ReDI School of Digital Integration

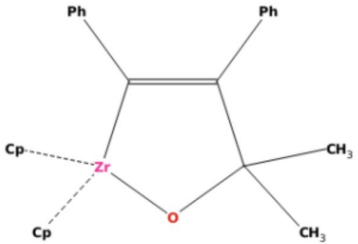
ML in Chemistry!

Data Analytics

Nesma Mousa
June 2023

We use tech to connect human potential and opportunity with dignity & humility

How does this look like?



(0D) Features

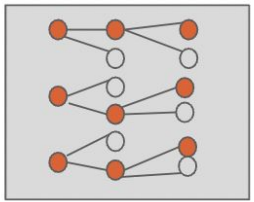
X1

X2

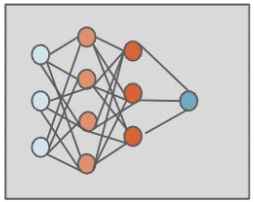
X3

Xn

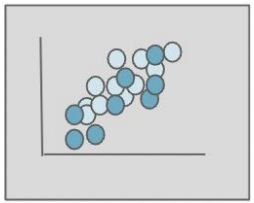
Tree_based



PyTorch_NN



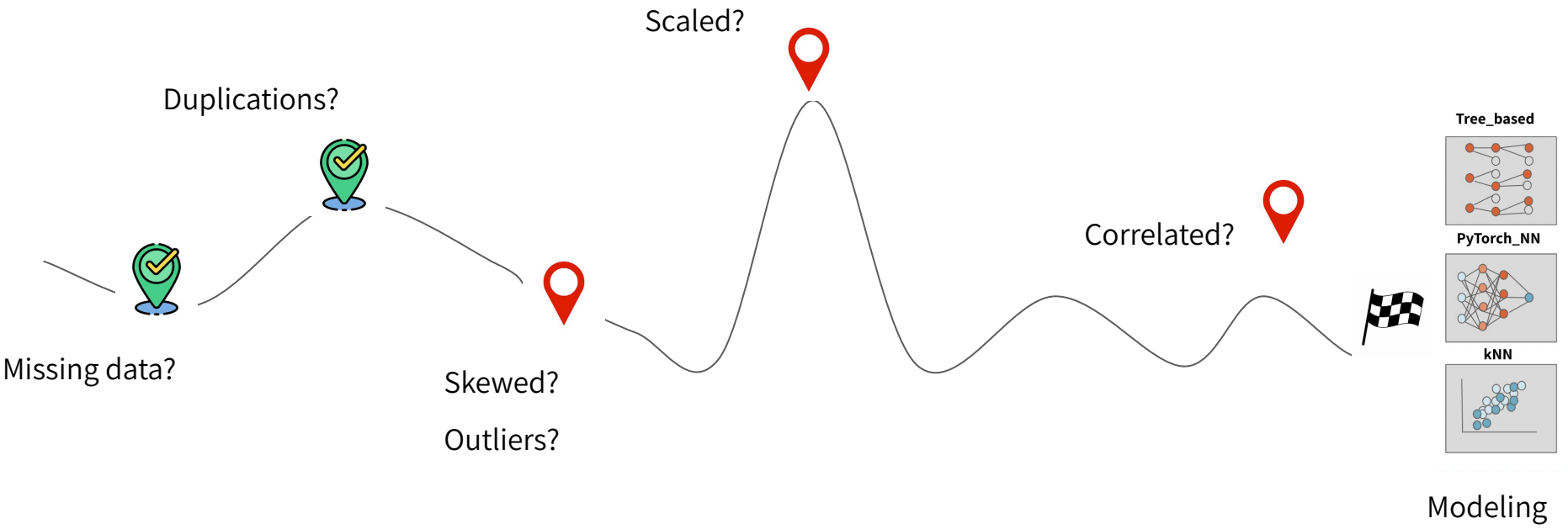
kNN



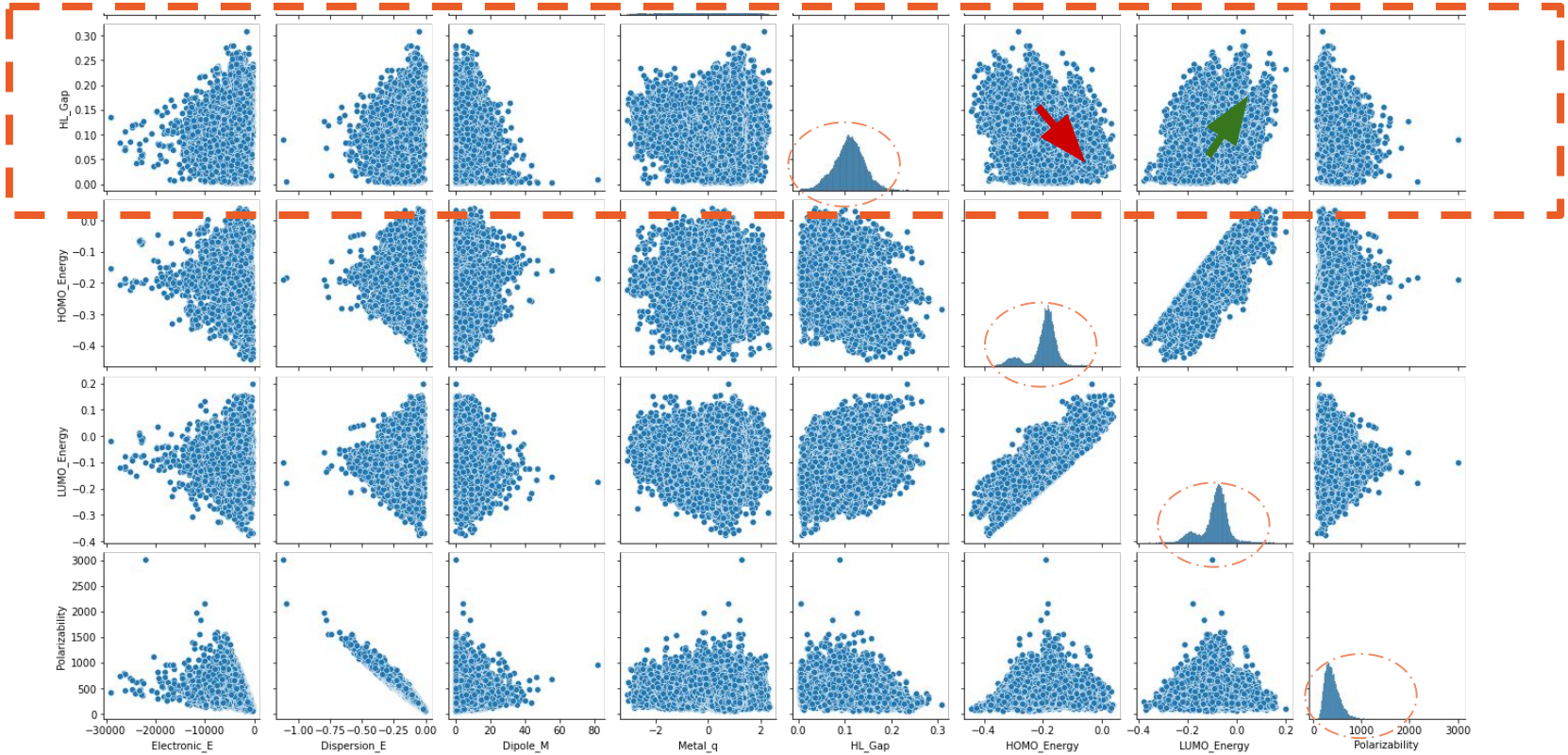
