Table of Contents

Introduction ***4***

Results and discussion ***7***

Dataset and descriptors **7**

Dataset pre-processing **7**

Data analysis **8**

Model selection **10**

Feature reduction and model evaluation **11**

Hyperparameter tuning **13**

Predictions **15**

Explainability **16**

Conclusion ***17***

Data availability ***18***

References ***19***

A Machine Learning Approach for Predicting Aqueous Solubility of Organic Molecules

Code to Discovery*

Email: codetodiscovery@gmail.com

Youtube: https://www.youtube.com/channel/UC6k1epfxt3gPVdQhE8gI1tQ

Saturday, 9 March 2024

A Machine Learning Approach for Predicting Aqueous Solubility of Organic Molecules

Abstract:

Understanding and predicting solubility is crucial across various scientific disciplines, influencing drug development, environmental risk assessments, and materials engineering. This study explores the application of machine learning (ML) models to predict the aqueous solubility of organic molecules. The dataset, comprising 1144 molecules, undergoes thorough pre-processing, feature reduction, and analysis. Machine learning models, including Random Forest (RF) and Extra Tree (ET), are evaluated, and their performances are compared. The study emphasizes the importance of interpretability and explains the impact of top descriptors on solubility predictions. Hyperparameter tuning and explainability techniques contribute to optimizing model performance and enhancing transparency. The results showcase the effectiveness of ML in predicting solubilities while addressing challenges related to complexity and interpretability.

Keywords: solubility; machine learning; random forest; Mordred

Introduction

Solubility, the ability of a substance to dissolve in a solvent, stands as a fundamental property governing a myriad of processes across diverse scientific disciplines. Understanding and predicting solubility is paramount, as it underpins crucial aspects in fields ranging from pharmaceuticals and environmental science to materials engineering. The capacity of organic molecules to dissolve influences their behavior, reactivity, and efficacy, making solubility a central parameter in the design and optimization of chemical processes.[1]

In the pharmaceutical industry, solubility plays a pivotal role in drug development and formulation. The efficacy of a drug is intricately linked to its solubility, as it directly affects bioavailability. Poorly soluble drugs often face challenges in absorption, limiting their therapeutic effectiveness. Consequently, the ability to predict and enhance solubility is a critical step in optimizing drug formulations and improving patient outcomes. In environmental science, solubility governs the fate and transport of chemicals in air, water, and soil. Understanding how organic molecules interact with different environmental matrices is essential for assessing the potential impact of pollutants, facilitating risk assessments, and guiding remediation strategies. The realm of materials engineering also relies heavily on solubility predictions. The solubility of organic molecules in various solvents is a key determinant in the development of coatings, adhesives, and other materials. Tailoring the solubility of components allows for precise control over the properties of the final product, enabling innovations in fields such as flexible electronics, biomaterials, and coatings technology.[2, 3]

In this context, machine learning (ML) models have emerged as powerful tools for predicting solubilities with high accuracy. The ability to predict solubilities computationally not only accelerates the drug discovery process but also facilitates the design of environmentally friendly chemicals and materials.

Numerous computational models capable of forecasting a molecule's aqueous solubility have been documented in the literature.[4] Recently, two distinct data preparation approaches, namely descriptor-based[5-7] and group contribution,[8-10] have been applied in various modelling methods to assess solubility measures. Descriptor-based models rely on parameters associated with physical properties, such as molecular topological indices. In contrast, group contribution methods establish a correlation between water solubility and various functional groups by breaking down molecular units into subunits and summing the estimated solubility of each subunit. Table 1 provides a comparison of noteworthy models developed and their respective performances.

While past studies have demonstrated the accessibility of predicting aqueous solubility, newcomers to the field may encounter challenges in grasping these algorithms due to their intricate physicochemical characteristics. Additionally, prevalent concerns in current research pertain to the validity of correlations, given their sensitivity to variations in the calibration conditions predefined in advance. Furthermore, the impact of chemical representations and their role in the performance of ML methods have not been thoroughly explored.

Table 1.Previous approaches utilizing ML approaches for predicting solubilities.

