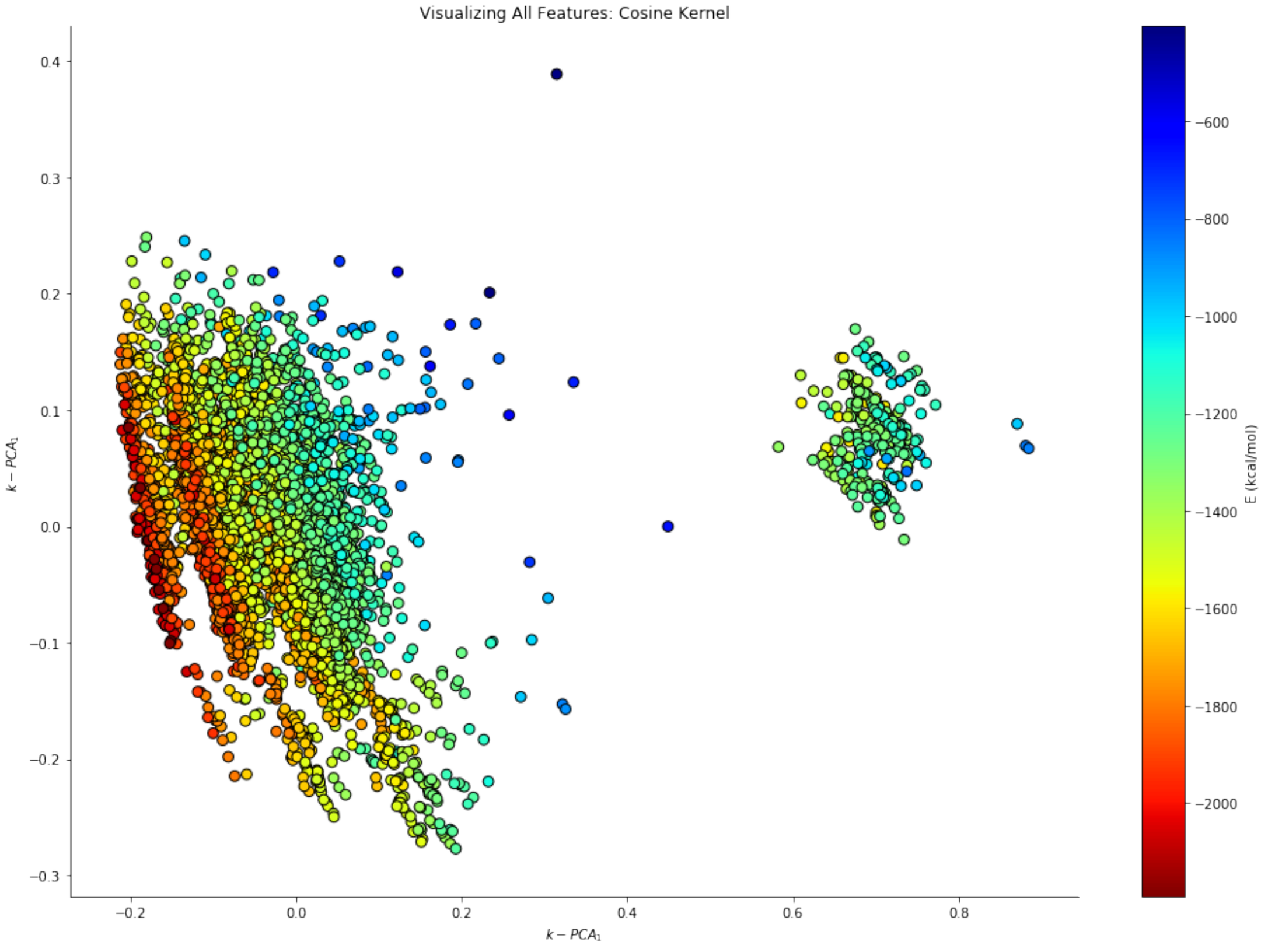




DEEP LEARNING FOR QUANTUM MECHANICS



DEEP LEARNING FOR QUANTUM MECHANICS

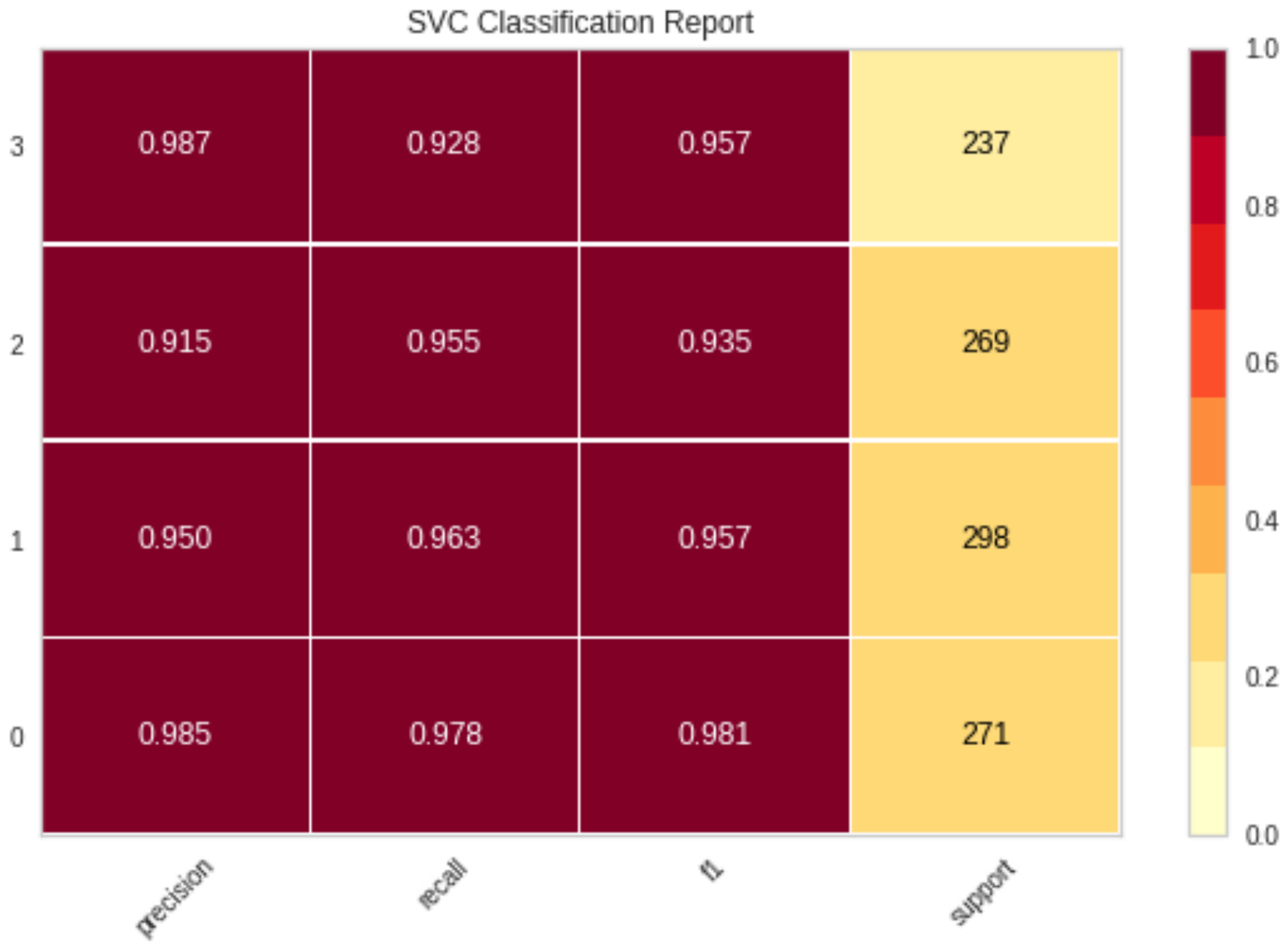


