Report on Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability?

Authors : Mario Lovric1,2 | Kristina Pavlovic1 | Petar Žuvela3 | Adrian Spataru1 |
Bono Lucˇic2 | Roman Kern1,4 | Ming Wah Wong3

This is a report of the work done on this research paper , the understanding and how we worked to reproduce the results of the original paper .

ABSTRACT :

This paper is about using machine learning algorithms to predict the intrinsic aqueous solubility of 829-drug-like compounds based on their chemical structures. The authors compared the performance of four different machine learning algorithms and found that the LASSO and random forests performed the best , LASSO algorithm had the best predictive ability. A consensus model combining the two best learners showed the best predictive ability and generalization. A ranking score considering generalization, number of features, and test performance was proposed for model selection.The results of this study could be useful in drug discovery and development by helping researchers prioritize compounds with better solubility early on in the process.

INTRODUCTION :

Solubility refers to the ability of a substance (the solute) to dissolve in a solvent to form a homogeneous solution. The solubility of a substance is typically expressed as the maximum amount of solute that can dissolve in a given amount of solvent at a specified temperature and pressure.The significance of predicting intrinsic aqueous solubility of drug-like compounds is that it can help in the early stages of drug discovery and development. Poor solubility is a common issue in drug development, and it can lead to low bioavailability, poor pharmacokinetics, and other problems. By predicting solubility early on, researchers can prioritize compounds with better solubility and potentially save time and resources in the drug development process.

According to this paper, intrinsic aqueous solubility refers to the solubility of a compound in water under standard conditions, without any other solutes present. The authors used a dataset of intrinsic aqueous solubility values for drug-like compounds to train and test their machine learning models.
Quantitative structure-property relationships (QSPRs) are mathematical models that relate the structural features of a molecule to its physical or chemical properties, such as solubility, melting point, or toxicity. QSPRs can be used to predict the properties of new molecules based on their

structural features, which can be useful in drug discovery, materials science, and other fields where the properties of molecules are important. Structural features are given priority in QSPRs because the properties of a molecule are largely determined by its molecular structure.

DATA COLLECTION :

The authors collected data from sources that included pH measurements between 22.5 and 25 degrees Celsius and used inert gases (argon, nitrogen) in their measurements. After collecting the data, the authors preprocessed the data by calculating and considering two types of predictive features: fingerprints (FPs) and molecular descriptors (DPs). The authors chose FPs with a comparatively short radius of 3 bonds and large vector length of 5120 bits, to avoid bit collision .From the available 5000 DRAGON molecular DPs, only a few groups of DPs were selected based on chemical intuition, specifically, constitutional, ring, topological DPs, functional group counts, and molecular properties. All DPs with missing values were removed. Such a preselection procedure yielded a total of 317 molecular DPs. A combination of FPs and DPs (FPDS) was also evaluated (5444 features in total).

Filtration of the compounds criteria followed:

- LogP55 in [3.6, 7.5],
- Molecular weight larger than 88 g/mol,
- Structures with more than six heavy atoms.
- The obtained logSw values in the extracted data were converted to logS0 based on their formal charges.

EVALUATED MACHINE LEARNING METHODS:

The four machine learning methods evaluated in this study are:

1. Partial Least Squares (PLS) 2. Random Forest (RF) 3. Light Gradient Boosting Machine (LGBM) 4. Least Absolute Shrinkage and Selection Operator (LASSO) .The authors used these algorithms to build models that predict the intrinsic aqueous solubility of drug-like compounds based on their chemical structural information. The models were trained and tested on a dataset of 829 compounds, and their performance was evaluated using various metrics such as root mean squared error (RMSE) and coefficient of determination (R^2).

MY AREA OF WORK:

I worked on the algorithm LGBM , stands for **Light Gradient Boosting Machine**. Lets discuss how this algorithm worked , and its limitations .