Entry	**Dataset Size**	**ML Model**	**Test R^2 score**	**Developer**	**Ref.**
1.	1297	Multiple linear regression (MLR)	0.88	Huuskonen	[11]
2.	1293	MLR	0.82	Yan	[12]
3.	2847	MLR	0.71	Delaney	[13]
4.	1294	MLR	0.90	Hou	[3]
5.	1290	MLR	0.73	Ali	[14]
6.	1290	Ensemble of ANN, RF, and XGB	0.93	Sorkun	[15]
7.	4376	MLR	0.89	Le	[16]

This manuscript presents a comprehensive exploration of the application of ML models in predicting the solubilities of organic molecules. By harnessing

the capabilities of ML, we aim to contribute to the advancement of research in fields where solubility is a critical parameter, fostering innovation and progress in pharmaceuticals, environmental science, and materials engineering. The general ML workflow used in this study has been depicted in Fig. 1.

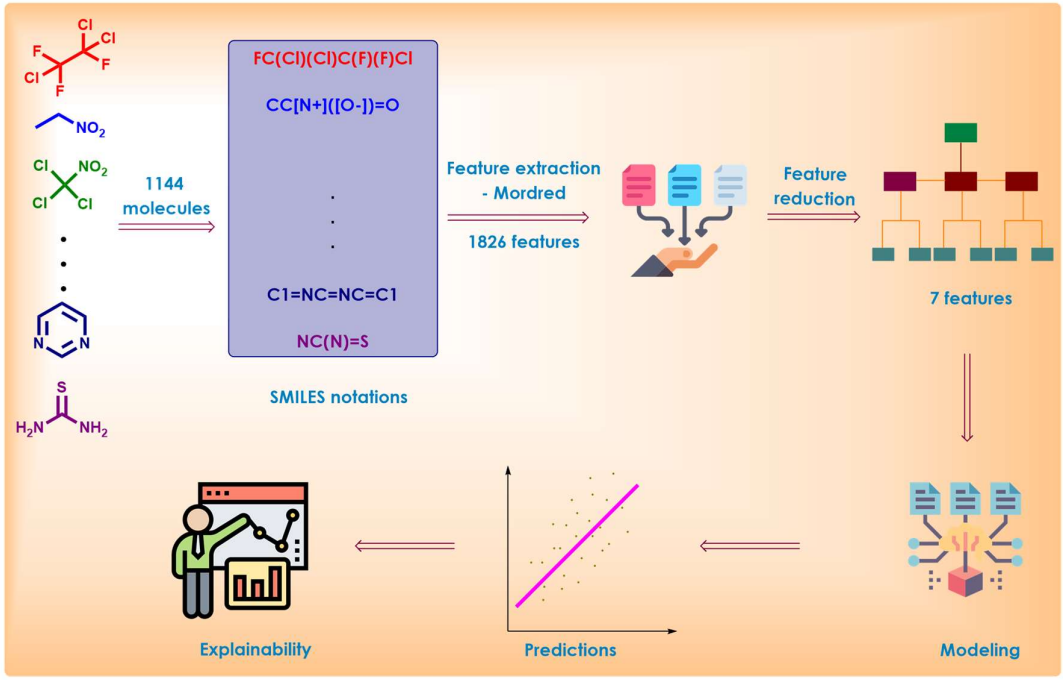

Figure 1. General machine learning workflow used in this study.

Results and discussion

Dataset and descriptors

The dataset utilized in this study comprises SMILES strings and experimentally determined logarithms of solubilities in mol/l for a total of 1144 organic molecules, sourced from reputable literature.[13] Descriptor computation was conducted employing an open-source Mordred script developed in Python.[17] This script facilitates the rapid calculation of 1826 topological, 1D, 2D, and 3D descriptors within a few minutes, even for datasets of the scale used in this investigation. Notably, Mordred exhibits a remarkable capability to compute descriptors for heavy and intricate molecules, a task typically considered challenging, and does so with superior speed compared to other established software tools. The manifold advantages offered by Mordred position it as a promising cheminformatics tool for diverse quantitative structure–activity relationship studies. The expeditious and effective application of Mordred resulted in the generation of 1826 molecular descriptors for the entire set of 1144 organic molecules.

Dataset pre-processing

The comprehensive dataset, encompassing 1826 descriptors for 1144 molecules, underwent thorough assessment for the presence of missing or erroneous values. Subsequently, the removal of features containing such values resulted in a refined set of descriptors, reducing their number to 1273. Further refinement involved the elimination of 153 features exhibiting constant values, resulting in a final count of 1120 numeric descriptors. To enhance model simplicity and interpretability, a correlation-based feature reduction process was implemented, with the objective of improving training efficiency and minimizing computational time. The Pearson correlation coefficient (r) served as the metric to identify redundancy, leading to the removal of features with $r \geq 0.80$ from the dataset,

ultimately yielding 281 descriptors. Within this subset, 279 descriptors were numerical, while 2 were boolean descriptors denoting true and false values. Consequently, label encoding was applied to convert these categorical descriptors into numerical equivalents, resulting in a dataset comprising 281 numerical descriptors and 1144 data points. Standardization was then applied to ensure normal distribution across all descriptors.

