Materials & Mathematics

Solubility

Zihui(Andy) Liu
Ruei-Lun(Johnson) Chiang
Owen Chang-Chien
Vinsensius
Raymond Doerr

Sections

1. Overview
2. Hypothesis
3. Linear Regression (brief) + Neural Network (NN)
4. PCA + NN
5. KMC + NN
6. Discussion
7. Conclusion

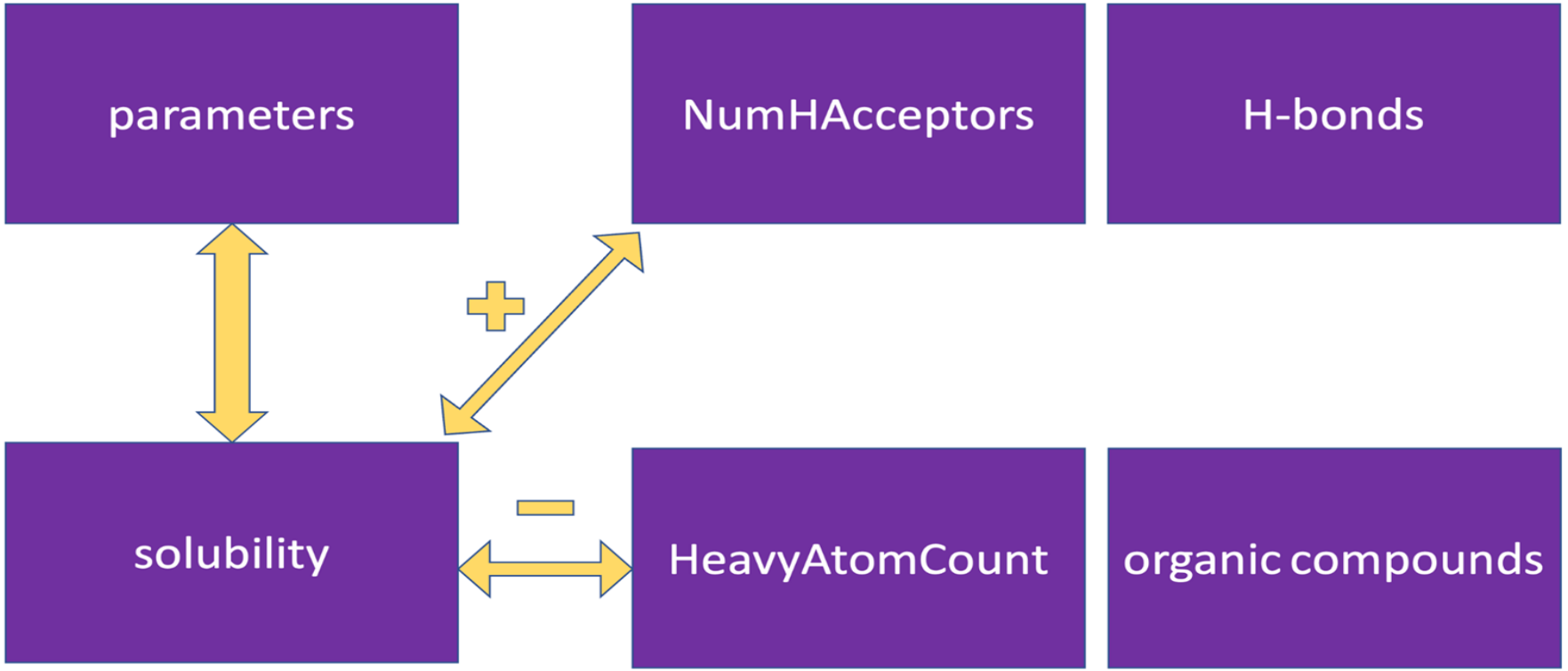
Overview

- The importance of solution?
  - We cannot live without talking about solutions

- Dataset: A curated aqueous solubility dataset
  - 9,982 unique compounds
  - 26 columns
  - 70% Training; 20% Validation; 10% Testing

