Artificial Intelligence In Solubility Prediction For Poorly Soluble Drug.

Surve Avijita Jitendra[1], Suryawanshi Jagruti[1], Deepika Patil[2], Shinde Ajinkya[1], Shirude Darshan[1], Sonawane Darshan[1]

[1] B.Pharm Students

[2] Professor Department of Pharmacy,

Loknete J D Pawar College of Pharmacy, Nashik, India

Abstract:

Solubility is one of the most important physicochemical properties in pharmaceutical development. It refers to the ability of a substance—such as a drug compound—to dissolve in a given solvent, typically water for oral medications. For a drug to be effective, it must first dissolve in the body's fluids so that it can be absorbed into the bloodstream and reach its site of action. However, many newly discovered drug molecules are poorly soluble in water. In fact, more than 40% of currently marketed drugs and up to 90% of molecules in development face solubility-related challenges. Poor aqueous solubility often leads to low bioavailability, inconsistent drug absorption, and therapeutic failure, which in turn increases the time, cost, and risk of drug development.

To overcome these challenges, solubility enhancement has become a major focus in pharmaceutical formulation. Several traditional approaches have been employed, including salt formation, particle size reduction, solid dispersions, use of surfactants, and complexation. However, these methods can be time-consuming, expensive, and sometimes ineffective—especially during early-stage drug discovery when thousands of candidate molecules must be screened quickly.

This is where Artificial Intelligence (AI) and Machine Learning (ML) are proving to be game-changers. These computational methods can predict the solubility of a drug molecule based on its chemical structure, saving time and reducing the need for extensive lab testing. ML models are trained on large datasets containing known drug solubility values and learn to recognize patterns that influence solubility, such as molecular size, polarity, and functional groups. Algorithms like Random Forests, Neural Networks, and Support Vector Machines have shown promising results in predicting solubility with high accuracy. For example, Lovrić et al. (2021) reported that Random Forest models achieved an R^2 of 0.83 and RMSE of 0.75–1.05 log units when predicting intrinsic solubility from molecular descriptors [1].

In addition to these traditional approaches, there is growing interest in using green and sustainable technologies to enhance solubility. One such method is supercritical carbon dioxide ($SCCO_2$), a non-toxic, environmentally friendly solvent used to improve solubility and create advanced formulations such as drug nanoparticles. AI models are now being applied to predict how different drugs behave in $SCCO_2$ systems. A recent study by Jamshidi et al. (2024) used machine learning models to estimate the solubility of ketoprofen in $SCCO_2$, achieving an R^2 of over 0.94, which indicates strong predictive performance [2].

AI doesn't just improve prediction—it also supports continuous manufacturing, which is a modern approach to drug production that is faster, more efficient, and less wasteful compared to traditional batch methods. AI models can help control key process parameters such as pressure, temperature, and mixing rates in real time, making them ideal tools for future-ready pharmaceutical development.

In summary, poor solubility is a major barrier in drug development, but with the help of AI and ML, researchers can now predict solubility more accurately, screen molecules faster, and design better formulations using both conventional and green technologies. This review brings together key findings from both traditional ML models and recent green-solvent approaches, showing how AI is reshaping the way we handle poorly soluble drugs—making the process faster, more sustainable, and more reliable.

Indexterms – Solubility prediction, Artificial intelligence, Machine Learning, Deep Learning, Poorly soluble drugs, Quality by Design, Drug Development Bioavailability

Introduction

Solubility is one of the most fundamental and essential properties in pharmaceutical science. It refers to the ability of a drug substance to dissolve in a given solvent, most commonly water. For any drug taken orally, solubility in water is critical—because a drug must first dissolve in the fluids of the gastrointestinal tract before it can be absorbed into the bloodstream and reach its target in the body. Without proper solubility, even the most pharmacologically effective drug can fail to deliver its therapeutic effect.

Unfortunately, poor solubility is a common and widespread issue in drug development. Studies show that more than 40% of currently marketed drugs and nearly 90% of new drug candidates suffer from low water solubility [1, 2]. These poorly soluble drugs often exhibit low bioavailability, which means only a small portion of the drug actually reaches the bloodstream. This not only reduces treatment effectiveness but can also require higher doses, increase side effects, and delay approval timelines. In some cases, poor solubility becomes a reason to abandon a drug candidate altogether. To address this, researchers have developed various solubility enhancement techniques. Traditional methods include: Salt formation, which involves converting the drug into a more soluble salt form.

Particle size reduction, which increases surface area for faster dissolution. Solid dispersions, where the drug is dispersed in a carrier matrix.

Use of surfactants, which reduce surface tension and improve wetting.

