1. Introduction:

This paper contains an comparative analysis of the reproduced graphs from the paper (https://pubs.acs.org/doi/10.1021/acs.jpcc.0c05993) using the ML models that our team created. Using our ML models, we reproduced important features the for Random Forest and Gradient Boosting models, reproduced predicted vs observed diameter plots for multilinear regression and reproduced the diameter distribution plot for our augmented dataset. Unlike the paper, we had extended the predictions to two more parameters: absorbance and photoluminescence (PL).

It is important to note that the models we created for this project were from a different dataset than the one explored in the JPCC paper. We used a subset of the original dataset, and we removed entries in the original dataset that did not have both absorbance and photoluminescence data entires.

2. Analysis
    a. Feature Importance

Feature importance is needed to identify the most important variable in the synthesis of CdSe. Figures 1 , 2 and 3 are plotted from our algorithms. Figure 1 shows the most important variable for our best prediction model: extra trees. It shows that time, growth temperature and cadmium acetate were the most influential parameters. For our computed random forest, time, growth temperature, and phosphines were the most influential ones. For our gradient boosting machine algorithm, growth temperature, time, and phosphines were the most influential. From the paper, both the random and gradient boosting machine algorithms had time of reaction, temperature, and metal precursors as the most influential parameters.

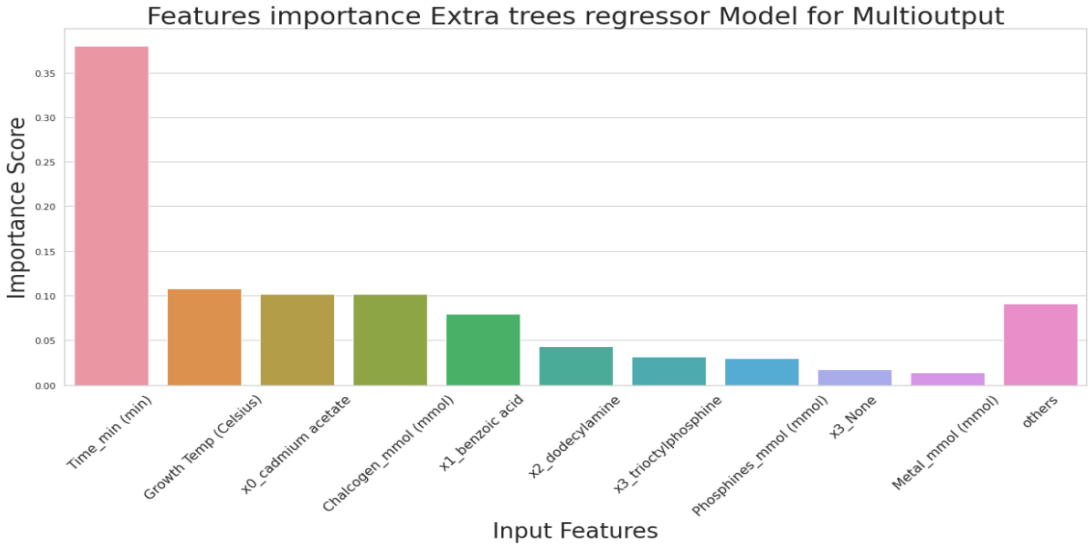

Figure 1: importance of the variable in the extra trees regression model

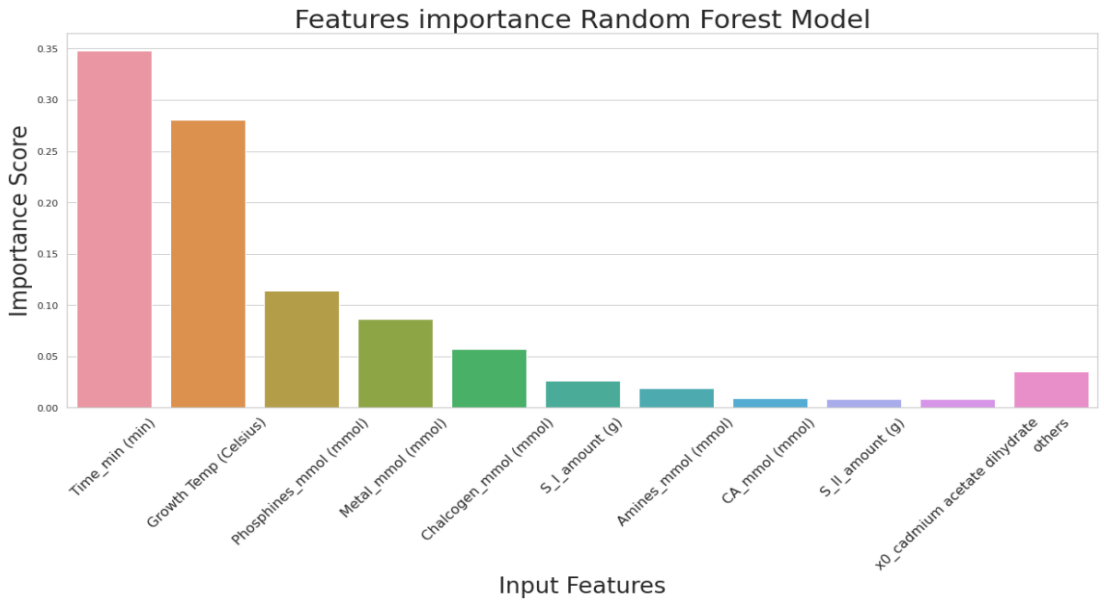

Figure 2: importance of the variable in the random forest regression model

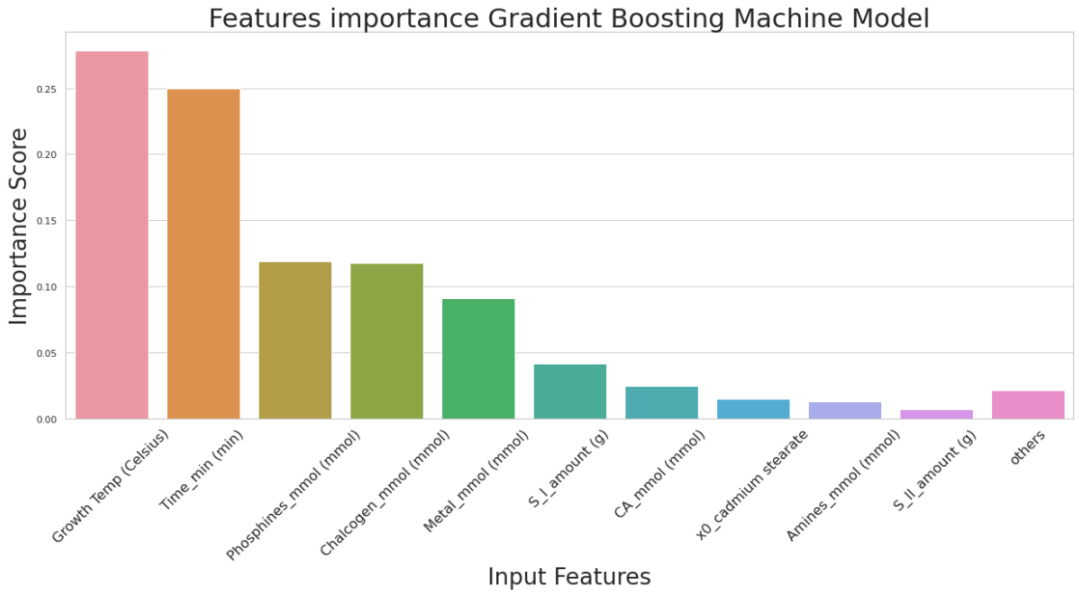

Figure 3: importance of the variable in the gradient boosting machine regression model

