DIRECT
Data Intensive Research
Enabling Clean Technologies

QSPR MODELING: APPLICATION OF MACHINE LEARNING ALOGRITHMS IN CLASSIFYING THE FAMILY AND PREDICTING FLASH POINTS AND CETANE NUMBER OF BIOFUEL COMPOUNDS

Contributors: Jingtian Zhang, Cheng Zeng, Renglong Zheng, Chenggang Xi

https://github.com/Zhangjt9317/Biofuel-Group-Project

Background

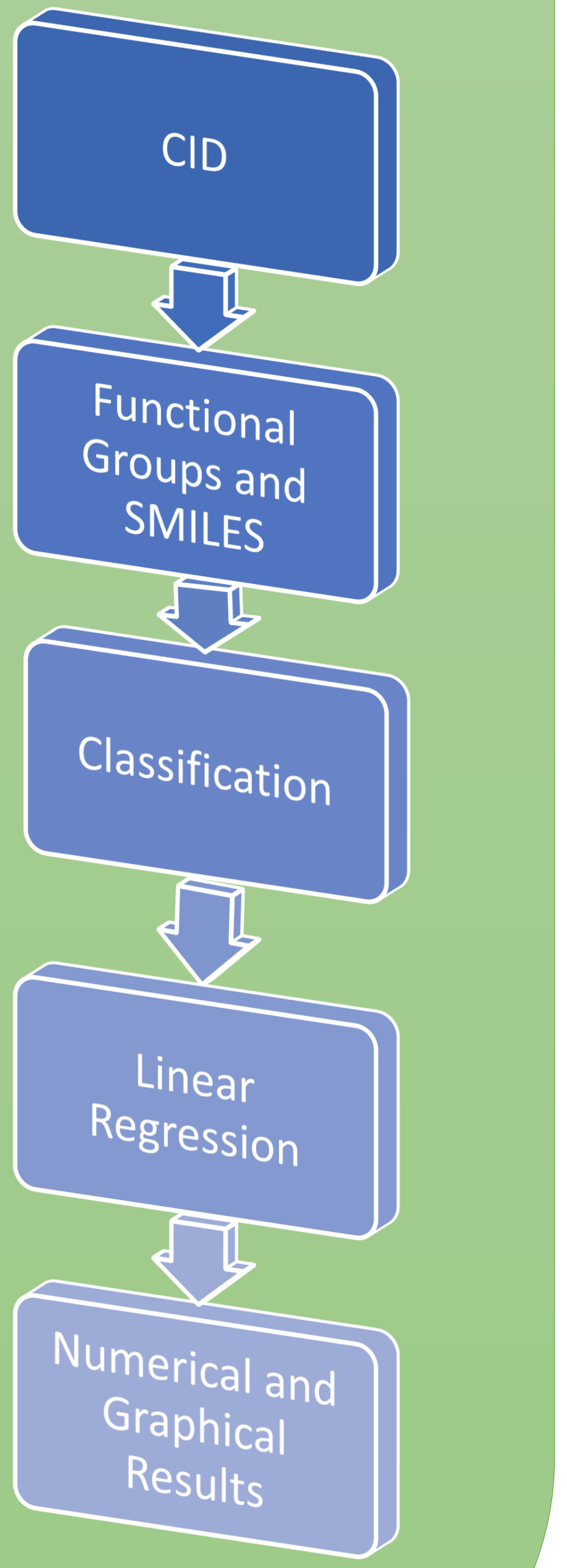
As the greenhouse emission rises in the past decades, **renewable energy technologies** have attracted attentions of many governments and private organizations, such as wind, hydro, solar and biomass are top charming choices to chase due to their availability.

Our group focuses on the prediction of properties of biofuel compounds by using **machine learning algorithms**. Molecular properties (flash points & cetane number), LUMO and HOMO can be predicted from training datasets if they are accessible are highly dependent on molecular structure.

