MACHINE LEARNING IN FORMULATION PREDICTION: EXCIPIENT SELECTION, DISSOLUTION, AND STABILITY ACROSS ORAL AND PARENTERAL DOSAGE FORMS

Pravallika Namepally[1], Keerthi Naga Deekshita[2], Krishnaveni Bommineni[3], Chandra Shekhar Reddy Bonepally[4]*

[1,2] Students, Department of Pharmaceutics, Sarojini Naidu Vanita Pharmacy Maha Vidyalaya, Affiliated to Osmania University.

[3]Assistant Professor, Department of Pharmacognosy, Sarojini Naidu Vanita Pharmacy Maha Vidyalaya, Affiliated to Osmania University.

[4]Professor and Head, Department of Pharmaceutics, Sarojini Naidu Vanita Pharmacy Maha Vidyalaya, Affiliated to Osmania University.

Corresponding author

Dr. Chandra Shekhar Reddy Bonepally

Student, Department of Pharmaceutics,

Sarojini Naidu Vanita Pharmacy Maha Vidyalaya,

Affiliated to Osmania University.

Tarnaka, Secunderabad,

Telangana -500017

Received: Aug 19, 2026 Accepted: Sep 16, 2026 Published: Sep 17, 2026

Abstract

Background. Historically, pharmaceutical formulation development relied on iterative, resource-intensive trial and error, making processes slow, costly, and dependent on informal expert knowledge. [2,3] Between 2018 and 2026, machine learning (ML) and deep learning (DL) have moved from theoretical possibilities to practical decision-support tools, learning quantitative relationships between formulation composition, process parameters, and product performance directly from experimental data. [1,2,4]

Scope. This review summarizes recent research on three interrelated prediction tasks: excipient selection and drug–excipient compatibility, dissolution and drug release, and physical and chemical stability, including shelf life, across oral solid dosage forms and parenteral products such as high-concentration biologics, lipid nanoparticles, microspheres, and polymeric long-acting injectables.

Main findings. For structured, tabular data, tree-ensemble methods (random forest, gradient boosting, XGBoost, LightGBM, extra-trees) and artificial neural networks are the most effective, regularly exceeding 80–96% predictive accuracy on preprocessed datasets.[6,12,22,23] Molecular-fingerprint and Mol2vec representations underpin compatibility classifiers, and SHAP-based interpretable modeling is increasingly common.[7,8,9] For parenterals, sequence- and structure-derived descriptors combined with molecular-dynamics features predict antibody viscosity and aggregation,[30,32,33]. At the same time, curated release-profile datasets enable prediction of microsphere and long-acting-injectable behavior. [39,41] Purpose-built tools—DE-INTERACT and FormulationDE for compatibility, ExPreSo for ranking biologic excipients, and FormulationLAI for long-acting injectables—demonstrate the translation of research models into deployable software. [7,9,11,42]

Outlook. The main bottleneck remains small, diverse, unbalanced, and often proprietary datasets. Near-term progress will be driven by generative augmentation, federated learning, hybrid mechanistic–ML modeling, and growing regulatory emphasis on model credibility expected in 2025. [19,45,46,47] While ML will not replace experimentation, it is fundamentally shifting formulation science from an experience-based discipline toward a data-driven one.

Keywords: *machine learning; pharmaceutical formulation; excipient compatibility; dissolution prediction; formulation stability; monoclonal antibodies; lipid nanoparticles; long-acting injectables; explainable AI; regulatory science.*

1 Introduction

Formulation development turns an active pharmaceutical ingredient (API) into a medicine that is safe, stable, manufacturable, and well absorbed by tuning solubility, bioavailability, release rate, and shelf life.[2] It is among the most experiment-intensive phases of drug development. Excipients can account for up to 90% of product mass, yet are usually chosen through experience-guided trial and error. [2,4] Each formulation must pass quality tests—hardness, content uniformity, disintegration, dissolution—and survive months of stability testing before a shelf life is assigned. This process is slow, costly, and captures expert intuition only indirectly, limiting its transferability and scalability. As excipient choices, formulation types, and quality-by-design (QbD) response surfaces multiply, the combinatorial space quickly outgrows what trial and error can examine.

