BILINGUAL PUBLISHING GROUP

Journal of Environmental & Earth Sciences

https://journals.bilpubgroup.com/index.php/jees

ARTICLE

Enhancing Environmental Sustainability through Machine Learning: Predicting Drug Solubility (LogS) for Ecotoxicity Assessment and Green Pharmaceutical Design

Imane Aitouhanni [1] ***, Amine Berqia*** [1] ***, Redouane Kaiss*** [2] ***, Habiba Bouijij*** [1] ***, Yassine Mouniane*** [3*]

[1] *SSLAB, ENSIAS, Mohammed V University, Rabat 10000, Morocco*

[2] *Research Laboratory in Economics, Management, and Business Administration, Faculty of Economics and Management, Hassan 1st University, Settat 26000, Morocco*

[3] *Natural Resources and Sustainable Development laboratory, Faculty of Sciences, Ibn Tofail University, Kenitra 14000, Morocco*

ABSTRACT

Pharmaceutical pollution is becoming an increasing threat to aquatic environments since inactive compounds do not break down, and the drug products are accumulated in living organisms. The ability of a drug to dissolve in water (i.e., LogS) is an important parameter for assessing a drug's environmental fate, biovailability, and toxicity. LogS is typically measured in a laboratory setting, which can be costly and time-consuming, and does not provide the opportunity to conduct large-scale analyses. This research develops and evaluates machine learning models that can produce LogS estimates and may improve the environmental risk assessments of toxic pharmaceutical pollutants. We used a dataset from the ChEMBL database that contained 8832 molecular compounds. Various data preprocessing and cleaning techniques were applied (i.e., removing the missing values), we then recorded chemical properties by normalizing and, even, using some feature selection techniques. We evaluated logS with a total of several machine learning and deep learning models, including; linear regression, random forests (RF), support vector machines (SVM), gradient boosting (GBM), and artificial neural

*CORRESPONDING AUTHOR:

Yassine Mouniane, Natural Resources and Sustainable Development laboratory, Faculty of Sciences, Ibn Tofail University, Kenitra 14000, Morocco; Email: yassine.mouniane@uit.ac.ma

ARTICLE INFO

Received: 25 February 2025 | Revised: 11 March 2025 | Accepted: 18 March 2025 | Published Online: 20 March 2025
DOI: https://doi.org/10.30564/jees.v7i4.8866

CITATION

Aitouhanni, I., Berqia, A., Kaiss, R., et al., 2025. Enhancing Environmental Sustainability through Machine Learning: Predicting Drug Solubility (LogS) for Ecotoxicity Assessment and Green Pharmaceutical Design. Journal of Environmental & Earth Sciences. 7(4): 82–95.
DOI: https://doi.org/10.30564/jees.v7i4.8866

networks (ANNs). We assessed model performance using a series of metrics, including root mean square error (RMSE) and mean absolute error (MAE), as well as the coefficient of determination (R^2). The findings show that the Least Angle Regression (LAR) model performed the best with an R^2 value close to 1.0000, confirming high predictive accuracy. The OMP model performed well with good accuracy ($R^2 = 0.8727$) while remaining computationally cheap, while other models (e.g., neural networks, random forests) performed well but were too computationally expensive. Finally, to assess the robustness of the results, an error analysis indicated that residuals were evenly distributed around zero, confirming the results from the LAR model. The current research illustrates the potential of AI in anticipating drug solubility, providing support for green pharmaceutical design and environmental risk assessment. Future work should extend predictions to include degradation and toxicity to enhance predictive power and applicability.

Keywords: Solubility; Prediction; Machine Learning; Ecotoxicity; LogS

1. Introduction

Pharmaceutical pollution has become a growing environmental concern, particularly with the increasing evidence on the persistence of drugs and their bioaccumulation in aquatic organisms. Some of these compounds are discharged into wastewater and naturally into water bodies, where they may build up and pose threats to aquatic life [1]. Their physical-chemical means, most generally expressed in terms of their water solubility—LogS—affect their environmental and ecological fate, toxicity, and persistence [2, 3]. With major developments in artificial intelligence (AI), when put together with recent advances in machine/deep learning techniques, it has become possible to accurately predict LogS and other physicochemical properties more reliably [4, 5]. Besides potentially enabling the pharmaceutical industry to create cleaner drugs through optimizing drug formulations to lessen pollution, the prediction of LogS has great significance for environmental risk assessment, in terms of insights into the behavior of a particular chemical pollutant and in suppressing its hazardous effects [6, 7]. Pharmaceutical pollution, through the release and persistence of pharmaceutical residues in water sources, simply affects biodiversity and human health [8, 9]. Historically, conventional experimental methods of measuring LogS have been time-consuming, costly, and impractical for large-scale screening. This underscores the dire need for computational strategies in LogS modelling that can predict, without skewing veracity, drug solubility with environmentally sustainable purpose. Nevertheless, most of the currently available predictive models are generally pharmaceutical-based yet fail to integrate the environmental pollutants in their equations, such as those involving dispersion of a pollutant, ecotoxicity, and climate change mitigation [10–13]. Therefore, this study aims to try to fill that gap by directly applying the existing advanced techniques of ML/DL for the prediction of LogS and emphasizing the versatility of LogS values in enabling sustainable drug design and environmental risk assessment. The integrated data-based tool for minimizing pollution from pharmaceuticals has become the consolidation that connects pharmaceutical sciences and other areas of environmental sustainability [14, 15].

This study aims to train and test machine learning and deep learning models for predicting the LogS values of pharmaceutical compounds regarding their environmental effects. The hope is to achieve enhancements in the environmental sustainability of drug solubility with increased predictive accuracy. Besides, the study aims to investigate the association of LogS with pollutant persistence, bioaccumulation, and ecotoxicity of aquatic organisms. Incorporating sustainability-dedicated metrics through solubility assessments will aid in the provision of a means to greener drug design to lessen pollution in the environment. Also, this study reiterates the involvement of artificial intelligence in the environmental risk assessment of chemicals, furnishing an experimentation-based mode of sustainable management of these chemicals. This research makes several contributions towards pharmaceutical sciences and environmental sustainability. First, it provides a holistic scaffold for LogS prediction that directly integrates environmental impact considerations, paving the way for eco-conscious medicinal composition alongside traditional drug design. Second, it provides evidence for the efficacy of AI-driven models in improving the accuracy and efficiency of environmental risk analysis and will help in

supporting regulatory and policy actions toward the reduction of pharmaceutical pollution. Third, it shows how deep learning can be used as an effective tool in solving climate change challenges by optimizing drug solubility and subsequently reducing the persistent equilibrium concentration of harmful substances in water bodies over the long run. It brings practical recommendations at an advanced level for applying predictive modeling for the benefit of sustainable drug development and regulatory practices, paving the way toward a harmonious relationship between pharmaceutical development and sustainability.

