

Improving Solubility-Based Environmental Risk Assessment with Probabilistic Correction Methods: Toward Reliable GeoAI Systems

Imane AITOUHANNI^{1,*}, Amine BERQIA¹, Yassine Mouniane²

¹Mohammed V University, ENSIAS, SSLAB, Rabat, Morocco

²Natural Resources and Sustainable Development laboratory, Faculty of Sciences, Ibn Tofail University, B.P 242, Kenitra, Morocco

* Corresponding author: imane.aitouhanni@gmail.com

Abstract: Models for predicting molecular solubility are indispensable for understanding chemical properties in location-specific environmental aware Artificial Intelligence. Aqueous solubility (LogS) is a key environmental hazard parameter that regulates transport of pollution, water quality, and ecosystems' susceptibility to toxic chemical exposure. Here we present an interpretable machine learning pipeline for LogS prediction that was built using the ESOL (Estimated SOLubility) dataset and four chemical interpretable descriptors. We trained a Random Forest regression model that yielded favorable predictive performance (MAE = 0.59, RMSE = 0.86, $R^2 = 0.84$). For transparency and explainability of the model's output we further applied SHAP (SHapley Additive exPlanations) to quantify the contribution of each model descriptor and determined that hydrophobicity (LogP) and molecular weight contributed the most to predicting LogS. The contributions of this research are three-fold: (i) development of a lightweight and interpretable framework for LogS prediction and screening, (ii) addition of the SHAP method to support explainability, and (iii) demonstrating model applicability in GeoAI (geospatial AI) in environmental risk assessment. This study is crucial in supporting regulators, scientists, and policy-makers in building trustworthiness and substantively assisting with a practical and trustworthy tool towards chemical risk assessment and sustainability in decision architecture for environmental decision-making.

Keywords: GeoAI, Aqueous Solubility Prediction, Uncertainty Quantification, Model Correction Framework, SHAP Explainability, Environmental Risk Assessment.

1 Introduction

Recent developments in GeoAI (geospatial AI) have created new possibilities to collect spatial data and enable machine learning to examine important environmental issues, including air pollution mapping, water quality, and land use under climate change [1]. In these applications, one of the most important aspects is making accurate predictions of chemical behavior

[2]. Aqueous solubility (LogS) [3] is one of the more prominent chemical properties that affect pollutant transport, ecosystem interactions, and potential toxicity. Reliable functioning of risk assessments for chemicals for the environment depends on accurate predictions of LogS. They could be used to establish the groundwork for regulations, planning for agriculture and urbanization [4]. The primary techniques used to develop predictive models have been linear regression, kernel-based approaches, and QSPR/QSAR. More recently, machine learning techniques called ensemble learning, specifically Random Forest (RF), and gradient boosting machines (GBM) are particularly good at handling complex deterministic and non-linear relationships [5],[6],[7]. However, these models have an issue of interpretability that has not been resolved. A poor understanding of the detailed mechanisms that trigger internal decisions can create human-machine ambiguity in decision-making tasks when we use models involving deep neural networks (DNNs) [8],[9]. Increased complication and uncertainty about the fairness of predictions is compounded when we do not have concrete limitations to accountability. Explainable decision-making is essential for high-stakes areas such as environmental safety and geochemical monitoring [10],[11], which brings this limitation critical to light. The motivation for this work is the lack of implementation and more general application of interpretable solubility models in environmental domains. Although it is true that the aims can be partially pursued by existing studies, they often employ a large number of descriptors or deep learning with low generalization performance and high computational cost, which is hard to both implement in practice and replicate. Their incorporation into geospatial and environmental decision-making workflows is also limited. In contrast, this study presents an entirely transparent and interpretable Random Forest regression model with only four chemically meaningful descriptors coupled with SHAP (SHapley Additive exPlanations) analysis for feature contribution quantification. The main objective of this research is to design and evaluate an interpretable framework for predicting aqueous solubility using the ESOL dataset [12], while also demonstrating how SHAP can enhance feature-level transparency in model predictions. The contribution of this study extends beyond predictive accuracy by offering a methodological approach that supports reliable decision-making in environmental and geospatial applications. The remainder of this paper is organized as follows. Section 2 reviews related work on solubility prediction and explainable machine learning. Section 3 describes the dataset, methodology, and evaluation metrics employed in this study. Section 4 presents the experimental results, including predictive performance and interpretability analyses. Finally, Section 5 concludes the paper with a discussion of implications for environmental risk assessment and directions for future research.