Machine learning offers an alternative: rather than testing every combination, ML models learn statistical relationships between inputs (API physicochemical properties, excipient identities/ratios, process parameters, packaging) and measured outputs (compatibility, release kinetics, degradation, viscosity), then predict how untested formulations will behave, enabling in silico screening, optimization, and design before any batch is produced.[2] Early work used expert systems and shallow artificial neural networks (ANNs); recent work uses deep networks and ensemble learners. A pioneering 2019 study showed deep learning exceeding 80% accuracy for oral disintegrating and sustained-release matrix tablets, outperforming six conventional ML methods and marking a shift from experience-dependent to data-based formulation science.[1]

Between 2020 and 2026, AI/ML moved from a future possibility to a practical part of the formulation process, covering solubility, excipient-ratio optimization, and process control. [3,4] However, most progress so far has come from academia, start-ups, and open-source tools, with industrial uptake still in early stages.[3]

This review is organized around three central prediction tasks—excipient selection, dissolution, and stability—examined in parallel for oral and parenteral dosage forms, which present distinct data and modeling challenges. Oral solid dosage forms typically offer larger tabular datasets of composition and process information. In contrast, parenteral products—particularly biologics and nanoparticle systems—require molecular representations of proteins and lipids and are supported by smaller experimental datasets. The discussion begins with modeling foundations common to all tasks, proceeds to individual prediction problems, highlights parenteral applications and cross-cutting enablers, and concludes with regulatory considerations and outstanding challenges.

2 Foundations: Data, Descriptors, and Algorithms

Formulation-prediction models depend on three factors: the numerical representation of the formulation, the algorithm that maps this representation to the outcome, and the approach used for validation and interpretability. Model quality is critically dependent on the input representation.

2.1 Representing formulations and molecules

Inputs fall into a few categories. Molecular descriptors and fingerprints capture API and excipient chemistry—property descriptors, Extended-Connectivity Fingerprints (ECFP), PubChem 881-bit substructure fingerprints, 2D descriptors, SMARTS-encoded substructures, and learned embeddings such as Mol2vec, converting structures into fixed-length vectors.[7,8,9] Formulation and process variables include excipient ratios, polymer type/molecular weight/composition, drug load, particle size, compression force, and manufacturing method; environmental conditions—pH, ionic strength, temperature, relative humidity—are also often included. Spectral and image data (NIR spectra, surface-dissolution imaging) enable real-time or vision-based prediction. [20,21] For proteins, sequence- and structure-derived features—spatial charge maps, solvent-accessible hydrophobic surface area, local surface curvature, and MD-derived descriptors—assess monoclonal antibody developability. [30,32,33]

2.2 Algorithm families in practice

For structured, tabular formulation data, tree-ensemble methods (random forest, gradient boosting, XGBoost, LightGBM, extra-trees) and artificial neural networks are most frequently used, often yielding superior results and serving as interpretable baselines.[6,22,23]Deep learning shows advantages with larger datasets or image/sequence inputs—convolutional networks for dissolution imaging, generative models for release-profile prediction—but its benefit narrows on small datasets.[19,21] AutoML frameworks (TPOT, H2O) and optimization algorithms (e.g., Grey Wolf Optimization) increasingly automate model exploration,[13,18] and consensus/stacking models that integrate multiple learners consistently outperform single methods for compatibility and release prediction.[8]

Prediction task	Common inputs	Typical algorithms	Reported performance
Drug–excipient compatibility	PubChem/Mol2vec fingerprints, 2D descriptors, SMARTS substructures	ANN, stacking ensembles, RF/SVM + RFE feature selection	Accuracy 0.75–0.98; AUC 0.82–0.93; MCC up to 0.86
Dissolution / drug release (oral)	Excipient ratios, compression force, NIR, chemical descriptors, kinetic-model terms	RF, XGBoost, ANN/DNN, AutoML, GAN augmentation	$R^2 \approx 0.60–0.96$; NRMSE ≈ 8% (ODT); F1 up to 0.87
Small-molecule / ASD stability	API structure, ECFP, drug loading, polymer ratio, RH, temperature, packaging	RF, LightGBM, XGBoost, DNN, kinetic + ML hybrids (ASAP-type)	Accuracy 0.82–0.96; shelf-life projection from short stressed studies
Biologic stability (aggregation/viscosity)	Fv sequence, spatial charge maps, MD-derived surface curvature, hydrophobicity	kNN, logistic/linear regression, deep spatial nets, ML+MD	$r \approx 0.85–0.91$ (aggregation); acc ≈ 0.85 (viscosity class)
Nanoparticle / LAI release	Lipid/polymer identity, LA/GA ratio, N/P ratio, MW, size, loading, LogP/TPSA	Consensus ML, ANN/SVM/PLS, GPR, TabPFN, explainable ML	$R^2 \approx 0.86–0.88$; RMSE ≈ 0.10; F1 ≈ 0.87 (profile class)

Table 1. Representative input–algorithm–performance combinations across formulation-prediction tasks (2018–2026) [7,8,9,12,13,18,22,23,30,32,33,39,41]. Metrics reflect individual studies and are not directly comparable across datasets, endpoints, and validation schemes.

