



**INTERNATIONAL JOURNAL OF
PHARMACEUTICAL SCIENCES**
[ISSN: 0975-4725; CODEN(USA): IJPS00]
Journal Homepage: <https://www.ijpsjournal.com>



Review Article

Artificial Intelligence and Machine Learning in Solubility Enhancement of Poorly Water-Soluble Drugs: From Solubility Prediction to Formulation Optimization

Dr. K. Narendra Kumar Reddy, K. Ravi*, Kodali. Ragini, Nannam. Ramabrahmam

Department of Pharmaceutics, Vikas Group of Institutions, Nunna, Vijayawada, Andhra Pradesh- 521212, India.

ARTICLE INFO

Published: 19 Oct 2026

Keywords:

Artificial intelligence, Machine learning, Poorly water-soluble drugs, Solubility prediction, Formulation optimization, Drug delivery

DOI:

10.5281/zenodo.23466730

ABSTRACT

Poor aqueous solubility limits the development of many orally administered drug candidates, particularly those in Biopharmaceutics Classification System (BCS) class II and, in some cases, class IV. Conventional solubility-enhancement approaches, including particle-size reduction, solid dispersions, amorphous systems, lipid-based formulations, cyclodextrin complexation, nanocrystals, salt formation and pharmaceutical cocrystals, are effective but often require extensive, formulation-specific experimentation. Growing pharmaceutical datasets and advances in artificial intelligence (AI) and machine learning (ML) now offer predictive, data-driven support. Models have been reported for predicting aqueous and formulation-phase solubility, screening drug-excipient combinations, estimating drug-polymer interaction, optimizing compositions and predicting dissolution or supersaturation behaviour. Bayesian optimization, active learning and surrogate modelling are being explored to reduce the number of formulation experiments, and integration with physiologically based pharmacokinetic (PBPK) modelling, Quality by Design (QbD) and automated experimentation may further support rational development. However, small and heterogeneous datasets, experimental variability, limited interpretability and external validation, data leakage and uncertain regulatory acceptance remain significant barriers. This review discusses AI and ML in the solubility enhancement of poorly water-soluble drugs, progressing from molecular solubility prediction to drug-excipient selection, advanced formulation design and formulation optimization. Particular emphasis is placed on experimentally validated applications and on the distinction between computational prediction and demonstrated pharmaceutical performance.

***Corresponding Author:** K. Ravi

Address: Department of Pharmaceutics, Vikas Group of Institutions, Nunna, Vijayawada, Andhra Pradesh- 521212, India

Email ✉: ravikorrapati11@gmail.com

Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.



Current challenges, research gaps and opportunities for reliable, explainable and experimentally integrated AI-driven formulation platforms are outlined.

INTRODUCTION

Poor aqueous solubility is one of the major challenges encountered during the development of orally administered drug products. A considerable number of drug candidates exhibit limited water solubility, which can result in slow or incomplete dissolution in gastrointestinal fluids and consequently restrict the amount of drug available for absorption. This problem is particularly important for BCS class II drugs, which exhibit low aqueous solubility despite relatively high permeability.¹ Therefore, improving the solubility and dissolution behaviour of poorly water-soluble drugs remains an important objective in pharmaceutical formulation development.^{2,3}

Drug solubility is influenced by several physicochemical characteristics, including molecular size, hydrophobicity, ionization, hydrogen-bonding capacity, melting point and crystal lattice energy.⁴ The solid-state form of the drug is also important: crystalline material generally possesses greater lattice stability, and therefore lower apparent solubility, than its amorphous counterpart, and poor solubility may arise either from unfavourable interaction with water, as in highly lipophilic molecules, or from strong interactions within the drug crystal. These factors make solubility enhancement a complex formulation problem that cannot be addressed effectively using a single universal approach.

Several conventional strategies have therefore been developed to improve the aqueous solubility and dissolution of poorly soluble drugs. These include particle-size reduction, micronization, nanocrystals, salt formation, pharmaceutical cocrystals, cyclodextrin complexation, solid dispersions, amorphous solid dispersions (ASDs), surfactant-based systems and lipid-based formulations.^{2,3,5-8} Although these approaches can substantially improve drug performance, formulation development commonly involves extensive experimental screening. Selection of suitable excipients, drug-to-excipient ratios and processing conditions may require numerous experiments because formulation

variables frequently interact in nonlinear and drug-specific ways.

The increasing application of artificial intelligence (AI) and machine learning (ML) in pharmaceutical research provides an opportunity to complement conventional formulation-development approaches. AI refers broadly to computational technologies capable of performing tasks associated with intelligent decision-making, whereas ML enables computational models to learn relationships from experimental or molecular data. In pharmaceutical science, these approaches can be used to predict physicochemical properties, identify important formulation variables, screen drug-excipient combinations and optimize formulation compositions.^{9,10}

One of the most established applications is the prediction of aqueous drug solubility. Early computational approaches, such as the ESOL model, demonstrated that molecular descriptors could be used to estimate aqueous solubility.¹¹ The subsequent development of large, curated datasets, including AqSolDB, has provided diverse molecular features for model training.¹² Recent studies have expanded this concept from simple aqueous-solubility prediction toward formulation-specific applications, including prediction of drug solubility in polymeric and lipidic environments.^{13,14}

AI/ML approaches are also increasingly being investigated for drug-excipient compatibility and formulation selection. Predictive models can potentially identify promising drug-excipient combinations before extensive experimental screening, thereby reducing the formulation search space.¹⁵ Similarly, machine learning has been applied to amorphous solid dispersion development, lipid-based formulations, dissolution modelling and formulation optimization.^{9,10,13,14} These applications are particularly relevant because solubility enhancement is rarely determined by a single variable; rather, drug properties, excipient characteristics, formulation composition and manufacturing conditions may collectively determine the final performance of a dosage form.