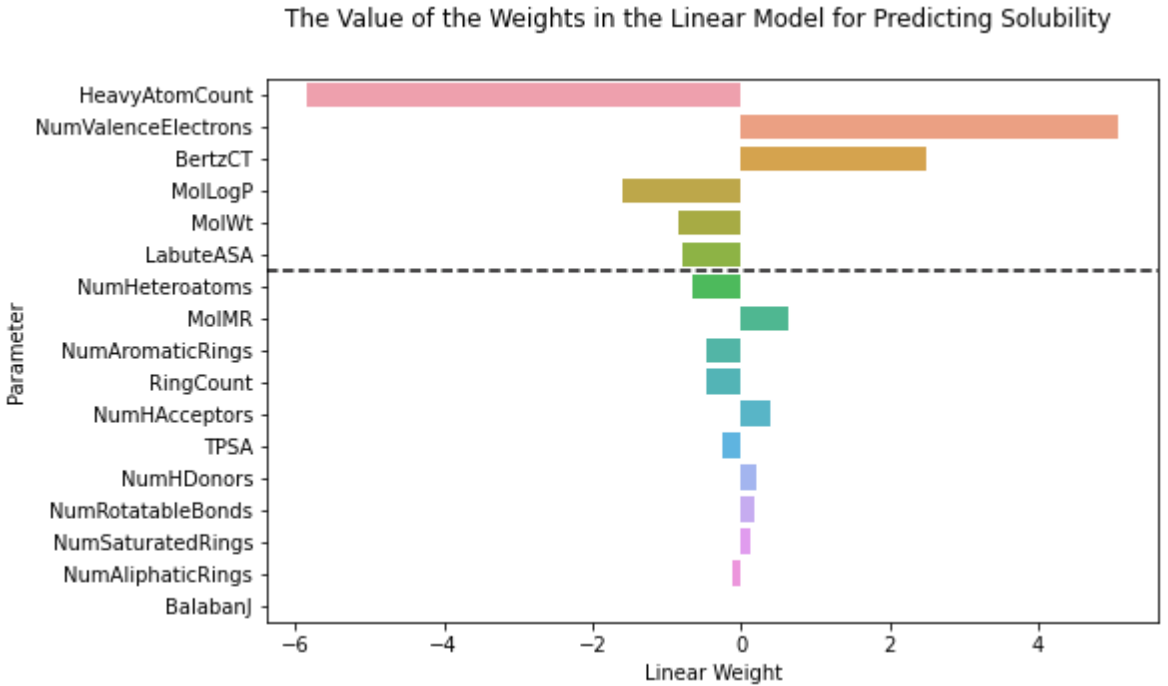
- Methods: Linear regression, Neural Network, PCA, and K-Means Clustering
  - graph comparison
  - MSE
  - Covariance

Hypothesis


parameters
NumHAcceptors
H-bonds
+
solubility
−
HeavyAtomCount
organic compounds

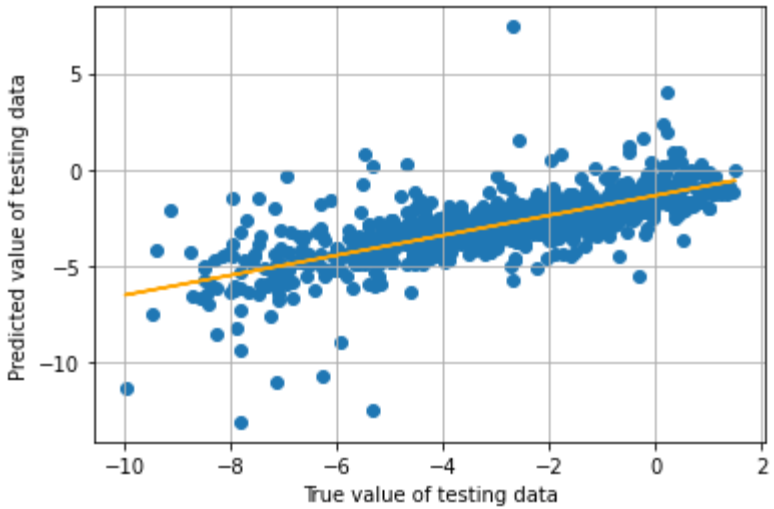
Linear Regression

HeavyAtomCount	-5.837487
NumValenceElectrons	5.074606
BertzCT	2.490288
MolLogP	-1.604043
MolWt	-0.847045
LabuteASA	-0.796701
NumHeteroatoms	-0.654110
MolMR	0.633050
NumAromaticRings	-0.481419
RingCount	-0.462177
NumHAcceptors	0.382826
TPSA	-0.254843
NumHDonors	0.207370
NumRotatableBonds	0.188156
NumSaturatedRings	0.124834
NumAliphaticRings	-0.124710
BalabanJ	-0.021462