Function and Overview

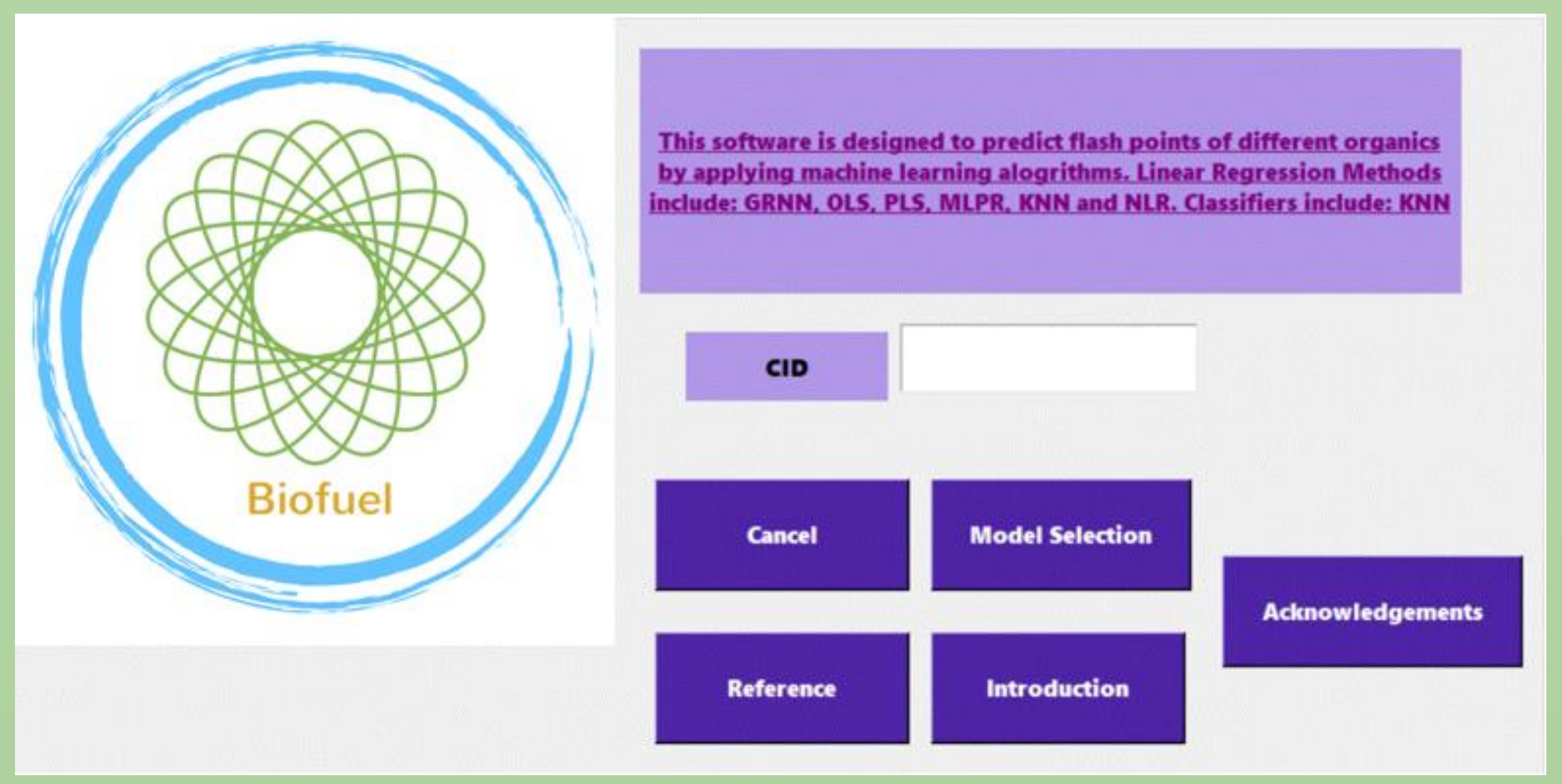
Our software applies the several **sklearn models** (K-Nearest-Neighbor, Support Vector Machine Regression, Linear Discrimination Analysis, Ordinary Linear Regression, Partial Least Square Regression, Polynormal Regression and Multi-Layer Perception Regression) and **neupy algorithm** (General Regression Neural Network) to **classify the input biofuel component** and **predict its physical properties** (Flash Point & Cetane Number).

In general, the users input the CID number for the biofuel component. The software uses **pubchempy** package to generate the SMILES (Simplified Molecular-Input Line-Entry System). Based on the **presence of the functional group**, the software could predict which family of the input biofuel belongs to. It is the classification step. Then, based on its family and physical property you are interested on, software will insert to regression models to predict its physical property, and give you the parity plot and mean-squared-error for the prediction result.