Data analysis

The logarithmic solubility measurements exhibit a range from –11.6 for decachlorobiphenyl to 1.58 for acetamide, indicating a diverse spectrum of solubility values within the dataset. The distribution plot of the target values displays a slight right skewness, suggesting an asymmetry in the distribution (Fig. 2A). The mean value of the distribution is –3.06, reflecting a central tendency slightly lower than the median value of –2.87.

To unravel the key features influencing the measured solubilities of organic compounds, a systematic approach was employed. Initially, the correlation of each descriptor with the target solubility values was calculated, allowing for the identification of four descriptors highly correlated with the solubility. The correlation matrix, as illustrated in Fig. 2B, showcases these influential descriptors, namely FilterItLogS, PEOE_VSA6, RNCG, and ABC. This step was crucial in pinpointing the descriptors most strongly associated with the solubility variations observed in the dataset.

The subsequent analysis delved deeper into the relationships between these identified descriptors and the measured solubilities through scatter plots, depicted in Fig. 2C. This visual exploration aids in understanding the nature and direction of correlations, providing valuable insights into how specific molecular

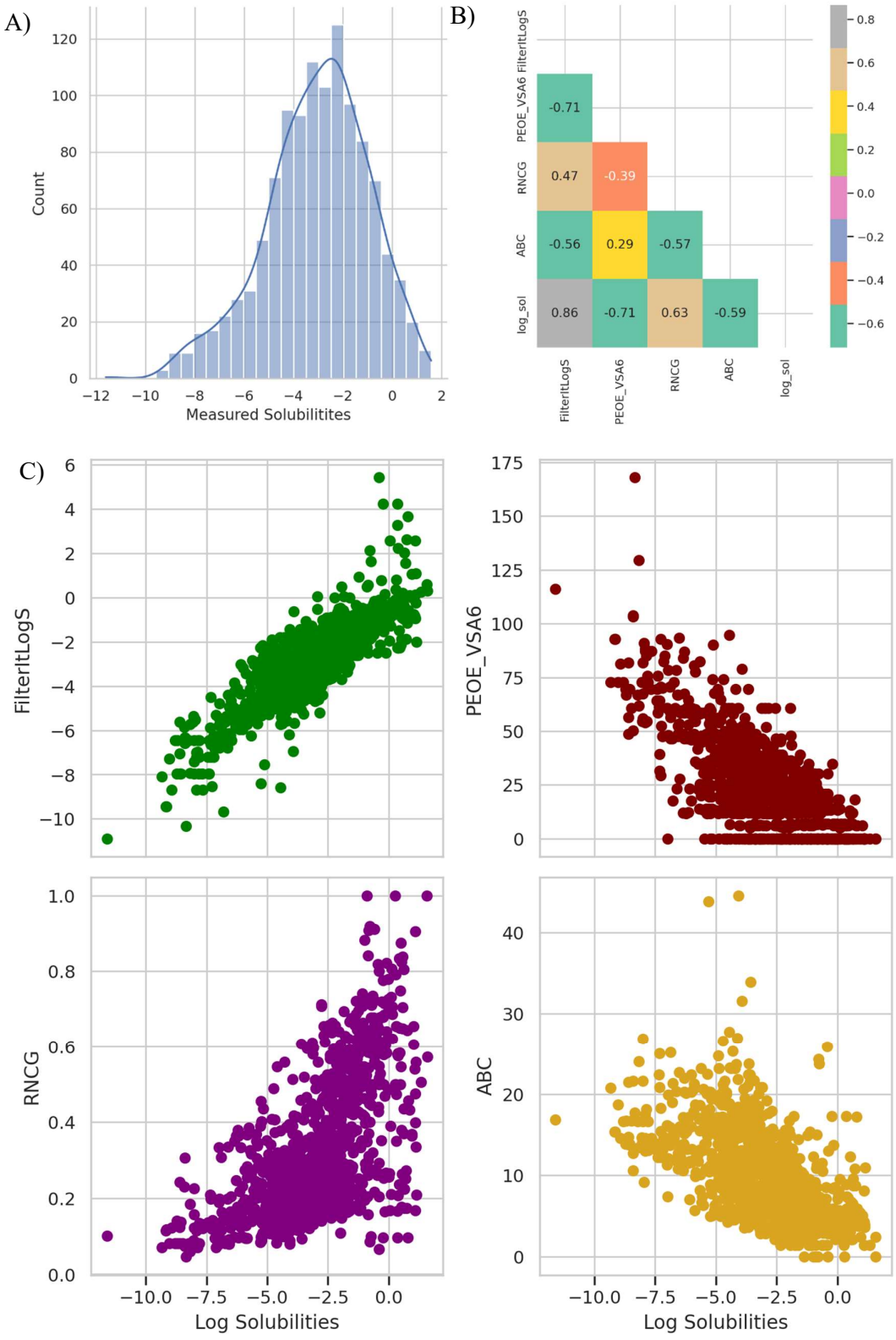

Figure 2. A) Distribution plot of measure log of solubilities. B) Correlation matrix of top 4 correlated descriptors. C) Scatter plot of all the 4 top correlated descriptors with measured log of solubilities (Log Solubilities).

characteristics contribute to the observed solubility trends.Within the array of descriptors employed in this comprehensive study, FilterItLogS emerged as the most influential, showcasing a robust correlation with a r value of 0.86. In close pursuit, PEOE_VSA6 and RNCG demonstrated significant correlations, with respective r values of -0.71 and 0.63. Notably, ABC exhibited the least correlation among the quartet of descriptors, registering an r value of -0.59.

Delving deeper into the relationships revealed by scatter plot analysis, a positive correlation surfaced between FilterItLogS and RNCG, indicating a tendency for these descriptors to vary in tandem. Conversely, PEOE_VSA6 and ABC exhibited a negative correlation, suggesting an inverse relationship between these descriptors. This nuanced exploration provides a more intricate understanding of the interplay among the descriptors, shedding light on their respective contributions to the solubility patterns observed in the organic molecules under scrutiny.

Model selection