- It is a gradient boosting algorithm that uses decision trees as a base algorithm. LGBM uses the first-order derivative information when optimizing the loss function, and its leaf growth strategy with depth limitation and multithread optimization contributes to solving

the excessive memory consumption with respect to other boosting-ensemble machine learning methods. This means that LGBM is designed to grow the decision tree in a way that limits its depth and uses multiple threads to optimize the process. By doing so, ***LGBM can reduce the amount of memory required to train the model and improve its computational efficiency. This is particularly important when dealing with large datasets or complex models that require a lot of memory to train.***

The authors found that ***LGBM had a complex hyperparameter space***, which was hard to optimize and was working in the overfitting regime in most cases. Despite these challenges, LGBM still produced results that were compared to the other algorithms used in the study. ***The results showed that LGBM had lower predictive ability on the test set compared to the other algorithms, such as LASSO and RF.***

LIMITATION IN LGBM : LGBM failed, the authors found that LGBM had a complex hyperparameter space, which was hard to optimize and was working in the overfitting regime in most cases. This means that ***LGBM was not able to generalize well to new data and was instead fitting too closely to the training data, resulting in poor performance on the test set.*** Additionally, the authors noted that LGBM showed a notably larger spread compared with other algorithms, which can be explained by evident overfitting on the train set and lower predictive ability on the test set.

HOW LGBM CAN BE IMPROVED : To improve LGBM's performance, the authors suggested ***exploring different hyperparameter settings and regularization techniques to prevent overfitting***. Additionally, the authors suggested using ***LGBM in combination with other machine learning algorithms to create an ensemble model that can improve predictive performance.*** Finally, the authors suggested using LGBM on larger datasets to further explore its potential for predicting solubility.

FEATURE SELECTION AND HYPERPARAMETER OPTIMIZATION:

Feature selection and hyperparameter optimization, these are two important steps in the machine learning process that can significantly impact the performance of the model.

FEATURE SELECTION:

Feature selection is the ***process of selecting a subset of relevant features from a larger set of features to improve the model's performance.*** In this work, the authors used a multistage post hoc feature selection strategy based on permutation importance to eliminate features. The method permutes the values of individual features one-by-one to assess the relevance of the features with respect to the response vector (logS0). By doing so, the authors were able to ***identify the most important features for predicting solubility and eliminate irrelevant or redundant features that could negatively impact the model's performance.***

HYPERPARAMETER OPTIMIZATION:

It is the ***process of selecting the best hyperparameters for a given machine learning algorithm to optimize its performance.*** Hyperparameters are parameters that are not learned during the training process but are set before training and can significantly impact the model's performance. In this work, the authors used Bayesian optimization (BO) to optimize the hyperparameters for each machine learning algorithm. BO aims to construct a posterior distribution of functions (Gaussian process) that best describes the loss function. By optimizing the hyperparameters, the authors were able to improve the performance of the machine learning algorithms and achieve better predictive ability on the test set.

MODEL TRAINING :

It is the process of using a dataset to train a machine learning algorithm to make predictions on new data. In this work, the authors split the datasets (logS0 and the predictive sets) following ***two strategies: randomly and by means of diversity picking.***

For both splits, the **external test set was set to 20% of the whole data ,**

Out of the remaining 80%(considered as 100%)

Training set (80%) **Validation set(20%)**

The authors applied a sequential set of steps to preprocess the data before training the machine learning models.
The steps involved in the preprocessing method are as follows:
1. Removing features with any missing values
2. Removing correlated features (Pearson correlation > 0.85)
3. Separating categorical features (from binary and continuous) and converting them to binary features (based on binning to four "dummy" bins)
4. Removing low variance binary features (lower than 1% variance)

By applying these preprocessing steps, the authors were able to improve the quality of the data and reduce the noise in the dataset, which in turn improved the performance of the machine learning models.

MODEL SELECTION PROCESS:

The authors used a ranking schema to objectively evaluate the performance of the machine learning models.The ranking schema was based on Equation(2).

$$RkM = 0.5RRMSE(test) + 0.3Rfeatures + 0.1R\Delta val + 0.1R\Delta train \rightarrow (2)$$

In this equation, RRMSE(test) is the rank of the root mean square error (RMSE) of the respective test set, Rfeatures is the rank based on the total number of features involved in the model, and Δval and Δtrain are defined with Equations 3 and 4, respectively. Both terms account for the generalizability of the models.

$$\Delta val = | RMSE(test) - RMSE(val)| \rightarrow (3)$$
$$\Delta train = | RMSE(train) - RMSE(val)| \rightarrow (4)$$

Equations 3 and 4, they are used to calculate the generalizability of the machine learning models. Equation 3 defines Δval, which is the absolute difference between the RMSE of the test set and the RMSE of the validation set, normalized by the absolute value of the difference. Equation 4 defines Δtrain, which is the absolute difference between the RMSE of the training set and the RMSE of the validation set, normalized by the absolute value of the difference. Both terms account for the generalizability of the models. By using these equations, the authors were able to objectively evaluate the performance of the machine learning models and identify the best-performing models for predicting intrinsic aqueous solubility of drug-like compounds.

