Applied Sciences, Computing and Energy, 2(1), 131-148

Advances in Machine Learning Approaches for Predicting Aqueous Solubility in Drug Discovery

Comfort M. Ngwu, Daniel C. Ottah*, Praise O. Otitodirichukwu and Chioma Harbor
Received : 20 December 2024/Accepted: 26 February 2025/Published: 11 March 2025

Abstract: *Nearly 90% of drug candidates and 40% of licensed medications have poor water solubility. Aqueous solubility is a critical factor in drug discovery and development, influencing absorption, bioavailability, and therapeutic effectiveness. Machine learning (ML) approaches, including Random Forest, Support Vector Machines, deep learning models, and hybrid techniques, have demonstrated superior accuracy in predicting solubility compared to traditional quantitative structure-property relationship (QSPR) models. These models utilize molecular descriptors, fingerprints, and advanced computational techniques to enhance predictive performance, enabling efficient screening of large chemical libraries. Challenges such as data quality, model interpretability, and overfitting persist, necessitating the adoption of explainable AI, active learning, and transfer learning to improve robustness and generalizability. The integration of ML-based solubility prediction into drug development pipelines has shown promise in optimizing formulation strategies and reducing late-stage failures. This study aims at providing a detailed review on the prediction of aqueous solubility in drug discovery, using machine learning approaches and also the advances found in them. Future research will focus on expanding high-quality datasets, refining hybrid ML-physics models, and leveraging quantum computing to further advance solubility prediction and accelerate pharmaceutical innovation.*

Keywords: *Aqueous solubility, Drug-Like compounds, Machine learning, Artificial Intelligence, Deep learning.*

Comfort M. Ngwu
TECHEM-360 Artificial Intelligence Research Group., Department of Chemistry, Michael Okpara University of Agriculture, Umudike, Umuahia, Abia State, Nigeria.
Email: ngwu.comfort@mouau.edu.ng
Orcid id: https://orcid.org/0000-0003-4723-2813

Daniel C. Ottah*
TECHEM-360 Artificial Intelligence Research Group, Department of Chemistry, Michael Okpara University of Agriculture, Umudike, Umuahia, Abia State, Nigeria.
Email: ottah.daniel@mouau.edu.ng
Orcid id: https://orcid.org/0009-0002-4599-6123

Praise O. Otitodirichukwu
TECHEM-360 Artificial Intelligence Research Group, Department of Chemistry, Michael Okpara University of Agriculture, Umudike, Umuahia, Abia State, Nigeria.
Email: Okafor.praiseo@mouau.edu.ng
Orcid id: https://orcid.org/0009-0003-9417-1213

Chioma Harbor
TECHEM-360 Artificial Intelligence Research Group, National Root Crops Research Institute, Umuahia, Abia State, Nigeria.
Email: chiomaharbor@gmail.com
Orcid id: https://orcid.org/0000-0002-5314-0280

1.0 Introduction

Aqueous solubility is a critical physicochemical property in pharmaceutical sciences, as it directly influences drug absorption, bioavailability, and therapeutic efficacy (Savjani *et al*., 2012). Poor solubility is a major challenge in drug discovery and

development, with nearly 40% of approved drugs and up to 90% of new chemical entities classified as poorly soluble (Di *et al*., 2009; Kalepu & Nekkanti, 2015). This limitation often leads to high attrition rates in drug development, increased formulation costs, and reduced therapeutic efficiency (Lipinski, 2000). Consequently, accurate solubility prediction is essential for optimizing drug candidates and improving the efficiency of pharmaceutical development.

Traditional solubility prediction approaches rely on experimental measurements, thermodynamic models, and Quantitative Structure-Property Relationship (QSPR) models. While these methods provide valuable insights, they are often time-consuming, expensive, and limited by the availability of high-quality experimental data (Yalkowsky & He, 2003). Machine learning (ML)-based approaches have emerged as powerful alternatives, leveraging computational algorithms to predict solubility based on molecular descriptors, topological indices, and structural fingerprints (Delaney, 2004; Tetko *et al*., 2001). ML models such as Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Deep Learning (DL) have demonstrated improved predictive accuracy over conventional QSPR methods (Goh *et al*., 2017; Wang *et al*., 2019).

Several physicochemical properties, including molecular weight, hydrogen bond donors and acceptors, logP, and polar surface area, significantly influence aqueous solubility (Manohar, 2019). ML-driven solubility prediction integrates these molecular features with computational algorithms to enhance predictive efficiency, reducing the reliance on experimental screening (Mohammad *et al*., 2024). Despite these advancements, challenges persist, including data scarcity, model interpretability, and generalizability across diverse chemical spaces (Lusci *et al*., 2013; Nguyen *et al*., 2022). Additionally, while many ML models excel in predictive accuracy, their black-box nature limits their applicability in regulatory and industrial settings, where explainability is crucial (Cortes-Ciriano *et al*., 2020).

Various formulation strategies have been employed to enhance the solubility of poorly water-soluble drugs, including micronization, solid dispersions, self-emulsifying drug delivery systems (SEDDS), and liposomal encapsulation (Kakran *et al*., 2012). However, these techniques often require extensive experimental optimization, further emphasizing the need for robust computational tools to guide formulation design (Jafari *et al*., 2022).

Knowledge Gap

Despite the progress in ML-based solubility prediction, key gaps remain. Many existing models are trained on limited datasets, reducing their generalizability across diverse drug-like molecules (Sorkun *et al*., 2019). Furthermore, while deep learning models have demonstrated superior predictive performance, their black-box nature hinders mechanistic understanding and practical deployment (Schwöbel *et al*., 2018). Additionally, there is a need for ML frameworks that integrate experimental uncertainty, interpretability, and domain-specific knowledge to improve real-world applicability (Wu *et al*., 2018).

Aim of the Present Study

This study aims to provide a comprehensive review of ML-based approaches for aqueous solubility prediction, highlighting key algorithms, datasets, challenges, and future directions. By addressing current limitations and exploring emerging trends, this review seeks to enhance the understanding of computational solubility prediction and its implications for drug discovery and formulation. The ultimate goal is to contribute to the development of more accurate, interpretable, and generalizable ML models that can streamline pharmaceutical research and innovation.

2.0. The emergence and importance of machine learning (ML) in solubility prediction

Machine learning (ML) has proved itself to be a vital tool used in predicting the solubility of molecules in organic and aqueous solvent; it has also addressed a critical challenge in drug development (Boobier *et al*., 2020). Different ML algorithms have been put into place to attain accurate predictions, and these include Random Forest, Support Vector Machine (SVM), and Artificial Neural Network (ANN). These algorithms, according to Lovrić *et al*. (2021), often outperformed traditional methods. These models work by combining physicochemical descriptors such as molecular weight, hydrogen bond ability, logP (partition coefficient), and polarity with advanced computational chemistry to deduce the complex relationship between solubility and molecular properties (Tayyebi *et al*., 2023). In a recent study, Tayyebi *et al*. (2023) showed that the comparison between descriptor-based and fingerprint-based models proved that both of them produced a high predictive accuracy. The incorporation of ML into the scientific study of how dissolution works, based on established principles or theories, rather than experiments has led to improved model accuracy and reproducibility of physicochemical relationships (Boobier *et al*., 2020). This advancement in solubility prediction using ML offers promising applications in various fields, including drug discovery. The aqueous solubility of drugs is a vital physicochemical property in drug development, affecting absorption, distribution and bioavailability of the drug (Wang & Hou, 2011).

