



Characterizing Molecular Property Variations Through Multivariate Machine Learning and Descriptor Analysis

Avik Kr. Basu

Assistant Professor, Department of Physics, Raja Peary Mohan College, Uttarpara, Hooghly, West Bengal - 712258

Email: avik08ju@gmail.com, ORCID ID: 0000-0001-9147-8322

Received: 06/01/2026

Revised: 10/03/2026

Accepted: 12/04/2026

Published: 27/04/2026

Abstract— The aqueous solubility of a molecule is one of the most important properties of a molecule that are determined by complex interactions among molecular and physicochemical characteristics. This study devised an interpretable multivariate machine-learning framework to characterize and predict solubility variation. AqSolDB enabled analysis of 9,881 compounds, following a quality screening and validation of structures. 208 two-dimensional descriptors were computed using RDKit, followed by variance, correlation and multicollinearity filtering to yield 58 descriptors. Application of Principal component analysis, correlation analysis, six regression models, locked-test validation and SHAP interpretation. XGBoost achieved the best performance, with cross-validated R^2 of 0.793 ± 0.011 and locked-test R^2 of 0.786, RMSE of 0.807, and MAE of 0.589. The variance of the descriptors was explained by the first five principal components (72.8%). MolLogP, MW, TPSA, aromaticity and complexity were the main drivers of solubility. The framework combines the ability of accurate prediction with chemically meaningful interpretation of the variation of molecular properties.

Keywords— aqueous solubility, molecular descriptors, machine learning, multivariate analysis, explainable artificial intelligence

1. INTRODUCTION

The ability to predict and calculate molecular properties is an important aspect of computational chemistry, cheminformatics and in particular, modern drug discovery, since experimental characterization of large chemical libraries is technically challenging, time consuming, and costly. Data-driven approaches relate molecular structure to measurable physicochemical and biological properties, offering an efficient alternative. However, the predictive value of such predictions is very sensitive to the representation of the molecular information. Descriptor-based representations are also important, as they convert structural data into variables that are easily understood and related to different chemical properties such as size, polarity, lipophilicity, topology, hydrogen bonding ability, aromaticity, flexibility, and others. [1] Integrated modelling pipelines have therefore been designed to integrate molecular representation, preprocessing of data, model construction, model validation, and model prediction to realize reproducible, computational workflows [2].

The quality and relevance of the molecular features are of key importance for quantitative structure–property and structure–activity modelling. Descriptor sets of high dimensions can include redundant, highly correlated, and weakly informative variables that can lead to high computational costs and model instability. To overcome multicollinearity and the excessive redundancy in the set of descriptors, feature extraction and dimensionality-reduction techniques must be applied to identify the molecular features that are informative for the desired purpose [3]. The aqueous solubility is a key physicochemical property that plays a critical role in compound formulation, bioavailability, environmental fate and chemical development, making it a specific topic of interest in these regards. Descriptor-based models have been shown to be useful in making estimates of aqueous solubility using machine learning-based methods, with the success of the model in predicting aqueous solubility being heavily dependent on the choice of descriptors, choice of algorithm, and choice of validation method [4].

Most of the studies on predicting molecular properties mentioned above pay attention to prediction accuracy but fail to explain the decisions of the models based on molecular factors. Nonlinear algorithms can be very successful at modeling the relationship between molecular structure and values for property, but the predictions are not easily interpreted directly. Explainable machine learning methods can be used to establish descriptors that are influential and quantify the impact of changes in molecular features by means of predicted properties to enhance transparency and chemical interpretability [5]. This is especially true if the goal is to predict rather than just to explain the variation in molecular properties as a function of structure.

There have been recent advances in molecular machine learning technologies that have expanded the modelling approaches available. Machine-learning techniques have already been applied to research complex molecular properties and phenomena that were previously studied with the aid of intensive physical and/or quantum-chemical calculations [6]. In the field of drug discovery and molecular design, artificial intelligence has found applications in various aspects, such as property prediction, virtual screening, molecular generation, optimization, and candidate prioritization, highlighting the growing significance of computational models in chemical decision-making processes [7]. The ability of large-scale chemical language models to represent meaningful relationships between molecular structure and physicochemical properties has additionally been demonstrated [8]. Also, graph-based learning approaches offer strong alternatives to modeling molecular interaction and property variation in a direct way [9].

However, even with the recent progress, there is a great benefit of using descriptor-based modelling in studies where interpretability and multivariate characterization are their main goals. This is illustrated by chemical-space visualization techniques, which show that representations of the molecule can be used to cluster molecules based on their structural and physicochemical similarity, and can highlight systematic property gradients across the space of molecules [10]. Concurrently, recent analyses of molecular representations have highlighted that none of them are fully optimal, as they all offer different degrees of interpretability, scalability and predictive power [11]. Interpretable descriptor-based approaches are thus very relevant to studies identifying why molecular properties vary, not simply to maximize predictive power.

In the present study, aqueous solubility has been selected as a representative molecular property (of many) and studied here as it varies with a multivariate descriptor-based machine-learning framework. It applies to two-dimensional molecular descriptors based on molecular structures, and does not include explicit three-dimensional conformations, molecular dynamics, quantum-mechanical calculations or variables associated with the experimental conditions, such as temperature, pH, ionic strength, and solid-state form. The framework is, therefore, not designed to mimic all the physicochemical phenomena that affect experimental solubility, but rather to describe the structural variation in the available descriptor space. In these limits, the study incorporates a descriptor screening, correlation analysis, principal component analysis, superposed regression, independent locked-test evaluation and explainable machine learning.

The novelty of the study is the combination of predictive modelling and multivariate chemical space analysis and

interpretation of descriptors. The framework does not address solubility prediction as a solely algorithmic benchmarking challenge but explores how the property variation is influenced by a combination of molecular size, lipophilicity, polarity, aromaticity, topology, hydrogen-bonding capacity, flexibility and structural complexity. The resulting analysis is designed to be predictive and to offer chemically interpretable insights into the relationships between the many dimensions of aqueous solubility. Hence the study is aimed at three specific objectives which are:

- To characterize multivariate molecular-property variation by examining the statistical structure, correlation patterns, and principal dimensions of a screened molecular descriptor space.
- To develop and compare linear and nonlinear machine-learning models for predicting experimental aqueous solubility and evaluate their generalization using an independent locked test set.
- To identify and interpret the principal molecular descriptors and nonlinear effects governing aqueous-solubility variation using explainable machine-learning analysis.