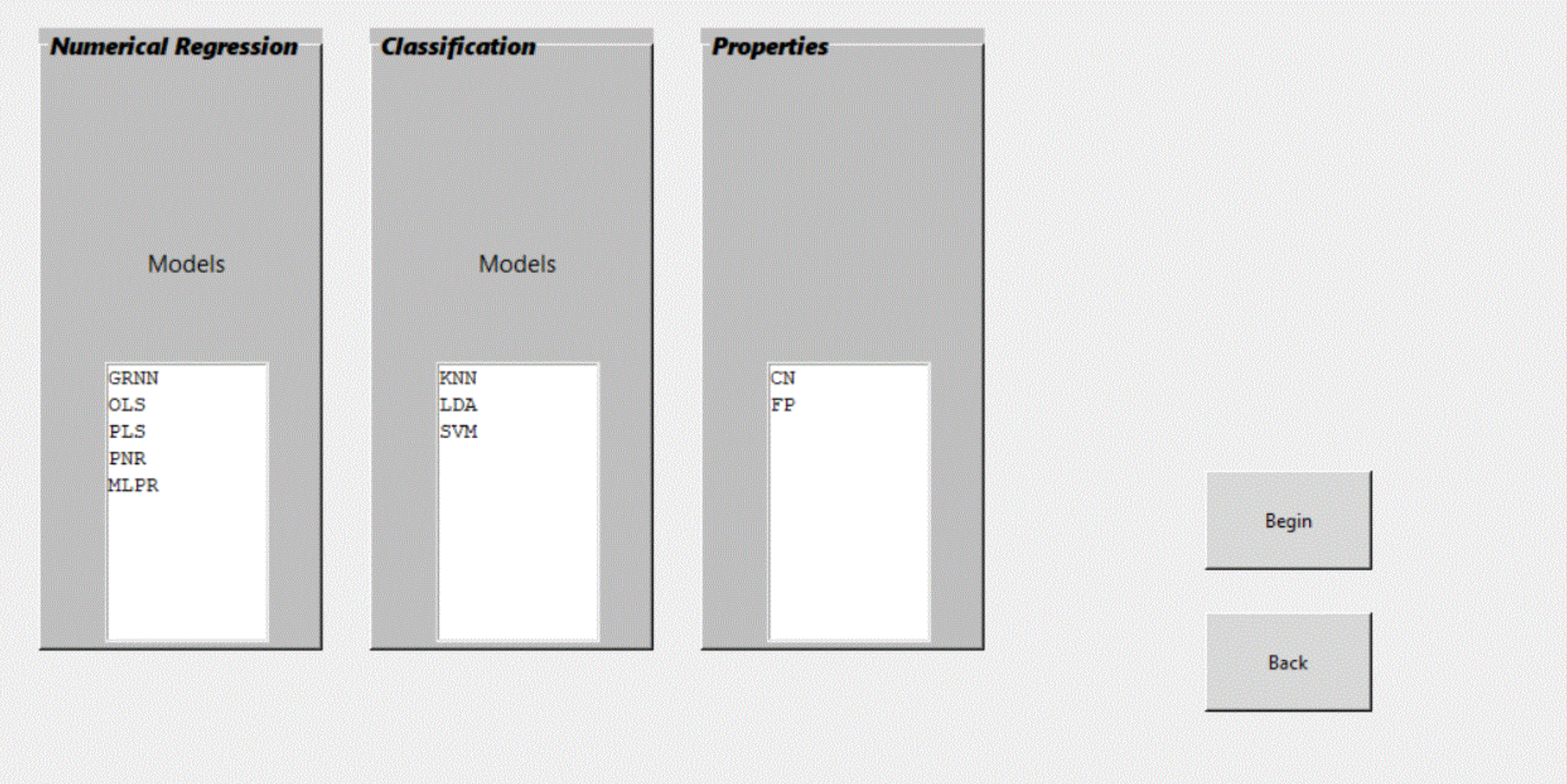

CID
Functional Groups and SMILES
Classification
Linear Regression
Numerical and Graphical Results

