DEEP LEARNING FOR QUANTUM MECHANICS

MITCH MURPHY

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

William Osler

$71.4 BILLION

DRUGMAKERS SAY R&D SPENDING HIT RECORD IN 2017

MOTIVATION

- Demand for new medications
- Lower health costs
- B. Himmetoglu (2016) Tree based machine learning framework for predicting ground state energies of molecules
- Hansen et al. (2012) Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies

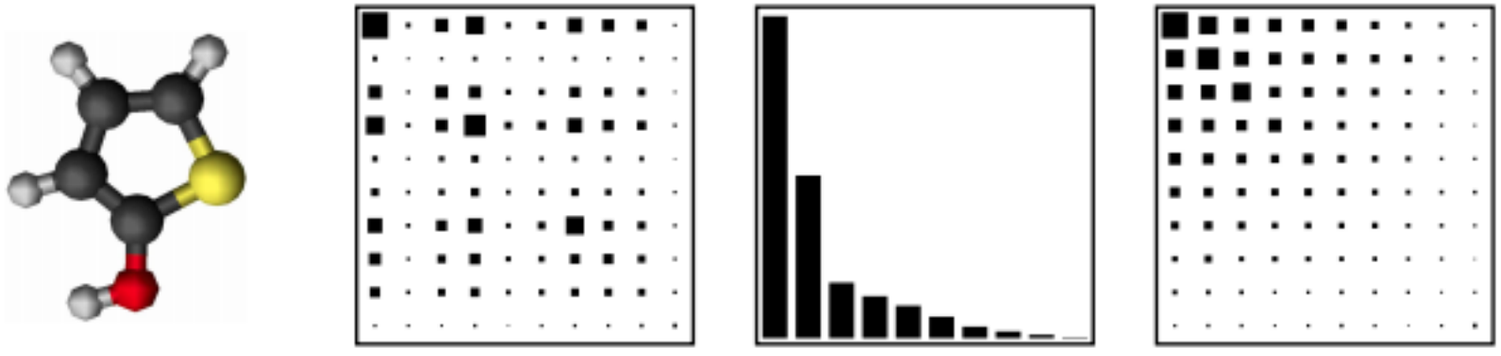
MoleculeNet

A Benchmark for Molecular Machine Learning

DATA

- QM7 dataset, Pande Group @ Stanford
  - Coulomb Matrix
  - Atomization energy
  - Atomic charges
  - Cartesian coordinates

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

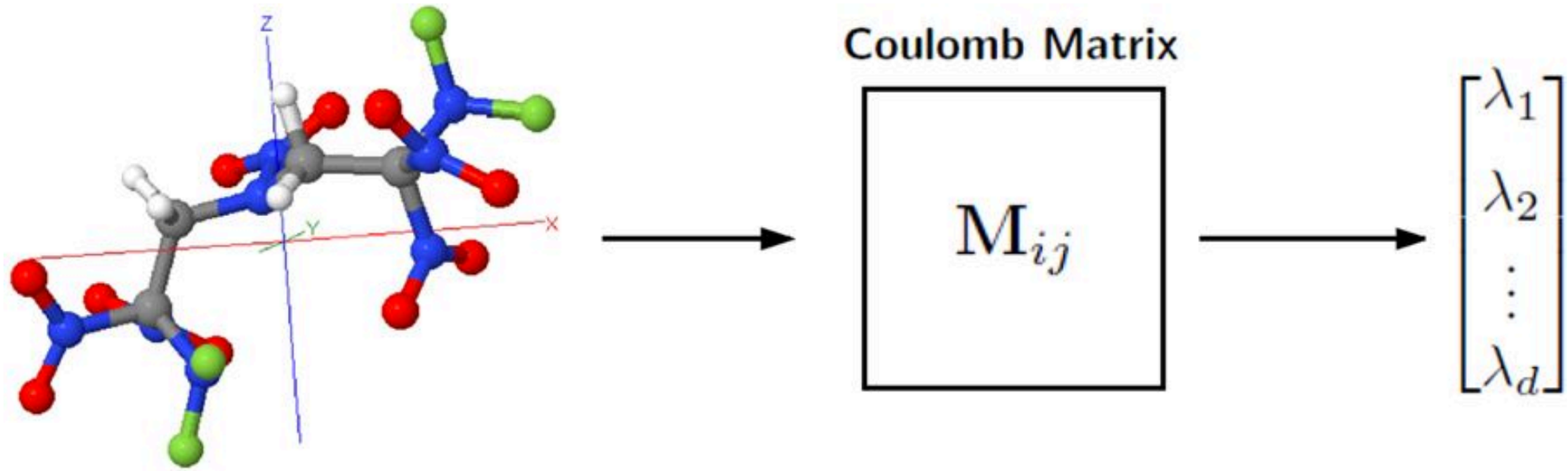


FEATURE ENGINEERING

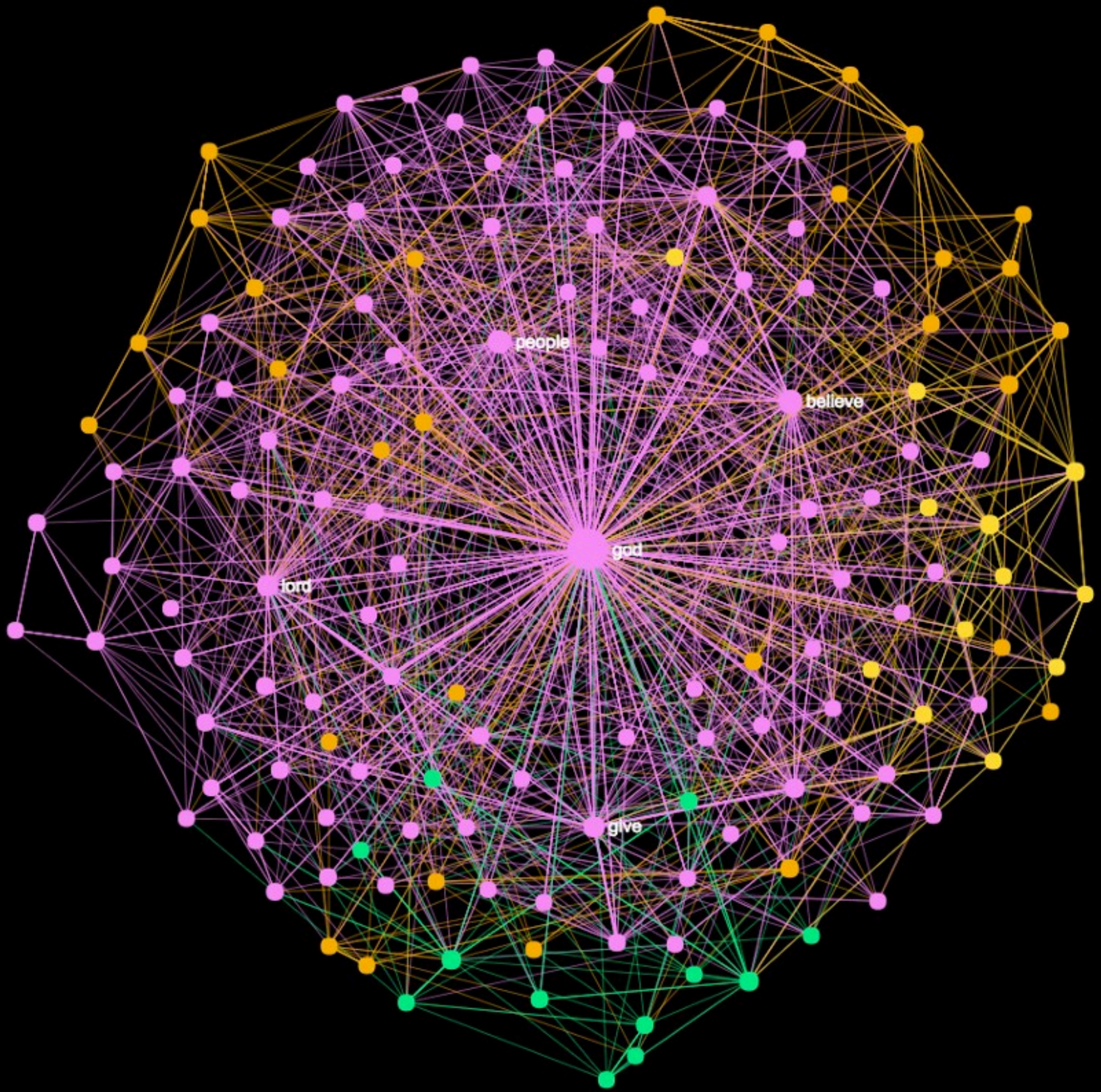
- Coulomb Matrix
  - upper triangle, unrolled and sorted
  - Eigenvectors