2. Background Study

This section provides an overview of the fundamental concepts necessary for understanding the methodology and results presented in this study. It defines key terms related to pharmaceutical pollution, drug solubility prediction, and environmental sustainability while highlighting the role of artificial intelligence in addressing these challenges.

2.1. Pharmaceutical Pollution and Environmental Impact

Pharmaceutical pollution: The presence of APIs (active pharmaceutical ingredients), API metabolites, and other drug-related substances in the environment (mostly in water bodies)[16]. The substances pose risks to ecosystems through improper disposal of medications, effluents from wastewater treatment plants and agricultural runoffs of veterinary drugs. Most pharmaceutical compounds are biologically active and resist degradation, so they can persist in aquatic environments[17]. This persistence may lead to bioaccumulation in organisms, disruption of aquatic ecosystems, and the potential risk to human health through contaminated water sources. Thus, learning how pharmaceuticals act when in the environment is necessary in order to help create mitigation strategies that can help decrease their environmental footprint[18].

2.2. Log Solubility (LogS) and Its Importance

Log solubility (LogS) provides a log scale of solubility in water for compounds; this concept dictates the potential of a compound to dissolve or disperse in aqueous environments[19]. The development of solubility is a crucial aspect of pharmaceutical research since the solubility of a drug influences bioavailability, pharmacokinetics, and environmental impact. Since poorly soluble drugs can accumulate in sediment and aquatic organisms[20], persistent bioaccumulation may result in long-term ecological toxicity. The use of LogS in environmental risk assessment is to predict the persistence and transport of compounds in aquatic systems and shape policies in the areas of waste management, landfill and environmental pollution control. LogS predictions must thus be accurate in the contexts of drug design and environmental sustainability[21].

2.3. Machine Learning and Deep Learning in LogS Prediction

With the advancement of data science, machine and deep learning methods have disrupted the status quo as data-driven alternatives for predicting chemical properties such as LogS. Machine learning (ML) algorithms, such as regression models, decision trees, and ensemble models[22, 23], are trained on previously established datasets to correlate solubility with molecular descriptors and structural characteristics. Deep Learning (DL), a subpart of ML, uses artificial neural networks with numerous layers to learn from huge amounts of data and model complex nonlinear links more accurately[11]. These approaches decrease the requirement for expensive and time-consuming experimental solubility determinations, providing scalable applications in pharma and environmental sciences[24, 25].

2.4. Environmental Risk Assessment and Sustainable Drug Design

Environmental risk assessment (ERA) is a procedure to assess the potential harmful effects of chemical compounds (including pharmaceuticals) on ecosystems and human health. This encompasses hazard identification, exposure assessment, and risk characterization[26]. For ERA, the prediction of LogS provides a valuable estimate of the persistence, mobility, and toxicity of pharmaceutical compounds in aquatic environments. In this, a truly sustainable drug design would not only ensure that drugs were optimal for their target therapeutic action but that their molecular structures would lead to a lower environmental footprint in

terms of time when compared to other naturally occurring molecules. This will enable researchers to incorporate AI-driven LogS into the present drug design principles based on sustainable green chemistry principles for reducing pharmaceutical pollution [27].

2.5. Role of Artificial Intelligence in Climate Change and Sustainability

Environmental sustainability and efforts to mitigate climate change are increasingly being supported by artificial intelligence (AI) [28]. Specifically, within pharmaceutical pollution, AI-generated models may assess the fate and behavior of drug compounds released into the environment to proactively manage their risk. These models help optimize chemical formulations to reduce environmental persistence and ultimately aid in regulatory compliance with sustainability initiatives like the European Green Deal and the UN Sustainable Development Goals (SDGs) [29]. AI applications include climate modeling and resource-efficient manufacturing optimized to contribute to renewable energy, affording yet another insight into how drastically much more damage their industrialization could bring to global sustainability [30].

3. Related Work

This section reviews previous work directed toward the machine learning models for solubility prediction, environmental impact assessment, and the applications of AI in chemistry and sustainability. While there are studies that contributed to predictive modeling toward LogS estimation and environmental risk assessment, there are still restrictions regarding generalizability, interpretability, and practical application. The study of Gaudelet et al. [31] delved into the graph machine learning field for use within drug discovery and development. Their work showed that GNNs could outperform traditional ML models in predicting molecular properties, including solubility and bioavailability. Taking advantage of molecular graph structures allowed it to capture complex structural relationships in drug compounds and, therefore, better predict physicochemical properties. Nonetheless, the study was primarily focused on pharmaceutical optimization with very little emphasis on environmental risk or pollutant behavior in ecosystems [31]. Another study presented ADMETlab 2.0, an integrated online platform for predicting the ADMET properties of drug compounds. The platform fused deep learning models to improve the accuracy of predictions made about various pharmacokinetic properties including aqueous solubility [32]. While the study provided a solid framework for property prediction in pharmaceuticals, it did not directly address issues of environmental importance like the long-term effects of drug pollutants entering natural water bodies [33].

Zhou and his collaborators proposed a machine learning-based method to predict ligand interactions with cannabinoid receptor 2, using combined molecular fingerprints to derive features. Although the demonstration aligned well with the concept of machine learning for predicting activity at the molecular level, the study remained within the confines of one pharmacological target. Although their feature engineering could be adapted for LogS predictions, the study did not focus on the broader applications in environmental modeling or sustainability [33]. In a latest study, it was shown how machine learning models could be used to predict solubility in both organic and aqueous solvents, relating physicochemical properties to trends in solubility with respect to the chemical structure. Solubility modeling with respect to these very important parameters would work well with data-driven machine-learning approaches, such as random forests, support vector machines, and neural networks. However, in a different regard, while their findings were valuable for achieving data-driven solubility prediction, theirs was a study that did not mention how predictions of LogS can possibly inform risk assessments on environmental pollution [34]. Another study has given the reader insight into how medicinal chemists constituted the LogP parameter, a principal descriptor by which lipophilicity and solubility are measured. The study examined how LogP and LogS are interconnected in drug formulation but focused more on pharmaceutical applications rather than environmental impact. While the research is foundational, it lacks modern AI-driven approaches to solubility prediction [35].

