Contents

Development of drug like compounds aqueous solubility prediction models using machine learning algorithms

Abstract.

Drug like compounds aqueous solubility is the ability of these compounds to dissolve in water. This study developed machine learning models that can predict drug like compounds aqueous solubility using five different machine learning algorithms (Random Forest Regressor, XGBOOST, LightGBM Regressor and MLP Regressor [ANN]). The dataset sourced from Kaggle was preprocessed and the models were designed as follows: the dataset was split into three sets; Training, Validation and Test in the ratio of 70%, 15% and 15% respectively. The training set underwent cross validation using Grid search and the k-fold method. The research results show that ANN achieved the best result with RMSE values that slightly decreased from 0.583 on Training set to 0.545 on Test set and R^2 values that slightly increased from 0.801(80.1%) on Training set to 0.810(81.0%) on Test set while Random Forest Regressor followed it with RMSE values that slightly increased from 0.418 on Training set to 0.514 on Test set and R^2 values that slightly decreased from 0.897(89.7%) on Training set to 0.831(83.1%) on Test set. Although certain limitations are recognized, the findings provide important insights for pharmaceutical companies. Additionally, they open up opportunities for further research. Future studies could explore more advanced algorithms, incorporate additional features, or collect a larger and more diverse dataset, which could improve the model's accuracy and relevance.

1. Introduction.

In a broad sense, solubility may be defined as the amount of a substance that dissolves in a given volume of solvent at a specified temperature (Alsenz & Kansy, 2007). According to Su and Herrero. 2023, one significant reason why many potential drug candidates fail in the early phase of developing new drugs is because they do not dissolve well in liquid. This poor solubility means they cannot be effectively tested or used, leading to their elimination from further consideration during initial testing. The solubility of a drug in water affects how well it can be absorbed into the bloodstream, distributed to where it's needed, broken down(metabolized), and excreted from the body. It also influences the drug's safety (toxicity). For drugs taken by mouth, good solubility is essential for ensuring that the drug can be absorbed and used effectively by the body.

Ideally, solubility would be measured directly for each compound, but this method is both time-consuming and costly, requiring the compound to be on hand for testing. Given the growing size of molecular screening libraries, which can contain up to 350 million compounds, experimentally determining the solubility of each one is impractical. Therefore, there is a demand for rapid and accurate solubility prediction to complement large-scale virtual screening efforts. (Francoeur & Koes, 2021).

Formally, there are two methods of measuring aqueous solubility of drug like compounds, they are:
Molecular Simulation and Quantum Calculations
According to Hossain et al (2019), Molecular simulation techniques employ statistical mechanics to determine solubility, either by directly computing the chemical potentials of the solute and water or by simulating the solute within explicit water molecules. The direct simulation of the solute can be conducted using various methods, all of which demand substantial computational resources and, in the case of direct calculation, extended periods are required to achieve equilibrium. while according to Ghasemi et al. (2007), quantum mechanics (QM)-based approaches are more advanced than simulation methods and are categorized based on whether the solvent is included in the calculation. Full QM methods incorporate water molecules using density functional theory, providing the most rigorous approach but tend to underestimate the solute's equilibrium density and require substantial computational power. In contrast, continuum solvent methods treat water as a bulk dielectric, saving computational resources but not accounting for the water’s degrees of freedom, and assuming the solute’s charge is confined within the cavity it forms in the solvent.
These approaches require a large amount of computing power in order to perform their calculations.

2. Literature Review

Lovric' et al 2021 in their study to accurately predict the solubility of drug- like compounds in water developed machine learning models using four machine learning algorithms: Random Forests (RF), Light Gradient Boosting Machine (LGBM), Partial Least Squares (PLS), and Least Absolute Shrinkage and Selection Operator (LASSO). According to their results, LASSO achieved the best results with a Root Mean Square Error (RMSE) (test) of 0.70 log points, an R^2 (test) of 0.80, and 105 features.

Dutta et al 2021 in their research to predict aqueous solubility of compounds directly from their molecular structure built machine learning models using three difference machine learning algorithms: Linear Regression, Ridge and Bayesian Ridge. From their result, the algorithms yielded the same results with Mean Squared Error (MSE) of 1.01 and R^2 of 0.77.

Palmer et al, 2007 in their analysis to correctly predict aqueous solubility of compounds developed machine learning models by applying four different machine learning algorithms: Random Forest regression (RF), Partial-Least-Squares (PLS) regression, Support Vector Machines (SVM) and Artificial Neural Networks (ANN). The RF out performed others with Root Mean Square Error (RMSE) value of 0.690 and R^2 value of 0.890 while SVM followed it with Root Mean Square Error (RMSE) value of 0.720 and R^2 value of 0.878.

3. Methodology.

3.1. Data Science Process

Problem identification ➡ Raw dataset collection ➡ Clean/pre-processing of raw dataset ➡ Exploratory data analysis ➡ Build model and Analyze result ➡ Presentation of result in a visual way.

3.1.1. Problem statement.

Studies had shown that Molecular Simulation and Quantum Calculations approaches of an aqueous solubility prediction require a large amount of computing power in order to perform their calculations. In order to overcome this problem, there is the need to apply machine learning method. Thus, this study.

3.1.2. Dataset collection.

The dataset is a secondary data that was gotten from Kaggle. It is aqueous solubility dataset and it consist of 1129 observations and 10 variables.