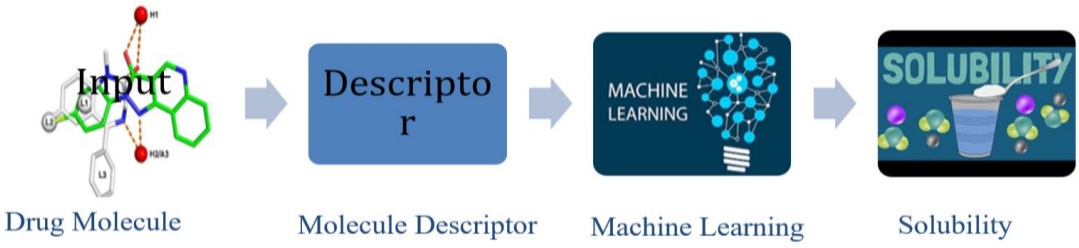

Fig 1. Aqueous solubility prediction procedure using machine learning

Despite the exploration of many approaches, the accurate prediction of solubility still remains a challenge. Models like simple linear regression, artificial neural networks, category formation, and in silico models have been compared, with simple linear regression surprisingly outperforming more complex methods in some cases (Hewitt *et al*., 2009). In a study conducted, Wang and Hou (2011), made it clear that key challenges experienced by the models include collecting high-quality, diverse solubility data and developing descriptors that account for lattice energy of solvation. Recent studies show how researchers have created quantitative structure-property relationship (QSPR) models, by combining descriptors for Gibbs energy of salvation and sublimation with octanol-water partition coefficients (LogP), these (QSPR) models achieved R2 > 0.7 and Q2 > 0.6 across multiple datasets (Meftahi *et al*., 2021). Additionally, commercial software programs have been evaluated for predicting solubility; this shows the effort put in by researchers to optimize the accuracy of solubility predictions. (Dearden, 2006). Despite the progress, the development of reliable models for predicting the solubility of drug-like compounds remains an active area of research.

2.1 Definition and measurement of aqueous solubility

Aqueous Solubility refers to how much of a substance that can dissolve in water before the

solution attains an equilibrium with the pure substance, at specific temperature and pressure ("Solubility," 2008). It is a vital property in drug discovery and development, affecting both environmental behavior and formulation selection (Elder & Holm, 2012). There are various methods of measuring solubility, which include simple and cost-effective dialysis tubing techniques for organic compounds (Etzweiler *et al*., 1995) and high-throughput automated methods for early drug discovery (Murdande *et al*., 2010). In a study, Murdande *et al*. (2010), highlighted that the challenges in solubility measurement arise due to the unstable nature of amorphous substances and the difficulty in precisely identifying and maintaining the solid-state form of the test substance throughout the experiment. However, predictive methods like both simple and complex in silicon (advanced computer-based) are used to derive aqueous solubility (Elder & Holm, 2012). The need to accurately determine the aqueous solubility of drugs, as seen in a study by Murdande *et al*. (2010), is crucial in drug discovery and development, with researchers emphasizing the importance of proper physico-chemical characterization and awareness of potential pitfalls in experimental methods.

2.2 Factors affecting solubility in drug-like compounds

Several factors affect the solubility of drug-like compounds. Factors such as molecular size, lipophilicity, and rigidity decrease solubility, while conformational flexibility and the presence of non-conjugated amine groups tend to increase it (Huuskonen *et al*., 2008). In a study, Walker (2013) emphasized that the general solubility equation (GSE) identifies logP and melting point as primary determinants of solubility. Improving solubility can be attained by reducing logP or melting point through increased polarity or disruption of intermolecular interactions. For lipophilic drugs, optimal oral bioavailability is associated with specific ranges of solubility parameters and hydrophilic-lipophilic balance (HLB) values, depending on molecular weight (Kuentz & Arnold, 2009). Aqueous solubility is key for a drug's effectiveness in the body, significantly impacting drug absorption, distribution, and elimination, while DMSO solubility plays an important role in the lab during the discovery and screening phases. In silico methods for predicting both water and DMSO solubility have become valuable tools in drug discovery processes. (Balakin *et al*., 2005).

2.3 Historical approaches to solubility prediction

Historically, two major approaches have been used to estimate solubility predictions, which are empirical models and quantitative structure-property relationships (QSPR) (Chen *et al*., 2002). In order to predict the aqueous solubility of various substances using molecular descriptors, several researchers have created QSPR models (Cheng & Merz, 2003b). These models employ various statistical methods, including multiple linear regression, genetic algorithms, and machine learning techniques (Raevsky *et al*., 2015). Chen *et al*. (2002) and Cheng & Merz (2003) detected and reported models with root-mean-square errors of 0.56-0.87 log units for training sets. Raevsky *et al*. (2015) compared global and local models, finding that global models offer better predictive power but less interpretability. Local models, while more interpretable, require structurally similar compounds for stability. Yousefinejad *et al*. (2016) explored both QSPR and linear solvation energy relationship (LSER) models for predicting pyrene solubility in organic solvents, highlighting the importance of both structural and empirical solvent properties in solubility prediction.

2.4 Limitations of traditional models and the need for ML-based methods

Conventional statistical and process-based models struggle to manage complex, high-dimensional data in areas such as healthcare,

transportation, and agriculture (Lam *et al*., 2024). The flexibility, scalability, and predictive accuracy of machine learning (ML) techniques make them appropriate for tasks including medication discovery, diagnostics, and individualized treatment (Rajula *et al*., 2020). Machine learning (ML) adaptation and hybridization with traditional econometric techniques show potential in transportation modeling (Vovsha, 2021). Lam *et al*. (2024), detected that for agroecosystem modeling, a hybrid approach combining process-based and ML models is proposed to enhance nitrogen loss prediction accuracy and scalability. In gasification research, ML algorithms like artificial neural networks (ANNs) and support vector machines (SVMs) provide alternatives to traditional modeling approaches, potentially fostering the development of novel catalysts (Umenweke *et al*., 2022). However, challenges remain, including the need for effective implementation and mitigating the black-box nature of some ML algorithms (Umenweke *et al*., 2022).

3.0 Machine Learning Techniques for Solubility Prediction

Machine learning (ML) techniques have shown strong potentials in predicting solubility of compounds in various solvents (Boobier *et al*., 2020b). Various studies have employed a hybridization of machine learning algorithms, including artificial neural networks (ANNs), support vector machines (SVMs), random forests, and gradient boosting methods, to predict solubility in organic solvents and water (Ye & Ouyang, 2021). These models have attained high accuracy, with predictions close to the expected level of noise in training data (LogS ± 0.7) (Boobier *et al*., 2020b).