Linear Regression

- Linear approach to modeling the relationship between a scalar response and one or more explanatory variables
- Covariance: 0.5071677649499687
- MSE: 2.6412435958194704



NN

- Both classification and regression

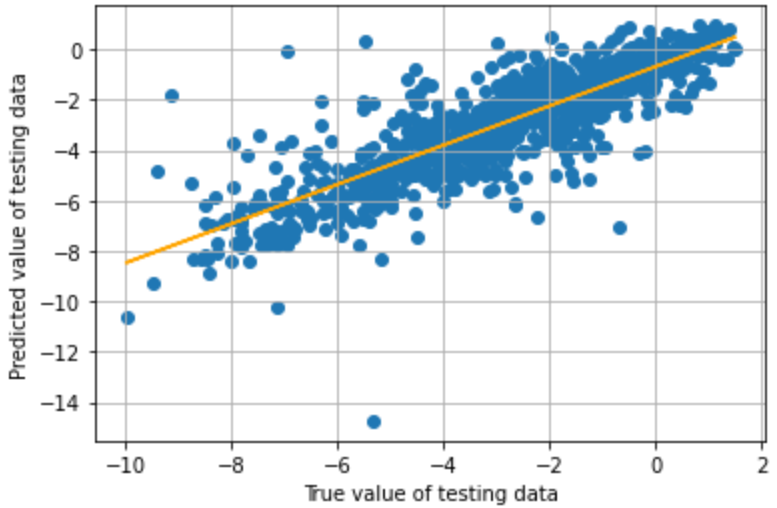
- Neural networking combines many different ideas

  - Linear and nonlinear regression
  - Ensemble methods
  - The input features often live in a