MODELING

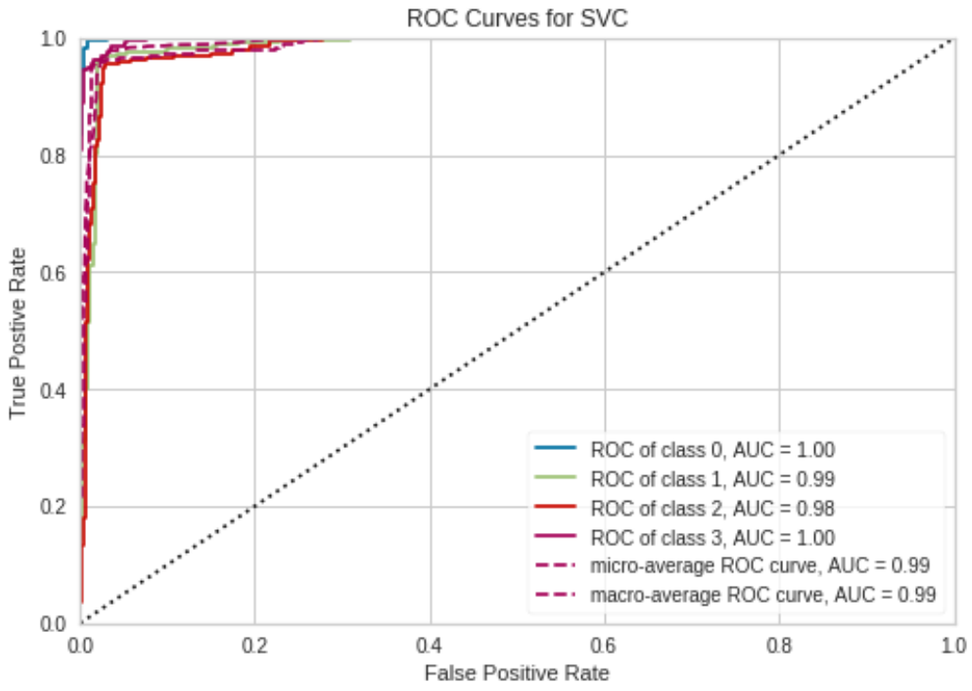
- Classification
- Regression (gradient boosting)
- MLP
- Convolutional Neural Network