GUI Design

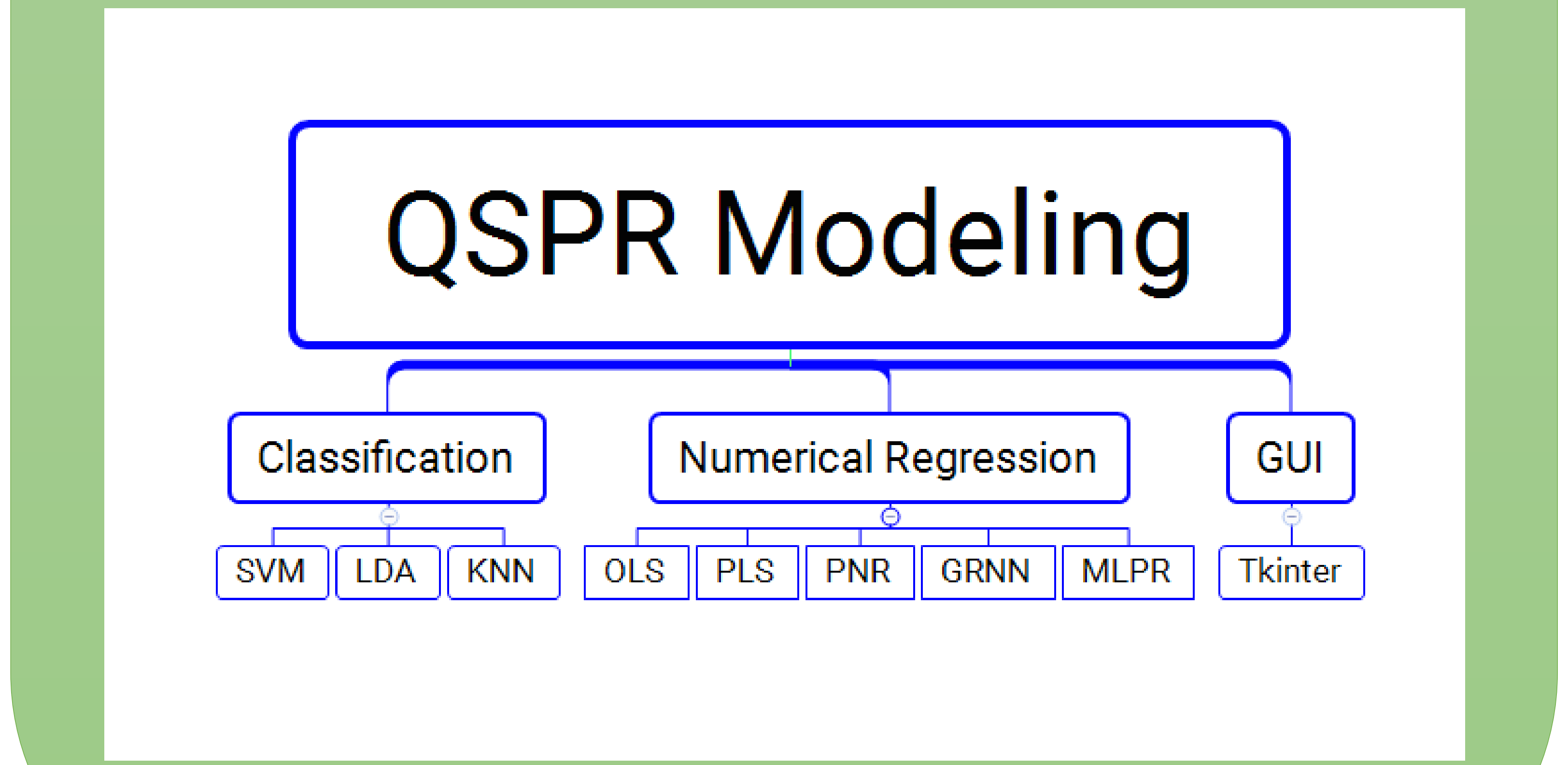
Our GUI is built by **Tkinter**. The first interface allows users to input the **CID number** of the biofuel.


This software is designed to predict flash points of different organics by applying machine learning alogrithms. Linear Regression Methods include: GRNN, OLS, PLS, MLPR, KNN and NLR. Classifiers include: KNN

At the second interface, users could start to select the different models to find the family of input biofuel component, and also predict the physical property using different regression models. The quality of the prediction model has been defined by the **parity plot**, **MSE value**, **R^2 value** and **accuracy**, which will be shown in the third interface.