    higher-dimensional space

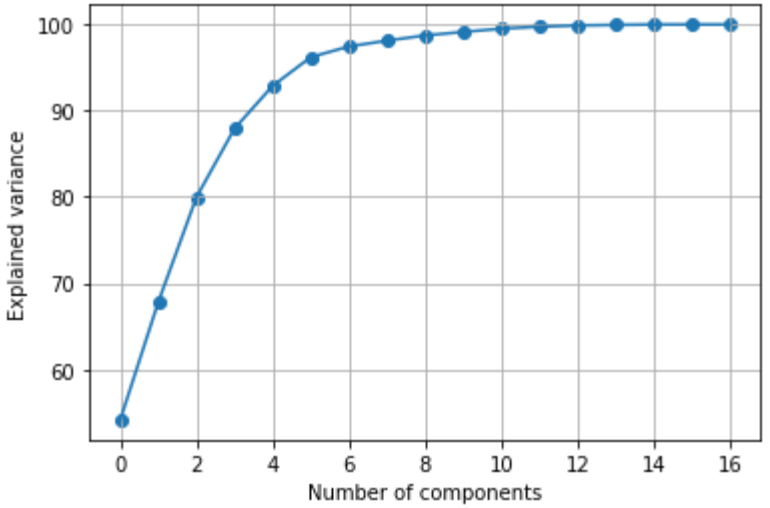
- Covariance: 0.6963961390293855

- MSE: 1.627108976289455

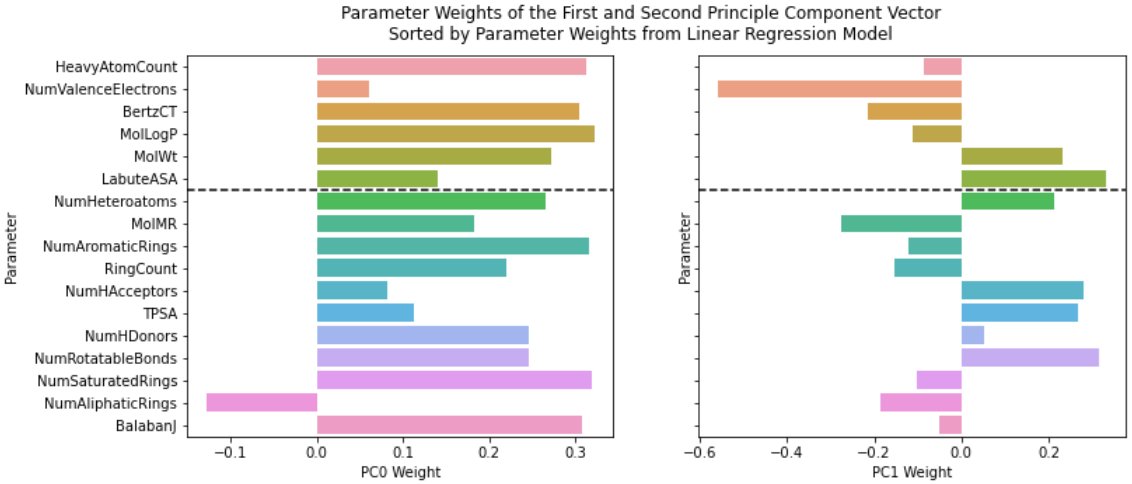


PCA + NN

- PCA is a unsupervised machine learning method
- Reduce the data so that we can focus on more important features of the data.
- To preserve 90% of the original data, we chose 5 components for PCA.
- After reducing the data size, we will train neural network using the reduced training data.

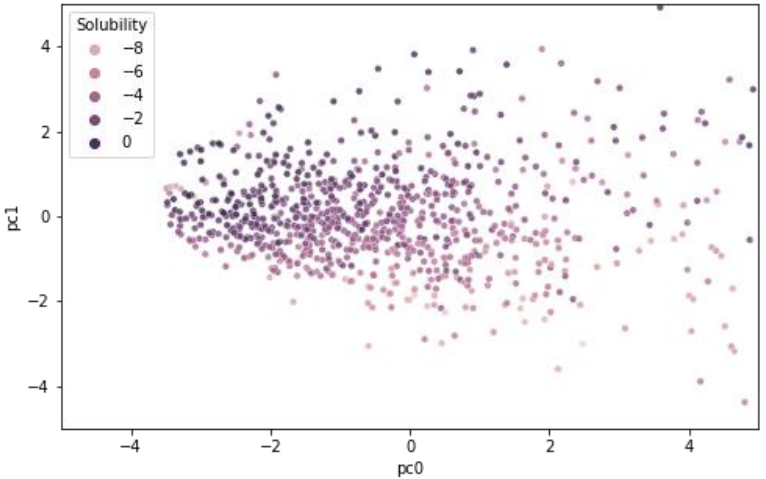


Interpretation: PCA result vs Regression Result



Plot of the first two components of PCA in terms of features

Plot of the first two components against the solubility

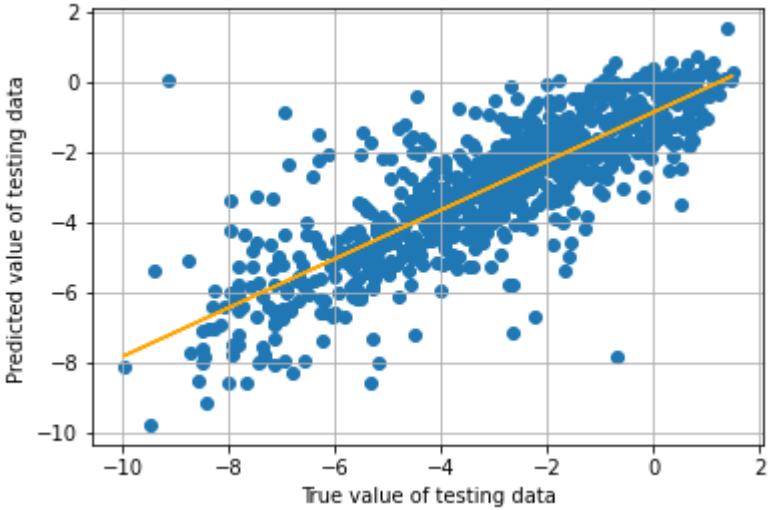


Discussion :

PCA+NN

Parity Data:

- Covariance: 0.6819
- MSE: 1.7048
- Most of the data is concentrated near the line.
- Few outliers are available in the plot due to loss of information from PCA.