To initiate the model screening process, the dataset underwent a random split, with 80% allocated to training and 20% to the test set. Model performance was assessed using the coefficient of determination (R^2) providing insights into accuracy and predictive power.[18] A comparative analysis of various linear and non-linear ML algorithms, including multiple linear regression (MLR), support vector machine (SVM), random forest (RF), and extra tree (ET), was conducted through cross-validation.[19]

The cross-validation technique involved randomly dividing the training data into k folds, training the model on k-1 folds, and validating on the remaining fold. The average metric across k runs was utilized to ascertain model accuracy. In this study, a 5-fold cross-validation approach was adopted, deemed sufficient for the dataset's size. The resulting model was then evaluated on the test set to

gauge its predictive capabilities, with the summarized outcomes presented in Table 2.

While classical linear models like MLR offer simplicity and interpretability, they may not capture non-linear dependencies between properties and structural features. In this study, MLR demonstrated average performance in predicting the solubility of organic molecules, as reflected in the test R^2 values.[20] In contrast, cluster-based regression techniques, such as SVM, exhibited an incremental increase in predictive power, evident in R^2 scores of 0.839, 0.938, and 0.830 for cross-validation, training, and test sets, respectively.

The utilization of tree-based models, such as RF and ET, capable of identifying complex patterns in the data, yielded improved accuracy. Notably, a substantial increase in test R^2 scores was observed, reaching 0.868 for RF and 0.895 for ET. RF, renowned for its ability to combine multiple decision trees, proved advantageous in minimizing bias and variance, thereby enhancing overall model accuracy.

Table 2. Cross-validation metrics of different models screened in this study.

Entry	**Model**	**Cross-val R^2 score**	**Train R^2 score**	**Test R^2 score**
1.	MLR	-8.8	0.953	0.624
2.	SVM	0.839	0.938	0.830
3.	RF	0.886	0.984	0.868
4.	ET	0.905	1.0	0.895

Feature reduction and model evaluation

While the RF and ET models, trained on 915 molecules with 281 descriptors, demonstrated remarkable accuracy in predicting log solubilities for organic molecules in the external test set, the extensive use of descriptors in this study

results in a "black box" quality, hindering interpretability.[21-24] In a pursuit to simplify and elucidate the Quantitative Structure-Property Relationship (QSPR) study and identify the most influential features shaping the model, a feature importance algorithm based on RF was employed. This algorithm quantifies the importance of each feature used in model training.