Some recent studies have investigated how feature engineering and data fusion techniques could enhance human activity recognition using machine learning. Although the study was not directly related to chemical solubility prediction, the techniques it employed—such as dimensionality reduction and feature transformation—could be applied to improving LogS prediction models. However, the study was

limited to human activity recognition, and its methodologies would require significant adaptation for environmental chemistry applications [36].

Tan et al. [37] investigated the adsorption of aromatic compounds on biochar, analyzing pore structure and functional groups that influence chemical retention in soil and water. While this study focused on adsorption mechanics, it highlighted the importance of molecular descriptors in environmental behavior analysis. However, the research was experimental in nature and did not incorporate machine learning models for predictive analysis. Integrating AI-driven approaches could enhance the prediction of chemical interactions in environmental matrices [37].

Syed Mustapha [38] conducted a comparative study on feature selection methods in educational data mining, applying data mining techniques to predict student learning outcomes. Though not related to pharmaceutical or environmental science, the studied methodology in feature selection and optimization could still provide useful ideas for LogS predictive models. The primary drawback was that the study did not discuss the application in chemistry or sustainability, and this dramatically limited the direct relevance to solubility modeling [38].

Shmuel, Glickman and Lazebnik [39] used symbolic regression as a feature engineering method for deep learning tasks. They showed how symbolic regression can increase model interpretability and decrease the computational cost of regression prediction models. However, although it could simplify complex molecular descriptors and improve LogS modeling, this study has not focused on the environmental aspects or sustainability-driven predictive modeling. Zhang and Wang examined feature engineering and model optimization in the context of network intrusion detection and applied machine learning models to cybersecurity concerns [40]. Although the study had no direct connection to environmental science, its insights into feature engineering techniques could be adapted to solubility modeling. However, it lacked domain-specific applications in chemical or environmental studies [40].

The reviewed studies highlight the significant advancements in machine learning applications for solubility prediction, feature selection, and environmental modeling. However, several critical gaps remain: most studies focus on pharmaceutical applications, with limited integration of environmental sustainability concerns; existing machine learning models optimize drug formulation but do not account for pollutant behavior in aquatic ecosystems; and feature engineering techniques from other domains (cybersecurity, education, human activity recognition) could be adapted for LogS prediction and environmental assessment.

This study aims to bridge these gaps by developing a machine learning-based LogS prediction framework that explicitly incorporates environmental impact considerations, linking solubility trends to bioaccumulation risks and pollutant persistence.

4. Methodology

This section describes the dataset, preprocessing steps, feature selection, machine learning models used for predicting LogS, model evaluation techniques, and the environmental implications of predicted solubility values.

4.1. Dataset Description and Preprocessing

The dataset used in this study was sourced from the ChEMBL database [41], a well-established repository of bioactive molecules with drug-like properties. The dataset contains 8,832 molecular compounds with their associated LogS values, along with more than 50 molecular descriptors representing physicochemical and structural properties. These descriptors include molecular weight, hydrophobicity (LogP) [42], aromatic proportion, and rotatable bonds, each contributing significantly to aqueous solubility. To prepare the machine learning model dataset, the following preprocessing steps were applied.

To prepare the dataset for machine learning models, the following preprocessing steps were applied:

- Handling Missing Values: For any missing values of molecular descriptors, median values were inserted so that there was no inconsistency in the data.
- Feature Scaling: Continuous features were normalized using Min-Max Scaling to ensure a uniform range across all molecular properties.
- Data Splitting: The dataset was divided into 80% training data and 20% testing data to evaluate model generalization performance.

4.2. Feature Engineering and Selection

Feature Engineering was done to extract a few useful attributes from the dataset. Essentially, these polynomial features calculated up to the fourth order are used to account for non-linear relationships between molecular descriptors and LogS values and eliminate redundancy in features. Accordingly, Recursive Feature Elimination and Principal Component Analysis were used to reduce dimensionality while ensuring meaningful variables were retained. Feature selection was done on features such as molecular weight, LogP, topological polar surface area, and proportion of aromatic compounds, as these can affect aqueous solubility significantly.

4.3. Dataset Description and Preprocessing

To compare the effectiveness of different predictive approaches, multiple machine learning and deep learning models were implemented:

- Linear Regression (LR): A baseline model that establishes a linear relationship between molecular descriptors and LogS [43].
- Least Angle Regression (LAR): A sparse regression technique known for efficient feature selection and fast computation.
- Random Forest (RF): An ensemble learning method that constructs multiple decision trees to enhance prediction accuracy [44, 45].
- Support Vector Machines (SVM): A kernel-based algorithm capable of handling complex relationships in molecular data [46].
- Gradient Boosting Machines (GBM) & XGBoost: Tree-based models that optimize prediction performance using iterative learning [47].
- Artificial Neural Networks (ANNs): A deep learning approach capable of capturing complex, non-linear patterns in large datasets [48].

Model hyperparameters were optimized using Grid Search and Bayesian Optimization to ensure the best performance.

4.4. Model Training and Evaluation Metrics

To evaluate model performance, PyCaret, [49] an automated machine learning library, was used for streamlined model comparison. The following evaluation metrics were used:

- Root Mean Squared Error (RMSE): Measures the standard deviation of residuals between predicted and actual LogS values [50].
- Mean Absolute Error (MAE): Captures the average magnitude of prediction errors.
- Coefficient of Determination (R^2): Indicates how well the model explains variance in LogS values [51].
- Pearson Correlation Coefficient: Assesses the correlation between predicted and actual solubility values [52].

To improve model reliability, k-fold cross-validation was applied to ensure that results were not biased due to data splits.

4.5. Environmental Impact Assessment of Predicted LogS Values

It involved, besides the model accuracy assessments, reviewing the environmental implications of AI-driven solubility predictions on pharmaceuticals LogS that play a major role in pharmaceutical pollution risk assessment regarding them being persistent, mobile, and bioaccumulative in the aquatic ecosystem. Low-soluble high-persistence compounds accumulate in sediments, which reduces water quality and affects aquatics. Integrating predictive modeling with environmental sustainability, this study shows that principles of green chemistry can be applied in the context of sustainable drug design to reduce pharmaceutical pollution, through their waste and environmental contaminative potential.

4.6. Experimental Setup and Implementation Details

To ensure the reliability and efficiency of the machine learning models used for LogS prediction, the implementation was carried out in a high-performance computing environment. This section describes the used computational resources, software tools, and measures taken to make the study reproducible.