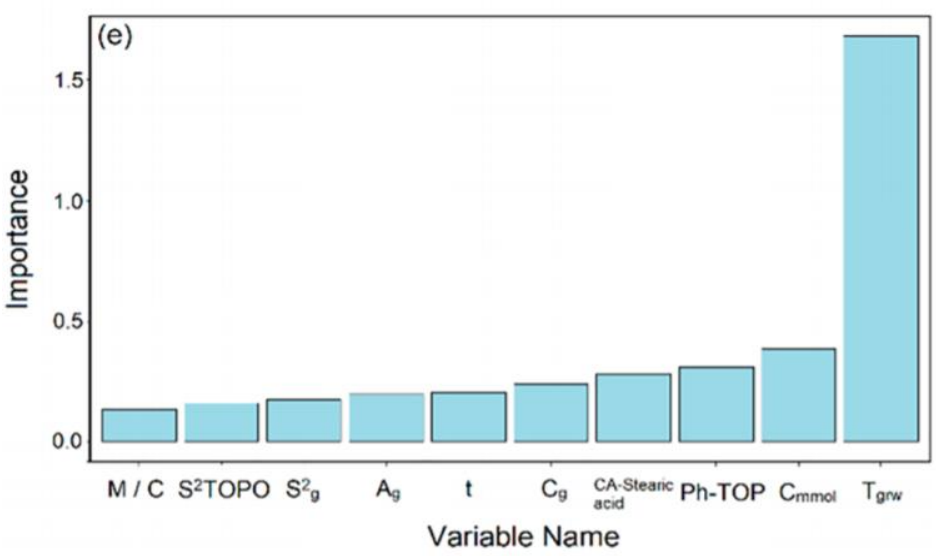

Figure 4: importance of the variable in the random forest regression model [1]

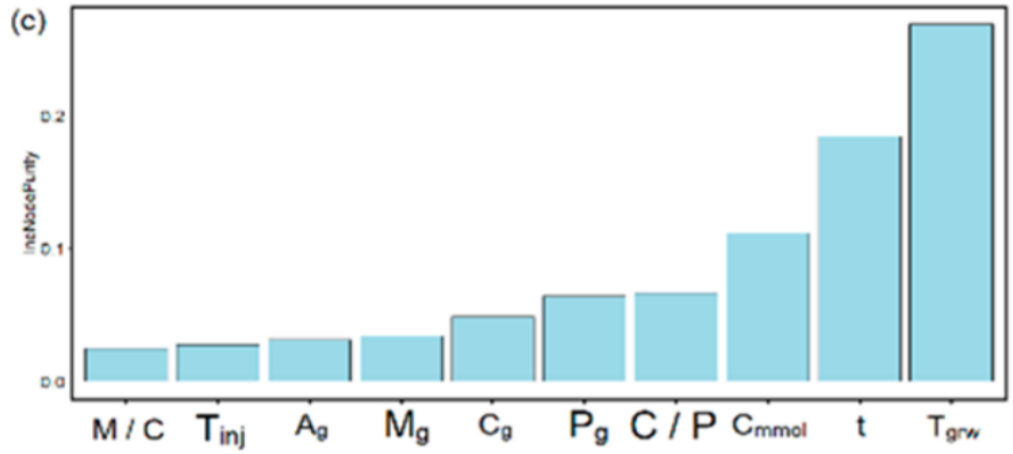

Figure 5: importance of the variable in the gradient boosting machine regression model [1]

A quick glance suggests that our top 3 most important features are not the same with what the JPCC paper found. We have 2 hypothesis why this may be the case. The first is that the JPCC paper's models used different parameter values than the ones that were used for our analysis. The second is that the testing/training datasets are different. Not only did we know the random seed that was used to split up the training and testing dataset, the dataset that we used for our project was also a subset of the original dataset as we removed the entries from the original dataset that did not have both absorbance and photoluminescence values.

b. Predicted vs observed parameter

We plotted the predicted vs observed diameter for our multilinear regression model (figure 7). Figure 6 represents the same plot from the paper. Compared to the plots from the paper, ours have the same behavior. The distribution, however, will not be the same since the plot is produce from random splitting of the training dataset. Unless we would use the same random seed than the paper, we cannot really reproduce the same distribution.

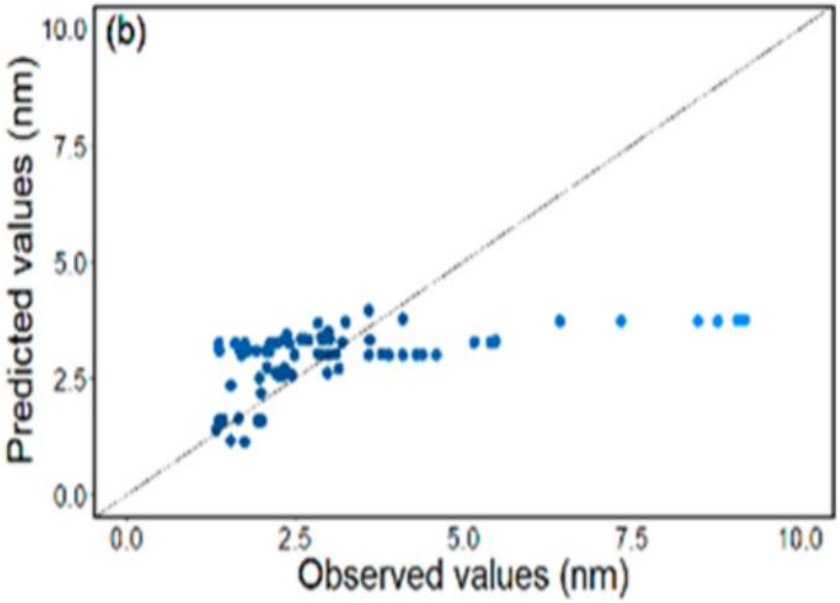

Figure 6: predicted vs observed plot of the multilinear regression model [1]

