Evaluating Featurizer-Model Combinations and Their Interpretability in Solubility Prediction

Moukthika Kuruva[1]

Received November 28, 2024
Accepted April 22, 2026
Electronic access May 15, 2026

This study evaluates the performance and interpretability of featurizer model combinations for predicting aqueous solubility in drug discovery. Using the AquaSolDB-clean dataset (about 10,000 molecules), the Light Gradient Boosting Machine (LightGBM), Random Forest, and Extreme Gradient Boosting Machine (XGBoost) models were assessed with Linear Regression as a baseline model, in conjunction with Extended-Connectivity Fingerprints (ECFP), Chem Bidirectional Encoder Representations from Transformers (ChemBERTa 2.0), and Graph Convolutional Networks (GCN). Among all combinations, LightGBM and ChemBERTa achieved the best performance with a root mean squared error (RMSE) of 1.181 (95% CI: 1.138–1.224). Paired t-tests against all other model-featurizer combinations confirmed that this combination performed significantly better than all others except XGBoost + ChemBERTa and Linear Regression + ChemBERTa, for which the differences were not statistically significant ($p = 0.12$ and $p = 0.17$, respectively). The study also investigates molecular substructures driving solubility predictions using Shapley Additive Explanations (SHAP), finding specific chemical structures contributing to accurate predictions. Confidence intervals are reported for each combination, statistical significance testing is performed using paired t-tests, and a critical analysis of featurizer contributions is provided, offering interpretability through feature attribution methods.

Introduction

Aqueous solubility is a critical physicochemical property that plays a key role in many scientific and industrial applications, particularly in drug discovery, chemical engineering, and environmental science[1]. In pharmaceutical development, solubility directly affects drug absorption and bioavailability. Poor solubility is one of the leading causes of failure in drug candidates, making accurate prediction of solubility an important step in early-stage screening. Traditional methods for measuring solubility rely on experimental techniques, which are often time-consuming and expensive[2]. As a result, there has been growing interest in developing computational approaches that can efficiently predict solubility from molecular structure. Machine learning models have become a popular solution due to their ability to learn complex relationships between chemical features and their corresponding solubility values.

Recent work has explored a wide range of modeling approaches, including basic linear regression models, ensemble learning methods, deep neural networks, and graph-based models. Common representations employed by these studies include molecular descriptors, fingerprints, SMILES-based embeddings, and graph structures, each capturing different aspects of chemical information. However, existing studies often focus on a focused or single set of models or representations, making it difficult to determine which approach among the many that exist is most efficient. In addition, many high-performing methods rely on complex feature engineering techniques or domain-specific knowledge, which can limit their practical applicability. There is a need for a comparative study of the three main featurization types: graph-based, SMILES-based, and fingerprint-based, across different model structures including gradient boosting, decision trees, and graphbased deep learning. This study aims to address these challenges by evaluating multiple machine learning models for solubility prediction using a consistent dataset and evaluation strategy. Different molecular representations, including sequence-based embeddings, graph-based features, and fingerprint-based features, are explored to assess their impact on performance. In addition, model performance is evaluated using cross-validation and statistical analysis to ensure reliability. The following question is aimed to be answered through the study: How do different molecular featurizers, from graph-based to language-based, affect the performance and interpretability of a machine learning model in predicting solubility?

Related Work

Solubility prediction has been widely studied using both traditional machine learning and deep learning approaches, with a strong focus on identifying effective model architectures and molecular representations. Early approaches relied on linear and statistical models[3]. More recent work has explored ensem-

[1] *Luminaire Education*
Washington High School , California, USA

ble methods, graph-based learning, transformer-based architectures, and common regression models such as LightGBM and Random Forest[4,5]. Several studies have demonstrated the effectiveness of ensemble and tree-based models for solubility prediction. LightGBM has been shown to outperform multiple baseline models including partial least squares, ridge regression, k-nearest neighbors, decision trees, extra trees, random forests, and support vector machines, while also revealing relationships between structural features and solubility[6]. Other work comparing random forests, LightGBM, Least Absolute Shrinkage and Selection Operator (LASSO), and partial least squares found that LASSO achieved the best predictive performance, while random forests offered a strong balance between model complexity and predictive ability[7]. Ensemble QSPR models combining random forest and XGBoost have also been shown to improve accuracy compared to individual models[8]. Additional studies using random forest, XGBoost, LightGBM, and CatBoost further highlight the importance of ensemble methods, while also exploring interpretability in these models[9]. These works establish ensemble learning as a strong baseline for solubility prediction.