Table 1: Developed aqueous solubility predictive models

Model	Accuracy on the Training set	Accuracy of Testing Set	Dataset	Machine learning model	References
ESOL	R^2 = 0.69	R^2 = 0.55	Curated dataset of 2,874 compounds	Linear Regression	(Delaney, 2004)
SolTran Net	Sensitivity = 94.8%	RMSE = 1.459	Second challenge dataset (SC2)	Molecule Attention Transformer (MAT)	(Francoeur & Koes, 2021)
-	RMSE = 0.665.	RMSE of 0.44 on the test set.	ESOL & ChemIDplus dataset.	Graph Convolutional Neural Network (GCN)	(Deng *et al*., 2022)
-	R^2 = 0.88	R^2 = 0.64	Dataset of 8,400 compounds	Random Forest using molecular descriptors	(Tayyebi *et al*., 2023)
	R^2 = 0.81	R^2 = 0.80	Dataset of 8,400 compounds	Random Forest using Morgan fingerprint.	(Tayyebi *et al*., 2023)

R^2 = R-squared; RMSE = Root Mean Square Error

LightGBM, in particular, demonstrated superior performance in predicting solubility in organic solvents (LogS ± 0.20) (Ye & Ouyang, 2021). A consensus model that included LASSO regression and random forests for predicting aqueous solubility struck the

optimum combination between predictive power and complexity (RMSE 0.67, R^2 0.81) (Lovrić *et al*., 2020). These approaches have also provided insights into the physicochemical relationships between solubility and molecular properties, enabling rational improvements in model accuracy (Nguyen, 2021; Boobier *et al*., 2020).

3.1 Overview of ML algorithms commonly used in solubility prediction

Machine learning (ML) algorithms have shown promising results in predicting drug solubility, a critical factor in drug discovery and development. Different supervised learning methods, including regression and classification techniques, have been applied successfully (Gider & Budak, 2023). Ensemble methods, particularly Random Forest (RF), have demonstrated superior performance in solubility prediction tasks (Gider & Budak, 2023; Lovrić *et al*., 2020). Neural networks and deep learning approaches, such as Graph Neural Networks (GNNs) and Multilayer Perceptron (MLP), have also been explored, though with varying degrees of success (Gider & Budak, 2023). Other effective algorithms include Support Vector Machines (SVMs), ExtraTrees, Bagging, and Gaussian Processes (GP) (Boobier *et al*., 2020; Nguyen, 2021). Combining machine learning with computational chemistry and careful feature selection has led to improved accuracy in solubility predictions across various solvents, including water and organic solvents (Boobier *et al*., 2020; Lovrić *et al*., 2020). These advancements contribute to more efficient drug development and chemical process design.

3.2 Feature extraction and representation

Molecular descriptors and fingerprints are very essential for predicting chemical properties and developing machine learning models in drug discovery (Stepisnik *et al*., 2021). Physicochemical traits, and fingerprints such as ECFP and MACCS, and 2D/3D attributes are examples of common representations (Rogers & Hahn, 2010). Neural network-based learnable representations outperform conventional techniques like fingerprints and chemical descriptors (Stepisnik *et al*., 2021). These features can be computationally demanding to calculate, leading to the development of various tools and platforms to streamline the process (Didachos *et al*., 2022; Dong *et al*., 2015). ChemDes, a web-based platform, integrates multiple packages to compute 3679 molecular descriptors and 59 types of fingerprints, offering a user-friendly interface and auxiliary tools for format conversion and similarity calculations (Dong *et al*., 2015). These advancements facilitate the use of molecular representations in various drug discovery applications.

3.2.1 Dataset preparation and preprocessing

This summary examines approaches to handling imbalanced datasets in biomedical and drug discovery research. Imbalanced data, common in high-throughput screening, can negatively impact classification performance (Korkmaz, 2020). Various techniques have been explored to address this issue. Resampling methods, including under-sampling and over-sampling, have shown promise when combined with machine learning algorithms (Xiang *et al*., 2020). The Synthetic Minority Over-sampling Technique (SMOTE) has been effectively coupled with algorithms like GLMBoost to improve classification of rare samples in PubChem BioAssay data (Hao *et al*., 2014). Feature selection has also demonstrated superior performance in many cases, particularly when used with Support Vector Machines (Xiang *et al*., 2020). Additionally, researchers have developed QSAR models using different descriptor types to handle imbalanced high-throughput screening data from PubChem (Zakharov *et al*., 2014). These approaches aim to enhance the detection and classification of active compounds in drug discovery studies.

3.2.2 Advances in Machine Learning Models for Solubility Prediction

Recent developments in machine learning algorithms for predicting solubility have produced promising results (Boobier *et al*.,). Researchers have integrated machine learning (ML) algorithms like Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Random Forests (RFs) with computational chemistry to improve the prediction of solubility in organic solvents and water (Nguyen, 2021). These models achieved accuracy close to experimental noise levels (LogS ± 0.7) and outperformed existing tools. LightGBM demonstrated superior performance (logS ± 0.20) compared to other algorithms for predicting solubility in organic solvents at different temperatures (Ye & Ouyang, 2021). The Second Solubility Challenge revealed that various algorithms could achieve near state-of-the-art performance, with a graph convolutional neural network achieving an RMSE of 0.86 log units (Conn *et al*., 2023). Critical factors for improving prediction accuracy include careful selection of high-quality training data from relevant chemical spaces and the use of appropriate feature sets. However, challenges remain in modeling complex chemical spaces from sparse training data.

3.3.3 Deep learning models for solubility prediction

Recently, various studies have explored deep learning approaches for predicting molecular solubility, a critical property in drug discovery and development (Panapitiya *et al*., 2021). Graph neural networks (GNNs) have shown encouraging results in capturing complex molecular features for solubility prediction (Ahmad *et al*., 2023). Attention-based GNNs and residual GNNs have demonstrated improved performance, with the latter achieving a Pearson correlation coefficient (R2) of 0.79 in cross-validation (Ahmad *et al*., 2024). Three-dimensional GNN models, such as ViSNet, have exhibited exceptional accuracy with an R^2 of 0.9980 for dipole moment prediction (Nguyen & Le, 2024). Transfer learning techniques, incorporating predicted dipole moments and molecular descriptors, have further enhanced solubility prediction accuracy (Nguyen & Le, 2024). While GNNs have shown good performance, models using molecular descriptors have achieved the best results in some studies (Panapitiya *et al*., 2021). These advancements in computational methods offer significant potential for accelerating drug development processes.

3.3.4 Hybrid approaches: Combining ML with physics-based models

Hybrid approaches combining machine learning (ML) with physics-based models are gaining significant grounds in various fields, including drilling automation (Aghito & Bjørkevoll, 2020), multiphase flow modeling (Soedarmo, 2023), cyber-physical systems (Rai & Sahu, 2020), and hydrology (Bhasme *et al*., 2021). These approaches aim to leverage the strengths of both ML and physics-based models, enhancing predictive accuracy while maintaining physical consistency. Hybrid models have shown promising results in predicting downhole pressure (Aghito & Bjørkevoll, 2020), liquid holdup in gas-liquid stratified flow (Soedarmo, 2023), and streamflow (Bhasme *et al*., 2021). They offer improved scalability, extrapolation capabilities, and robustness to noisy data compared to pure ML models (Soedarmo, 2023). However, challenges remain in effectively combining ML and physics-based approaches, and further research is needed to harness the full potential of hybrid models (Rai & Sahu, 2020). Despite these challenges, hybrid approaches show potentials in advancing modeling capabilities across various domains.

