(ISSN: 0973-7049(P))

The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(3): S.I (3), 779-798, 2026 www.thebioscan.com

Artificial Intelligence and Machine Learning for Drug Solubility Enhancement: Predictive Modelling, Formulation Optimization, and Translational Challenges

Running title: AI and ML for drug solubility enhancement

Zankhana Sheth[1], Chirag Parmar[1], Niraj Pandya[1], Adil Patel[2*]

[1] Shivam Pharmaceutical Studies and Research Center, Valasan, Anand, Gujarat, India
[2]Department of Pharmaceutics & Pharmaceutical Technology, Ramanbhai Patel College of Pharmacy, Charotar University of Science and Technology, Changa, Gujarat, India.
*Corresponding Author: Dr. Adil Patel, Email: adilpatel.ph@charusat.ac.in
DOI: https://doi.org/10.63001/tbs.2026.v21.i03.S.I(3).pp779-798

KEYWORDS

Artificial intelligence; Machine learning; Solubility; Drug delivery systems; Formulation design; Pharmaceutical preparations

Article Info
Received:
21.06.2026
Accepted:
19.07.2026
Published:
06-08-2026

Abstract

Background: Poor aqueous solubility remains a major constraint in drug development because it can limit dissolution, oral absorption, dose feasibility and formulation robustness. Artificial intelligence (AI) and machine learning (ML) are increasingly used to predict solubility-related properties and guide enabling-formulation design. **Objective:** This review critically evaluates AI/ML approaches for drug-solubility enhancement, with particular emphasis on the relationship between predicted endpoints and experimentally meaningful pharmaceutical outcomes. **Methods:** Peer-reviewed literature published from January 2010 to 16 July 2026 was identified through structured searches of PubMed/MEDLINE and major scholarly publisher databases, supplemented by backward and forward citation tracking. Evidence concerning aqueous-solubility prediction, amorphous solid dispersions, cocrystals and salts, nanocrystals and nanoparticulate systems, cyclodextrin complexes, solvent and excipient selection, generative design, high-throughput screening and automated experimentation was synthesised critically. **Key findings:** Evidence is strongest for molecular-solubility prediction, amorphous solid-dispersion development, cocrystal screening and nanoformulation optimization. AI/ML can narrow experimental search spaces, integrate molecular, material and process variables, and support adaptive experiment selection. However, most studies remain retrospective and optimise surrogate endpoints such as logS, miscibility, cocrystal formation, particle size or complexation affinity rather than prospectively verified biorelevant dissolution or systemic exposure. Heterogeneous datasets, inconsistent endpoint definitions, information leakage, limited external validation, uncertain applicability domains, incomplete uncertainty reporting and restricted interpretability constrain translation. **Conclusion:** AI is most defensible as an uncertainty-aware decision-support layer integrated with mechanistic modelling, Quality by Design and prospective experimentation. Regulatory credibility will require predefined contexts of use, transparent data provenance, independent validation, human oversight and controlled lifecycle governance.

Abbreviations

AI, artificial intelligence; ASD, amorphous solid dispersion; BCS, Biopharmaceutics Classification System; DoE, design of experiments; ML, machine learning; PBPK, physiologically based pharmacokinetic; QbD, Quality by Design; QSPR, quantitative structure-property relationship.

1. Introduction

Aqueous solubility is a fundamental determinant of whether a pharmacologically active molecule can be converted into a practical drug product. Approximately 40% of approved drugs and a substantially larger proportion of

development candidates have been estimated to exhibit poor aqueous solubility, although prevalence varies with dataset composition and the definition applied.[1] Insufficient solubility may restrict the concentration available for absorption, delay dissolution, increase dose variability and complicate preclinical, clinical and manufacturing development. Solubility is not an invariant molecular property: reported values depend on pH, ionic strength, temperature, ionization state, solid form, equilibration time and analytical procedure.

The Biopharmaceutics Classification System links solubility and intestinal permeability to oral absorption.[2] Class II compounds have low solubility and high permeability, whereas class IV compounds are limited by both properties. The Developability Classification System extends this framework by incorporating dose, intestinal conditions, dissolution-rate limitation and the approximate particle size required to overcome dissolution-controlled absorption.[3] These classifications emphasis that a formulation should not be judged from solubility alone. Solubilizing excipients can increase total drug concentration without a proportional increase in the freely dissolved fraction, and the resulting change in permeability may be favorable, neutral or adverse.[4]

Established enabling approaches include salt formation, crystallization, amorphous solid dispersions (ASDs), particle-size reduction, nanocrystals, cyclodextrin inclusion, surfactant and cosolvent systems, lipid-based formulations and inorganic carriers.[1] Their selection is not interchangeable because each modifies a different combination of lattice energy, surface area, wetting, molecular association, supersaturation and precipitation. Conventional development combines expert judgement, preformulation studies, screening and design of experiments (DoE), but the number of possible drugs, excipients, ratios, processes and storage conditions can create a high-dimensional optimization problem.

AI and ML provide tools for classification, regression, ranking, generation and adaptive experiment selection. In pharmaceutical development they have been proposed for candidate prioritization, excipient selection, critical-quality-attribute prediction and formulation-process optimization.[5,6] Their value, however, depends on whether the modelled endpoint represents the intended pharmaceutical

decision. This review evaluates AI/ML approaches across major solubility-enhancement technologies and examines data provenance, validation, interpretability, applicability domains, experimental confirmation and translational readiness.

2. Review scope and literature identification

This article was developed as a critical narrative review informed by a structured literature search. Peer-reviewed publications dated from 1 January 2010 to 16 July 2026 were identified through PubMed/MEDLINE and searches of ScienceDirect, ACS Publications, SpringerLink, Royal Society of Chemistry, Wiley and other publisher platforms. Backward and forward citation tracking was used to identify additional primary studies. Search concepts combined artificial intelligence, machine learning, deep learning, neural networks, random forests, support-vector machines, gradient boosting, graph neural networks, transformers, QSPR/QSAR and generative models with aqueous solubility, dissolution, supersaturation, amorphous solid dispersion, drug-polymer miscibility, cocrystal, salt screening, nanocrystal, nanoparticle, cyclodextrin, excipient compatibility, formulation optimization, active learning, robotics and digital twins.