2. LITERATURE REVIEW

In recent years, molecular property prediction studies have increasingly moved away from the typical linear QSPR approach to data-driven approaches that are able to model nonlinear relationships between molecular descriptors. In the field of aqueous solubility, multi-algorithm QSPR studies have revealed that performing several simpler statistical models simultaneously is superior to simpler models when various physicochemical and structural descriptors are used, as an ensemble of simpler models could be superior to the nonlinear machine-learning models [12]. These studies prove the benefit of comparing several learners, but numerous studies are limited to predictive ranking and not to identifying how descriptor interactions can affect the property landscape.

The predictive power and chemical interpretability of descriptor-based machine learning remains good. A thermodynamic-property prediction development framework showed that the process of collecting data, creating descriptors, preprocessing the data, screening the features, selecting the model, and model validation are interdependent, and not just as individual modelling steps [13]. This workflow is especially applicable to the problem of molecular solubility since high dimensional descriptions spaces are prone to contain redundant or collinear variables. But descriptor-based pipelines could still suffer from instability for reasons of applicability domain restrictions and atypical compounds if these factors are not taken into consideration. Since then, there has been a focus on outlier detection and domain analysis to determine whether the predictions are being made within chemically represented

areas of the descriptor space [14].

In parallel, the method space for molecular property prediction has been extended with representation learning. Learned representations from molecular sequences have been shown to encode structural and contextual information that is not necessarily captured manually engineered descriptors [15] by transformer-based models. They have the advantage of automatic feature learning from large molecular corpora, but at the same time have to deal with a higher computational cost and less direct interpretability than descriptor-based methods.

Another major alternative is the graph-oriented models. Multichannel substructure graph approaches explicitly consider local structural motifs and molecular connectivity, which allows for explicit representation of nonlinear inter-substructural relationships in a more direct manner than is possible in conventional tabular descriptor spaces [16]. Such approaches are appropriate for the more sophisticated molecular property applications but generally do not offer a clear mapping between the model input (and its parameters) and more familiar physicochemical properties like lipophilicity, polarity, or molecular size.

Labelling data has been further reduced by self-supervised molecular learning. In fact, a number of global molecular representation frameworks have demonstrated that it is possible to learn meaningful structural representations prior to the task specific fine-tuning process, benefiting downstream property prediction tasks [17]. This has been extended to chemistry-aware fragmentation strategies, where the local molecular information is also preserved as part of the pretraining process, but chemically meaningful substructures are also provided to the model [18]. These methods are very effective when a large collection of molecules has been collected, but without labels, and their latent features cannot be easily mapped directly to meaningful chemical mechanisms or to interpretable descriptor effects.

New architectures have been proposed for enhancing representation, such as the attention-based ones. In focused-attention models, like TransFoxMol, the focus on the relevant molecular tokens/structural regions has shown to enhance property-prediction performance [19]. However, like cascaded-attention and graph-contrastive methods, the combination of pretraining, graph representation learning, and task-specific fine-tuning has been used to capture richer molecular information from structural data [20]. These architectures offer high predictive flexibility but can be less appropriate for studies where the primary goal is an explicit description of the physicochemical variation.

The general drug-discovery literature also shows that machine learning is rapidly gaining relevance in candidate prioritisation, molecular screening, property prediction and in preclinical decision making [21]. The usefulness of the model in a translation, however, will not only be dependent

on the accuracy of the model, but also on its robustness, repeatability, interpretability, and ability to link the learned patterns to chemically relevant variables.

Recently, several multimodal self-supervised methods have been proposed, which have tried to integrate multiple molecular perspectives while minimizing the overlap between the different representations [22]. These kinds of frameworks can also enhance generalization by using multiple structural modalities but they also increase the number of features in the architecture and make it harder to see the individual contributions.

Overall, all of these studies have demonstrated clear advances in nonlinear QSPR modelling, graph learning, transformers, self-supervision, and multimodal molecular representations. They possess greater predictive flexibility and more elaborate structural encoding as their key strengths. They share common drawbacks of lack of interpretability, computational complexity and lack of focus on the explicit characterization of the way the variation of the properties is regulated by the values of traditional molecular descriptors. To this end, the present study combines all the above-mentioned molecular descriptors extracted from 2D structures, multivariate statistical analysis, multiple machine-learning regressors, independent locked-test validation, and SHAP-based interpretation into one. In this design, simultaneous prediction and chemical characterization of the variation in aqueous solubility is possible across the molecular descriptor space.

3. METHODOLOGY

3.1 Research Design and Analytical Framework

A quantitative, computational, and predictive research design was used in this study to characterize aqueous-solubility variation, by using molecular descriptors based on multivariate analysis and machine learning techniques. The analytical framework employed molecular-structure validation, generation of descriptors, multivariate exploration, supervised regression, model validation, and explainable artificial intelligence. Experimental LogS was regarded as a dependent molecular property, while structural, constitutional, physicochemical, topological, surface and hydrogen-bonding descriptors were considered as a predictor space.

The workflow was broken down into five steps. Before calculation of the descriptors, the molecular structure was validated and standardised. Secondly, a larger collection of two-dimensional molecular descriptors was obtained using RDKit from valid molecular structures. Third, the quality control and dimensionality reduction of the descriptors were carried out to eliminate undefined, low-information and highly correlated descriptors that are redundant. Moreover, the kept DES was analyzed using descriptive statistics, correlation analysis, principal component analysis (PCA), and supervised machine-learning regression. Finally, the best model was interpreted through the SHAP-based global and local descriptor-effect analysis to gain insight into the molecular features responsible for the variation in solubility.

All model selection and other pre-processing choices were made on the development data only, to avoid optimistic performance estimates. The independent locked test set was not used for the modelling pipeline or hyperparameter tuning, but was only used once the pipeline and hyperparameters were finalised.