3.3.5 Transfer learning for solubility prediction in low-data scenarios

Various studies have recently explored transfer learning approaches to enhance solubility prediction in low-data scenarios. Graph Neural Networks (GNNs) have shown promise in capturing complex molecular features, with 3D GNN models like ViSNet demonstrating exceptional performance (Nguyen & Le, 2024). Pre-trained GNN representations, such as those from SchNet, can be effectively transferred to predict various molecular properties, including solubility, with relatively small training datasets (El-Samman *et al*., 2024). Combining transfer learning with transformer attention mechanisms has also proven competitive for aqueous solubility prediction (Wiercioch & Kirchmair, 2021). However, the success of transfer learning may depend on the similarity between pre-training and fine-tuning tasks, as demonstrated in a study on molecular property predictions (Kirschbaum & Bande, 2024). These approaches offer promising strategies for improving solubility predictions in data-limited scenarios, potentially benefiting drug discovery and materials science applications.

3.3.6 Explainable AI (XAI) for solubility prediction models

Explainable Artificial Intelligence (XAI) has been explored by recent studies as a technique to enhance the interpretability of molecular prediction models, particularly for solubility prediction. Various approaches have been employed, including chemical counterfactuals and descriptor explanations (Wellawatte *et al*., 2023), Shapley Additive exPlanations (SHAP) analysis (Shokrollahi *et al*., 2024; Su & Herrero, 2023), and backpropagation of relevance information (Hödl *et al*., 2023). These methods have been applied to different machine learning models, such as Random Forest (RF) and Transformers, to provide insights into structure-property relationships and feature contributions. Studies have demonstrated that XAI techniques can effectively explain predictions while maintaining high model performance (Su & Herrero, 2023; Shokrollahi *et al*., 2024). The integration of XAI in solubility prediction models has shown promise in guiding compound optimization (Su & Herrero, 2023) and supporting sustainable CO2 storage strategies (Shokrollahi *et al*., 2024), highlighting the potential of XAI in advancing chemical research and decision-making processes.

3.3.7 Evaluation of Machine Learning Models

Machine learning model evaluation is crucial for selecting optimal solutions and assessing performance. Common metrics include classification accuracy, though it has limitations such as neglecting error types and class distribution dependence (Novakovic *et al*., 2017). For regression models, root mean squared error and R^2 are popular statistical metrics (Mishra, 2019). Various techniques exist for model evaluation, selection, and algorithm comparison, including holdout methods, bootstrapping, and cross-validation (Motwani *et al*., 2016). When comparing algorithms, statistical tests and strategies for multiple comparisons are essential, especially for small datasets (Motwani *et al*., 2016). Recent research has focused on evaluation metrics for specific tasks like binary, multi-class, and multi-label classification, regression, image segmentation, object detection, and information retrieval (Rainio *et al*., 2024). Choosing appropriate statistical tests, obtaining sufficient metric values, and correctly interpreting results are crucial for comparing models effectively, particularly in medical imaging applications (Rainio *et al*., 2024).

3.3.8 Common performance metrics

Recent studies have explored various machine learning approaches for predicting drug solubility, a crucial factor in drug discovery and development. Common performance metrics used include mean squared error (MSE), mean absolute error (MAE), and R^2

score (Imane & Amine, 2024; Alobaida & Huwaimel, 2022 Lovrić *et al*., 2020). Avdeef (2020) highlights the importance of understanding different definitions of these statistics, as they can indicate systematic errors in prediction models. Multiple algorithms have been evaluated, including Random Forest, Gradient Boosting, Support Vector Machine, and Multilayer Perceptron (Imane & Amine, 2024; Alobaida & Huwaimel, 2022) Lovrić *et al*. (2020) compared four algorithms and found that LASSO yielded the best predictive ability on an external test set, while Random Forest achieved the best balance between complexity and predictive ability. They also proposed a ranking score combining generalization, number of features, and test performance to select the best model.

3.3.9 Importance of external validation datasets

External validation is crucial for assessing the generalizability and robustness of machine learning models in medicine (Cabitza *et al*., 2021; Rosenblatt *et al*., 2023). While internal validation methods like cross-validation are necessary, they cannot guarantee model quality due to potential biases in training data (Ho *et al*., 2020). External validation, using truly independent datasets, provides a more rigorous evaluation of model performance across different populations (Taylor *et al*., 2008). However, the statistical power of external validation studies is influenced by sample sizes in both training and external datasets, with many prior studies prone to low power (Rosenblatt *et al*., 2023). To assess the soundness of external validation procedures, researchers should consider dataset cardinality and similarity to the training set (Cabitza *et al*., 2021). Multiple performance measures, including discrimination, calibration, and utility, should be used to comprehensively evaluate model performance during external validation (Taylor *et al*., 2008; Cabitza *et al*., 2021).

4.0 Comparison of ML models with traditional QSPR models

Recently, different studies have compared machine learning (ML) approaches with traditional quantitative structure-property relationship (QSPR) models for predicting molecular properties. ML-enhanced QSAR models, particularly tree-based algorithms, have shown improved predictive accuracy and efficiency in drug discovery (Phimonjit *et al*., 2024). For ADME-Tox targets, traditional 1D, 2D, and 3D descriptors outperformed fingerprint-based methods when using XGBoost (Orosz *et al*., 2022). A scalable ML framework integrating multiple steps of QSPR modeling demonstrated superior performance and efficiency compared to literature-reported models for various molecular properties (Chen *et al*., 2023). In predicting human intrinsic clearance, ML models using in vitro inputs achieved comparable or superior performance to mechanistic in vitro-in vivo extrapolation models, with the best ML model showing an average absolute fold error of 2.5 compared to 2.8 for the best mechanistic model (Keefer *et al*., 2023). These findings highlight the potential of ML approaches in enhancing molecular property prediction.

4.1 Applications and Case Studies

AI is frequently used to forecast a drug-like compound's water solubility, eliminating the need for expensive and time-consuming tests (Fatemi *et al*., 2010). Early in the drug development process, it helps find interesting drug candidates by enabling high-throughput screening of sizable chemical libraries (Lovrić *et al*., 2021). By forecasting solubility at various pH levels, AI models also help with drug formulation optimization (Huuskonen *et al*., 1998). Additionally, by altering chemical structures to improve solubility while preserving efficacy, AI-driven methods assist property-based medication design (Lovrić *et al*., 2021). Drug development can be accelerated by using deep learning and

machine learning models based on experimental solubility data, as demonstrated by case studies (Keefer *et al*., 2023).