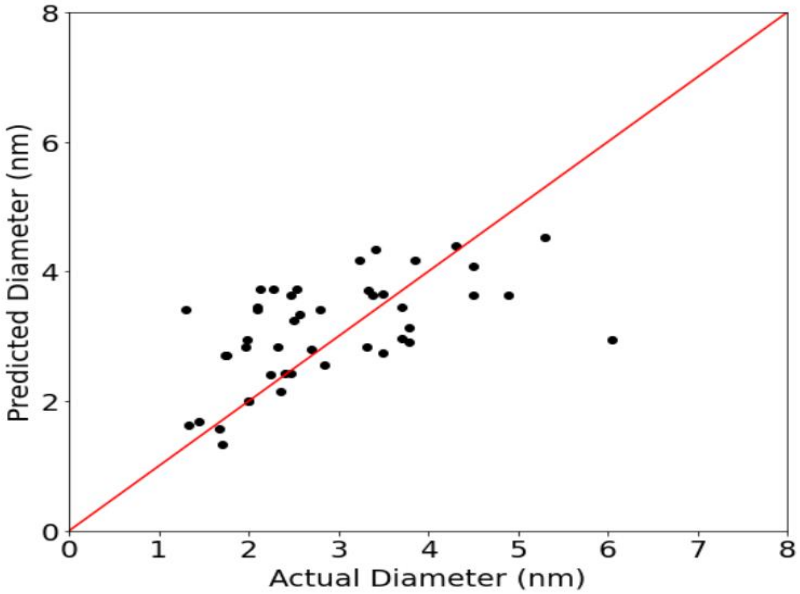

Figure 7: predicted vs observed plot of the multilinear regression model [1]

c. Performance metric for the algorithm

Here we compare the R^2, RMSE and MAE errors of our models with those reported in the JPCC paper. Only the best performance that optimize all 3 output was reported. Figure 8 represents the comparison of single output vs multioutput. When compared to figure 9 a and b, which is from the paper, our performance did just as good. For single output, prediction of absorbance performed better than prediction of diameter and PL. From figure 9b, we found that the gradient boosting R2 score to predict diameter is 0.93 while from our gradient boosting using multioutput is 0.70 (figure 12). For the paper, the gradient boosting model was the best performer while for our models, the extra trees using multioutput variable is. Using extra tree were able to predict absorbance, diameter, and PL with a R2 score of 0.94, 0.81, 0.78, respectively. To represent the best performance, we opted to optimize parameter. This explained why our R2 score for diameter prediction isn't as high as the paper performance.

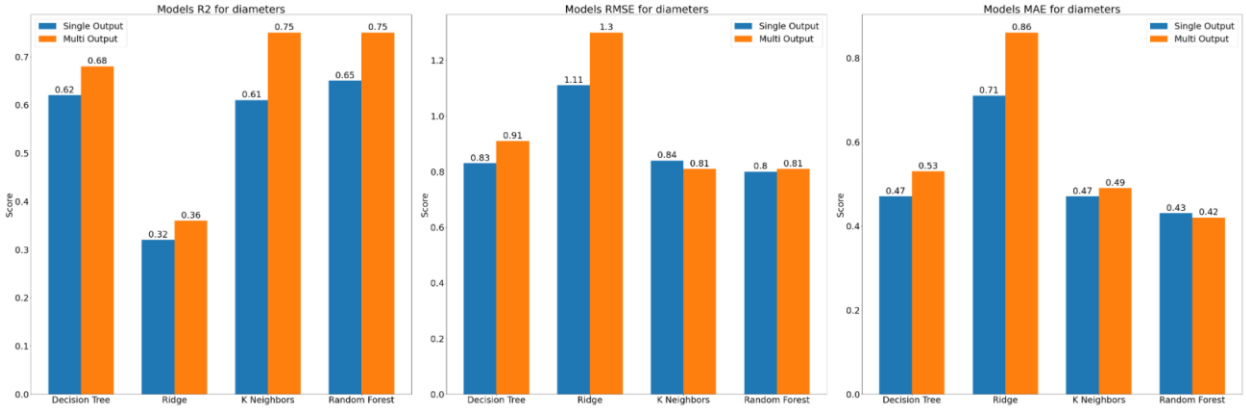

Figure 8: Performance metrics for the algorithm tested in the CdSe data set for diameter