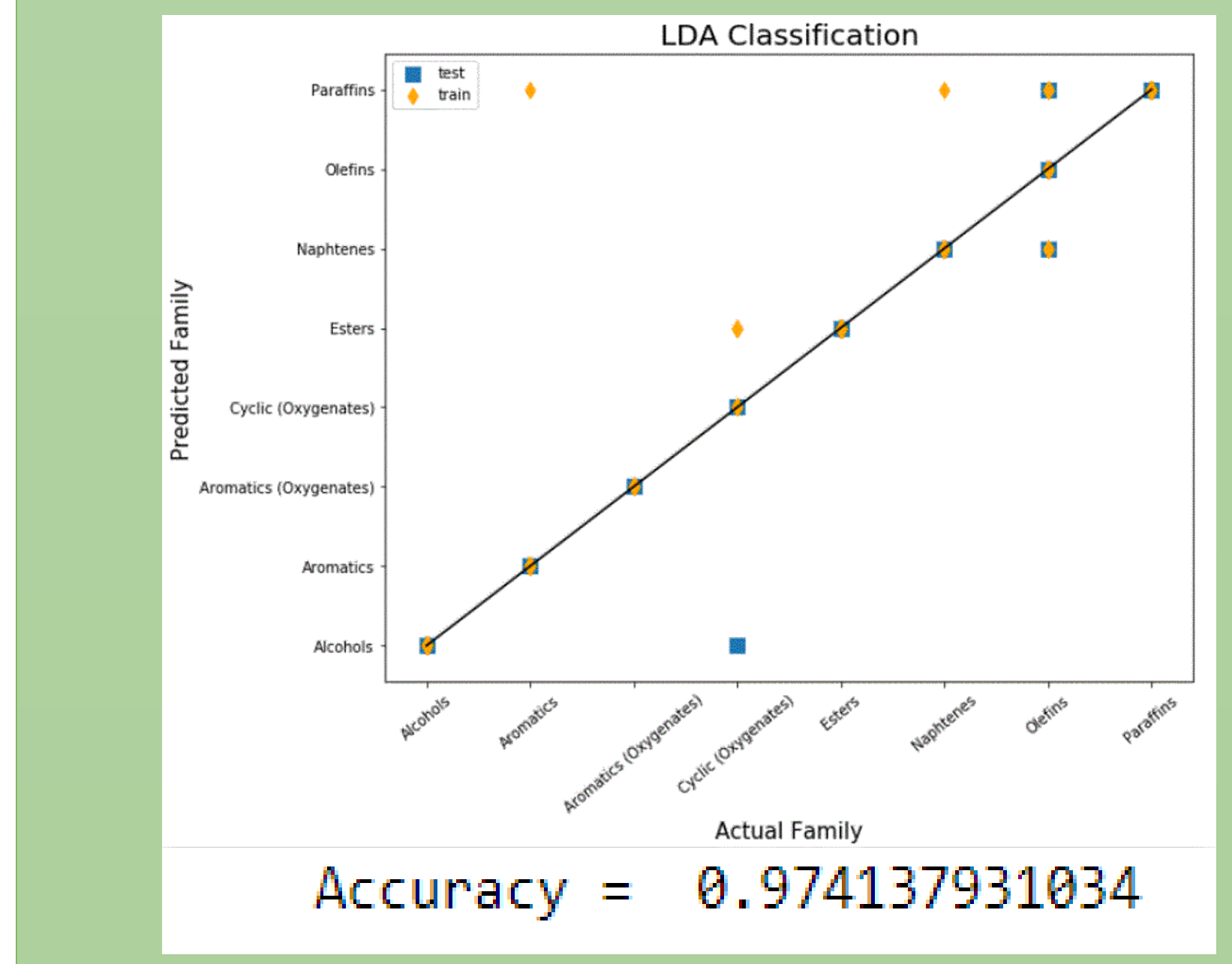


The following are choices contained in the machine learning portion, and the input number will go through the entire modeling process and generate **numerical** and **graphical results**.