4.2 Real-world applications of ML-based solubility prediction in drug discovery

Drug development is being revolutionized by machine learning (ML)-based solubility prediction, which improves formulation techniques, boosts high-throughput screening, and optimizes lead compounds (Popova *et al*., 2018). A study that trained deep learning models like EdgeConv and Graph Convolutional Networks (GCN) on 3,942 molecules achieved great predicted accuracy (R2 = 0.918) (Ghanavati *et al*., 2024). Researchers have also created machine learning models utilizing PubChem and OCHEM dataset (Ye & Ouyang, 2021). Another study outperformed conventional approaches with near-experimental accuracy ($\log S \pm 0.20$) by using LightGBM and deep learning to predict solubility in organic solvents. By removing poorly soluble molecules early on, these models aid in cost reduction, drug development acceleration, and bioavailability improvement (Ghanavati *et al*., 2024). ML-based solubility prediction, which makes use of large chemical datasets, guarantees improved drug candidate selection, lowers failure rates, and boosts the effectiveness of pharmaceutical research (Ye & Ouyang, 2021).

4.2 Integration of solubility prediction models into drug development pipelines

By incorporating solubility prediction models into drug development pipelines, researchers can improve early-stage decision-making by identifying compounds with favorable solubility profiles, thereby reducing late-stage failures (Boobier *et al*., 2020). By utilizing machine learning and computational chemistry, these models estimate solubility based on molecular properties, which aids in compound selection and formulation design (Ghanavati *et al*., 2024). They also help tailor formulation strategies, ensuring bioavailability and therapeutic efficacy. In summary, the integration of solubility prediction models streamlines drug discovery, increasing efficiency and success rates (Conn *et al*., 2023).

4.3 Case studies showcasing successful ML models for solubility prediction

Aqueous solubility has been accurately predicted by a number of machine learning models utilizing datasets such as OCHEM and PubChem. ESP maps, molecular graphs, and tabular features were used by Ghanavati *et al*. (2024) to create deep learning models. They used XGBoost to get an R2 of 0.918 and ensemble models to further improve their results (pubs.rsc.org). In order to evaluate several machine learning algorithms and interpretation strategies for solubility prediction, another study combined datasets, including PubChem (explorationpub.com). In their comparison of fingerprint-based and descriptor-based models, Tayyebi *et al*. (2023) found that the former performed better than the latter, with an RMSE of 0.64 and an R2 of 0.88 (jcheminf.biomedcentral.com). These studies show how well various chemical representations and machine learning algorithms work for predicting solubility. Predictive accuracy and interpretability are greatly impacted by the selection of descriptors and machine learning models.

4.4 Challenges and Limitations

Aqueous solubility prediction can be challenging for machine learning (ML) due to the lack of reliable, balanced, and consistent experimental data (Boobier *et al*., 2020). Drug-like compounds contain intricate interactions that simple features might fall short, making it challenging to choose the best molecular descriptors (Mohammad *et al*.,. 2024). Overfitting, poor generalization to new compounds, and interpretability issues are common problems with machine learning models (Xiang *et al*., 2020). Adoption is further hampered by high computing costs and

sluggish incorporation into current drug research pipelines. Model reliability is decreased by challenges with data uniformity, benchmarking, and prediction uncertainty. Improved uncertainty quantification, explainable AI, hybrid physics-ML techniques, and richer datasets are all needed to meet these problems (Panapitiya *et al*., 2021).

4.4.1 Data availability and quality issues

Due to the scarcity of costly, proprietary, and constrained experimental data, machine learning methods for solubility prediction face difficulties (Avdeef, 2020). Data integration is challenging due to inconsistencies in solubility measurements caused by changes in pH, temperature, and solvent conditions (Meng *et al*., 2022). Biased predictions result from the imbalance in many datasets, which include a higher percentage of poorly soluble chemicals (Avdeef, 2020). The lack of standardized databases, experimental noise, and molecular structure errors all contribute to the decreased trustworthiness of the model (Sorkun *et al*., 2020). Additionally, ML models have trouble generalizing to new drug-like compounds due to the low chemical diversity in the datasets that are currently available (Avdeef, 2020). Standardizing data gathering, using physics-based simulations to increase forecast accuracy and dependability, and growing high-quality datasets are all necessary to address these problems (Meng *et al*., 2022).

4.4.2 Overfitting in ML models and lack of generalizability

Models that learn patterns specific to training data but fail to generalize to new compounds are said to be overfitting in machine learning models for predicting aqueous solubility (Zhu, 2024). This is often due to small or biased datasets, excessive model complexity, and redundant or irrelevant features (Charilaou & Battat, 2022). Techniques such as feature selection, regularization (L1/L2), and cross-validation can be used to improve generalizability, while ensemble learning, external validation, and interpretable models further improve robustness and reliability (Bu & Zhang, 2020). Finally, according to Bu and Zhang (2020), striking a balance between model complexity and dataset diversity ensures better predictive performance in real-world drug discovery applications.

4.4.3 Interpretability of complex ML models

Regulatory compliance, trust, and error detection in pharmacology depend on interpretability in machine learning (ML) (Liu *et al*., 2022). However, complex ML models, such as deep learning, frequently function as "black boxes," making their predictions hard to understand (Preuer *et al*., 2019). To improve interpretability, techniques like SHAP, LIME, and attention mechanisms highlight important features that influence drug behavior, which helps with drug discovery, adverse drug reaction prediction, and personalized medicine by providing an explanation for model decisions (Rodríguez-Pérez & Bajorath, 2019). Future developments in causal inference, hybrid models, and standardization will further expand the role of ML in pharmacology (Liu *et al*., 2022).

4.4.4 Computational cost and resource requirements

The computational cost of utilizing machine learning to predict the aqueous solubility of drug-like molecules varies based on feature selection, model complexity, and dataset size (Xue *et al*., 2024). While deep learning models like GNNs and transformers require powerful GPUs and a lot of memory, simple models like decision trees and linear regression require very little (Mitchell & Jurs, 1998). Simple models take minutes to train, while deep models take days or weeks, and cloud computing is frequently needed for scalability (Ghasemi & Saaidpour, 2007). Techniques for feature selection, transfer learning, and model compression can all help to maximize computing cost (Lovrić *et al*., 2021). Scalable solutions for effectively managing high-

resource machine learning activities are offered by cloud services like AWS, Google Cloud, and Colab (Xue *et al*., 2024).

4.4.5 Future Perspectives and Research Directions

Predicting aqueous solubility has been greatly enhanced by machine learning (ML); yet, issues with data quality, model accuracy, and generalizability still exist. Future studies should concentrate on creating standardized, high-quality datasets and use cutting-edge designs such as transformer-based models and graph neural networks (GNNs) (Mitchell & Jurs, 1998). Predictive power can be increased by using better molecular representations, such as deep embeddings and quantum-informed features. Hybrid rule-based models and explainability strategies like SHAP will aid in interpreting ML predictions for increased acceptance and confidence (Rodríguez-Pérez & Bajorath, 2019).. With little experimental data, data-efficient learning strategies like few-shot learning and active learning can boost performance (Liu *et al*., 2022). Lastly, the development of soluble, drug-like molecules will be accelerated by combining machine learning (ML) with multi-task drug discovery pipelines and physics-based simulations (Liu *et al*., 2022).