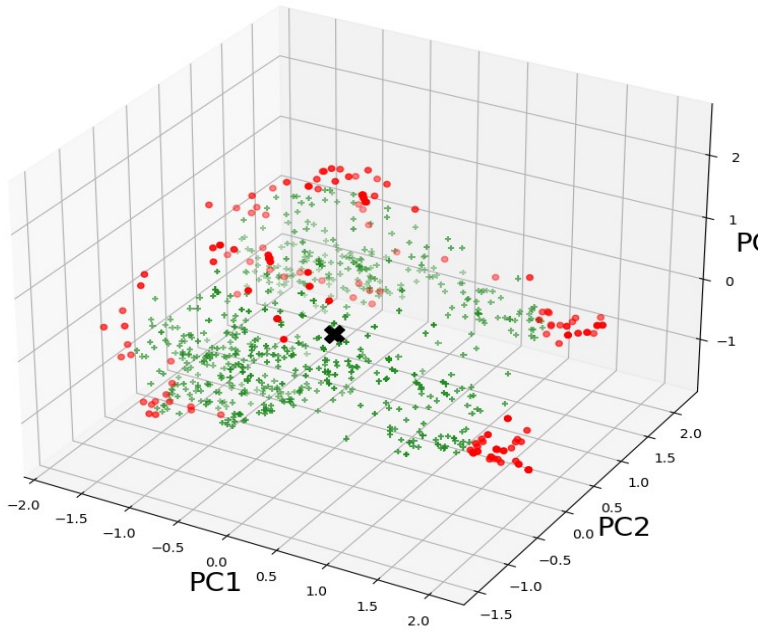
PCA-split :

The study used Principal Component Analysis (PCA) to create a challenging test scenario for machine learning models predicting solubility. Here are the key steps: PCA Transformation: Fingerprint (FP) data was transformed into three principal components (LVs) using PCA.

1. Data Preprocessing: Low-variance FPs were removed before PCA.
2. Centroid and Distances: The centroid of the PCA space was calculated, and distances of compounds from the centroid were determined.
3. Train-Test Split: Compounds were split based on their distances; those below the 80th percentile formed the train set (PCA-train), and those above it formed the test set (PCA-test). The validation set was excluded.
4. Model Evaluation: The two winning models (LASSO and RF) were retrained on PCA-train and tested on PCA-test.

Results showed that the RF model performed better in this challenging scenario. The choice of the RF model was based on a comprehensive quality estimation metric (RkM). This emphasized the importance of testing models with extreme scenarios to assess their

generalization capabilities and using quality estimation methods in model selection. The study also suggested that ensemble models can enhance predictions in cases of descriptor redundancy.
This method helps assess model performance and generalization .



MODEL OPTIMIZATION RESULTS :

The models were trained using different feature sets (FP, DS, and FPDS), splitting strategies (random and diversity picking), and with or without feature preprocessing .

- The results showed that LASSO had the best score by root mean square error (RMSE) on the test set, but it had a high number of features .
- RF models exhibited overfitting but to a lesser extent than LGBM, and they showed a smaller spread in performance .
- The study also compared the performance of the models on different data-splitting strategies, with diversity picking leading to overly optimistic results and lower generalization robustness .
- The best models were chosen based on a ranking schema that considered test performance, complexity (number of features), and generalization .
- A consensus model was built, outperforming all other models, with a RMSE of 0.67 log points and an R2 of 0.81 .

RESULTS AND CONCLUSION :

The study compared four machine learning algorithms (random forests, LightGBM, partial least squares, and LASSO) for predicting the intrinsic aqueous solubility of drug-like compounds.

- LASSO yielded the best predictive ability on an external test set, while an RF model achieved a good balance between complexity and predictive ability with a smaller number of features.
- The RF model also showed better generalization on a more aggressive test set, indicating its robustness .

The authors proposed a ranking score for model selection, considering generalization, number of features, and test performance. A consensus model was built, which exhibited the best predictive ability and generalization .