3.1.2.1. Dataset Attribute Information.

S/N	Variables	Data type	meaning
1	Compound ID	Categorical	The name or identifier of the chemical compound.
2	ESOL predicted log solubility in mols per litre	Numeric	The logarithmic value of the solubility of the compound in water as predicted by the ESOL model, measured in moles per liter
3	Minimum Degree	Numeric	The minimum degree of atoms in the molecule, likely indicating the minimum number of bonds to an atom in the structure.
4	Molecular Weight	Numeric	The molecular weight of the compound in atomic mass units (amu)
5	Number of H-Bond Donors	Numeric	The number of hydrogen bond donors in the molecule, typically the number of hydrogen atoms bonded to electronegative atoms like oxygen or nitrogen
6	Number of Rings	Numeric	The number of ring structures present within the molecule
7	Number of Rotatable Bonds	Numeric	The number of bonds in the molecule that can rotate, affecting the molecule's flexibility.
8	Polar Surface Area	Numeric	The surface area of the molecule that is polar, usually measured in square angstroms.
9	Measured log solubility in mols per litre	Numeric	The logarithmic value of the experimentally measured solubility of the compound in water, measured in moles per liter
10	smiles	Categorical	The Simplified Molecular Input Line Entry System (SMILES) notation, a textual representation of the chemical structure of the compound

4. Objectives.

The main objective of this study and selecting this dataset is to study the dataset, analyze the dataset and use it to develop effective and efficient drug like compound's aqueous solubility prediction machine learning models that can help pharmaceutical companies' decision during drug discovery and production**.**

5. Features of the dataset.

The dataset is a supervised dataset. This is because it has dependent variable: Measured log solubility in mols per litre. It is also a regression task because the target variable (Measured log solubility in mols per litre) is numeric. Furthermore, it is a linear regression for the target variable is continuous. Based on these features, the models will be developed using regression machine learning algorithms.

6. Dataset Importation.

This project was developed using python language and the development environment was Jupyter Notebook. The dataset was imported into the environment using pandas; a python library.

7. Data Cleaning/ preprocess.

Data cleaning is a process in machine learning modeling which focus on detecting and handling corrupt data such as wrong data type, duplicate rows, missing values and outliers to improve the quality of the dataset for effective and efficiency decision making.

First, the dataset was examined for wrong data type, duplicate rows and missing values. From the examination, there was no wrong datatype, duplicate row and missing values.

Then, each numerical variable was examined for outliers and normality using the following:

. Interquartile Range (IQR) Method to view the numbers and the percentage of the outliers.

. Shapiro-Wilk test to analyze normality of the data.

. Data Description to view the statistical summary of the data

. Boxplot and Histogram to visualize outliers and normality.

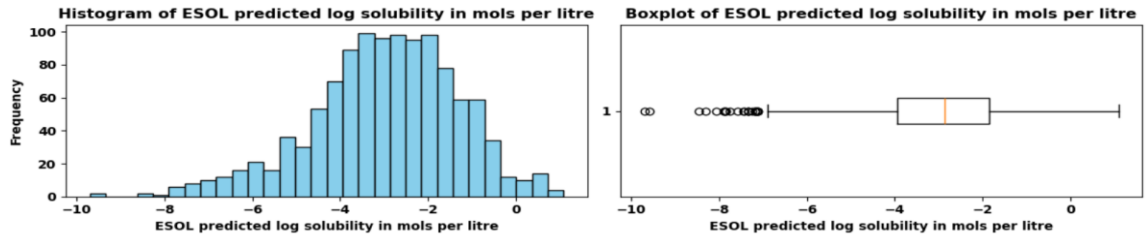
After these examinations and analysis, it was observed that the variables; ESOL predicted log solubility in mols per litre, Minimum Degree, Molecular Weight, Number of H-Bond Donors, Number of Rings, Number of Rotatable Bonds, Polar Surface Area and measured log solubility in mols per litre had outliers and show non normal distribution.

8. Outlier handling, Transformation and Normalization

8.1. ESOL predicted log solubility in mols per litre.

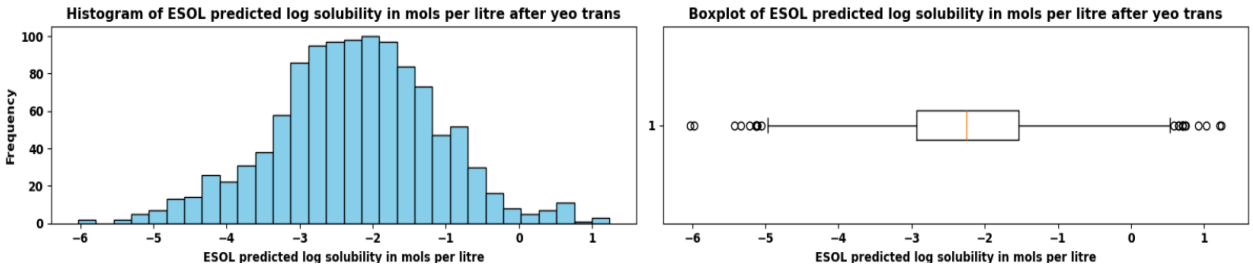
From ESOL predicted log solubility in mols per litre, it was observed that the numbers of outliers were 24 and this made 2.13% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are -9.702000 and 1.091000 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.9845110177993774 and p-value is 1.3833133527541008e-09 as shown by Shapiro-wilk test. Finaly, it was visualizes using boxplot and histogram as shown in Figure a bellow:

Fig.a.



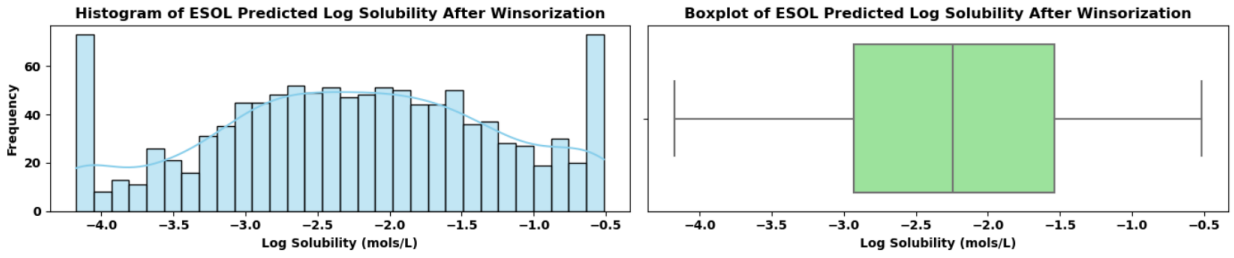
First, the variable (ESOL predicted log solubility in mols per litre) was transformed because the Shapiro-wilk test p-value is less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is -9.702000. Figure b bellow shows boxplot and histogram of ESOL predicted log solubility in mols per litre after transformation.

Fig.b.



Then the outliers were handled using winsorization method. Figure c bellow shows boxplot and histogram of ESOL predicted log solubility in mols per litre after winsorization.