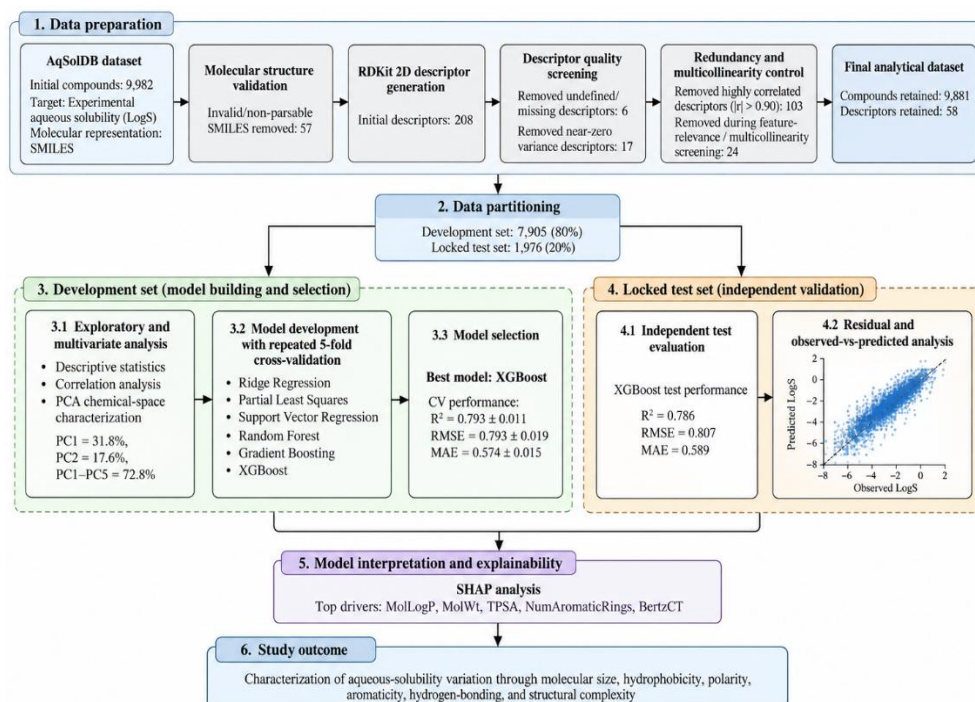


Figure 1. Workflow for Molecular Descriptor Analysis and Solubility Prediction

3.2 Data Source and Molecular Dataset

AqSolDB, an aqueous-solubility database containing a curated collection of molecular structures and aqueous-solubility measurements [23] was used for the study. The first set of data used was 9,982 compounds. In each observation, an individual chemical compound was used as the observation, the predicted value was LogS of the compound and the molecular structure was represented mainly by SMILES notation.

Pre-computation parsing of molecular structures was performed before computing the descriptors. The dataset was culled of 57 records that could not be represented in molecular form and 44 records which were removed after the descriptors were assessed for quality, leaving 9,881 compounds for analysis. There was large variability in the values of LogS observed; the mean was -2.91 , the standard deviation was 1.74 , the median was -2.76 , and the range was from -11.60 to 1.58 .

Two hundred and eight 2D descriptors were computed with RDKit. These encompassed various aspects of molecular structure such as molecular size, lipophilicity, polarity, hydrogen-bonding ability, aromaticity, flexibility, surface area, topology and structural complexity. The representative descriptors were MolLogP, molecular weight, topological polar surface area (TPSA), hydrogen-bond donors and acceptors, aromatic-ring count, rotatable-bond count, FractionCSP3, BertzCT, HeavyAtomCount, RingCount and LabuteASA.

3.3 Study Population and Data Partitioning

Chemically diverse compounds listed in AqSolDB with available valid molecular structure and experimental aqueous-solubility measurements were used for the analytical population. Conventional probabilistic participant sampling was not possible since the study did not rely on biological or human subjects to recruit participants but instead relied on an existing curated molecular database.

After a descriptor-quality and eligibility filter, 9,881 compounds retained were then divided into an 80% development set and a 20% locked test set. The development set consisted of 7,905 compounds that were employed for preprocessing the descriptors, multivariate analysis, cross-validation, selection of hyperparameters, and comparison of models. The remaining 1,976 compounds constituted an independent test set and were solely used for final generalization evaluation. In the development set, the model performance was estimated by repeated 5-fold cross-validation, to minimize the dependence on one training-validation split and to provide stable estimates.

3.4 Descriptor Processing and Multivariate Analysis

Descriptor preprocessing was designed to remove the noise, redundancy and the multicollinearity from the descriptors while retaining the chemically meaningful information. Among the 208 original descriptors, six were discarded for having undefined or missing descriptors and 17 for poorly varying descriptors, leaving 185 variables. A total of 103 highly redundant descriptors were removed with the help of pairwise correlation filtering with an absolute correlation threshold of $|r| > 0.90$. Using a combination of subsequent multicollinearity and feature-relevance screening, a set of 58 molecular descriptors were obtained.

To avoid information leakage, all data-dependent operations were fitted with development data solely and applied to the validation or test data. For descriptors that needed scaling, the values were scaled based solely on parameters derived from the training fold.

Pearson correlation coefficients were used to statistically describe the associations between descriptors and LogS and inter-descriptor relationship. The remaining space was then projected onto the principal components which were used to describe the major axes of molecular variation and to visualize the chemical space. The first five principal components accounted for 72.8% of the total descriptor variance, giving a reduced multivariate representation of the molecular size, hydrophobicity, complexity, polarity and hydrogen-bonding properties.

3.5 Machine-Learning Modelling and Performance Evaluation

Ridge Regression, Partial Least Squares Regression, Support Vector Regression, Random Forest, Gradient Boosting, and XGBoost regression methods were compared to quantify linear and nonlinear structural-property relationships. The linear and latent-variable models were Ridge and Partial Least Squares while nonlinear models were Support Vector Regression and tree-based ensemble models, which were able to model descriptor interactions and complex molecular-property relationships.

Only repeated five-fold cross-validation performed in the development partition to develop models. Hyperparameter optimization was done in the cross-validation framework and the preprocessing was done independently in each fold. The coefficient of determination (R^2), root mean square error (RMSE) and mean absolute error (MAE) were used for evaluating performance. To evaluate the predictive performance and stability of cross validation, mean $\pm$ standard deviation was reported.

Once model is selected, the final performance was measured only on the locked test set. XGBoost also had the best performance across all folds of the cross validation and was thus used as the main predictive model. Systematic prediction error and observed-versus-predicted comparisons were examined using residual analysis and observed-versus-predicted comparisons.

To evaluate the contribution of multivariate information, progressively refined representations, such as the initial quality-controlled descriptor pool, the correlation filtered space, and the final 58-descriptor set were compared by means of descriptor analyses. This enabled the evaluation of the potential value of dimensionality reduction to gain parsimony without compromising the predictive information.

