In [ ]:

```
import numpy as np
import random
import matplotlib.pyplot as plt

from sklearn.model_selection import train_test_split
from sklearn import preprocessing
from sklearn.utils import shuffle

from sklearn.gaussian_process import GaussianProcessRegressor
from sklearn.gaussian_process.kernels import Matern, RBF,WhiteKernel
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score

from modAL.models import BayesianOptimizer, ActiveLearner, CommitteeRegressor
from modAL.acquisition import max_EI
from modAL.disagreement import max_std_sampling

import seqlogo

import copy

### Set random seed
seed = 5
random.seed(seed)
np.random.seed(seed)
```

In [ ]:

```
data = np.loadtxt('hw3_data.csv', dtype = str, delimiter = ',')[1:]
### TO DO
```

In [ ]:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
```

In [ ]:

```
### TO DO
### please edit this block, and feel free to remove any code we provided for you.

regressor = ActiveLearner(
    X_training=X_train[0].reshape(1, -1), y_training=y_train[0].reshape(1, -1)
)
n_queries = 200
for idx in range(n_queries):
    query_idx = np.random.randint(len(X_train))
    X_train, y_train = (np.delete(X_train, query_idx, axis=0), np.delete(y_train, query_idx))
    regressor.teach(X_train[query_idx].reshape(1,-1), y_train[query_idx].reshape(1,-1))

y_pred_final = regressor.predict(X_test, return_std=False)
y_train_pred=regressor.predict(X_train, return_std=False)
r2=r2_score(y_test,y_pred_final)
r2_train=r2_score(y_train,y_train_pred)
print(y_test,y_pred_final,y_train_pred)
print("R2",r2,r2_train)
```