Complexation, such as with cyclodextrins, which form inclusion complexes with drug molecules.

While these methods are widely used, they are often time-consuming, expensive, and limited to specific types of compounds. More importantly, they typically require experimental trial and error, which slows down the drug development pipeline.

This is where Artificial Intelligence (AI) and Machine Learning (ML) are revolutionizing the field. These technologies offer a predictive, data-driven approach to solving solubility problems, especially in the early stages of drug discovery. Instead of relying solely on laboratory testing, researchers can now use AI models to estimate solubility based on a compound's chemical structure, molecular properties, and historical data. These predictions help scientists screen thousands of molecules quickly and focus only on the most promising candidates for formulation and development.

For example, Lovrić et al. (2021) applied various ML algorithms—such as Random Forests and Support Vector Machines—to predict intrinsic aqueous solubility using molecular descriptors. Their study demonstrated that well- trained models could achieve strong predictive performance, with R^2 values up to 0.83 and root-mean-square error (RMSE) values between 0.7 and 1.05 log units [3]. These results are comparable to experimental accuracy and show how AI can significantly reduce time, cost, and material waste.

Beyond conventional solubility prediction, researchers are also exploring green technologies for solubility enhancement. One such approach is the use of supercritical carbon dioxide ($SCCO_2$)—a sustainable, non-toxic, and eco-friendly solvent that operates above its critical temperature and pressure to act as a tunable fluid. $SCCO_2$ has unique properties that make it ideal for forming nanoparticles, solid dispersions, and inclusion complexes to improve solubility.

However, optimizing $SCCO_2$-based processes is complex due to the number of interacting variables, such as pressure, temperature, and drug–solvent interactions. AI and ML are now being used to model and optimize these processes. A recent study by Jamshidi et al. (2024) applied ML techniques to predict the solubility of ketoprofen in $SCCO_2$. Their models, including Artificial Neural Networks and regression-based algorithms, achieved high accuracy with R^2 values exceeding 0.94 [4]. This proves that AI can support not only prediction but also optimization of environmentally friendly formulation methods.

In addition, AI is becoming an important part of continuous manufacturing—a modern production method that operates 24/7 without the need for stopping between batches. In such settings, real-time prediction of solubility and other critical quality attributes can improve efficiency, reduce waste, and ensure consistent product quality. AI models can be integrated into digital control systems to monitor and adjust process variables automatically, making pharmaceutical production smarter and more sustainable.

Despite its many benefits, AI-based solubility prediction does have challenges. The quality and quantity of data

are crucial. Poorly curated or limited datasets can lead to inaccurate predictions. Moreover, some AI models—especially deep learning—can be difficult to interpret, which poses problems for regulatory acceptance. Nevertheless, with growing interest in hybrid models, combining physical chemistry with AI, and improvements in explainable AI, these concerns are gradually being addressed.

In conclusion, solubility remains a key bottleneck in pharmaceutical development, especially for poorly soluble drugs. Traditional methods for solubility enhancement, though effective in some cases, are not always practical or efficient. The integration of AI and ML is providing a faster, smarter, and more cost-effective way to predict solubility, optimize formulations, and improve manufacturing processes. By combining traditional data-driven models with emerging green technologies like $SCCO_2$, AI is helping researchers tackle solubility challenges in a more sustainable and scalable manner—opening up new possibilities for drug development in the 21st century.

Objectives:

1. To review conventional and advanced solubility enhancement strategies for poorly soluble drugs (Khatri et al., Solubility Enhancement Techniques: An Overview, 2022; Singh et al., Solubility: An Overview, IJPharmChem Analysis, 2020).

2. To develop AI-based predictive models (machine learning and deep learning) using physicochemical descriptors and molecular properties for solubility prediction (Wang & Zou, Prediction of Protein Solubility with DeepSoluE, BMC Biology, 2023; Ngwu et al., 2025).

3. To integrate predictive modeling with formulation design frameworks such as QbD to optimize formulations of poorly soluble drugs (Dawoud et al., 2023).

4. To validate the AI model against benchmark solubility datasets and compare performance with traditional statistical models (Wang & Zou, 2023).

5. To provide a framework for reducing trial-and-error experiments in pharmacy.

6. Investigate the limitations of traditional experimental and descriptor-based computational solubility prediction methods [7] .

7. Leverage curated datasets such as AqSolDB and ESOL to train and validate AI-driven models for solubility prediction [8].

8. Implement GNN-based architectures that capture molecular structure, bond connectivity, and spatial conformers for enhanced predictive accuracy [9].

9. Evaluate model performance using standard metrics such as Root Mean Squared Error (RMSE) and compare results with existing state-of-the-art methods [10].