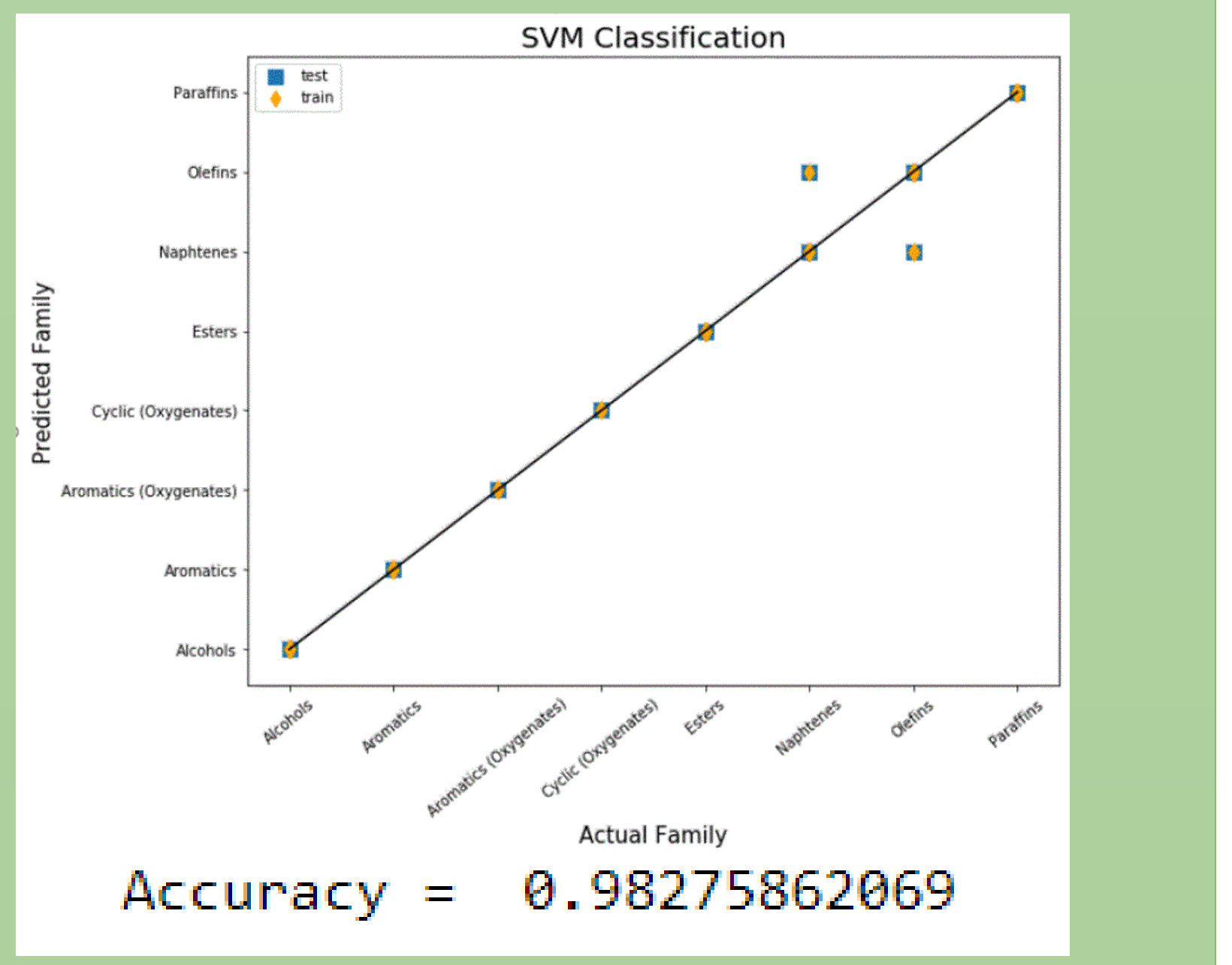

QSPR Modeling
Classification
Numerical Regression
GUI
SVM
LDA
KNN
OLS
PLS
PNR
GRNN
MLPR
Tkinter

Output Results

In the classification step, two different outstanding classification models have been displayed here. By comparing between the actual family and the predicted family, we could find our classification have a **good prediction performance**, based on their accuracy calculation.