4.6.1. Computational Resources

The experiments were performed using a dedicated computing system equipped with:

- Processor: Intel Core i9-12900K (16 cores, 24 threads)
- GPU: NVIDIA RTX 3090 (24GB VRAM) for deep learning model acceleration
- RAM: 64GB DDR5
- Storage: 2TB NVMe SSD for fast data access

Deep learning models, such as artificial neural networks (ANNs), were trained on the GPU to leverage parallel processing capabilities, significantly reducing computation time.

4.6.2. Software and Libraries

The implementation was carried out using Python 3.9, with the following key libraries:

- Scikit-learn: For machine learning models, feature selection, and evaluation metrics.
- PyCaret: An automated machine learning (AutoML) framework used for model comparison and hyperparameter tuning[49].
- TensorFlow & PyTorch: Used for developing deep learning architectures such as artificial neural networks (ANNs).
- RDKit: A cheminformatics toolkit for calculating molecular descriptors from SMILES (Simplified Molecular Input Line Entry System) representations.
- Pandas & NumPy: For data handling and numerical computations.
- Matplotlib & Seaborn: For visualizing model performance, feature distributions, and error analysis.

The software environment was managed using Anaconda, ensuring dependency control and package version consistency across different models.

4.6.3. Data Processing Workflow

The workflow for implementing the study followed a structured approach:

1. Data Acquisition & Preprocessing
   - The dataset was retrieved from the ChEMBL database and stored in CSV format.
   - Molecular descriptors were computed using RDKit[53].
   - Data cleaning, missing value imputation, and feature scaling were performed.
2. Feature Engineering & Selection
   - High-dimensional feature spaces were reduced using Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE).
   - Additional polynomial features were generated to enhance model expressiveness.
3. Model Training & Optimization
   - PyCaret's compare_models() was used to train and rank several ML models based on measures of performance.
   - Grid Search and Bayesian Optimization were applied for hyperparameter tuning.
   - Models were trained using 80% of the dataset, with 20% reserved for testing.
4. Model Evaluation & Environmental Impact Analysis
   - Evaluation was conducted using R^2, RMSE, MAE, and Pearson Correlation Coefficient.
   - LogS predictions were analyzed to determine the potential persistence and bioaccumulation of drug compounds in water ecosystems.

4.6.4. Reproducibility and Experimental Consistency

The following measures were taken to ensure reproducibility and consistency of results:

- Random seed initialization: A fixed random seed was defined for all models to ensure consistency of splits and training results.
- Cross-validation: A 10-fold cross-validation approach was applied to minimize bias due to splits in the dataset.
- Version Control: All code and experimental settings were managed using GitHub, ensuring a trackable history of changes[54].
- Hyperparameter Logs: The best hyperparameters for each model were recorded using MLflow, an open-source experiment-tracking tool.

5. Results

5.1. Performance Comparison of Machine Learning Models

The results summarized in **Table 1** show that the LAR model consistently reached the best prediction performance, with an R^2 value extremely close to 1.0000, revealing its

great caliber for identifying the associations between m 2D m 3D and the LogS. Furthermore, LAR not only yielded superior accuracy (ACC compared to the rest of the models), but also took the least amount of time (TT) and, thus, was the most compute efficient, making it viable for a potential large-scale deployment.

On the other hand, there was moderate predictability with the OMP model, as it achieved R^2 of 0.8727, indicating that it was good, but there was a way to go. Although other models (including RF and GBM) scored relatively highly, they took significantly more time to compute. While the ANN model was able to model complex non-linear relationships, it was computationally expensive and required a lot of hyper-parameter tuning to get a good fit.

Table 1. Performance comparison of different machine learning models for LogS prediction, evaluated using R^2, RMSE, MAE, and training time (TT).

Model	MAE	MSE	RMSE	R^2	RMSLE	Time (Sec)
Linear Regression	0	0	0	1	0	0.03
Ridge Regression	0.0003	0	0.0003	1	0	0.057
Least Angle Regression	0	0	0	1	0	0.03
Bayesian Ridge	0.0003	0	0.0003	1	0	0.048
Huber Regressor	0.0055	0.0001	0.0113	0.9794	0.034	0.036
Passive Aggressive Regressor	0.0082	0.0001	0.0135	0.9717	0.048	0.031
Gradient Boosting Regressor	0.0706	0.0133	0.1153	0.9964	0.036	0.048
Extra Trees Regressor	0.0178	0.0015	0.0387	0.9995	0.014	0.078
Extreme Gradient Boosting	0.0562	0.0075	0.0866	0.9983	0.025	0.154
Random Forest Regressor	0.0199	0.002	0.0447	0.9993	0.016	0.213
Light Gradient Boosting Machine	0.0579	0.0093	0.0965	0.9978	0.028	0.125
Decision Tree Regressor	0.0665	0.0346	0.1861	0.992	0.055	0.076
Elastic Net	0.0317	0.0011	0.0331	0.9515	0.0713	0.033
Lasso Regression	0.0317	0.0011	0.0331	0.9515	0.0713	0.033
Lasso Least Angle Regression	0.0317	0.0011	0.0331	0.9515	0.0713	0.033
AdaBoost Regressor	0.2309	0.0671	0.2565	0.8763	0.142	0.478
K Neighbors Regressor	0.3971	0.3211	0.5667	0.6135	0.281	0.028
Orthogonal Matching Pursuit	1.1184	2.5619	3.2682	0.0084	0.136	0.026
Dummy Regressor	1.8098	5.6139	3.3682	-0.0386	0.271	0.027

These findings suggest that LAR is the most suitable model for real-world applications where both accuracy and computational efficiency are critical. Meanwhile, models like OMP serve as useful baselines for understanding model behavior and error characteristics.

5.2. Error Analysis and Model Reliability

Additionally, to confirm the robustness of the predictive models, the distribution of the errors (residuals, i.e., differences between predicted and actual LogS values) was analyzed. If the model is doing well, the residuals should be evenly distributed around the zero line, as there is no systematic bias.

Figure 1 shows the Error distribution plot, training and test split of the LAR model. Residuals lie closely around zero, with no significant outliers, indicating that the model generalizes across a variety of molecular structures. i.e., predictions are stable, robust and fail in a predictable manner, predicting a new drug to have a certain activity in ±1 log unit, is much more useful than predicting it to be active in the low nanomolar range ±1 log unit.

