Final Project Report

WPI Data Science

Graduate Qualifying Project Spring 2021

Group: Aramco Americas (AA)

Mentors:

Dr. Sibo Lin - Sibo.lin@aramcoamericas.com

Dr. Yagnaseni Ghosh - Yagnaseni.ghosh@aramcoamericas.com

Address: 400 Technology Sq., Suite 300

Cambridge, MA 02139

Phone Number: 857-998-6902

Website: http://americas.aramco.com/

Group Members

Jannik Haas
jbhaas@wpi.edu
408-887-5208
WPI Data Science

Pascal Bakker
psbakker@wpi.edu
603-359-8328
WPI Data Science

Qiaochu Song
qsong2@wpi.edu
508-373-6255
WPI Data Science

Xinyuan Yang
xyang7@wpi.edu
512-661-4505
WPI Data Science

Table of Contents

1. Introduction

1.1 Motivation and Background

Quantitative structure-property relationships (QSPRs) mathematically link physical or chemical properties with the structure of a molecule. Similarly, quantitative structure–activity relationships (QSARs) link activities with the molecular structure. To model a QSAR/QSPR, a chemical property of a molecule is modeled as the response variable for regression or classification models. Catalyst QSAR/QSPR studies employ models for identifying effective catalysts and designing new catalysts in chemical sciences and engineering.

Quantitative structure properties are used as descriptors for a molecule and as the input features for models. These properties (descriptors) can be obtained from computational chemical software. For instance, Mordred is a developed descriptor-calculation software application that can calculate more than 1800 two- and three-dimensional descriptors. SambVca is a web tool for analyzing catalytic pockets with topographic steric maps. For a given molecule the number of descriptors can be very large and hard to analyze. Machine learning is now a powerful tool in determining the relationship between a molecules' descriptors and its catalytic activity in the research area of cheminformatics.

1.2 Research Challenges/Gaps

There has been a lot of research done on the different features that influence the different chemical and physical properties and activities of different chemical molecules. Some research papers have been able to use density functional theorem (DFT) derived features to produce predictive models that are able to map the relationship between the molecule and its different activity preferences. One major setback in this field is the cost required to produce the data to analyze these relationships. Since all the data is experimental data it takes a lot of time and money to produce this data and create replicable experiments. Using more sophisticated data science methods we hope to use the data that we are provided to produce high performing models and gain insights into the different features that hold predictive power when looking at the different chemical targets.

1.3 Problem Statement

In the petrochemical industry, tetramerization and trimerization of ethylene are valuable transformations. Current catalysts for this transformation are plagued by formation of polyethylene byproduct, which fouls the reactor and requires troublesome cleaning. In this project, we propose to model QSAR/QSPR of catalysts and their performance in tetramerization experiments using machine learning techniques. We hope to gain insight on the features that have the highest correlation with increased 1-octene and 1-hexene production and decreased polyethylene production. The final goal is to use our findings to help our mentors discover or create a new catalyst that will have the desired activities outlined above.

1.4 Project objective and summary of contributions

Over the course of this project we hope to experiment with several different data cleaning and manipulation methods as well as statistical and machine learning methods to extract important features from the dataset that correlate with the different targets. We hope to gain greater insight into the data science pipeline that is required for a complex data science problem in a real world application and to learn more about the petrochemical industry and different chemical properties and how they affect the activity and outcome of chemical reactions.

2. Methods

As our project goal requires models chosen to have high interpretability, we plan to employ regression models, with (S-hexene + S-Octene) * activity / S-HDPE as well as its transformations, e.g. logarithmic, as our ultimate predictor. Meanwhile, separated terms, including (1) Activity*S-hexene, (2) Activity * S-octene, (3) Activity * S-HDPE are also chosen as predictors. The ultimate goal is to increase S-hexene and S-octene activity and decrease S-HDPE activity which is why the combined result is our ultimate target, however we will also separate the targets to see if we can achieve better results.

Besides interpretability, because of small dataset and high feature dimension, feature selection is also an important part in our project.

Our model choices include:

1. Single variable linear regression
2. Multiple linear regression

3. Lasso regression
4. Stepwise linear regression
5. Decision Tree
6. Random Forest regression
7. CatBoost
8. XGBoost
9. Ensemble model
10. ANN

These models are relatively highly interpretable, and can be used for feature selection which will be crucial due to the high feature space and collinear nature of the data.

Besides these methods, we also use the Boruta package in R to perform feature selection. Besides applying the above-mentioned models on thousands of features, we also plan to use principal component analysis (PCA) to analyse the feature composition of the principal components, as well as the relationship between these principal components and the target.

Feature transformations were performed to model more complex nonlinear relationships. For each feature x these included:

1. x
2. x^2 (Scaling and translation was added to avoid numerical overflow)
3. log(x-min (x) + 1)
4. Box-Cox(x - min (x) + 1)

The box-cox feature transformation creates normal features from non-normal distributions. Since normality is an assumption of many machine learning and statistical learning models this transformation was picked.

To handle missing values, we tried both simple and complex imputation by applying the following rules:

1. Drop missing values > 50%
2. Input missing values with following rulesets:
    a. Mean/mode imputation

i. with mode if num of unique <= 5

ii. with mean if num of unique > 5

b. KNN imputation

c. Iterative imputation (MICE)

Then the model performances after imputation are compared with those without imputation.

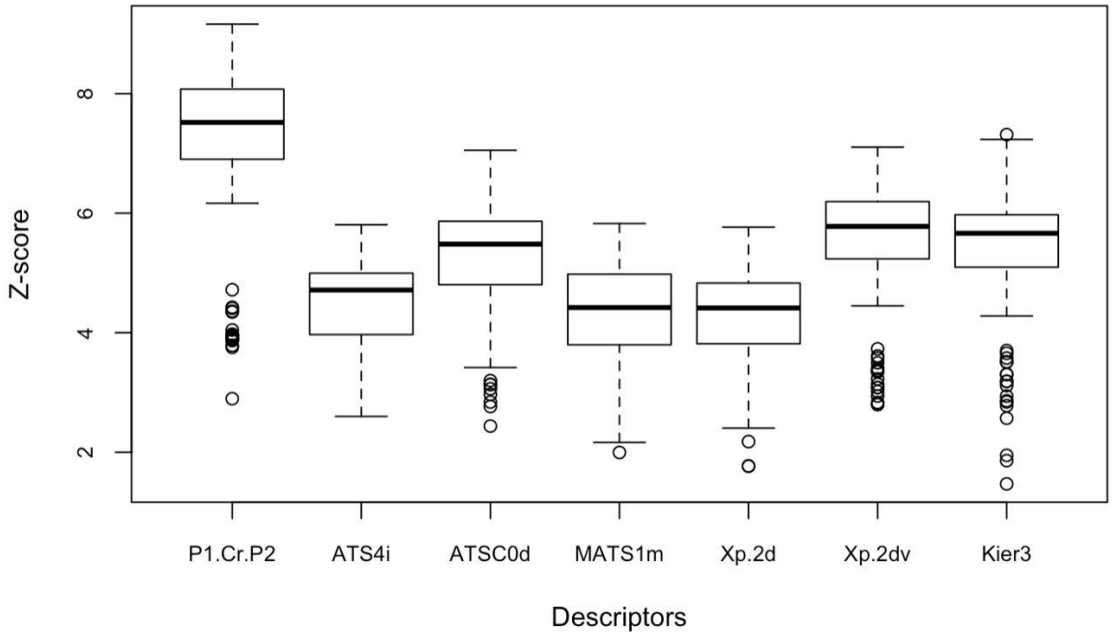
3. Results

3.1 Boruta for feature selection:

Boruta package is an R package based on Random Forest for feature selection. Boruta has been utilized in other relevant QSPR research. Here we have applied this method on the Dow dataset for feature selection, with log_target as the response variable. We run with nTree = 1000 and the model gave seven confirmed important descriptors, along with 35 tentative descriptors, after 99 iterations. The suggested formula would be as follows if we fit a MLR model with the result.

log_target ~ P1.Cr.P2 + ATS4i + ATSC0d + MATS1m + Xp.2d + Xp.2dv + Kier3

Figure: Boruta descriptor selection results. Z-score quantifies the frequency of a descriptor considered in the Random Forest with shadow variables.