Eight key features emerged as the most relevant for model training, as identified by the mentioned algorithm: FilterItLogS, Lipinski, SIC0, RNCG, RPCG, ATS0Z, AATS0i, and AETA_eta (Fig. 3A). To assess the performance of RF model with a reduced number of features, it was trained using only these top 8 descriptors. This was achieved by incrementally adding descriptors, and the corresponding metrics were recorded for comparative analysis. In Fig. 3B, the graph illustrates the train and test R^2 scores with the sequential addition of descriptors for the RF model.

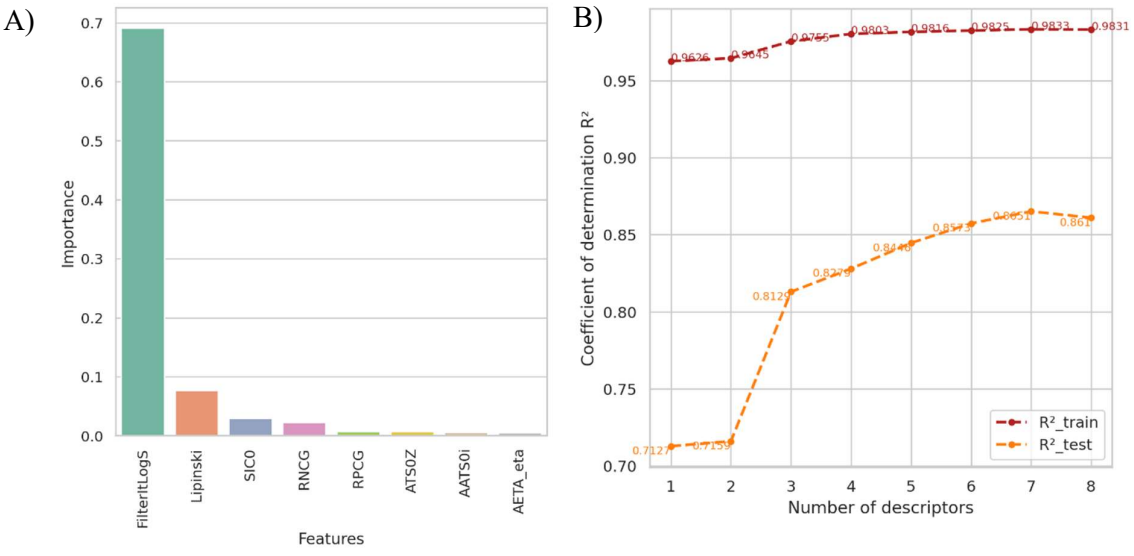

Figure 3. A) Feature importance chart. B) Line chart showing increment in R2 values with sequential addition of descriptors; No descriptors: 1 (FilterItLogS), 2 (FilterItLogS, Lipinski), 3 (FilterItLogS, Lipinski, SIC0) ,4 (FilterItLogS, Lipinski, SIC0, RNCG), 5 (FilterItLogS, Lipinski, SIC0, RNCG, RPCG), 6 (FilterItLogS, Lipinski, SIC0, RNCG, RPCG, ATS0Z), 7 (FilterItLogS, Lipinski, SIC0, RNCG, RPCG, ATS0Z, AATS0i), 8 (FilterItLogS, Lipinski, SIC0, RNCG, RPCG, ATS0Z, AATS0i, AETA_eta).

Notably, utilizing a single most contributing feature, FilterItLogS, yielded a high train R^2 score but a mediocre test R^2 score, indicative of an overfitting scenario. The sequential addition of Lipinski had a marginal impact on the scores, but the inclusion of SIC0 led to a substantial enhancement in the test scores. Subsequent addition of descriptors such as RNCG, RPCG, ATS0Z, and AATS0i contributed to a gradual increment in the test score until it plateaued with the addition of the last descriptor, AATS0i. An observed slight decrease in the test R^2 score was noted with the addition of the 8^{th} descriptor, AETA_eta.

These results highlight the optimal nature of the top 7 descriptors for achieving the best model performance, as any further addition of descriptors resulted in a decline in the predictive power of the RF model. Consequently, only the top 7 descriptors were retained for subsequent studies. Table 3 provides representative cross-validation, train, and test metrics for both the RF and ET models trained on both the 7 and 281 descriptors, revealing no significant decrease in model performance with the reduction in the number of descriptors.