Eligible studies developed, validated or applied an identifiable data-driven method to a small-molecule drug, excipient or pharmaceutical formulation and assessed an endpoint directly relevant to solubility enhancement or formulation selection. Primary evidence included intrinsic, apparent or kinetic solubility; dissolution and supersaturation; ASD formation or stability; cocrystal or salt formation; cyclodextrin complexation; nanoformulation attributes; excipient selection and compatibility; and formulation or process optimization. General ADMET studies without independent analysis of solubility, non-pharmaceutical materials studies, conference abstracts, patents, non-peer-reviewed preprints and purely conventional DoE studies without a learning component were excluded from the primary synthesis. Reviews were used for context and citation discovery but were not treated as independent confirmation of model performance.

The literature is heterogeneous in endpoint definitions, datasets, algorithms and validation

practices; therefore, quantitative pooling was not appropriate. Evidence was synthesized by formulation technique, with particular attention to independent chemical systems, external or prospective validation, experimental confirmation, applicability-domain assessment and the difference between authors' conclusions and the present critical synthesis.

3. Conceptual basis of AI-assisted solubility enhancement

Thermodynamic solubility describes the equilibrium concentration in contact with a defined solid phase. Intrinsic solubility concerns the neutral species of an ionisable compound, whereas apparent or total solubility may include ionised, complexed or colloid-associated drug. Kinetic solubility is commonly measured over a restricted period after dilution of a concentrated stock and may reflect transient supersaturation. Dissolution is a rate process influenced by solubility, surface area, hydrodynamics, wetting and solid-state properties. These endpoints are related but not interchangeable; heterogeneous labels impose an upper limit on model reliability.[7–10]

Input representations range from calculated physicochemical descriptors and molecular fingerprints to simplified molecular-input line-entry system strings, molecular graphs, three-dimensional coordinates, formulation variables, spectra, images and process time series. Descriptor-based models are often interpretable and data-efficient. Deep models can learn task-specific representations directly from strings or graphs, but greater architectural complexity does not ensure better prediction. Recursive graph models, molecular transformers and graph-attention networks have all been applied to solubility prediction.[11,12] Direct comparisons have shown that descriptor-based neural networks or tree ensembles can equal or outperform graph and three-dimensional architectures in some datasets.[13–16]

Validation must reproduce the intended context of use. Random splitting can produce optimistic estimates when duplicate measurements, close analogues or records from the same formulation appear in both training and test sets. Drug-wise, scaffold-wise, pair-wise, temporal and institution-wise separation provide more demanding evidence. Performance metrics should be accompanied by uncertainty, calibration and applicability-domain assessment. Feature attribution can explain model behaviour,

but it does not establish causality. The scientifically appropriate workflow is therefore to define an experimentally meaningful endpoint, compare simple and complex baselines, quantify uncertainty, assess domain membership and prospectively test model-selected candidates.

4. AI and ML applications across solubility-enhancement techniques

4.1. Molecular aqueous-solubility prediction

Aqueous-solubility prediction is the most mature application. AqSolDB consolidated 9,982 unique compounds from nine public sources and enabled broad benchmarking, but it also exposed conflicting values and incomplete experimental metadata.[7] Quality-oriented curation and consensus modelling subsequently demonstrated that a smaller, better-qualified dataset may outperform indiscriminate aggregation of heterogeneous records.[8] Curated thermodynamic and kinetic datasets with quality weighting improved the performance of Chemprop and AttentiveFP models and achieved strong rank correlations in selected pharmaceutical series, although performance remained series-dependent.[10]

Representation learning has expanded from recursive graph networks to attention-based transformers and graph neural networks.[11,12] Comparative studies indicate that molecular descriptors and fingerprints remain competitive, particularly when data are noisy or limited.[13–16] The principal unresolved problem is prospective reliability: published models can perform substantially worse on newly curated external compounds than on familiar benchmark datasets.[9] Molecular logS prediction should therefore be regarded as an upstream risk-assessment layer. Selection among salts, cocrystals, ASDs, nanocrystals, cyclodextrins and lipid systems requires additional information on pKa, dose, melting point, polymorphism, permeability, precipitation, compatibility and manufacturability.

Structure-only prediction has a defined ceiling in formulation development. Molecular graphs, fingerprints or descriptors can encode ionisable groups and lipophilicity, but they do not identify the experimental pH, polymorphic state, crystal habit or residual solvent unless those variables are supplied explicitly. Consequently, models trained on unqualified logS values may return apparently precise

predictions for physically different endpoints. For candidate prioritisation, the preferred output is a range or calibrated probability accompanied by domain membership and a statement of the measurement definition. Prospective value should be tested using compounds and assay conditions concealed before model locking, rather than by repeated optimization against established public benchmarks.[8–10]

4.2. Amorphous solid dispersions and drug-polymer systems

ASDs enhance dissolution by maintaining a drug in a high-energy non-crystalline state within a carrier that may improve wetting and inhibit nucleation or crystal growth. ML studies have addressed physical stability, manufacturability, dissolution and pharmacokinetic translation. Han et al. modelled three- and six-month stability using 646 literature-derived formulations involving 50 drugs and 25 polymers; random-forest and deep-neural-network models achieved approximately 82.5% test accuracy under the study's joint stability criterion.[17] PharmSD extended the programme into a web platform that predicts stability and dissolution-related attributes, although the underlying evidence overlaps with the earlier dataset.[18]

Hybrid frameworks have linked ML with molecular dynamics and physiologically based pharmacokinetic (PBPK) modelling. Gao et al. used 674 dissolution profiles to classify supersaturation behaviour and estimate time-dependent dissolution, then applied molecular simulation and PBPK modelling to selected systems.[19] Jiang et al. incorporated molecular fingerprints, drug loading, polymer properties and hot-melt-extrusion variables to classify amorphisation and chemical stability, reporting accuracies of 92.8% and 96.0% for the best models within the study dataset.[20] High-throughput microscale preparation has also generated standardised positive and negative examples: a 1,272-formulation study combined powder X-ray diffraction with nested cross-validation and reported 96.7% random-forest accuracy for ASD formation.[21]

The most clinically oriented work integrated molecular-dissolution prediction with PBPK models for marketed ASDs. A dataset of 238 formulations and 1,693 time points was used to model molecularly dissolved drug, and most pharmacokinetic predictions for 37 marketed formulations were reported within two-fold of observed values.[22] These studies show

progression from isolated classification towards multiscale decision support. Nevertheless, amorphisation, physical stability, dissolution and systemic exposure remain distinct endpoints. Drug-wise prospective validation, long-term storage and direct comparison with expert-led or DoE development remain necessary.

