XGBoost, mordred and RDKit for the prediction of glass transition temperature of polymers

Goh, Kai Leong

2021

Goh, K. L. (2021). XGBoost, mordred and RDKit for the prediction of glass transition temperature of polymers. Student Research Paper, Nanyang Technological University, Singapore. https://hdl.handle.net/10356/155298

https://hdl.handle.net/10356/155298

XGBoost, *Mordred* and *RDKit* for the Prediction of Glass Transition Temperature of Polymers

Goh Kai Leong
School of Physical and Mathematical Sciences

Dr Lu Yunpeng
Asst Prof Xia Kelin
Wee Jun Jie
School of Physical and Mathematical Sciences

Abstract - Glass transition temperature (T_g) is the temperature at which a polymer changes from crystalline state to rubbery state. This change in the property below and above T_g is very important in food science and pharmaceutical industries. In recent decades, there has been a growth in using machine learning (ML) to develop quantitative structure–property relationship (QSPR) models. QSPR uses molecular descriptors and molecular fingerprints as features to predict the properties of chemical compounds. As a result, numerous works have been dedicated to creating a good QSPR model to predict T_g. However, to the best of our knowledge, there was no previous research work that involved the use of the *Mordred* molecular descriptors library or the Extreme Gradient Boosting (*XGBoost*) regression algorithm to predict T_g. Therefore, this project employed *Mordred* and *XGBoost*, together with the *RDKit* cheminformatics library to predict T_g of 640 polymers. A total of 12 sets of features were generated by *RDKit* and *Mordred* as inputs for *XGBoost* to predict T_g. The scoring metrics from the *Scikit-learn* and *Numpy* libraries showed that the 2D molecular descriptors of *Mordred* (Mordred-2D) and the Extended-Connectivity Fingerprint with a diameter of 4 bonds (ECFP4) had the best performances. The results further improved when Mordred-2D and ECFP4 were combined to form a new set of features. Future work aims to increase the number of polymer data points and explore better methods to represent the polymer repeating units for the calculation of descriptors and fingerprints.

Keywords - Glass Transition Temperature (T_g), Polymers, Machine Learning (ML), *RDKit*, *Mordred*, Extreme Gradient Boosting (*XGBoost*), Molecular descriptors, Molecular fingerprints, Quantitative Structure-Property Relationship (QSPR)

1 INTRODUCTION

1.1 BACKGROUND INFORMATION

Glass transition temperature (T_g) is the temperature at which a polymer begins to change from being hard and brittle to become more soft and pliable. In other words, below T_g the polymer behaves like a glass or crystal, and above T_g the polymer behaves like a rubbery material. Therefore, T_g reflects the flexibility of the polymer chain, and it generally increases with the rigidity of the polymer chain. [1 - 3]

The relationship of T_g with crystalline state and rubbery state of polymers has very important industrial applications. For example, considering food stability during food processing, an increase in molecular mobility around T_g will increase the diffusion of the food molecules, resulting in physical and chemical changes in the food systems. Such changes include agglomeration, caking, collapse and nonenzymatic browning. [4] Another important application of T_g is found in the modification of the physical properties of polymeric drug molecules. Increasing the T_g allows the polymeric drug molecules to be maintained in amorphous solid forms at ambient or body temperatures. This modification improves the solubility, bioavailability and physical stability of the drug, which also improves its effectiveness in the human body. [5]

1.2 MACHINE LEARNING

The field of machine learning (ML) has grown significantly over the past decade, as a result of fundamental advances in ML algorithms and broader accessibility of computational hardware that made ML an integral part of many tools and services worldwide. One important ML technique is supervised learning (SL). An SL model learns from the inputs and outputs in the training dataset and make predictions on the unknown data. [6]

An example of SL in computational chemistry is the quantitative structure–property relationship (QSPR) modelling. A QSPR model quantifies and relates the determining factors for a particular measured property with molecular features of a given system of chemical compounds. It is a mathematical model that connects experimental property values (output) with a set of features (input) derived from the molecular structures. [7]

1.3 OBJECTIVE

There had been projects dedicated to building QSPR models for the prediction of T_g with varying levels of success. Some similarities in the projects include: the features involved were molecular descriptors and molecular fingerprints, and regression models were evaluated for their abilities to predict T_g. [8 - 11]

Similarly, in this project a QSPR model was developed to predict T_g of polymers. The features were calculated based on the 3D structures of the monomer molecules of the polymers. The *RDKit* [12] cheminformatics library and the *Mordred* [13] molecular descriptors library were used to calculate the features. The ML model employed in this project was Extreme Gradient Boosting (*XGBoost* [14]) regression algorithm. What differentiated this project from the others was that, to the best of our knowledge, there were no previous projects that employed *Mordred* and *XGBoost* to predict T_g.

2 METHODOLOGY

2.1 DATASET

A dataset of 640 polymers was used. The dataset was provided by professor Atsushi Goto and his research team. [15] Majority of polymers in the dataset were classified under four main parent structures: 24 polyacrylamides, 147 polyacrylates, 39 polyitaconates, 242 polymethacrylates, and 172 polystyrenes. The remaining 16 polymers were classified as 'others'. The chemical space of the polymers in the dataset comprised 11 elements (H, B, C, N, O, F, Si, P, S, Cl, Br). The distribution plot of the T_g values of the dataset is visualized in figure 1. The T_g values range from -92.0°C to 260.0°C, with a mean of 67.5°C and a standard deviation of 61.2°C.

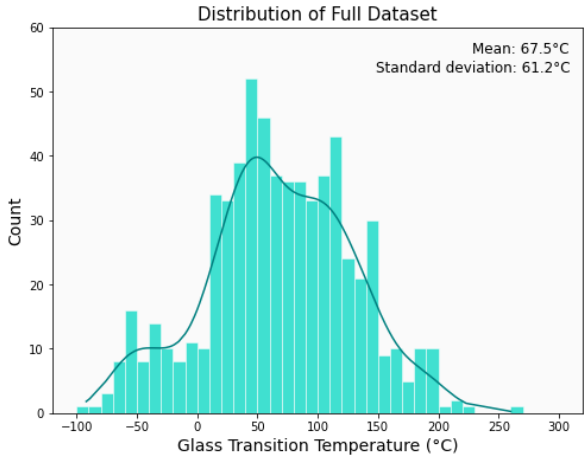

Figure 1. Distribution of T_g values for 640 polymers

In order for the *RDKit* and *Mordred* to be able to process the molecular structures of the polymers, the polymers needed to be in the *MDL SDfile* (*.mol* file extension) format. The first step was to use the *GaussView* [16] software to draw the 3D structure of the monomer molecules, which were saved as *Gaussian Input* files (*.gjf* file extension). This was followed by the geometry optimization using the *Gaussian* [17] software, with the *.gjf* files as inputs, and the outputs being the *Gaussian Output* files (*.out* file extension), which contained the molecules with optimized 3D geometries. The optimization job was performed with the following settings: ground-state, discrete Fourier transform (DFT), default spin (B3LYP), 6-31G (d,p) basis set. The *.out* files were converted to the *Sybyl MOL2* (*.mol2* file extension) format after inspection using *GaussView*. The *.mol2* files were further processed in the *Avogadro* [18,19] software to assign any formal charges to atoms wherever necessary and applicable (e.g. in the NO_2 group the N has a formal charge of +1 and one of the O has a formal charge of -1). Finally, the *.mol2* files were converted to *.mol* format using *Avogadro*.

2.2 FEATURE SELECTION

After the *.mol* files with optimized geometries were created, they were converted into *Mol* objects using *RDKit*. [20] The *Mol* objects were used to calculate molecular descriptors and molecular fingerprints.

RDKit was also used for the calculation of 200 2D descriptors, 10 3D descriptors and 8 fingerprints. [20-22] The 200 2D descriptors were considered 1 set of features (RDKit-2D) and the 10 3D descriptors another set (RDKit-3D). The 8 molecular fingerprints were each considered a set of features, and they were: Avalon, Daylight, Extended-Connectivity Fingerprint (ECFP) with a diameter of 4 bonds (ECFP4), ECFP with a diameter of 6 bonds (ECFP6), 2 electro-topological state fingerprints (Estate1 and Estate2), Extended-reduced Graph (ErG) and Molecular Access System (MACCS) keys. [20, 23-29]

Mordred was used to calculate 1613 2D descriptors and 213 3D descriptors. [13] Upon inspection, it was found that there were descriptors containing invalid values ("*encountered in double_scalars*"), which were removed. As a result, there were 1273 2D descriptors and 56 3D descriptors remaining. The 1273 2D descriptors were considered a set (Mordred-2D) and the 56 3D descriptors another set (Mordred-3D).

In total, there were 12 sets of features in this project.

Feature selection was then performed to remove redundant features, with the aim of computational runtime reduction and also to improve the performance of the QSPR model. [30] The features with zero variance were removed by the feature

selector module in the *scikit-learn* (*sklearn*) library, *sklearn.feature_selection.VarianceThreshold*. [31]

2.3 QSPR MODEL

The *XGBoost* regression model was applied. *XGBoost* is the enhanced version of the gradient boosting (GB) algorithm. GB works by building an additive model in a forward stage-wise fashion, which allows the arbitrary differentiable loss functions to be optimized. [32] It was selected for this project for its high scalability and that it runs 10 times faster than existing popular machine learning algorithms. It was also the number one choice of algorithm for 17 out of 29 winning solutions in an ML competition organized by Kaggle. [33]

Before applying *XGBoost*, the dataset of 640 polymers was rearranged in ascending order of the T_g values, followed by being divided into the train-validation dataset containing 600 data points and the test dataset containing 40 data points. The test dataset was obtained by extracting a data point every 16 steps from the rearranged 640 data points. This step was performed to ensure that both the train-validation dataset and the test dataset were representative of the original unsplit dataset.

The distribution plot of the T_g values of the train-validation dataset is visualized in figure 2. The T_g values range from -89.0°C to 260.0°C, with a mean of 67.7°C and a standard deviation of 61.2°C.

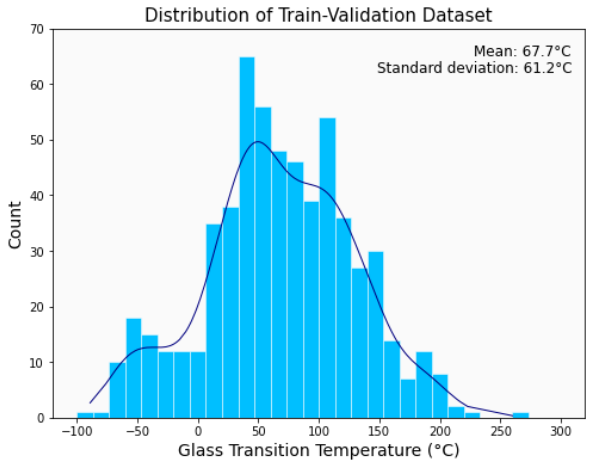

Figure 2. Combined distribution of T_g values for train and validation datasets.

Similarly, the distribution plot of the T_g values of the test dataset is visualized in figure 3. The T_g values range from -92.0°C to 189.0°C, with a mean of 63.6°C and a standard deviation of 62.0°C.

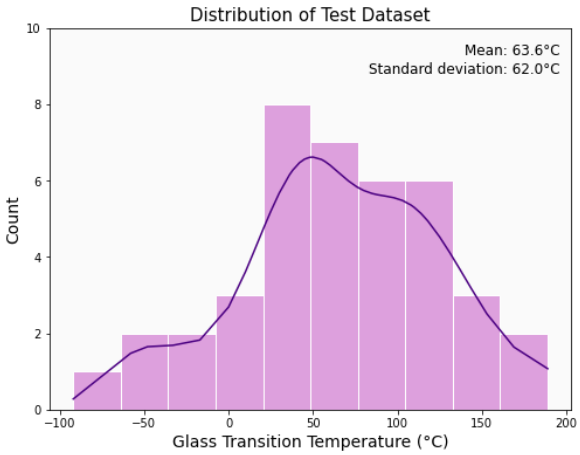

Figure 3. Distribution of T_g values for test dataset

Based on figures 2 and 3, it was concluded that both the train-validation dataset and the test dataset were representative of the original unsplit dataset.

The train-validation data was shuffled by rows pseudo-randomly 50 times to obtain 50 unique permutations, each corresponding to one random state from 1 to 50. These permutations were split into train and validation datasets at a ratio of 80:20, using the *sklearn.model_selection.train_test_split* module. [34] Each of the 50 train datasets were used to train the *XGBoost* regressor. The trained regressor was then validated on the corresponding 50 validation datasets.

2.4 CROSS VALIDATION

Cross-validation (CV) is a method to assess and avoid the overfitting of a ML model. Overfitting is when a model achieves a perfect score when trained on the training data, but fail to predict anything useful on an unseen data. [35]

Fivefold CV was conducted in this project. Manually, 1/5 of data points were extracted from the train-validation dataset of 600 data points. This was performed by extracting a data point every 5 steps. The remaining 4/5 were used to train the *XGBoost* regressor. The trained regressor was validated using the extracted 1/5 CV test data. This process was conducted 5 times, where in each round, the first, second, third, fourth and fifth data point for every 5 steps was extracted respectively. The average result of the 5 runs was obtained. Manual cross-validation was performed to ensure the reproducibility of the results.

2.5 SCORING METRICS

The scoring metrics used in this project to evaluate the performance of the potential QSPR models

were the coefficient of determination (r^2), the root-mean-square error (RMSE), the standard deviation (SD) of the 50 runs and the best random state out of the 50 random states.

The r^2 score represents the proportion of variance in the outcome variable which is explained by the predictor variables in the sample. [36] The best possible score is 1.0. A negative score indicates that the model is arbitrarily worse. A constant model that always predicts the expected value of output, disregarding the input features, would yield a r^2 score of 0.0. The r^2 score was calculated using the *sklearn.metrics.r2_score* module. [37]

The mean-square error (MSE) is a measure of closeness of the estimator to the true value. MSE reflects the accuracy of the estimator, i.e., how much the expected values deviate systematically from the actual values, and also the precision of the estimator, which indicates how much variation about the expected value due to sampling variability. [38] The RMSE is simply the square root of the MSE. Both MSE and RMSE can be calculated using the *sklearn.metrics.mean_squared_error module*. [39]

SD is a measure of the spread of a distribution and in this case, measuring the extent of variation in the r^2 or RMSE scores of 50 runs. A lower SD means a more consistent performance. SD was calculated with the *numpy.std* module in the *Numpy* library. [40]

The best random state is the random state that produces either the best maximum r^2 score or the best minimum RMSE score.

3 RESULTS

All the procedures described in the Methodology section were conducted for all the 12 sets of features. A preliminary evaluation was conducted on the 50 validation datasets for the 12 sets of features. The 4 best performing sets were further evaluated using the test dataset. The 2 best performing sets out of the 4 would be combined to assess whether there were any improvements in the performance of the QSPR model.

For a QSPR model to be considered good for prediction, it should at least achieve $r^2 > 0.6$, q^2 (cross-validation r^2) > 0.5, and RMSE should be as low as possible [41, 42]

3.1 PRELIMINARY EVALUATION

Preliminary evaluation was performed on all the 12 sets of features individually. The results for the maximum scores, minimum scores, mean scores, SDs, CV scores, and best random states are shown in tables 1a, 1b, 2a, and 2b.

Table 1a. The r^2 scores of the 12 sets of features in descending order in terms of maximum r^2.

Descriptors/ Fingerprints	Max r^2	Min r^2	Mean r^2
ECFP6	0.82583	0.62438	0.72435
ECFP4	0.81459	0.62213	0.72724
Estate1	0.81003	0.47866	0.65660
Mordred-2D	0.80037	0.54036	0.69826
MACCS	0.78781	0.50316	0.67297
Daylight	0.78714	0.50704	0.65422
RDKit-2D	0.77881	0.55823	0.68100
Avalon	0.77787	0.45570	0.66888
Estate2	0.76953	0.48375	0.65749
Mordred-3D	0.67761	0.35213	0.52007
ErG	0.53861	0.13926	0.35186
RDKit-3D	0.25782	-0.29200	0.07192

Table 1b. The SD, mean q^2 and best random states. Features are arranged based on table 1a.

Descriptors/ Fingerprints	Standard deviation of r^2	Mean 5-fold CV q^2	Best random state
ECFP6	0.04864	0.71290	7
ECFP4	0.04721	0.72765	7
Estate1	0.07265	0.66794	7
Mordred-2D	0.06249	0.68458	28
MACCS	0.06683	0.66809	7
Daylight	0.06431	0.64635	7
RDKit-2D	0.05225	0.67959	24
Avalon	0.06242	0.69544	24
Estate2	0.06265	0.70898	47
Mordred-3D	0.07172	0.54282	50
ErG	0.10330	0.35122	47
RDKit-3D	0.10846	0.12901	34

According to tables 1a and 1b, RDKit-3D and ErG were the 2 set of features with the worst performances. They did not meet the minimum requirement for r^2 and q^2 scores for a good QSPR

model. The maximum and mean r^2 scores for both were below 0.6. They also had the highest SDs for the 50 runs of r^2 scores. On the other hand, the 4 set of features with the best performances were ECFP6, ECFP4, Estate1 and Mordred-2D. Their maximum r^2, mean r^2 and q^2 scores satisfied the minimum requirements for a good QSPR model. Their maximum r^2 scores were above 0.8.

Table 2a. RMSE scores of the 12 sets of features in ascending order in terms of minimum RMSE.

Descriptors/ Fingerprints	Min RMSE	Max RMSE	Mean RMSE
Daylight	26.152	40.741	35.246
ECFP6	26.573	36.413	31.510
Mordred-2D	26.796	40.281	32.937
ECFP4	27.479	36.081	31.340
RDKit-2D	27.693	38.847	33.922
Avalon	28.441	42.808	34.541
Estate1	29.185	40.972	35.063
Estate2	30.010	40.278	35.090
MACCS	30.139	41.878	34.235
Mordred-3D	35.238	49.102	41.658
ErG	41.430	55.635	48.359
RDKit-3D	52.635	67.447	58.010

Table 2b. The SD, mean CV RMSE and best random states. Features are arranged based on table 2a.

Descriptors/ Fingerprints	Standard deviation of RMSE	Mean 5-fold CV RMSE	Best random state
Daylight	2.708	36.290	40
ECFP6	2.452	32.573	37
Mordred-2D	3.204	34.157	28
ECFP4	2.211	31.664	37
RDKit-2D	2.483	34.364	40
Avalon	3.000	33.662	41
Estate1	3.039	35.072	47
Estate2	2.517	32.675	28
MACCS	2.624	34.850	33
Mordred-3D	3.152	41.827	37
ErG	3.467	49.055	41
RDKit-3D	3.680	57.044	35

Based on tables 2a and 2b, RDKit-3D and ErG were once again the 2 set of features with the worst performances. Both set of features had the highest RMSE scores and SDs. On the other hand, the 4 set of features with the best performances were Daylight, ECFP6, ECFP4, and Mordred-2D. They had the lowest minimum RMSE scores.

After the evaluation of the 12 sets of features, feature selection was performed. The comparison of algorithm runtimes and number of features before and after the feature selection are shown in tables 3a and 3b.

Table 3a. Runtime comparison in descending order before and after feature selection.

Descriptors/ Fingerprints	Runtime/s (Before)	Runtime/s (After)
Mordred-2D	123.772	117.245
Daylight	59.436	59.616
ECFP4	56.025	33.801
ECFP6	54.935	45.661
Avalon	18.979	19.329
RDKit-2D	18.547	15.479
Mordred-3D	15.242	15.243
ErG	13.427	11.528
MACCS	10.953	10.562
RDKit-3D	9.165	9.165
Estate2	8.886	8.262
Estate1	8.492	7.330

Table 3b. Comparison of the number of features in descending order of runtime (from table 3a) before and after feature selection.

Descriptors/ Fingerprints	Number of features (Before)	Number of features (After)
Mordred-2D	1273	1102
Daylight	2048	2046
ECFP4	2048	1149
ECFP6	2048	1696
Avalon	512	506
RDKit-2D	200	165
Mordred-3D	56	56
ErG	315	203
MACCS	167	146
RDKit-3D	10	10
Estate2	79	36
Estate1	79	36

Based on tables 3a and 3b, reduction in the runtime was observed for the sets of features with reduced number of features. ECFP4 had the most reductions in both number of features and runtime. However, it was found that feature selection did not change the results of the scoring metrics at all.

3.2 TEST DATASET EVALUATION

Based on the preliminary evaluation using the 50 validation datasets, it was found that there were 5 sets of features that could potentially provide good results in the evaluation of test dataset. The 5 sets were: Daylight, ECFP4, ECFP6, Estate1 and Mordred-2D. Daylight performed well for RMSE. Estate1 performed well for r^2. ECFP4, ECFP6 and Mordred-2D performed well in terms of both r^2 and RMSE scores.

Hence, further evaluation was conducted on the 5 sets of features. The test dataset evaluation was divided into 2 groups of 4 sets of features. One group had r^2 as the main scoring metric, comprising ECFP4, ECFP6, Estate1 and Mordred-2D. The other group had RMSE as the main scoring metric, comprising Daylight, ECFP4, ECFP6 and Mordred-2D. The results are shown in tables 4 and 5.

Table 4. Results for test dataset, with r^2 as the main scoring metric, arranged in descending order of r^2 scores.

Descriptors/ Fingerprints	Best Random state	r^2 of test data	RMSE of test data
Mordred-2D	28	0.80732	27.200
ECFP4	7	0.80228	27.554
ECFP6	7	0.73235	32.058
Estate1	7	0.67807	35.159

Table 5. Results for test dataset, with RMSE as the main scoring metric, arranged in ascending order of RMSE scores.

Descriptors/ Fingerprints	Best Random state	RMSE of outsider data	r^2 of outsider data
Mordred-2D	28	27.200	0.80732
ECFP4	37	30.370	0.75980
ECFP6	37	36.733	0.64860
Daylight	40	38.849	0.60695

Using r^2 as the main scoring metric, it was found that ECFP4 and Mordred-2D had consistent performances with $r^2 > 0.8$ for both the 50 validation datasets and the test dataset. On the contrary, ECFP6 and Estate1 did not perform as well, with both of their r^2 were below 0.8 for the test dataset. The RMSE scores for ECFP4 and Mordred 2D were also lower than those of ECFP6 and Estate1.

Using RMSE as the main scoring metric, it was found that only Mordred-2D achieved an RMSE score < 30, and also an r^2 score > 0.8. Nevertheless, ECFP4 was still the second best performing set. Comparison of results between tables 4 and 5 suggested that using r^2 as the main scoring metric resulted in better performances of the potential QSPR models than using RMSE.

3.3 COMBINING FEATURES

Based on the test dataset evaluation, it was found that ECFP4 and Mordred-2D were the 2 best performing sets. As a result, they were selected to undergo further evaluation. They were combined to form a new set of features (Mordred-2D + ECFP4). This set of features was subjected to the same train-validation evaluation similar to sub-section 3.1 and also the test dataset evaluation in sub-section 3.2. Results are shown in tables 6a, 6b, 7a, 7b and 8.

Table 6a. Comparison of combined and non-combined features. The r^2 scores arranged in descending order in terms of maximum r^2.

Descriptors/ Fingerprints	Max r^2	Min r^2	Mean r^2
ECFP4	0.81459	0.62213	0.72724
Mordred-2D + ECFP4	0.81334	0.62749	0.72006
Mordred 2D	0.80037	0.54036	0.69826

Table 6b. The SD, mean q^2 and best random states. Features are arranged based on table 6a.

Descriptors/ Fingerprints	Standard deviation of r^2	Mean 5-fold CV q^2	Best random state
ECFP4	0.04721	0.72765	7
Mordred-2 + ECFP4	0.04539	0.71207	40
Mordred 2D	0.06249	0.68458	28

Based on tables 6a and 6b, evaluation of the 50 validation datasets revealed that the combined set of features not only satisfied the minimum requirement for r^2 and q^2 scores for a good QSPR model, it also achieved a maximum $r^2 > 0.8$. However, it is noteworthy that ECFP4 on its own produced slightly better results in terms of the maximum r^2 and mean r^2, although the combined set performed better than ECFP4 in terms of minimum r^2. Additionally, the combined set had the lowest SD for the 50 runs of r^2 scores, which suggests that combining Mordred-2D and ECFP4 sets improved the consistency of the results.

Table 7a. Comparison of combined and non-combined features. The RMSE scores arranged in ascending order in terms of minimum RMSE.

Descriptors/ Fingerprints	Min RMSE	Max RMSE	Mean RMSE
Mordred-2D + ECFP4	24.489	39.548	31.813
Mordred-2D	26.796	40.281	32.937
ECFP4	27.479	36.081	31.340

Table 7b. The SD, mean CV RMSE and best random states. Features arranged as in table 7b.

Descriptors/ Fingerprints	Standard deviation of RMSE	Mean 5-fold CV RMSE	Best random state
Mordred-2D + ECFP4	2.747	32.619	40
Mordred-2D	3.204	34.157	28
ECFP4	2.211	31.664	37

Based on tables 7a and 7b, evaluation of the 50 validation datasets revealed the combined set had the best performance with the lowest minimum RMSE. In terms of maximum RMSE, mean RMSE, SD and CV RMSE, Mordred-2D performed the worst and ECFP4 the best. It was noticed that in both cases of using r^2 and RMSE as the main scoring metrics of evaluation, the same best random states of Mordred-2D (random state 28) and the combined set (random state 40) accounted for the corresponding best performances of these 2 sets, with the combined set performing better than Mordred-2D in both cases. As for ECFP4, 2 different random states accounted for the r^2 and RMSE. This suggests that ECFP4 might not be as consistent in its performance compared to those of Mordred-2D and the combined set.

Table 8. Results for test dataset, with r^2 as the main scoring metric, features are arranged in descending order of r^2 scores.

Descriptors/ Fingerprints	Best random state	r^2 of outsider data	RMSE of outsider data
Mordred-2D + ECFP4	40	0.81187	26.877
Mordred-2D	28	0.80732	27.200
ECFP4	7	0.80228	27.554
ECFP4	37	0.75980	30.370

The results in table 8 are further visualized in graphical format in figures 4, 5, 6 and 7.

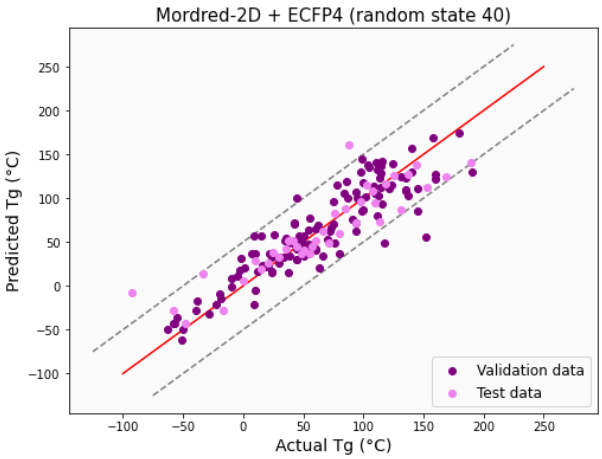

Figure 4. Plot of predicted T_g vs. actual T_g values for combined set (Mordred-2D + ECFP4), at random state 40.

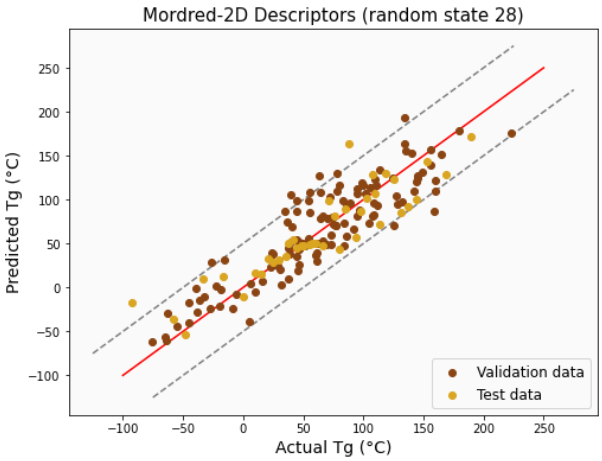

Figure 5. Plot of predicted T_g vs. actual T_g values for Mordred-2D, at random state 28.

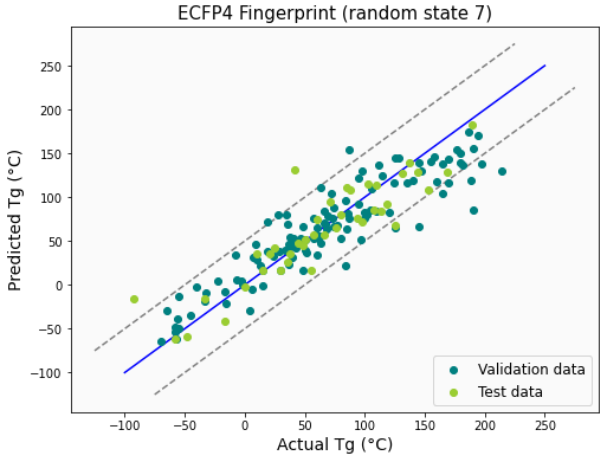

Figure 6. Plot of predicted T_g vs. actual T_g values for ECFP4, at random state 7.

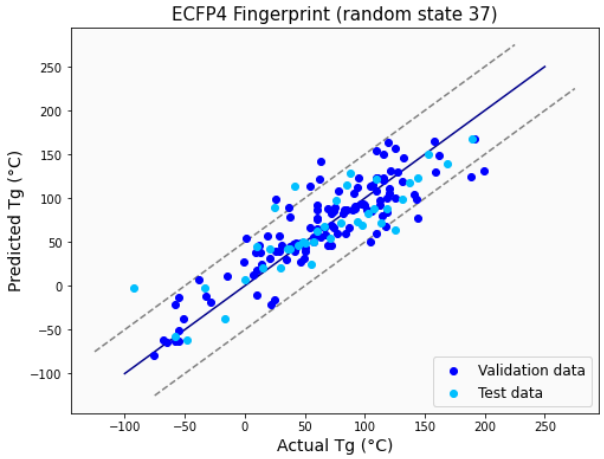

Figure 7. Plot of predicted T_g vs. actual T_g values for ECFP4, at random state 37.

Comparison of figures 4, 5, 6 and 7 revealed that the data points for both the 50 validation datasets and the test dataset were generally close to the identity function line for all the 4 sets of features listed in table 8. Mostly importantly, the combined set represented by figure 4 had the least number of data points outside of the 2 dotted boundary lines.

Based on table 8, evaluation of the test dataset had revealed the combined set had the best performance in terms of both r^2 and RMSE score. The results had also supported the suggestion that the combined set had a more consistent performance than Mordred-2D and ECFP4. The comparison of ECFP4 with the 2 different best random states (random state 7 for best r^2 and random state 37 for best RMSE) supported the suggestion that r^2 was a better scoring metric than RMSE in this project for deciding the best set of features and its random state for the QSPR model.

4 CONCLUSION

In summary, a QSPR model has been successfully developed to predict T_g of polymers, based on the 3D molecular structures of 640 polymers.

Firstly, the monomer molecules of the polymers were drawn using the *GaussView* software and their molecular geometries were optimized by the *Gaussian* software.

Secondly, the *Avogadro* software was used to make minor adjustments to the optimized molecules and also to convert the molecule files into a suitable file extension for the *RDKit* and *Mordred* libraries to iterate and calculate the relevant molecular descriptors and molecular fingerprints. There were 12 sets of descriptors and fingerprints involved in this project. These descriptors and fingerprints were the input features intended for training the *XGBoost* regression model.

Next, the 640 polymer data were split into the train-validation dataset and the test dataset. The train-validation dataset was shuffled pseudo-randomly into 50 unique permutations before a further 80/20 split. *XGBoost* was applied on the 50 validation datasets and the test dataset. The performances of the potential QSPR models were evaluated by the r^2, q^2, RMSE and SD scoring metrics.

Preliminary evaluation was conducted on the 50 validation datasets with the 12 sets of features. Feature selection was performed, and it only reduced the algorithm runtimes, while results of the scoring metrics remained unchanged. A further evaluation was conducted on the test dataset with the best performing sets features in the preliminary evaluation. The 2 best performing sets of features in the evaluation of test dataset were found to be Mordred-2D and ECFP4. These 2 sets were combined and there were improvements in results.

In conclusion, a combined set of features containing the 2D molecular descriptors of the *Mordred* library and the Extended-Connectivity Fingerprint with a diameter of 4 bonds was found to perform the best with the *XGBoost* regression algorithm for the prediction of T_g.

In the future work, the number of polymers in the original raw dataset will be increased and similar ML procedures will be repeated to see if the results can be improved. Next, the method of 3D molecular structures representation will also be reviewed. In this project, the polymers were represented in their monomeric forms. However, the monomer structure may not accurately represent repeating unit of the polymer. For example, the alkene functional group in the monomer is absent in the polymer after the polymerization. In other cases, the monomer is a cyclic compound like epoxide.

ACKNOWLEDGMENT

We would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project.

REFERENCES

[1] Ebnesajjad, S. Introduction to Plastic. In *Chemical Resistance of Commodity Thermoplastics*; William Andrew, **2016**; pp xiii-xxv.

[2] Dominguez, J. C.; Rheology and curing process of thermosets. In *Thermosets*, 2nd Edition; Elsevier, **2018**; pp 115-146.

[3] Wang, R.; Zheng, S.; Zheng, Y. Matrix materials. In *Polymer Matrix Composites and Technology*; Woodhead, **2011**; pp 101-167.

[4] Balasubramanian, S.; Devi, A.; Singh, K. K.; D Bosco, S. J.; Mohite, A. M. Application of Glass Transition in Food Processing. *Crit. Rev. Food Sci. Nutr.* **2016**, *56(6)*, 919-936.

[5] Jadhav, N. R.; Gaikwad, V. L.; Nair, K. J.; Kadam, H. M. Glass transition temperature: Basics and application in pharmaceutical sector. *Asian J. Pharm*. **2009**, *3(2)*, 82-89.

[6] Prezhdo, O. V. Advancing Physical Chemistry with Machine Learning. *J. Phys. Chem. Lett*. **2020**, *11(22)*, 9656–9658.

[7] Katritzky, A. R.; Kuanar, M.; Slavov, S.; Hall, C. D.; Karelson, M.; Kahn, I.; Dobchev, D. A. Quantitative Correlation of Physical and Chemical Properties with Chemical Structure: Utility for Prediction. *Chem. Rev.* **2010**, *110(10)*, 5714–5789.

[8] Jha, A.; Chandrasekaran, A.; Kim, C.; Ramprasad, R. Impact of dataset uncertainties on machine learning model predictions: the example of polymer glass transition temperatures. *Modelling Simul. Mater. Sci. Eng.* **2019**, *27*, 024002.

[9] Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition Temperatures: A case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model*. **2019**, *59*, 5013-5025.

[10] Jiang, Z.; Hu, J.; Marrone, B. L.; Pilania, G.; Yu, X. A Deep Neural Network for Accurate and Robust Prediction of Glass Transition Temperature of Polyhydroxyalkanoate Homo- and Copolymers. *Materials*. **2020**, *13*, 5701.

[11] Wen, C.; Liu, B.; Wolfgang, J.; Long, T. E.; Odle, R.; Cheng, S. Determination of glass transition temperature of polyimides from atomistic molecular dynamics simulations and machine-learning algorithms. *J. Polym. Sci.* **2020**, *58*: 1521-1534.

[12] Landrum, G. An overview of the RDKit. http://www.rdkit.org/docs/Overview.html (accessed May 12, 2020).

[13] Moriwaki, H.; Tian, Y.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *J. Cheminformatics*. **2018**, *10*, 4.

[14] XGBoost Developers. XGBoost Tutorials. Introduction to Boosted Trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html (accessed May 12, 2020).

[15] Academic Profile: Prof Atsushi Goto. https://dr.ntu.edu.sg/cris/rp/rp01024 (accessed May 12, 2020).

[16] Dennington, R.; Keith, T. A.; Millam, J. M. *GaussView, Version 6*, Semichem Inc., Shawnee Mission, KS, **2016**.

[17] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.; Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.; Hratchian, H. P.; Ortiz, J. V.; Izmaylov, A. F.; Sonnenberg, J. L.; Williams-Young, D.; Ding, F.; Lipparini, F.; Egidi, F.; Goings, J.; Peng, B.; Petrone, A.; Henderson, T.; Ranasinghe, D.; Zakrzewski, V. G.; Gao, J.; Rega, N.; Zheng, G.; Liang, W.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Throssell, K.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M. J.; Heyd, J. J.; Brothers, E. N.; Kudin, K. N.; Staroverov, V. N.; Keith, T. A.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A. P.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Millam, J. M.; Klene, M.; Adamo, C.; Cammi, R.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Farkas, O.; Foresman, J. B.; Fox, D. J. *Gaussian 16, Revision C.01*, Gaussian, Inc., Wallingford CT, **2016**.

[18] Hanwell, M.D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R. Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Cheminformatics*. **2012**, *4*, 17.

[19] Avogadro: an open-source molecular builder and visualization tool. Version 1.2.0. http://avogadro.cc (accessed May 12, 2021).

[20] Landrum, G. Getting Started with the RDKit in Python. https://www.rdkit.org/docs/GettingStartedInPython.html (accessed May 12, 2021).

[21] Landrum, G. rdkit.Chem.Descriptors3Dmodule. https://www.rdkit.org/docs/source/rdkit.Chem.Descriptors3D.html (accessed May 12, 2021).

[22] Wicker, J. G. P.; Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm*, **2015**, *17*, 1927-1934.

[23] Landrum, G. Fingerprints in the RDKit. https://www.rdkit.org/UGM/2012/Landrum_RDKit_UGM.Fingerprints.Final.pptx.pdf

[24] Daylight Fingerprints. https://www.daylight.com/meetings/summerschool01/course/basics/fp.html (accessed May 12, 2021).

[25] Landrum, G. rdkit.Chem.EState.Fingerprinter module https://www.rdkit.org/docs/source/rdkit.Chem.EState.Fingerprinter.html (accessed May 12, 2021).

[26] Stiefl, N.; Watson, I. A.; Baumann, K.; Zaliani, A. ErG: 2D Pharmacophore Descriptions for Scaffold Hopping. *J. Chem. Inf. Model.* **2006**, *46(1)*, 208–220.

[27] Gedeck, P.; Rohde, B.; Bartels, C. QSAR − How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *J. Chem. Inf. Model.* **2006**, *46(5)*, 1924–1936.

[28] Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model*. **2010**, *50(5)*, 742–754.

[29] Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42(6)*, 1273–1280.

[30] Bommert, A.; Sun, X.; Bischl, B.; Rahnenfuhrer, J.; Lang, M. Benchmark for filter methods for feature selection in high-dimensional classification data. *Computational Statistics and Data Analysis*. **2020**, *143*, 106839.

[31] sklearn.feature_selection.VarianceThreshold. https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html (accessed May 12, 2021).

[32] sklearn.ensemble.GradientBoostingRegressor. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html (accessed May 12, 2021).

[33] Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. **2016**, 785-794.

[34] sklearn.model_selection.train_test_split https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (accessed May 12, 2020).

[35] 3.1. Cross-validation: evaluating estimator performance. https://scikit-learn.org/stable/modules/cross_validation.html (accessed May 12, 2021).

[36] Miles, J. R Squared, Adjusted R Squared. *Wiley StatsRef: Statistics Reference Online*. **2014**. https://doi.org/10.1002/9781118445112.stat06627 (accessed May 12, 2020).

[37] sklearn.metrics.r2_score https://scikit-learn.org/stable/modules/generated/sklearn.metrics.r2_score.html (accessed May 12, 2021).

[38] Schluchter, M. D. Mean Square Error. *Wiley StatsRef: Statistics Reference Online*. **2014**. https://doi.org/10.1002/9781118445112.stat05906 (accessed May 12, 2020).

[39] sklearn.metrics.mean_squared_error https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html (accessed May 12, 2021).

[40] The SciPy community. numpy.std. https://numpy.org/doc/stable/reference/generated/numpy.std.html (accessed May 12, 2021).

[41] Tropsha, A.; Gramatica, P.; Gombar, V. K. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci*. **2003**, *22*, 69-77.

[42] Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model*. **2015**, *55*, 1316-1322.