Fig.c

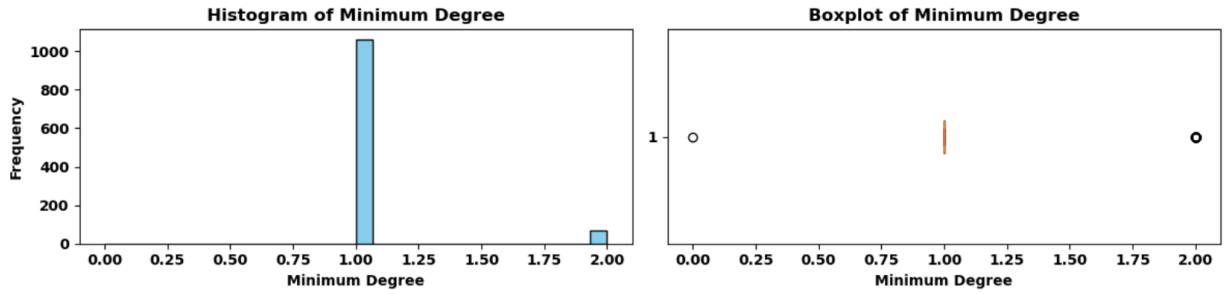


After handling the outliers, the variable was normalized using min-max method. This was done to set the range of its mean and standard deviation between 0 and 1 since many algorithms are sensitive to range(scale). This ensures that the data points have comparable influence on the algorithm.

8.2. Minimum Degree

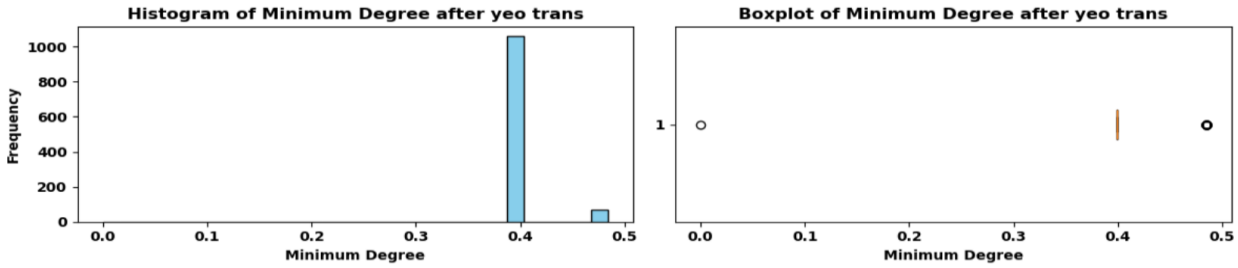
From Minimum Degree, it was observed that there are 68 numbers of outliers and these made 6.03% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are 0 and 2 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.25854331254959106 and p-value is 0.0 as shown by Shapiro-wilk test. Finaly. It was visualizes using boxplot and histogram as shown in Figure d bellow:

Fig.d



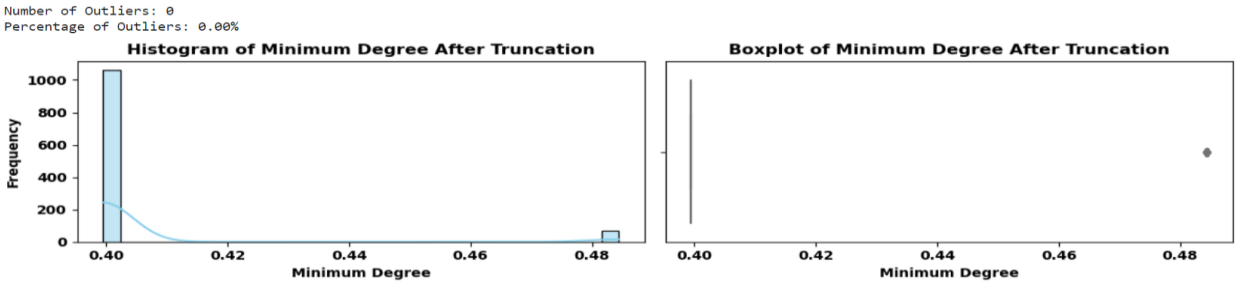
First, the variable (Minimum Degree) was transformed because the Shapiro-wilk test p-value is 0.0. This is less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is 0. Figure e bellow shows boxplot and histogram of Minimum Degree after transformation.

Fig.e



Then the outliers were handled using truncation method after winsorization failed to handle the outliers. Figure f bellow shows interquartile range result, boxplot and histogram of after Minimum Degree truncation.

Fig.f

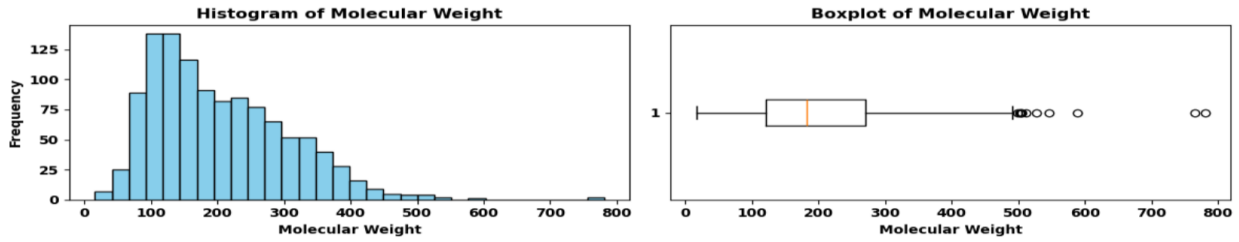


After handling the outliers, the variable was normalized using min-max method

8.3. Molecular Weight

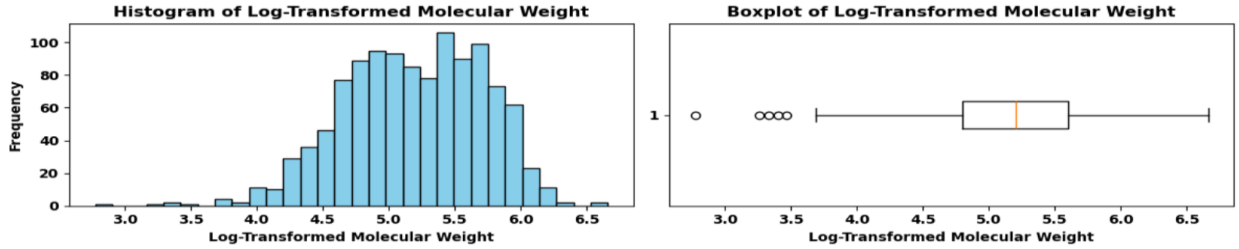
From Molecular Weight, it was observed that there are 10 numbers of outliers and these made 0.89% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are 16.043 and 780.949 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.9423 and p-value is 1.548×10^{-20} as shown by Shapiro-wilk test. Finaly. It was visualizes using boxplot and histogram as shown in fig. 'g' bellow:

Fig.g.