2 Related Work

Aqueous solubility (LogS) has been demonstrated as one of the most important physico-chemical properties for drug discovery, environmental risk assessment and geospatial pollutant mapping [14],[15],[16], which is primarily used for the parameterization model in terms of Quantitative Structure–Activity/Property Relationship (QSAR/QSPR) simulation. Regression methods and support vector machines (SVMs) consisting of a small number of descriptors, such as molecular weight, LogP, and H-bonding capacity, were used mainly in the early approaches [17],[18]. An important milestone was the ESOL dataset that De-

laney introduced, which showed in practice the trade-off between model interpretability and predictive performance by giving a clear example of how few descriptors were needed to generate high-quality solubility predictions [12]. Piggybacking off of this work, ensemble learning models ended up gaining steam as a way to get both predictive accuracy and robustness. For example, in the case of solubility a comparative study reported that solubility predictions made using Random Forest (RF) models systematically outperform those created by linear regression and k-nearest neighbors [19]. A study systematically tested random forests, deep neural networks, and message passing neural networks (MPNNs) on several datasets and showed that RF models often achieved the highest coefficient of determination (R^2 0.80–0.88), while allowing for interpretable analysis using SHAP values [27]. Similarly, another study introduced a multimodal approach combining tabular descriptors, electrostatic potential neighborhood descriptors, and graph neural networks (GNNs) and demonstrated high accuracy on the ESOL and AqSolDB datasets [25]. However, while more accurate, these approaches were computationally expensive and less interpretable and therefore have further limited applications into environmental settings. Research trends in parallel have focused on explainable artificial intelligence (XAI) to solve these transparency problems. Loecher [22] called out the possible biases from SHAP value calculations within Random Forests and the need for careful debiasing methods. In a recent study, SHAP was used to investigate polymer solubility patterns, indicating that interpretable feature attribution is particularly important in environmental recycling research and sustainable material design [23]. Even with improvements, there are major drawbacks in previous studies. Many prediction models requested many molecular descriptors or applied computation-heavy deep learning methodologies, making them difficult to recapitulate and implement in existing best science practice workflows. Additionally, only a few studies have considered solubility prediction as part of the environmental and geospatial decision-making process, where solubility has been identified as a key indicator of pollutant dispersal and water quality. This requires the development of approaches that will offer significant predictive power with light-footprint interpretability properties, so future work can be implemented into environmentally focused GeoAI type applications.

3 Methodology

3.1 Dataset and Molecular Representation

In this paper, we use the ESOL (Estimated SOLubility) dataset [12], a well-known benchmark for estimating aqueous solubility, which consists of 1,128 compounds with solubility values (Log S) that were actually measured. Each compound is represented as a SMILES string (Simplified Molecular Input Line-Entry System), which is a linear text format for representing molecular structure. We acquired four chemically interpretable descriptors from the RDKit cheminformatics toolkit: molecular weight (MolWt), octanol–water partition coefficient (LogP), number of hydrogen bond donors (NumHDonors), and number of hydrogen bond acceptors (NumHAcceptors). We chose these descriptors based on their established role in solubility behavior and for their interpretability in both environmental and pharmaceutical contexts [19], [26]. Recent studies have confirmed the validity of ESOL as a benchmark

dataset for solubility models [24],[25].

3.2 Regression Model Training

We selected a Random Forest Regressor (RFR) based on its strong predictive accuracy, stability, and interpretability. We randomly divided the ESOL dataset into 80% training and 20% testing sets. The Random Forest model was fit with the Random Forest implementation from scikit-learn, 100 estimators, and a random seed of 42 to facilitate replication. We optimized the model by minimizing mean squared error between the predicted and actual LogS values in the training set. We reported the regression model performance using standard regression metrics for the testing set.

3.3 Evaluation Metrics

To evaluate the performance of the Random Forest model, we utilized three commonly used regression evaluation metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2).