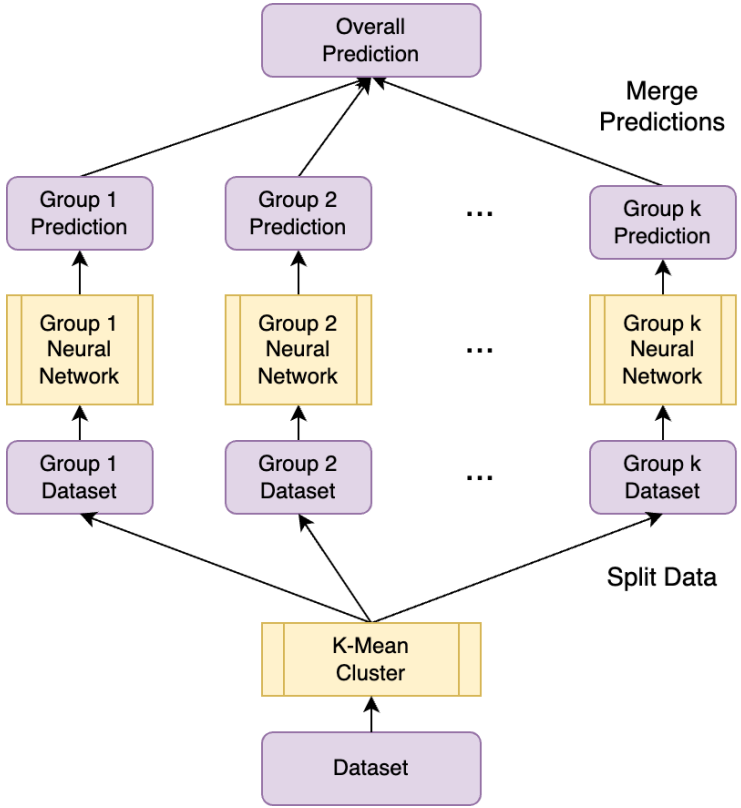
KMC + NN

- Neural network prediction over grouped data by K-Mean Cluster (KMC+NN)

- Motivation

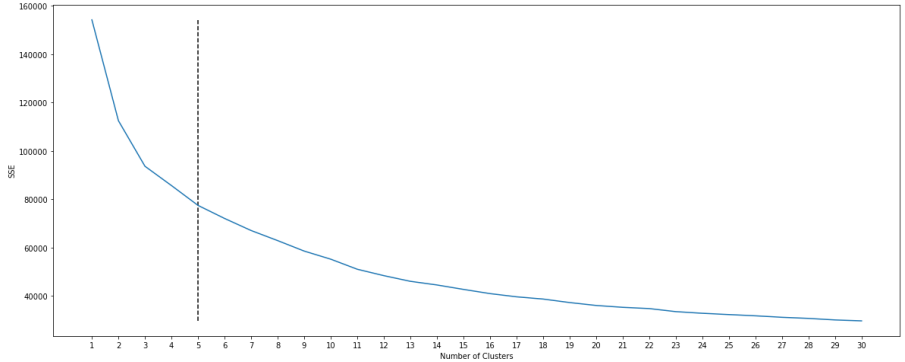
$$\forall i \ \mathrm{Var}(\text{Group } i) < \mathrm{Var}(\text{All Data})$$

  - Easier to make prediction on grouped data
- Pipeline
  1. Using K-Mean Cluster to group data. (k=5)
  2. Run Regular Neural Network on each Group
     a. Different Hidden Layer for each Group
  3. Merge the Group Prediction



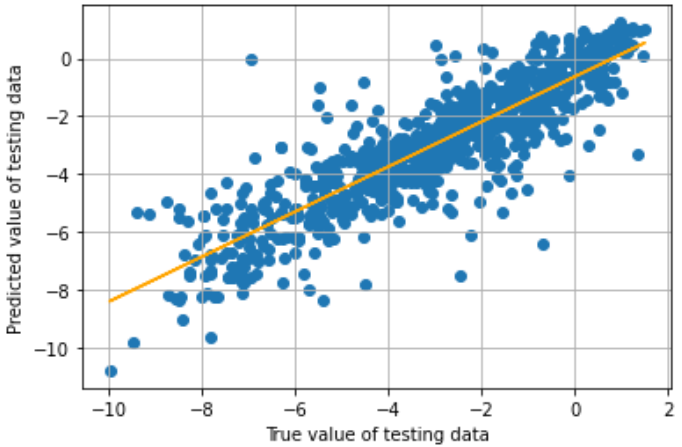
KMC + NN Result

- By SSE vs k plot, k=5 is a good trade-off



Sum of Standard Error (SSE) vs. Number of Clusters (k)