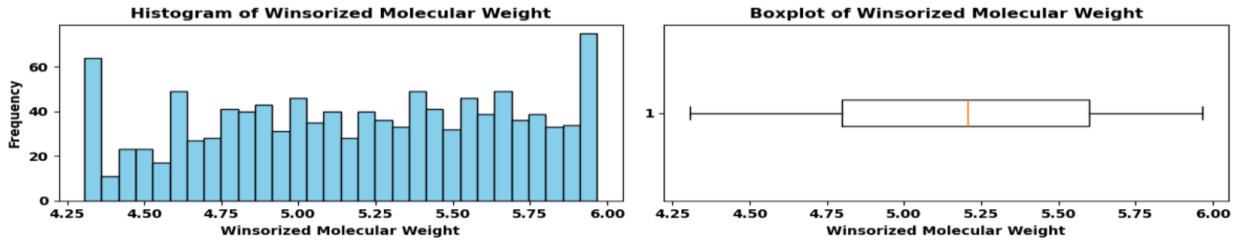
First, the variable (Molecular Weight) was transformed because the Shapiro-wilk test p-value is 1.5478 x^{-20}. This is less than 0.05 which means it does not follow normal distribution and Log transformation method that can handle positive values was used because its minimum and maximum values are 16.0430 and 780.9490 respectively. Figure h bellow shows boxplot and histogram of Molecular Weight after log transformation.

Fig.h.



Then the outliers were handled using winsorization method to handle the outliers. Figure i bellow shows boxplot and histogram of after Molecular Weight winsorization.

Fig.i.

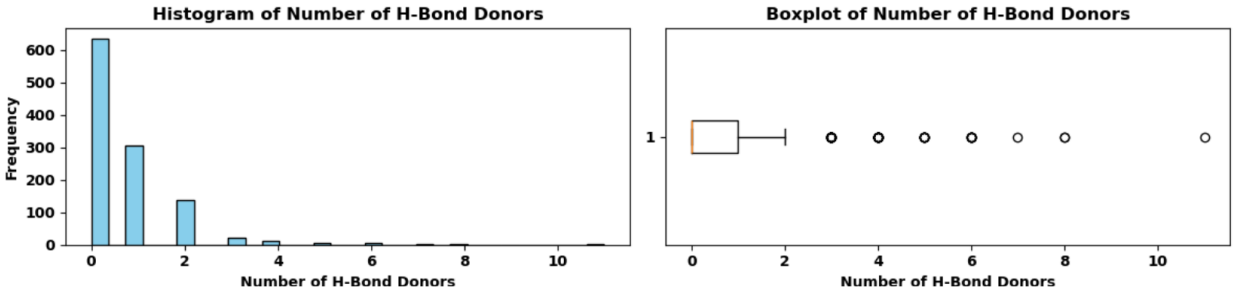


After handling the outliers, the variable was normalized using min-max method.

8.4. Number of H-Bond Donors.

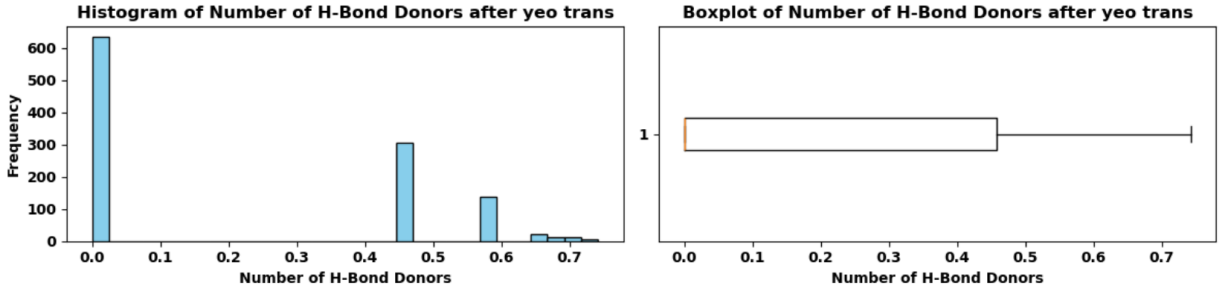
From Number of H-Bond Donors, it was observed that there are 49 numbers of outliers and these made 4.34% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are 0 and 11 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.6510 and p-value is 5.3389×10^{-43} as shown by Shapiro-wilk test. Finaly, it was visualizes using boxplot and histogram as shown in figure j bellow:

Fig.j.



First, the variable (Number of H-Bond Donors) was transformed because the Shapiro-wilk test p-value is5.3389×10^{-43}. This is less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is 0. Figure k bellow shows boxplot and histogram of Minimum Degree after transformation.

Fig.k.

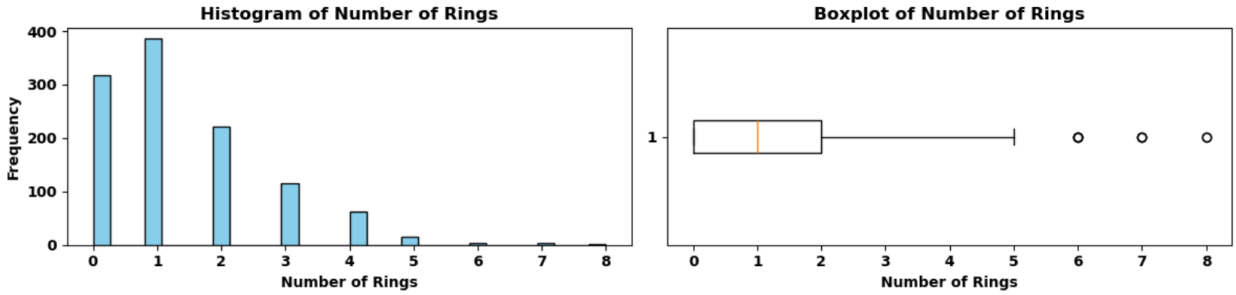


From the boxplot above, it is seen that the application of the Yeo- Johnson transformation method handled the outliers. As a result, the variable was normalized using min – max method after transformation.