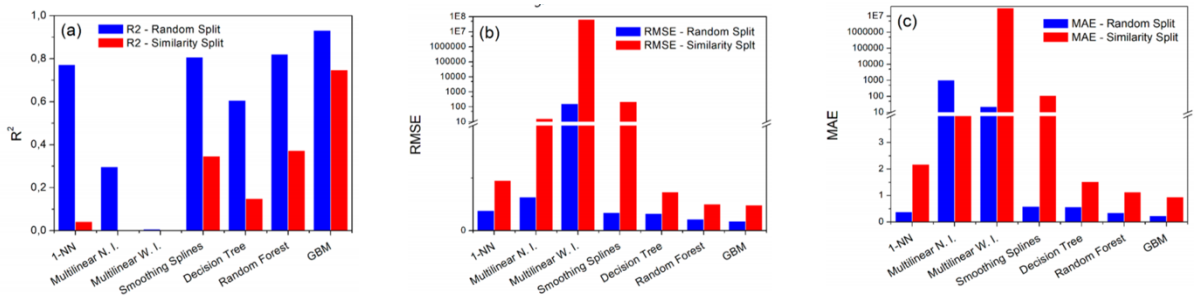

Figure 9 a: Performance metrics for the single output vs multioutput algorithm tested in the CdSe data set for diameter [1]

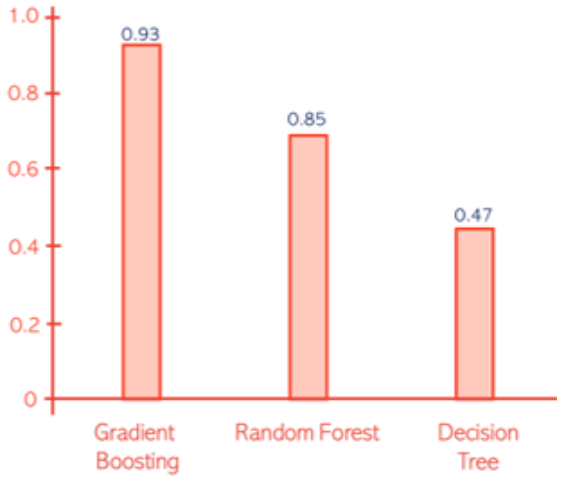

Figure 9 b: Performance metrics for the single output vs multioutput algorithm tested in the CdSe data set for diameter

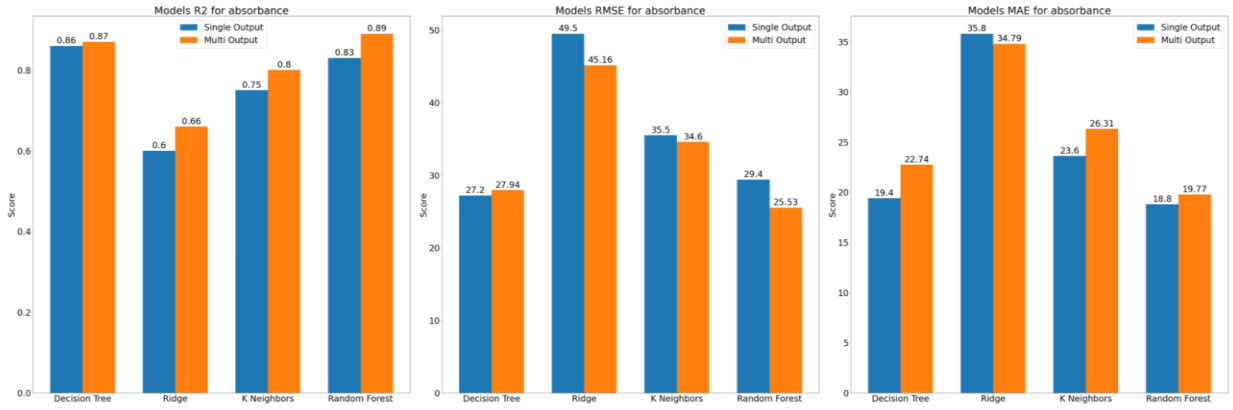

Figure 10: Performance metrics for the single output vs multioutput algorithm tested in the CdSe data set for absorbance