Table 3. Metrics for RF and ET models trained on different number of descriptors

Model	No. of Features	Cross-val R^2 score	Train R^2 score	Test R^2 score
RF	**281**	0.886	0.984	0.868
	7	0.877	0.983	0.865
ET	**281**	0.905	1.0	0.895
	7	0.882	0.999	0.865

Hyperparameter tuning

Hyperparameter tuning is a critical step in optimizing the performance of ML models, and GridSearchCV is a valuable tool for accomplishing this task, particularly when applied to RF algorithms. GridSearchCV systematically explores a predefined hyperparameter grid, searching for the combination that yields the

best model performance based on specified evaluation metrics. For RF, essential hyperparameters include the number of trees in the forest, the maximum depth of each tree, and the minimum number of samples required to split an internal node. By exhaustively testing various parameter combinations, GridSearchCV helps identify the optimal configuration, leading to a RF model that generalizes well to new data and achieves superior predictive capabilities. This systematic approach streamlines the hyperparameter tuning process, enabling data scientists to find the best model parameters efficiently and enhance the overall effectiveness of RF in solving diverse ML problems. The exploration of various parameter values through GridSearchCV resulted in the identification of final parameters, as outlined in Table 4. Remarkably, the adoption of these new parameters contributed to a reduction in overfitting, with training R^2 scores decreasing from 0.983 to 0.962. Meanwhile, only a marginal decline in test R^2 scores was noted, moving from 0.865 to 0.854.

Table 4. Parameters and their values screened using GridSearchCV

Entry	**Parameter**	**Values screened**	**Final value**
1.	max_depth	[30, 45, 60, 75, 100, None]	30
2.	max_features	[log2, sqrt]	log2
3.	min_samples_leaf	[2, 4]	2
4.	min_samples_split	[5, 10]	5
5.	n_estimators	[400, 600, 700, 800, 900]	900

Predictions

Fig. 4A represents the regression and residual plots for RF model trained using top 7 descriptors. Thiophenol (A), pinacolone (B) and fenfuram (C) were predicted with high accuracy with absolute error zero or close to zero, while some complex molecules such as dimecron (D), sucrose (E) and etofenprox (F) were predicted with relatively high absolute error of 2.38, 2.45, and 2.81, respectively (Fig. 4B). The variation in prediction errors for different molecules in the test set can be attributed to several factors. Molecules with complex structures or unique features such as the ones discussed above are more challenging for the RF model to accurately predict probably because the training data does not cover similar structures well, and the model may struggle to generalize to such cases. The other reason could be that the RF model is slightly overfitting the training data;