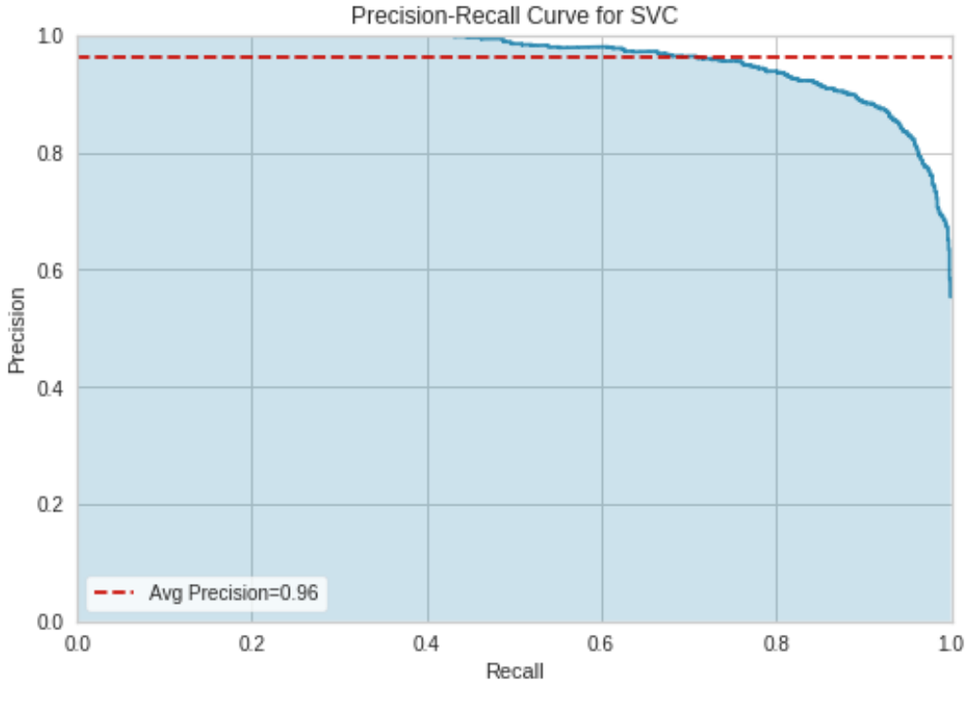
CLASSIFICATION

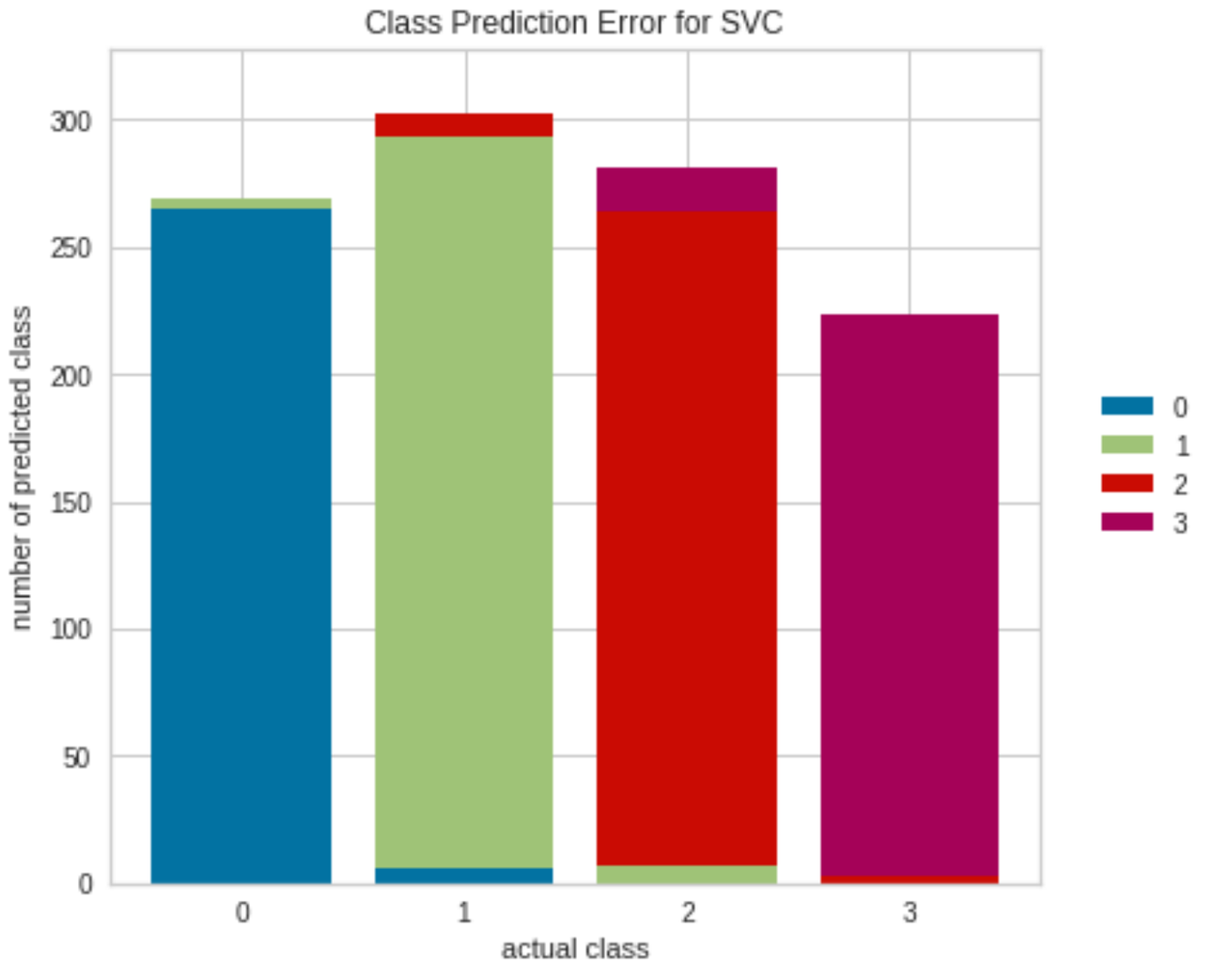
- Linear SVM
- 70/15/15 splits



CLASSIFICATION