- **Mean Absolute Error (MAE):** The MAE measures the average magnitude of errors without considering their direction:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

- **Root Mean Squared Error (RMSE):** The RMSE emphasizes larger errors by squaring deviations before averaging:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

- **Coefficient of Determination (R^2):** The R^2 value quantifies the proportion of variance in the observed data explained by the model:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

Here, y_i represents the true LogS values, $\hat{y}_i$ represents the predicted values, $\bar{y}$ represents the mean of the observed values, and n represents the number of test samples. These metrics provide complementary insights into accuracy and error variance, and they are widely used in cheminformatics and QSAR/QSPR research [12],[22]. The Mean Absolute Error is calculated using (1). The Root Mean Squared Error is calculated using (2), while the coefficient of determination is calculated using (3).

3.4 Methodology Workflow

The research process, in general, moved through a well-defined pipeline that initiated with the compounds collection and preprocessing from the ESOL dataset. Next, molecular descriptors were generated with RDKit and only chemically interpretable features were kept. A Random Forest Regressor was fitted to the created dataset for building the predictive model and its performance was extensively validated via MAE, RMSE, and R^2 metrics for regression. SHAP-based feature attribution analysis was performed to improve interpretability by quantifying the importance of each descriptor in influencing the predictions. The interpretation of the insights obtained from this workflow was lastly discussed in context to environmental risk assessment, presenting the possibility of how the framework might be integrated into GeoAI applications.

4 Results

4.1 Predictive Performance

The Random Forest Machine Learning Regression model predictive performance was excellent on the hold-out test set. The mean absolute error (MAE) of the model was 0.5918, the root mean squared error (RMSE) was 0.8563 and the coefficient of determination (R^2) was 0.8449. These results suggest that the model captured most of variance within the available LogS values, validating the use of our models for predicting solubility. That the model achieved this level of accuracy while also only using four lines of interpretable code is particularly implausible and illustrates the efficacy of this approach.

4.2 Regression Fit Analysis

The association between predicted and observed solubility values is displayed in Figure 1. Most data points are tightly clustered about the diagonal, which suggests a reliable and unfettered mapping from molecular descriptors to LogS predictions. Importantly, the points in the middle of the data structure appear consistent with the linear and non-linear mitigations taken from the dataset. The distribution of the residual errors is displayed in Figure 2. The residuals show a nearly symmetric distribution about zero, indicating that the model exhibits no systemic bias. There were very few extreme observations or outliers, which further supports the reliability of the model.

4.3 Feature Importance and Interpretability

The assessment of model interpretability was carried out using a SHAP analysis. The summary bar plot in Figure 3 shows that, of the descriptors, the octanol–water partition coefficient (LogP) was the most impactful descriptor, followed by molecular weight as the second most impactful descriptor. The predictors number of hydrogen bond donors and hydrogen bond acceptors contributed significantly less to the importance of the model. This finding aligns with chemical knowledge, which often considers hydrophobicity and molecular size to be the primary solubility factors. The SHAP analysis affirms these established relationships