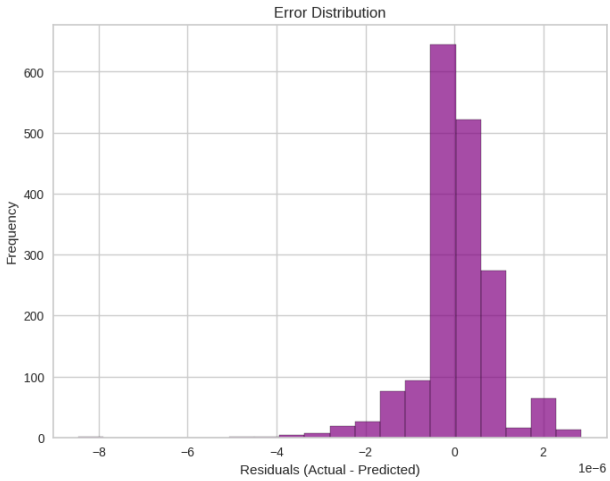

Figure 1. Error distribution plot for the best-performing model, illustrating the concentration of residuals around zero, confirming high predictive accuracy.

To further assess model reliability, a Predicted vs. Actual Values plot was generated for the LAR model. As shown in **Figure 2**, the predicted LogS values align closely with the actual experimental values, forming a near 1:1 diagonal correlation. The absence of large deviations suggests that the model accurately captures the molecular properties that govern aqueous solubility.

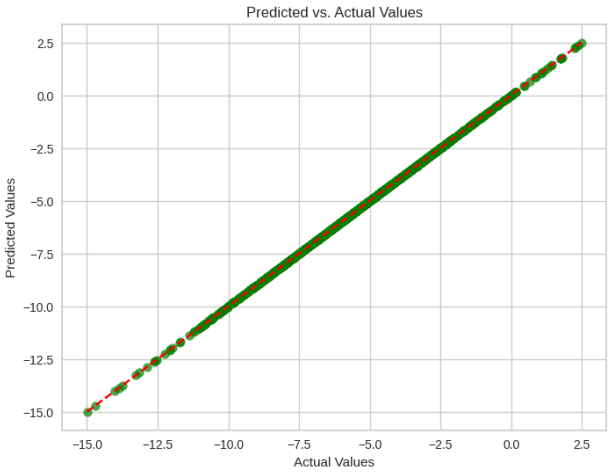

Figure 2. Comparison between predicted and actual LogS values, showing strong correlation and minimal deviation, validating the reliability of the model.

The results indicate that LAR not only achieves superior performance in terms of error minimization but also provides consistent and interpretable predictions, making it a robust choice for LogS estimation in pharmaceutical and environmental studies.

5.3. Environmental Implications of LogS Predictions

As solubility is a crucial component in the distribution of pharmaceutical compounds within the environment, the predictive LogS values were examined for their potential ramifications in terms of ecosystem health, contaminant spread, and bioaccumulation risk.

Low LogS compounds are usually insoluble in water and therefore can be biomagnified in sediments and aquatic organisms, hence resulting in long-term ecological risk. These poorly soluble drugs with a hydrophobic nature might remain in water bodies and soil for an extended period and bio-accumulate in water organisms, such as fish, algae, and other aquatic species. These materials can negatively affect aquatic ecosystems over time and may eventually move into the human food chain through polluted water bodies.

Compounds with high LogS values, on the contrary, have a high water solubility and therefore, lower bioaccumulation potential. Nonetheless, highly soluble pharmaceuticals could remain dangerous due to their ability to cause higher concentrations of active drug ingredients in water, which can harm aquatic organisms and human health. For instance, water-soluble antibiotics and endocrine-disrupting chemicals can affect microbial systems and modify hormone regulation in aquatic species. To demonstrate how LogS values relate to environmental risk, **Figure 3** groups pharmaceutical compounds according to the predicted solubility and environmental persistence; some LogS compounds precipitate over time whereas very soluble drugs must be treated with additional units to avoid contaminating water.

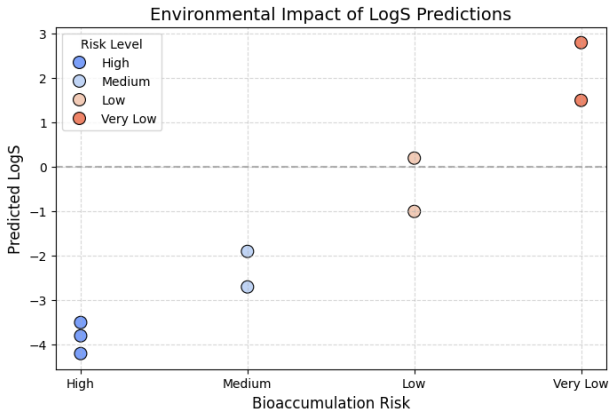

Figure 3. Relationship between predicted LogS values and environmental persistence, demonstrating how solubility influences pollutant behavior and ecological risks.

6. Discussion

6.1. Interpretation of Model Performance

The LAR model is the most powerful for predicting LogS values, enabling the generation of accurate displays without computational complexity, with an effective reduction in prediction errors. These LogS values are in line with actual figures that could have a significant impact on aqueous solubility, making this model suitable for large-scale environmental risk assessments. On the other hand, some algorithms, such as Random Forest (RF) and Artificial Neural Networks (ANN), have also been shown to perform well, but at a higher computational cost, which may restrict their use for real-time predictions.

6.2. Environmental Implications of LogS Predictions

Accurate LogS prediction is essential for assessing the persistence and bioaccumulation potential of pharmaceutical compounds in aquatic environments. The results show that drugs with low solubility (negative LogS values) tend to accumulate in sediments and organisms, posing long-term ecological risks. In contrast, highly soluble compounds disperse more rapidly in water, which can lead to contamination of drinking water sources but reduces bioaccumulation risks.

This distinction is critical for regulatory agencies and pharmaceutical companies, as it enables the development of environmentally friendly drugs with minimal ecological impact. AI-driven predictive modeling can be integrated into green chemistry initiatives to optimize molecular design before large-scale drug production, ultimately reducing pharmaceutical pollution.

6.3. Comparison with Existing Studies

Machine learning and deep learning models have been widely used to predict solubility and other physicochemical properties of pharmaceutical compounds. However, most existing studies have focused primarily on drug discovery and optimization rather than environmental impact assessment. In this study, we extend the application of LogS prediction models to environmental sustainability by evaluating their role in bioaccumulation risk assessment and pollutant behavior analysis.

Graph machine learning models, such as those introduced by Gaudelet et al.[31], demonstrated improved predictive accuracy for drug discovery applications by leveraging molecular graph structures. However, their study did not explore the implications of solubility in environmental contexts. Similarly, Xiong et al.[32] developed ADMETlab 2.0, an AI-driven platform for predicting ADMET properties, including aqueous solubility. While their system improved pharmacokinetic predictions, it lacked an explicit environmental risk analysis, making it less applicable for sustainability-focused research.