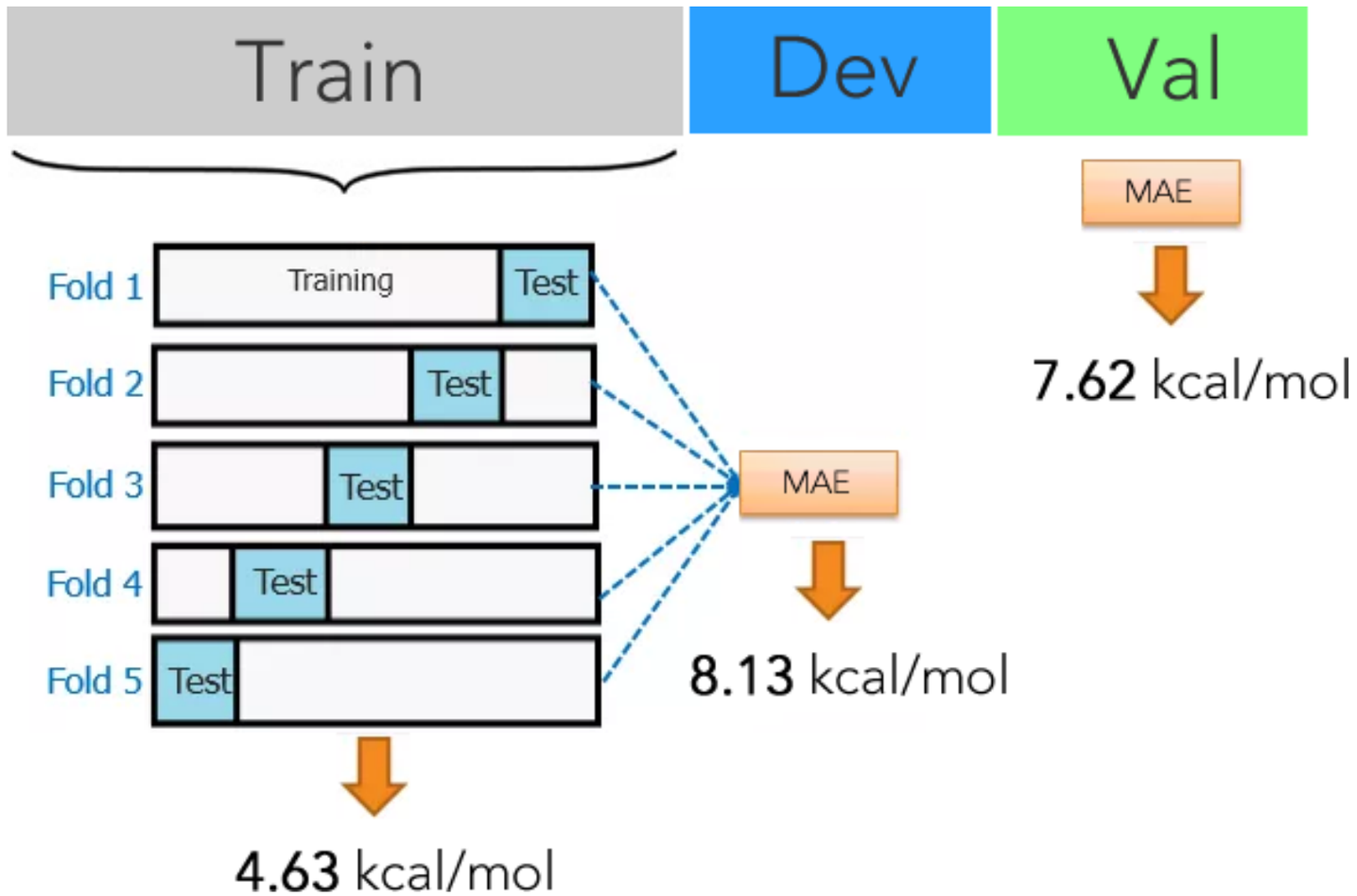




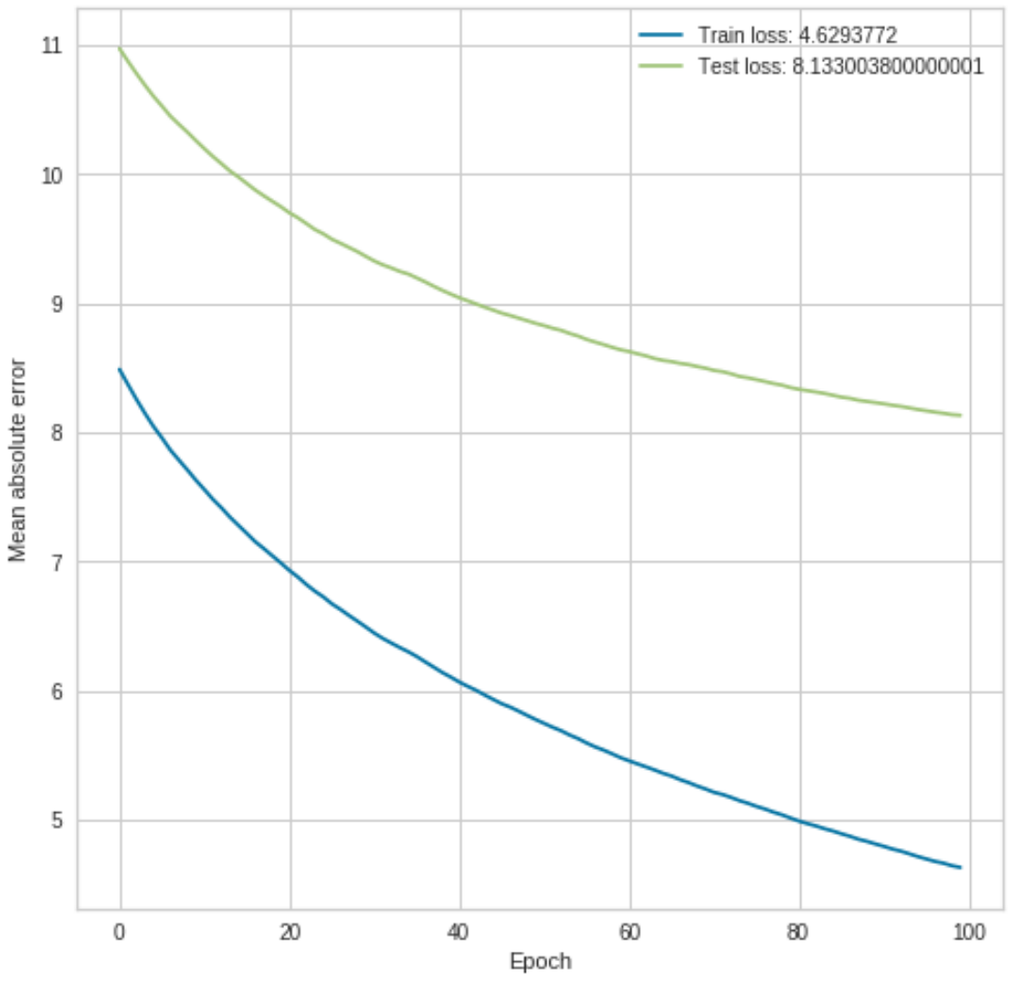
REGRESSION

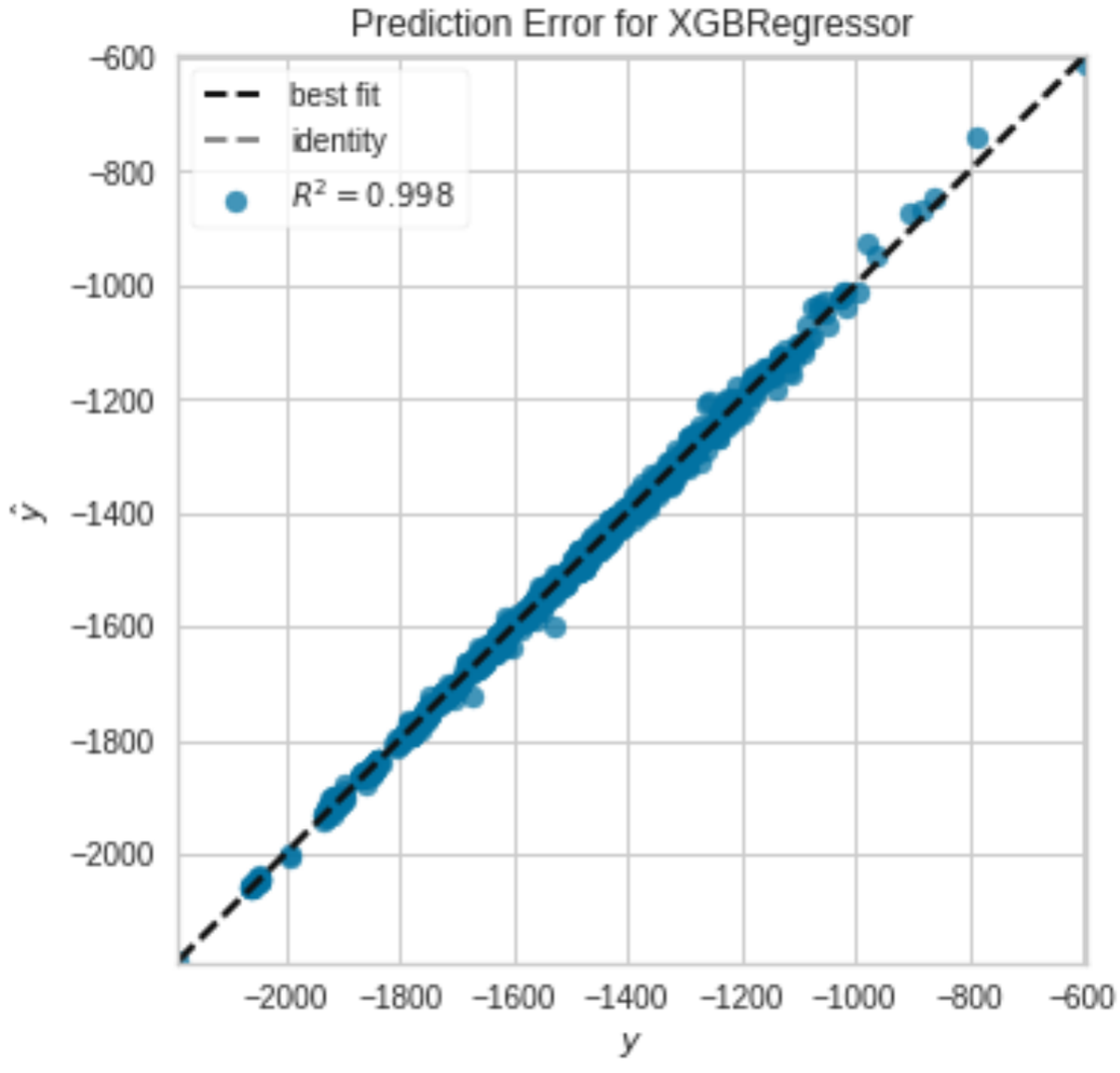
XGBoost

- Gradient boosting (XGBoost)
- 70|15|15 splits