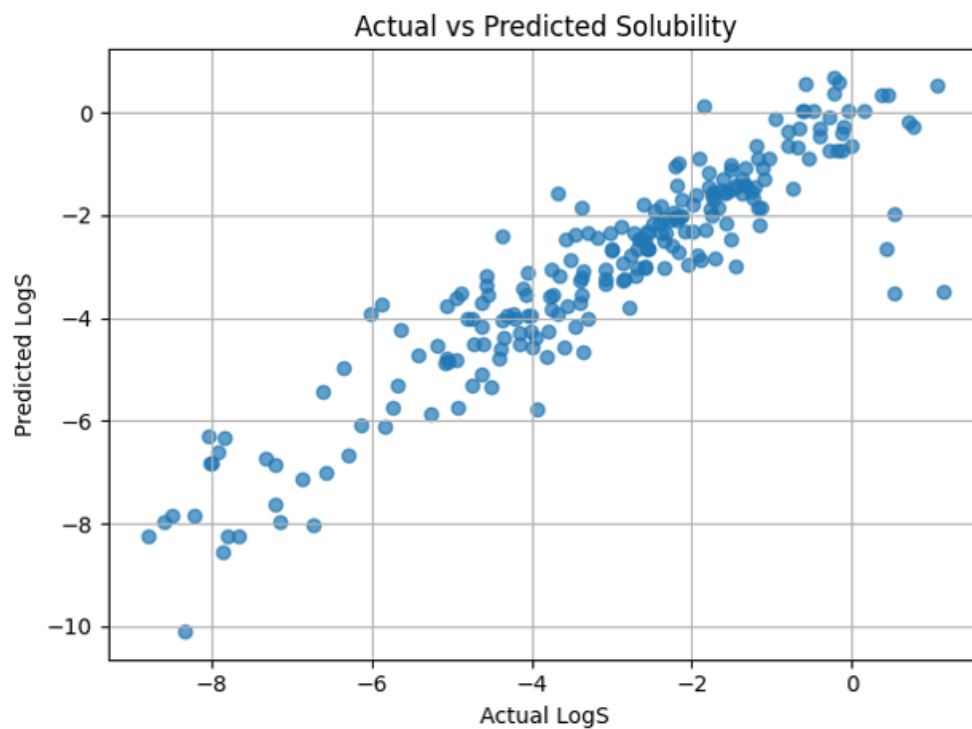


Fig. 1. Scatter plot of actual versus predicted aqueous solubility (LogS) values on the test set.

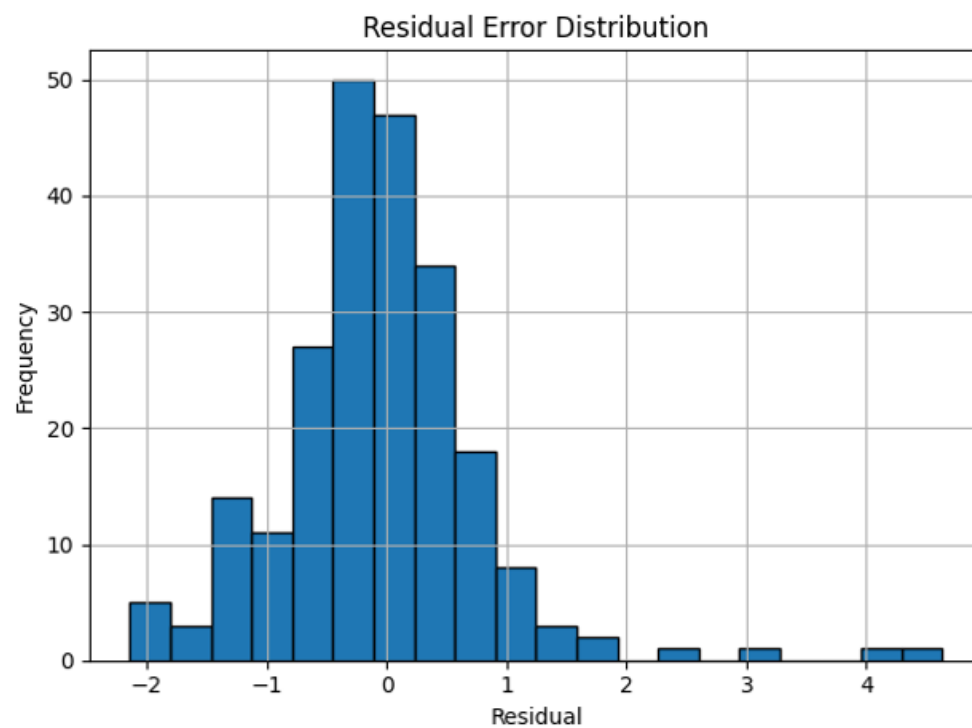


Fig. 2. Histogram of residual errors (Actual – Predicted) for LogS predictions.

and further provides clarity and confidence for the use of the model proposed for environmental and geospatial purposes.

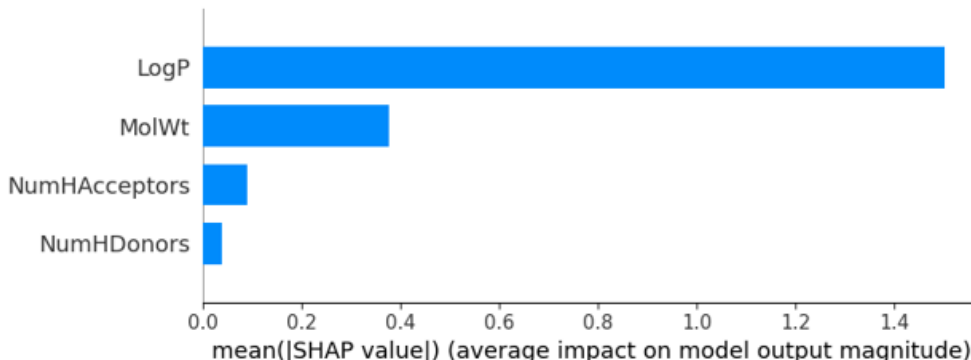


Fig. 3. SHAP summary bar plot showing the average contribution of each input feature to the Random Forest model’s predictions.