8.5. Number of Rings.

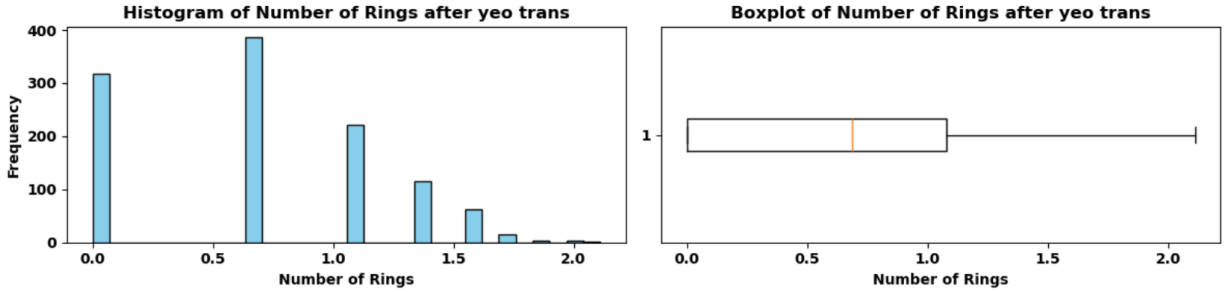
From Number of Rings, it was observed that there are 9 numbers of outliers and these made 0.80% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are 0 and 8 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.8564 and p-value is 7.7575 x 10^{-31} as shown by Shapiro-wilk test. Finaly, it was visualizes using boxplot and histogram as shown in figure l bellow:

Fig.l.



First, the variable (Number of Rings) was transformed because the Shapiro-wilk test p-value is 7.7575 x 10^{-31}. This is less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is 0. Figure m bellow shows boxplot and histogram of Minimum Degree after transformation.

Fig.m.

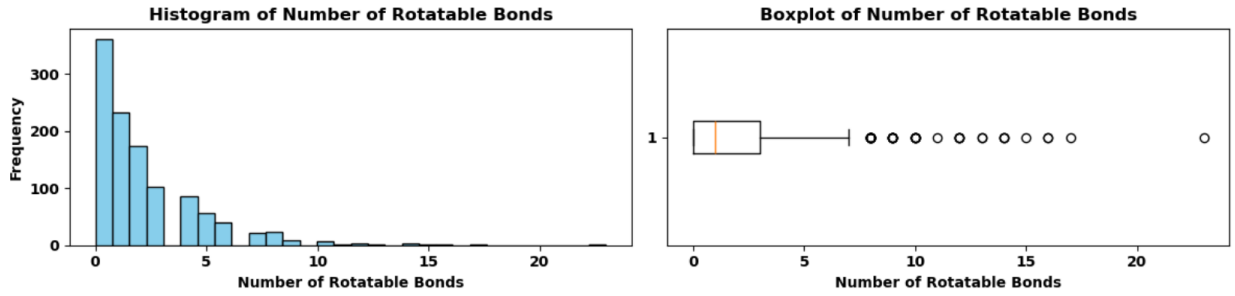


From the boxplot above, it is seen that the application of the Yeo- Johnson transformation method handled the outliers. As a result, the variable was normalized using min – max method after transformation.

8.6. Number of Rotatable Bonds

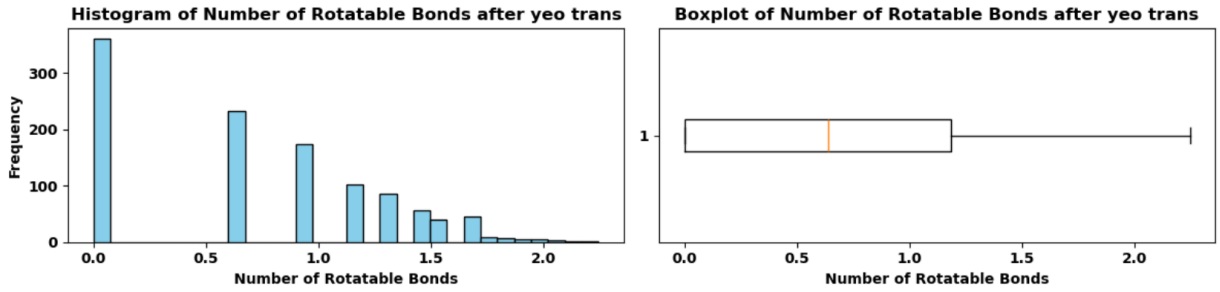
From Number of Rotatable Bonds, it was observed that there are 54 numbers of outliers and these made 4.79% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are 0 and 23 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.7774 and p-value is 1.4015×10^{-36} as shown by Shapiro-wilk test. Finaly, it was visualizes using boxplot and histogram as shown in figure n bellow:

Fig.n.



First, the variable (Number of Rotatable Bonds) was transformed because the Shapiro-wilk test p-value is1.4015×10^{-36}. This is less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is 0. Figure o bellow shows boxplot and histogram of Minimum Degree after transformation.

Fig.o.

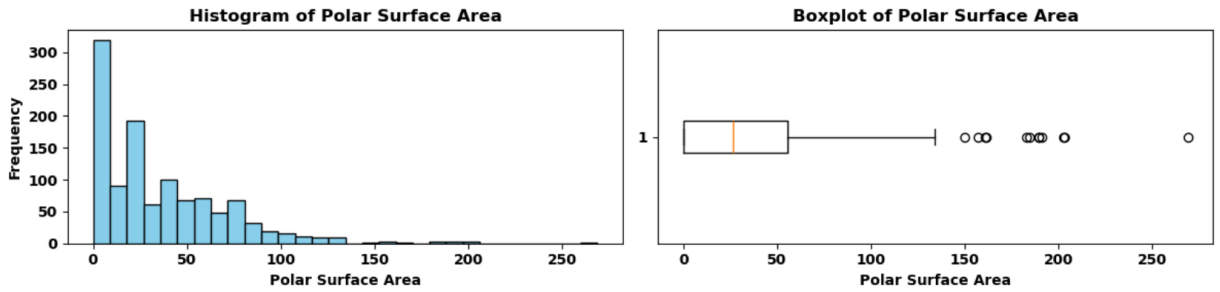


From the boxplot above, it is seen that the application of the Yeo- Johnson transformation method handled the outliers. As a result, the variable was normalized using min – max method after transformation.