  - 5-fold cross validation
- Hyperparameter tuning
  - learning rate: 0.06
  - max depth: 5
  - subsample: 0.9
  - feature subsample: 0.2
  - lambda: 0.9
  - alpha: 0.01

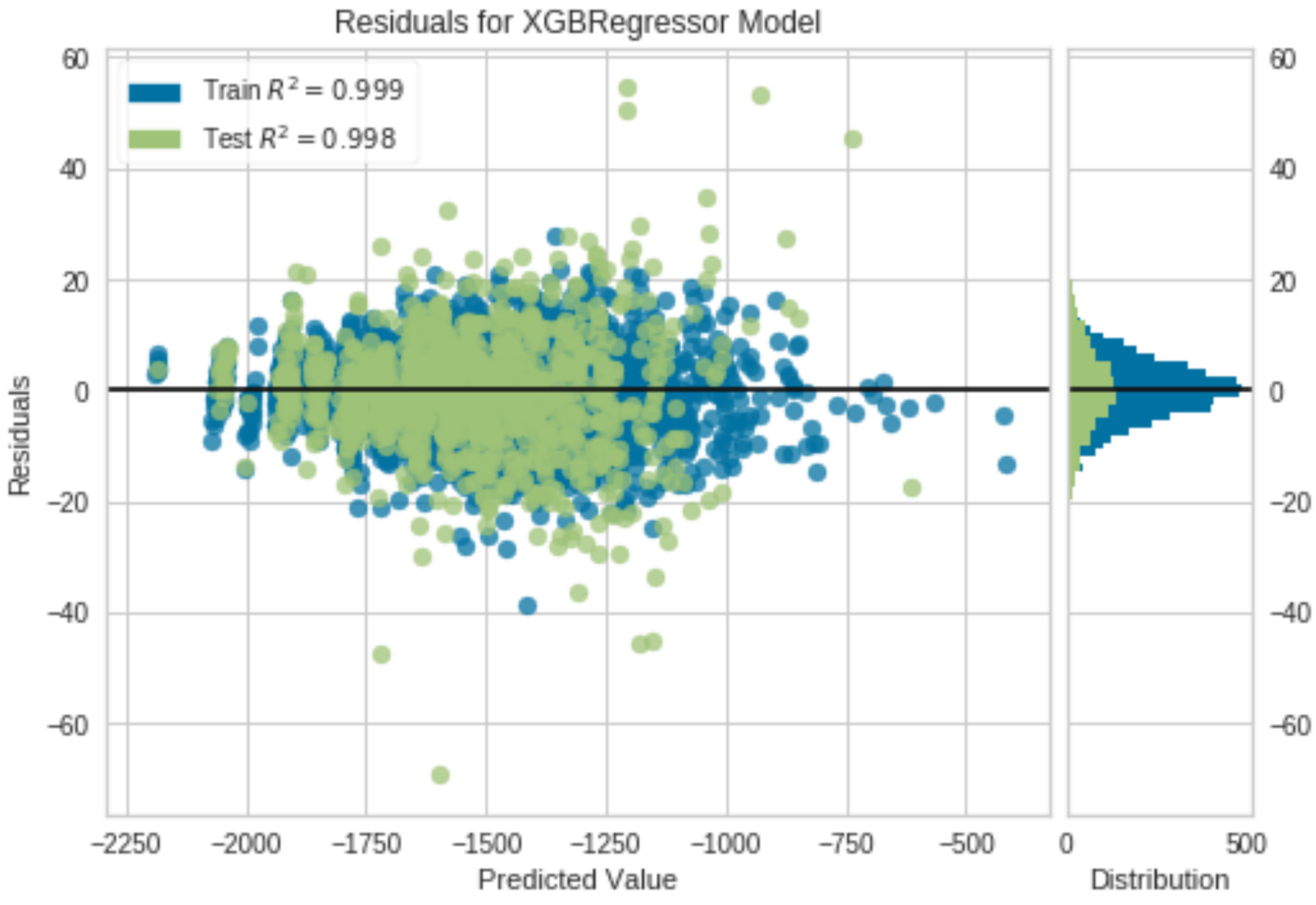


XGBOOST


Train loss: 4.6293772
Test loss: 8.133003800000001
Mean absolute error
Epoch
11
10
9
8
7
6
5
0
20
40
60
80
100