Despite these opportunities, the pharmaceutical application of AI is associated with important challenges. Pharmaceutical datasets are often relatively small and heterogeneous, while solubility measurements can vary



according to experimental conditions such as pH, temperature, medium composition and analytical methodology. Inadequate external validation, overfitting, data leakage and limited model interpretability may further restrict the generalizability of AI models.¹⁶ Therefore, computational predictions should be considered complementary to, rather than replacements for, experimental pharmaceutical evaluation.

The objective of this review is to examine critically the role of AI and ML in the solubility enhancement of poorly water-soluble drugs. The review progresses from fundamental solubility prediction to drug-excipient interaction modelling, advanced formulation systems, dissolution and supersaturation prediction, PBPK integration and formulation optimization. Particular emphasis is placed on experimentally supported applications and on the integration of AI with QbD and experimental formulation development.

The central premise of this review is that AI/ML may enable a transition from predominantly empirical solubility-enhancement development toward a more predictive and data-driven workflow. However, its practical value depends on the quality of pharmaceutical data, appropriate model validation, mechanistic interpretation, and experimental confirmation of predicted formulations.

2. LITERATURE SEARCH APPROACH

This is a narrative review and not a systematic review. Sources were identified through targeted searches of PubMed-indexed and publisher-hosted literature and of

regulatory agency publications, using combinations of the terms artificial intelligence, machine learning, aqueous solubility prediction, poorly water-soluble drugs, amorphous solid dispersion, lipid-based formulation, drug-excipient compatibility, dissolution, PBPK, Quality by Design and formulation optimization. Foundational papers on solubility enhancement and on solubility datasets were combined with recent studies and reviews on ML in formulation science. The reference set is selective, no performance figures have been pooled across studies, and numerical results are quoted only where they were reported by the original authors for their own data.

3. CONVENTIONAL SOLUBILITY-ENHANCEMENT STRATEGIES AND THEIR DATA REQUIREMENTS

Solubility-enhancement strategies can be grouped according to whether they modify the drug solid, the drug environment or the delivery system. Physical modifications, such as particle-size reduction, nanocrystals and amorphization, increase the surface area or the free energy of the solid. Chemical modifications, such as salt formation and cocrystallization, alter the solid-state structure or ionization behaviour of the molecule. Carrier-based approaches, such as solid dispersions, cyclodextrin complexes and lipid-based systems, change the microenvironment in which the drug dissolves or is presented to the gastrointestinal tract.^{2,3} Table 1 summarizes representative strategies and the variables that a data-driven model would need to capture.

Table 1: Representative Solubility-Enhancement Strategies and Variables Relevant to ml

Strategy	Principal mechanism	Typical limitations	Variables a model must capture
Particle-size reduction / nanocrystals ³	Larger surface area; faster dissolution rate	Does not raise equilibrium solubility; aggregation and physical instability	Particle size, stabilizer type and level, milling or precipitation conditions
Salt formation ⁵	Ionization; altered solid-state properties	Requires ionizable groups; possible conversion back to free form	pKa, counter-ion, pH of medium, hygroscopicity
Cyclodextrin complexation ⁶	Inclusion of lipophilic moiety in a hydrophilic host	Limited drug loading; large excipient mass	Cyclodextrin type, molecular fit, stoichiometry, stability constant
Amorphous solid dispersions ⁸	Higher free energy; polymer-stabilized supersaturation	Recrystallization on storage or dissolution; hygroscopicity	Polymer, drug loading, miscibility, glass-transition behaviour, process route
Lipid-based formulations ⁷	Pre-dissolved drug; dispersion and digestion in the gut	Drug precipitation on dispersion; excipient variability	Drug solubility in lipids and surfactants, composition ratios, dispersion behaviour



Strategy	Principal mechanism	Typical limitations	Variables a model must capture
Pharmaceutical cocrystals ³	New crystalline phase with altered lattice energy	Coformer screening is empirical; possible dissociation	Coformer choice, hydrogen-bonding motifs, solvent and process

Compiled by the author from the cited references.

Two observations follow from this overview. First, each strategy is governed by a different set of interacting variables, so a single model is unlikely to serve all of them; a solubility predictor for ASD polymer screening has little in common with a model for cyclodextrin binding. Second, because the experimental search space is large and the cost per experiment (material, time, analytical effort) is substantial, even modest reductions in the number of experiments needed to reach a robust formulation are attractive. This is the practical rationale for data-driven methods, and also the reason they must be judged by how much experimental work they save without compromising performance.