4.4.6 Advancements in data collection and generation for solubility prediction

Recent advancements in data collection for solubility prediction using machine learning (ML) include high-throughput experimental techniques, standardized measurements, and expanded public and proprietary datasets (Meng *et al*., 2022). Computational methods such as molecular dynamics simulations, quantum chemistry calculations, and hybrid ML-physics models generate additional solubility data (Vermeire *et al*., 2022). Generative AI techniques, including GANs and self-supervised learning, help augment datasets and improve model generalization. Improved molecular representations, such as graph-based models and transformer-driven embeddings, enhance predictive accuracy (Sorkun *et al*., 2020). These advancements collectively enable more reliable solubility prediction models, accelerating drug discovery and formulation (Meng *et al*., 2022).

4.4.7 Development of interpretable and robust ML models

Systematic data collection and feature extraction from carefully selected datasets, like ESOL and AQUASOL, are necessary for the creation of interpretable and reliable machine learning models for forecasting the aqueous solubility of drug-like molecules (Alvarez-Melis & Jaakkola, 2018). Traditional methods like Random Forests and XGBoost are used in the model selection process, as are more sophisticated structures like Graph Neural Networks (GNNs) (Mitchell & Jurs, 1998). Robustness is attained using uncertainty quantification techniques, and interpretability is guaranteed by SHAP values (Amirian *et al*., 2020). To guarantee generalizability, the model's performance is thoroughly assessed using metrics like RMSE and MAE in addition to cross-validation and external validation (Linardatos *et al*., 2020). Lastly, the trained models can be made available as web apps or APIs, which would make it easier to include them into ADMET predictions and other drug development processes (Linardatos *et al*., 2020).

4.4.8 Leveraging AI-driven methods like active learning for efficient data utilization

An AI-driven method called active learning (AL) lessens reliance on massive annotated datasets by deciding on probing the most informative examples for labeling, improving data efficiency (Dwivedi *et al*., 2019). To maximize model training, AL can strategically prioritize diverse, unclear, or representative chemical structures when predicting the aqueous solubility of drug-like molecules (Nedelkoski *et al*., 2019). While reducing experimental and computational expenses, this

focused selection procedure makes it easier to create machine learning (ML) models that are more stable and interpretable (Verrelst *et al*., 2016). Researchers may create models more quickly and with less resource usage by combining AL with solubility prediction frameworks, which allows for high predictive accuracy with less data points (Nedelkoski *et al*., 2019). As a result, the use of AL in this field may speed up medication formulation and discovery by making solubility forecasts more accurate and data-efficient (Verrelst *et al*., 2016).

4.4.9 Potential of quantum computing in enhancing predictive performance

By improving linear algebra operations and speeding up computer processes through quantum parallelism, quantum computing holds promise for improving machine learning (ML) predictive performance (Panyaram, 2024). By embedding classical data into high-dimensional quantum states, quantum feature maps enhance the ability to identify patterns in intricate datasets (Palle, 2020). Furthermore, by effectively resolving high-dimensional optimization issues, quantum optimization algorithms—like the Quantum Approximate Optimization Algorithm (QAOA)—can improve model training (Li *et al*., 2024). These developments may be especially helpful in fields where strong predictive modeling is necessary due to molecular complexity, such as drug discovery (Palle, 2020). Efficiency, scalability, and predictive accuracy across a range of applications are anticipated to increase with the maturation of quantum hardware and its integration with machine learning (Li *et al*., 2024).

5.0 Conclusion

Machine learning (ML) has significantly advanced aqueous solubility prediction by leveraging algorithms like Random Forest, Support Vector Machines, deep learning models, and hybrid approaches, improving predictive accuracy and efficiency. These advancements have enabled rapid screening of large chemical libraries, optimizing drug formulation and reducing costly experimental failures. The integration of ML-based solubility prediction into drug discovery pipelines enhances early-stage decision-making, minimizing late-stage attrition and improving bioavailability. Despite these successes, challenges like data quality, model interpretability, and generalizability persist, requiring further research into explainable AI, active learning, and quantum computing techniques. Future solubility prediction models will likely incorporate more robust datasets, physics-informed ML, and transfer learning to refine accuracy and applicability. Ultimately, the continued evolution of ML-driven solubility prediction holds great potential to revolutionize pharmaceutical research, accelerating the development of effective and well-soluble drug candidates.

6.0 Reference

Aghito, M., & Bjørkevoll, K. S. (2020). Hybrid approach for drilling automation. *Day 3 Tue, November 30, 2021*. https://doi.org/10.2118/200734-ms

Ahmad, W., Tayara, H., & Chong, K. T. (2023). Attention-based graph neural network for molecular solubility prediction. *ACS Omega, 8*(3), 3236–3244. https://doi.org/10.1021/acsomega.2c06702

Ahmad, W., Tayara, H., Shim, H., & Chong, K. T. (2024). SolPredictor: Predicting solubility with residual gated graph neural network. *International Journal of Molecular Sciences, 25*(2), 715. https://doi.org/10.3390/ijms25020715

Alobaida, A., & Huwaimel, B. (2022). Analysis of enhancing drug bioavailability via nanomedicine production approach using green chemistry route: Systematic assessment of drug candidacy. *Journal of Molecular Liquids, 370*, 120980. https://doi.org/10.1016/j.molliq.2022.120980

Alvarez-Melis, D., & Jaakkola, T. S. (2018). On the robustness of interpretability methods. *arXiv*. https://www.microsoft.com/en-us/research/publication/on-the-robustness-of-interpretability-methods/

Amirian, M., Tuggener, L., Chavarriaga, R., Satyawan, Y. P., Schilling, F., Schwenker, F., & Stadelmann, T. (2020). Two to trust: AutoML for safe modelling and interpretable deep learning for robustness. *ACS*, 268–275. https://doi.org/10.21256/zhaw-20217

Avdeef, A. (2020). Do you know your r²? *ADMET & DMPK*. https://doi.org/10.5599/admet.888

Balakin, K., Savchuk, N., & Tetko, I. (2005). In silico approaches to prediction of aqueous and DMSO solubility of drug-like compounds: Trends, problems, and solutions. *Current Medicinal Chemistry*, *13*(2), 223–241. https://doi.org/10.2174/092986706775197917

Bhasme, P., Vagadiya, J., & Bhatia, U. (2022). Enhancing predictive skills in a physically consistent way: Physics-informed machine learning for hydrological processes. *Journal of Hydrology*, *615*, 128618. https://doi.org/10.1016/j.jhydrol.2022.128618

Boobier, S., Hose, D. R. J., Blacker, A. J., & Nguyen, B. N. (2020a). Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19594-z

Boobier, S., Hose, D. R. J., Blacker, A. J., & Nguyen, B. N. (2020b). Machine learning with physicochemical relationships: Solubility prediction in organic solvents and water. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19594-z

Bosselmann, S., & Williams, R. O. (2011). Route-specific challenges in the delivery of poorly water-soluble drugs. In *AAPS Advances in the Pharmaceutical Sciences Series* (pp. 1–26). https://doi.org/10.1007/978-1-4614-1144-4_1

Bu, C., & Zhang, Z. (2020). Research on overfitting problem and correction in machine learning. *Journal of Physics Conference Series*, *1693*(1), 012100. https://doi.org/10.1088/1742-6596/1693/1/012100.