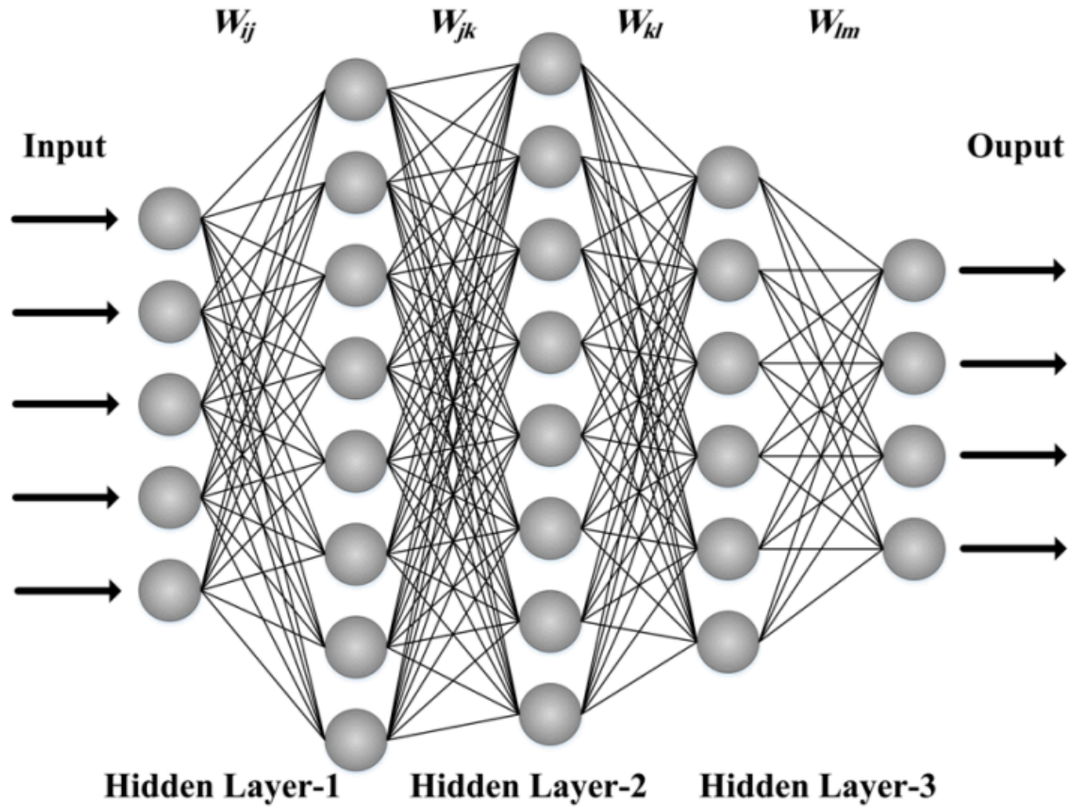

Prediction Error for XGBRegressor
best fit
identity
$R^2 = 0.998$
$\hat{y}$
y
−600
−800
−1000
−1200
−1400
−1600
−1800
−2000

XGBOOST


Residuals for XGBRegressor Model
Train $R^2 = 0.999$
Test $R^2 = 0.998$
Residuals
Predicted Value
Distribution
60
40
20
0
−20
−40
−60
−2250
−2000
−1750
−1500
−1250
−1000
−750
−500
0
500

MULTILAYER PERCEPTRON

- 2 hidden layers
  - 256 hidden units
  - 512 hidden units
  - initializer: normal
- Adam optimizer
  - learning rate: 0.001
- Loss: mean absolute error
- 50 epochs, early stopping of 5
- batch size of 8