3.6 Model Explainability and Descriptor-Effect Analysis

The final XGBoost model was explained using SHapley Additive exPlanations (SHAP). Global descriptor importance was summarized by the mean absolute SHAP values, and individual descriptor effects by SHAP summary and dependence analyses, which were non-linear. The mean absolute SHAP values quantified the importance of individual descriptors, while SHAP summary and dependence analyses characterized the direction, magnitude and non-linearity of individual descriptor effects. Special focus was put on MolLogP, MW, TPSA, aromatic-ring count, BertzCT, FractionCSP3, HeavyAtomCount, hydrogen-bond donors, RingCount and rotatable bonds. The interpretability stage allowed this predictive model to be applied not only as an estimation tool, but also as a chemical meaningful characterization tool for molecular-property variation.

3.7 Ethical and Responsible Computational Considerations

The study did not involve human participants, identifiable personal information, clinical records or animal experimentation but only used a previously compiled molecular database. As a result, the concepts of personal consent and individual privacy were not pertinent. However, the responsible uses of calculations were respected, with transparent preprocessing, independent testing, reproducible generation of the descriptors, and

explicit distinction between the development of the models and their final evaluation.

Bias can occur due to the lack of equal representation of chemical classes, experimental conditions, solubility measurement methods, or areas of chemical space in the source database. Model predictions should thus be used only in the range of applicability of the training data and should not be used to replace experimental measurement. Model interpretations instead of causal evidence of physicochemical mechanisms were used for SHAP explanations. These precautions were put in place to minimize overinterpretation and ensure clear and responsible use of machine learning for molecular-property prediction.

4. RESULTS

4.1 Dataset Characteristics and Experimental Solubility Variation

After validation of molecular structure and screening the data from the descriptors, 9,881 compounds were selected from the initial 9,982 compounds in the AqSolDB. 57 compounds failed to be successfully parsed into the molecular structure and another 44 compounds were discarded during the process of assessing the quality of the descriptors. The final dataset was split into 7,905 compounds used to build the model and 1,976 compounds used to evaluate the model with a locked test set.

Experimental aqueous solubility data varied greatly within the chemical space analyzed (Table 1). The mean LogS was -2.91 (with a standard deviation of 1.74) and the median was -2.76 . The values ranged from -11.60 to 1.58 reflecting a wide range of aqueous solubility behavior. The interquartile range was between -3.92 and -1.71 . A total of 2,511 compounds (25.4%) had LogS values below -4 , 4,076 compounds (41.3%) occurred between -4 and -2 , and 3,294 compounds (33.3%) had values of -2 or above.

Table 1. Dataset Characteristics and Experimental Solubility Distribution

Characteristic	Value
Initial compounds	9,982
Final analytical dataset	9,881
Development set	7,905
Locked test set	1,976
Mean LogS	-2.91
Standard deviation	1.74
Median LogS	-2.76
Range	-11.60 to 1.58
LogS < -4	2,511 (25.4%)

$-4 \leq \text{LogS} < -2$	4,076 (41.3%)
$\text{LogS} \geq -2$	3,294 (33.3%)

The following figure (Fig. 2) further depicts the distribution of LogS and its organisation in molecular descriptor space. Solubility distribution was wide and diverse; not focused in a confined range. PCA-based mapping showed that LogS

was continuous along the chemical space with different regions corresponding to low and high solubility. This pattern allowed for the use of multivariate analysis, which was not dependent on the use of any particular molecular descriptor.

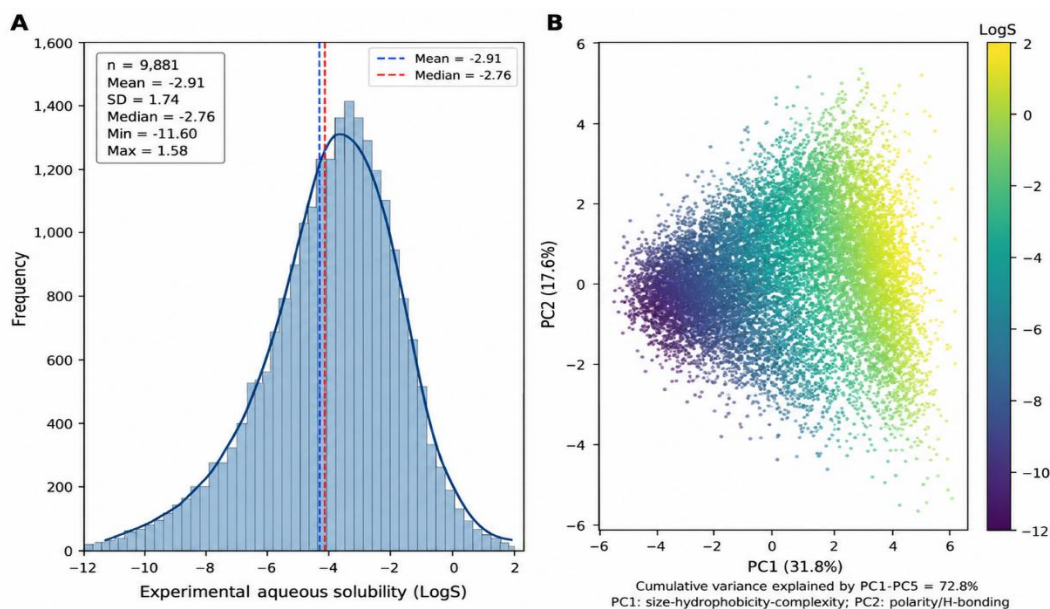


Figure 2. Experimental Solubility Distribution and Molecular Chemical-Space Variation

4.2 Multivariate Molecular Descriptor Structure

The dimensionality of the molecular representation was significantly reduced by the descriptor screening. The 208 descriptors generated initially were reduced to 17 due to a near zero variance and 6 were discarded due to undefined or missing values. The remaining 185 descriptors were subjected to pairwise correlation between descriptors ($|r| > 0.90$) and to the removal of descriptors that were not

relevant to the features, with 103 descriptors being excluded and 24 others being removed. There were thus 58 descriptors in the final modelling space.

The best correlation with aqueous solubility was found for MolLogP ($r = -0.61$), followed by molecular weight ($r = -0.47$), HeavyAtomCount ($r = -0.44$), LabuteASA ($r = -0.42$), NumAromaticRings ($r = -0.39$), and BertzCT ($r = -0.38$). TPSA was correlated with a positive correlation of 0.29, 0.23, 0.21 and 0.17, for NumHDonors, FractionCSP3 and NumHAcceptors, respectively.