Cabitza, F., Campagner, A., Soares, F., De Guadiana-Romualdo, L. G., Challa, F., Sulejmani, A., Seghezzi, M., & Carobene, A. (2021). The importance of being external: Methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, *208*, 106288. https://doi.org/10.1016/j.cmpb.2021.106288

Charilaou, P., & Battat, R. (2022). Machine learning models and over-fitting considerations. *World Journal of Gastroenterology*, *28(5)*, 605–607. https://doi.org/10.3748/wjg.v28.i5.605

Chen, G., Song, Z., Qi, Z., & Sundmacher, K. (2023). A scalable and integrated machine learning framework for molecular properties prediction. *AIChE Journal*, *69(10)*. https://doi.org/10.1002/aic.18185

Chen, X., Cho, S. J., Li, Y., & Venkatesh, S. (2002). Prediction of aqueous solubility of organic compounds using a quantitative structure–property relationship. *Journal of Pharmaceutical Sciences*, *91(8)*, 1838–1852. https://doi.org/10.1002/jps.10178

Cheng, A., & Merz, K. M. (2003). Prediction of aqueous solubility of a diverse set of compounds using quantitative structure−property relationships. *Journal of Medicinal Chemistry*, *46(17)*, 3572–3580. https://doi.org/10.1021/jm020266b

Conn, J. G. M., Carter, J. W., Conn, J. J. A., Subramanian, V., Baxter, A., Engkvist, O., Llinas, A., Ratkova, E. L., Pickett, S. D., McDonagh, J. L., & Palmer, D. S. (2023).

Blinded predictions and post hoc analysis of the second solubility challenge data: Exploring training data and feature set selection for machine and deep learning models. *Journal of Chemical Information and Modeling, 63(4),* 1099–1113. https://doi.org/10.1021/acs.jcim.2c01189

Dearden, J. C. (2006). *In silico* prediction of aqueous solubility. *Expert Opinion on Drug Discovery, 1(1),* 31–52. https://doi.org/10.1517/17460441.1.1.31

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

Deng, C., Liang, L., Xing, G., Hua, Y., Lu, T., Zhang, Y., Chen, Y., & Liu, H. (2022). 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

Didachos, C., Kintos, D. P., Fousteris, M., Gerogiannis, V. C., Son, L. H., & Kanavos, A. (2022). A cloud-based distributed computing approach for extracting molecular descriptors. *Chemometrics, 75,* 1–6. https://doi.org/10.1145/3564982.3565004

Dong, J., Cao, D., Miao, H., Liu, S., Deng, B., Yun, Y., Wang, N., Lu, A., Zeng, W., & Chen, A. F. (2015). ChemDes: An integrated web-based platform for molecular descriptor and fingerprint computation. *Journal of Cheminformatics, 7(1).* https://doi.org/10.1186/s13321-015-0109-z

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., ... Williams, M. D. (2019). Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management, 57,* 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Elder, D., & Holm, R. (2012). Aqueous solubility: Simple predictive methods (*in silico*, *in vitro*, and bio-relevant approaches). *International Journal of Pharmaceutics, 453(1),* 3–11. https://doi.org/10.1016/j.ijpharm.2012.10.041

El-Samman, A., De Castro, S., Morton, B., & De Baerdemacker, S. (2024). Transfer learning graph representations of molecules for pKa, 13C-NMR, and solubility. *Canadian Journal of Chemistry, 102(4),* 275–288. https://doi.org/10.1139/cjc-2023-0152

Etzweiler, F., Senn, E., & Schmidt, H. W. H. (1995). Method for measuring aqueous solubilities of organic compounds. *Analytical Chemistry, 67(3),* 655–658. https://doi.org/10.1021/ac00099a025

Fatemi, M. H., Heidari, A., & Ghorbanzade, M. (2010). Prediction of aqueous solubility of drug-like compounds by using an artificial neural network and least-squares support vector machine. *Bulletin of the Chemical Society of Japan, 83(11),* 1338–1345. https://doi.org/10.1246/bcsj.20100074

Francoeur, P. G., & Koes, D. R. (2021). SolTranNet–A machine learning tool for fast aqueous solubility prediction. *Journal of Chemical Information and Modeling, 61(6),* 2530–2536. https://doi.org/10.1021/acs.jcim.1c00331.

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.* https://doi.org/10.1039/d4dd00065j

Ghasemi, J., & Saaidpour, S. (2007). QSPR prediction of aqueous solubility of drug-like organic compounds. *Chemical and Pharmaceutical Bulletin*, *55*(4), 669–674. https://doi.org/10.1248/cpb.55.669

Gider, V., & Budak, C. (2024). Drug solubility prediction: A comparative analysis of GNN, MLP, and traditional machine learning algorithms. *Gazi Üniversitesi Fen Bilimleri Dergisi Part C Tasarım ve Teknoloji*, *12*(1), 164–175. https://doi.org/10.29109/gujsc.1371519

Hao, M., Wang, Y., & Bryant, S. H. (2013). An efficient algorithm coupled with synthetic minority over-sampling technique to classify imbalanced PubChem BioAssay data. *Analytica Chimica Acta*, *806*, 117–127. https://doi.org/10.1016/j.aca.2013.10.050

Hewitt, M., Cronin, M. T. D., Enoch, S. J., Madden, J. C., Roberts, D. W., & Dearden, J. C. (2009). In silico prediction of aqueous solubility: The solubility challenge. *Journal of Chemical Information and Modeling*, *49*(11), 2572–2587. https://doi.org/10.1021/ci900286s

Ho, S. Y., Phua, K., Wong, L., & Goh, W. W. B. (2020). Extensions of the external validation for checking learned model interpretability and generalizability. *Patterns*, *1*(8), 100129. https://doi.org/10.1016/j.patter.2020.100129

Hödl, S., Robinson, W., Bachrach, Y., Huck, W., & Kachman, T. (2023). Explainability techniques for chemical language models. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2305.16192

Hörter, D., & Dressman, J. B. (2001). Influence of physicochemical properties on dissolution of drugs in the gastrointestinal tract. *Advanced Drug Delivery Reviews*, *46*(1–3), 75–87. https://doi.org/10.1016/S0169-409X(00)00130-7

Huuskonen, J., Livingstone, D., & Manallack, D. (2008). Prediction of drug solubility from molecular structure using a drug-like training set. *SAR and QSAR in Environmental Research*, *19*(3–4), 191–212. https://doi.org/10.1080/10629360802083855

Huuskonen, J., Salo, M., & Taskinen, J. (1998). Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. *Journal of Chemical Information and Computer Sciences*, *38*(3), 450–456. https://doi.org/10.1021/ci97010 -0x

Imane, A., & Amine, B. (2024). SolvPredict: A comprehensive exploration of predictive models for molecule solubility. *Chemometrics*, 1–8. https://doi.org/10.1109/iscv60512.2024.10620130

Kakran, M., Shegokar, R., Sahoo, N. G., Al Shaal, L., Li, L., & Müller, R. H. (2012). Overcoming the challenge of poor drug solubility. *Pharmaceutical Engineering*, *32*(2), 1–8.