We computed the VIF score to avoid collinearity. As shown in the table below, there are two pairs of collinear terms with high VIF score. The names of the paired descriptors indicate that each pair might come from the same characteristics of a molecule.

Table: VIF score for the seven selected descriptors

P1.Cr.P2	ATS4i	ATSC0d	MATS1m	Xp.2d	Xp.2dv	Kier3
1.736151	101.353707	42.819372	3.219064	61.941558	85.297263	3.927692

We fit a MLR model based on the Boruta result and the VIF score (select "ATS4i", "Xp-2dv" from the two pairs respectively). The result is shown in the table below. Two of the five descriptors, i.e. "P1-Cr-P2" and "Kier3" are considered to be of good significance. The adjusted R squared value of this model is as low as 0.4163. We also computed the LOOCV MSE = 12.061, compared to the mean guess baseline MSE = 15.464.

Table: linear regression on log-transformed target, collinear terms eliminated.

```
lm(formula = log_target ~ P1.Cr.P2 + ATS4i + MATS1m + Xp.2dv +
    Kier3, data = dowdata)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8052 -1.6151 -0.3082  1.7370  6.5049

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.9596     0.4166  23.904  < 2e-16 ***
P1.Cr.P2     -1.5866     0.5080  -3.123  0.00306 **
ATS4i         2.4489     3.6914   0.663  0.51032
MATS1m       -1.1694     0.7491  -1.561  0.12522
Xp.2dv       -3.1270     3.7668  -0.830  0.41065
Kier3        -1.6717     0.7574  -2.207  0.03223 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.033 on 47 degrees of freedom
Multiple R-squared:  0.4724,    Adjusted R-squared:  0.4163
F-statistic: 8.417 on 5 and 47 DF,  p-value: 9.581e-06
```

We also run the Boruta method on the Sasol datasets.The model has rejected all of the descriptors when the nTree is set relatively large, i.e. nTree = 1000. If nTree decreased by half, only one ATS family descriptor confirmed as useful.

3.2 Modeling results with data imputation:

3.2.1 Mean and mode imputation

For this method simple mean and mode imputation was performed on all three datasets after any column with more than 50% missing values was removed. Columns with less than or equal to 5 unique values were imputed with the mode since these were most likely categorical values and the rest of the columns were imputed with the mean of the column. The feature transformations outlined in section 2.

Dow Data Results:

With the original dataset:

	RMSE	MAPE	adj r2	feature selected	feature count
lr	3.299433	0.318615	-0.728222	all	848
dt	4.100517	0.372114	-1.300334	[AATS2v_boxcox, ATSC8c_boxcox, ATSC3v_boxcox, ...	7
lasso	3.024880	0.295431	-0.395624	[AATS4p_boxcox, ATSC7dv_boxcox, ATSC8se_boxcox...	7
stepwise	2.262977	0.209096	0.356863	[nBridgehead, ATSC1c_boxcox, ATSC3dv_log, ATSC...	12
rf	3.679538	0.375757	-1.103592	all	848
catboost	3.401158	0.344110	-0.754666	all	848

With simple mean and mode imputation:

	RMSE	MAPE	adj r2	feature selected	feature count
lr	3.194509	0.311171	-0.307042	all	2291
dt	4.725128	0.454509	-2.193262	[AATSC7c_boxcox, MATS2pe_log, GATS4d_boxcox, A...	7
lasso	2.902498	0.289651	-0.190533	[ATSC8se_boxcox, AATSC8c_boxcox, GATS4i_boxcox...	7
stepwise	2.099891	0.198942	0.453183	[nBridgehead_boxcox, AATSC4Z_boxcox, AATSC3pe_...	12
rf	3.377657	0.341012	-0.669909	all	2291
catboost	3.385988	0.333129	-0.664441	all	2291

With simple mean and mode imputation and square feature transformation:

	RMSE	MAPE	adj r2	feature selected	feature count
lr	4.978078	0.509869	-3.430079	all	2291
dt	4.445099	0.437169	-1.818505	[AATS3se_boxcox, AATSC5Z_boxcox, AXp-3d_boxcox...	7
lasso	3.047012	0.302986	-0.405461	[AATS4p_boxcox, ATSC8se_squared, GATS6Z_square...	5
stepwise	1.630005	0.143777	0.666323	[nBase_boxcox, AATS4p_boxcox, ATSC3i, AATSC6c_...	12
rf	3.394810	0.346372	-0.761125	all	2291
catboost	3.432421	0.340901	-0.745196	all	2291

Sasol Data-2 Results:

With the original dataset:

	RMSE	MAPE	adj r2	feature selected	feature count
lr	15.586823	2.463732	-100.392766	all	2223
dt	2.654000	0.667075	-0.995924	[AATS7s, ATSC7i, AATS4i_CrCl2, ATSC7se_CrCl2, ...	6
lasso	1.807345	0.508078	-0.056896	[ATSC3c_boxcox, C2SP3_boxcox, AXp-1d, MINsssN_...	7
stepwise	1.258670	0.248287	0.485727	[nBase, ATS8d, AATS6p_boxcox, ATSC2dv, ATSC4d,...	12
rf	2.794573	0.702442	-1.214997	all	2223
catboost	2.648025	0.676094	-0.972244	all	2223

With simple mean and mode imputation:

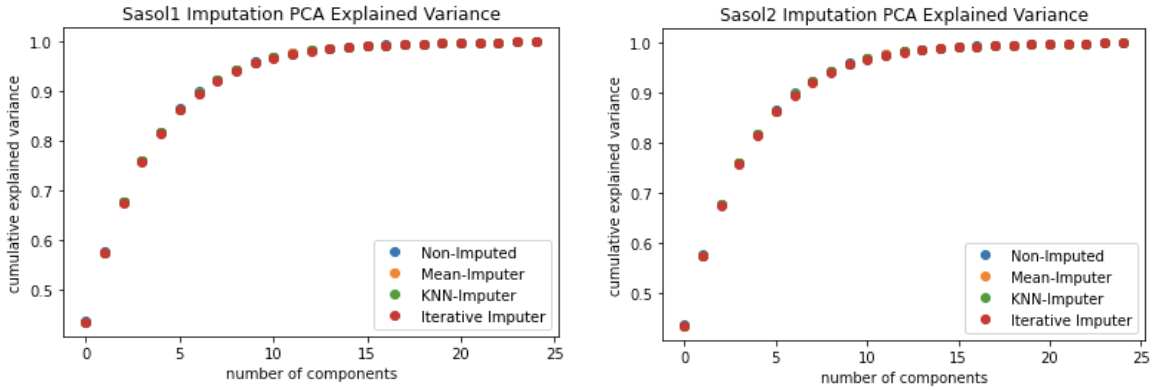
	RMSE	MAPE	adj r2	feature selected	feature count
lr	5.840562	1.635621	-14.778230	all	2237
dt	2.909837	0.648083	-1.419292	[ABCGG_log, ATSC7v, GATS1d_boxcox, AATS4i_CrCl...	6
lasso	1.807345	0.508078	-0.056896	[ATSC3c_boxcox, C2SP3_boxcox, AXp-1d, MINsssN_...	7
stepwise	0.966746	0.196430	0.575980	[nBase, MATS7dv_boxcox, GATS6c_boxcox, GATS3d_...	12
rf	2.878394	0.703986	-1.411143	all	2237
catboost	2.838602	0.746989	-1.197132	all	2237