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

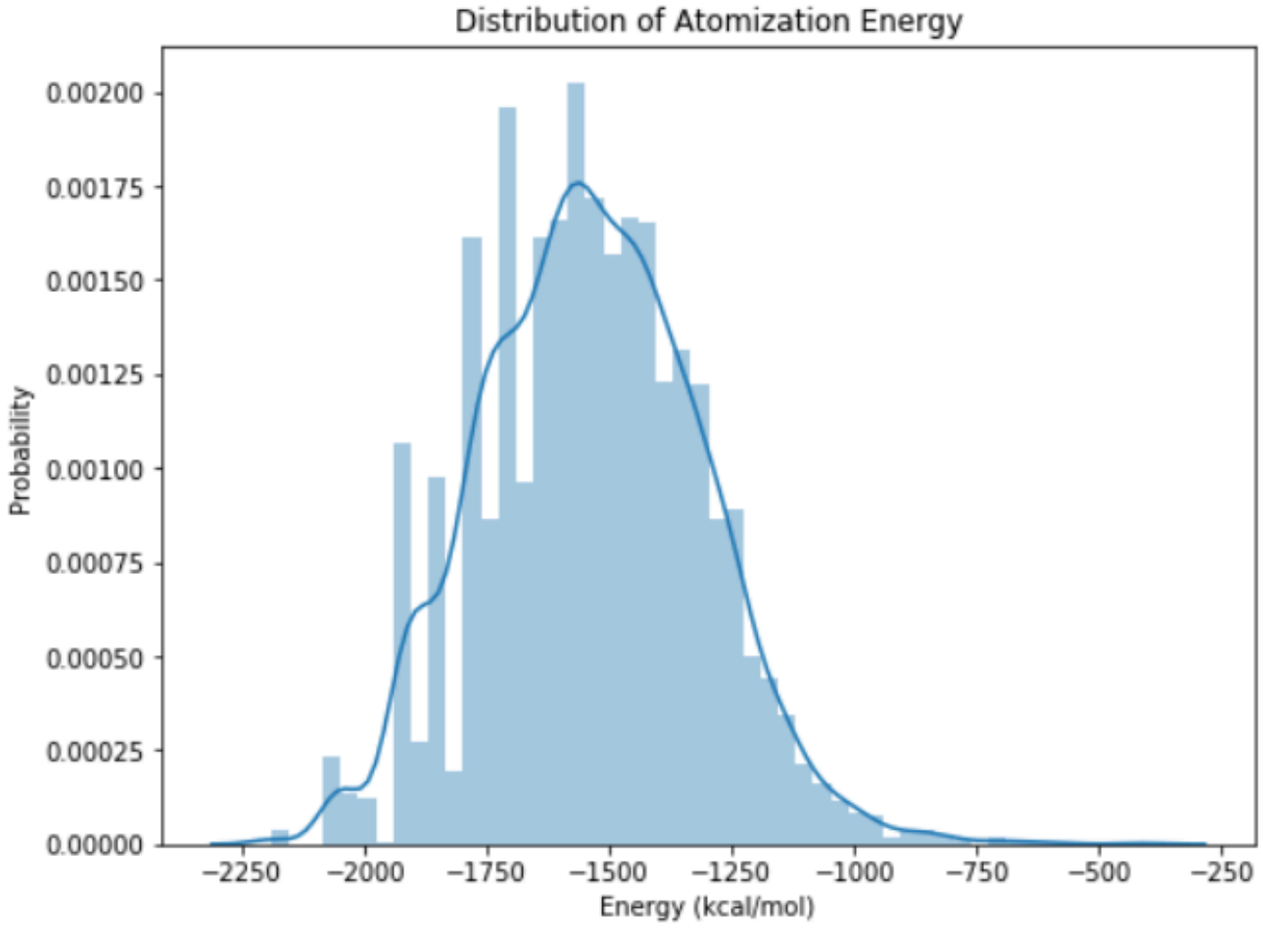
- Interatomic distance matrix
  - Eigenvector centrality