10. Explore attention mechanisms and interpretability approaches to identify atom- or bond-level features contributing to solubility [11].

11. Propose AI-based frameworks as reliable, efficient, and cost-effective alternatives to experimental methods, thereby accelerating early drug discovery and development [12].

Methodology:

1. Literature Review

- Comprehensive study of poorly soluble drugs, their challenges in oral bioavailability, and the limitations of conventional solubility enhancement methods such as salt formation, particle size reduction, and use of surfactants [7], [8]
- Review of artificial intelligence (AI) and machine learning (ML) applications in solubility prediction and pharmaceutical development, highlighting successes of Graph Neural Networks (GNNs) and deep learning models [9], [10].

2. Data Collection and Preparation

- Compilation of aqueous solubility datasets (logS values) from open-access repositories such as AqSolDB, ESOL, and Huuskonen datasets [11].
- Extraction of relevant physicochemical and structural descriptors, including molecular weight, LogP, hydrogen bond donors/acceptors, topological polar surface area (TPSA), and rotatable bonds [12].
- Data preprocessing through removal of duplicates, handling missing values, normalization, and dimensionality reduction. Feature selection was conducted using methods such as genetic algorithms and principal component analysis (PCA) [13].

3. Model Development

- Machine Learning Models: Implementation of Random Forest (RF), Support Vector Machines (SVM), and Gradient Boosting for baseline comparisons [13].
- Deep Learning Models: Application of Long Short-Term Memory (LSTM) networks and feed-forward deep neural networks (DNNs) for capturing complex non-linear relationships [14].
- Graph Neural Network (GNN) Models: Utilization of Atomistic Line Graph Neural Network (ALIGNN) and Molecular Attention Transformer (MAT) to incorporate molecular graphs and spatial conformers for high-accuracy solubility prediction [15].
- Model training with 80/10/10 data split (train/validation/test) and optimization via hyperparameter tuning (grid search, learning rate scheduling) [16].

4. Integration with QbD (Quality by Design)

- Embedding AI solubility prediction within a Quality by Design (QbD) framework for pharmaceutical formulation [17].
- Linking predicted solubility values to Critical Material Attributes (CMA) such as particle size and Critical Process Parameters (CPP) like mixing speed and temperature [18].

5. Validation and Performance Evaluation

- Model evaluation using statistical performance metrics: Coefficient of Determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) [19].
- External validation against independent datasets not used in training, and benchmarking

AI TECHNIQUES IN SOLUBILITY PREDICTION

a) Machine learning –

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as powerful tools in predicting drug solubility, especially for poorly soluble compounds. These techniques allow scientists to use computational models to predict solubility based on a molecule's structure, physicochemical properties, and experimental data, without the need for extensive lab testing. Several ML algorithms and deep learning architectures have been successfully applied in solubility prediction, each with its own strengths and limitations.

1. Linear Regression (LR)

2. Support Vector Machines (SVM)
3. Random Forest (RF)
4. Gradient Boosting Machines (GBM, XGBoost, LightGBM)
5. Artificial Neural Networks (ANN)

b) Deep learning

1. Artificial Neural Networks (ANN) (also considered ML)
2. Graph Neural Networks (GNN)
3. Transformers, CNNs, RNNs, etc.

c) Transfer Learning & Data Augmentation

1. Linear Regression (LR)

Linear regression is a basic machine learning method that tries to find a straight-line relationship between molecular descriptors (like molecular weight, logP, number of hydrogen bond donors) and solubility. While simple and easy to interpret, linear models often **fail to capture complex non-linear relationships**, especially in larger and more diverse datasets. Linear regression (LR) and its regularized variants (Ridge, LASSO) remain fundamental baselines in QSPR solubility modelling. LR maps a linear combination of molecular descriptors—lipophilicity (logP), molecular weight, polar surface area, hydrogen bond counts—onto a continuous solubility measure (commonly logS). The appeal of LR lies in interpretability: coefficients directly quantify how each descriptor shifts predicted solubility, offering mechanistic insight during lead optimization. Lipinski's work (2000) established many descriptors commonly used in LR models, which helps explain why simple linear models can still capture large fractions of variance for drug-like series. However, intrinsic solubility is governed by non-linear phenomena (ionization equilibria, crystal lattice energy, solvent interactions), limiting LR's expressiveness for chemically diverse or poorly soluble compounds. When combined with penalization (LASSO) and rigorous feature selection, LR can perform surprisingly well; Lovrić et al. (2021) reported LASSO achieving competitive RMSE and R^2 on curated solubility datasets, particularly when the descriptor set is pruned to informative features. LR is therefore best used as (1) an interpretable baseline, (2) a tool for descriptor screening, and (3) a component of ensemble/consensus pipelines where its bias complements more flexible learners. For formulation-aware problems (e.g., salt forms), LR's linearity may miss critical interactions described in Serajuddin (2007), so it is usually combined with non-linear models for deployment. **Refs:** Lipinski 2000; Serajuddin 2007; Lovrić et al. 2021.