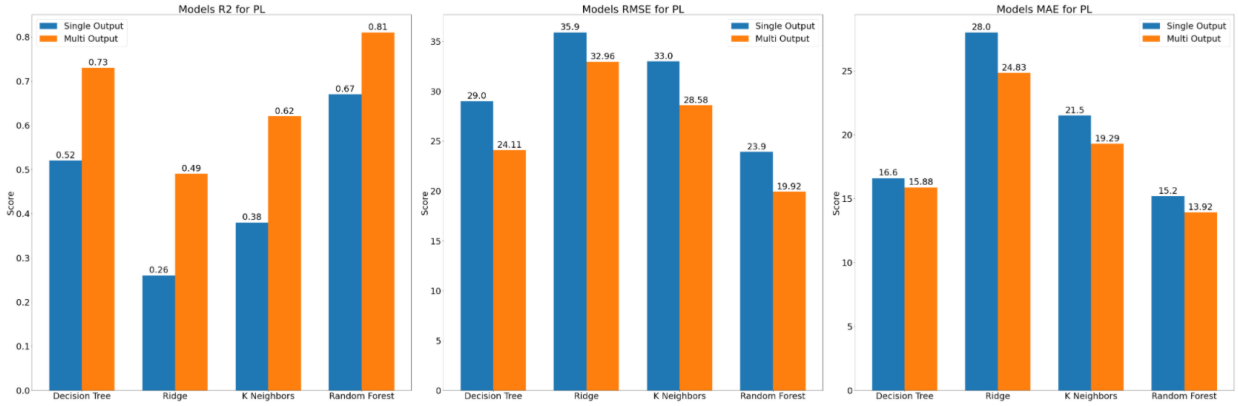

Figure 11: Performance metrics for the single output vs multioutput algorithm tested in the CdSe data set for PL