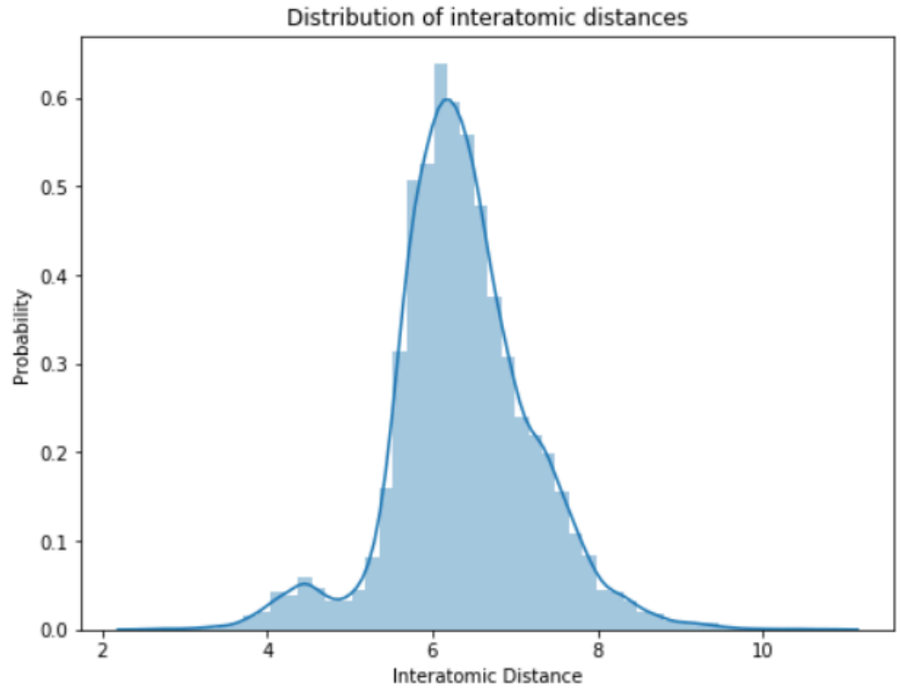


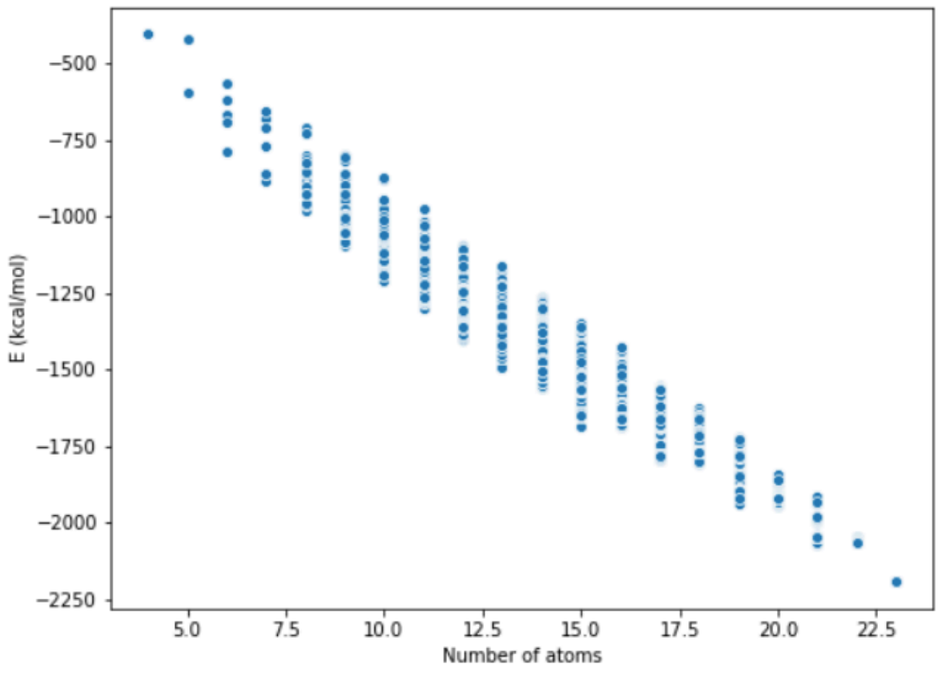
NETWORK CENTRALITY



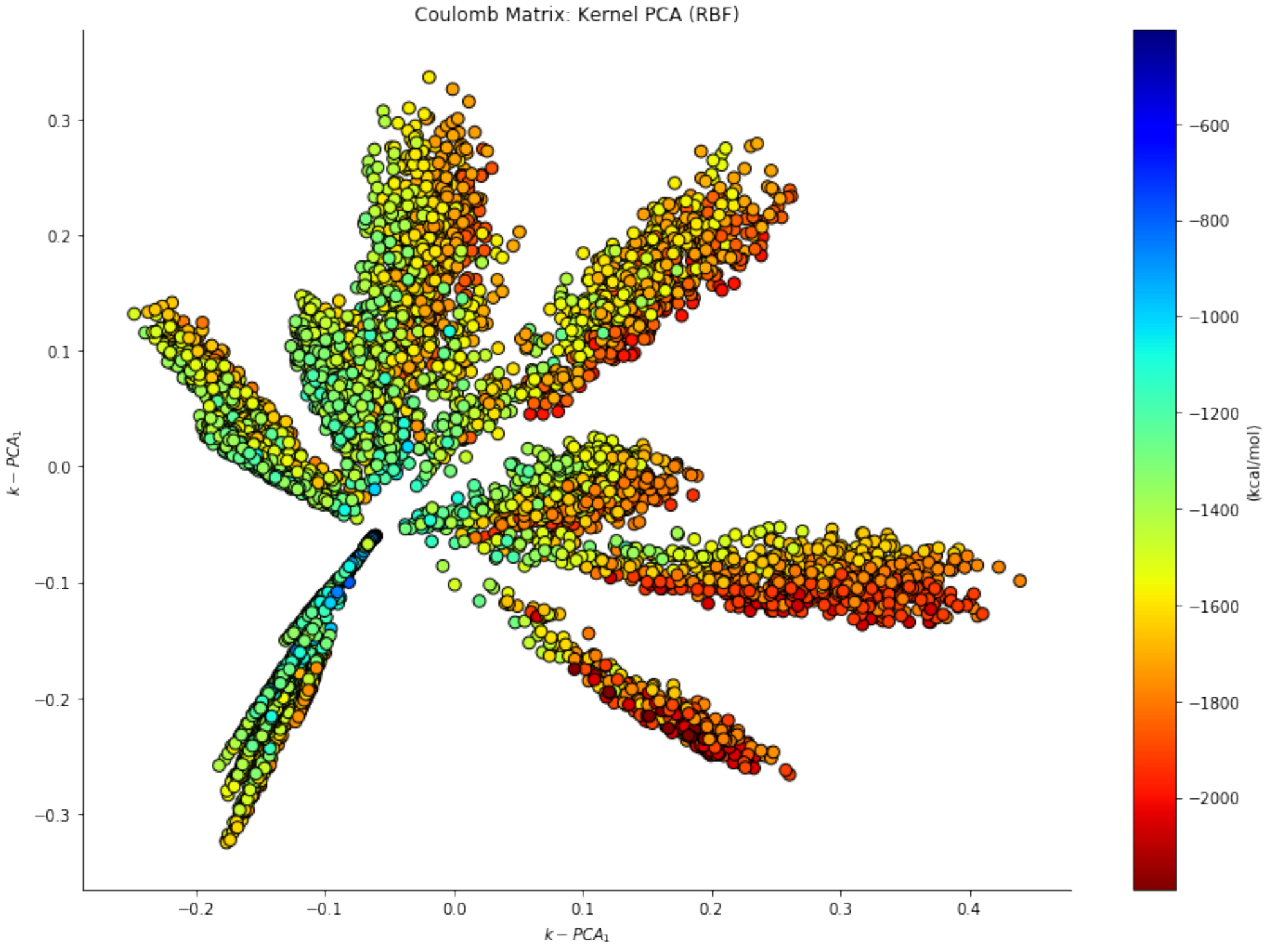
VISUALIZATION



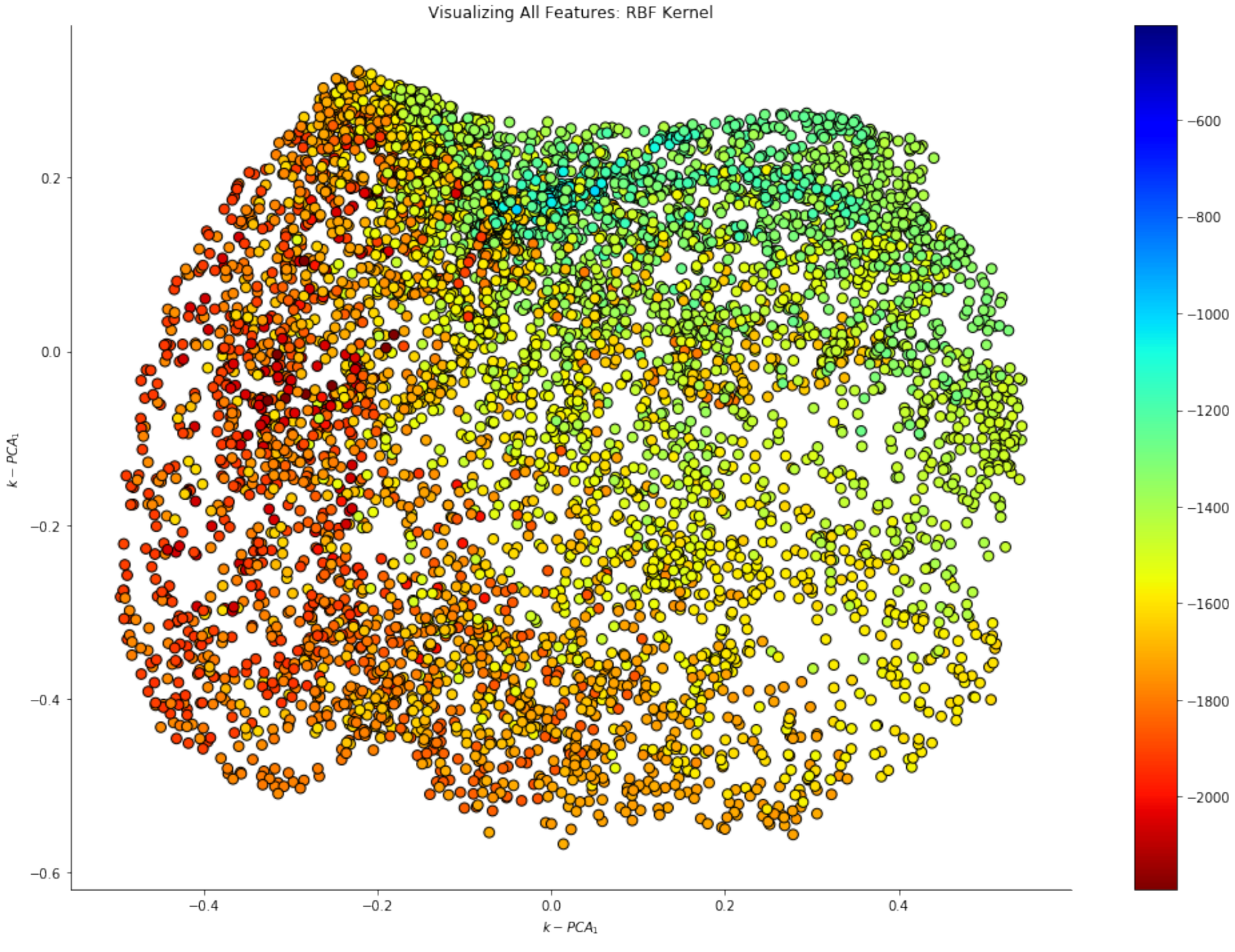




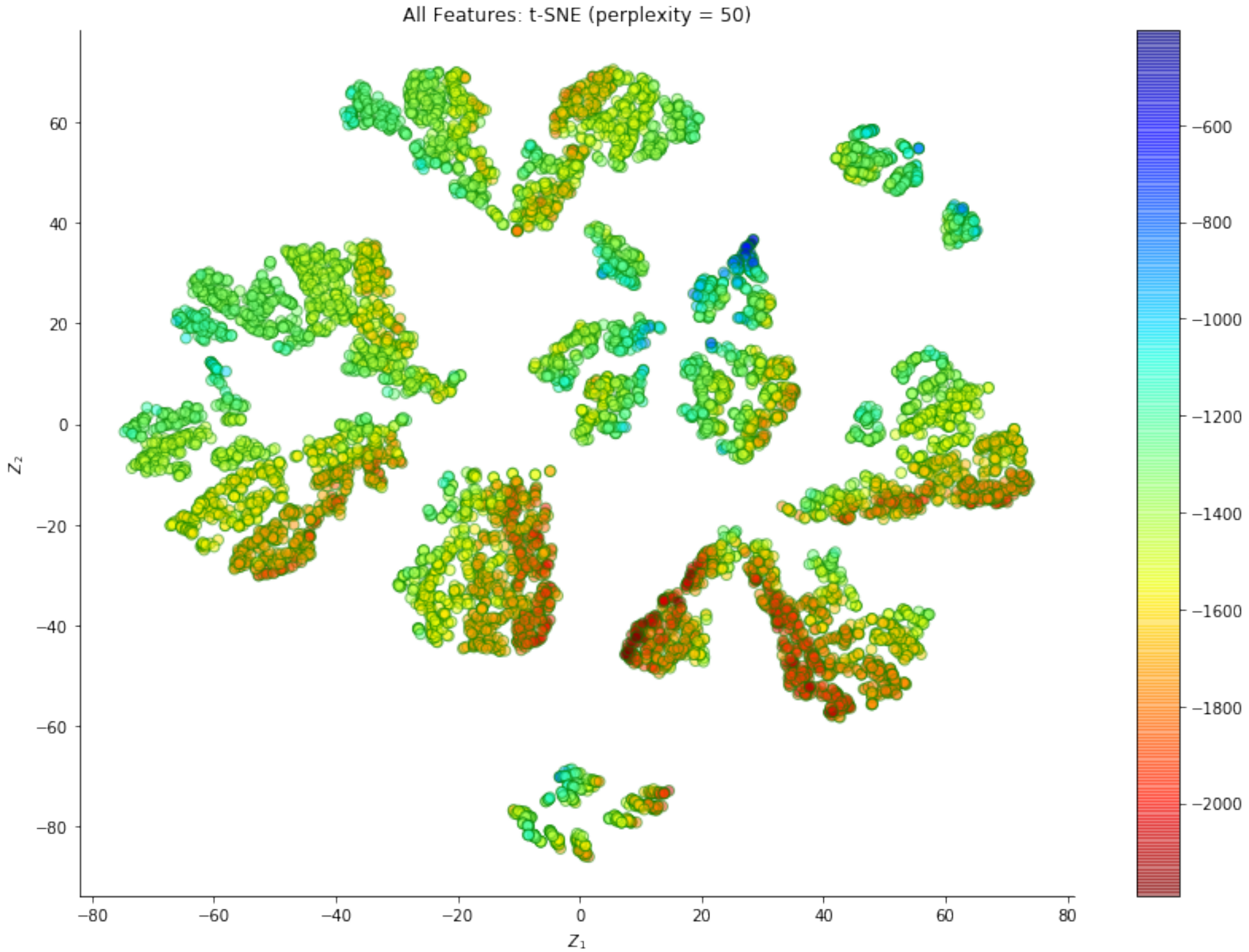
VISUALIZATION



VISUALIZATION



VISUALIZATION



MODELING

- Classification
- Regression
- Deep Learning

CLASSIFICATION

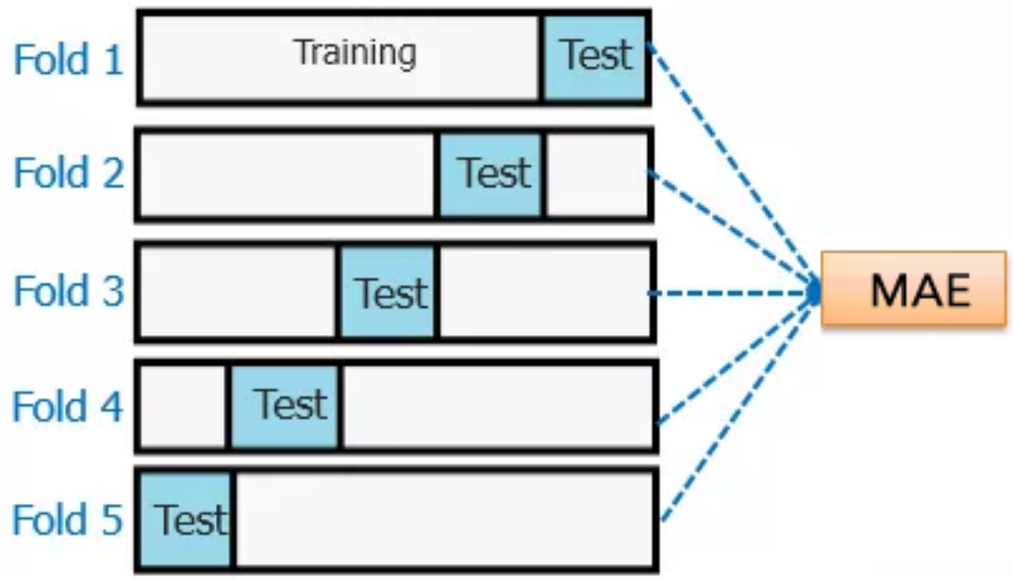
- 10 classes, bin size of 178
- Linear SVM
- 70/15/15 splits
- Class weights

```
Train score:  0.971159623472862
Test score:  0.9033728350045579
--------
EVALUATE on validation set
--------
Validation score:  0.9237209302325582
              precision    recall  f1-score   support

           3       1.00      1.00      1.00         1
           4       1.00      1.00      1.00         3
           5       0.88      1.00      0.93        14
           6       0.87      0.87      0.87        83
           7       0.91      0.88      0.90       214
           8       0.90      0.95      0.93       318
           9       0.97      0.91      0.94       307
          10       0.93      0.97      0.94       115
          11       1.00      1.00      1.00        19
          12       1.00      1.00      1.00         1

   micro avg       0.92      0.92      0.92      1075
   macro avg       0.94      0.96      0.95      1075
weighted avg       0.92      0.92      0.92      1075
```