8.7. Polar Surface Area.

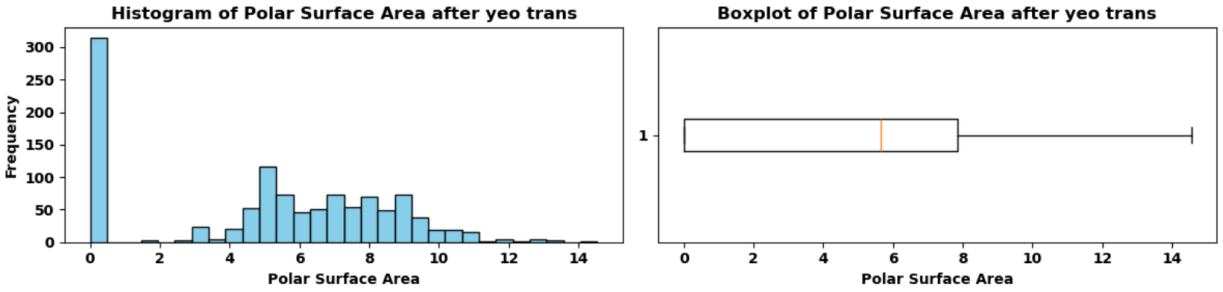
From Polar Surface Area, it was observed that there are 12 numbers of outliers and these made 1.06% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are 0 and 268.6800 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.8624 and p-value is 2.6194×10^{-30} as shown by Shapiro-wilk test. Finaly, it was visualizes using boxplot and histogram as shown in figure p bellow:

Fig.p.



First, the variable (Polar Surface Area) was transformed because the Shapiro-wilk test p-value is 2.6194×10^{-30}. This is less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is 0. Figure r bellow shows boxplot and histogram of Minimum Degree after transformation.

Fig.r.

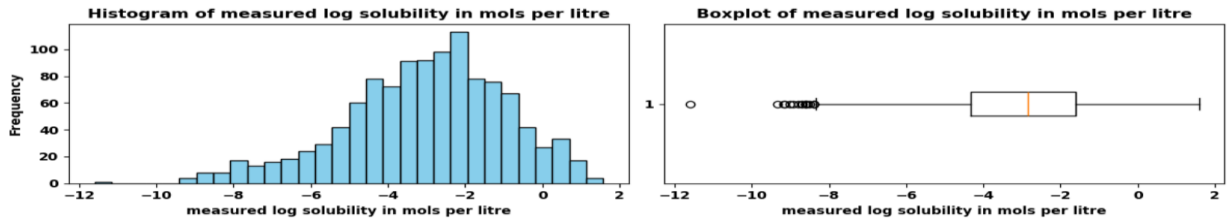


From the boxplot above, it is seen that the application of the Yeo- Johnson transformation method handled the outliers. As a result, the variable was normalized using min – max method after transformation.

8.8. measured log solubility in mols per litre .

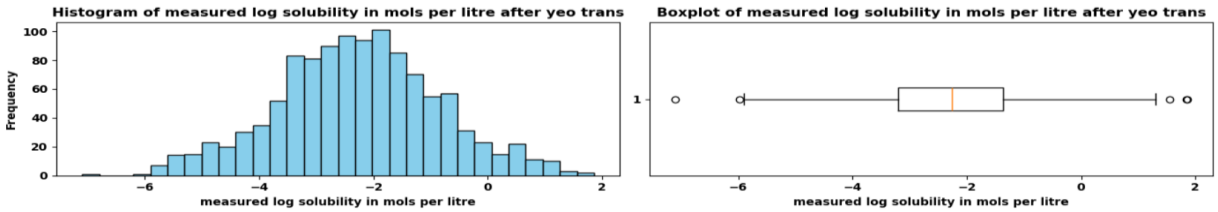
From measured log solubility in mols per litre, it was observed that there are 17 numbers of outliers and these made 1.51% of the data as shown by Interquartile Range (IQR) Method. Also, its minimum and maximum values are -11.6000 and 1.5800 respectively as shown by its statistical summary. In addition, its Test Statistic is 0.9835 and p-value is 5.1853 x 10^{-10} as shown by Shapiro-wilk test. Finaly, it was visualizes using boxplot and histogram as shown in figure s bellow:

Fig.s.



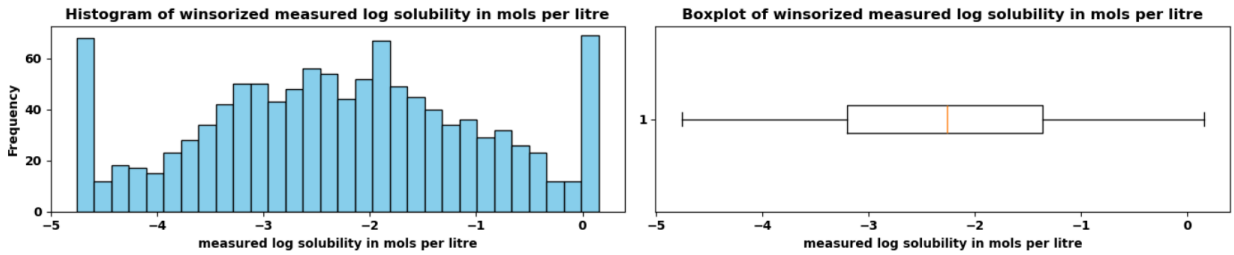
First, the variable (Polar Surface Area) was transformed because the Shapiro-wilk test p-value is. This is 5.1853 x 10^{-10} less than 0.05 which means it does not follow normal distribution and Yeo- Johnson transformation method that can handle both zeros (0) and negative values was used because its minimum value is -11.6000. Figure t bellow shows boxplot and histogram of Minimum Degree after transformation.

Fig.t.



Then the outliers were handled using winsorization method. Figure u bellow shows boxplot and histogram of after Molecular Weight winsorization.

Fig.u.