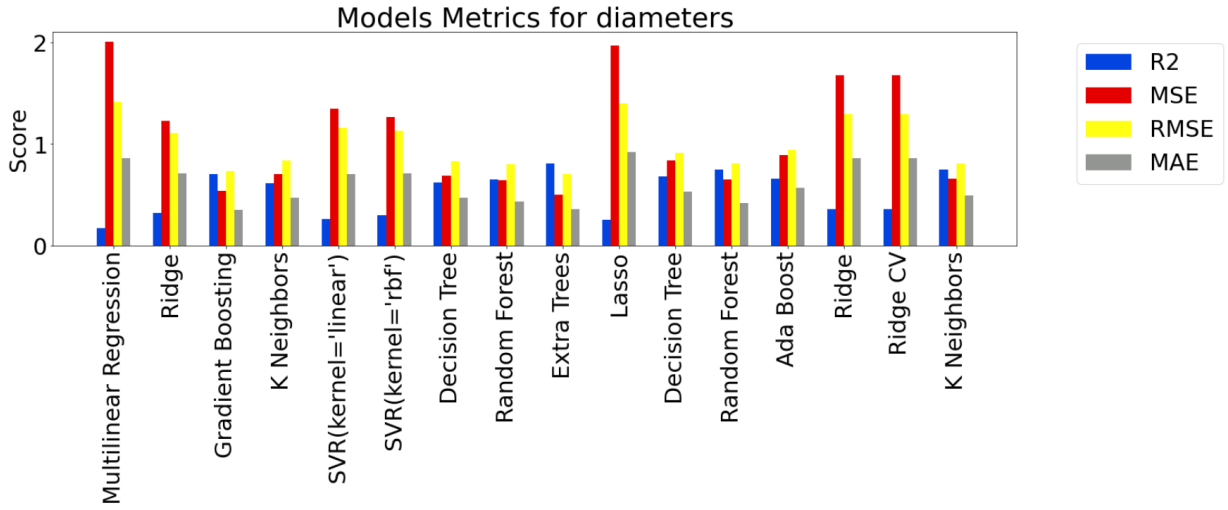

Figure 12 a: Performance metrics for 16 algorithm model

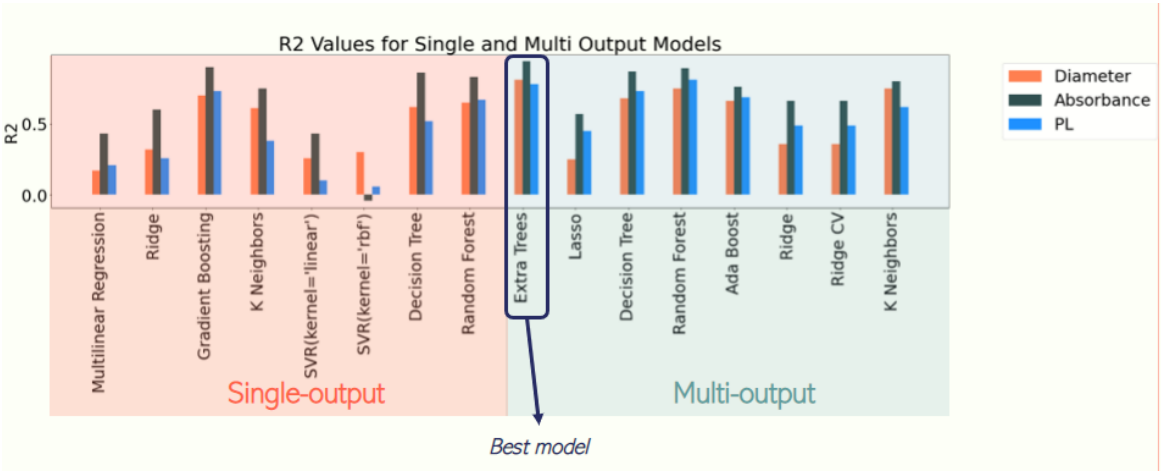

Figure 12 b: Performance metrics for 16 algorithm model

d. Diameter distribution

Like the paper, we plotted the diameter distribution, figure 13. We have successfully recovered the paper trend which consisted of an asymmetric distribution. The observation range is between 1 and 12 nm which is the maximum of twice the CdSe Bohr radius making. Our diameter distribution validate that nanoparticles found in our dataset were quantum dot. We also plotted the injection temperature and growth temperature of CdSe only. We couldn't conclude to a conclusion when comparing our plot figure 15 to the paper plot figure 16.

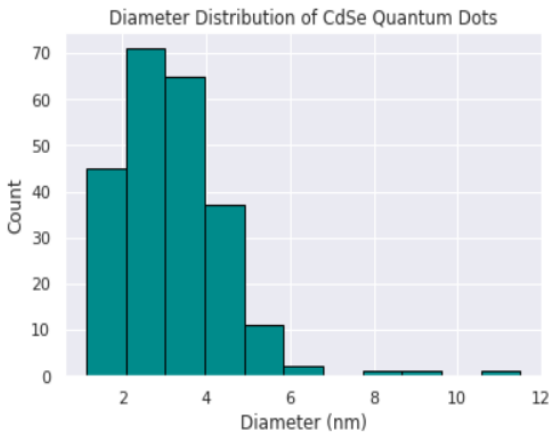

Figure 13: Diameter distribution of CdSe quantum dots from augmented dataset

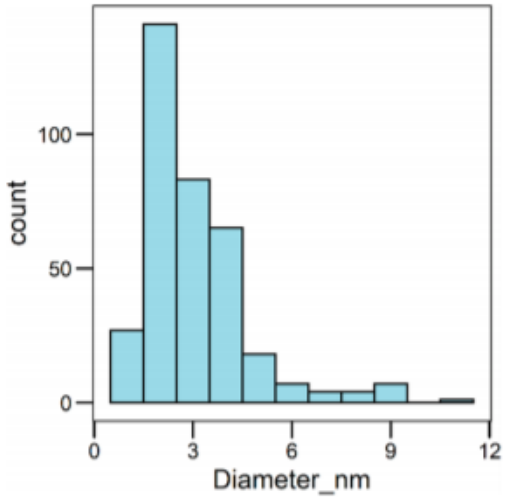

Figure 14: Diameter distribution of CdSe quantum dots [1]