- ❖ The paper also discussed the physical interpretation of the top five important molecular descriptors for aqueous solubility, providing insights into the relationship between molecular parameters and solubility.

SCBO (sum of conventional bond orders)
DDtr 06 (descriptor for cyclic character)
AMR (molecular refraction)
MLOGP (octanol-water partition coefficient)
TPSA (total polar surface area).

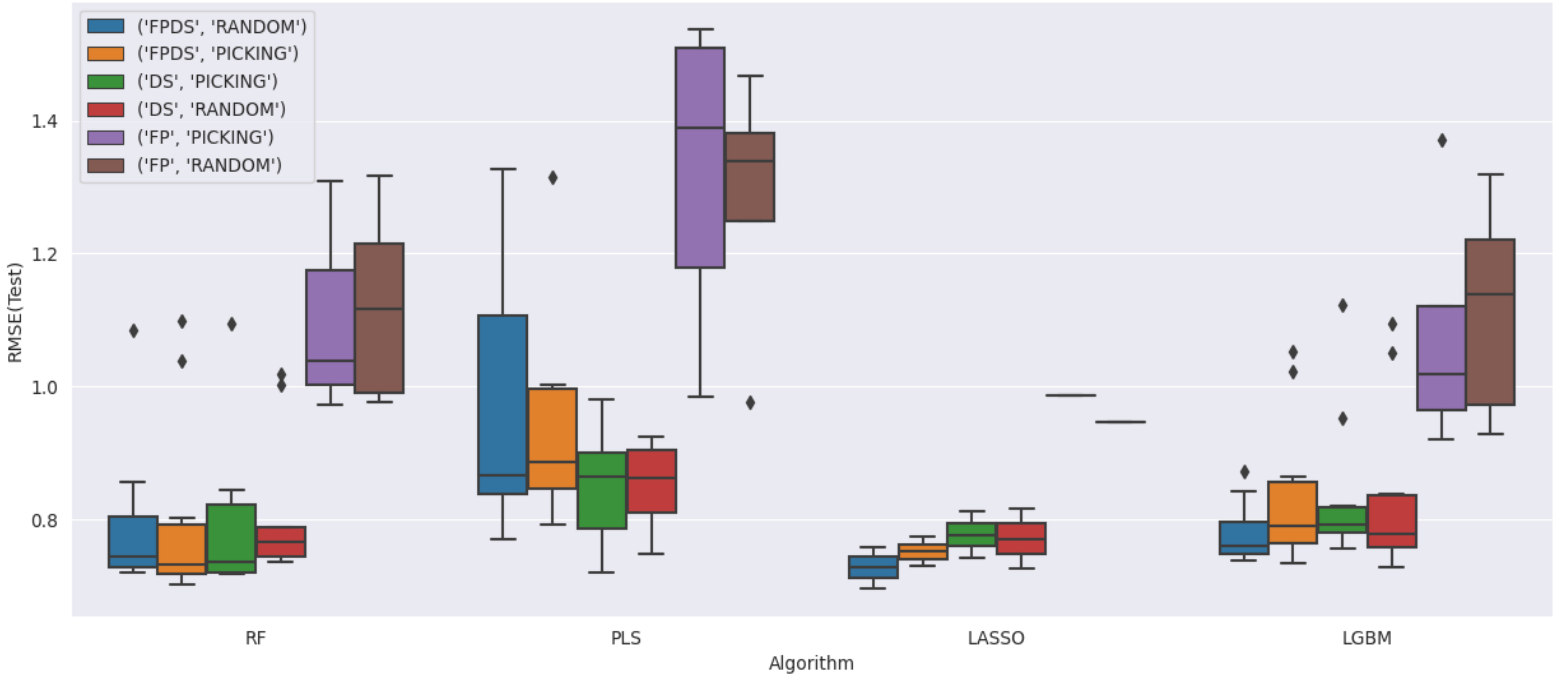
- WE HAVE WORKED ON THE CODING PART OF THE PROJECT IN THE PYTHON LANGUAGE IN JUPYTER NOTEBOOK . WE WERE ABLE TO GET THE RESULTS CLOSE TO THE RESULTS MENTIONED IN THE PAPER :

FOR LASSO:
{'Train': 0.6625462197225582, 'Validation': 0.957961151629105, 'Test': 0.6963947677484571}.