With simple mean and mode imputation and square feature transformation:

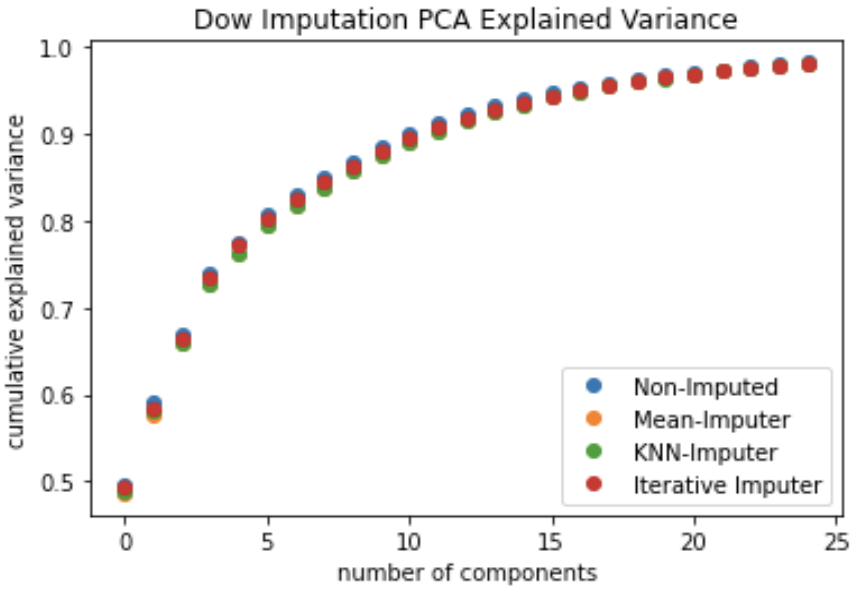
	RMSE	MAPE	adj r2	feature selected	feature count
lr	22.252332	5.661757	-166.501102	all	2237
dt	2.012394	0.475151	-0.315189	[ATSC3d_squared, ATSC7p_squared, GATS8d_square...	6
lasso	2.069020	0.497131	-0.310989	[AATS6p_squared, ATSC4i_squared, AXp-1d_square...	5
stepwise	1.414585	0.306881	0.396392	[nBase, ATSC4Z_squared, ATSC8Z_squared, AATSC2...	12
rf	2.935732	0.719795	-1.538423	all	2237
catboost	2.679755	0.679318	-1.039187	all	2237

3.2.2 Complex Imputation Methods

Complex imputation methods including KNN imputation and MICE imputation were also used to handle missing values. After dropping columns with more than 50% of missing values we found that the Dow data had the largest number of columns that required imputation with the imputed dataset having 2357 columns and the dataset in which we dropped all columns having any missing values had 914 columns. Sasol1 and Sasol2 had 2340 and 2299 columns in the imputation set respectively and 2319 and 2285 columns in the non imputed dataset. PCA was performed on all the datasets and the results are shown below:



Sasol1 and Sasol2 PCA Explained Variance Plots



Dow PCA Explained Variance Plots

In the initial attempts for this methodology, the complex imputation had a large impact in the amount of variance few PCs were able to explain in the Dow dataset. After discussion with our mentors we discovered several outlier ligands that, due to their chemical structures, had a large number of descriptors that were not applicable to them and were therefore missing values in the dataset. After removing these outliers, we see in the plot above that the cumulative variance explained did not vary greatly among the non-imputed and the imputed datasets. Looking closely at the principal components however, we saw that all the coefficients of the features were still very small and almost identical. This was a similar finding as our initial PCA exploration. We attribute this to the collinearity and small size of the data. Even though PCA is known for being able to deal with wide and collinear datasets, we did not see good results. We chose the first 12

principal components to feed into different models to see the results. There was no significant increase in performance with the best performing model producing an adjusted r^2 of 0.570 on the log of the combined target.
An ensemble model was created using linear regression, decision tree, and random forest. For the final prediction we took the average of each of the individual models. We also experimented with learning the weights for each of the models but this did not give us better results since each model produced very similar final results individually.

3.3 PCA

The PCA and PCR is performed on Sasol 1 (containing 29 molecules) on all molecular descriptors, i.e. XTB+Mordred+SambVca+Geometric. Compute PCs from 2338 dims, using all of the Sasol 1 data records. The same procedure is done for Sasol 2.
Given the number of variables/descriptors (>2000), it is impossible to give visually-friendly loading plots to show the influences of each variable/descriptor on each PC. As a result, we instead checked some top descriptors' names, by taking the absolute values of the coefficients, and then sort in descending order. We then examined the descriptors along with their absolute coefficients. We found that the collinearity problem still exists in PCs, take the top 30 descriptors of PC2 for example. All coefficients are small values. Note that there are pairs of similar descriptor names with exactly the same value in their coefficients, which are possibly collinear descriptors in the same PC.

We plotted PC1 vs. PC2, PC1 vs. PC3, in the same graph respectively, with a colormap on the data points to indicate the value of the log_target. There is no obvious pattern in the two plots. Yet we found that on PC2, most of the data points have values around zero, and the most of the variance on PC2 is introduced by three outliers:

- Sasol02_22: the catalyst we do NOT want because it has the lowest activity (20100 g/gCr-h) in Sasol 1 (>= 1 digit less) and the lowest S_alpha value(52.9). As a result, its alpha * activity is the lowest. Meanwhile it has high S_PE = 12.9 (3rd high).
- Sasol11_15:activity = 932800, low S_alpha = 59.6, high S_PE=16.2 (2nd high)
- Sasol11_16: activity = 385900, S_alpha = 65.2, moderate S_PE=4.6