After handling the outliers, the variable was normalized using min-max method.

9. Categorical variable preprocessing.

Smiles is the only categorical variable in the dataset. This is the textual representation of the chemical structure of the compound. So, in order to convert it to numerical variable so that it can be used for linear and non-linear relationship analysis, its morgan fingerprint was Calculated, stored as bit vector and the bit vector were aggregated as single numerical value.

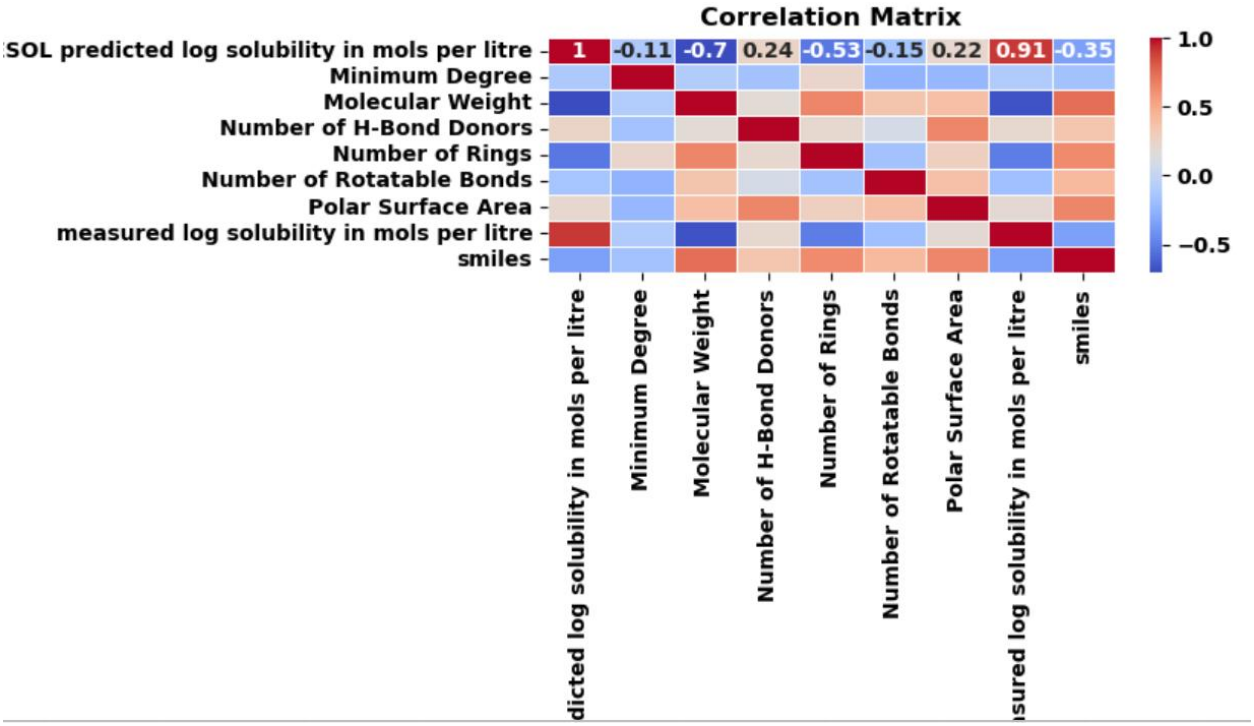
10. Exploratory data analysis.

10.1. Non-linear relationship analysis.

scatter plot and a polynomial regression were used to analyze non-linear relationship between each variable and the target variable.

10.2. linear relationship analysis.

Correlation metrix was used to analyze linear relationship among the variable and this was visualized using heatmap as shown below:



From the correlation metrix above, it can be seen that there is a strong linear relationship (0.91) between ESOL predicted log solubility in mols per litre and measured log solubility in mols per litre (the target variable). As a result, ESOL predicted log solubility in mols per litre was dropped in order to avoid collinearity effect on the model.

Also, Compound ID was dropped from the dataset because this is a unique identifier. As a result, it has no relationship with the target variable.

11. Reason for choosing Algorithms.

From analysis using Shapiro Wilk test, it was observed that the variables do not follow normal distribution.

Also, from non-linear relationship analysis using scatter plot and polynomial regression fit, it was observed that non-linear relationship exists between each variable and the target variable.

Further analysis using correlation matrix indicates that there is linear relationship among the variables.

12. Algorithms.

Based on these observations above, the following algorithms that do not assume normally distributed data and can handle both linear and non-linear relationships were selected.

12.1. Random Forest Regressor.

This is a machine learning algorithm that makes prediction by combining the decision of different decision trees. These decision trees are formed from different subset that were formed by bootstrap sampling method, that's sampling with replacement. It uses loss function as a splitting criterion.

12.2. XGBoost (Extreme Gradient Boosting).

This is a machine learning algorithm that makes decision by combining the decision of different decision trees that were built sequentially. In this case, the decision trees are made from the original dataset, but next decision tree is built to correct the error made by the previous decision tree. It uses loss function as a splitting criterion.

12.3. LightGBM (Light Gradient Boosting Machine)

This is a machine learning algorithm that makes decision the same way XGBoost makes its decision, but instead of building its trees level-wise like the XGBoost, it builds its decision trees leaf-wise. It uses loss function as a splitting criterion.

12.4. KNeighborsRegressor

This is a machine learning algorithm that makes decision using knearest neighbor. It does this by calculating the distance of every data point using different methods (e.g Euclidean method). From the calculated distance, it finds nearest neighbors based on the number of K and take the average of the target variable values of the nearest neighbor.

12.5. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANNs) make decisions through a process that involves multiple layers of interconnected nodes (or neurons) that process input data and generate an output.

13. Model Design.

The dataset was divided into three sets:

13.1. Training set.

70% of the dataset was used to train the models. This was done so that the models will have enough data points to learn underlying pattern effectively.

13.2. Cross validation.

The training set underwent cross validation using Grid search and k-fold method. This process tuned the hyperparameter of the algorithm to produce the best model's configuration, ensuring that the model generalize well to new and unseen data points.

13.3. Validation set.

15% of the dataset was used to further fine tune the hyperparameters of the models. This was done to prevent overfitting and underfitting.

13.4. Test set.

15% of the dataset was reserved and used to examine the performance of the models on new and unseen data points.