Drug-polymer miscibility deserves separate consideration from amorphisation. A dispersion can appear X-ray amorphous immediately after manufacture while containing composition fluctuations or drug-rich domains that later nucleate. Predictors therefore need to integrate drug and polymer chemistry with drug loading, glass-transition behaviour, processing temperature, residual solvent or moisture and the storage protocol. Polymer identity encoded as a simple category is unlikely to transfer to an unseen carrier; molecular or material descriptors are preferable when extrapolation is claimed. Equally, dissolution models should distinguish total dispersed drug from the molecularly dissolved fraction and should preserve the medium, drug loading and time scale of precipitation. The most informative workflow would rank polymers jointly for manufacturability, physical stability and sustained supersaturation rather than optimise any one response in isolation.[17,19,20,22]

4.3. Cocrystal and salt screening

Machine learning has been widely used to prioritise cocrystal coformers from molecular or crystallographic information.[23–25] The main methodological weakness is construction of the negative class. Randomly generated molecular pairs are not equivalent to combinations experimentally shown to fail under defined crystallisation conditions, and random pair splitting may place the same drug or common coformer in both training and test sets.

Hybrid approaches combine empirical classification with physics-based descriptors. A combined ML/COSMO-RS workflow used cocrystallisation probability and calculated excess mixing enthalpy, with improved performance over either component in the reported evaluation.[26] An XGBoost model based on molecular descriptors reported a success rate above 90% within its development framework.[27] Prospective imatinib screening identified eight new multicomponent solid forms and selected products with improved physicochemical performance.[28] Conversely, direct comparison in a caffeine system found COSMO-RS more

predictive than the evaluated ML models, demonstrating that algorithmic novelty does not guarantee superiority.[29]

Recent models have expanded from binary cocrystal classification to richer solid-form decisions. A hybrid graph-isomorphism-network/Mordred model reported balanced accuracy of 0.916 across 7,395 successful and unsuccessful pairs.[30] DualNet was trained on 22,298 experimentally validated entries and simultaneously classified salts, cocrystals and physical mixtures, with calibrated uncertainty and prospective ciprofloxacin ranking.[31] Formation prediction is nevertheless only an upstream step. The resulting form must be characterised and evaluated for solubility, dissolution, hygroscopicity, stability and manufacturability.

Salt-specific machine-learning evidence remains less mature than cocrystal prediction. Counterion selection requires ionisation and proton-transfer information together with solid-form propensity, but the development decision also depends on hygroscopicity, disproportionation, chemical stability and performance across gastrointestinal pH. DualNet represents an important progression because it distinguishes salts, neutral cocrystals and physical mixtures, yet its classification endpoint cannot identify the pharmaceutically preferred form. A defensible platform should therefore use sequential models: formation and identity prediction followed by experimentally anchored property ranking, with uncertainty retained at both stages.[31]

4.4. Nanocrystals and nanoparticulate systems

Nanocrystals enhance dissolution mainly through increased surface area and, at very small dimensions, curvature-related increases in saturation solubility. Carrier-based nanoparticles introduce additional objectives such as drug loading, encapsulation, release and colloidal stability. He et al. compiled 910 particle-size and 341 polydispersity records across wet bead milling, high-pressure homogenisation and antisolvent precipitation. Light-gradient-boosting models performed best for the two top-down processes, whereas precipitation was less predictable, plausibly reflecting unrepresented mixing and nucleation dynamics.[32]

Formulation-specific studies have modelled solid-lipid-nanoparticle size, lecithin/chitosan nanoparticle release and microfluidic liposome manufacture.[33–35] More than 1,300

systematically generated liposome formulations were used to predict formation, size and process settings with XGBoost and inverse design, supported by wet-laboratory validation.[36] Prospective ML-assisted microfluidics produced liposomes close to target sizes of approximately 100 and 600 nm.[37] TuNa-AI extended optimization to drug-excipient combinations and ratios, increasing venetoclax loading and reducing an excipient in a trametinib system while preserving evaluated performance.[38]

These results establish that ML can narrow process and composition spaces, but particle size, polydispersity and encapsulation efficiency are surrogate quality attributes rather than direct evidence of improved molecular solubility or exposure. Future nanoformulation studies should optimise multicomponent target-product profiles and report biorelevant dissolution, free-drug concentration, stability, pharmacokinetics and process-scale robustness.

Nanoformulation development is inherently multi-objective. Reducing particle size may accelerate dissolution while increasing interfacial energy, aggregation and Ostwald-ripening risk; greater carrier content may improve encapsulation while reducing drug loading and increasing dose volume. Model-selected optima should therefore be reported as trade-off surfaces under explicit route, dose and excipient constraints rather than as a single desirability score. Prospective studies also need to test redispersibility after drying, stability during storage and the concentration of freely available drug after dilution or digestion. These measurements determine whether improved manufacturing attributes translate into useful biopharmaceutical performance.[32,36,38]

4.5. Cyclodextrin complexation

Cyclodextrins can increase apparent solubility by forming non-covalent host-guest associations, but performance depends on cavity complementarity, substitution, ionisation, stoichiometry, concentration and auxiliary components.[39] QSPR models for beta-cyclodextrin and sulfobutylether-beta-cyclodextrin used ensemble regression to predict complexation free energies.[40] A larger dataset of approximately 3,000 observations and 1,320 guest structures was modelled using LightGBM, random forests and deep learning; LightGBM achieved a mean absolute error of 1.38 kJ/mol and an R-squared value of 0.86 within the reported evaluation.[41]

Formulation-level work has moved beyond binary host-guest affinity. Random-forest models based on 596 ternary drug-cyclodextrin-polymer records predicted total solubility and ternary-to-binary enhancement, with reported R-squared values of 0.887 and 0.815.[42] Classification models have also been developed to prioritise inclusion-complex formation using recall- or precision-oriented thresholds according to the cost of false negatives and false positives.[43] Stronger binding should not automatically be equated with better product performance because excessive association may reduce freely dissolved drug or slow release. Prospective formulation selection should therefore measure total and unbound drug, phase-solubility behaviour, dissociation, precipitation and permeability.

The formulation target should not be maximum host-guest affinity alone. Weak binding may provide inadequate solublization, whereas very strong association can reduce the freely dissolved fraction or delay release. Cyclodextrin concentration, substitution pattern, complex stoichiometry and self-association further influence the phase-solubility response. Future models should therefore predict both total and unbound drug, identify the relevant host grade and incorporate permeability or dissociation measurements when oral absorption is the objective. Ternary drug-cyclodextrin-polymer models are promising because they move closer to formulation-level solubility, but their transferability requires leave-one-drug-out and leave-one-host-class-out validation.[4,41,42]

4.6. Solvent, excipient and component selection

Data-driven component selection links molecular prediction to practical formulation design. Deep-learning models have learned recurring active-ingredient/excipient patterns from historical formulation datasets, but such outputs reproduce published practice rather than prove optimality.[44] High-throughput experiments generated 2,304 measurements for three BCS class II drugs, 11 excipients and multiple biorelevant media, showing that solublization depends jointly on drug type, pH, excipient concentration and time.[45]

Compatibility is an essential constraint. PharmDE combines 532 literature-derived incompatibility records with 60 expert rules to support traceable risk assessment.[46] DE-INTERACT used molecular fingerprints to classify 179 incompatible and 3,549 compatible

pairs and experimentally examined selected predicted incompatibilities.[47] A subsequent stacking ensemble reported improved class-specific performance and identified 10 of 12 experimental incompatibilities compared with three detected by DE-INTERACT.[48] These systems should be treated as preformulation triage because compatibility also depends on component ratio, impurities, water activity, processing and storage.