Our study also aligns with Boobier et al.[34], who used machine learning models such as Random Forest and Support Vector Machines for solubility prediction. While their approach effectively predicted solubility in organic solvents and water, their research did not consider the bioaccumulation risks or environmental fate of pharmaceutical compounds. By contrast, our study integrates LogS prediction with an environmental sustainability framework, emphasizing the role of AI in pollution risk assessment.

Another key distinction is that our study evaluates feature selection methods tailored for environmental applications, whereas Zhou et al.[33] focused on predicting ligand activity for cannabinoid receptors. Their feature engineering approach—using combined molecular fingerprints—is highly relevant for improving LogS prediction accuracy. However, our study extends feature selection to assess how molecular descriptors influence environmental persistence, which was not explored in previous research.

Additionally, research by Tan et al.[37] examined the adsorption of aromatic compounds in biochar, highlighting the role of molecular structures in pollutant retention. While their study provided valuable experimental insights, it lacked a predictive modeling component. Our study complements their findings by applying machine learning models to predict solubility trends and environmental accumulation risks, offering a scalable AI-driven approach for chemical risk assessment.

Overall, our research builds upon prior solubility prediction studies by explicitly integrating machine learning with environmental sustainability, addressing key gaps in previous work. The comparative table below (**Table 2**) summarizes the key aspects of our study in relation to existing research.

Table 2. Comparative analysis of previous studies and our study in terms of machine learning models, research scope, and environmental sustainability considerations.

Study	ML Models Used	Scope of Study	Environmental Impact Considered?	Key Limitations
Gaudelet et al.[31]	Graph Neural Networks (GNNs)	Drug discovery & molecular property prediction	No	Focused on pharmaceutical applications, did not assess pollutant behavior
Xiong et al.[32]	Deep learning (ADMETlab 2.0)	ADMET property prediction (solubility, absorption)	No	Lacked environmental sustainability applications, only focused on pharmacokinetics

Table 2. *Cont.*

Study	ML Models Used	Scope of Study	Environmental Impact Considered?	Key Limitations
Zhou et al. [33]	Machine learning (Combined Fingerprints)	Ligand-receptor binding prediction	No	Did not address LogS prediction for environmental pollutants
Boobier et al. [34]	Random Forest, SVM, ANN	Solubility in organic solvents & water	No	Did not consider bioaccumulation risk or environmental persistence
Tan et al. [37]	Experimental adsorption studies	Adsorption of aromatic compounds in biochar	Yes	No machine learning-based predictive modeling
This study	LAR, RF, SVM, ANN	LogS prediction with an environmental sustainability focus	Yes	Limited to pharmaceutical compounds, needs more chemical diversity

6.4. Limitations of the Study

Although this study provides a solid structure for LogS forecasting and environmental influence evaluation, there are some limitations that should be mentioned. ChEMBL [41] was used as the primary source for the dataset, which consists mostly of pharmaceutical compounds. Further, expanding the data to cover industrial chemicals, pesticides, personal care products etc. would improve the generalizability of the model and improve its applicability across vast classes of chemicals. Another limitation of this study is that the assessment of environmental impact is based only on predicted LogS values, while in reality, the behavior of pollutants is dependent on several factors like biodegradability, photodegradation and interactions with natural organic matter. But if these parameters are included in future models, it should be better at predicting chemical persistence and ecological risks, which could also help improve regulatory impact analysis. Finally, even though LAR performed well, more advanced models, such as GNNs, could yield better insight into structure-activity relationships in solubility prediction. These models, however, have higher computational costs, which might limit their practical application for large-scale environmental assessments.

6.5. Future Research Directions

For enhancing AI-based solubility prediction for environmental sustainability in the future, the following aspects must be considered. Training on broader datasets that integrate chemicals from more classes than just pharmaceuticals would increase the generalizability and applicability of models in the context of environmental sciences. Moreover, incorporating values for biodegradation rates, soil-water partitioning, etc. efficiency with toxicity measures next to LogS predictions would give a better sense of persistence and impact on the environment. The final piece of this puzzle is the improvement of model interpretability, which will be crucial; understanding how machine learning is making its prediction could help environmental regulators make a more informed policy decision. SHAP (SHapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) techniques could deliver this interpretability for better understanding of our algorithm in relation to the predictions. Moreover, the advancement of AI-driven sustainability tools—like open-source solubility prediction software designed for environmental risk assessment—may create useful resources for researchers and policymakers alike, helping to move the needle towards data-driven approaches to pollution reduction and sustainable chemical management.

7. Conclusions

In this study, we applied machine learning models to the task above, and we showed essentially that LAR achieved higher accuracy (both in terms of Pearson correlation coefficient and R-accuracy) with lower computational consumption than many other models. This study also showed that LogS predictions could be utilized as an environmental risk assessment tool, providing critical insights into pharmaceutical bioaccumulation and pollutant behavior in aquatic ecosystems. AI-assisted solubility prediction opens doors for sustainable drug design as evidenced by our results. Challenging high-LogS compounds are more prone to be deposited in sediments and accumulate in predatorial organisms, thus reflecting long-term ecological effects and high bioconcentration risk, whereas extreme water-solubility drugs will be quickly dispersed as well as can contaminate the environment and water bodies. Using machine learning to refine environmental sustainability models, the study demonstrates how predictive models can inform regulatory and pollution control

policies. Despite its contributions, this research has certain limitations. The dataset used primarily consists of pharmaceutical compounds, limiting its generalizability to other chemical pollutants such as pesticides, industrial chemicals, and personal care products. Additionally, while LogS is a key factor in environmental impact assessment, other properties such as biodegradability, adsorption potential, and toxicity should be considered in future studies. To enhance the applicability of AI in pharmaceutical and environmental sciences, future research should focus on expanding datasets to include a wider range of chemical pollutants, developing hybrid AI models that incorporate biodegradability and toxicity predictions, and building open-source AI-powered sustainability tools for environmental regulators and researchers. By integrating machine learning with environmental risk assessment, this study paves the way for greener pharmaceutical innovations and more sustainable chemical management. The findings underscore the potential of AI-driven solutions to mitigate pharmaceutical pollution and contribute to global sustainability efforts.