4. MACHINE LEARNING FOUNDATIONS FOR PHARMACEUTICAL FORMULATION DATA

4.1 Data and molecular representation

The quality of any ML model is limited by the data from which it learns. Pharmaceutical formulation data may be of several types: molecular structures and computed descriptors for the drug and excipients, measured physicochemical properties (solubility, melting point, logP, pKa), formulation composition and process parameters, and performance outcomes

such as dissolution profile, physical stability and bioavailability. Molecules are commonly encoded as numerical descriptors (for example molecular weight, lipophilicity, counts of hydrogen-bond donors and acceptors, topological indices), as binary substructure fingerprints, or as learned representations derived from molecular graphs.^{9,17} The choice of representation influences what the model can learn and how well it can be interpreted.

4.2 Algorithms commonly used

Regression approaches, including multiple linear regression and penalized regression, are transparent and often serve as baselines. Tree-based ensembles (random forest, gradient boosting) handle nonlinear relationships and mixed variable types and are frequently used on tabular formulation data. Support vector machines and Gaussian process models are popular for small datasets, the latter having the added advantage of providing an uncertainty estimate that can guide the choice of the next experiment. Neural networks, including deep and graph-based architectures, can model complex structure-property relationships but typically need larger datasets and careful regularization.^{9,10,18} Table 2 summarizes the main approaches and the situations in which each is most useful.

Table 2: Commonly Used ml Approaches in Solubility and Formulation Studies

Approach	Typical pharmaceutical use	Strengths	Cautions
Linear and penalized regression	Baseline solubility models from a few descriptors	Transparent; stable in small datasets	Cannot capture strong non-linearity
Random forest / gradient boosting	Solubility, stability and compatibility classification	Handles nonlinear effects and mixed inputs	Can overfit if tuned repeatedly; limited extrapolation
Support vector machines	Classification of formulation outcomes	Effective in high-dimensional, small datasets	Sensitive to scaling and kernel choice
Gaussian process models	Data-efficient modelling; active learning	Gives uncertainty estimates	Computationally heavier for large data
Neural networks / graph networks	Molecular fingerprints to solubility; complex endpoints	Flexible; learns representations	Data hungry; opaque; may not generalize outside training space

Compiled by the author from general ML methodology and the cited reviews.

4.3 Working with small and heterogeneous datasets

A defining feature of formulation science is that datasets are typically small relative to the number of



variables, and that data produced in different laboratories are not strictly comparable because of differences in methods, media and analysts. Several strategies are used to manage this. Simple models with few parameters are preferred; domain knowledge is encoded through physically meaningful descriptors; related datasets may be combined, provided differences in measurement conditions are recorded and accounted for; and active learning, in which the model selects the most informative next experiment, reduces the number of experiments required to define a response.^{9,10} None of these strategies removes the need for well-documented, reproducible experimental data, which remains the most valuable resource.

4.4 Model validation

Reported performance depends heavily on how data are split. Random splits of molecules or formulations that are chemically near-identical give optimistic estimates, whereas splits by scaffold, by drug, or by time give a more realistic picture of how the model will behave on a genuinely new compound. Data leakage, in which information from the test set influences model development through preprocessing, feature selection or repeated tuning, is a recognized source of inflated performance across ML-based science.¹⁶ The applicability domain, meaning the region of chemical and formulation space in which the model has been trained and can be trusted, should be defined, and predictions outside it should be flagged.¹⁹

5. PREDICTION OF AQUEOUS SOLUBILITY

5.1 From simple descriptors to curated datasets

Computational estimation of aqueous solubility has a long history. The ESOL approach showed that aqueous solubility could be estimated directly from molecular structure using a small number of calculated descriptors in a simple regression model, providing a transparent baseline that remains widely used for comparison.¹¹ Later work benefited from the assembly of larger and more carefully curated datasets. AqSolDB brought together solubility values from multiple sources into a single curated reference set with calculated two-dimensional descriptors for a diverse range of compounds, enabling more systematic benchmarking of models.¹² Other work emphasized dataset

integrity, reporting predictive models of aqueous solubility built on a large dataset described as being of high integrity.²⁰

5.2 Machine learning, physicochemical knowledge and generalization

More recent work has examined how ML can be combined with physicochemical relationships to predict solubility in organic solvents and in water, an approach that is relevant to formulation because many processes involve non-aqueous media.²¹ A further study has specifically examined the generalization, complexity and predictive ability of ML models of intrinsic aqueous solubility, which underlines the importance of testing models beyond the data on which they were built.¹⁸ Reviews of computational solubility prediction in water-based systems describe qualitative and quantitative approaches and discuss their roles in discovery and development settings.¹⁹

5.3 Data quality is the limiting factor

Experimental solubility is not a single fixed number. It depends on solid form, pH, temperature, ionic composition, equilibration time and analytical method, and values reported for the same compound may differ between laboratories. When such values are pooled without harmonization, part of the model error reflects the noise in the data and not the model itself, which places an upper limit on the accuracy that can be demonstrated.^{12,20} Intrinsic solubility (of the neutral species) should be distinguished from apparent solubility at a given pH, and thermodynamic from kinetic solubility measurements; mixing them in a training set is a common source of avoidable error.



5.4 What solubility prediction does and does not deliver

Even an accurate predictor of intrinsic aqueous solubility does not by itself yield a formulation. It helps to prioritize candidates, flag likely BCS class II or IV behaviour early and select which enhancement strategy to attempt first, but the solubility that matters in a dosage form is determined by the formulation environment (polymer, lipid, surfactant, pH, and the dynamic conditions of the gut), not by the molecule alone. This is why the field has moved toward formulation-phase predictions, discussed next.

6. SOLUBILITY AND MISCIBILITY IN FORMULATION-RELEVANT MEDIA

6.1 Lipid-based formulations

Lipid-based formulations keep the drug in a dissolved or solubilized state in the gastrointestinal tract, and selection of lipid and surfactant excipients depends strongly on the drug's solubility in them.⁷ Screening this solubility experimentally for many excipients is laborious. The Sol_ME tool addressed this by using an artificial neural network with molecular fingerprints trained on 1,379 drug-solvent entries to predict drug solubility in lipidic excipients.¹³ The authors reported a correlation coefficient of 0.998 for the model and, in a case study with apalutamide, identified cinnamon oil as a suitable carrier, reporting a 75% reduction in formulation volume and about 80% agreement between predicted and experimentally measured solubilities across 35 drug-solvent combinations.¹³

This study is instructive for the central theme of this review. The very high correlation on the modelling data describes how well the model fits compounds drawn

from the same collection, while the 80% agreement on experimental testing is a more modest but more relevant measure of how the model performs in the laboratory. The gap between the two illustrates why computational metrics and experimentally demonstrated performance should be reported separately. A subsequent study described the formulation, pharmacokinetic and pharmacodynamic evaluation of a unit-dose lipid formulation for apalutamide developed with the assistance of Sol_ME,²² extending the evidence from solubility prediction towards in vivo performance, although the strength of that evidence should be judged from the original report.

6.2 Polymers and amorphous solid dispersions

ASDs depend on selecting a polymer that is miscible with the drug and that maintains the amorphous state and supersaturation during storage and dissolution.⁸ Because many drug-polymer pairs and ratios are possible, experimental screening is costly. One study combined high-throughput screening with ML to predict the formation of both binary and ternary ASD formulations for early discovery and development, a design that addresses the scarcity of consistent data by generating it systematically.¹⁴ Combining automated experimentation with ML in this way is attractive because the same experimental platform provides both the training data and the means of validating predictions, reducing the heterogeneity that limits models trained on literature data.

Table 3 lists the studies discussed so far and what each contributes. The table separates the nature of the evidence, because a model that reports prediction accuracy on a dataset and a study that experimentally confirms formulations answer different questions.

Table 3: Selected AI/ml Studies Discussed and the Nature of the Evidence

Study	Task	Approach and data (as reported)	Nature of evidence
ESOL ¹¹	Aqueous solubility from structure	Regression on calculated molecular descriptors	Computational benchmark; transparent baseline
AqSolDB ¹²	Curated aqueous solubility reference set	Multi-source compilation with 2D descriptors	Data resource enabling benchmarking
Sun et al. ²⁰	Aqueous solubility models	Models built on a large dataset of controlled integrity	Computational; emphasizes data quality
Boobier et al. ²¹	Solubility in organic solvents and water	ML combined with physicochemical relationships	Computational; broader solvent scope



Study	Task	Approach and data (as reported)	Nature of evidence
Sol_ME ¹³	Drug solubility in lipid excipients	Neural network, fingerprints, 1,379 entries; apalutamide case	Computational plus experimental confirmation (35 combinations)
Lu et al. ¹⁴	Binary and ternary ASD formation	High-throughput screening plus ML	Computational and experimental screening data
Nguyen et al. ¹⁵	Drug-excipient compatibility	ML prediction of compatibility	Computational; assess against experimental stress data

Compiled by the author from the cited sources; details are limited to what the sources report.

7. DRUG-EXCIPIENT COMPATIBILITY AND EXCIPIENT SELECTION

Incompatibility between drug and excipient, through chemical interaction, moisture-mediated degradation or physical change, can compromise stability and must be excluded early in development. Conventional compatibility studies use thermal and spectroscopic methods and stressed storage, which are time-consuming when many excipients are considered. ML models trained on documented compatibility outcomes offer a way to rank excipients before experiments, and have been explored with the aim of supporting safer and more efficient formulations.¹⁵

Several points deserve emphasis. Compatibility data are scarce, imbalanced (compatible pairs usually outnumber incompatible ones) and reported inconsistently, and an incompatibility may depend on moisture, temperature, impurities and processing, none of which is captured by structure alone. Models should therefore be treated as screening tools that prioritize which pairs to test, with experimental confirmation by the usual analytical methods, and false negatives (incompatible pairs predicted as compatible) are more costly than false positives. Sensitivity for the incompatible class and the risk-based use of predictions should be reported alongside overall accuracy.

A related application is excipient selection for solubility enhancement more generally, in which the model proposes polymers, surfactants, lipids or cyclodextrins most likely to support a given drug. Here, as in compatibility, the practical benefit is a smaller, better-chosen experimental set; the value of a model is measured by how often its top-ranked suggestions are experimentally confirmed, compared with a conventional screening strategy.^{9,10}

8. DISSOLUTION, SUPERSATURATION AND STABILITY PREDICTION

For many solubility-enhanced systems, the key performance attribute is not equilibrium solubility but the extent and duration of supersaturation during dissolution. Amorphous systems and some lipid and cocrystal formulations generate supersaturated solutions that may precipitate before absorption, so the formulation must maintain supersaturation long enough for absorption, often with the help of precipitation-inhibiting polymers.^{3,8} Modelling this behaviour requires time-resolved data rather than a single solubility value, which makes datasets harder to obtain and compare.

ML approaches to dissolution use formulation composition, drug properties, process parameters and test conditions to predict dissolution profiles, and related approaches aim to predict physical stability, such as the tendency of an amorphous system to crystallize on storage. Reviews of ML in formulation science identify dissolution and stability prediction as active but less mature areas than solubility prediction, in part because of the limited availability of standardized datasets.^{9,10} Mechanistic models of dissolution and precipitation can be combined with ML so that the data-driven component learns residual effects not captured by the mechanism, an approach sometimes described as hybrid or physics-informed modelling.

For stability, the relevant prediction horizon is long (months to years) while available experimental data come from accelerated conditions. Extrapolation from accelerated to real-time storage is itself uncertain, and a model trained on short-term data should not be assumed to predict shelf life without supporting real-time results.



9. FORMULATION OPTIMIZATION: FROM EXPERIMENTAL DESIGN TO ACTIVE LEARNING

9.1 Quality by Design and design of experiments

Quality by Design is a systematic approach in which product and process understanding, based on sound science and risk management, is built into development, with critical quality attributes linked to critical material attributes and process parameters.²³ In practice, QbD commonly uses design of experiments (DoE) and response-surface modelling to map the relationship between formulation variables and responses such as dissolution, particle size or stability. These are, in a broad sense, early forms of data-driven modelling, and ML can be seen as an extension that handles larger variable sets and more complex nonlinear responses.

9.2 Surrogate modelling, Bayesian optimization and active learning

In surrogate modelling, a computationally cheap statistical model approximates the relationship between formulation inputs and outputs, and is used to explore the design space and select experiments. Bayesian optimization builds on this by using the surrogate's predictions and uncertainty to choose the next experiment that is most likely to improve the target response, balancing exploration of uncertain regions with exploitation of promising ones. Active learning applies the same logic to model building, selecting the experiments that are most informative for improving the model. The common aim is to reach a satisfactory formulation or a reliable model with fewer experiments than a conventional factorial or screening design; reviews of ML in formulation science discuss this aim in more detail.^{9,10}

These methods work best when the response can be measured quickly and reproducibly, the number of variables is moderate, and constraints (for example, acceptable excipient levels) are encoded. They can be coupled with automated formulation and testing platforms to form closed-loop systems, in which an algorithm proposes formulations, a robot prepares and tests them and results are fed back. Closed-loop systems are still developing, and their value for solubility enhancement will depend on the availability of reliable

automated assays for solubility, dissolution and stability.⁹

9.3 Multi-objective problems

Real formulations must satisfy several objectives at once: adequate solubility and dissolution, physical and chemical stability, manufacturability, acceptable excipient load and cost. Multi-objective optimization identifies the set of trade-off solutions (the Pareto front) so that the formulator can choose among them. Constraints derived from regulatory acceptability of excipients and from process capability should be part of the optimization problem, and not applied afterwards.

10. INTEGRATION WITH PBPK MODELLING AND IN VIVO PERFORMANCE

Improved in vitro solubility or dissolution does not automatically translate into improved bioavailability. Absorption depends on the interplay of dissolution, supersaturation, precipitation, gastrointestinal transit, permeability and first-pass metabolism. PBPK models represent these processes mechanistically and can estimate the effect of formulation on plasma exposure, and so are a natural partner for ML, which can supply inputs (such as predicted solubility or dissolution parameters) that are difficult to measure.^{9,10}

A recent study described integrating AI with PBPK modelling to predict the in vitro and in vivo fate of amorphous solid dispersions, linking formulation-level predictions to systemic exposure.²⁴ The attraction of such pipelines is that they connect a formulation decision to a clinically relevant outcome; the risk is that errors compound, since uncertainty in the ML-predicted input propagates through the mechanistic model. Credibility therefore requires verification against measured pharmacokinetic data, uncertainty quantification and, where possible, testing in a compound that was not used to build the model.

The overall chain from molecule to patient can be summarized as prediction of solubility, selection of excipients, optimization of composition, prediction of dissolution and supersaturation, PBPK simulation and, finally, experimental confirmation including in vivo studies where justified. Evidence of value is strongest at the ends of this chain that have been experimentally



tested, and weakest where predictions have only been compared with other predictions.

10.1 An integrated, experiment-anchored workflow

Bringing the preceding sections together, an AI-supported solubility-enhancement programme can be

organized as a sequence in which each computational step is followed by an experimental check that determines whether the programme proceeds. Table 4 outlines such a workflow. It is the author's synthesis of the approaches discussed in this review and is not a validated protocol.

Table 4: Outline of an Experiment-Anchored, Ai-Supported Workflow for a Poorly Soluble Drug

Stage	Decision	Role of AI/ML	Experimental anchor
1. Characterize	Is solubility or dissolution limiting absorption?	Predict aqueous solubility and BCS tendency from structure ^{11,12}	Measure equilibrium solubility, pKa, melting point and solid form
2. Choose strategy	Which enhancement approach is most promising?	Rank strategies using drug properties; flag candidates for ASD, lipid or salt routes ^{9,10}	Small feasibility experiments for the top two or three strategies
3. Screen excipients	Which polymers, lipids or surfactants?	Predict formulation-phase solubility or miscibility ^{13,14} ; screen compatibility ¹⁵	Confirm top-ranked excipients; test compatibility by standard methods
4. Optimize	Which composition and process settings?	Surrogate modelling, Bayesian optimization or active learning within QbD limits ²³	Sequential experiments chosen by the algorithm; verify against DoE where feasible
5. Predict performance	Will dissolution and supersaturation be adequate?	Dissolution and stability models; hybrid mechanistic-ML models	Dissolution testing in biorelevant media; accelerated and real-time stability
6. Translate	Is in vivo benefit expected?	PBPK simulation with ML-derived inputs ²⁴	Pharmacokinetic study where justified; compare with model prediction
7. Document	Is the evidence credible and auditable?	Record data provenance, model version and limits of use ^{25,26}	Change control and periodic re-verification

Prepared by the author as a synthesis of the approaches discussed in this review.

Two features of this workflow deserve emphasis. First, the experimental anchors define the stopping rules: a model whose top-ranked suggestions are not confirmed at stage 3 should be re-examined or abandoned before effort is invested in later stages. Second, the benefit of the workflow can be measured, for example as the number of experiments or the quantity of drug substance needed to reach a target performance compared with a conventional approach, and such measurement should accompany any claim that AI has accelerated development.

For a student or early-career researcher, a practical starting point is stage 3, applied to a drug already

available in the laboratory: measure its solubility in a small panel of candidate excipients, build a simple baseline model, and test whether the model's ranking of a few additional excipients is confirmed by measurement. Such a small project teaches data curation, validation and honest reporting, and yields the kind of experimentally confirmed result that the literature still lacks.

11. VALIDATION, REPRODUCIBILITY AND DATA QUALITY

Table 5 lists the main weaknesses that affect AI studies in pharmaceutical formulation and ways to address them, and Table 6 offers a short checklist for authors, reviewers and students appraising such studies.

Table 5: Common Pitfalls in AI/ml Formulation Studies and Mitigation

Pitfall	How it arises	Mitigation
Small dataset / overfitting	Many descriptors, few formulations; repeated model tuning	Simple models; regularization; nested cross-validation; held-out test set



Pitfall	How it arises	Mitigation
Data leakage	Preprocessing or feature selection on all data; near-duplicate compounds in train and test sets ¹⁶	Fit all preprocessing within training folds; split by scaffold, drug or time
Inconsistent experimental data	Different pH, temperature, solid form and methods across sources	Record metadata; harmonize or model conditions explicitly; prefer internally consistent data
Publication bias	Failed formulations rarely reported	Include negative results; generate balanced data with screening platforms
Limited applicability domain	Model used outside its chemical or formulation space ¹⁹	Define the domain; flag extrapolation; report uncertainty
Computation without confirmation	Accuracy on dataset presented as formulation success	Experimental validation of top predictions; report prospective hit rate
Opaque models	Black-box predictions without rationale	Report interpretable baselines; use explanation tools with caution

Compiled by the author from the cited literature.

Table 6: Appraisal Checklist for AI/ml Formulation Studies

Question	Why it matters
Is the training data source, size and measurement protocol described?	Data quality limits attainable accuracy ^{12,20}
Was the test set truly independent (by compound, scaffold or time)?	Guards against leakage and optimistic estimates ¹⁶
Is a simple baseline model reported for comparison?	Shows whether complexity adds value
Is the applicability domain stated, and are uncertainties given?	Prevents unjustified extrapolation ¹⁹
Were the best predictions confirmed experimentally, and what was the hit rate?	Distinguishes prediction from demonstrated performance ¹³
Are code, data and processing steps available?	Enables reproducibility and independent testing
Is there a link to in vivo or clinically relevant performance?	Solubility is a means, not an end ²⁴

Compiled by the author from the cited literature.

Data sharing and standardization deserve emphasis. Formulation science would benefit from shared, well-annotated datasets that include negative results and complete metadata (solid form, pH, temperature, method, equilibration time), as has happened with curated solubility collections.¹² Without such resources each group works with small, idiosyncratic datasets and models cannot be compared fairly.

12. REGULATORY AND QUALITY PERSPECTIVES

As AI enters development, regulators have begun to set expectations. The US Food and Drug Administration has issued draft guidance on considerations for the use of AI to support regulatory decision-making for drug and biological products, which adopts a risk-based approach in which the credibility of a model is assessed for its specific context of use.²⁵ The European Medicines Agency has published a reflection paper on the use of AI across the medicinal product lifecycle,

addressing matters such as data quality, model performance and the need for human oversight.²⁶ These documents are broader than formulation, and neither is specific to solubility enhancement, but their principles are directly applicable: the model's intended use should be defined, its performance and limits documented, and its impact on product quality assessed in proportion to risk.

Within the QbD framework, a model that informs development decisions but whose outputs are confirmed by experiments carries lower regulatory risk than one that is used to justify a control strategy or to replace testing.²³ Documentation of data provenance, model version, validation and change control is part of good practice. Developers should also be aware that regulatory guidance on AI is evolving, and that current national and regional documents, including those of Indian authorities, should be consulted before AI-derived evidence is included in a submission.



13. CHALLENGES AND RESEARCH GAPS

The main challenges can be summarized as follows. Data are limited, heterogeneous and often proprietary, and negative results are under-reported. Many studies stop at computational validation, so the number of experimentally confirmed, prospectively tested formulations remains small compared with the number of models published. Interpretability is limited for complex models, which weakens trust and makes it hard to extract mechanistic insight. External validation on compounds structurally distinct from the training set is rare. Formulation-phase predictions (solubility in polymers, lipids and mixed systems) lag behind aqueous solubility prediction in data availability and benchmark standardization. Linking predictions to in vivo performance is complicated by the propagation of uncertainty through PBPK models. Finally, uncertainty about regulatory acceptance, and limited access to computational expertise and automation in many academic and smaller industrial settings, slow translation.^{9,10,16}

Several research gaps follow. Shared benchmark datasets with complete metadata and negative results are needed. Prospective studies should report how often top-ranked predictions are confirmed and how many experiments were saved compared with conventional development. Methods for uncertainty quantification and applicability-domain assessment should become standard. Hybrid mechanistic-ML models deserve further development for dissolution, precipitation and stability. Finally, systematic comparisons between AI-guided and conventional development of the same drug would give the most convincing evidence of value.

14. FUTURE PERSPECTIVES

The most likely near-term development is closer integration of prediction and experiment. Automated or high-throughput platforms can generate consistent data at scale, active learning can direct what is measured, and ML can in turn shorten the cycle between hypothesis and test.^{9,14} Multi-scale integration, in which molecular predictions feed formulation-phase models and these feed PBPK simulation, is a longer-term goal that will depend on better uncertainty handling at each step. Explainable models that connect predictions to

recognizable physical factors (lipophilicity, lattice energy, hydrogen-bonding, miscibility) will be more acceptable to formulators and regulators than purely statistical ones.

Large language models and other generative tools may help with literature extraction and with the structuring of formulation databases, but their outputs require verification against primary sources, since errors in extracted data would propagate into models. For students and early-career researchers in pharmaceuticals, practical skills in experimental design, data curation and basic ML, together with an understanding of validation, are likely to become part of routine formulation training.

CONCLUSION

AI and ML are increasingly applicable to the solubility enhancement of poorly water-soluble drugs, from predicting aqueous and formulation-phase solubility to selecting excipients, optimizing compositions and connecting formulation to in vivo exposure through PBPK modelling. The strongest studies combine computational prediction with experimental confirmation, as illustrated by tools that were tested against measured solubilities and by approaches that couple automated screening with ML. The principal limitations are the size, consistency and availability of data, limited interpretability, infrequent external and prospective validation, and evolving regulatory expectations. AI should therefore be regarded as a complement to, and not a replacement for, experimental formulation science. Its value will be determined by whether predictions are confirmed in the laboratory, reduce the experimental effort needed, and ultimately yield formulations with demonstrated improvement in pharmaceutical performance.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the Department of Pharmacy, Vikas Group of Institutions, Nunna, Vijayawada, for providing the academic environment and resources that supported the preparation of this review article.

CONFLICT OF INTEREST

The authors declare no conflict of interest.



REFERENCES

1. Amidon GL, Lennernäs H, Shah VP, Crison JR. A theoretical basis for a biopharmaceutical drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability. *Pharm Res.* 1995;12 (3):413-420. doi:10.1023/A:1016212804288
2. Savjani KT, Gajjar AK, Savjani JK. Drug solubility: importance and enhancement techniques. *ISRN Pharm.* 2012;2012:195727. doi:10.5402/2012/195727
3. Williams HD, Trevaskis NL, Charman SA, et al. Strategies to address low drug solubility in discovery and development. *Pharmacol Rev.* 2013;65 (1):315-499. doi:10.1124/pr.112.005660
4. Lipinski CA. Drug-like properties and the causes of poor solubility and poor permeability. *J Pharmacol Toxicol Methods.* 2000;44 (1):235-249. doi:10.1016/S1056-8719(00)00107-6
5. Serajuddin ATM. Salt formation to improve drug solubility. *Adv Drug Deliv Rev.* 2007;59 (7):603-616.
6. Davis ME, Brewster ME. Cyclodextrin-based pharmaceuticals: past, present and future. *Nat Rev Drug Discov.* 2004;3 (12):1023-1035. doi:10.1038/nrd1576
7. Pouton CW. Formulation of poorly water-soluble drugs for oral administration: physicochemical and physiological issues and the lipid formulation classification system. *Eur J Pharm Sci.* 2006;29 (3-4):278-287.
8. Baghel S, Cathcart H, O'Reilly NJ. Polymeric amorphous solid dispersions: a review of amorphization, crystallization, stabilization, solid-state characterization, and aqueous solubilization of biopharmaceutical classification system class II drugs. *J Pharm Sci.* 2016;105 (9):2527-2544.
9. Bao Z, Huynh S, Le Devedec F, et al. The growing impact of machine learning on drug formulation science. *Adv Drug Deliv Rev.* 2026;229:115761. doi:10.1016/j.addr.2025.115761
10. Lucas I, Sousa J, Vitorino C. Integrating artificial intelligence into drug delivery systems: formulation development and current challenges. *Eur J Pharm Biopharm.* 2026;226:115133. doi:10.1016/j.ejpb.2026.115133
11. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. *J Chem Inf Comput Sci.* 2004;44 (3):1000-1005. doi:10.1021/ci034243x
12. Sorkun MC, Khetan A, Er S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Sci Data.* 2019;6:143. doi:10.1038/s41597-019-0151-1
13. Patel S, Kalasariya A, Desai J, Patel M, Patel A, Shah U. Machine learning-based prediction of drug solubility in lipidic environments: the Sol_ME tool for optimizing lipid-based formulations with a preliminary apalutamide case study. *AAPS PharmSciTech.* 2025;26 (2):50. doi:10.1208/s12249-025-03051-5
14. Lu T, Wu Y, Xiong P, et al. Combining high-throughput screening and machine learning to predict the formation of both binary and ternary amorphous solid dispersion formulations for early drug discovery and development. *Pharm Res.* 2025;42 (4):697-709. doi:10.1007/s11095-025-03853-z
15. Nguyen TH, Nguyen TL, Nguyen DD, et al. (author list to be confirmed against the journal page). Towards safer and efficient formulations: machine learning approaches to predict drug-excipient compatibility. *Int J Pharm.* 2024;653:123884. doi:10.1016/j.ijpharm.2024.123884
16. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns (N Y).* 2023;4 (9):100804. doi:10.1016/j.patter.2023.100804
17. Wang W, Ye Z, Gao H, Ouyang D. Computational pharmaceuticals - a new paradigm of drug delivery. *J Control Release.* 2021;338:119-136. doi:10.1016/j.jconrel.2021.08.030
18. Lovrić M, et al. Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: generalization, complexity, or predictive ability? *J Chemom.* 2021;35 (7):e3349. doi:10.1002/cem.3349



19. Bergström CAS, Larsson P. Computational prediction of drug solubility in water-based systems: qualitative and quantitative approaches used in the current drug discovery and development setting. *Int J Pharm.* 2018;540 (1-2):185-193.
20. Sun H, Shah P, Nguyen K, et al. Predictive models of aqueous solubility of organic compounds built on a large dataset of high integrity. *Bioorg Med Chem.* 2019;27 (14):3110-3114. doi:10.1016/j.bmc.2019.05.037
21. Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat Commun.* 2020;11:5753. doi:10.1038/s41467-020-19594-z
22. Authors to be added from the journal page. A machine learning model (Sol_ME) assisted development of a unit-dose lipid formulation for apalutamide: formulation, pharmacokinetics, and pharmacodynamics evaluation. *Int J Pharm.* 2025. (ScienceDirect PII S0378517325009500)
23. Yu LX. Pharmaceutical quality by design: product and process development, understanding, and control. *Pharm Res.* 2008;25 (4):781-791. doi:10.1007/s11095-007-9511-1
24. Authors to be added from PubMed. Integrating artificial intelligence and physiologically based pharmacokinetic modeling to predict in vitro and in vivo fate of amorphous solid dispersions. 2025. PubMed PMID: 40825400
25. US Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products: draft guidance for industry. Silver Spring (MD): US FDA; 2025. Available from: <https://www.fda.gov/media/184830/download>
26. European Medicines Agency. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Amsterdam: EMA; 2024. Available from: <https://www.ema.europa.eu/system/files/documents/scientific-guideline/reflection-paper-use-artificial-intelligence-ai-medicinal-product-lifecycle-en.pdf>

HOW TO CITE: Dr. K. Narendra Kumar Reddy, K. Ravi, Kodali. Ragini, Nannam. Ramabrahmam, Artificial Intelligence and Machine Learning in Solubility Enhancement of Poorly Water-Soluble Drugs: From Solubility Prediction to Formulation Optimization, *Int. J. of Pharm. Sci.*, 2026, Vol 4, Issue 10, 2903-2915. <https://doi.org/10.5281/zenodo.23466730>