Target

y

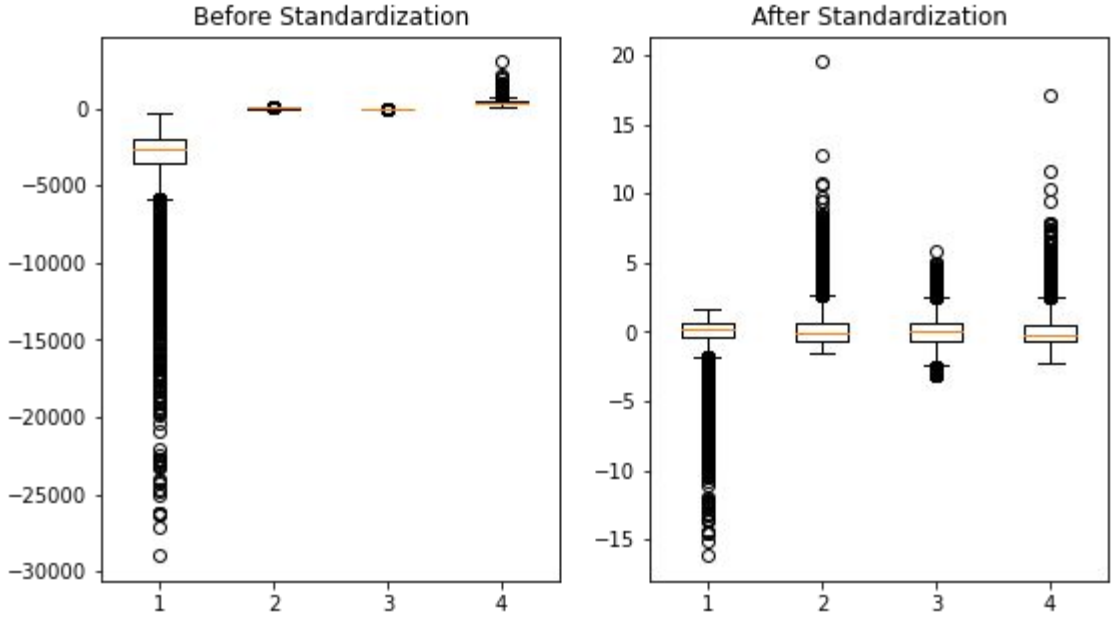
Data Checkpoints!



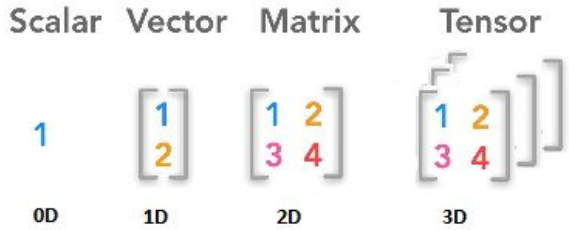
Zooming in!


HL_Gap
HOMO_Energy
LUMO_Energy
Polarizability
Electronic_E
Dispersion_E
Dipole_M
Metal_q
HL_Gap
HOMO_Energy
LUMO_Energy
Polarizability

Scaled!, Should we really bother?



Is that all? Happy with the Dimensions?



https://www.databricks.com/blog/2021/04/14/7-reasons-to-learn-pytorch-on-databricks.html

```
# Convert arrays to PyTorch tensors
x = torch.tensor(x, dtype=torch.float32)
y = torch.tensor(y, dtype=torch.float32)
```

Fitting | GridSearch-CV Time!

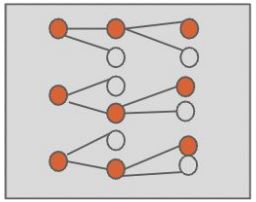
2D array
(86665, 7)

Scaled tensor-ed
(86665, 7)

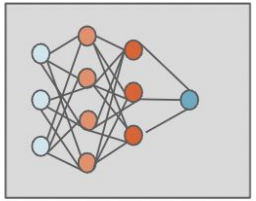
Scaled 2D array
(86665, 7)

80 / 20

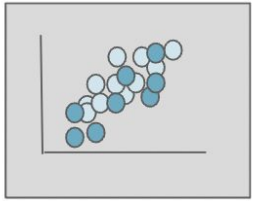
Tree_based



PyTorch_NN



kNN



RF: `n_trees in [50, 100, 300, 500, 1000, 2000, 5000, 10000]`

GBoost: `{'learning_rate': [0.1, 0.01, 0.001], 'subsample': [0.5], 'max_features': [6, 7]}`