REGRESSION

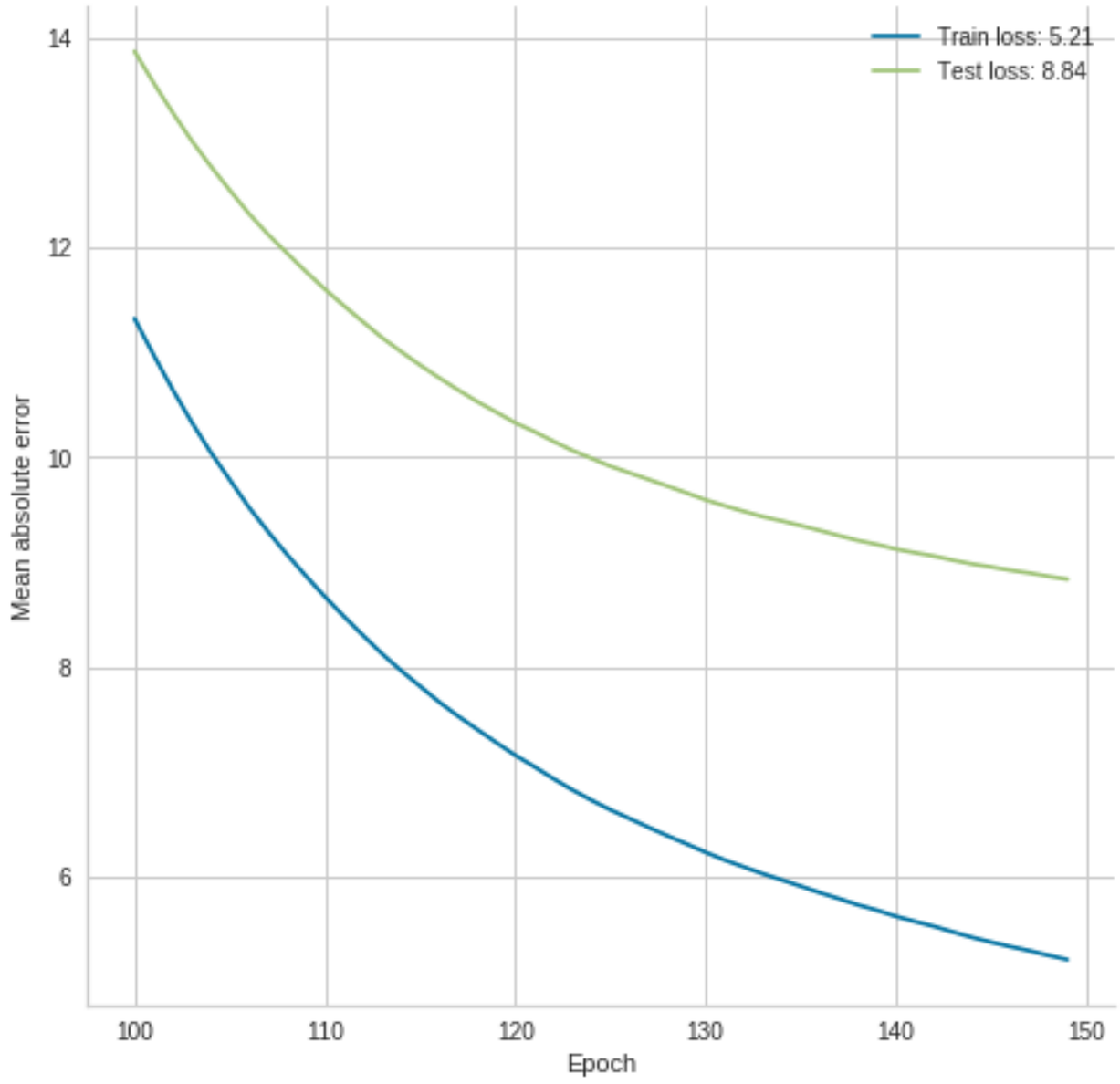
XGBoost

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

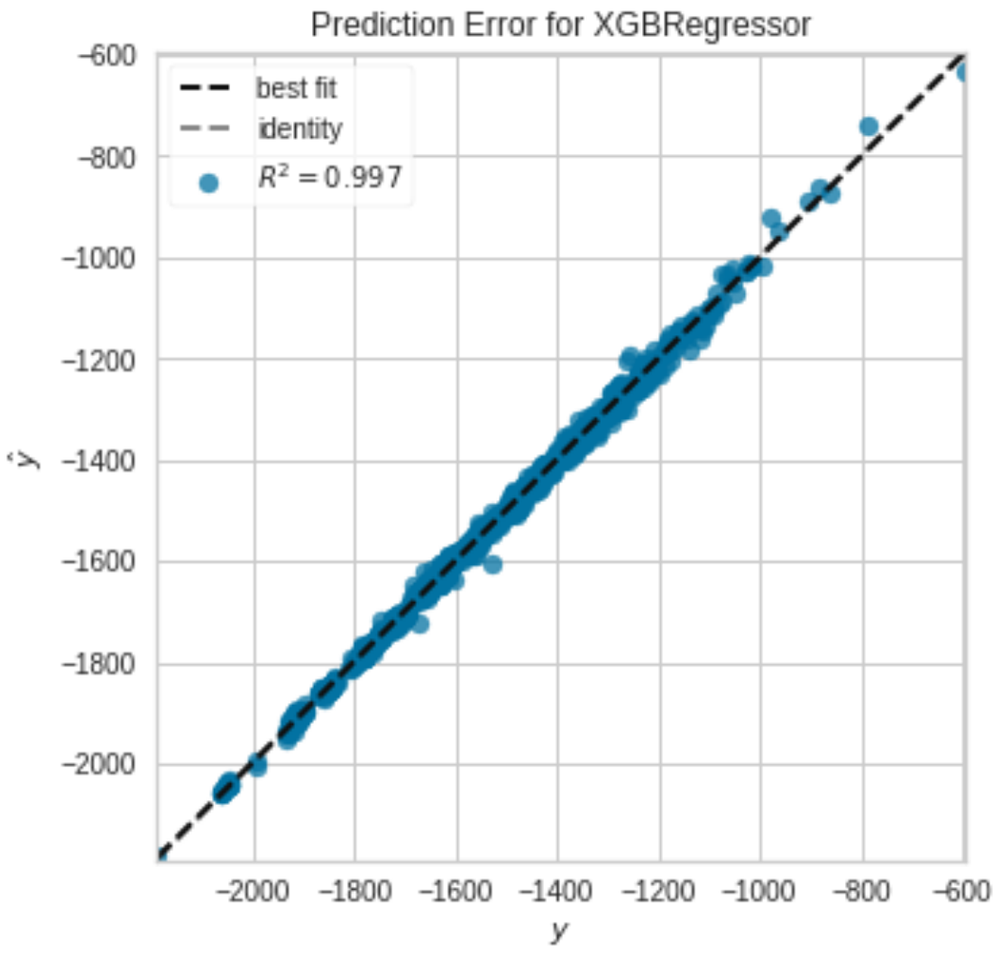


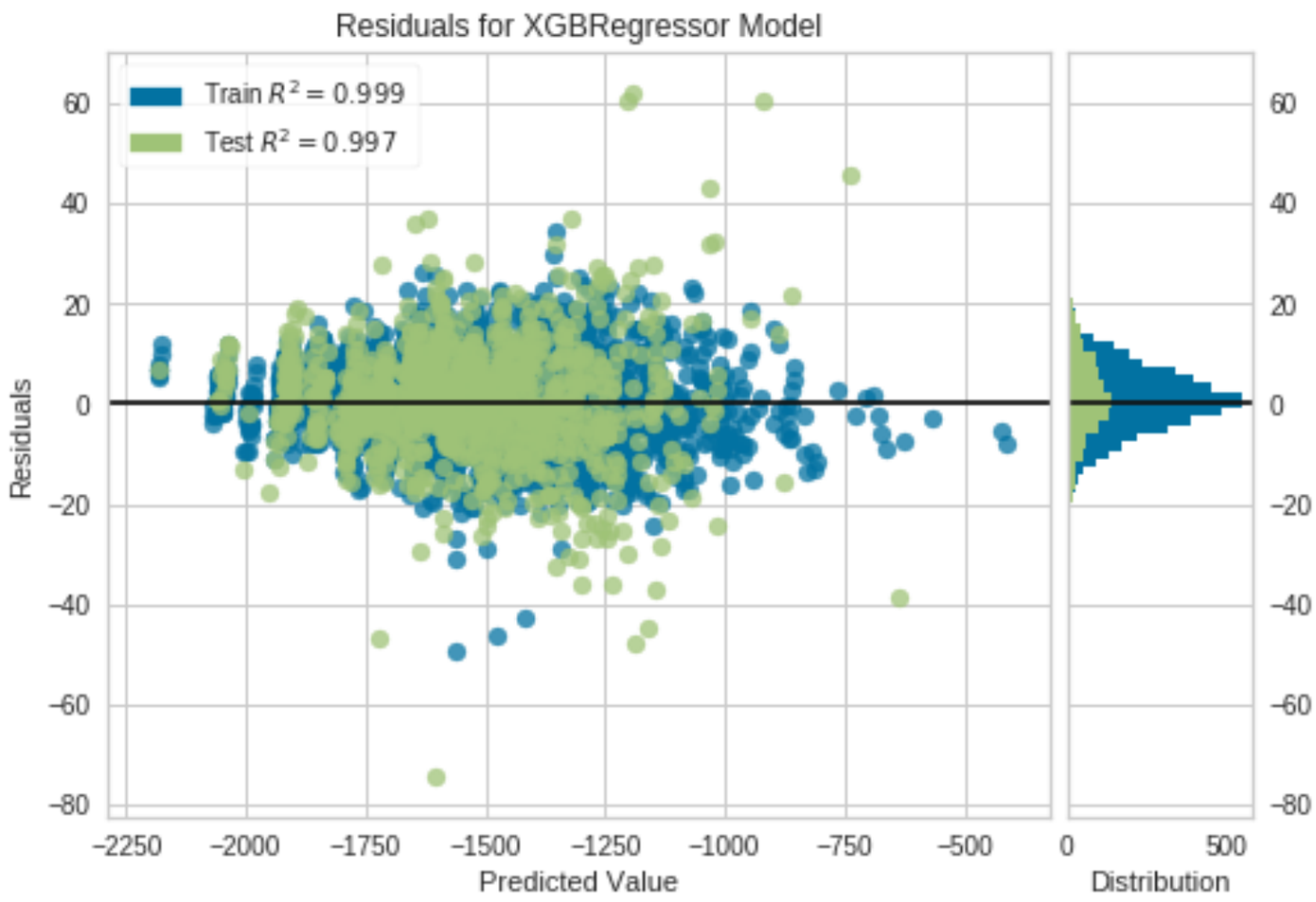
XGBOOST: ERRORS

- Train loss: **5.21**
- Dev loss: **8.84**
- Val loss: **8.31**
- Val R^2: 0.99726