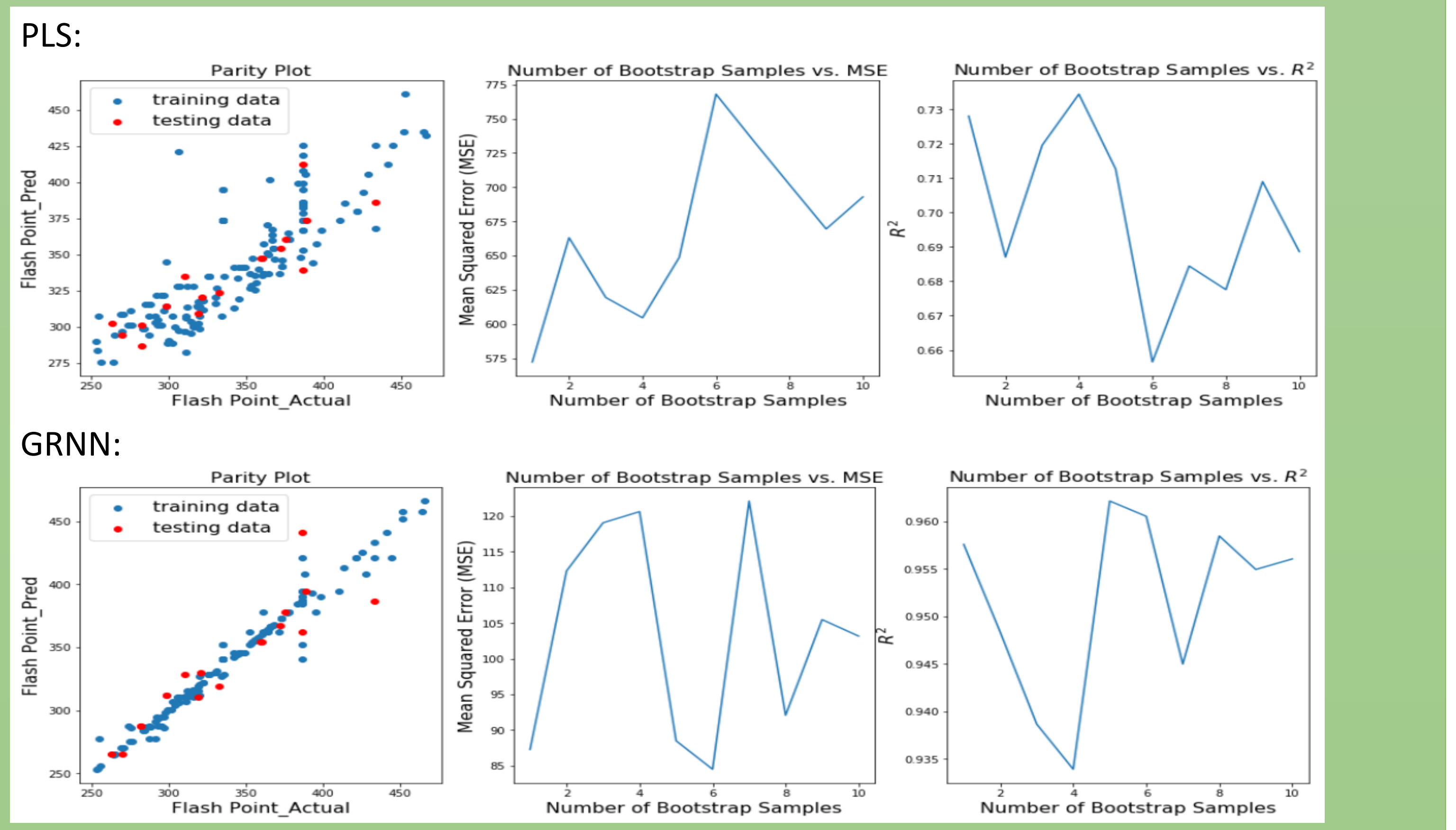


```
Accuracy =  0.974137931034
```



```
Accuracy =  0.98275862069
```

In the physical property prediction step, two different outstanding prediction models have been shown below. From the parity plots, we could see that our prediction models have a **good linear correlation behavior**. Also, we applied the bootstrap method to normalize our input data. We could see the mean-squared-error is **reverse proportional** to the R^2 value. From the bootstrap vs. MSE plot, it could tell us the best prediction in each prediction iteration.



Reference

Saldana, D. A., Starck, L., Mougin, P., Rousseau, B., & Creton, B. (2013). Prediction of flash points for fuel mixtures using machine learning and a novel equation. *Energy and Fuels*, *27*(7), 3811–3820.