- Final MSE: 1.3191 (not always)
- Parity plot
  - Covariance: 0.7539
  - Data distributed evenly
  - No extreme outliers

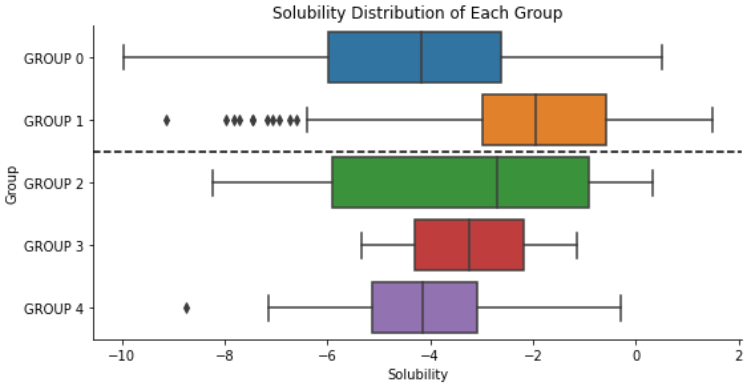


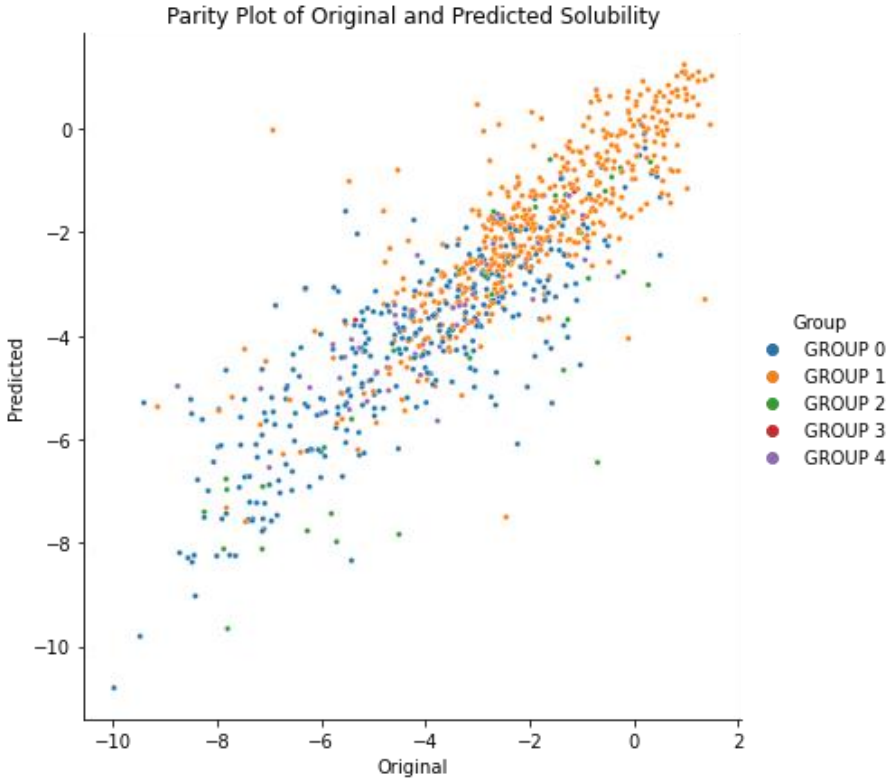
KMC+NN Parity Plot

KMC + NN, Interpretation On Clusters

- The group contribute part of the parity plot

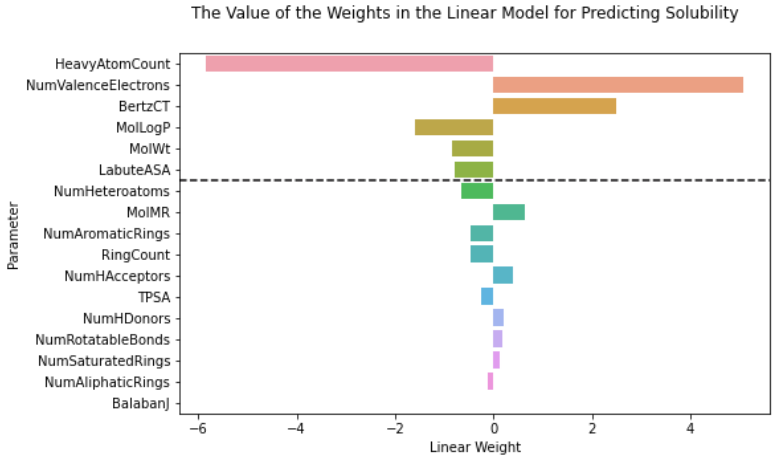
- **Group 0** and **Group 1** are important
  - **Group 0:** More Negative Region
  - **Group 1:** Less Negative Region

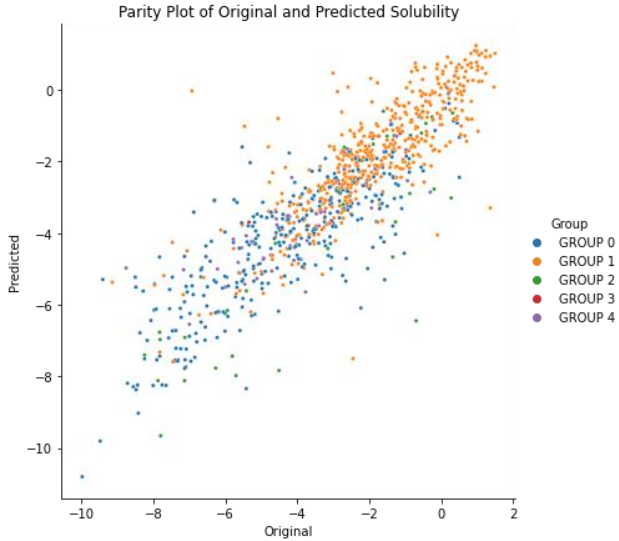




KMC + NN, Interpretation On Clusters

- First 6 parameter from Linear Regression
- **Group 0** and **Group 1** opposite parameter mean values
- Results are consistent with the linear regression





Discussion: Chemistry

Based on the value of the weights in the linear model for predicting the solubility I ruled these factors important and decided to evaluate these:

There is a trend in which positive correlation between **number of valence electrons** and **solubility**, the more full an electrons shell is the more stable that element is, and the less likely it will dissolve in water.

Additionally, with the topological complexity, **BertzCT**, the more surface area that a molecule has exposed the more water molecules that are available to interact, causing an increase in **solubility**.

Heavy Atom Count - with non H atoms, there are no hydrogen bonds, and with water and solubility, like dissolves like, and with this, this is a strong factor against **solubility**.

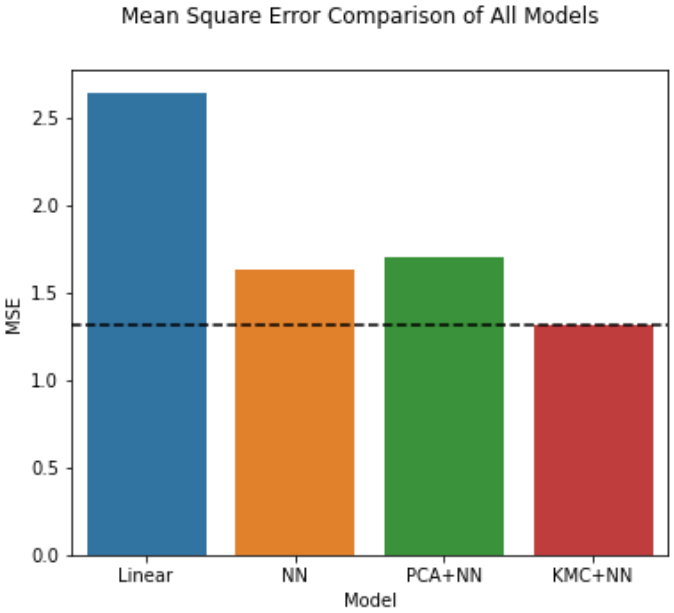
The **Octanol water partition coefficient(MolLogP)** serves as a relationship between fat solubility and water solubility of a substance. Greater than one if it is more soluble in fat like solvents and less than one if it is more soluble in water. This tells us that because there is a negative correlation some of the compounds are more soluble in Octanol than water.