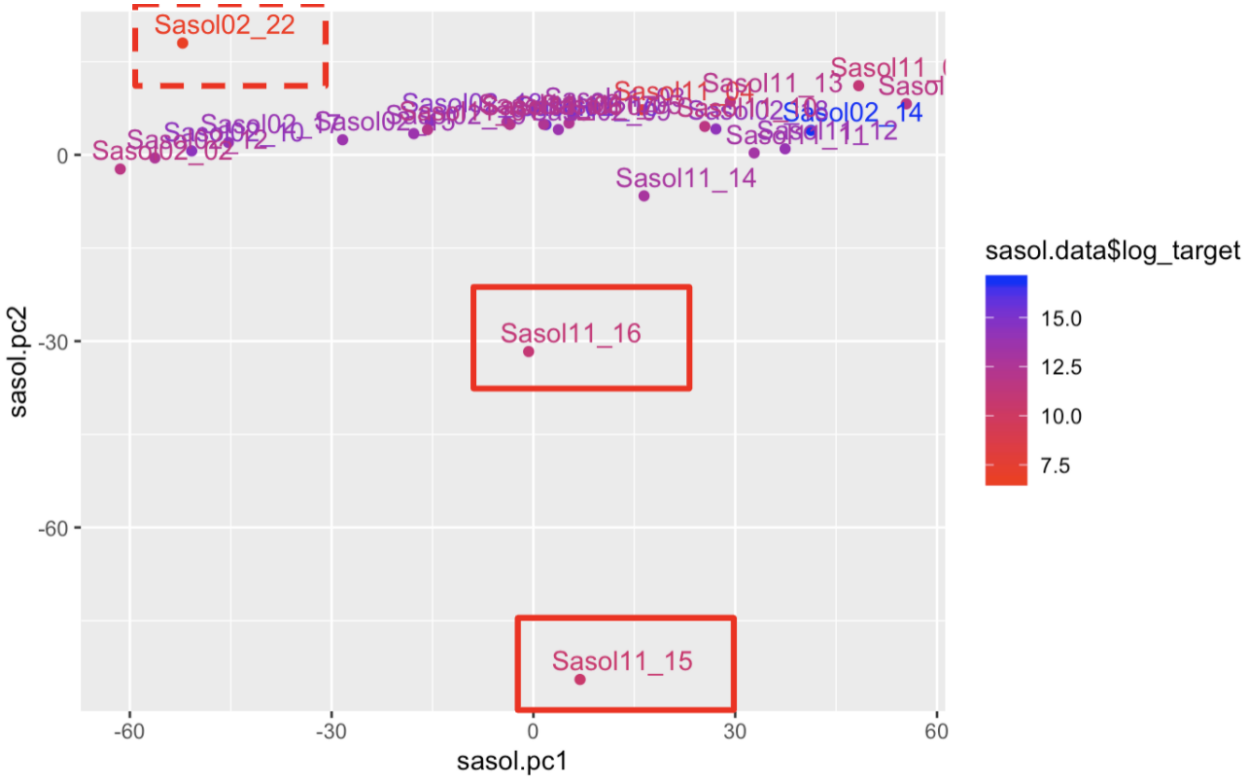


Figure: PC1 vs. PC2, Sasol 1

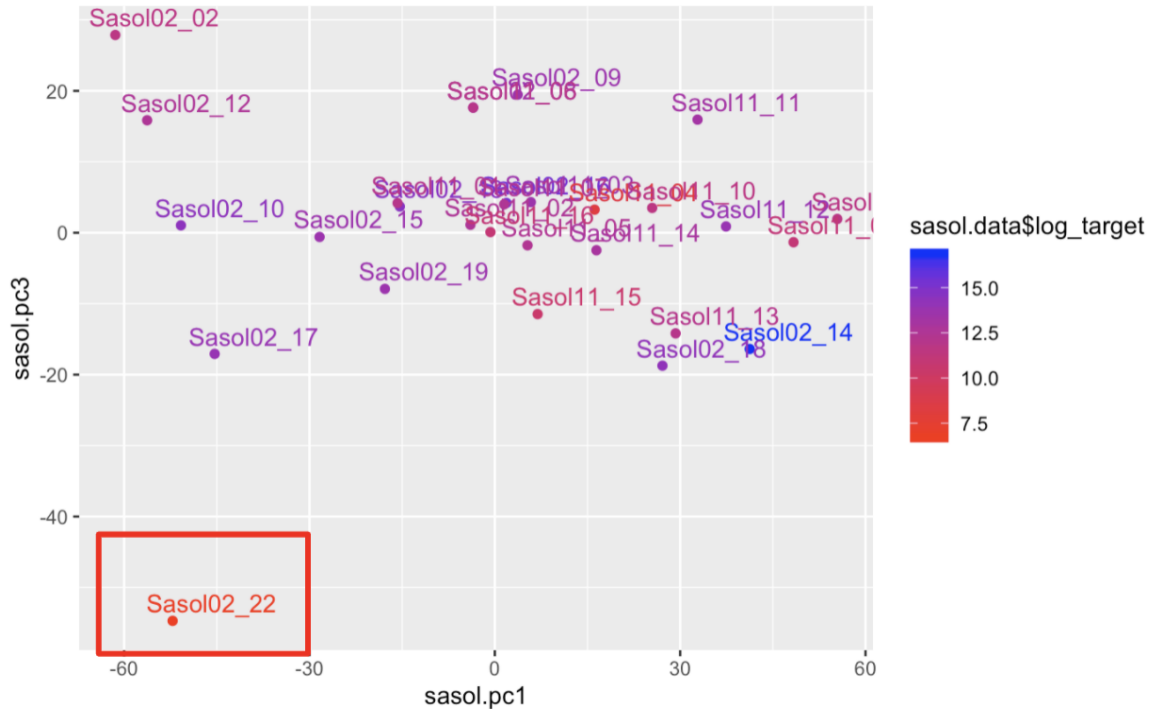


Figure: PC1 vs. PC3, Sasol 1

We then removed these three points and then redid the PCA. Here are the new PCs and data

scatterplots. The data points became better scattered, mainly because the scale range of PC2 changed. Here we also noticed there are some close/overlapping pairs of ligands in PC1 to PC3.

- Overlapping pair: Sasol02_08 & Sasol11_06
- Very close pairs:
  Sasol02_13 & Sasol11_01
  Sasol02_16 & Sasol11_07

We examined the overlapping pair of ligands which have nearly the same values for each PC. As explained by our mentor, this pair of ligands are very similar in their chemical structures.

Note that P = 45 for Sasol02_08 and P = 50 for Sasol11_06.

Table: PC values (1 to 3) and experimental property values of Sasol02_08 & Sasol11_06

Ligand	PC1	PC2	PC3	activity	S_alpha	S_alpha * activity	PE
Sasol02_08	-6.4285486	6.936189	14.3368780	1070000	70.5	754350	1.3
Sasol11_06	-6.4284259	6.936884	14.3365882	1065300	70.4	750340	4.8