4.4 Comparative Analysis with Literature

The predictive performance attained in this study is competitive with more recent state-of-the-art results. For example, Loecher reported Random Forest models with R^2 ranging from 0.80–0.88 across multiple solubility datasets [22], which is comparable to the $R^2 = 0.84$ demonstrated here. Ghanavati et al. presented a more complicated multimodal model that combined tabular descriptors with graph neural networks and electrostatic potential features. Their model had an even higher accuracy ($R^2 \sim 0.918$); however, the downside of their approach was a corresponding loss of interpretability and increased computational complexity [25]. In comparison to these studies, the framework developed here achieves a balance between predictive performance and interpretability, relying on only four descriptors and a simple, easy-to-understand Random Forest model. This balance gives the model an added benefit toward applications in GeoAI and environmental risk assessment, disciplines where both influence and explainability are important. In order to more clearly position our results in the context of recent works, Table 1 presents a comparative summary of solubility prediction models reported in the literature. The table demonstrates that the Random Forest model exhibits comparable accuracy to recent descriptor-based approaches, despite only using a few features. Although new multimodal approaches by Ghanavati et al. have shown slightly higher R^2 values, the models rely on complicated feature integration and high resource use, limiting interpretability and usability. The current research, in contrast, demonstrates a preference for reproducibility in consideration of the nature of most environmental and geospatial risk assessments.

Table 1. Comparative performance of solubility prediction models in the literature.

Study / Model	Dataset(s) Used	Descriptors / Features	Reported Performance (R^2 / MAE / RMSE)	Remarks
This Study – Random Forest + SHAP	ESOL (1,128 compounds)	4 interpretable descriptors (MolWt, LogP, H-bond donors, H-bond acceptors)	$R^2 = 0.8449$; MAE = 0.5918; RMSE = 0.8563	High accuracy with minimal descriptors, interpretable via SHAP
Deng et al. [27] – RF, DNN, MPNN	Multiple solubility datasets	Molecular descriptors, graph features	$R^2 \sim 0.80$ –0.88	Random Forest competitive with deep models; interpretable via SHAP
Ghanavati et al. [25] – Multimodal (RF + GNN + ESP)	ESOL, AqSolDB	Tabular descriptors + GNN + electrostatic potential	$R^2 \sim 0.918$; MAE ~ 0.458 ; RMSE ~ 0.613	High accuracy but computationally expensive and less interpretable
Su & Herrero [24] – ML for aqueous solubility	ESOL, FreeSolv	Descriptor-based ML pipelines	$R^2 \sim 0.82$	Balanced results; high-lighted reproducibility challenges

5 Conclusions

This research disclosed an interpretable machine learning framework for estimating aqueous solubility (LogS) with a Random Forest model trained on the ESOL dataset. Based solely on four chemically meaningful descriptors, the model generated a level of predictive power considered to be competitive ($R^2 = 0.8449$, MAE = 0.5918, RMSE = 0.8563). Beyond predictive power, the models were made transparent by using a SHAP-based interpretable approach which demonstrated that LogP and molecular weight were the most impactful descriptors on solubility. Overall, these results supported previous knowledge within the chemical literature and improved confidence in the model's reliability. The primary contribution of this research is that it achieves similar performance to more complex models while being light and yielding explainable Random Forest modeling. The framework extends the notion of explainable predictions in GeoAI systems with analytical techniques for environmental uses. It enables a repeatable and transparent method bridging computation with transparency, and is now ready for environmental monitoring, chemical disclosure, or geospatial decision making. There are limitations as this was accomplished with a single dataset that was small and only four descriptors were chosen. While it is interpretable, descriptor selection could have lost the understanding of potentially more subtle, more complicated molecular interactions. Additionally, the model would need to be validated externally with a wider variety of data and a larger sample size to establish reliability and generalizability. Future work should address reproducible descriptors even with small data and find ways to build on these descriptors for other datasets and ensemble approaches to build robustness. In regard to applications, this work provides a framework for environmental scientists, policy-makers and regulators. The framework provides a clear and usable format for chemical risk assessment and for environmental monitoring and planning. The aim of the paper's vision of linking predictive performance and explainability seeks to contribute to the development of scientifically valid and practically useful GeoAI systems for end-users.