Author Contributions

Conceptualization, I.A. and Y.M.; methodology, I.A. and R.K.; software, I.A. and H.B.; validation, I.A., A.B. and H.B.; formal analysis, I.A.; investigation, Y.M.; resources, I.A.; data curation, I.A.; writing—original draft preparation, I.A.; writing—review and editing, Y.M.; visualization, H.B. and R.K; supervision, A.B.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data supporting the findings of this study are available upon reasonable request from the corresponding author.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] Vaudreuil, M.A., Munoz, G., Duy, S.V., et al., 2024. Tracking down pharmaceutical pollution in surface waters of the St. Lawrence River and its major tributaries. Science of the Total Environment. 912, 168680. DOI: https://doi.org/10.1016/j.scitotenv.2023.168680

[2] Aus der Beek, T., Weber, F.A., Bergmann, A., et al., 2016. Pharmaceuticals in the environment—Global occurrences and perspectives. Environmental toxicology and chemistry. 35(4), 823–835. DOI: https://doi.org/10.1002/etc.3339

[3] Ortúzar, M., Esterhuizen, M., Olicón-Hernández, D.R., et al., 2022. Pharmaceutical pollution in aquatic environments: A concise review of environmental impacts and bioremediation systems. Frontiers in Microbiology. 13, 869332. DOI: https://doi.org/10.3389/fmicb.2022.869332

[4] Alshehri, F., Rahman, A., 2023. Coupling Machine and deep learning with explainable artificial intelligence for improving prediction of groundwater quality and decision-making in Arid Region, Saudi Arabia. Water. 15, 2298

[5] Aitouhanni, I., Mouniane, Y., Berqia, A., 2024. Machine learning-powered prediction of molecule solubility: Paving the way for environmental, and energy applications. BIO Web of Conferences. 109, 01037. DOI: https://doi.org/10.1051/bioconf/202410901037

[6] Schwarzenbach, R.P., Escher, B.I., Fenner, K., et al., 2006. The challenge of micropollutants in aquatic systems. Science. 313(5790), 1072–1077. DOI: http://doi.org/10.1126/science.1127291

[7] Goswami, D., Mukherjee, J., Mondal, C., et al., 2024. Bioremediation of azo dye: A review on strategies, toxicity assessment, mechanisms, bottlenecks and prospects. Science of The Total Environment. 954, 176426. DOI: https://doi.org/10.1016/j.scitotenv.2024.176426

[8] Kayode-Afolayan, S.D., Ahuekwe, E.F., Nwinyi, O.C., 2022. Impacts of pharmaceutical effluents on aquatic ecosystems. Scientific African. 17, e01288. DOI: https://doi.org/10.1016/j.sciaf.2022.e01288

[9] Pérez-Lucas, G., Navarro, S., 2024. How Pharmaceutical residues occur, behave, and affect the soil environment. Journal of Xenobiotics. 14, 1343–1377. DOI: https://doi.org/10.3390/jox14040076

[10] Paillet, F.L., 2000. A field technique for estimating aquifer parameters using flow log data. Groundwater. 38(4), 510–521. DOI: https://doi.org/10.1111/j.1745-6584.2000.tb00243.x

[11] Lovrić, M., Pavlović, K., Žuvela, P., et al., 2021. Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? Journal of Chemometrics. 35(7–8), e3349. DOI: https://doi.org/10.1002/cem.3349

[12] Singh, A.K., Bilal, M., Iqbal, H.M.N., et al., 2021. Trends in predictive biodegradation for sustainable mitigation of environmental pollutants: Recent progress and future outlook. Science of The Total Environment. 770, 144561. DOI: https://doi.org/10.1016/j.scitotenv.2020.144561

[13] Tong, X., Mohapatra, S., Zhang, J., et al., 2022. Source, fate, transport and modelling of selected emerging contaminants in the aquatic environment: Current status and future perspectives. Water Research. 217, 118418. DOI: https://doi.org/10.1016/j.watres.2022.118418

[14] Selvaraj, C., Chandra, I., Singh, S.K., 2022. Artificial intelligence and machine learning approaches for drug design: Challenges and opportunities for the pharmaceutical industries. Molecular Diversity. 26, 1893–1913. DOI: https://doi.org/10.1007/s11030-021-10326-z

[15] Cui, S., Gao, Y., Huang, Y., et al., 2023. Advances and applications of machine learning and deep learning in environmental ecology and health. Environmental Pollution. 335, 122358. DOI: https://doi.org/10.1016/j.envpol.2023.122358

[16] Daughton, C.G., Ternes, T.A., 1999. Pharmaceuticals and personal care products in the environment: Agents of subtle change?. Environmental Health Perspectives. 107(6), 907–938. DOI: https://doi.org/10.1289/ehp.99107s6907

[17] Wang, H., Xi, H., Xu, L., et al., 2021. Ecotoxicological effects, environmental fate and risks of pharmaceutical and personal care products in the water environment: A review. Science of The Total Environment. 788, 147819. DOI: https://doi.org/10.1016/j.scitotenv.2021.147819

[18] Domínguez-García, P., Fernández-Ruano, L., Báguena, J., et al., 2024. Assessing the pharmaceutical residues as hotspots of the main rivers of Catalonia, Spain. Environmental Science and Pollution Research. 31, 44080–44095. DOI: https://doi.org/10.1007/s11356-024-33967-7

[19] Avdeef, A., 2015. Suggested improvements for measurement of equilibrium solubility-pH of ionizable drugs. ADMET & DMPK. 3(2), 84–109. DOI: https://doi.org/10.5599/admet.3.2.193

[20] Florence, A.T., Attwood, D., 1998. The solubility of drugs. In: Attwood, D., Florence, A.T., B. (eds.). Physicochemical Principles of Pharmacy. Palgrave: London, UK. pp. 152–198. DOI: https://doi.org/10.1007/978-1-349-14416-7_6

[21] Kramer, R.M., Shende, V.R., Motl, N., et al., 2012. Toward a molecular understanding of protein solubility: Increased negative surface charge correlates with increased solubility. Biophysj. 102, 1907–1915. DOI: https://doi.org/10.1016/j.bpj.2012.01.060

[22] Boobier, S., Hose, D.R.J., Blacker, A.J., et al., 2020. Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. Nature Communications. 11, 5753. DOI: https://doi.org/10.1038/s41467-020-19594-z

[23] Cenci, F., Diab, S., Ferrini, P., et al., 2024. Predicting drug solubility in organic solvents mixtures: A machine-learning approach supported by high-throughput experimentation. International Journal of Pharmaceutics. 660, 124233. DOI: https://doi.org/10.1016/j.ijpharm.2024.124233

[24] Aitouhanni, I., Berqia, A., 2024. SolvPredict: A comprehensive exploration of predictive models for molecule solubility. International Conference on Intelligent Systems and Computer Vision, ISCV, Fez. DOI: http://doi.org/10.1109/ISCV60512.2024.10620130