Deep learning approaches have also been widely explored to capture more complex relationships. Transformer based models such as SolTranNet have demonstrated improved performance over traditional linear models by learning contextual representations from SMILES strings, although larger models were not always beneficial for prediction performance[10]. Other work using convolutional neural networks, deep neural networks, and generalized regression neural networks found that GRNN achieved the best performance when modeling solubility under varying temperature and pressure conditions[11]. These studies show that deep learning models can effectively learn nonlinear relationships from molecular data. Graph-based learning methods have gained attention due to their ability to directly model molecular structure. Studies using graph convolutional networks and graph attention mechanisms have shown strong performance, especially on high-quality datasets, although these models were sensitive to noise and errors[12]. In contrast, descriptor-based approaches have been shown to be more stable, with large-scale analysis identifying hundreds of molecular descriptors that significantly contribute to solubility prediction[12]. Additional work comparing molecular descriptors and Morgan fingerprints with random forest models showed that descriptor-based features outperform fingerprint-based representations, while also providing interpretability through SHAP analysis[13]. More recent work has also explored combining multiple representations, including electrostatic potential maps, molecular graphs, and tabular descriptors, where tabular features combined with XGBoost achieved the best performance[14]. These findings highlight the importance of feature representation in solubility prediction.

Beyond purely data-driven approaches, hybrid methods that incorporate thermodynamic modeling have also been explored. COSMO-SAC and COSMO-RS models have been combined with machine learning techniques to improve predictive accuracy. For example, Gaussian process regression has been used to correct deviations in COSMO-SAC predictions, significantly improving alignment with experimental data[15]. Similarly, neural networks combined with COSMO-RS achieved strong predictive performance while requiring less data[16]. Other work has proposed integrating machine learning predictions with thermodynamic properties such as activity coefficients, fusion enthalpy, and melting point to improve solubility estimation[17]. These approaches show that incorporating physical chemistry knowledge can improve prediction quality. Additional studies have explored specialized feature engineering and modeling techniques. Regression methods combining LASSO and support vector regression achieved improved accuracy for high-temperature solubility prediction[18]. Molecular dynamics simulations have also been used to extract physicochemical features, which were then used in models such as random forest, extra trees, XGBoost, and gradient boosting, with gradient boosting achieving the best performance[19]. Other work using COSMO-RS with descriptors derived from quantum chemical methods applied models such as association rule mining, decision trees, and neural networks, showing that specific descriptor types have varying levels of influence on solubility prediction[20]. While these approaches improve performance, they often require complex feature extraction processes.

Despite the large body of work, several important gaps remain. First, many studies focus on a limited set of model types, rather than performing a direct comparison across fundamentally different model structures under the same experimental setup. Second, there is no clear consensus on the most effective molecular representation, with different studies favoring SMILES-based embeddings, graph-based features, molecular descriptors, or fingerprints. Third, many high-performing approaches rely on complex feature engineering techniques such as molecular dynamics simulations or thermodynamic modeling, which limits their practical applicability. Finally, interpretability is not consistently applied across different modeling approaches, making it difficult to fully understand model behavior.

To address these limitations, three models of varying structures, paired with molecular fingerprints like ECFP, ChemBERTa 2.0, and GCN have been tested under the same conditions on the same dataset to find optimal model-featurizer pairings[21–23].