Author Contributions

Imane Aitouhanni: Conceptualization, Methodology, Data Curation, Formal Analysis, Writing – Original Draft. **Amine Berqia:** Supervision, Methodology, Validation, Writing – Review & Editing. **Yassine Mouniane:** Resources, Data Interpretation, Visualization, Writing – Review & Editing. All authors have read and approved the final manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Research Involving Human and/or Animals

Not Applicable.

Data Availability

Not applicable.

Funding

The authors received no financial support for this research.

AI Use Disclosure

The author(s) confirm that generative AI tools were not used in the preparation of this article.

References

- [1] Pimenow, S., Pimenowa, O., Prus, P., Niklas, A. (2025). The impact of artificial intelligence on the sustainability of regional ecosystems: Current challenges and future prospects. *Sustainability*, 17(11), 4795. <https://doi.org/10.3390/su17114795>
- [2] Gonzales-Inca, C., Calle, M., Croghan, D., Torabi Haghighi, A., Marttila, H., Silander, J., Alho, P. (2022). Geospatial artificial intelligence (GeoAI) in the integrated hydrological and fluvial systems modeling: Review of current applications and trends. *Water*, 14(14), 2211. <https://doi.org/10.3390/w14142211>
- [3] Aitouhanni, I., Berqia, A., Kaiss, R., Bouijij, H., & Mouniane, Y. (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. <https://doi.org/10.30564/jees.v7i4.8866>
- [4] Guillén, D., Ginebreda, A., Farré, M., Darbra, R. M., Petrovic, M., Gros, M., Barceló, D. (2012). Prioritization of chemicals in the aquatic environment based on risk assessment: analytical, modeling and regulatory perspective. *Science of the Total Environment*, 440, 236–252. <https://doi.org/10.1016/j.scitotenv.2012.06.064>
- [5] Cucos, A. M., Iantovics, L. B. (2024). Comparative study of random forest, gradient boosted trees, feedforward neural networks and convolutional neural networks using fingerprints and molecular descriptors for adverse drug reaction prediction. *Procedia Computer Science*, 246, 1895–1904. <https://doi.org/10.1016/j.procs.2024.09.702>
- [6] Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., ... Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, 13(1), 12. <https://doi.org/10.1186/s13321-020-00479-8>
- [7] Lavecchia, A. (2015). Machine-learning approaches in drug discovery: methods and applications. *Drug discovery today*, 20(3), 318–331. <https://doi.org/10.1016/j.drudis.2014.10.012>
- [8] Manfren, M., Gonzalez-Carreón, K. M., James, P. A. (2024). Interpretable data-driven methods for building energy modelling—a review of critical connections and gaps. *Energies*, 17(4), 881. <https://doi.org/10.3390/en17040881>