e. Injection temperature distribution

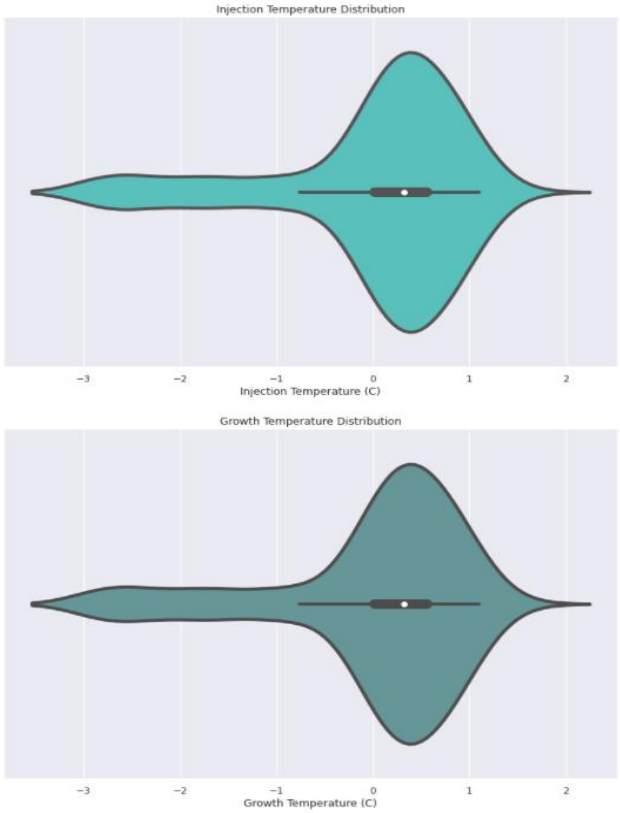

Figure 15: plot of injection temperature and growth temperature as a function of CdSe from augmented dataset

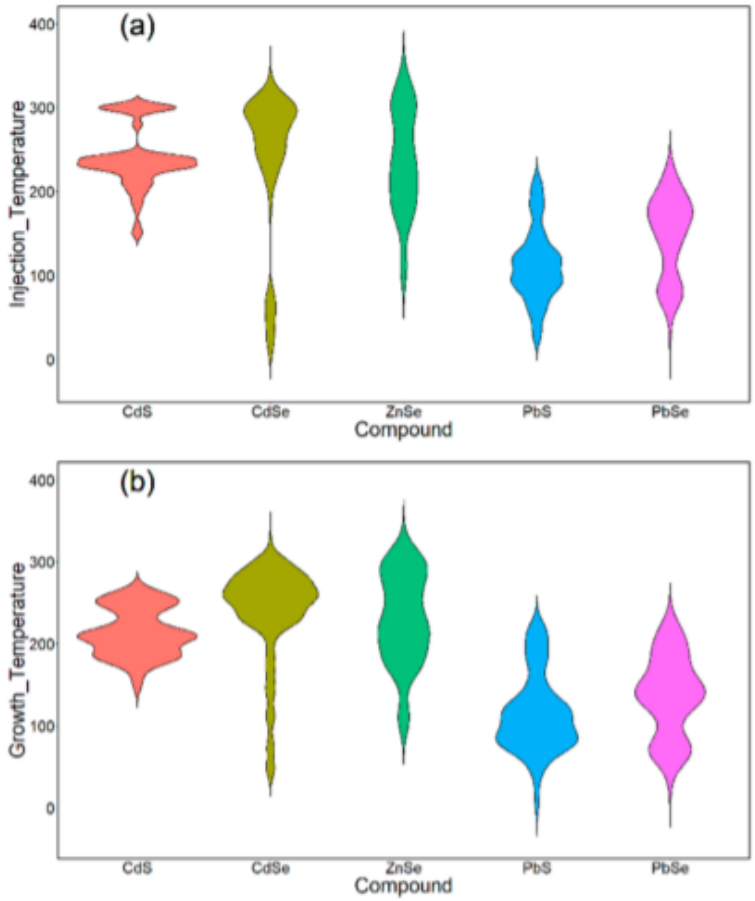

Figure 16: Plot of injection temperature (a) and growth temperature (b) as a function of MC (M = Cd, Zn, or Pb; C = S, Se) from reported syntheses

Reference:

[1] Baum, F., 2018. *Machine Learning Tools to Predict Hot Injection Syntheses Outcomes for II−VI and IV−VI Quantum Dots*. [online] Available at: <https://pubs.acs.org/doi/pdf/10.1021/acs.jpcc.0c05993> [Accessed 17 March 2021].