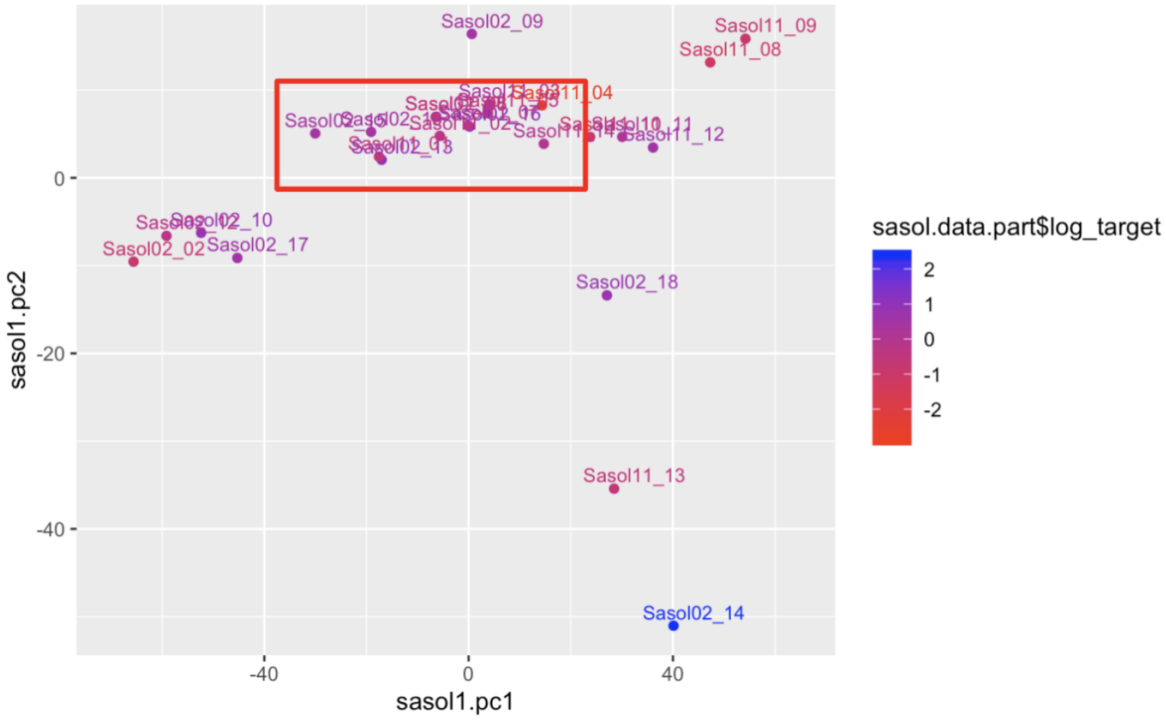


Figure: PC1 vs. PC2, Sasol 1, three outlier ligands removed

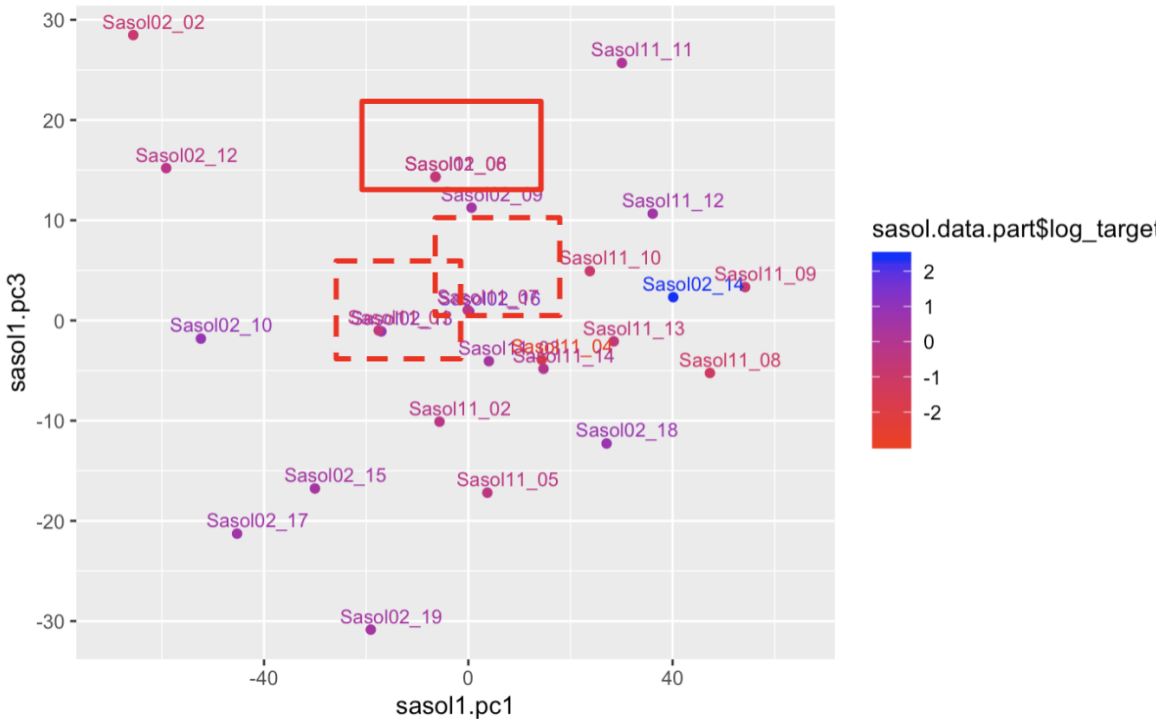


Figure: PC1 vs. PC3, Sasol 1, three outlier ligands removed

The top PCs can explain more variance after removing the three outlier ligands (from 0.4 to more than 0.5 for the PC1). As the number of PCs increases, the cumulative % variance explained keeps increasing, by a decreasing positive value. In case of overfitting, it is likely that no more than 10 PCs can be a good representation of all descriptors.

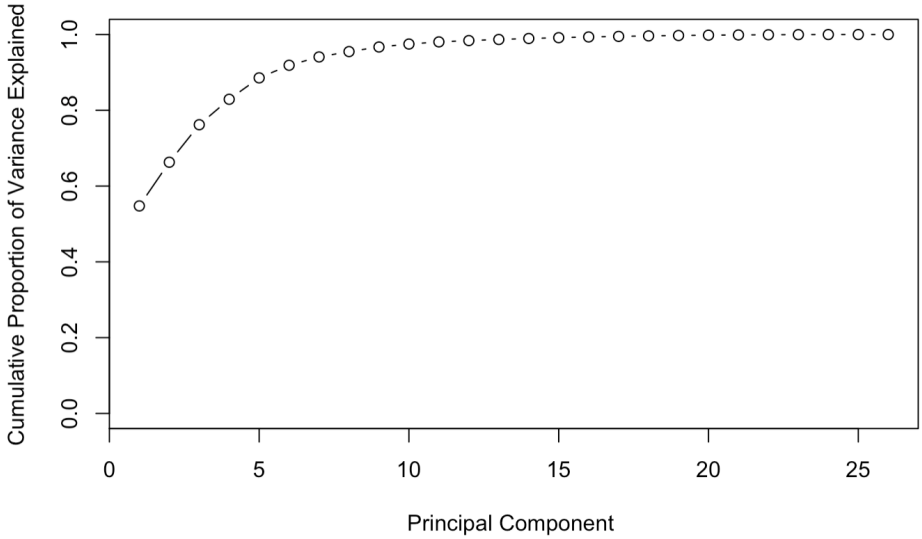


Figure: cumulative % variance explained (with the increasing of number of PCs), Sasol 1

The PCR performance is not good. Among the top 3 PCs only the PC2 is of statistical significance. The adjuster R-squared is only 0.1334 when using PC2.

Table: PCR using PC2, Sasol 1

```
lm(formula = sasol.data$log_target[-c(13, 28, 29)] ~ sasol.pr.out.partial$x[,
    2])

Residuals:
    Min      1Q  Median      3Q     Max
-4.3313 -0.7051  0.2966  0.9660  2.6302

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 12.97935    0.29681  43.730   <2e-16 ***
sasol.pr.out.partial$x[, 2] -0.04492    0.02040  -2.202   0.0375 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.513 on 24 degrees of freedom
Multiple R-squared:  0.1681,    Adjusted R-squared:  0.1334
F-statistic: 4.848 on 1 and 24 DF,  p-value: 0.03753
```

For Sasol 2, the data points scattered more evenly than Sasol 1, and no obvious outliers.