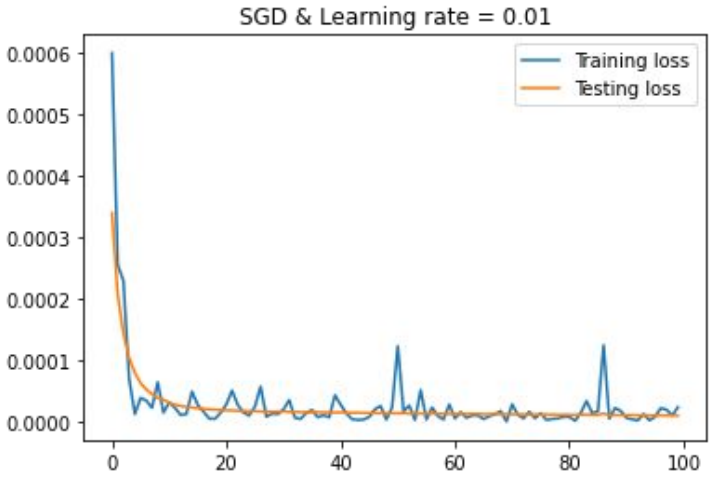
`Optimizer: Adam, SGD`
`Lr: 0.01, 0.001`

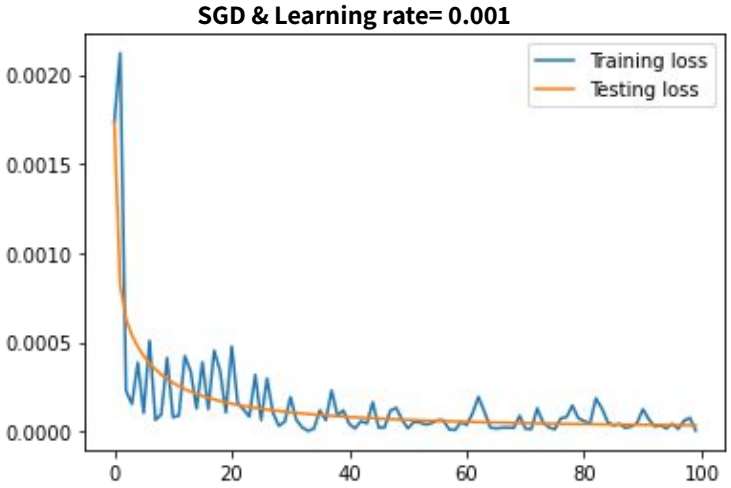
`k=1,...5,...,10`

Numbers?... Error Evaluation!

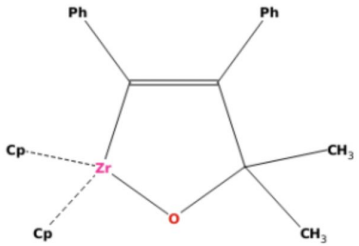
	RF	GBoost	NN_PyTorch	kNN
MAE	0.0375	0.0009		0.0064
MSE	5.362	1.389	Epoch 100 4.5907e-14	8.266
RMSE	0.0007	0.0011		0.0090
R2	0.9995	0.9987		0.927

Graphs?


SGD & Learning rate = 0.01
Training loss
Testing loss
0.0006
0.0005
0.0004
0.0003
0.0002
0.0001
0.0000
0
20
40
60
80
100


SGD & Learning rate= 0.001
Training loss
Testing loss
0.0020
0.0015
0.0010
0.0005
0.0000
0
20
40
60
80
100

Conclusion!



(0D) Features

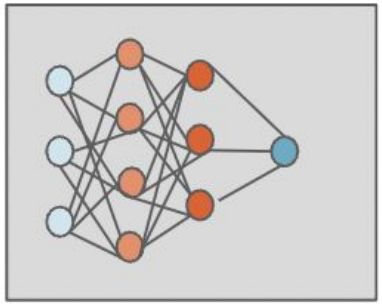
X1

X2

X3

Xn

PyTorch_NN



Target

y

(HL_GAP)

Thanks a lot!

Contact

Nesma Mousa