[25] Kandhare, P., Kurlekar, M., Deshpande, T., et al., 2025. A review on revolutionizing healthcare technologies with AI and ML applications in pharmaceutical sciences. Drugs Drug Candidates. 4(1), 9. DOI: https://doi.org/10.3390/ddc4010009

[26] Rathi, B.S., Kumar, P.S., Vo, D.V.N., 2021. Critical review on hazardous pollutants in water environment: Occurrence, monitoring, fate, removal technologies and risk assessment. Science of The Total Environment. 797, 149134. DOI: https://doi.org/10.1016/j.scitotenv.2021.149134

[27] Tickner, J.A., Geiser, K., Baima, S., 2022. Transitioning the chemical industry: Elements of a roadmap toward sustainable chemicals and materials. Environment: Science and Policy for Sustainable Development. 64(2), 22–36. DOI: https://doi.org/10.1080/00139157.2022.2021793

[28] Ueda, D., Walston, S.L., Fujita, S., et al., 2024. Climate change and artificial intelligence in healthcare: Review and recommendations towards a sustainable future. Diagnostic and Interventional Imaging. 105(11), 453–459. DOI: https://doi.org/10.1016/j.diii.2024.06.002

[29] Chen, T.L., Kim, H., Pan, S.Y., et al., 2020. Implementation of green chemistry principles in circular economy system towards sustainable development goals: Challenges and perspectives. Science of The Total Environment. 716, 136998. DOI: https://doi.org/10.1016/j.scitotenv.2020.136998

[30] Regona, M., Yigitcanlar, T., Hon, C., et al., 2024. Artificial intelligence and sustainable development goals: Systematic literature review of the construction industry. Sustainable Cities and Society. 108, 105499. DOI: https://doi.org/10.1016/j.scs.2024.105499

[31] Gaudelet, T., Day, B., Jamasb, A.R., et al., 2021. Utilizing graph machine learning within drug discovery

and development. Briefings in Bioinformatics. 22(6), 1–22. DOI: https://doi.org/10.1093/bib/bbab159

[32] Xiong, G., Wu, Z., Yi, J., et al., 2021. ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Research. 49(W1), W5–W14. DOI: https://doi.org/10.1093/nar/gkab255

[33] Zhou, H., Shan, M., Qin, L.P., et al., 2023. Reliable prediction of cannabinoid receptor 2 ligand by machine learning based on combined fingerprints. Computers in Biology and Medicine. 152, 106379. DOI: https://doi.org/10.1016/j.compbiomed.2022.106379

[34] Boobier, S., Hose, D.R.J., Blacker, A.J., et al., 2020. Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. Nature Communications. 11, 5753. DOI: http://doi.org/10.1038/s41467-020-19594-z

[35] Martin, Y.C., 2018. How medicinal chemists learned about log P. Journal of Computer-Aided Molecular Design. 32, 809–819. DOI: http://doi.org/10.1007/S10822-018-0127-9

[36] Chen, J., Sun, Y., Sun, S., 2021. Improving human activity recognition performance by data fusion and feature engineering. Sensors. 21(3), 692. DOI: http://doi.org/10.3390/S21030692

[37] Tan, X.F., Zhu, S.S., Wang, R.P., et al., 2021. Role of biochar surface characteristics in the adsorption of aromatic compounds: Pore structure and functional groups. Chinese Chemical Letters. 32(10), 2939–2946. DOI: https://doi.org/10.1016/j.cclet.2021.04.059

[38] Syed Mustapha, S., 2023. Predictive analysis of students' learning performance using data mining techniques: A comparative study of feature selection methods. Applied System Innovation. 6(5), 86. DOI: https://doi.org/10.3390/asi6050086

[39] Shmuel, A., Glickman, O., Lazebnik, T., 2024. Symbolic regression as a feature engineering method for machine and deep learning regression tasks. Machine Learning: Science and Technology. 5, 025065. DOI: https://doi.org/10.1088/2632-2153/AD513A

[40] Zhang, Y., Wang, Z., 2023. Feature engineering and model optimization based classification method for network intrusion detection. Applied Sciences. 13(16), 9363. DOI: http://doi.org/10.3390/APP13169363

[41] ChEMBL Database. Available from: https://www.ebi.ac.uk/chembl/ (cited 28 June 2023).

[42] Bhal, S.K., Year. Application Note LogP-Making Sense of the Value. Available from: www.acdlabs.com (cited 12 July 2024).

[43] Linear Regression. Available from: http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm (cited 10 February 2024).

[44] Stouch, T.R., Kenyon, J., Ball, R., et al., 2003. In silico ADME/Tox: Why models fail. Journal of Computer-Aided Molecular Design. 17, 83–92. DOI: http://doi.org/10.1023/A:1010933404324

[45] Wang, X., Liu, Y., Huang, X., et al., 2024. Exploring the co-occurrence patterns and sources of emerging contaminants in urban rivers: A case study in Shenzhen, China. Environmental Pollution. 338, 123771. DOI: https://doi.org/10.1016/j.envpol.2024.123771

[46] Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimal Experimental Design, with SAS Applications. Chapman and Hall/CRC: Boca Raton, FL, USA. DOI: https://doi.org/10.1017/CBO9780511801389

[47] Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM: San Francisco, CA, USA. pp. 785–794. DOI: https://xgboost.readthedocs.io/en/stable/python/python_api.html

[48] Khan, M.A., Raza, S., 2024. A review of machine learning algorithms for predicting chemical properties of organic compounds: A cheminformatics approach. Ecological Informatics. 79, 102500. DOI: http://doi.org/10.1016/J.ECOINF.2024.102500

[49] PyCaret Documentation. Available from: https://pycaret.readthedocs.io/en/latest/ (cited 15 August 2024).

[50] Hempel, S., Bock, H., Neuberger, B., et al., 2014. Observation operators for a comprehensive water quality model and their impact on the state and flux estimation. Geoscientific Model Development. 7, 1247–1267. DOI: https://doi.org/10.5194/gmd-7-1247-2014

[51] Corporate Finance Institute. R-squared. Available from: https://corporatefinanceinstitute.com/resources/data-science/r-squared/ (cited 14 September 2024).

[52] Winkler, D.A., 2019. PeerJ Computer Science. 5, e623. DOI: http://doi.org/10.7717/PEERJ-CS.623/SUPP-1

[53] The RDKit. Available from: https://www.rdkit.org/ (cited 17 October 2024).

[54] PyCaret. Available from: https://github.com/pycaret/pycaret (cited 20 November 2024).