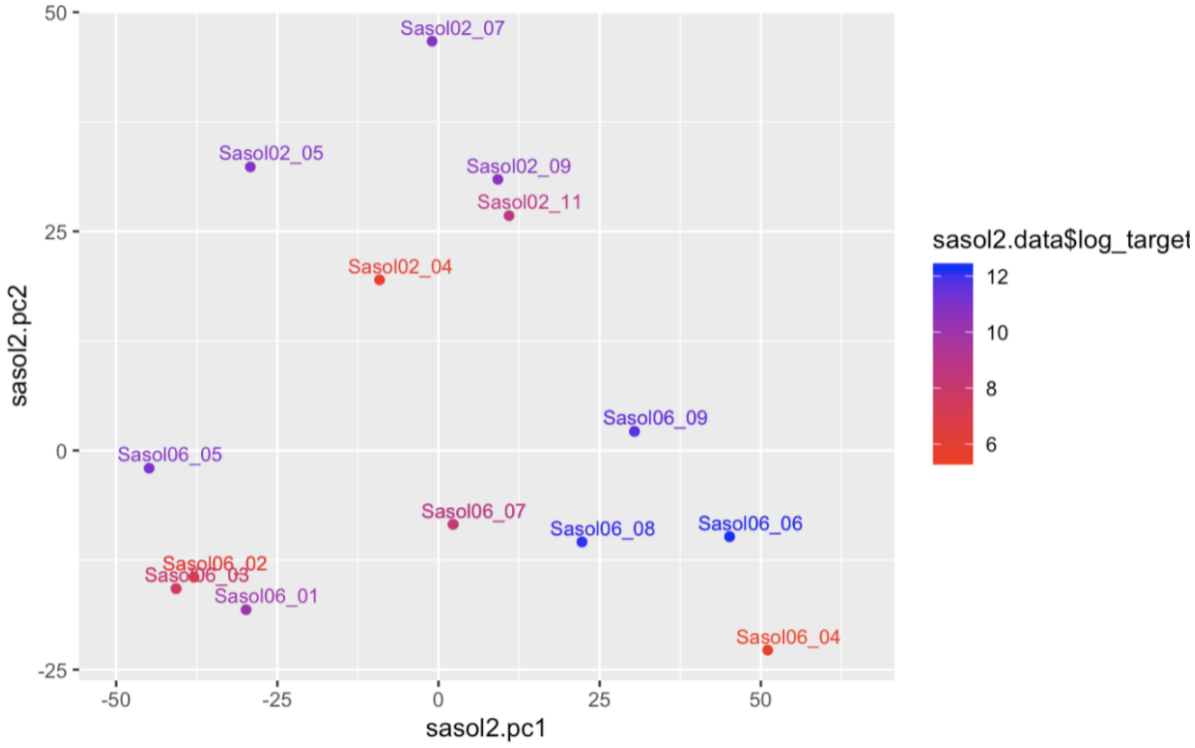


Figure: PC1 vs. PC2, Sasol 2

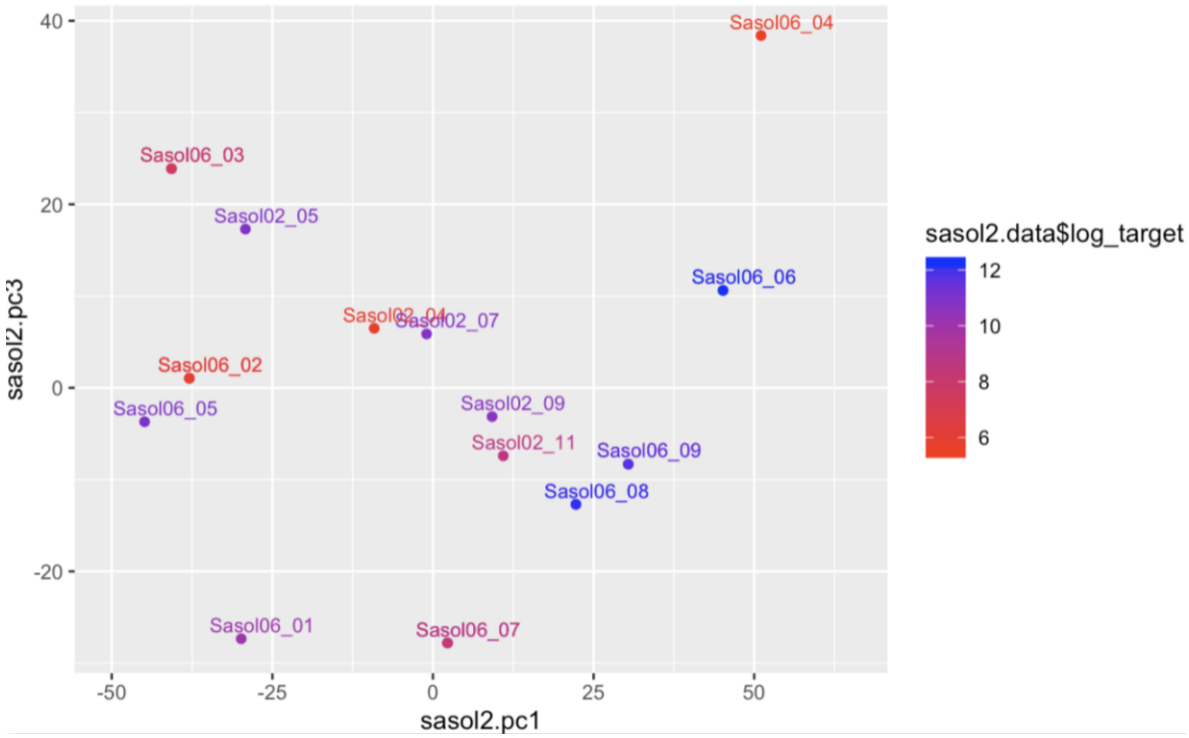


Figure: PC1 vs. PC3, Sasol 2

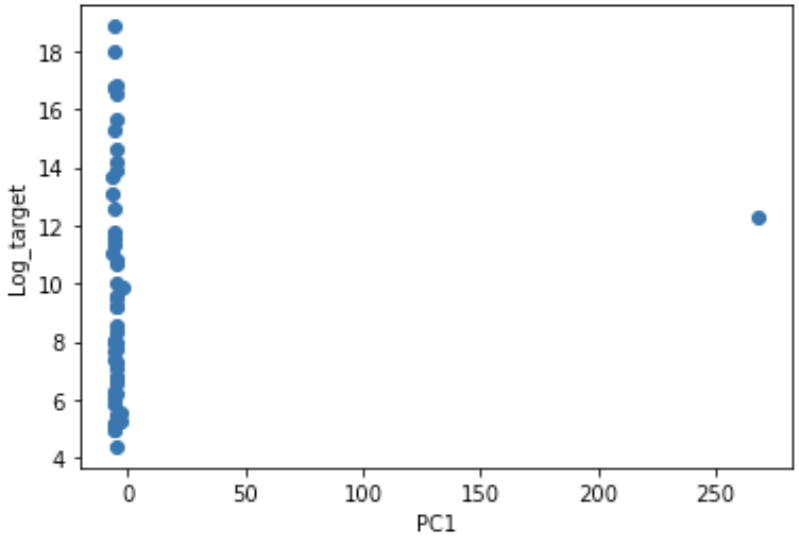
PCR is not performing well on Sasol 2, either, even worse than Sasol 1, with no significant PC.

Table: PCR using PC1&2, Sasol 2

```
Call:
lm(formula = sasol2.data$log_target[-11] ~ sasol2.pr.out.partial$x[,
    1] + sasol2.pr.out.partial$x[, 2])

Residuals:
    Min      1Q  Median      3Q     Max
-3.8477 -1.6036  0.6309  1.5101  2.5921

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.668551   0.455064  21.247 3.76e-13 ***
sasol2.pr.out.partial$x[, 1] 0.028181   0.014404   1.957   0.0681 .
sasol2.pr.out.partial$x[, 2] 0.003969   0.022190   0.179   0.8603
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.984 on 16 degrees of freedom
Multiple R-squared:  0.1944,    Adjusted R-squared:  0.09365
F-statistic:  1.93 on 2 and 16 DF,  p-value: 0.1775
```

We also gave an attempt of PCA on the Dow dataset. One outlier ligand (Dow2_17) was found on PC1.