14. Results and Discussion.

1. **Random Forest Regressor with pruning.**

S/N	Root Mean Squared Error (RMSE)	Coefficient of Determination(R^2)
Training	0.41899947763455625	0.8974414846161014
Cross validation	0.5910454939421562	0.7932463808551831
Validation	0.5729180726686316	0.8292516011578004
Test	0.514990136156866	0.8310834014293869

Random Forest Regressor performs well across **training**, **cross-validation**, **validation**, and **test sets**, with good consistency between the RMSE and R^2 scores. The slight increase in RMSE from training to cross-validation, validation, and test sets is expected. This indicates that the model is not overfitting and generalizes well to new, unseen data. The consistent performance across these stages shows that the model is robust and reliable.

2. **XGBoost Regressor with pruning.**

S/N	Root Mean Squared Error (RMSE)	Coefficient of Determination(R^2)
Training	0.30255111361218734	0.9465260489717249
Cross validation	0.5870830931703321	0.7959032941693109
Validation	0.5513410622276241	0.8418707276035038
Test	0.5454146820454665	0.810535366138648

XGBoost Regressor performs very well on the training set, with a high R^2 and low RMSE, indicating a good fit. However, there is a larger gap between training, cross-validation and test RMS, suggesting some potential overfitting.

Despite this, the performance on cross-validation, validation, and test sets is still strong, with relatively consistent RMSE and R^2 scores, indicating that the model generalizes well to new data, though slightly less robustly than it performs on training data.

3. **LightGBM Regressor with pruning.**

S/N	Root Mean Squared Error (RMSE)	Coefficient of Determination(R^2)
Training	0.35881311085255324	0.9247889988804663
Cross validation	0.5862647972690272	0.7963048765918032
Validation	0.563249862590254	0.834965863860551
Test	0.5398224216668108	0.8144006956599403

LightGBM Regressor performs well on training set with a low RMSE and high R^2, indicating a good fit. However, there is a significant increase in RMSE across cross validation set, validation set and test set, suggesting a potential overfitting.

Despite this, the model still shows good performance on cross validation set, validation set and test set with relatively consistent R^2 and RMSE, indicating that it generalizes well on unseen and new data, but less robust compare to its performance on the training set.

4. **KNeighbors Regressor**.

S/N	Root Mean Squared Error (RMSE)	Coefficient of Determination(R^2)
Training	0.08379658181184067	0.9958979838174864
Cross validation	0.8097742226954208	0.6123126383233067
Validation	0.7874068282174307	0.6774701248271855
Test	0.6882104312301078	0.698340499414581

KNeighborsRegressor performs very well on training set with low RMSE and high R^2, indicating a good fit. However, a significant increase in RMSE and a significant decrease in R^2 across cross validation set, validation set and test set is a poor fit, indicating overfitting.

5**. MLPRegressor(ANN).**

S/N	Root Mean Squared Error (RMSE)	Coefficient of Determination(R^2)
Training	0.5833587251904213	0.801200101545535
Cross validation	0.6792739427651299	0.7274023971018311
Validation	0.6132140937939524	0.8043878389015562
Test	0.5454623470969796	0.8105022491858291

MLPRegressor(ANN) performed very good across training set, cross validation set, validation set and test set with a good consistency between RMSE and R^2. Slight increase in RMSR across cross validation set and validation set is expected, indicating that the model generalizes well on new and unseen data. Also, a decrease in RMSE and increase in R^2 values of the test set is a good sign.

15. Conclusion.

In summary, this research effectively tackled the prediction of drug like compounds aqueous solubility through machine learning models (Random Forest Regressor, XGBOOST, LightGBM Regressor and MLP Regressor [ANN]), achieving high Coefficients of Determination (R^2) and low RMSE scores on training, cross validation, validation and test sets. Despite acknowledging certain limitations, the results offer valuable insights for pharmaceutical companies. Furthermore, it presents opportunities for future investigation. Subsequent studies could delve into advanced algorithms, integrate more features, or gather a broader and more diverse dataset, potentially enhancing the model's accuracy and applicability

References.

Alsenz, J., & Kansy, M. (2007). High throughput solubility measurement in drug discovery and development. Advanced Drug Delivery Reviews 59 (2007), 546–567. https://www.sciencedirect.com/science/article/pii/S0169409X07000786

Dutta, A., & Karmakar, R. (2021). Estimating aqueous solubility directly from molecular structure using machine learning approach. In *2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (pp. 467-473). IEEE. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9640774

Francoeur, P. G., & Koes, D. R. (2021). SolTranNet–A machine learning tool for fast aqueous solubility prediction. *Journal of Chemical Information and Modeling, 61*(6), 2530-2536. https://pubs.acs.org/doi/epdf/10.1021/acs.jcim.1c00331

Ghasemi, J., & Saaidpour, S. (2007). QSPR prediction of aqueous solubility of drug-like organic compounds. Chemical and Pharmaceutical Bulletin, 55(4) 669—674, https://www.jstage.jst.go.jp/article/cpb/55/4/55_4_669/_pdf/-char/ja

Hossain, S., Kabedev, A., Parrow, A., Bergström, C. A. S., & Larsson, P. (2019). Molecular simulation as a computational pharmaceutics tool to predict drug solubility, solubilization processes, and partitioning. European Journal of Pharmaceutics and Biopharmaceutics, *137(2019)*, 46-56. https://www.sciencedirect.com/science/article/pii/S0939641118313675?ref=pdf_download&fr=RR-2&rr=8ae8ea07ed89d180

Lovrić, M., Pavlović, K., Žuvela, P., Spataru, A., Lucić, B., Kern, R., & Wong, M. W. (2021). *Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? Journal of Chemometrics, 35*(e3349). https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/epdf/10.1002/cem.3349

Palmer, D. S., O'Boyle, N. M., Glen, R. C., & Mitchell, J. B. O. (2007*). Random Forest Models to Predict Aqueous Solubility.* Journal of Chemical Information and Modeling, 47(1), 150-158. https://pubs.acs.org/doi/epdf/10.1021/ci060164k

Su, M., & Herrero, E. (2023). Creation and interpretation of machine learning models for aqueous solubility prediction. *Exploratory Drug Science*, *1*, 388–404. https://www.explorationpub.com/Journals/eds/Article/100826