- Execution time: **22.1** seconds



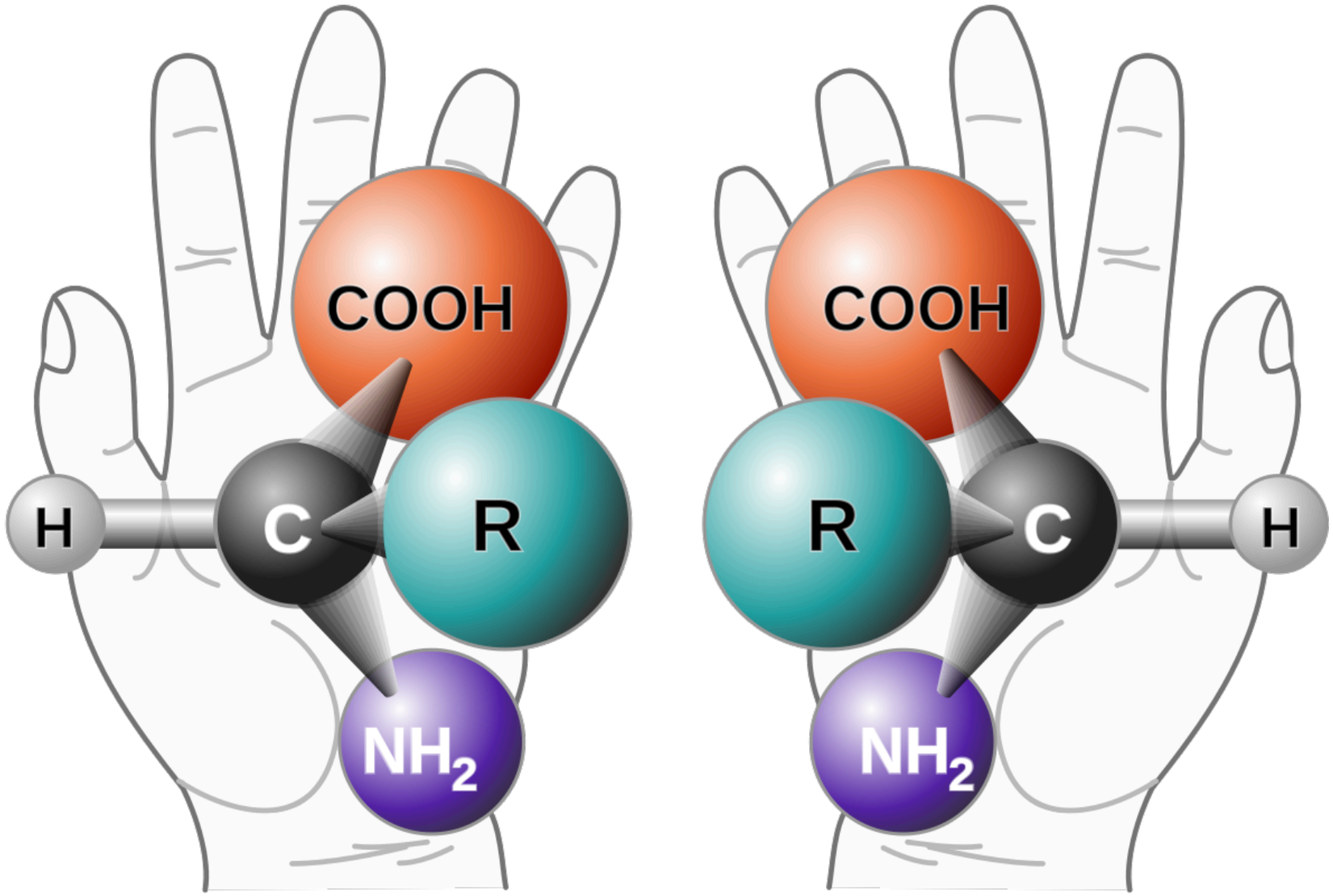
XGBOOST: RESIDUALS





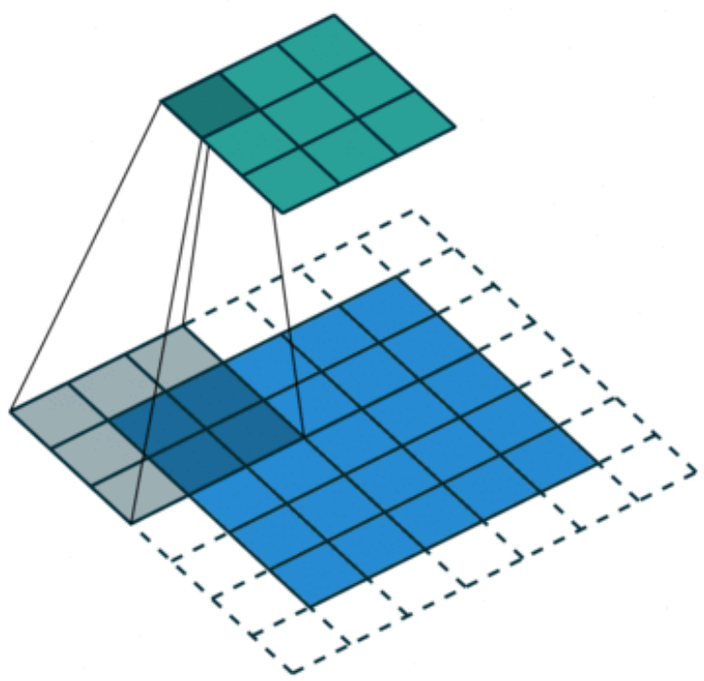
CONVOLUTIONAL NEURAL NETWORK

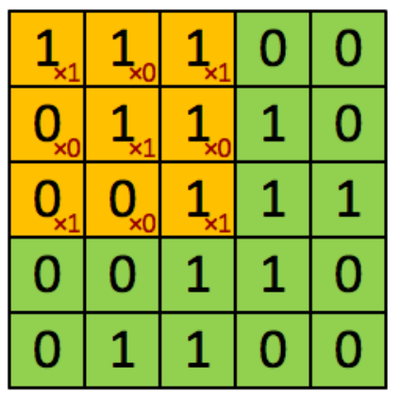
- Translation invariance
- 1D & 2D
- Optimizers
- Loss
- Initializers
- Learning rate
- Batch size