Kalepu, S., & Nekkanti, V. (2015). Insoluble drug delivery strategies: Review of recent advances and business prospects. *Acta Pharmaceutica Sinica B*, *5*(5), 442–453. https://doi.org/10.1016/j.apsb.2015.07.003

Keefer, C. E., Chang, G., Di, L., Woody, N. A., Tess, D. A., Osgood, S. M., Kapinos, B., Racich, J., Carlo, A. A., Balesano, A., Ferguson, N., Orozco, C., Zueva, L., & Luo, L. (2023). The comparison of machine learning and mechanistic in vitro–in vivo extrapolation models for the prediction of human intrinsic clearance. *Molecular Pharmaceutics*, *20*(11), 5616–5630. https://doi.org/10.1021/acs.molpharmaceut.3c00502

Kirschbaum, T., & Bande, A. (2024). Transfer learning for molecular property predictions from small data sets. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2404.13393

Könczöl, Á., & Dargó, G. (2018). Brief overview of solubility methods: Recent trends in equilibrium solubility

measurement and predictive models. *Drug Discovery Today Technologies, 27*, 3–10. https://doi.org/10.1016/j.ddtec.2018.06.001

Korkmaz, S. (2020). Deep learning-based imbalanced data classification for drug discovery. *Journal of Chemical Information and Modeling, 60*(9), 4180–4190. https://doi.org/10.1021/acs.jcim.9b01162

Kuentz, M. T., & Arnold, Y. (2009). Influence of molecular properties on oral bioavailability of lipophilic drugs – Mapping of bulkiness and different measures of polarity. *Pharmaceutical Development and Technology, 14*(3), 312–320. https://doi.org/10.1080/10837450802626296

Kumar, S., Bhargava, D., Thakkar, A., & Arora, S. (2013). Drug carrier systems for solubility enhancement of BCS Class II drugs: A critical review. *Critical Reviews in Therapeutic Drug Carrier Systems, 30*(3), 217–256. https://doi.org/10.1615/critrevtherdrugcarriersyst.2013005964

Lam, S. K., Pan, B., Qin, A., & Chen, D. (2024). Advancing agroecosystem modelling of nitrogen losses with machine learning. *Deleted Journal, 1*(1), 100006. https://doi.org/10.1016/j.ecz.2024.100006

Li, K., Zhao, P., Dai, S., Zhu, A., Hong, B., Liu, J., Wei, C., Huang, W., & Zhang, Y. (2024). Exploring the impact of quantum computing on machine learning performance. *Middle East Journal of Applied Science & Technology, 07*(02), 145–161. https://doi.org/10.46431/mejast.2024.7215

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), 18. https://doi.org/10.3390/e23010018

Liu, G., Lu, J., Lim, H. S., Jin, J. Y., & Lu, D. (2022). Applying interpretable machine learning workflow to evaluate exposure–response relationships for large-molecule oncology drugs. *CPT Pharmacometrics & Systems Pharmacology, 11*(12), 1614–1627. https://doi.org/10.1002/psp4.12871

Lovrić, M., Pavlović, K., Žuvela, P., Spataru, A., Lučić, B., Kern, R., & Wong, M. W. (2021). Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *Journal of Chemometrics, 35*(7–8). https://doi.org/10.1002/cem.3349.

Verrelst, J., Dethier, S., Rivera, J. P., Muñoz-Marí, J., Camps-Valls, G., & Moreno, J. (2016). Active learning methods for efficient hybrid biophysical variable retrieval. *IEEE Geoscience and Remote Sensing Letters, 13*(7), 1012–1016. https://doi.org/10.1109/lgrs.2016.2560799

Vovsha, P. (2021). Comparison of traditional econometric models and machine learning methods in the context of travel decision making and perspectives for synergy. In *Studies in Computational Intelligence* (pp. 177–186). Springer. https://doi.org/10.1007/978-3-030-75583-6_18

Walker, M. A. (2013). Improving solubility via structural modification. In *Topics in Medicinal Chemistry* (pp. 69–106). Springer. https://doi.org/10.1007/7355_2013_32

Wang, J., & Hou, T. (2011). Recent advances on aqueous solubility prediction. *Combinatorial Chemistry & High Throughput Screening, 14*(5), 328–338. https://doi.org/10.2174/138620711795508331

Wellawatte, G. P., Gandhi, H. A., Seshadri, A., & White, A. D. (2023). A perspective on explanations of molecular prediction models. *Journal of Chemical Theory and Computation, 19*(8), 2149–2160. https://doi.org/10.1021/acs.jctc.2c01235

Wiercioch, M., & Kirchmair, J. (2021). Dealing with a data-limited regime:

Combining transfer learning and transformer attention mechanism to increase aqueous solubility prediction performance. *Artificial Intelligence in the Life Sciences*, *1*, 100021. https://doi.org/10.1016/j.ailsci.2021.100021

Xiang, D., Zhou, W., Xiang, Y., Zhang, J., & Liu, S. (2020). A study of data pre-processing techniques for imbalanced biomedical data classification. *International Journal of Bioinformatics Research and Applications*, *16*(3), 290. https://doi.org/10.1504/ijbra.2020.10031330

Xie, B., Liu, Y., Li, X., Yang, P., & He, W. (2024). Solubilization techniques used for poorly water-soluble drugs. *Acta Pharmaceutica Sinica B*, *14*(11), 4683–4716. https://doi.org/10.1016/j.apsb.2024.08.027

Xue, N., Zhang, Y., & Liu, S. (2024). Evaluation of machine learning models for aqueous solubility prediction in drug discovery. *bioRxiv*. https://doi.org/10.1101/2024.06.10.598383

Ye, Z., & Ouyang, D. (2021). Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms. *Journal of Cheminformatics*, *13*(1). https://doi.org/10.1186/s13321-021-00575-3

Yousefinejad, S., Honarasa, F., & Chaabi, M. (2016). New relationship models for solvent–pyrene solubility based on molecular structure and empirical properties. *New Journal of Chemistry*, *40*(12), 10197–10207. https://doi.org/10.1039/c6nj02319c

Zakharov, A. V., Peach, M. L., Sitzmann, M., & Nicklaus, M. C. (2014). QSAR modeling of imbalanced high-throughput screening data in PubChEM. *Journal of Chemical Information and Modeling*, *54*(3), 705–712. https://doi.org/10.1021/ci400737s

Zhu, Z. (2024). Systematic optimization of overfitting problem in machine learning. *Highlights in Science Engineering and Technology*, *111*, 353–359. https://doi.org/10.54097/3tkzrj84

Compliance with Ethical Standards

Declaration

Ethical Approval

Not Applicable

Competing interests

The authors declare that they have no known competing financial interests

Funding

The authors declared no external source of funding

Author's Contributions

Comfort M. Ngwu Supervision, Conceptualization, Writing—original draft Daniel C. Ottah Conceptualization, Writing-original draft, Praise O. Otitodirichukwu Writing-original draft, Writing-review & editing, Chioma Harbor Writing—original draft., Writing-review & editing. All authors read and approved the final manuscript.