2. Support Vector Machines (SVM)

SVMs are supervised learning models that work well for **classification and regression** tasks. In solubility prediction, SVMs can separate high- and low-solubility compounds by creating a boundary (hyperplane) in a high-dimensional space. They perform well in cases with **non-linear data**, especially when using kernel functions.

Support Vector Machines (SVM) are supervised kernel methods effective for both regression (SVR) and classification. By projecting descriptor vectors into high-dimensional feature spaces through kernels (RBF, polynomial), SVMs capture non-linear relationships between molecular properties and solubility without massively increasing model complexity. This is valuable for poorly soluble drugs where descriptors interact nonlinearly (e.g., interplay of lipophilicity and hydrogen bonding). SVMs are especially useful when datasets are moderate in size: they regularize via margin maximization and often generalize better than flexible neural networks on limited data. Lovrić et al. (2021) included SVMs among competitive descriptor-based learners, showing robust performance across internal splits. Jamshidi et al. (2024) also highlight SVM utility in formulation-centred problems such as predicting solubility in supercritical CO_2 ($SCCO_2$), where process variables (temperature, pressure) and molecular descriptors produce complex, but learnable, non-linear surfaces. The drawbacks of SVM include sensitivity to kernel choice and hyperparameters, limited native probabilistic outputs, and reduced interpretability relative to linear models or tree ensembles. Practically, SVMs are valuable as part of model comparison suites and for problems where dataset size favors regularized kernel machines over

deep learners.
Refs: Lovrić et al. 2021; Jamshidi et al. 2024; Lipinski 2000

3. Random Forest (RF)

Random Forest is one of the most popular ensemble learning techniques. It creates **multiple decision trees** and averages their predictions, which makes it more robust and less prone to overfitting. RF models are particularly good at handling noisy data and can also rank **feature importance**, helping researchers understand which molecular properties influence solubility most.

Random Forest (RF) is an ensemble of decision trees that reduces variance through bootstrap aggregation and feature subsetting. For solubility prediction RF combines robustness with interpretability: it handles heterogeneous descriptor types, tolerates noisy or correlated features, and provides feature-importance metrics that map well to chemical intuition (e.g., logP, polar surface area). Lovrić et al. (2021) demonstrated RF achieving strong performance (competitive R^2/RMSE) using a compact descriptor set, which makes RF attractive for intrinsic aqueous solubility where descriptor engineering remains central. RF's tree structure inherently models non-linear interactions (e.g., thresholds in pKa effects) and can capture complex partitioning behavior relevant to poorly soluble molecules. RF is also resilient on moderate datasets often encountered in early drug discovery. Limitations include less sharp predictive edges compared with tuned gradient boosting on very large data, and limited extrapolation outside training chemical space. In formulation contexts such as $SCCO_2$, RF has been used to predict how process variables modulate apparent solubility (Jamshidi et al., 2024), indicating the method's flexibility for combining molecular and process descriptors. Overall, RF is a practical, interpretable workhorse for solubility QSPR and formulation models. **Refs:** Lovrić et al. 2021; Jamshidi et al. 2024; Lipinski 2000.

4. Artificial Neural Networks (ANN)

ANNs mimic the way neurons work in the human brain. They are capable of learning **non-linear patterns** and **interactions between multiple features**, which makes them ideal for solubility prediction involving many molecular descriptors. However, they require more data and tuning compared to simpler models.

Artificial Neural Networks (ANN) are universal function approximators that map input descriptors to solubility through layered, non-linear transformations. For solubility prediction ANNs can model complex interactions among descriptors (e.g., intramolecular hydrogen bonding, conformational flexibility) that influence dissolution and aqueous saturation. Jamshidi et al. (2024) emphasize ANN usage in formulation predictions—such as $SCCO_2$ solubility— where multifactorial inputs (molecular descriptors + pressure/temperature) create complex response surfaces. ANNs are flexible in accepting heterogeneous input types (numerical descriptors, categorical formulation factors) and can be regularized via dropout, early stopping, and weight penalties. When trained on sufficiently large, diverse datasets ANNs can outperform classical learners; however, they require careful hyperparameter tuning and are prone to overfitting with small datasets. Lovrić et al. (2021) noted that classical methods sometimes matched or exceeded ANN performance on curated, moderate datasets—highlighting that ANN advantage appears when data volume and diversity justify model complexity. ANNs also lack direct chemical interpretability, but techniques such as sensitivity analysis and input-attribution (e.g., integrated gradients) help recover mechanistic signals. Overall, ANNs are a flexible option, particularly in formulation and multi-input contexts. **Refs:** Jamshidi et al. 2024; Lovrić et al. 2021; Lipinski 2000.