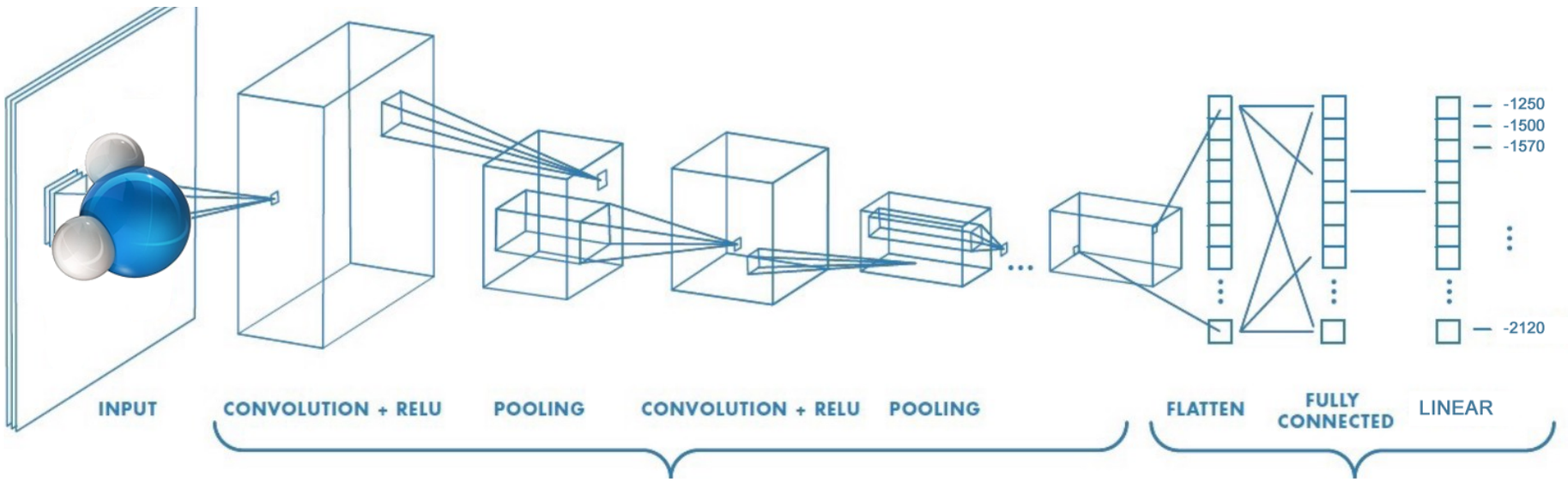


TRAIN LOSS: 24.9

TEST LOSS: 26.9

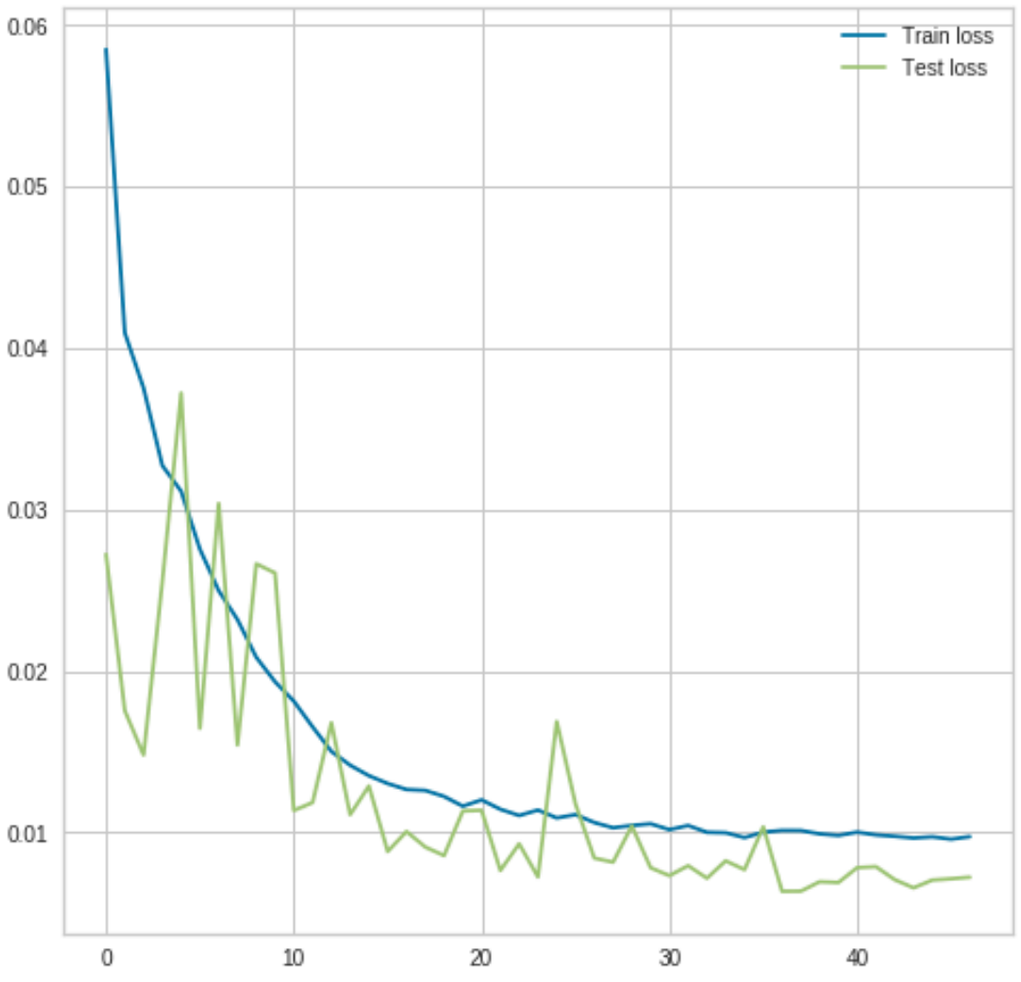
CONVOLUTIONAL NEURAL NETWORK

- Translational invariance
- 4 convolutional layers
- 1 max pooling
- 1 fully connected and 1 dropout



CONVNET

- Train loss: 19.18 kcal/mol
- Test loss: 12.72 kcal/mol



- FURTHERMORE
- SHORTCOMINGS
- QUESTIONS?