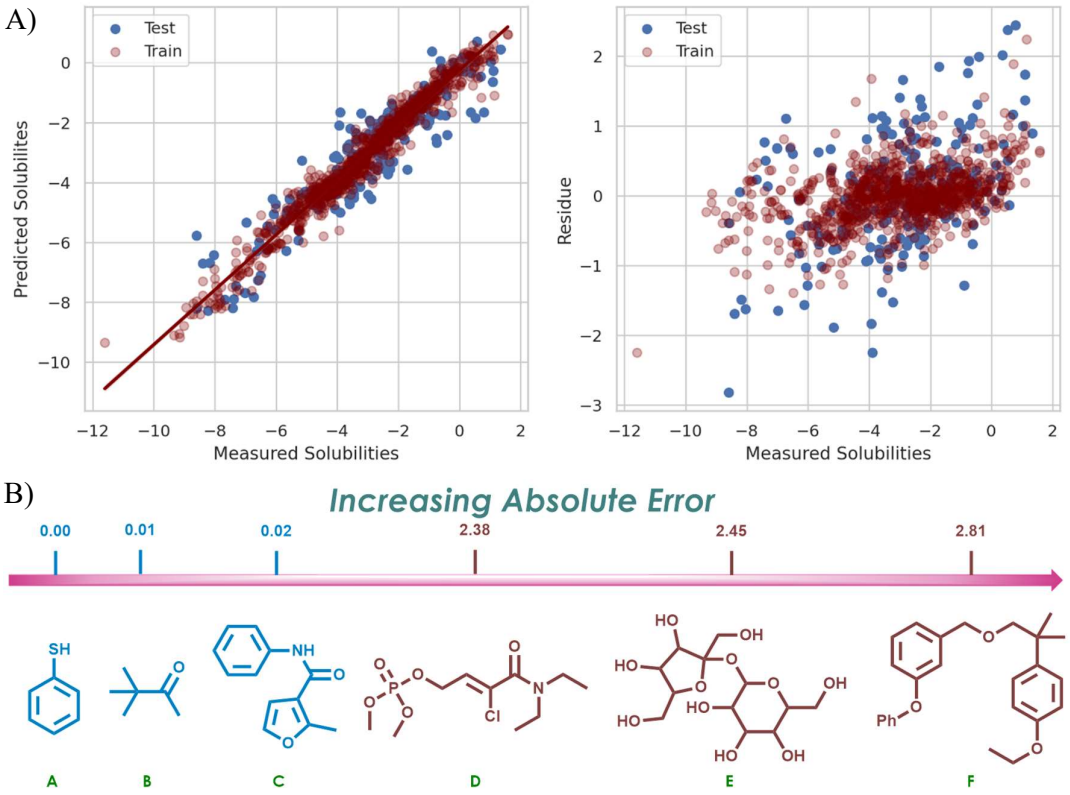

Figure 4. A) Regression and residual plots for the RF model trained on 7 descriptors. B) Structure and absolute prediction error (predicted – measured) for top 3 and bottom three organic molecules.

therefore the possibility of performing poorly on some molecules cannot be disregarded.

Explainability

The benefits offered by these ML algorithms, when compared to simpler and more interpretable linear models like MLR and PLS, come at the expense of increased complexity. This complexity introduces uncertainty regarding their functioning and decision-making capabilities. Scientists, who prioritize results that align with their cognitive abilities and knowledge gained from experimentation, are hesitant to adopt "black box" models for real-world applications due to their ambiguous nature. Consequently, explainable artificial intelligence (XAI), a burgeoning field, has gained significant popularity in the scientific community. XAI focuses on approaches that simplify ML models without sacrificing accuracy, resulting in inherently interpretable models.

Our primary goal has always been to make the QSPR study more straightforward and transparent for human comprehension. To achieve this, various variable reduction techniques were employed. We implemented an RF-based feature importance algorithm, which identified seven numerical descriptors out of 281 as the most influential. The key question arises: can the impact of these seven descriptors, crucial for the inner workings and mechanism of the tree-based architecture, be explained using a chemist's heuristic definition of solubility? Table 5 provides a concise description of all seven descriptors, along with their relative feature importance scores.[17]

Notably, FilterItLogS, representing the computed log of solubilities of molecules, emerged as the most vital quantitative descriptor in contributing to the development of the RF model. The second most crucial descriptor, Lipinski, is a Boolean variable with true and false values. Molecules adhering to the four rules of Lipinski were assigned a true value, while others received a false value. This binary representation, converted to 1s and 0s during data pre-processing, is

significant as it considers molecular mass, hydrogen bond acceptor/donor abilities, and octanol-water partition coefficient log P—a measure of solubilities.

SIC0, accounting for the structural information content of atoms in the molecule, plays a crucial role in predicting solubilities. RNCG and RPCG consider the relative negative and positive charges on molecules, reflecting non-covalent interactions between solute-solvent and solute-solute. These interactions are essential contributors to solubility variations, justifying their inclusion as top descriptors. The remaining two descriptors, ATS0Z and AATS0i, make minor contributions and are autocorrelation descriptors, representing the spatial arrangement of atoms or molecular properties in a molecule. Despite their minor role in the RF model, these descriptors involve calculating correlations between properties and their values at different interatomic distances within a molecule.

Table 5. Top 7 features, their description, and feature importance value

Entry	Feature label	Feature description	Feature importance
1.	FilterItLogS	Filter-it™ LogS	0.690
2.	Lipinski	Lipinski rule of five	0.076
3.	SIC0	0-ordered structural information content	0.030
4.	RNCG	Relative negative charge	0.022
5.	RPCG	Relative positive charge	0.007
6.	ATS0Z	Moreau-broto autocorrelation of lag 0 weighted by atomic number	0.007
7.	AATS0i	Averaged moreau-broto autocorrelation of lag 0 weighted by ionization potential	0.005

Conclusion

In conclusion, this manuscript provides an exploration of ML models for predicting organic molecule solubilities. The study emphasizes the significance of solubility predictions in pharmaceuticals, environmental science, and materials

engineering. Through careful data preprocessing, feature reduction, and model evaluation, we demonstrate the efficacy of ML models, specifically RF and ET. The top seven descriptors identified, including FilterItLogS, Lipinski, SIC0, RNCG, RPCG, ATS0Z, and AATS0i, play crucial roles in predicting solubilities and offer insights into molecular interactions. The study also highlights the importance of model interpretability, paving the way for the adoption of ML approaches in real-world applications. Overall, our findings contribute to advancing research in solubility prediction, fostering innovation in diverse scientific fields.

Data availability

The dataset and model algorithms can be accessed from this link:

https://github.com/codetodiscovery/Predict-Solubilities.git

References

1. Jorgensen, W. L.; Duffy, E. M., Prediction of drug solubility from structure. *Adv. Drug Deliv. Rev.* **2002**, 54, 355-366.
2. Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N., Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, 11, 5753.
3. Hou, T. J.; Xia, K.; Zhang, W.; Xu, X. J., ADME Evaluation in Drug Discovery. 4. Prediction of Aqueous Solubility Based on Atom Contribution Approach. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 266-275.
4. Tayyebi, A.; Alshami, A. S.; Rabiei, Z.; Yu, X.; Ismail, N.; Talukder, M. J.; Power, J., Prediction of organic compound aqueous solubility using machine learning: a comparison study of descriptor-based and fingerprints-based models. *J. Cheminform.* **2023**, 15, 99.
5. Patil, G. S., Prediction of aqueous solubility and octanol—water partition coefficient for pesticides based on their molecular structure. *J. Hazard. Mater.* **1994**, 36, 34-43.
6. Nirmalakhandan, N. N.; Speece, R. E., Prediction of aqueous solubility of organic chemicals based on molecular structure. *Environ. Sci. Technol.* **1988**, 22, 328-338.
7. Mitchell, B. E.; Jurs, P. C., Prediction of Aqueous Solubility of Organic Compounds from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 489-496.
8. Kühne, R.; Ebert, R. U.; Kleint, F.; Schmidt, G.; Schüürmann, G., Group contribution methods to estimate water solubility of organic chemicals. *Chemosphere* **1995**, 30, 2061-2077.
9. Klopman, G.; Wang, S.; Balthasar, D. M., Estimation of aqueous solubility of organic molecules by the group contribution approach. Application to the study of biodegradation. *J. Chem. Inf. Comput. Sci.* **1992**, 32, 474-482.
10. Lee, Y.-C.; Myrdal, P. B.; Yalkowsky, S. H., Aqueous functional group activity coefficients (AQUAFAC) 4: Applications to complex organic compounds. *Chemosphere* **1996**, 33, 2129-2144.
11. Huuskonen, J., Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, 40, 773-777.
12. Yan, A.; Gasteiger, J., Prediction of Aqueous Solubility of Organic Compounds Based on a 3D Structure Representation. *J. Chem. Inf. Comput. Sci.* **2003**, 43, 429-434.
13. Delaney, J. S., ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **2004**, 44, 1000-1005.
14. Ali, J.; Camilleri, P.; Brown, M. B.; Hutt, A. J.; Kirton, S. B., In Silico Prediction of Aqueous Solubility Using Simple QSPR Models: The Importance of Phenol and Phenol-like Moieties. *J. Chem. Inf. Model.* **2012**, 52, 2950-2957.

15. Sorkun, M. C.; Koelman, J. M. V. A.; Er, S., Pushing the limits of solubility prediction via quality-oriented data selection. *iScience* **2021**, 24.
16. Salahinejad, M.; Le, T. C.; Winkler, D. A., Aqueous Solubility Prediction: Do Crystal Lattice Interactions Help? *Mol. Pharmaceutics* **2013**, 10, 2757-2766.
17. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T., Mordred: a molecular descriptor calculator. *J. Cheminform.* **2018**, 10, 4.
18. Mitchell, J. B. O., Machine learning methods in chemoinformatics. *WIREs Comput. Mol. Sci.* **2014**, 4, 468-481.
19. Lever, J.; Krzywinski, M.; Altman, N., Model selection and overfitting. *Nat. Methods* **2016**, 13, 703-704.
20. Krzywinski, M.; Altman, N., Multiple linear regression. *Nat. Methods* **2015**, 12, 1103-1104.
21. Dybowski, R., Interpretable machine learning as a tool for scientific discovery in chemistry. *New J. Chem.* **2020**, 44, 20914-20920.
22. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, 1, 206-215.
23. Lipton, Z. C., The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* **2018**, 16, 31–57.
24. Murdoch, W. J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B., Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci.* **2019**, 116, 22071-22080.