Table 2. Multivariate Descriptor Profile and Selected Molecular Features

Descriptor	Mean $\pm$ SD	Correlation with LogS
MolLogP	2.31 ± 2.04	-0.61
MolWt	326.7 ± 151.4	-0.47
TPSA	67.8 ± 44.1	+0.29
NumAromaticRings	1.72 ± 1.41	-0.39
BertzCT	631.5 ± 447.8	-0.38
HeavyAtomCount	22.7 ± 10.4	-0.44
NumHDonors	1.54 ± 1.42	+0.23
FractionCSP3	0.39 ± 0.24	+0.21
RingCount	2.42 ± 1.75	-0.33

PCA was performed to confirm that the space retained was structured multivariate chemical space. PC1 explained 31.8% of total variance, PC2 17.6%, PC3 10.9%, PC4 7.3%, and PC5 5.2%, giving a cumulative explained variance of 72.8%. PC1 was mainly dominated by

molecular size, hydrophobicity and structural complexity, while PC2 was related to polarity and hydrogen-bonding properties. Descriptors with molecular size and complexity properties were grouped together, and the descriptors with polarity and hydrogen-bonding properties were grouped together in multivariate pattern (Fig. 3).

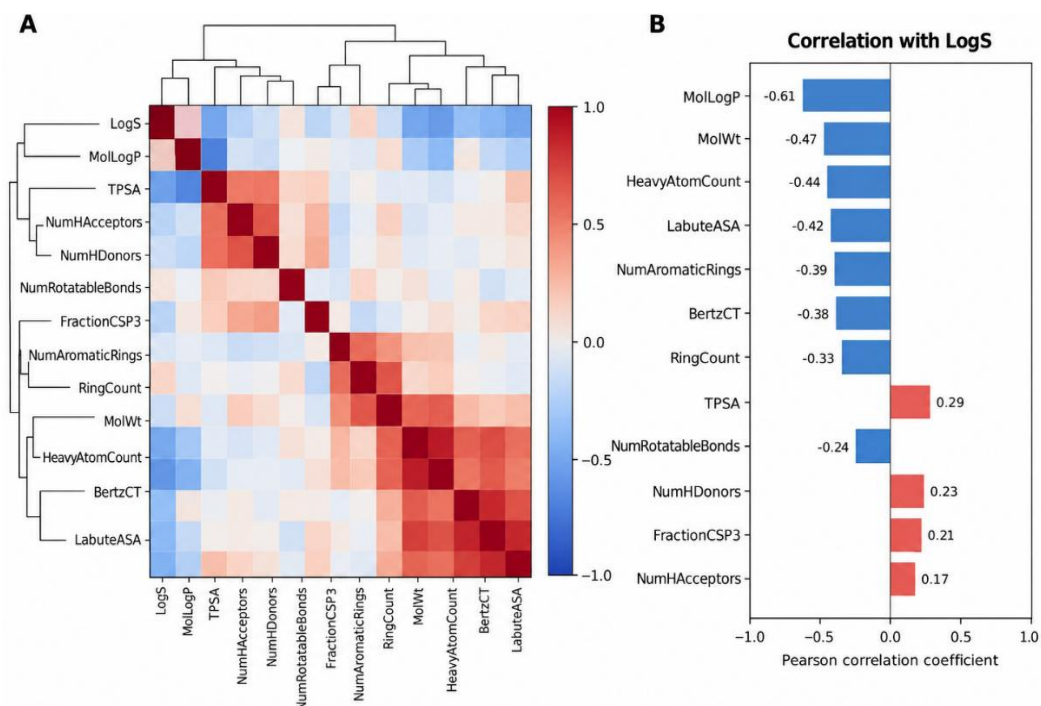


Figure 3. Multivariate Correlation Structure of Molecular Descriptors and Solubility

4.3 Comparative Performance of Machine-Learning Models

The six regression methods tested, all using repeated five-fold cross validation (5FCV), showed significant differences. The linear and latent-variable models performed the poorest in predicting. Ridge Regression achieved an R^2 of 0.612 ± 0.018 , while Partial Least Squares

produced an R^2 of 0.628 ± 0.017 .

Substantially improved performance was achieved by using the nonlinear approaches. Support Vector Regression achieved an R^2 of 0.742 ± 0.014 , Random Forest reached 0.756 ± 0.013 , and Gradient Boosting achieved 0.771 ± 0.012 . XGBoost provided the strongest overall performance, yielding an R^2 of 0.793 ± 0.011 , RMSE of 0.793 ± 0.019 , and MAE of 0.574 ± 0.015 .

Table 3. Cross-Validated Performance of Multivariate Machine-Learning Models

Model	R^2	RMSE	MAE
Ridge Regression	0.612 ± 0.018	1.086 ± 0.031	0.817 ± 0.024
Partial Least Squares	0.628 ± 0.017	1.063 ± 0.028	0.796 ± 0.021
Support Vector Regression	0.742 ± 0.014	0.884 ± 0.024	0.648 ± 0.018
Random Forest	0.756 ± 0.013	0.860 ± 0.022	0.629 ± 0.017
Gradient Boosting	0.771 ± 0.012	0.834 ± 0.021	0.611 ± 0.016
XGBoost	0.793 ± 0.011	0.793 ± 0.019	0.574 ± 0.015

The progression from linear to nonlinear ensemble methods suggests that there are more complex relationships between the descriptors when the descriptors are considered to be aqueous solubility variation.