Figure: PC1 in Dow



3.4 Influence of experimental variable

As mentioned in the proposal, the Dow, Sasol 1 and Sasol 2 dataset are different not only in ligands, but also the experimental conditions. The Dow dataset has not included variables of experimental conditions. For the two Sasol dataset, the average react pressure (P (bar)), temperature (T) and time length (time) are included. In Sasol 1, the temperature T is fixed at 60, while time and pressure are not fixed. In Sasol 2, the time variable is 30 for all records; temperature T = 65 for all P=30 and T = 45 for all P = 45. Besides the difference in temperature and time, the solvents used are also different in Dow and Sasol.

We also found that the ligand “Sasol02_09” appeared in both Sasol 1 and 2, with differences in experimental conditions: time, temperature, pressure, and solvents. This ligand is often used as a reference in such research. By comparing these two records, we can see that the same ligand can lead to great difference in the target value, which is mainly because of the great difference in the value of activity.

Sasol 1:

molecule	P (bar)	T	Solvent	Time	Cr complex	Cr (umol)	S_alpha (wt%)	PE (wt%)	S_alpha*Activity	1-octene activity	PE activity	log_target
Sasol02_09	45	60	MCH	18.0	Cr(acac)3	2.5	73.8	1.5	1306260.0	1129226.37	26550.0	13.677214

Sasol 2:

molecule	P (bar)	T	Solvent	Time	Cr complex	Cr (umol)	S_alpha (wt%)	PE (wt%)	S_alpha*Activity	1-octene activity	PE activity	log_target
Sasol02_09	30	65	Toluene	30	Cr(acac)3	3.3	74.1	0.8	36457.2	27627.768	393.6	10.727038

3.5 Neural Networks

In an attempt to ignore interpretability, and to build a high performing model, neural networks with a variety of hyperparameters were tried. Between one and five layered networks were attempted with the dataset, ignoring features with missing values. However, despite tinkering, the networks would not converge on a small loss value. This is logical, as the amount of data is too small for a deep network to have an impact.

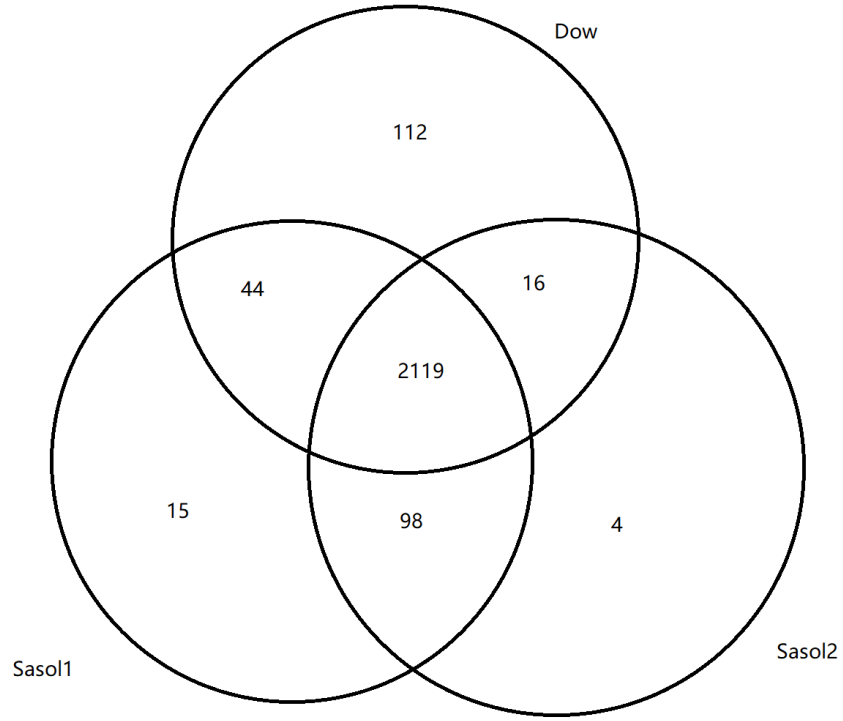
```
Epoch 998/1000
3/3 [==============================] - 0s 13ms/step - loss: 64616247853056.0000 - accuracy: 0.0000e+00 - val_los
s: 3760971864604672.0000 - val_accuracy: 0.0000e+00
Epoch 999/1000
3/3 [==============================] - 0s 12ms/step - loss: 70793280356352.0000 - accuracy: 0.0000e+00 - val_los
s: 3796814171996160.0000 - val_accuracy: 0.0000e+00
Epoch 1000/1000
3/3 [==============================] - 0s 16ms/step - loss: 54834880315392.0000 - accuracy: 0.0000e+00 - val_los
s: 3811072389677056.0000 - val_accuracy: 0.0000e+00
```

3.6 Feature Clustering

The PCA results motivated us to find cluster patterns in the descriptors. We did agglomerative clustering on descriptors. Our main goal in this part is to find similar descriptors and thus decrease collinearity.

We use correlation coefficients as a similarity metric, and set the threshold as 0.75. Here in the Venne diagram, we can see the overlaps among the features of the three datasets.

Figure: Venne diagram of number of features and overlaps in the three datasets



In the table below we listed the number of features in the top three largest clusters in each dataset. The three dataset have many descriptors in common. Dow is slightly more different from the other two datasets. There are about 300 clusters for each dataset. On average each cluster has about 20 features, but there are some large clusters containing about 300 descriptors.

Table: number of features top three largest clusters for each dataset

Dataset	Number of features in 3 largest clusters
Dow	240, 161, 148
Sasol 1	296, 113, 84
Sasol 2	154, 126, 118

We then fit regression models on each dataset, compare the results of using all features versus the results of using representative features. The best model is stepwise linear regression. Although in

some cases using representative features did help improve the performance.but for stepwise regression, it made the performance worse.

Table: comparisons of model performances, using complete dataset vs. representatives of features

Dow

Complete dataset

	adj r2	feature count
lr	-0.556438	2291
dt	-1.672527	7
lasso	-0.284069	4
stepwise	0.608422	18
rf	-0.655971	2291
catboost	-0.802144	2291

Representatives of features

	adj r2	feature count
lr	-0.284156	388
dt	-1.847728	7
lasso	-0.021813	7
stepwise	0.508785	17
rf	-0.655990	388
catboost	-0.386057	388

Sasol 1

Complete dataset

	adj r2	feature count
lr	-10.052530	269
dt	-5.648989	7
lasso	-2.588226	4
stepwise	0.016569	7
rf	-1.490553	269
catboost	-1.640641	269
svm_rbf	-1.891547	269

Representatives of features

	adj r2	feature count
lr	-6.461331	2276
dt	-3.109554	5
lasso	-1.430883	1
stepwise	0.690799	11
rf	-2.097015	2276
catboost	-2.034938	2276
svm_rbf	-2.039857	2276

Sasol 2

Complete dataset

	adj r2	feature count
lr	-20.675313	2237
dt	-1.004703	6
lasso	-1.072551	3
stepwise	0.377070	5
rf	-0.935786	2237
catboost	-1.041131	2237
svm_rbf	-1.165291	2237

Representatives of features

	adj r2	feature count
lr	-64.805417	198
dt	-1.256847	6
lasso	-1.166980	1
stepwise	0.291829	3
rf	-0.750050	198
catboost	-0.658497	198
svm_rbf	-1.166757	198

3.7 Final Modeling Results

The best model is stepwise linear regression and the performances on each dataset is listed in the table below.