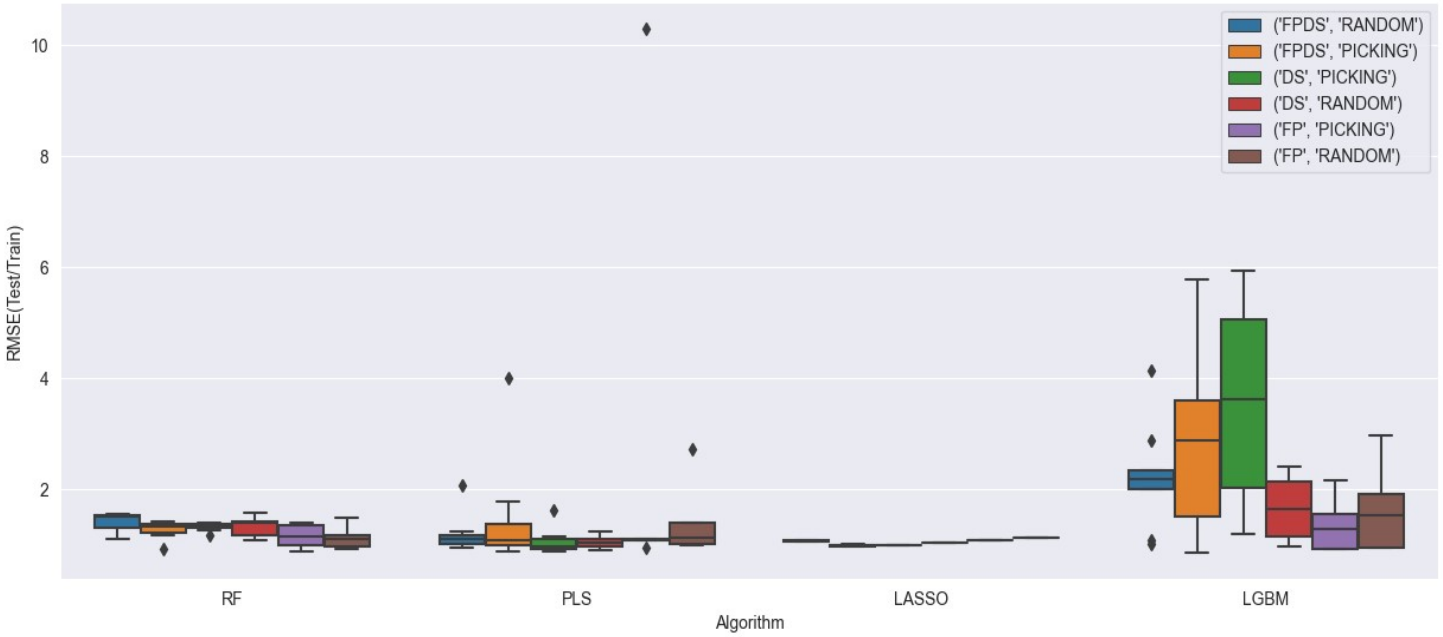
FOR RF:
{'Train': 0.47263630176807325, 'Validation': 0.9390128644625835, 'Test': 0.7212356018397095}

FROM THE ABOVE RESULTS , WE SUCCESSFULLY DERIVED THE RESULTS AND STUIDED THAT

- **LASSO showed the best predictive ability on an external test set**, while
- **RF achieved a good balance between complexity and predictive ability.**



- The provided image illustrates the distribution of testing errors for four machine learning algorithms (PLS, LASSO, LGBM, and RF) under two different training/test set partition methods: random split (depicted with green ascending lines) and diversity picking (depicted with red descending lines).
- This figure is valuable for understanding how these machine learning algorithms perform and how their performance varies depending on the choice of training/test set partition method. It provides insights into the reliability and consistency of these algorithms in different partition scenarios.



- The figure provided offers insights into the generalization ability and robustness of machine learning models developed in this study. It presents the RMSE(test) / RMSE(train) ratio for models grouped by the method used (RF, PLS, LASSO, LGBM) and three sets of predictive variables.
- This figure is valuable for understanding how the machine learning models perform in terms of generalization and robustness. It allows for a comparison of different model development methods and the impact of predictive variables on model performance under different partition scenarios.

- Prepared by
O.Pravallika
(Intern)