For **LabuteASA**, it describes the approximate accessible surface area of the molecule. It is slightly positive correlation because the more surface area accessible the higher the solubility is.

Conclusion: Models

Final note about All models

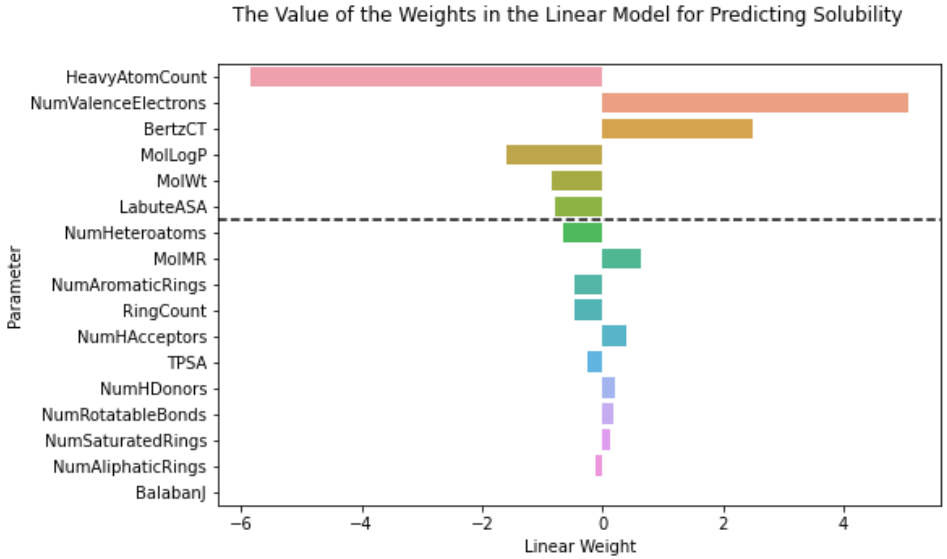
- Linear Model
  - Underperforming due to its model simplicity
- Neural Network
  - Doing as good as we expected
- PCA+NN
  - Performing slightly worse than plain neural network due to reduced dimensional data
- KMC+NN
  - Sometimes performs better, but highly depending on the group separation by KMC

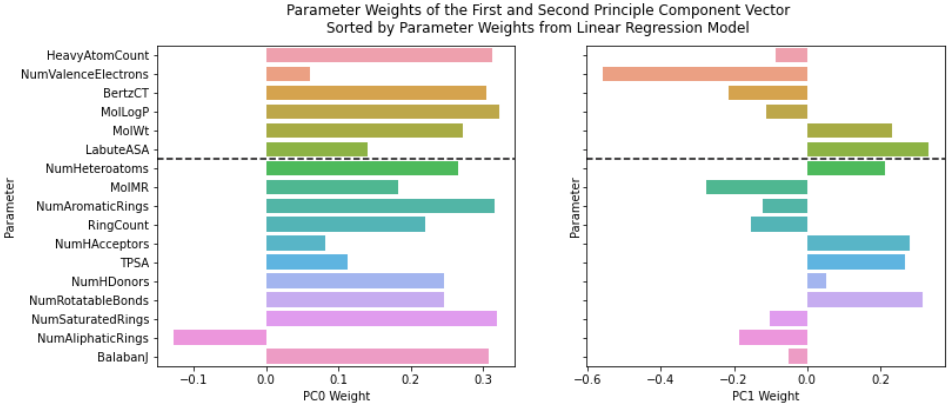


Conclusion: Predictors

Most important parameters

- HeavyAtomCount
- NumValanceElectrons
- BertzCT
- MolLogP
- MolWt
- LabuteASA





Thank you

Reference:

AqSolDB: A curated aqueous solubility dataset https://www.kaggle.com/datasets/sorkun/aqsoldb-a-curated-aqueous-solubility-datase

Python: Feature/Variable importance after a PCA analysis https://pyquestions.com/feature-variable-importance-after-a-pca-analysis

Keras API reference: https://keras.io/api/

Derivation and Applications of Molecular Descriptors Based on Approximate Surface Area: https://link.springer.com/protocol/10.1385/1-59259-802-1:261