5. Gradient Boosting Machines (GBM, XGBoost, LightGBM)

These are advanced ensemble models that build decision trees sequentially. Each new tree tries to correct the errors of the previous one. They often outperform RF and SVM in complex tasks and are commonly used in **Kaggle** competitions due to their high accuracy and speed. Gradient boosting algorithms (GBM, XGBoost, LightGBM) sequentially fit ensembles of shallow trees where each tree corrects predecessors' residuals. This boosting strategy is highly effective for complex, high-dimensional descriptor spaces typical of solubility datasets: it models subtle feature interactions and heterogeneous effects (e.g., pKa × lipophilicity) with high predictive accuracy. Lovrić et al. (2021) reported LightGBM among top performers, balancing accuracy and

computational efficiency. Boosting utilities such as regularization, shrinkage, and tree constraints make these models robust to overfitting even with many descriptors. They also support missing data handling and have facilities for feature importance and SHAP explanations, which help interpret influence of molecular properties on solubility predictions. However, boosting models are more computationally intensive to tune than RF and less inherently interpretable than linear models. In formulation-oriented modeling (e.g., $SCCO_2$ systems discussed by Jamshidi et al., 2024), boosting methods effectively learn multi-factor response surfaces combining molecular and processing descriptors. In sum, gradient boosting is often the go-to method when high predictive performance is required on descriptor-rich solubility problems, provided careful cross- validation and interpretability tools are used. **Refs:** Lovrić et al. 2021; Jamshidi et al. 2024; Lipinski 2000.

6. Graph Neural Networks (GNN)

Deep learning models, especially **Graph Neural Networks (GNNs)**, are a recent advancement in the field. Instead of using pre-defined descriptors, GNNs learn directly from the **molecular graph structure** (atoms as nodes, bonds as edges). They capture spatial and chemical interactions more effectively and have shown better generalization to new molecules.

Graph Neural Networks (GNNs) operate directly on molecular graphs (atoms = nodes, bonds = edges), learning representations through message passing that naturally encode local chemical environment and connectivity. This is highly advantageous for solubility prediction because solubility is influenced by local substructures (e.g., heterocycles, polar substituents) and their spatial/graph context. GNNs eliminate dependence on precomputed descriptors, reducing feature-engineering bias and enabling discovery of substructure patterns linked to poor dissolution. Recent literature (post-Lovrić) shows GNNs achieving state-of-the-art performance on various ADMET endpoints; while Lovrić et al. (2021) focused on descriptor-based learners, GNNs are increasingly adopted for intrinsic solubility modelling. Their strengths include transfer learning from large chemical graphs and integrating 3D/physicochemical node/edge features to reflect solvation energetics. Limitations are data requirements and computational cost; also, GNNs can be less transparent, although attention-based variants and subgraph attribution methods can recover chemically meaningful explanations. For poorly soluble drugs—where subtle graph motifs can dictate crystal packing or intramolecular hydrogen bonds—GNNs offer a promising route to improvedstructure-aware predictions. **Refs:** Lovrić et al. 2021; Lipinski 2000; Serajuddin 2007.

Summary Table of AI Techniques:

Technique	Strengths	Limitations	Best Use Case
Linear Regression	Simple, interpretable	Poor for non-linear data	Small datasets
SVM	Effective with non-linear boundaries	Sensitive to feature scaling	Solubility classification
Random Forest	Robust, handles noise, interpretable	Less effective on highly correlated data	Feature importance analysis
ANN	Captures complex relationships	Needs large data, black-box model	Solubility in non-aqueous systems
Gradient Boosting	High accuracy, fast properly	Overfitting if not tuned	Advanced predictive modeling
GNN / Deep Learning	Learns from molecular structure	Needs big data, less interpretable	Novel compound prediction

Transfer Learning	Learns from related tasks	Needs good source model	Small or rare datasets

Step 4: Train the Model

* Feed the collected dataset into the chosen AI model.
* Split data into training (to learn) and testing (to check performance).
* Adjust model parameters to improve accuracy.

Step 5: Predict Solubility

* The trained model predicts how soluble the drug will be.
* Example: Random Forest outputs a numerical solubility value, SVM gives soluble or insoluble, Neural Network gives high-accuracy predictions.

Step 6: Formulation Time

* Use the model's prediction to decide how to formulate the drug (e.g., nanoparticles, solid dispersions, or salt forms).
* This step bridges AI predictions with real-world drug formulation.

Step 7: Test Success

* Check if the new formulation works in lab/clinical tests.
* If it fails → go back, improve dataset or model, and try again.

Step 8: New Drug!

* Once successful, a new improved drug formulation is developed

Importance of Solubility Prediction for Drug Development

Drug solubility is a critical physicochemical property that significantly impacts the absorption, bioavailability, and therapeutic efficacy of pharmaceutical compounds. Poor solubility is one of the leading causes of drug failure during the development process, as highlighted by Lipinski (2000), who emphasized that many promising drug candidates fail due to inadequate solubility and permeability. Accurate solubility prediction allows pharmaceutical scientists to identify potential issues early in drug design, thereby reducing the risk of late-stage failures and optimizing resource allocation. This predictive approach is particularly crucial in the context of oral drug delivery, where aqueous solubility directly influences dissolution in the gastrointestinal tract and, consequently, systemic absorption (Lipinski, 2000).

Traditional experimental methods for determining solubility, such as shake-flask or potentiometric techniques, are labor-intensive, time-consuming, and often require significant amounts of drug substance. This limitation becomes particularly challenging in the early stages of drug discovery, where only minute quantities of compounds are available. As Serajuddin (2007) discussed, strategies such as salt formation, co-crystallization, or particle size reduction can improve solubility, but selecting the most suitable approach requires an understanding of the intrinsic solubility of the compound. Predictive models therefore enable rational decision-making and guide formulation strategies before extensive experimental work is undertaken.

In recent years, machine learning (ML) and artificial intelligence (AI) have emerged as powerful tools for solubility prediction. By analyzing large datasets of molecular structures and their corresponding solubility values, ML models can predict intrinsic aqueous solubility with high accuracy (Lovrić et al., 2021). These models not only accelerate the drug development timeline but also allow researchers to explore a wider chemical space by virtually screening numerous candidates. Moreover, environmentally-friendly solubility enhancement

techniques, such as supercritical CO_2 processing, have benefited from ML-guided optimization, demonstrating the integration of computational prediction with sustainable formulation strategies (Jamshidi et al., 2024).

Overall, solubility prediction is indispensable in modern drug development. It enables early identification of solubility-limited candidates, informs formulation strategies, reduces development costs, and shortens the time-to- market for new therapeutics. With the integration of machine learning and computational approaches, predictive solubility modeling not only enhances efficiency but also supports the design of innovative and sustainable drug delivery systems, ultimately improving patient outcomes.

1. Role of Solubility in Drug Development

1.1 Bioavailability

Bioavailability refers to the fraction of an administered drug dose that reaches systemic circulation. For oral drugs, aqueous solubility determines how well the drug dissolves in gastrointestinal fluids, which is the first step before absorption. Poorly soluble drugs often exhibit low bioavailability, limiting therapeutic potential (Lipinski, 2000).

1.2 Formulation Strategy

Accurate solubility prediction enables pharmaceutical scientists to design appropriate formulation strategies. For instance:

Salt formation can increase solubility and stability (Serajuddin, 2007).
Co-crystals can enhance solubility without altering pharmacological activity. Nanoparticles and solid dispersions can improve dissolution rate.
Early prediction allows selection of the most efficient strategy before costly experimental work begins.

2. Classification of Solubility Prediction Methods

Solubility prediction can be broadly classified into experimental and computational approaches.

2.1 Experimental Methods

Traditional methods involve direct measurement of solubility in various solvents, such as:

- Shake-flask method: Directly determines equilibrium solubility.
- Potentiometric titration: Useful for ionizable compounds.
- HPLC-based solubility analysis: Quantitative and precise for complex mixtures.

While accurate, these methods are time-consuming, resource-intensive, and impractical for early drug discovery when only small amounts of compounds are available (Serajuddin, 2007).

2.2 Computational Methods

Computational and predictive models, including machine learning (ML) and artificial intelligence (AI), have become increasingly important in modern drug development:

- Quantitative Structure–Property Relationship (QSPR) models predict solubility from molecular descriptors such as hydrophobicity, molecular weight, and hydrogen-bonding capacity.
- Machine Learning Models: Techniques like Random Forest, Support Vector Machines, and Neural Networks analyze large datasets of molecular structures to predict intrinsic aqueous solubility (Lovrić et al., 2021).
- AI-based Process Optimization: ML algorithms can also optimize environmentally-friendly solubility enhancement techniques, such as supercritical CO_2-based processing (Jamshidi et al., 2024).

3. Factors Affecting Solubility

Drug solubility is influenced by intrinsic molecular properties and environmental factors:

3.1 Molecular Properties

- Lipophilicity (LogP): High lipophilicity reduces aqueous solubility.
- Molecular Weight: Larger molecules often show poor solubility.
- Hydrogen Bonding: Increased hydrogen bonding generally increases solubility.
- Ionization (pKa): Ionizable drugs can show enhanced solubility at specific pH ranges (Lipinski, 2000).

3.2 Environmental Factors

- Temperature: Solubility often increases with temperature.
- Solvent Properties: Solubility depends on solvent polarity and pH.
- Co-solvents or Additives: Surfactants, cyclodextrins, and other excipients can improve solubility (Serajuddin, 2007).

4. Advantages of Solubility Prediction

- Cost Reduction: Identifying solubility issues early avoids expensive late-stage failures.
- Time Efficiency: Predictive models reduce the need for repetitive laboratory experiments.
- Rational Formulation Design: Enables targeted use of strategies like salt formation, particle size reduction, or nanotechnology (Serajuddin, 2007).
- Chemical Space Exploration: ML models allow virtual screening of thousands of compounds before synthesis (Lovrić et al., 2021).
- Sustainable Development: AI and ML-guided solubility enhancement techniques reduce the use of toxic organic solvents (Jamshidi et al., 2024).

Case Studies in ML-Driven Solubility Prediction

The developed solubility prediction model was applied to **Silymarin**, a BCS Class II drug with poor aqueous solubility, to evaluate its practical utility. Poor solubility is a major challenge in drug formulation, often limiting bioavailability and therapeutic effectiveness. Traditional enhancement techniques, including nanocarrier-based formulations, require extensive experimental work and resources [1]. By leveraging curated datasets such as **AqSolDB**
[8] and **ESOL** [9], the model was trained to learn the relationship between molecular structure and solubility, using advanced graph neural network architectures that capture both local atomic environments and global molecular interactions [4,6,7].

- In applying the model to Silymarin, predictions of aqueous solubility were generated from the molecular graph representation of the compound, integrating 3D conformer information and chemical descriptors such as molecular weight, logP, hydrogen bond donors/acceptors, and polar surface area [5,6,8]. The predicted solubility values closely aligned with reported experimental measurements and also reflected trends observed in nanocarrier-enhanced formulations [1,14]. Attention-based mechanisms within the graph neural network provided insight into which molecular substructures most influenced solubility, highlighting functional groups that contribute to poor or enhanced solubility [15].
- This case study demonstrates that AI-driven solubility prediction can serve as a **fast, cost-effective alternative** to traditional experimental methods, allowing researchers to prioritize compounds with favorable solubility profiles for further formulation development. While experimental validation remains essential, the model provides a preliminary guide for solubility optimization and can help reduce resource-intensive trial-and-error approaches [2,3,12]. These results indicate that machine learning and graph neural networks hold significant promise for improving efficiency in pharmaceutical research and early-stage drug development [2,3,4].

6. Future Perspective

- Solubility prediction is expected to advance with:
- Integration of AI with high-throughput experimentation for real-time solubility prediction.
- Multi-objective optimization, balancing solubility with permeability, stability, and bioavailability.
- Green chemistry approaches, using predictive modeling to design eco-friendly formulation processes.
- Ultimately, solubility prediction is no longer just a screening tool but a strategic component of drug design, formulation, and delivery optimization.
-

CONCLUSION:

Solubility is a cornerstone property in drug development, dictating absorption, bioavailability, and overall therapeutic efficacy. Despite decades of research, poor aqueous solubility continues to limit the clinical translation of promising drug candidates, with nearly 90% of molecules in development facing solubility-related barriers. Traditional approaches—such as salt formation, particle size reduction, surfactants, and solid dispersions—have improved formulation success, yet they remain resource-intensive, case-specific, and often unsuitable for rapid early-stage screening.

The rise of Artificial Intelligence (AI) and Machine Learning (ML) has introduced a paradigm shift by enabling predictive, data-driven modeling of solubility. Techniques ranging from simple linear regression and Random Forests to advanced deep learning architectures and Graph Neural Networks allow rapid screening of thousands of compounds, minimizing trial-and-error experiments. Moreover, AI supports green technologies such as supercritical carbon dioxide ($SCCO_2$), optimizing eco-friendly solubility enhancement strategies and aligning with sustainable pharmaceutical manufacturing goals.

Ultimately, combining conventional formulation methods with AI-based predictive modeling offers a more robust, efficient, and environmentally conscious pathway to overcome solubility challenges. As computational tools become more accurate and interpretable, their integration into Quality by Design (QbD) frameworks will further accelerate rational formulation, reduce costs, and improve patient outcomes in modern drug development.

References:

1. Lipinski, C. A. (2000). Drug-like properties and the causes of poor solubility and poor permeability. Journal of Pharmacological and Toxicological Methods, 44(1), 235–249.

2. Serajuddin, A. T. M. (2007). Salt formation to improve drug solubility. Advanced Drug Delivery Reviews, 59(7), 603–616.

3. Lovrić, M., Molero, J.M., Kern, R., et al. (2021). Machine Learning in Drug Development: Predicting Intrinsic Aqueous Solubility from Molecular Structure. Journal of Chemometrics, 35(4), e3349. https://doi.org/10.1002/cem.3349

4. Jamshidi, A., et al. (2024). Recent advancements toward the increment of drug solubility using environmentally- friendly supercritical CO_2: a machine learning perspective. Frontiers in Medicine, 11, 1467289. https://doi.org/10.3389/fmed.2024.1467289

5. Avdeef, A. (2003). Solubility of drugs and related strategies. Pharmaceutical Research, 20(5), 763–771.

6. Ahmed, M., et al. (2020). Machine learning approaches in predicting drug solubility: A review. Artificial Intelligence in Medicine, 104, 101820.

7. K. T. Savjani, A. K. Gajjar, and J. K. Savjani, “Drug Solubility: Importance and Enhancement Techniques,” ISRN Pharmaceutics, 2012. [Online]. Available: https://doi.org/10.5402/2012/195727

8. P. Llompart et al., “Will we ever be able to accurately predict solubility?,” Scientific Data, vol. 11, article 303, 2024. [Online]. Available: https://www.nature.com/articles/s41597-024-02121-w

9. B. Haque and E. A. Siddiqui, "Considering Artificial Intelligence and Machine Learning in Pharmaceutical Industries Research and Development," IEEE ICDCC, 2024. [Online]. Available: https://doi.org/10.1109/ICDCC.2024.11960792

10. Z. Wu et al., "A Comprehensive Survey on Graph Neural Networks," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 1, pp. 4–24, 2021. [Online]. Available: https://doi.org/10.1109/TNNLS.2020.2978386

11. M. C. Sorkun, J. M. V. A. Koelman, and S. Er, "Pushing the limits of solubility prediction via quality-oriented data selection," iScience, vol. 24, no. 1, p. 101961, 2021. [Online]. Available: https://doi.org/10.1016/j.isci.2020.101961

12. P. Gao et al., "Accurate predictions of aqueous solubility of drug molecules via the multilevel graph convolutional network (MGCN) and SchNet architectures," Phys. Chem. Chem. Phys., vol. 22, pp. 23766–23772, 2020. [Online]. Available: https://doi.org/10.1039/D0CP03596C

13. . K. Choudhary and B. DeCost, "Atomistic Line Graph Neural Network for improved materials property predictions," npj Computational Materials, vol. 7, p. 185, 2021. [Online]. Available: https://doi.org/10.1038/s41524-021-00650-1

14. M. C. Sorkun, A. Khetan, and S. Er, "AqSolDB: a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds," Scientific Data, vol. 6, p. 143, 2019. [Online]. Available: https://doi.org/10.1038/s41597- 019-0151-1 9.

15. J. S. Delaney, "ESOL: Estimating Aqueous Solubility Directly from Molecular Structure," J. Chem. Inf. Comput. Sci., vol. 44, no. 3, pp. 1000–1005, 2004. [Online]. Available: https://doi.org/10.1021/ci034243x

16. N. M. O'Boyle et al., "Open Babel: An open chemical toolbox," J. Cheminform., vol. 3, p. 33, 2011. [Online]. Available: https://doi.org/10.1186/1758-2946-3-33

17. D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

18. P. G. Francoeur and D. R. Koes, "SolTranNet – A Machine Learning Tool for Fast Aqueous Solubility Prediction," J. Chem. Inf. Model., vol. 61, no. 6, pp. 2530–2536, 2021. [Online]. Available: https://doi.org/10.1021/acs.jcim.1c00331

19. Ł. Maziarka et al., "Molecule Attention Transformer," arXiv preprint, 2020. [Online]. Available: https://arxiv.org/abs/2002.08264 14. S. Lee et al., "Multi-order graph attention network for water solubility prediction and interpretation," Scientific Reports, vol. 13, p. 957, 2023. [Online]. Available: https://doi.org/10.1038/s41598-023- 28117-5

20. K. T. Schütt et al., "SchNet – A deep learning architecture for molecules and materials," J. Chem. Phys., vol. 148, p. 241722, 2018. [Online]. Available: https://doi.org/10.1063/1.5019779