Table: best model cross validating and training set R^2 on each dataset

Dataset	5-fold CV R^2	Training set R^2
Dow	0.60	0.82
Sasol 1	0.69	0.94
Sasol 2	0.37	0.75

We listed the feature component of each model, along with the coefficients. The significant positive features are marked with a red star mark, while the significant negative features are marked with a blue star mark. The feature names with suffix “_log”, “_boxcox” indicates a transformed version of the original feature.

Table: features and coefficients, Dow

	coef	std err	t	P>|t|	[0.025	0.975]
ATSC1v	-0.2815	0.641	-0.439	0.663	-1.584	1.021
★ AATSC4Z_boxcox	-1.1636	0.435	-2.677	0.011	-2.047	-0.280
★ AATSC0i_boxcox	-1.3661	0.515	-2.653	0.012	-2.413	-0.320
MATS2s_boxcox	-0.8943	0.507	-1.765	0.086	-1.924	0.135
★ GATS6dv_boxcox	1.5180	0.657	2.311	0.027	0.183	2.853
★ GATS4se	-1.1603	0.514	-2.258	0.030	-2.204	-0.116
C3SP3_boxcox	-0.6615	0.545	-1.214	0.233	-1.769	0.446
EState_VSA2	0.0866	0.406	0.214	0.832	-0.738	0.911
EState_VSA7	0.3776	0.461	0.820	0.418	-0.558	1.314
★ n10FRing	0.8881	0.399	2.227	0.033	0.078	1.699
nFHRing_log	0.4459	0.561	0.795	0.432	-0.693	1.585
nG12FHRing	0.4190	0.587	0.714	0.480	-0.774	1.612
★ AATSC7dv_Cr_boxcox	2.0747	0.688	3.017	0.005	0.677	3.472
AATSC7v_Cr_boxcox	0.1841	0.513	0.359	0.722	-0.858	1.226
★ Xch-5dv_Cr_boxcox	2.1939	0.710	3.092	0.004	0.752	3.636
MDEC-23_Cr	-0.4220	0.423	-0.997	0.326	-1.282	0.438
nG12FRing_Cr	0.0998	0.739	0.135	0.893	-1.401	1.601
GGI10_Cr_log	0.1582	0.473	0.334	0.740	-0.803	1.120
const	9.9609	0.283	35.206	0.000	9.386	10.536

Table: features and coefficients, Sasol 1

	coef	std err	t	P>|t|	[0.025	0.975]
★ AATS6se_log	73.8378	13.462	5.485	0.000	45.436	102.240
★ AATS6i_log	1.5652	0.700	2.237	0.039	0.089	3.042
ATSC7dv_log	0.3912	0.198	1.973	0.065	-0.027	0.810
GATS5Z_log	-0.1771	2.595	-0.068	0.946	-5.651	5.297
GATS2v	1.4901	6.063	0.246	0.809	-11.302	14.282
★ PNSA3_boxcox	2.242e-05	2.44e-06	9.177	0.000	1.73e-05	2.76e-05
ETA_shape_y_log	18.6378	9.828	1.896	0.075	-2.097	39.372
★ GeomDiameter_log	5.6302	1.140	4.939	0.000	3.225	8.035
GGI10_log	4.4470	3.401	1.307	0.208	-2.729	11.623
★ AATSC3se_CrCl2_log	-1307.8174	542.830	-2.409	0.028	-2453.090	-162.545
Vbur_P1_boxcox	-0.1373	0.075	-1.833	0.084	-0.295	0.021
const	-7.1945	8.957	-0.803	0.433	-26.091	11.702

Table: features and coefficients, Sasol 2

	coef	std err	t	P>|t|	[0.025	0.975]
ATSC4p_log	-0.0026	0.505	-0.005	0.996	-1.087	1.081
ATSC5i_log	-0.6108	0.386	-1.581	0.136	-1.439	0.218
MATS7s_boxcox	17.4878	12.441	1.406	0.182	-9.196	44.171
GATS8c	2.1538	2.424	0.889	0.389	-3.045	7.353
★ BCUTi-1h_boxcox	-1.299e+04	2173.746	-5.976	0.000	-1.77e+04	-8328.703
const	10.2825	2.851	3.607	0.003	4.168	16.397

4. Challenges

1. Experimental variables contribute greatly to the target, but are limited in experimental settings.
2. Conclusions not consistent across the dataset: useful descriptors in Dow and Sasol are too different.
3. Collinearity is still a problem. Single variable models may help.
4. Still no great results in modeling the overall targets, partly because the influence of experimental variables is strong, yet these variables are excluded when modeling.

5. Conclusions

- Catalysts across datasets vary greatly. Performances of models also vary.
- For Dow & Sasol 1 - our model can explain >60% variance in test data.
- The family of ATS (Autocorrelation of Topological Structure) descriptors is useful across all three datasets. This includes ATS…, ATSC…, MATS…, AATS... , AATSC…, GATS….

References

1. Moriwaki, H., Tian, YS., Kawashita, N. *et al.* Mordred: a molecular descriptor calculator. *J Cheminform* 2018 10(4). https://doi.org/10.1186/s13321-018-0258-y, https://github.com/mordred-descriptor.
2. Falivene, L.; Credendino, R.; Poater, A.; Petta, A.; Serra, L.; Oliva, R.; Scarano, V.; Cavallo, L. SambVca 2. A Web Tool for Analyzing Catalytic Pockets with Topographic Steric Maps. Organometallics 2016, 35, 2286−2293.
3. Tang S, Liu Z, Zhan X, Cheng R, He X, Liu B. 2D-QSPR/DFT studies of aryl-substituted PNP-Cr-based catalyst systems for highly selective ethylene oligomerization. J Mol Model. 2014 Mar;20(3):2129. doi: 10.1007/s00894-014-2129-4. Epub 2014 Feb 20. PMID: 24554126.
4. Maley S, Kwon D-H, Rollins N. Quantum-Mechanical Transition-State Model Combined with Machine Learning Provides Catalyst Design Features for Selective Cr Olefin Oligomerization. 2020, 11, 9665. doi: 10.1039/d0sc03552a.
5. Sanchez-Lengeling, Benjamin; Outeiral, Carlos; Guimaraes, Gabriel L.; Aspuru-Guzik, Alan (2017): Optimizing distributions over molecular space. An Objective-Reinforced Generative Adversarial Network for Inverse-design Chemistry (ORGANIC). ChemRxiv. Preprint. (https://github.com/aspuru-guzik-group/ORGANIC)
6. Esaki T., Watanabe R., Kawashima H., Ohashi R., Natsume-Kitatani Y., Nagao C., Mizuguchi K., Data Curation can Improve the Prediction Accuracy of Metabolic Intrinsic Clearance, Mol. Inf. 2019, 38, 1800086.

7. Kursa, M., & Rudnicki, W. 2010 Sep 16. Feature Selection with the Boruta Package. Journal of Statistical Software. [Online] 36:11.
8. S. Kurth, M.A.L. Marques, E.K.U. Gross, Density-Functional Theory, Editor(s): Franco Bassani, Gerald L. Liedl, Peter Wyder, Encyclopedia of Condensed Matter Physics, Elsevier, 2005, Pages 395-402, ISBN 9780123694010, (https://www.sciencedirect.com/science/article/pii/B0123694019004459)
9. Stef van Buuren, Flexible Imputation of Missing Data, second edition, Chapman & Hall/CRC, Boca Raton, 2012.
10. Dr. Cox, An Analysis of Transformations, 1964. https://www.ime.usp.br/~abe/lista/pdfQWaCMboK68.pdf
11. Liudmila Prokhorenkova, Gleb Fusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin, CatBoost: unbiased boosting with categorical features, 2018. https://papers.nips.cc/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf