APPLICATION OF MACHINE LEARNING IN MOLECULAR SOLUBILITY PREDICTION USING MOLECULAR FINGERPRINT AND PHYSICOCHEMICAL PROPERTIES MODELS

Dinh Long Huynh – 000624-8397
Course: Preclinical and Clinical Data Analysis in Predictive Drug Discovery and Development – 3FG289

ABSTRACT

This study analyzed three different approaches in the dataset used for building a predictive model: the physicochemical-properties-based (PCP) dataset, the molecular fingerprint-based (MF) dataset, and the combined (MFPCP) dataset. The raw data contained 122 inorganic compounds and complex, which could cause the RF model to perform poorly, as reported previously. To address this issue, the project filtered out inorganic compounds and used the filtered dataset for all three approaches. The results revealed that the MFPCP approach outperformed the other two, with an RMSE of 0.57 an MAE of 0.40, and an R2 of 0.69.

(part 7-8-9-10-11 in Python file)

Keywords: *solubility, predict, model, MLR, KNN, RF, machine learning*

INTRODUCTION

In the previous report, a predictive model was created to forecast molecular solubility using descriptors. Three models were used, including Multiple Linear Regression, K-Nearest Neighbor, and Random Forest Regressor. However, the Random Forest model, which performed better than the other two models, still did not meet the desired performance level. The testing set's RMSE, MAE, and R2 were 0.58, 0.43, and 0.66, respectively. Therefore, I began exploring alternative approaches to improve the model's performance with limited data.

Over the past century, various computational models have been developed to predict molecular solubility before conducting experiments. Two common approaches are the physiochemical feature-based descriptors model (PCP) and the molecular fingerprint-based (MF) model[1,2]. Additionally, Sumin Lee et al. proposed the MFPCP approach, which combines both models. In their report, they showed that the combined dataset can outperform each of these models[3].

My study utilized the same dataset from the previous report and added compounds' fingerprints as new descriptors. Then, I used these descriptors alone and in combination with physicochemical properties to build the Random Forest Model.

METHODS

1. Standardized data

The initial data (df_raw) was preprocessed to remove categorical variables, such as ID, Name, InChI, InChIKey, SMILES, Group, and Occurrences. The resulting data frame, named "df_drop," emerged from this elimination process.

Subsequently, the “df_drop” underwent standardization using mean centering and unit standard deviation methods. Mean centering involves subtracting each observation from the mean of the observations of the variable, effectively centering the new means around zero. Following this, unit standard deviation divides each observation by the standard deviation of the observations of the variable. This process ensures that the derived standard deviation becomes equal to 1. The new observation values are determined through these two methods, providing a standardized representation of the data.

$$\frac{x_i - \mu_i}{\sigma_i}$$

x_i: observation of variable i

μ_i: mean of observations of variable i

σ_i: standard deviation of observations of variable i

In the project, I used the StandardScaler() function in Python to handle these processes. The aim of the standardization step is to improve the calculation rate in the machine learning algorithm. This also ensures similarity in each dimension for distance similarity metrics, which are related to the k-nearest neighbor regressor that we plan to use.

2. Data preparation

First, the data was filtered to eliminate the inorganic compound by removing those that do not contain carbon atoms in their structure. This resulted in the PCP dataset. To create the MF dataset, the RDkit package was used to convert molecules from InChI to Morgan and MACCS fingerprints. Since the 2D fingerprint has been proven to be as effective as 3D fingerprints in predicting solubility[4], the 2D fingerprints were used for simplicity and to optimize calculation time. The Morgan and MACCS fingerprints were transformed into dataframes, with each digital bit assigned to a separate column. These two dataframes were then merged to obtain the total fingerprint, resulting in a table with 2216 columns named 'df_fp' that was used for the MF approach. Finally, the PCP and MF datasets were merged to create the MFPCP dataset.

All three datasets were divided into two subsets: a training set and a testing set, with proportions of 0.8 and 0.2, respectively. The training set was used to build the model, serving as the basis for model fitting and cross-validation. The testing set was used for plotting and comparing between models. The accuracy of the model was also tested using the testing set to assess its performance for future predictions. Another point worth mentioning is that when splitting data, the random_sate was set as a constant number, ensuring 3 models was trained and tested on the same split sets.

3. Random Forest Regressor

According to the previous report, Random Forest (RF) was proven that performed best among MLR, KNN and RF regressor, thus it was used for predicting solubility in this report. To ensure reproducibility of the model, the random_state of RF was set to a constant number. The same approach as the previous report was used to optimize the number of trees in the RF model, using 10-fold cross-validation with the GridSearchCV function in Python. Evaluation scores, including RMSE, MAE, and R2 were collected for each iteration, and their mean and standard deviation were calculated. The 95% confidence interval (CI) was used to compare the performance of the model between different parameters. Based on the evaluation scores, the optimal number of trees was chosen for each MF, PCP, and MFPCP approach.

RESULTS – DICUSSION

1. Data examination

The dataset contains 112 inorganic compounds, which lack a carbon atom in their structure. Several of these inorganic compounds have a strong lattice structure due to their ionic interaction with each other. This makes them relatively had solid state limit in case of solubility. Moreover, d orbitals from the 4th-period elements can cause complexation with ligands, and coincidentally, some of these compounds are complexes. The solubility of these compounds is affected by the ligands and can sometimes lead to non-pattern and misleading results in the models. Therefore, it is recommended to filter out these compounds before training the RF model.

The selected combination of both MACC and Morgan fingerprint is preferred over using each of them alone because they complement each other. MACC offers structural key descriptors, which are represented by a 166-bit string. Each bit corresponds to a pre-defined substructure, so only 166 constant substructures are considered when analyzing compounds. However, there may be

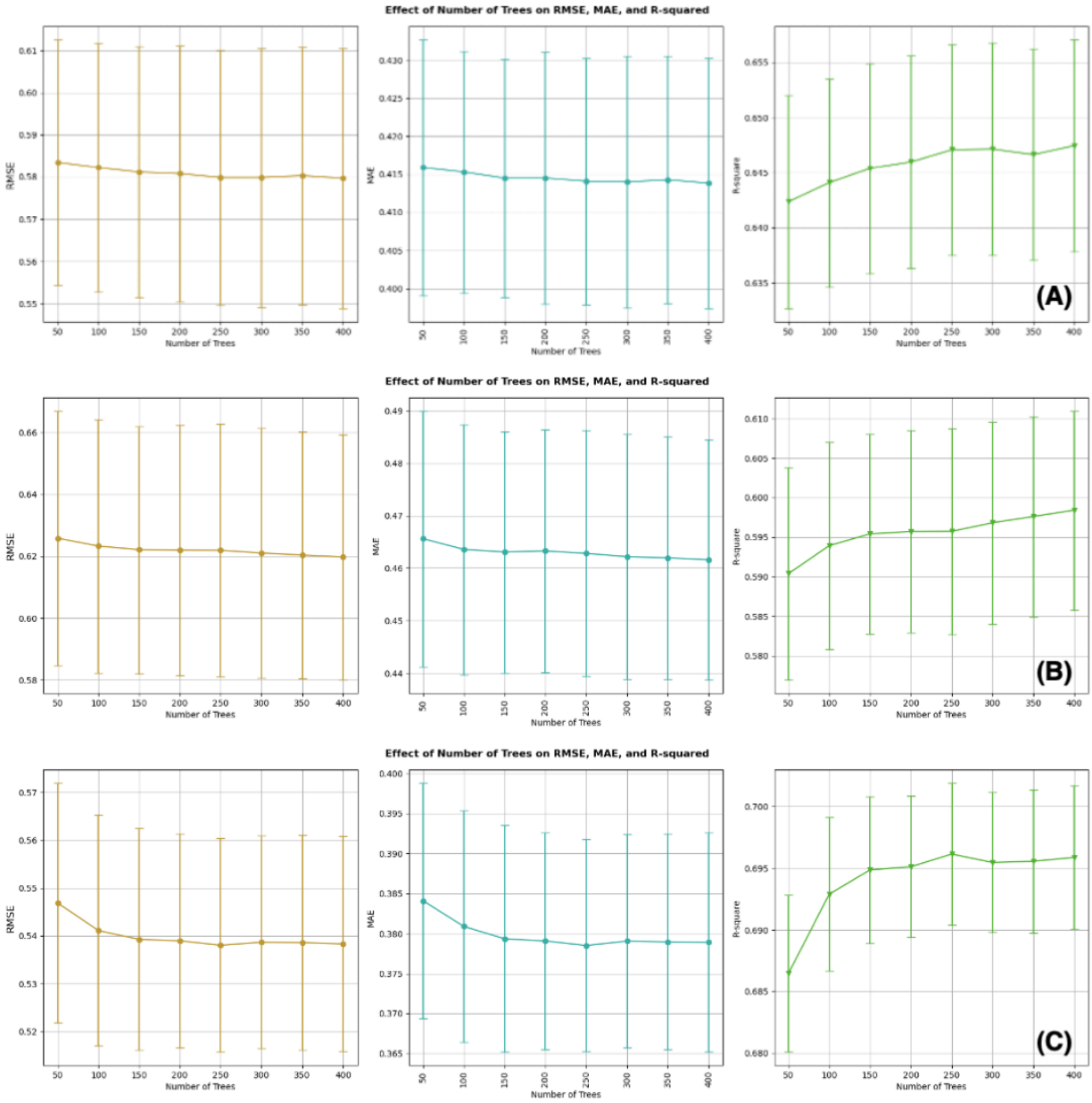

Figure 1. (A) The effect of Tree Number on evaluation scores of PCP approach
(B) The effect of Tree Number on evaluation scores of MF approach
(C) The effect of Tree Number on evaluation scores of MFPCP approach

doubts when there is an external substructure in the molecule. To address this issue, the Morgan fingerprint comes into play. This circular fingerprint is not predefined and can represent an infinite number of different molecular features, which is the perfect complement to MACC keys[5]. However, it is not used alone because it also has a drawback. The Morgan fingerprint cannot be reversely translated back into a molecule, and there is a risk of hash collision, resulting in higher bias models. The combination of MACC and Morgan fingerprint was previously used by Minjian et. al. in the report of QSAR analysis for finding JAK2 Inhibitors, resulting in an R2 of 94% for the training set and 80% for the testing set[6].

2. Random Forest Regression

The cross-validation process showed that the optimal number of trees for the PCP, MF, and MFPCP approaches were 250, 400, and 250, respectively (**Fig. 1**). It is noteworthy that when considering the evaluation scores' 95% CI, there was no significant difference between the different tree numbers for each approach (**Fig. 1**). This suggests that the number of trees can be chosen flexibly within the range of 50 to 400 trees, which is

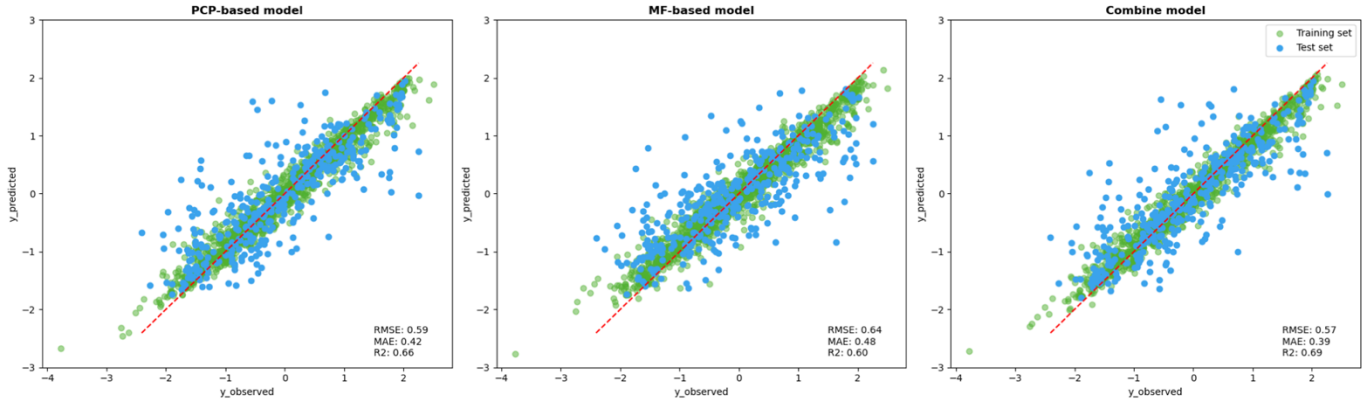

Figure 2. Models performance on testing set

Table 1. Summary of models performance

Datasets	Optimal Estimator (n_estimator)	Training set			Testing set		
		RMSE	MAE	R2	RMSE	MAE	R2
PCP	300	0.58 ± 0.03	0.41 ± 0.02	0.65 ± 0.01	0.59	0.42	0.66
MF	400	0.62 ± 0.04	0.46 ± 0.02	0.60 ± 0.01	0.64	0.48	0.60
MFPCP	250	0.54 ± 0.02	0.38 ± 0.01	0.70 ± 0.01	0.57	0.39	0.69

the investigated range, depending on the data size and computational time.

The trial results on the testing set revealed that the MF approach had a lower performance compared to the PCP approach. For the testing set, the RMSE of the model using MF data increased to 0.64, while for the PCP model, using the same testing set, the score was only 0.59 (**Table 1**). This observation is consistent with the report by Arash et al., where the circular fingerprint-based approach also showed a higher RMSE compared to the descriptor-based approach, which were 0.80 and 0.64, respectively[2]. Furthermore, the same trend was also observed with the greater MAE and lower R2 of MF in comparison to the PCP approach (**Table 1**).

By using both PCP and MF data as descriptors, the performance of PCPMF model was significantly improved as compared to using only one of these descriptors. In fact, the RMSE of the MFPCP approach was much lower, at only 0.57, while it was 0.59 and 0.64 for PCP and MF, respectively (**Table 1**). This is a significant achievement for my dataset as even more complicated models such as SVM or lightGBM have not been able to achieve an RMSE lower than 0.59 in the last report. Additionally, the MFPCP approach had the lowest MAE value at 0.39 and the highest R2 value at 0.69 (**Table 1**). Furthermore, when analyzing the training set, it was found that the MAE and R2 values of PCPMF were significantly lower and higher, respectively, than those of PCP and MF, even when considering the 95% confidence interval (**Table 1**). Additionally, PCPMF was proven to have a significantly lower RMSE than MF with values of 0.54 ± 0.02 and 0.62 ± 0.04, respectively (**Table 1**). This can be attributed to the combination of the two datasets, which resulted in a larger number of features capable of capturing the characteristics of molecules.

CONCLUSION

During my follow-up study, I applied the RF models to predicit molecular solubility on PCP, MF, and MFPCP datasets. Firstly, the inorganic compounds was removed out of data as they can affect the accuracy of the predictive model. Then, the cross-validation was applied to find the optimal tree number for each training dataset. Finally, I used the optimal model to predict solubility in the testing set and calculated performance metrics such as RMSE, MAE, and R2. The results indicated that MFPCP performed better than both

PCF and MF, across all evaluation scores. This suggests that the combination of physicochemical properties and molecular fingerprint is the best predictors of solubility for my dataset.

REFERENCES

1. Ye, Z. & Ouyang, D. Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *J. Cheminformatics* **13**, 98 (2021).
2. Tayyebi, A. *et al.* Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models. *J. Cheminformatics* **15**, 99 (2023).
3. Lee, S. *et al.* Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* **7**, 12268–12277 (2022).
4. Gao, K. *et al.* Are 2D fingerprints still valuable for drug discovery? *Phys. Chem. Chem. Phys. PCCP* **22**, 8373–8390 (2020).
5. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
6. Yang, M. *et al.* Machine Learning Models Based on Molecular Fingerprints and an Extreme Gradient Boosting Method Lead to the Discovery of JAK2 Inhibitors. *J. Chem. Inf. Model.* **59**, 5002–5012 (2019).