Methods

Dataset and Featurizers

AquaSolDB-clean or AqSolDBc is a curated version of the widely-used AquaSolDB dataset[24]. By removing repeating data points and values below the limit of detection or quantification, the probability of error is reduced. The clean dataset contains

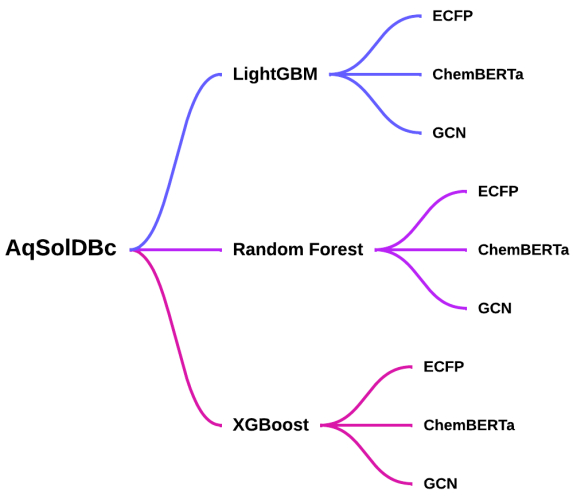

Fig. 1 Overview of the Research Workflow: Illustrating the process of evaluating different ML models with featurizers.

approximately ten thousand values and includes many molecular features, including SMILES strings. Duplicates and extreme outliers were removed. The used feature from the dataset includes the SMILES strings and the predictions were compared to the experimentally measured logS values. Potential limitations include its limited size and chemical diversity relative to other datasets (e.g., ESOL, ZINC).

Featurizers- In this study, a diverse set of molecular featurizers were employed to capture complex chemical information. ECFP converts molecular structures into binary vectors, encoding substructural features important for solubility prediction. This fingerprint-based approach is commonly used due to its ability to represent the detailed connectivity patterns of atoms. ChemBERTa 2.0, a language based featurizer, applies transformer models to interpret molecular structures formatted as SMILES strings. It is an experienced featurizer, benefiting from extensive pre-training on vast chemical datasets, enabling it to record complex patterns within molecular sequences. Lastly, the GCN represents molecules as graphs, where the nodes and edges correspond to atoms and bonds. This graph-based method provides a detailed understanding of the influence of structural configurations on chemical behavior. These featurizers were chosen to span symbolic, language-based, and structural paradigms. Future work could expand this to include DeepChem's featurizers such as Mol2Vec, MACCS, and others.

Models

To evaluate the influence of model architecture on solubility prediction, four different machine learning models were implemented. These included tree-based ensemble methods like Random Forest, LightGBM, and XGBoost, and a baseline Linear Regression model representing a traditional QSAR-style approach. These models were chosen to reflect a range of capacities for capturing non-linear relationships and molecular complexity. However, each model type comes with its limitations. For example, linear regression may underfit complex patterns, while tree-based models can overfit small datasets or become biased toward dominant substructures.

Experimental Setup

The input features for the model included only the SMILES strings representing molecular structure, and no variables that were direct indicators of solubility were utilized to prevent data leakage. To evaluate generalization to novel chemical scaffolds, a scaffold splitting strategy with an 80/10/10 split for training, validation, and testing was adopted. This ratio is widely accepted as a strong balance between training capacity and realistic generalization. Scaffold splitting ensures that molecules with similar core structures are grouped together, making the test set more challenging and reflective of real-world generalization scenarios compared to random splits. To enhance statistical reliability, 5-fold cross-validation was applied across all model–featurizer combinations. For each fold, RMSE was computed, reporting both the mean and standard deviation to quantify model performance and variance. To determine whether differences between models were statistically significant, RMSE distributions were simulated and pairwise t-tests were conducted against the best performing model. Finally, for interpretability, a SHAP analysis was performed on the featurizers to highlight feature contributions.

Feature Attribution

Feature attribution was performed using multiple complementary techniques tailored to the molecular representation. For ECFP-based models, global importance was derived from gain-based rankings averaged across cross-validation folds. Fingerprint bits were then mapped back to chemical substructures using RDKit, and atom-level heatmaps were generated to visualize molecular fragments[25]. For ChemBERTa embeddings, SHAP was applied to the XGBoost classifier trained over pooled embeddings. This produced global feature rankings across embedding dimensions. To provide chemical interpretability, token-level attributions were projected back onto atoms, and color-coded heatmaps were generated to highlight contributions that increased or decreased solubility predictions. For the GCN models, important atoms were identified using a combination of SHAP analysis and per-atom occlusion, where each atom was temporarily masked to observe its effect on the prediction. The resulting importance scores were visualized as red and blue highlights on each molecule, illustrating the functional groups

that most strongly influenced the model's decision.

Results

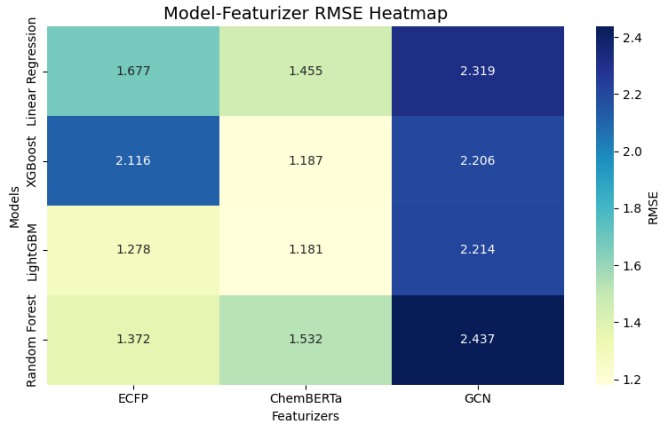

Fig. 2 Heatmap displaying the RMSE for all model-featurizer combinations tested

Performance Comparison

The experiments revealed substantial variation in predictive performance across model–featurizer combinations (Fig. 2). The best-performing configurations were LightGBM combined with ChemBERTa and XGBoost combined with ChemBERTa, achieving mean RMSE values of 1.181 (95% CI: 1.138–1.224) and 1.187 (95% CI: 1.147–1.227), respectively. The overlapping confidence intervals indicate that the difference between these two models is not statistically significant.

The statistical analysis further confirms ChemBERTa's effectiveness as a featurizer. Pairwise t-tests using LightGBM + ChemBERTa as the reference model show that every non-ChemBERTa combination, including GCN-based models, ECFP-based models, and linear baselines, performed significantly worse ($p < 0.01$ for all comparisons). Only the two ChemBERTa combinations, XGBoost + ChemBERTa ($p = 0.12468$) and Linear Regression + ChemBERTa ($p = 0.167114$), failed to show a statistically significant difference.

Among traditional fingerprint approaches, ECFP showed competitive performance, particularly when paired with LightGBM (RMSE 1.278, 95% CI: 1.230–1.326) and Random Forest (RMSE 1.372, 95% CI: 1.312–1.432). Linear Regression models generally lagged behind tree-based methods, even when paired with ChemBERTa. In contrast, models using the GCN featurizer consistently underperformed, with RMSE values ranging from 2.206 (95% CI: 2.153–2.259) to 2.437 (95% CI: 2.418–2.456).

Although Graph Convolutional Networks (GCNs) are theoretically well suited for molecular property prediction, the specific GCN implementation used in this study underperformed due to several architectural and data-related limitations. The model employed only atomic number as a single node feature, lacking important chemical information such as valence, aromaticity, hybridization, and formal charge. Additionally, bond features were not encoded, preventing the model from distinguishing between different bond types. These limitations restrict the ability of the GCN to learn meaningful structural representations.

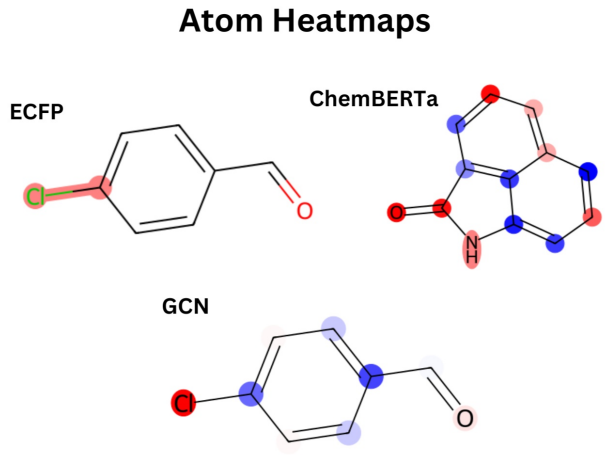

Fig. 3 Example substructures contributing to feature index.

Important Substructures

For the ECFP models, averaging feature importance across folds shows that only a few bits matter most, with bit 561 being the top signal. From Figure 3, it can be seen that when this bit is mapped back to chemical space, it corresponds strongly to an aryl chloride. This suggests that halogenated aromatic systems strongly drive the model's predictions.

It can be seen in Figure 4, that for the ChemBERTa + XGBoost model, global SHAP values on the embeddings reveal that just a handful of dimensions (such as features 248, 80, and 116) carry most of the signal. To connect these features back to chemistry, the attributions are projected onto the molecule itself. The resulting atom-level heatmap (Figure 3) shows that the carbonyl oxygen and nearby nitrogen atoms are the main drivers, while several aromatic carbons instead reduce the prediction.

For the GCN pipelines, the global SHAP bar plot (Figure 5) over GCN embeddings shows a distributed profile. Instead of a single dominant feature, numerous embedding dimensions showed moderate importance. These signals were localized through per-atom occlusion on the XGBoost model and generated corresponding atom-level heatmaps. In this example, the chlorine and the oxygen are the most important atoms (shown

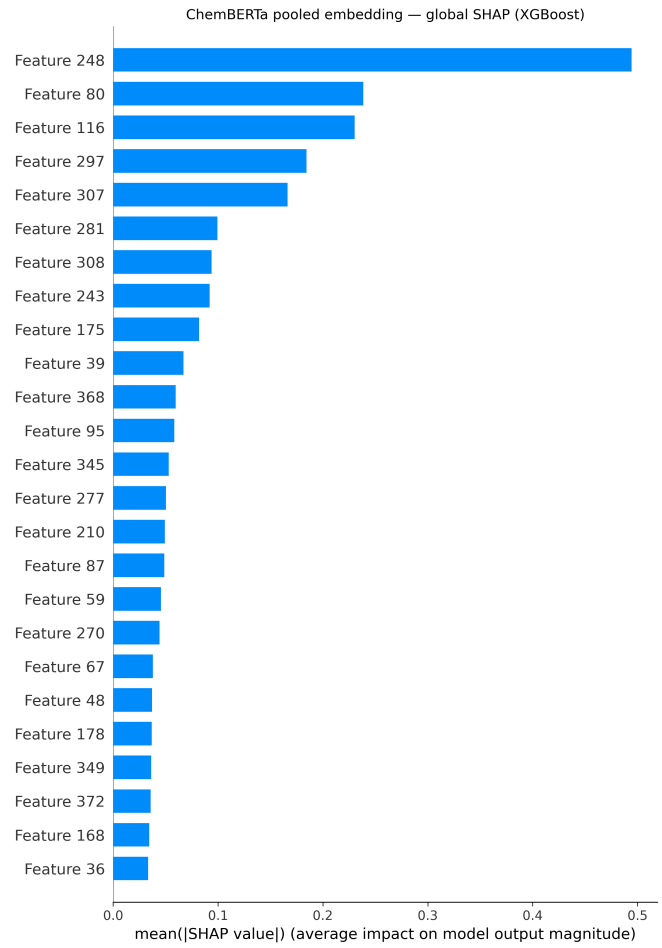

Fig. 4 SHAP bar showing ChemBERTa feature importance.

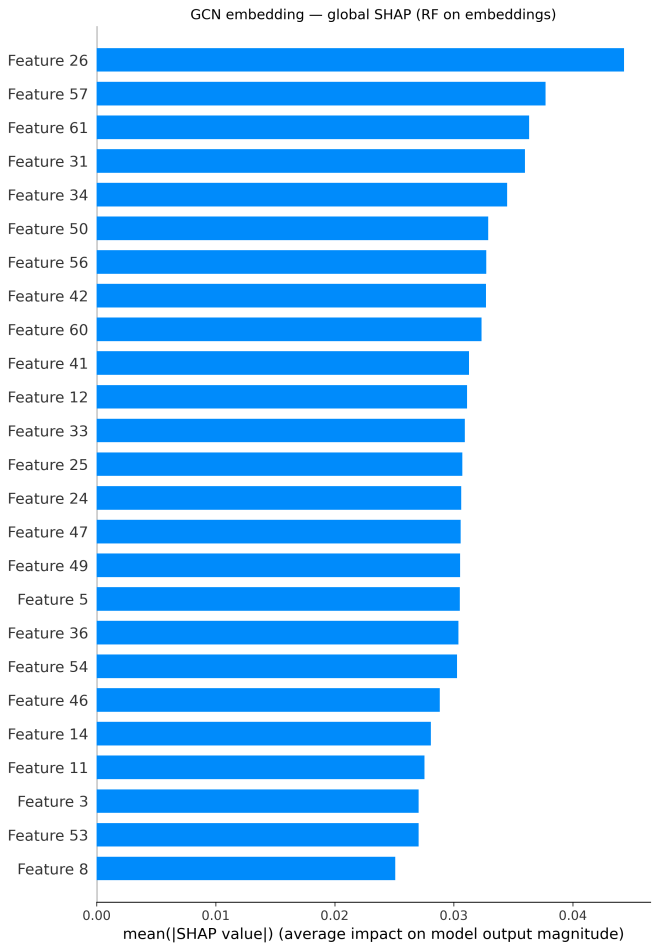

Fig. 5 SHAP bar showing GCN feature importance.

in red), while some of the ring carbons reduce the prediction (shown in blue). This pattern aligns with the trends observed in the ECFP and ChemBERTa models, with heteroatoms and halogens driving solubility-related signals.

Discussion

The results demonstrate that model performance in solubility prediction is strongly influenced by the choice of molecular representation and model architecture. Across all experiments, ChemBERTa-based molecular representations consistently produced the lowest RMSE values when combined with models such as XGBoost and LightGBM. The narrow confidence intervals observed for these models indicate stable performance across cross validation folds, suggesting that these representations capture meaningful structural information consistently. In contrast, other featurization approaches, particularly graph-based representations, resulted in higher prediction errors despite having stable confidence intervals. When it comes to feature representation, studies combining multiple representations, such as electrostatic potential maps, graph features, and tabular descriptors, have shown that representation choice plays a central role in predictive performance [14]. The results of this study reinforce the idea that richer representations generally lead to better model outcomes.

These findings are consistent with prior work showing that representations from SMILES strings can effectively encode molecular structure. For example, transformer-based models such as SolTranNet have demonstrated improved performance over traditional linear models by capturing contextual relationships within SMILES sequences [10]. The strong performance of ChemBERTa in this study supports this idea, indicating that pretrained language models can provide features helpful for downstream prediction tasks. In addition, the effectiveness of gradient boosting methods aligns with previous studies where LightGBM and other ensemble models outperformed a range of baseline approaches [6,11,14]. These models are able to recognize complex feature interactions while maintaining strong

generalization, contributing to their consistent performance.

The underperformance of graph-based models in this study contrasts with prior work where graph neural networks achieved strong results for molecular property prediction [12]. This difference can be explained by the simplicity of the GCN implementation used in this work. The model relied only on atomic number as the node feature and did not include bond features or additional chemical descriptors. As a result, the model lacked important information such as bond type, aromaticity, and hybridization, which are known to be critical for capturing molecular structure. This limitation likely restricted the model's ability to learn meaningful graph representations, leading to reduced predictive performance. This observation suggests that while graph-based models have strong theoretical advantages, their effectiveness depends heavily on the quality and richness of the input features.

Despite these findings, several limitations should be noted. First, the study was conducted on a single, small dataset, which may limit the generalizability of the results to other chemical domains or datasets with different distributions. Second, the GCN architecture used in this work was relatively simple and did not incorporate advanced features such as edge attributes, attention mechanisms, or deeper network structures. Third, ChemBERTa embeddings were used without fine-tuning, possibly affecting their performance slightly. Finally, while multiple model types were compared, hyperparameter optimization was not exhaustively explored, which could influence performance outcomes.

These limitations also point to directions for future work. Incorporating richer graph features, experimenting with a larger dataset and more advanced graph neural network architectures could improve the performance of graph-based models. In addition, combining multiple representations, such as graph features and learned embeddings, could provide deeper information and improve predictive accuracy.

Overall, this study demonstrates that representation choice is a key factor in solubility prediction. ChemBERTa-based embeddings combined with gradient boosting methods provide a strong balance between accuracy and stability, while simpler graph-based approaches may require more advanced feature engineering to achieve competitive performance. These results provide useful guidance for selecting models and representations in molecular property prediction tasks.

Conclusion

This study shows that molecular representation plays a central role in solubility prediction, with ChemBERTa features combined with gradient-boosting models achieving the best overall performance. The results highlight that selecting an optimal pairing of featurizer and model can significantly improve predictive accuracy. In addition, interpretability analysis confirms that the models capture chemically meaningful patterns, consistent with known factors such as hydrogen bonding and aromaticity. Some limitations should be noted. The study was conducted on a single dataset, which may limit generalizability to other chemical spaces. The performance of graph-based models was also worse than expected, likely due to simplified feature design and architectural choices rather than inherent limitations of graph neural networks. Future work can explore more advanced representations, including improved graph-based models and hybrid approaches that combine multiple feature types. Fine-tuning pretrained models and evaluating performance across more diverse datasets may further improve generalizability and applicability.

References

1 R. Qing, S. Hao, E. Smorodina, D. Jin, A. Zalevsky, S. Zhang, Protein design: From the aspect of water solubility and stability. Chemical Reviews 122, 14085–14179 (2022).

2 S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, M. J. Ahsan, Machine learning in drug discovery: a review. Artificial Intelligence Review 55, 1947–1999 (2022).

3 R. Gozalbes, A. Pineda-Lucena, QSAR-based solubility model for drug-like compounds. Bioorganic & Medicinal Chemistry 18, 7078–7084 (2010).

4 M. Li, H. Chen, H. Zhang, M. Zeng, B. Chen, L. Guan, Prediction of the aqueous solubility of compounds based on light gradient boosting machines. ACS Omega 7, 42027–42035 (2022).

5 D. S. Palmer, N. M. O'Boyle, R. C. Glen, J. B. O. Mitchell, Random forest models to predict aqueous solubility. Journal of Chemical Information and Modeling 47, 150–158 (2007).

6 Z. Ye, D. Ouyang, Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. Journal of Cheminformatics 13 (2021). doi:10.1186/s13321-021-00575-3.

7 M. Lovrić, K. Pavlović, P. Žuvela, A. Spataru, B. Lučić, R. Kern, M. W. Wong, Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds. Journal of Chemometrics 35, e3349 (2021). doi:10.1002/cem.3349.

8 P. Hu, Z. Jiao, Z. Zhang, Q. Wang, Development of solubility prediction models with ensemble learning. Industrial & Engineering Chemistry Research 60, 11627–11635 (2021). doi:10.1021/acs.iecr.1c02142.

9 N. Xue, Y. Zhang, S. Liu, Evaluation of Machine Learning Models for Aqueous Solubility Prediction in Drug Discovery. ICAIBD Conference, 26–33 (2024). doi:10.1109/ICAIBD62003.2024.10604556.

10 P. G. Francoeur, D. R. Koes, SolTranNet–a machine learning tool for fast aqueous solubility prediction. Journal of Chemical Information and Modeling 61, 2530–2536 (2021). doi:10.1021/acs.jcim.1c00331.

11 F. An, B. T. Sayed, R. M. Parra, M. H. Hamad, R. Sivaraman, Z. Z. Foumani, A. A. Rushchitc, E. El-Maghawry, R. M. Alzhrani, S. Alshehri, et al., Machine learning model for prediction of drug solubility in supercritical solvent. Journal of Molecular Liquids 363, 119901 (2022). doi:10.1016/j.molliq.2022.119901.

12 T. Zheng, J. B. Mitchell, S. Dobson, Revisiting the application of machine learning approaches in predicting aqueous solubility. ACS Omega 9, 35209–35222 (2024). doi:10.1021/acsomega.4c06163.

13 A. Tayyebi, A. S. Alshami, Z. Rabiei, X. Yu, N. Ismail, M. J. Talukder, J. Power, Prediction of organic compound aqueous solubility using machine learning. Journal of Cheminformatics 15 (2023). doi:10.1186/s13321-023-00752-6.

14 M. A. Ghanavati, S. Ahmadi, S. Rohani, A machine learning approach for the prediction of aqueous solubility of pharmaceuticals. Digital Discovery 3, 2085–2104 (2024). doi:10.1039/d4dd00065j.

15 G. Oliveira, P. H. Wegner, P. V. de Lima Carvalho, F. A. Voll, A. P. Scheer, R. P. Soares, F. O. Farias, Machine learning-enhanced Cosmo-SAC for accurate solubility predictions. Fluid Phase Equilibria 600, 114535 (2026). doi:10.1016/j.fluid.2025.114535.

16 N. Mac Fhionnlaoich, J. Zeglinski, M. Simon, B. Wood, S. Davin, B. Glennon, A hybrid approach to aqueous solubility prediction using COSMO-RS and machine learning. Chemical Engineering Research and Design 209, 67–71 (2024). doi:10.1016/j.cherd.2024.07.050.

17 E. Al Ibrahim, N. Morgan, S. Müller, S. Motati, W. H. Green, Accurately predicting solubility curves via thermodynamic cycle and machine learning. Journal of the American Chemical Society 147, 45057–45069 (2025). doi:10.1021/jacs.5c13746.

18 M. Osada, K. Tamura, I. Shimada, Prediction of the solubility of organic compounds in high-temperature water using machine learning. Journal of Supercritical Fluids 190, 105733 (2022). doi:10.1016/j.supflu.2022.105733.

19 Z. Sodaei, S. Ekrami, S. M. Hashemianzadeh, Machine learning analysis of molecular dynamics properties influencing drug solubility. Scientific Reports 15 (2025). doi:10.1038/s41598-025-11392-1.

20 E. Can, A. Jalal, I. G. Zirhlioglu, A. Uzun, R. Yildirim, Predicting water solubility in ionic liquids using machine learning. Journal of Molecular Liquids 332, 115848 (2021). doi:10.1016/j.molliq.2021.115848.

21 S. D. Axen, X.-P. Huang, E. L. Cáceres, L. Gendelev, B. L. Roth, M. J. Keiser, A simple representation of three-dimensional molecular structure. Journal of Medicinal Chemistry 60, 7393–7409 (2017).

22 A. Lang, et al., Fine-Tuning ChemBERTa-2 for Aqueous Solubility Prediction. ResearchGate (2023).

23 C. Deng, L. Liang, G. Xing, Y. Hua, T. Lu, Y. Zhang, Y. Chen, H. Liu, Multi-channel GCN ensembled machine learning model for molecular aqueous solubility prediction. Molecular Diversity 27, 1023–1035 (2023).

24 P. Llompart, C. Minoletti, S. Baybekov, D. Horvath, G. Marcou, A. Varnek, Will we ever be able to accurately predict solubility? Scientific Data 11, 303 (2024).

25 V. F. Scalfani, V. D. Patel, A. M. Fernandez, Visualizing chemical space networks with RDKit and NetworkX. Journal of Cheminformatics 14, 87 (2022).