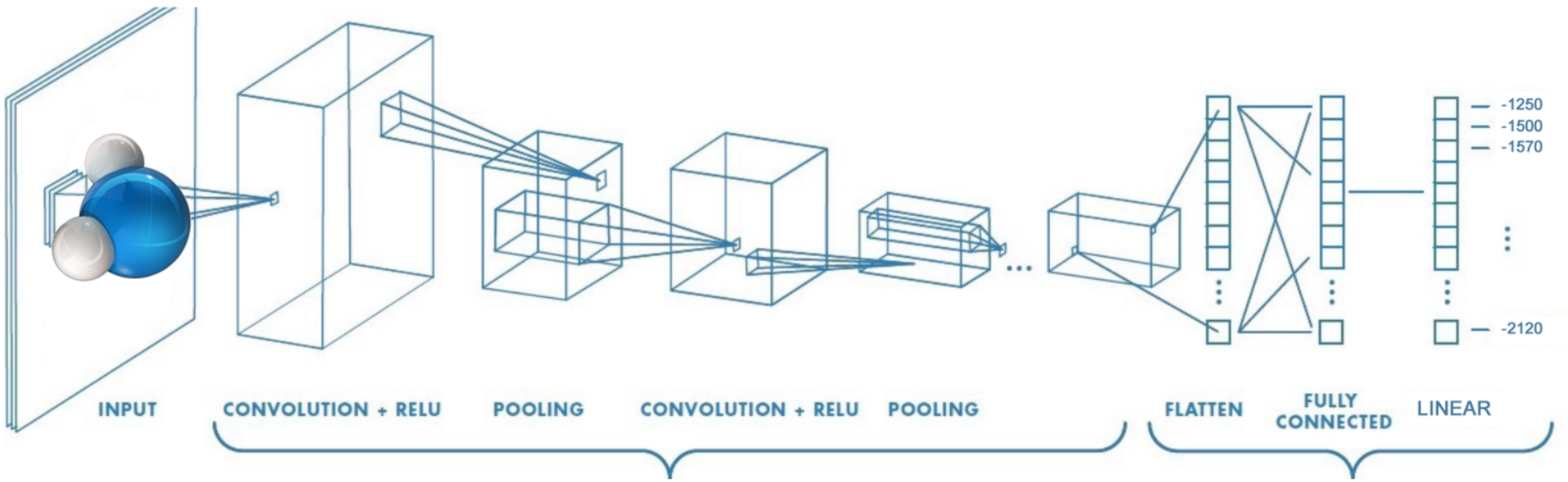


2D CONVNET

- Coulomb Matrix
  - (23, 23)
- Filters?







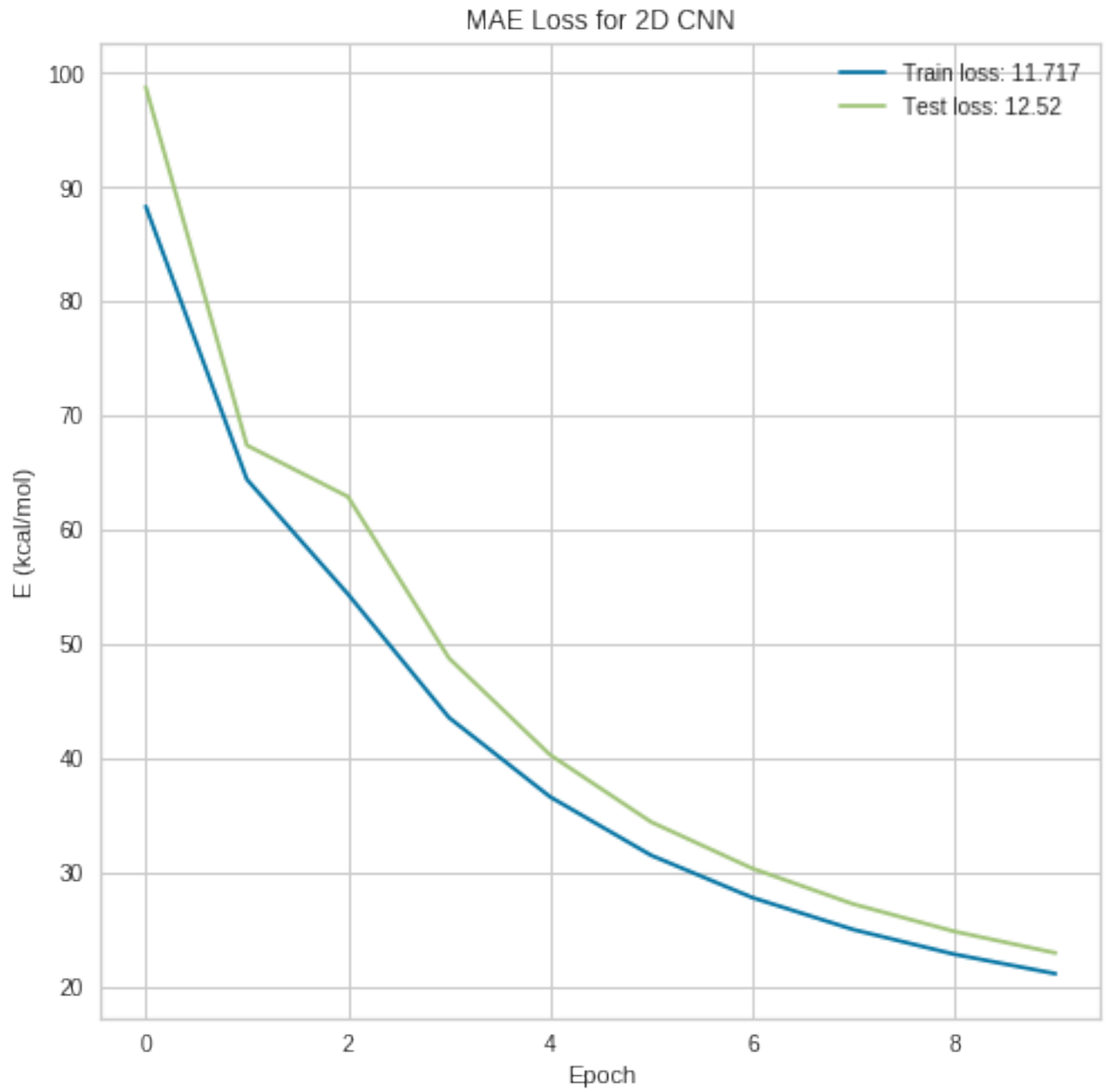
2D CONVNET

Layer (type)	Output Shape	Param #
conv2d_50 (Conv2D)	(None, 23, 23, 32)	1600
conv2d_51 (Conv2D)	(None, 19, 19, 64)	51264
conv2d_52 (Conv2D)	(None, 17, 17, 64)	36928
conv2d_53 (Conv2D)	(None, 15, 15, 64)	36928
conv2d_54 (Conv2D)	(None, 13, 13, 128)	73856
conv2d_55 (Conv2D)	(None, 11, 11, 128)	147584
flatten_27 (Flatten)	(None, 15488)	0
dense_44 (Dense)	(None, 1)	15489

Total params: 363,649
Trainable params: 363,649
Non-trainable params: 0

None
Execution time: 103.60028743743896
Epochs: 10

Train loss: 11.71735034630337
Test loss: 12.52040726607776



1D CONVNET

- Coulomb Matrix, unrolled
  - (529, 1)