4.4 Locked-Test Predictive Performance

www.aimpjournals.com

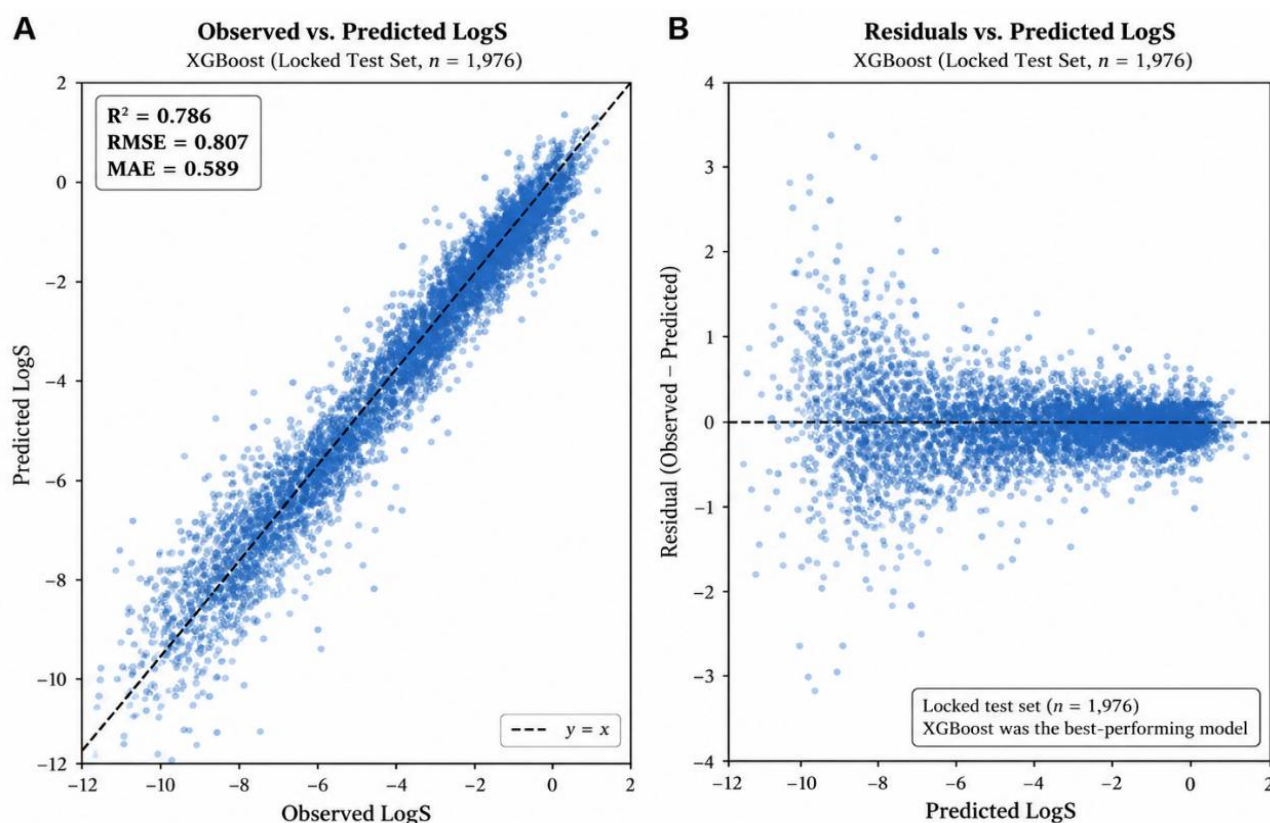
The independent set of 1,976 compounds was used to test the final models. The best performance was seen by XGBoost with a locked test R^2 of 0.786, RMSE of 0.807 and MAE of 0.589. The Gradient Boosting model had an R^2 of 0.764, Random Forest with an R^2 of 0.748 and Support Vector Regression with an R^2 of 0.731.

Table 4. Locked-Test Performance and Principal Descriptor Drivers of Solubility

Model	Test R ²	RMSE	MAE
SVR	0.731	0.905	0.668
Random Forest	0.748	0.875	0.641
Gradient Boosting	0.764	0.846	0.621
XGBoost	0.786	0.807	0.589

The slight drop in R² from cross validation to test (0.793 to 0.786), suggested good generalization. Close agreement between the predicted and experimentally obtained LogS

values is seen throughout most of the test range in figure 4. Dispersion was mostly concentrated around zero, suggesting a higher prediction uncertainty for the molecules with low solubility at the extremes.

**Figure 4. Observed and Predicted Aqueous Solubility on the Locked Test Set**

4.5 Explainable Characterization of Molecular Property Drivers

A clear hierarchy of molecular descriptors contributing to the XGBoost predictions was found using SHAP analysis. The most important variable is MolLogP with a mean absolute SHAP value of 0.472, followed by MolWt (0.316), TPSA (0.244), NumAromaticRings (0.201) and BertzCT (0.174). The mean absolute SHAP values for these models were: FractionCSP3 0.148, HeavyAtomCount 0.132, NumHDonors 0.109, RingCount 0.096, and NumRotatableBonds 0.081.

Figure 5 showed that increase in MolLogP value was

related mainly to decrease in predicted LogS, thus lipophilicity being the major factor affecting lower aqueous solubility. Predictions were also driven towards lower solubility by increasing the number of aromatic rings, increasing the molecular complexity, and increasing the number of heavy atoms, with increasing numbers of all of these also decreasing solubility. In contrast, the higher the TPSA and hydrogen-bond donor capacity, the better the predicted LogS was in general, which correlates with more polar interactions with water. The effect of rotatable bonds (molecular flexibility) was nonlinear, while FractionCSP3 had a moderate positive contribution.

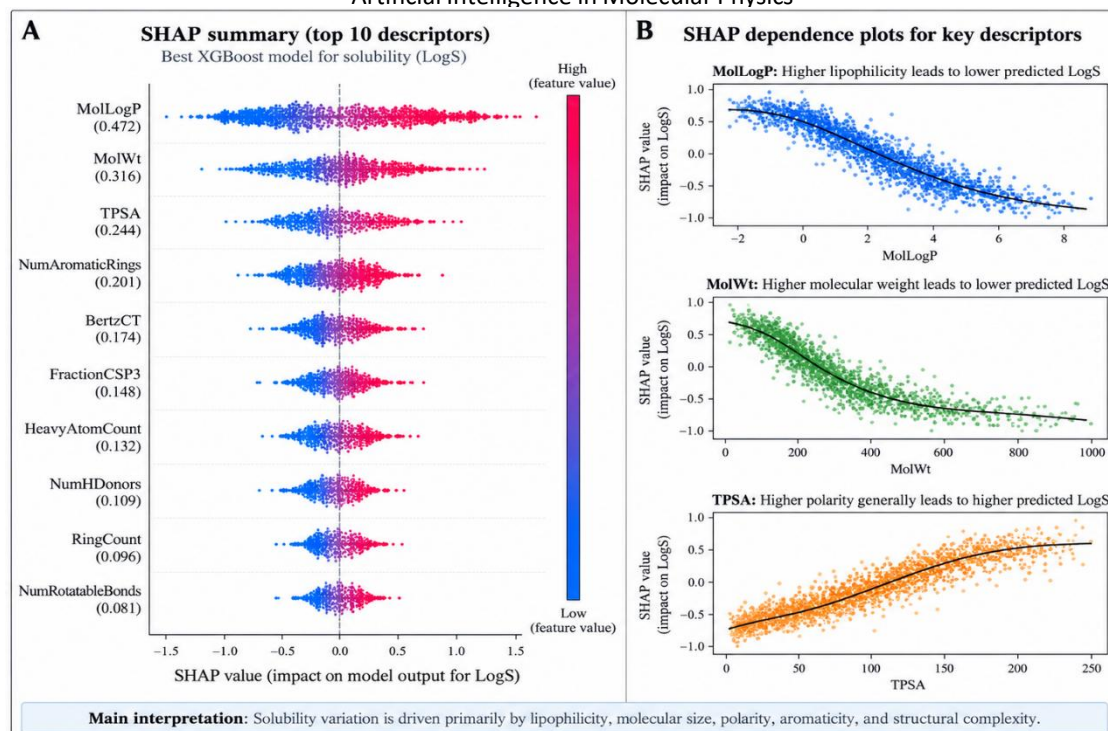


Figure 5. SHAP-Based Characterization of Molecular Descriptor Effects on Solubility

The overall explainability analysis indicated that molecular-property variation is the result of a combined effect of hydrophobicity, molecular size, polarity, aromaticity, hydrogen-bonding capacity, and molecular complexity. The convergence of the analysis of the agreement of correlation analysis, PCA, predictive modelling and SHAP interpretation led to the conclusion that the descriptor landscape of aqueous solubility is multidimensional and significantly nonlinear.

5. DISCUSSION

The results indicate that the variation of aqueous solubility is under the control of more than one single descriptor, but a combination of multidimensional physicochemical, structural and topological molecular properties. The broad LogS range from -11.60 to 1.58 revealed significant differences in the chemical space being analyzed, and the PCA revealed that the first five principal components were able to explain 72.8% of the total variance in the descriptors. The molecular size, hydrophobicity, and structural complexity of the molecules dominated the PC1, while the polarity and hydrogen-bonding properties dominated the PC2. This separation reflects the existence of different but coupled physicochemical dimensions in the molecular space, both of which had an impact on solubility.

The predictive results also confirmed the nonlinear relationships. XGBoost achieved the strongest cross-validated performance, with $R^2 = 0.793 \pm 0.011$, $RMSE = 0.793 \pm 0.019$, and $MAE = 0.574 \pm 0.015$, and maintained similar performance on the locked test set with $R^2 = 0.786$, $RMSE = 0.807$, and $MAE = 0.589$. The model was not overly dependent on the development partition, with minimal loss of accuracy between cross validation and

independent testing. The results of the nonlinear ensemble methods over Ridge Regression and Partial Least Squares are comparable to those of other QSPR studies which indicated that the treatment with multi-algorithm and nonlinear machine learning methods gave better results for the aqueous solubility parameter [12].

The descriptor-screening strategy is also consistent with results from previous descriptor-based thermodynamic-property studies, which highlight the need to consider proper descriptor preprocessing, feature construction, redundancy control and model validation to achieve stable predictive models [13]. In this present study, the number of descriptors was reduced from 208 descriptors to 58 descriptors that still retained the important chemical dimensions for the prediction. In addition, the observed residual widening for very low-solubility compounds further emphasizes the need for applicability-domain assessment, as it will be increasingly uncertain for structurally unusual or extreme compounds [14].

The main mechanistic interpretation for the model was obtained by using the SHAP approach. MolLogP was the most prominent descriptor followed by molecular weight, TPSA, aromatic-ring count and BertzCT. The greater the hydrophobicity (log P), molecular size, aromaticity, and structural complexity, the more likely the predictions were to be lower in LogS, while higher polarity and hydrogen-bonding capacity tended to lead to higher predicted solubility. It is the chemically interpretable nonlinear interactions between important molecular properties that provided the predictive advantage to XGBoost, not simply from a high dimensional structure. Even if the molecular-property models are learned automatically using transformer architecture, they can produce more complex

representations of molecules but may not be more intuitive for conventional molecular descriptors [15]. The present framework thus provides a complementary benefit, by making use of competitive nonlinear prediction and a specific chemical interpretation.

Results have implications for molecular screening and early-stage compound prioritization. Descriptor-level interpretation can be used to pinpoint the combinations of hydrophobicity, size, polarity and aromaticity that are correlated with bad aqueous solubility, ahead of potentially expensive experimental testing, and is a tool for property-driven molecular design.

Several limitations remain. Only 2D descriptors were used and no explicit 3D conformational information, crystal-state effects, pH, temperature, ionic strength or experimental protocol variation were included. The study also employed a single curated database of solubility, resulting in possible differences in the external domain of chemical problems. To go further, external validation, applicability-domain modelling, uncertainty quantification, 3D and quantum-derived descriptor, as well as comparisons with graph neural networks or transformer representations should be considered in future works. The incorporation of these methods in a descriptor analysis that can be interpreted may further enhance predictive robustness and mechanistic understanding of variation in molecular properties.

6. CONCLUSION

This study aimed to tackle the problem of aqueous-solubility variation by offering an approach that not only provides good predictive ability but also generates an interpretative analysis of the molecular descriptors. For this purpose, a multivariate machine-learning workflow was designed with AqSolDB, with molecular-structure validation, generation of molecular descriptors with RDKit, reduction of dimensionality, PCA, multiple regression algorithms, independent locked-test validation and explainability with SHAP. A total of 9,982 compounds were selected from the initial compounds, 9,881 of which could be analyzed, and 208 initial descriptors were narrowed down to informative variables of 58. The first five PCs accounted for 72.8% of the variance in the descriptors, indicating that molecular size, hydrophobicity, polarity, hydrogen bonding, aromaticity, and structural complexity were the most important factors in the variation of solubility. XGBoost gave the best predictive results (CV R2 = 0.793; LTV R2 = 0.786) among the evaluated ones with low RMSE and MAE. The top molecular drivers of LogS variation identified by SHAP, which were the most significant in their contribution to the model, were MolLogP, molecular weight, TPSA, aromatic-ring count, and BertzCT. The primary contribution is an integrated framework which goes beyond prediction to multivariate characterization of the behavior of molecular properties.

The main drawback is the use of 2D descriptors and a limited solubility database that has been manually curated. Future research should include external validation, uncertainty estimation, applicability-domain analysis, 3D descriptors, quantum descriptors and comparisons with graph neural networks and molecular representations based on transformer networks.

REFERENCES

1. K. V. Chuang, L. M. Gunsalus, and M. J. Keiser, "Learning molecular representations for medicinal chemistry," *J. Med. Chem.*, vol. 63, no. 16, pp. 8705–8722, 2020, doi: 10.1021/acs.jmedchem.0c00385.
2. A. J. Minnich et al., "AMPL: A data-driven modeling pipeline for drug discovery," *J. Chem. Inf. Model.*, vol. 60, no. 4, pp. 1955–1968, 2020, doi: 10.1021/acs.jcim.9b01053.
3. S. A. Alsenan, I. M. Al-Turaiki, and A. M. Hafez, "Feature extraction methods in quantitative structure–activity relationship modeling: A comparative study," *IEEE Access*, vol. 8, pp. 78737–78752, 2020, doi: 10.1109/ACCESS.2020.2990375.
4. G. Falcón-Cano, C. Molina, and M. Á. Cabrera-Pérez, "ADME prediction with KNIME: In silico aqueous solubility consensus model based on supervised recursive random forest approaches," *ADMET DMPK*, vol. 8, no. 3, pp. 251–273, 2020, doi: 10.5599/admet.852.
5. R. Rodríguez-Pérez and J. Bajorath, "Explainable machine learning for property predictions in compound optimization," *J. Med. Chem.*, vol. 64, no. 24, pp. 17744–17752, 2021, doi: 10.1021/acs.jmedchem.1c01789.
6. P. O. Dral and M. Barbatti, "Molecular excited states through a machine learning lens," *Nat. Rev. Chem.*, vol. 5, pp. 388–405, 2021, doi: 10.1038/s41570-021-00278-1.
7. J. Deng, Z. Yang, I. Ojima, D. Samaras, and F. Wang, "Artificial intelligence in drug discovery: Applications and techniques," *Brief. Bioinform.*, vol. 23, no. 1, Art. no. bbab430, 2022, doi: 10.1093/bib/bbab430.
8. J. Ross, B. Belgodere, V. Chenthamarakshan, I. Padhi, Y. Mroueh, and P. Das, "Large-scale chemical language representations capture molecular structure and properties," *Nat. Mach. Intell.*, vol. 4, pp. 1256–1264, 2022, doi: 10.1038/s42256-022-00580-7.
9. Y. Li et al., "An adaptive graph learning method for automated molecular interactions and properties predictions," *Nat. Mach. Intell.*, vol. 4, no. 7, pp. 645–651, 2022, doi: 10.1038/s42256-022-00501-8.
10. M. C. Sorkun, D. Mullaj, J. M. V. A. Koelman, and

- S. Er, "ChemPlot, a Python library for chemical space visualization," *Chemistry-Methods*, vol. 2, no. 7, Art. no. e202200005, 2022, doi: 10.1002/cmt.202200005.
11. M. McGibbon, S. Shave, J. Dong, Y. Gao, D. R. Houston, J. Xie, Y. Yang, P. Schwaller, and V. Blay, "From intuition to AI: Evolution of small molecule representations in drug discovery," *Brief. Bioinform.*, vol. 25, no. 1, Art. no. bbad422, 2024, doi: 10.1093/bib/bbad422.
 12. T. Zhu, Y. Chen, and C. Tao, "Multiple machine learning algorithms assisted QSPR models for aqueous solubility: Comprehensive assessment with CRITIC-TOPSIS," *Sci. Total Environ.*, vol. 857, Art. no. 159448, 2023, doi: 10.1016/j.scitotenv.2022.159448.
 13. C. Trinh, Y. Tbatou, S. Lasala, O. Herbinet, and D. Meimaroglou, "On the development of descriptor-based machine learning models for thermodynamic properties: Part 1—From data collection to model construction: Understanding of the methods and their effects," *Processes*, vol. 11, no. 12, Art. no. 3325, 2023, doi: 10.3390/pr11123325.
 14. C. Trinh, S. Lasala, O. Herbinet, and D. Meimaroglou, "On the development of descriptor-based machine learning models for thermodynamic properties: Part 2—Applicability domain and outliers," *Algorithms*, vol. 16, no. 12, Art. no. 573, 2023, doi: 10.3390/a16120573.
 15. A. Sultan, J. Sieg, M. Mathea, and A. Volkamer, "Transformers for molecular property prediction: Lessons learned from the past five years," *J. Chem. Inf. Model.*, vol. 64, no. 16, pp. 6259–6280, 2024, doi: 10.1021/acs.jcim.4c00747.
 16. S. Wang, Z. Li, S. Zhang, M. Jiang, X. Wang, and Z. Wei, "Molecular property prediction based on a multichannel substructure graph," *IEEE Access*, vol. 8, pp. 18601–18614, 2020, doi: 10.1109/ACCESS.2020.2968535.
 17. P. Li, J. Wang, Y. Qiao, H. Chen, Y. Yu, X. Yao, P. Gao, G. Xie, and S. Song, "An effective self-supervised framework for learning expressive molecular global representations to drug discovery," *Brief. Bioinform.*, vol. 22, Art. no. bbab109, 2021, doi: 10.1093/bib/bbab109.
 18. A. Xie, Z. Zhang, J. Guan, and S. Zhou, "Self-supervised learning with chemistry-aware fragmentation for effective molecular property prediction," *Brief. Bioinform.*, vol. 24, Art. no. bbad296, 2023, doi: 10.1093/bib/bbad296.
 19. J. Gao et al., "TransFoxMol: Predicting molecular property with focused attention," *Brief. Bioinform.*, vol. 24, Art. no. bbad306, 2023, doi: 10.1093/bib/bbad306.
 20. Z. Zheng, Y. Tan, H. Wang, S. Yu, T. Liu, and C. Liang, "CasANGCL: Pre-training and fine-tuning model based on cascaded attention network and graph contrastive learning for molecular property prediction," *Brief. Bioinform.*, vol. 24, Art. no. bbac566, 2023, doi: 10.1093/bib/bbac566.
 21. D. B. Catacutan, J. Alexander, A. Arnold, and J. M. Stokes, "Machine learning in preclinical drug discovery," *Nat. Chem. Biol.*, vol. 20, pp. 960–973, 2024.
 22. A. Shen, M. Yuan, Y. Ma, J. Du, and M. Wang, "Complementary multi-modality molecular self-supervised learning via non-overlapping masking for property prediction," *Brief. Bioinform.*, vol. 25, Art. no. bbae256, 2024, doi: 10.1093/bib/bbae256.
 23. M. C. Sorkun, A. Khetan, and S. Er, "AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds," *Sci. Data*, vol. 6, Art. no. 143, 2019, doi: 10.1038/s41597-019-0151-1.