Sol_ME addressed drug solubility in lipidic environments using 1,379 drug-solvent records and 35 experimental combinations; a preliminary apalutamide case study reduced formulation volume by 75%.[49] Retrospective analysis of registered ibuprofen products linked formulation patterns and drug forms with human pharmacokinetic characteristics.[50] Such observational associations are translationally relevant but are vulnerable to confounding and survivorship bias. The preferred output is a ranked list constrained by safety, dose, compatibility, precipitation risk, processability and uncertainty rather than a single unconstrained optimum.

Compatibility predictions are conditional rather than absolute. Component ratio, water activity, excipient impurities, residual solvent, processing stress and packaging can convert an apparently compatible molecular pair into an unstable dosage form. Similarly, high solubility in an oil or surfactant does not demonstrate maintenance of solublization after aqueous dilution or gastrointestinal digestion. Component-selection tools should therefore operate as linked filters: solubility ranking, incompatibility-risk assessment, precipitation testing, permitted-exposure constraints and process feasibility. Negative predictions should trigger targeted experiments rather than automatic exclusion, particularly when the candidate lies outside the modelled excipient or chemical domain.[46,48,49]

Table 1. Representative AI/ML applications relevant to drug-solubility enhancement.

Application	Representative evidence	Modelled endpoint or decision	Principal translational limitation
Aqueous solubility	AqSolDB, quality-oriented curation and deep/ensemble models.[7,8,10]	Intrinsic/apparent logS, regression or rank prediction	Heterogeneous assay conditions; solid form and pH often absent
Amorphous solid dispersions	Stability, dissolution and PBPK-linked models.[17,19,22]	Physical stability, dissolution class/time course, exposure	Dataset overlap; limited drug-wise prospective validation
Cocrystals and salts	ML, COSMO-RS hybrids and multiclass solid-form prediction.[26,29,31]	Formation probability; salt/cocrystal/mixture ranking	Formation does not establish solubility, stability or manufacturability

Application	Representative evidence	Modelled endpoint or decision	Principal translational limitation
Nanocrystals and nanoparticles	Process models, inverse design and tunable composition systems.[32,36,38]	Size, polydispersity, formation, loading and process settings	Critical quality attributes are not direct measures of free drug or exposure
Cyclodextrin systems	Affinity, inclusion classification and ternary solubility models.[40–42]	Complexation free energy, formation and total solubility	Binding strength may not maximise free-drug concentration or absorption
Excipient and compatibility selection	Expert systems, fingerprint models and lipid-solubility tools.[46,48,49]	Compatibility risk, solvent/excipient ranking	Concentration, impurities, process and storage context incompletely represented
Generative and autonomous systems	Digital microstructures, generative formulations and self-driving manufacture.[51,52,56]	Candidate generation, process selection and adaptive optimization	Direct end-to-end solubility-enhancement validation remains absent

5. Generative AI, high-throughput experimentation and robotics

Generative models propose new compositions or product structures rather than only scoring supplied candidates. A generative structure-synthesis method created three-dimensional digital tablet and implant microstructures conditioned on particle size and drug loading and reproduced selected transport and percolation characteristics of experimental samples.[51] Conditional generative adversarial networks trained on 1,437 three-dimensional-printing formulations generated 270 candidates, selected formulations were fabricated and at least one AI-generated formulation was successfully printed.[52] These studies establish technical feasibility but do not yet demonstrate autonomous solubility-enhancement design.

Integrated platforms and automated laboratories provide the infrastructure for closed-loop development. FormulationAI combines predictive models across cyclodextrins, solid dispersions, phospholipid complexes, nanocrystals, self-emulsifying systems and liposomes.[53] A spray-drying robot automated combinatorial preparation of multicomponent solid dispersions, reduced processing time approximately eight-fold and limited cross-contamination.[54] Active-learning systems coupled with automated small-angle scattering have mapped soft-material formulation spaces, although not specifically pharmaceutical solubility systems.[55]

A self-driving tableting data factory integrated a digital formulator, automated powder handling, tablet manufacture, testing and Bayesian optimization; the reported workflow reduced development time to approximately 6 h and active-ingredient use by 65% for a manufacturability-focused metformin case study.[56] Proposed self-driving biomaterials architectures and current optimization reviews describe the transition from automated execution to active selection of the next experiment.[57,58]

Direct pharmaceutical examples include active-learning optimization of an enormous aceclofenac nanoformulation space and batch Bayesian optimization of oral microemulsions.[59,60]

No reviewed study demonstrated a fully autonomous workflow that selected an enabling strategy for a previously unseen poorly soluble drug and prospectively confirmed improved biorelevant dissolution or systemic exposure. The near-term role of generative and autonomous methods is therefore constrained candidate proposal, uncertainty-aware experiment selection and standardised data generation under expert-defined pharmaceutical limits.

6. Traditional versus AI-assisted formulation development

AI should be compared with the strongest relevant conventional alternative rather than with an oversimplified concept of trial and error. Expert-led screening efficiently uses tacit knowledge for small or unfamiliar design spaces. DoE provides prospectively balanced data, interpretable effects and statistically defined local design spaces. Mechanistic models represent thermodynamics, nucleation, dissolution, diffusion or pharmacokinetics and can extrapolate when their assumptions remain valid. ML accommodates higher-dimensional nonlinear relationships but may reproduce dataset-specific correlations.

The approaches are complementary. DoE can generate balanced training data; ML can model relationships that exceed low-order response surfaces; Bayesian optimization can adaptively select experiments; and mechanistic models can provide constraints or informative predictors. Hybrid ASD, cocrystal and cyclodextrin studies demonstrate the value of linking statistical predictions with molecular simulation, thermodynamic calculations or PBPK models.[19,22,26,41] Physics-based cocrystal screening outperformed selected ML models in a defined system, underscoring the need for direct comparison rather than assumed algorithmic hierarchy.[29]

The largest advantage of AI arises when it changes the sequence and information value of experiments. Active-learning and Bayesian-optimization studies show that large constrained formulation spaces can be explored without exhaustive testing.[59,60] Efficiency should nevertheless be measured as the total resources

required to reach a verified formulation decision, including data curation, model development, software maintenance and experimental confirmation. Table 2 compares the principal approaches.