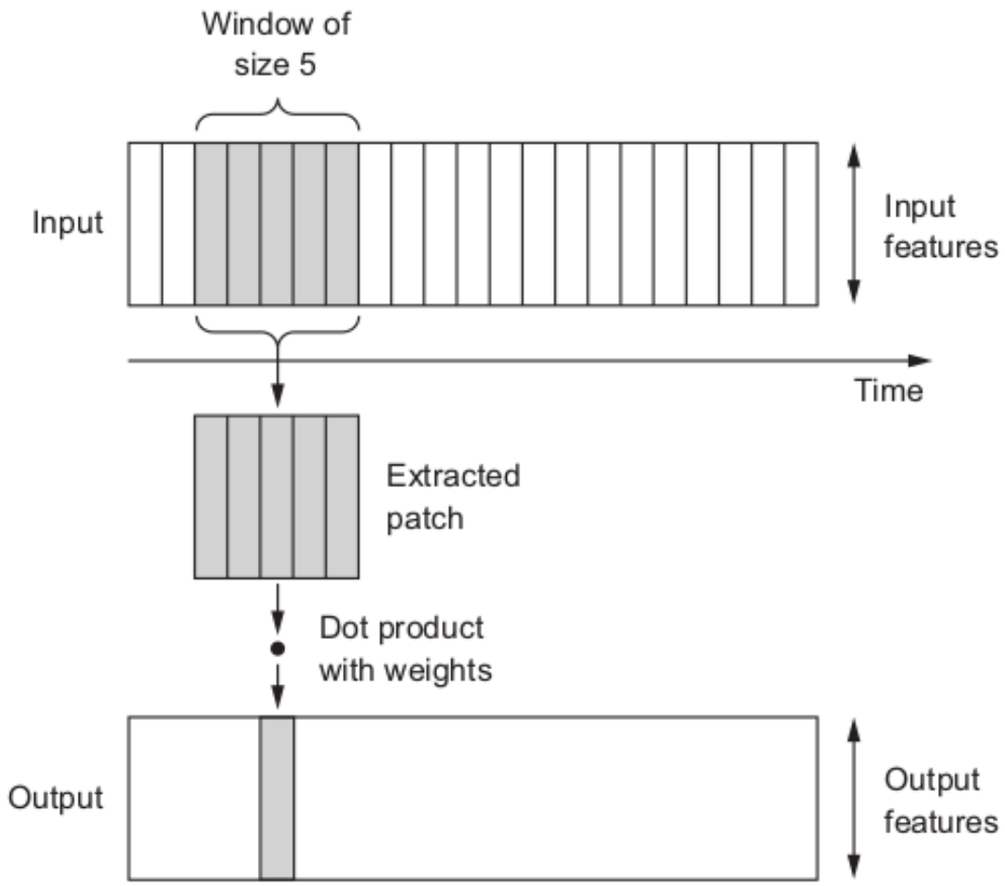


Figure 6.26 How 1D convolution works: each output timestep is obtained from a temporal patch in the input sequence.

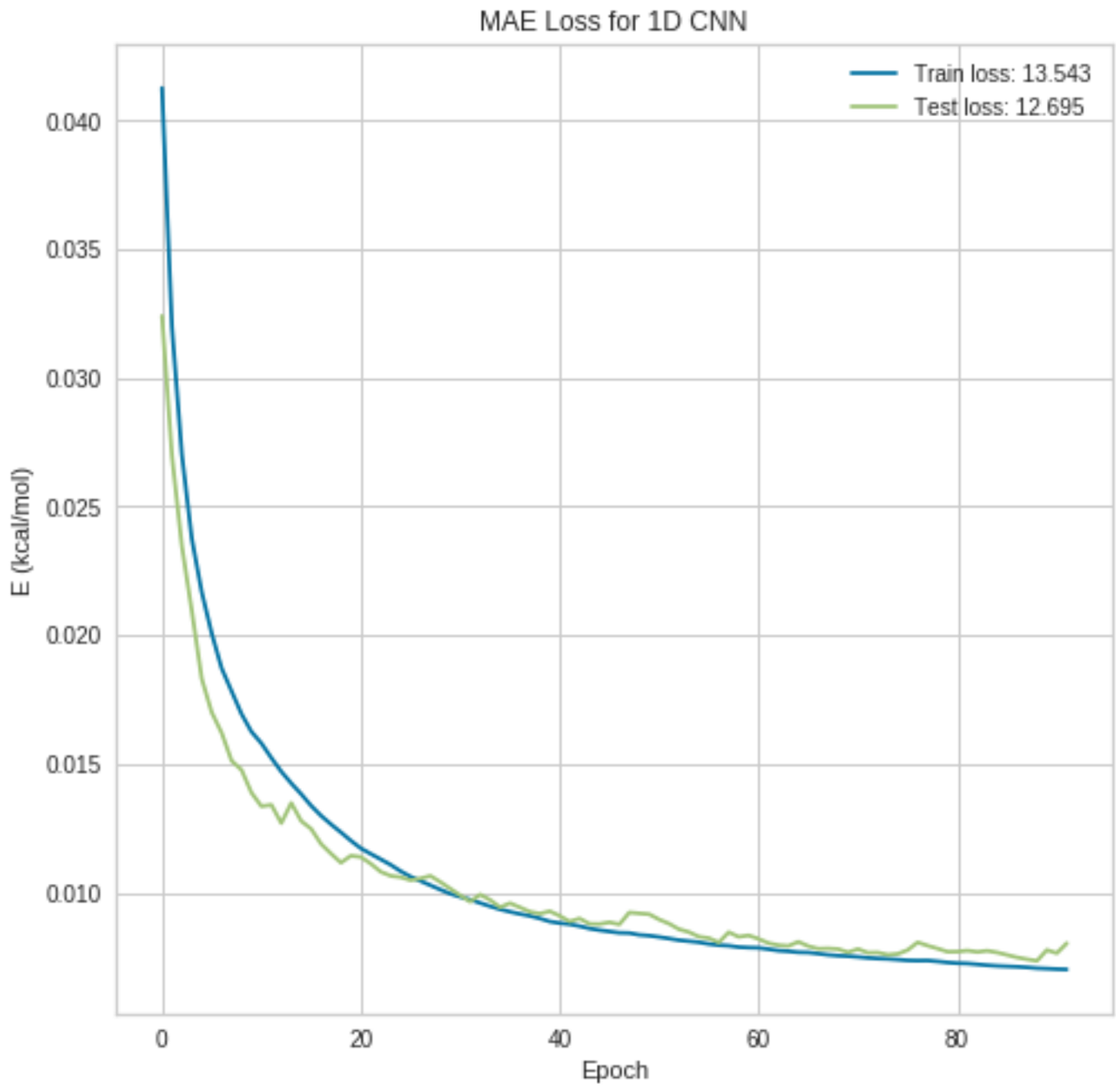
1D CONVNET

Layer (type)	Output Shape	Param #
conv1d_53 (Conv1D)	(None, 520, 32)	352
conv1d_54 (Conv1D)	(None, 514, 128)	28800
conv1d_55 (Conv1D)	(None, 510, 128)	82048
conv1d_56 (Conv1D)	(None, 508, 256)	98560
flatten_14 (Flatten)	(None, 130048)	0
dense_25 (Dense)	(None, 1)	130049

Total params: 339,809
Trainable params: 339,809
Non-trainable params: 0

Execution time: 81.10480189323425
Epochs: 10

Train loss: 10.986873955758139
Test loss: 11.813746383218408
Validation error: 11.457974091172218



CONCLUSION

- Results
  - Gradient Boosting: **8.31**
  - 1D CNN: **11.45**
  - 2D CNN: **12.52**
- Comparison
  - **8.56** - Linear Regression, Pande Group @ Stanford
  - **~10** - Rupp et al. 2012
  - **13.18** - Hansen et al. 2012
  - **41.81** - Himmetoglu 2016

QUESTIONS

- Shortcomings
- Furthermore