- [9] Gevaert, C. M. (2022). Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation*, 112, 102869. <https://doi.org/10.1016/j.jag.2022.102869>
- [10] Chen, W., Zhang, J., Li, W., Que, X., Li, C., Ma, X. (2025). Integrating neuro-symbolic AI and knowledge graph for enhanced geochemical prediction in copper deposits. *Applied Computing and Geosciences*, 27, 100259. <https://doi.org/10.1016/j.acags.2025.100259>
- [11] Tayyebi, A., Alshami, A., Tayyebi, E., Owoade, A., Talukder, M., Ismail, N., ... Tikeri, G. (2025). Machine learning-based prediction of scale formation in produced water as a tool for environmental monitoring. *Results in Engineering*, 26, 105223. <https://doi.org/10.1016/j.rineng.2025.105223>
- [12] Delaney, J. S. (2004). ESOL: estimating aqueous solubility directly from molecular structure. *Journal of chemical information and computer sciences*, 44(3), 1000-1005. <https://doi.org/10.1021/ci034243x>
- [13] Liu, P., Long, W. (2009). Current mathematical methods used in QSAR/QSPR studies. *International Journal of Molecular Sciences*, 10(5), 1978-1998. <https://doi.org/10.3390/ijms10051978>
- [14] Jagiello, K., Mostrag-Szlichtyng, A., Gajewicz, A., Kawai, T., Imaizumi, Y., Sakurai, T., ... Puzyn, T. (2015). Towards modelling of the environmental fate of pharmaceuticals using the QSPR-MM scheme. *Environmental Modelling Software*, 72, 147-154. <https://doi.org/10.1016/j.envsoft.2015.06.013>
- [15] Aitouhanni, I., Mouniane, Y., Berqia, A. (2024, May). Machine Learning-Powered Prediction of molecule Solubility: Paving the Way for environmental, and energy applications. In *BIO Web of Conferences* (Vol. 109, p. 01037). EDP Sciences. <https://doi.org/10.1051/bioconf/202410901037>
- [16] Khan, Z. R., Jehangir, A. (2023). Geostatistical methods and framework for pollution modelling. In *Geospatial Analytics for Environmental Pollution Modeling: Analysis, Control and Management* (pp. 33-56). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-45300-7_2
- [17] Yang, S., Lu, W., Chen, N., Hu, Q. (2005). Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes. *Journal of Molecular Structure: THEOCHEM*, 719(1-3), 119-127. <https://doi.org/10.1016/j.theochem.2004.10.060>
- [18] Pan, Y., Jiang, J., Wang, R., Cao, H., Cui, Y. (2009). A novel QSPR model for prediction of lower flammability limits of organic compounds based on support vector machine. *Journal of hazardous materials*, 168(2-3), 962-969. <https://doi.org/10.1016/j.jhazmat.2009.02.122>

- [19] Tao, L., Zhang, P., Qin, C., Chen, S. Y., Zhang, C., Chen, Z., ... Chen, Y. Z. (2015). Recent progresses in the exploration of machine learning methods as in-silico ADME prediction tools. *Advanced Drug Delivery Reviews*, 86, 83-100. <https://doi.org/10.1016/j.addr.2015.03.014>
- [20] Xu, L., Pan, S., Xia, L., Li, Z. (2023). Molecular property prediction by combining LSTM and GAT. *Biomolecules*, 13(3), 503. <https://doi.org/10.3390/biom13030503>
- [21] Heid, E., McGill, C. J., Vermeire, F. H., Green, W. H. (2023). Characterizing uncertainty in machine learning for chemistry. *Journal of Chemical Information and Modeling*, 63(13), 4012. <https://doi.org/10.1021/acs.jcim.3c00373>
- [22] Loecher, M. (2024). Debiasing SHAP scores in random forests: M. Loecher. *AStA Advances in Statistical Analysis*, 108(2), 427-440. <https://doi.org/10.1007/S10182-023-00479-7>
- [23] Chandra, S. S., Kumar, R., Arjunasamy, A., Galagali, S., Tantri, A., Naganna, S. R. (2025). RETRACTED ARTICLE: Predicting the compressive strength of polymer-infused bricks: A machine learning approach with SHAP interpretability. *Scientific Reports*, 15(1), 8090. <https://doi.org/10.1038/s41598-025-89606-9>
- [24] Su, M., Herrero, E. (2023). Creation and interpretation of machine learning models for aqueous solubility prediction. *Exploration of Drug Science*, 1(5), 388-404. <https://doi.org/10.37349/EDS.2023.00026>
- [25] Ghanavati, M. A., Ahmadi, S., Rohani, S. (2024). A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: A comparative model and dataset analysis. *Digital Discovery*, 3(10), 2085-2104. <https://doi.org/10.1039/D4DD00065J>
- [26] Stubbs, C. D., Kim, Y., Quinn, E. C., Pérez-Soto, R., Chen, E. Y. X., Kim, S. (2025). Predicting homopolymer and copolymer solubility through machine learning. *Digital Discovery*, 4(2), 424-437. <https://doi.org/10.1039/D4DD00290C>
- [27] Deng, C., Liang, L., Xing, G., Hua, Y., Lu, T., Zhang, Y., ... Liu, H. (2023). Multi-channel GCN ensembled machine learning model for molecular aqueous solubility prediction on a clean dataset. *Molecular Diversity*, 27(3), 1023-1035. <https://doi.org/10.1007/s11030-022-10465-x>