Comparative studies should predefine the same drug-substance budget, candidate materials, experimental time and acceptance criteria for AI-assisted, DoE and expert-led arms. Reporting only the number of formulations tested favours computational approaches while overlooking curation, software and model-maintenance costs. Conversely, comparing AI with exhaustive grid screening understates the efficiency of established sequential experimentation. A meaningful endpoint is the probability of reaching a prospectively verified formulation within the complete resource constraint, together with the uncertainty and reproducibility of that decision.

Table 2. Comparative characteristics of formulation-development approaches.

Approach	Strengths	Limitations	Most appropriate use
Expert-led screening	Uses tacit pharmaceutical knowledge; flexible with unfamiliar compounds	Subjective; difficult to scale and reproduce	Initial strategy selection and small design spaces
High-throughput screening	Direct, standardised measurement; retains negative results	Resource-intensive if candidate space is very large	Generation of formulation-level training data
Design of experiments	Prospective balance, interpretable effects and local design-space definition	Limited dimensionality; low-order response surfaces may be inadequate	Defined factor ranges with a moderate number of variables
Mechanistic modelling	Physical interpretability and potential extrapolation	Assumptions and uncertain parameters may dominate	Processes with established governing relationships
Conventional ML	Flexible nonlinear prediction and rapid screening	Data-dependent; leakage and domain-shift risks	Relevant historical or high-throughput datasets
Deep learning	Learns representations from graphs, strings, images or time series	High data demand and lower transparency	Large molecular or multimodal datasets
Bayesian optimization/active learning	Adaptive and experiment-efficient	Can efficiently optimise an incomplete surrogate objective	Expensive experiments in constrained spaces
Hybrid ML-mechanistic models	Combines empirical flexibility with scientific structure	Complex validation and uncertainty propagation	Sparse data with reliable mechanistic knowledge
Autonomous experimentation	Closed-loop execution and systematic metadata capture	Infrastructure-intensive; requires human governance	Repeated workflows with machine-readable endpoints

7. Challenges and translational limitations

The dominant barriers are data-related. Formulation datasets are small, fragmented and selectively published; unsuccessful formulations, long-term stability results and scale-up failures are particularly scarce. Multiple time points or concentrations from one system inflate row counts without adding independent chemical diversity. Endpoint ambiguity compounds this problem: intrinsic, apparent, kinetic and thermodynamic solubility, total dispersed drug and molecularly dissolved drug

are frequently pooled despite different meanings.[7–10]

Information leakage occurs when test data contain exact duplicates, close analogues, shared drugs or repeated formulations represented during training. Similarity-aware splitting methods show that conventional random partitions can inflate performance in out-of-distribution applications.[61] Molecular benchmarks also exhibit coverage bias, and scaffold splitting does not guarantee representation of the wider pharmaceutical chemical space.[62] Validation should be matched to the claim: drug-wise separation for new active ingredients, excipient-wise separation for new materials, laboratory-wise separation for process transfer and prospective independent testing for broad deployment.

Interpretability, calibration and applicability domains remain inconsistently reported. Feature importance and Shapley values describe model behaviour but do not establish causality. Point predictions should be accompanied by uncertainty and a mechanism for abstention when the candidate lies outside the training domain. Reporting should include data provenance, endpoint definitions, preprocessing, partitions, hyperparameters, software versions, trained models and full performance outputs. FAIR data principles provide a foundation for reuse, while PROBAST+AI offers transferable domains for assessing bias and applicability even though it was not designed specifically for pharmaceutics.[63,64]

Regulatory credibility will depend on context of use and the consequence of model error. The US Food and Drug Administration's 2025 draft guidance proposes a risk-based framework for AI models used to support regulatory decisions, and the European Medicines Agency's reflection paper addresses AI across the medicinal-product lifecycle.[65,66] Models used only to rank exploratory experiments require less evidence than models used to define a design space, replace testing or control commercial manufacture. Lifecycle governance should include version control, drift detection, documented retraining rules, revalidation and auditable links between predictions, experiments and decisions.

Table 3. Major methodological challenges and minimum corrective practices.

Challenge	Risk to interpretation	Minimum corrective practice
Endpoint heterogeneity	The model learns mixed or source-specific labels	Define solubility type, pH, temperature, solid form, units and analytical method
Publication and survivorship bias	Successful formulations dominate training data	Retain experimentally documented failures and discontinued candidates
Information leakage	Internal test performance is inflated	Use drug-, pair-, batch-, laboratory- or time-grouped splitting
Restricted applicability domain	Precise predictions are returned for unsupported candidates	Report domain membership, uncertainty and abstention criteria
Class imbalance	Accuracy obscures failure to detect rare adverse outcomes	Report recall, precision, calibration and decision-specific thresholds
Interpretability overreach	Associations are mistaken for mechanisms	Confirm feature-based hypotheses through controlled experiments
Incomplete reproducibility	Models cannot be independently checked	Share data/code or controlled-access validation; report versions and partitions
Lifecycle drift	Performance degrades after material, process or software changes	Monitor drift, control retraining and revalidate changes

8. Future perspectives

Progress should be judged by prospective improvement in pharmaceutical decisions rather than marginal benchmark gains. Explainable AI should provide testable hypotheses, indicate domain membership and identify the most informative confirmatory experiment. Uncertainty should distinguish measurement error, model uncertainty, limited chemical coverage and omitted process variables. Physics-informed models can incorporate thermodynamics, mass balance, composition constraints, dissolution, precipitation and particle-growth relationships, while data-driven components can estimate uncertain parameters or correct systematic model error.

Integration with Quality by Design (QbD) offers a practical development framework. AI can link quality target product profiles, critical material attributes, critical process parameters and adaptive experimentation. Quality by Computational Design has been proposed as a broader paradigm combining AI, multiscale simulation and clinically guided inverse design.[67] A physiology-based ML/PBPK/molecular-dynamics framework for long-acting injectables illustrates how target exposure can be translated backwards into release and formulation requirements, although the platform is not direct oral-solubility evidence.[68]

Digital twins could connect material state, formulation microstructure, process measurements, dissolution and predicted exposure, provided that the model is operationally updated from the physical system.

Current reviews identify data integration, equipment transfer, uncertainty and regulatory complexity as major barriers.[69,70] Multimodal models may combine structures, spectra, diffraction, thermal data, images, dissolution curves and process time series, but they must preserve source provenance and endpoint context.

Federated learning may enable collaborative use of proprietary failed-formulation and stability data. MELLODDY demonstrated cross-company learning across more than 21 million compounds, over 40,000 assays and more than 2.6 billion activity records without centralising proprietary discovery data.[71] Transfer to formulation development will require common ontologies, harmonised experimental methods, privacy safeguards and partner-held external validation. Priority studies include multicentre solubility challenges with complete metadata, drug-wise ASD validation, two-stage cocrystal formation/property models, direct AI-versus-DoE nanoformulation benchmarks, cyclodextrin studies measuring free drug and end-to-end autonomous optimization using biorelevant dissolution or exposure as the objective.

Table 4. Priority studies required for translational validation.

Priority study	Essential design features	Outcome required for translation
Multicentre solubility challenge	Undisclosed compounds; full pH, temperature, ionisation and solid-form metadata	External error, calibration and chemical-domain performance
Drug-wise ASD validation	New drugs and polymers; at least six months of stability; biorelevant dissolution	Prospective stability and dissolution versus expert/DoE controls
Two-stage solid-form platform	Formation model followed by property model and experimental characterisation	Solubility, dissolution, hygroscopicity, stability and manufacturability
Nanoformulation benchmark	Identical drug, excipient list, budget and endpoints for AI, DoE and experts	Resource use, free drug, dissolution, stability and pharmacokinetics
Cyclodextrin external study	Total and unbound drug, phase solubility, dissociation and permeability	Host selection that improves biopharmaceutical performance
Federated formulation learning	Harmonised metadata and partner-held independent tests	Cross-organisation generalisation without centralising proprietary records
Autonomous enabling-formulation study	Closed-loop preparation, characterisation and active learning	Prospective improvement in biorelevant dissolution or exposure

9. Conclusion

AI and ML can reduce experimental search spaces, identify nonlinear formulation relationships and support candidate prioritization across molecular solubility prediction, ASDs, cocrystals and salts, nanoformulations, cyclodextrin systems and excipient selection. The most mature evidence concerns retrospective molecular and formulation prediction, while prospective experimental confirmation and independent replication remain less common.

The central limitation is the widespread optimization of surrogate endpoints. Accurate prediction of logS, miscibility, cocrystal formation, particle size or complexation affinity does not by itself establish stable biorelevant dissolution, manufacturability or improved systemic exposure. No modelling architecture is universally superior; descriptor-based ensembles, deep networks and physics-based methods each perform differently according to data quality and chemical domain.

The most scientifically defensible role of AI is therefore as an uncertainty-aware decision-support layer within mechanistically informed and experimentally closed-loop development. Credible translation requires context-rich data, explicit unsuccessful examples, leakage-resistant validation, prospective testing, applicability-domain assessment, transparent reporting and lifecycle control. The decisive measure of progress will be whether AI-assisted workflows reproducibly deliver better formulation decisions with fewer resources while maintaining experimentally verified pharmaceutical performance.

References

1. Xie B, Liu Y, Li X, Yang P, He W. Solubilization techniques used for poorly water-soluble drugs. Acta Pharm Sin B. 2024;14(11):4683-4716. doi:10.1016/j.apsb.2024.08.027.
2. Amidon GL, Lennernas H, Shah VP, Crison JR. A theoretical basis for a biopharmaceutic 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.
3. Butler JM, Dressman JB. The Developability Classification System: application of biopharmaceutics concepts to formulation development. J Pharm Sci. 2010;99(12):4940-4954. doi:10.1002/jps.22217.
4. Dahan A, Miller JM, Amidon GL. The solubility-permeability interplay and oral drug formulation design: two heads are better than one. Adv Drug Deliv Rev. 2016;101:99-107. doi:10.1016/j.addr.2016.04.018.
5. Bannigan P, Aldeghi M, Bao Z, Hase F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Adv Drug Deliv Rev. 2021;175:113806. doi:10.1016/j.addr.2021.05.016.
6. Bao Z, Bufton J, Hickman RJ, Aspuru-Guzik A, Bannigan P, Allen C. Revolutionizing drug formulation development: the increasing impact of machine learning. Adv Drug Deliv Rev. 2023;202:115108. doi:10.1016/j.addr.2023.115108.
7. 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.
8. Sorkun MC, Koelman JMVA, Er S. Pushing the limits of solubility prediction via quality-oriented data selection. iScience. 2021;24(1):101961. doi:10.1016/j.isci.2020.101961.
9. Llompart P, Minoletti C, Baybekov S, Horvath D, Marcou G, Varnek A, et al. Will we ever be able to accurately predict solubility? Sci Data. 2024;11:303. doi:10.1038/s41597-024-03105-6.
10. Meng J, Chen P, Wahib M, Yang M, Zheng L, Wei Y, et al. Boosting the predictive performance with aqueous solubility dataset curation. Sci Data. 2022;9:71. doi:10.1038/s41597-022-01154-3.
11. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model. 2013;53(7):1563-1575. doi:10.1021/ci400187y.
12. Francoeur PG, Koes DR. SolTranNet-A machine learning tool for fast aqueous solubility prediction. J Chem Inf Model. 2021;61(6):2530-2536. doi:10.1021/acs.jcim.1c00331.
13. Panapitiya G, Girard M, Hollas A, Sepulveda J, Murugesan V, Wang W, et al. Evaluation of deep learning architectures for aqueous solubility prediction. ACS Omega. 2022;7(18):15695-15710. doi:10.1021/acsomega.2c00642.
14. Lee S, Lee M, Gyak KW, Kim SD, Kim MJ, Min K. Novel solubility prediction models: molecular fingerprints and physicochemical features versus graph convolutional neural networks. ACS Omega. 2022;7(14):12268-12277. doi:10.1021/acsomega.2c00697.
15. Zheng T, Mitchell JBO, Dobson S. Revisiting the application of machine learning approaches in predicting aqueous solubility. ACS Omega. 2024;9:35209-35222. doi:10.1021/acsomega.4c06163.
16. Ghanavati MA, Ahmadi S, Rohani S. A machine learning approach for the prediction of aqueous solubility of pharmaceuticals: a comparative model and dataset analysis. Digit Discov. 2024;3:2085-2104. doi:10.1039/D4DD00065J.

17. Han R, Xiong H, Ye Z, Yang Y, Huang T, Jing Q, et al. Predicting physical stability of solid dispersions by machine learning techniques. J Control Release. 2019;311-312:16-25. doi:10.1016/j.jconrel.2019.08.030.
18. Dong J, Gao H, Ouyang D. PharmSD: a novel AI-based computational platform for solid dispersion formulation design. Int J Pharm. 2021;604:120705. doi:10.1016/j.ijpharm.2021.120705.
19. Gao H, Wang W, Dong J, Ye Z, Ouyang D. An integrated computational methodology with data-driven machine learning, molecular modeling and PBPK modeling to accelerate solid dispersion formulation design. Eur J Pharm Biopharm. 2021;158:336-346. doi:10.1016/j.ejpb.2020.12.001.
20. Jiang J, Lu A, Ma X, Ouyang D, Williams RO 3rd. The applications of machine learning to predict the forming of chemically stable amorphous solid dispersions prepared by hot-melt extrusion. Int J Pharm X. 2023;5:100164. doi:10.1016/j.ijpx.2023.100164.
21. Lu T, Wu Y, Xiong P, Zhong H, Ding Y, Li H, 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;42:697-709. doi:10.1007/s11095-025-03853-z.
22. Zhu J, Xiong P, Wang W, Lu T, Ouyang D. Integrating artificial intelligence and physiologically based pharmacokinetic modeling to predict in vitro and in vivo fate of amorphous solid dispersions. J Control Release. 2025;386:114123. doi:10.1016/j.jconrel.2025.114123.
23. Wang D, Yang Z, Zhu B, Mei X, Luo X. Machine-learning-guided cocrystal prediction based on large database. Cryst Growth Des. 2020;20(10):6610-6621. doi:10.1021/acs.cgd.0c00767.
24. Devogelaer JJ, Meekes H, Tinnemans P, Vlieg E, de Gelder R. Co-crystal prediction by artificial neural networks. Angew Chem Int Ed Engl. 2020;59(48):21711-21718. doi:10.1002/anie.202009467.
25. Mswahili ME, Lee MJ, Martin GL, Kim J, Kim P, Choi GJ, et al. Cocrystal prediction using machine learning models and descriptors. Appl Sci. 2021;11(3):1323. doi:10.3390/app11031323.
26. Yuan J, Liu X, Wang S, Chang C, Zeng Q, Song Z, et al. Virtual coformer screening by a combined machine learning and physics-based approach. CrystEngComm. 2021;23(35):6039-6044. doi:10.1039/D1CE00587A.
27. Yang D, Wang L, Yuan P, An Q, Su B, Yu M, et al. Cocrystal virtual screening based on the XGBoost machine learning model. Chin Chem Lett. 2023;34(8):107964. doi:10.1016/j.cclet.2022.107964.
28. Liang X, Liu S, Li Z, Deng Y, Jiang Y, Yang H. Efficient cocrystal coformer screening based on a machine learning strategy: a case study for the preparation of imatinib cocrystal with enhanced physicochemical properties. Eur J Pharm Biopharm. 2024;196:114201. doi:10.1016/j.ejpb.2024.114201.
29. Molajafari F, Li T, Abbasichaleshtori M, Hajian ZD, Cozzolino AF, Fandrick DR, et al. Computational screening for prediction of co-crystals: method comparison and experimental validation. CrystEngComm. 2024;26(11):1620-1636. doi:10.1039/D3CE01252B.
30. Ghanavati MA, Rohani S. Cocrystal formation prediction: hybrid GIN-Mordred model outperforms DFT-based methods. Cryst Growth Des. 2025;25(8):2717-2729. doi:10.1021/acs.cgd.5c00347.
31. Ghanavati MA, Khalili B, Alberico D, Rohani S. Machine learning-driven discovery of multicomponent pharmaceutical solid forms via DualNet: confidence-aware prediction and ranking of salts and cocrystals. Int J Pharm. 2025;684:126117. doi:10.1016/j.ijpharm.2025.126117.
32. He Y, Ye Z, Liu X, Wei Z, Qiu F, Li HF, et al. Can machine learning predict drug nanocrystals? J Control Release. 2020;322:274-285. doi:10.1016/j.jconrel.2020.03.043.
33. Ozturk AA, Gunduz AB, Ozisik O. Supervised machine learning algorithms for evaluation of solid lipid nanoparticles and particle size. Comb Chem High Throughput Screen. 2018;21(9):693-699. doi:10.2174/1386207322666181218160704.
34. Dawoud MHS, Mannaa IS, Abdel-Daim A, Sweed NM. Integrating artificial intelligence with Quality by Design in the formulation of lecithin/chitosan nanoparticles of a poorly water-soluble drug. AAPS PharmSciTech. 2023;24(6):169. doi:10.1208/s12249-023-02609-5.
35. Rebollo R, Oyoun F, Corvis Y, El-Hammadi MM, Saubamea B, Andrieux K, et al. Microfluidic manufacturing of liposomes: development and optimization by design of experiment and machine learning. ACS Appl Mater Interfaces. 2022;14(35):39736-39745. doi:10.1021/acsami.2c06627.
36. Eugster R, Orsi M, Buttitta G, Serafini N, Tiboni M, Casettari L, et al. Leveraging machine learning to streamline the development of liposomal drug delivery systems. J Control Release. 2024;376:1025-1038. doi:10.1016/j.jconrel.2024.10.065.
37. Jia Y, Liang X, Zhang L, Zhang J, Zafar H, Huang S, et al. Machine learning-assisted microfluidic approach for broad-spectrum liposome size control. J Pharm Anal. 2025;15(6):101221. doi:10.1016/j.jpha.2025.101221.
38. Zhang Z, Xiang Y, Laforet J, Spasojevic I, Fan P, Heffernan A, et al. TuNa-AI: a hybrid kernel machine to design tunable nanoparticles for drug delivery. ACS Nano. 2025;19(37):33288-33296. doi:10.1021/acsnano.5c09066.
39. Sarabia-Vallejo A, Caja MDM, Olives AI, Martin MA, Menendez JC. Cyclodextrin inclusion complexes for improved drug bioavailability and activity: synthetic and analytical aspects. Pharmaceutics. 2023;15(9):2345. doi:10.3390/pharmaceutics15092345.
40. Merzlikine A, Abramov YA, Kowsz SJ, Thomas VH, Mano T. Development of machine learning models of beta-cyclodextrin and sulfobutylether-beta-cyclodextrin complexation free energies. Int J Pharm. 2011;418(2):207-216. doi:10.1016/j.ijpharm.2011.03.065.
41. Zhao Q, Ye Z, Su Y, Ouyang D. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. Acta Pharm Sin B. 2019;9(6):1241-1252. doi:10.1016/j.apsb.2019.04.004.
42. Li J, Gao H, Ye Z, Deng J, Ouyang D. In silico formulation prediction of drug/cyclodextrin/polymer ternary complexes by machine learning and molecular modeling techniques. Carbohydr Polym. 2022;275:118712. doi:10.1016/j.carbpol.2021.118712.
43. Ma Y, Niu Y, Yang H, Dai J, Lin J, Wang H, et al. Prediction and design of cyclodextrin inclusion complexes formation via machine learning-based strategies. Chem Eng Sci. 2022;261:117946. doi:10.1016/j.ces.2022.117946.
44. Yang Y, Ye Z, Su Y, Zhao Q, Li X, Ouyang D. Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharm Sin B. 2019;9(1):177-185. doi:10.1016/j.apsb.2018.09.010.

45. Bahr MN, Matamoros SV, Campbell GA. A high throughput approach of selecting excipients for solubility enhancement of BCS Class II active pharmaceutical ingredients for oral dosage forms. Chem Eng Res Des. 2023;193:751-758. doi:10.1016/j.cherd.2023.04.011.
46. Wang N, Sun H, Dong J, Ouyang D. PharmDE: a new expert system for drug-excipient compatibility evaluation. Int J Pharm. 2021;607:120962. doi:10.1016/j.ijpharm.2021.120962.
47. Patel S, Patel M, Kulkarni M, Patel MS. DE-INTERACT: a machine-learning-based predictive tool for the drug-excipient interaction study during product development-validation through paracetamol and vanillin as a case study. Int J Pharm. 2023;637:122839. doi:10.1016/j.ijpharm.2023.122839.
48. Hang NT, Long NT, Duy ND, Chien NN, Van Phuong N. 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.
49. 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;26:50. doi:10.1208/s12249-025-03051-5.
50. Garamani D, Sjogren E, Mihranyan A. Rational formulation design through retrospective machine learning methodology: case study ibuprofen. Int J Pharm. 2026;693:126667. doi:10.1016/j.ijpharm.2026.126667.
51. Hornick T, Mao C, Koynov A, Yawman P, Thool P, Salish K, et al. In silico formulation optimization and particle engineering of pharmaceutical products using a generative artificial intelligence structure synthesis method. Nat Commun. 2024;15:9622. doi:10.1038/s41467-024-54011-9.
52. Elbadawi M, Li H, Sun S, Alkahtani ME, Basit AW, Gaisford S. Artificial intelligence generates novel 3D printing formulations. Appl Mater Today. 2024;36:102061. doi:10.1016/j.apmt.2024.102061.
53. Dong J, Wu Z, Xu H, Ouyang D. FormulationAI: a novel web-based platform for drug formulation design driven by artificial intelligence. Brief Bioinform. 2024;25(1):bbad419. doi:10.1093/bib/bbad419.
54. Klimsa V, Ruphuy G, Jonas J, Maskova L, Kaspar O, Zvatora P, et al. Spray drying robot for high-throughput combinatorial fabrication of multicomponent solid dispersions. Powder Technol. 2023;428:118872. doi:10.1016/j.powtec.2023.118872.
55. Martin TB, Sutherland DR, McDannald A, Kusne AG, Beaucage PA. Autonomous small-angle scattering for accelerated soft material formulation optimization. Chem Mater. 2025;37(12):4272-4281. doi:10.1021/acs.chemmater.5c00860.
56. Abbas F, Salehian M, Hou P, Moores J, Goldie J, Tsioutsios A, et al. Accelerated drug development using a digital formulator and a self-driving tableting data factory. Nat Commun. 2026;17:4739. doi:10.1038/s41467-026-71204-6.
57. Tamasi MJ, Gormley AJ. Biologic formulation in a self-driving biomaterials lab. Cell Rep Phys Sci. 2022;3(9):101041. doi:10.1016/j.xcrp.2022.101041.
58. Ros H, Chan N, Cook MT, Shorthouse D. Artificial intelligence and machine learning guided optimization in drug delivery. Adv Drug Deliv Rev. 2026;232:115781. doi:10.1016/j.addr.2026.115781.
59. Bao Z, Le Devedec F, Huynh S. A data-driven workflow for nanomedicine optimization using active learning and automated experimentation. Mol Pharm. 2025;22(12):7465-7477. doi:10.1021/acs.molpharmaceut.5c00933.
60. Gunawardena M, Chau B, Nothacker H, Pangeni R, Roper T, Xu Q, et al. Machine learning guided optimization of an oral microemulsion system: a Bayesian optimization approach. AAPS Open. 2026;12:34. doi:10.1186/s41120-026-00176-0.
61. Joeres R, Blumenthal DB, Kalinina OV. Data splitting to avoid information leakage with DataSAIL. Nat Commun. 2025;16:3337. doi:10.1038/s41467-025-58606-8.
62. Kretschmer F, Seipp J, Ludwig M, Klau GW, Bocker S. Coverage bias in small molecule machine learning. Nat Commun. 2025;16:554. doi:10.1038/s41467-024-55462-w.
63. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18.
64. Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.
65. 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 and other interested parties. Silver Spring (MD): US Food and Drug Administration; 2025. Accessed July 17, 2026.
66. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023. Amsterdam: European Medicines Agency; 2024. Accessed July 17, 2026.
67. Wang N, Zhong H, Xiong P, Zhu J, Ouyang D. Quality by Computational Design: harnessing AI to advance rational drug development. Adv Drug Deliv Rev. 2026;229:115764. doi:10.1016/j.addr.2025.115764.
68. Xiong P, Zhu J, Zhong H, Wang W, Wang N, Lu T, et al. FormulationLAI: a physiology-based machine learning framework for accelerated development of long-acting injectable formulations. J Control Release. 2026;389:114418. doi:10.1016/j.jconrel.2025.114418.
69. Rezaeizadeh M, Razavi SM, Muzzio FJ. Current state of machine learning implementation in pharmaceutical process modeling for oral solid dosage forms. Int J Pharm. 2026;689:126479. doi:10.1016/j.ijpharm.2025.126479.
70. Maharjan R, Kim NA, Kim KH, Jeong SH. Transformative roles of digital twins from drug discovery to continuous manufacturing: pharmaceutical and biopharmaceutical perspectives. Int J Pharm X. 2025;10:100409. doi:10.1016/j.ijpx.2025.100409.
71. Heyndrickx W, Mervin L, Morawietz T, Sturm N, Friedrich L, Zalewski A, et al. MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without compromising proprietary information. J Chem Inf Model. 2024;64(7):2331-2344. doi:10.1021/acs.jcim.3c00799.