2.3 Validation and interpretability

Given small formulation datasets, rigorous validation—k-fold cross-validation, genuinely held-out external test sets, and, where possible, prospective experimental confirmation—should take precedence over headline accuracy. The field has also moved toward explainable AI: SHAP is now routinely applied to identify functional groups behind predicted incompatibilities, material characteristics that shape release profiles, and sequence features affecting antibody viscosity, turning predictions into mechanism-based, regulator-friendly guidance. [9,18,23]

3 Excipient Selection and Drug–Excipient Compatibility

Excipient choice affects manufacturability, release, and stability, yet incompatibilities are typically identified late via DSC, FTIR, NMR, chromatography, and extended stability studies per pair. Machine learning reframes compatibility as a pairwise classification problem, letting experimental effort focus on genuine risks.

3.1 From expert systems to learned classifiers

Early tools were knowledge-based expert systems—Ouyang and colleagues' PharmDE encoded curated rules for likely drug–excipient interactions.[10] Learned models began with DE-

INTERACT, an ANN trained on over 3,500 drug–excipient examples, each represented by concatenated 881-bit PubChem fingerprints (1,762 inputs), achieving high training/validation accuracy; its incompatibility calls—paracetamol with vanillin, paracetamol with methylparaben, and brinzolamide with polyethylene glycol—were confirmed by DSC, FTIR, HPTLC, and HPLC.[7] A subsequent stacking model combining Mol2vec embeddings with 2D descriptors achieved accuracy/precision/recall/AUC/MCC of 0.98/0.87/0.88/0.93/0.86, detecting incompatibility in ten of twelve validation cases versus three of twelve for DE-INTERACT, and was deployed as a public web application.[8] More recently, FormulationDE merged an ML classifier with PharmDE's rules over ~1,105 balanced compatibility records (579 compatible, 526 incompatible), using SMARTS-based "DE-des" descriptors, recursive feature elimination, and SHAP force plots to flag the functional groups driving each risk assessment—an interpretable, risk-flagging tool.[9] Broader reviews report ML compatibility accuracies above 90% using solubility, polarity, and hydrogen-bonding descriptors.[3,4]

3.2 Implications for oral and parenteral products

For oral solids, where excipients dominate product mass, these classifiers pre-screen fillers, binders, disintegrants, and lubricants before compatibility studies begin. For parenterals, the same logic covers buffers, tonicity agents, surfactants, and cryoprotectants—though compatibility is often inseparable from stability, judged by measured effects on aggregation, oxidation, or particle formation rather than a binary flag. For biologics, ExPreSo (Excipient Prediction Software) ranks excipients that stabilize protein drug products, accelerating formulation when drug substance is limited; because it does not model excipient combinations, top-ranked picks can be mechanistically redundant (e.g., polysorbate 80/20, sucrose/trehalose), so outputs should seed design-of-experiments (DoE) screening rather than serve as final recommendations.[11]

4 Dissolution and Drug-Release Prediction (Oral Dosage Forms)

Dissolution governs the rate at which an API becomes available for absorption and is among the most consequential, laboriously measured critical quality attributes. Two complementary ML strategies have emerged: predicting the complete release profile directly from composition, and predicting kinetic parameters of a mechanistic model (e.g., Weibull or first-order) used to reconstruct the curve. Approaches also differ depending on whether they predict a single summary value (dissolution or disintegration time) or the entire time-resolved profile.

4.1 Full release profiles and sustained-release systems

For sustained-release tablets, data-driven models built on raw-material databases and AutoML (TPOT) pipelines compared decision-tree, gradient-boosting, random-forest, extra-tree, XGBoost, and deep-learning regressors; random forest was consistently the strongest, matched only by a heavily tuned deep network.[13] Predicting the complete release profile is harder: across 377 in-house direct-compression formulations measured at eleven time points, random forest and XGBoost achieved five-fold cross-validated R^2 near 0.64 and 0.60, respectively, while also yielding release-kinetic parameters for mechanistic insight.[12] In a QbD context, polyethylene-oxide extended-release hydrophilic matrix tablets modeled via both ANN and functional-data-

analysis-plus-DoE achieved high accuracy, with the latter approach also offering interpretability alongside predictive power.[14]

4.2 Disintegration, hardness, and data-scarcity techniques

For orally disintegrating tablets (ODTs), an H2O AutoML study enriched a public database with API chemical descriptors, yielding a deep-learning model with ~8% cross-validated NRMSE and R^2 near 0.84; SHAP identified the key disintegration drivers, and an open-source ODT calculator was released.[18] Related deep-learning models reached ~99% accuracy for tablet hardness and 73% for disintegration time,[17] while optimized ensembles and Bayesian models minimized RMSE and MAPE for disintegration prediction.[16] Data-scarcity solutions include PCA-reduced networks for oral fast-disintegrating film disintegration and, notably, Wasserstein GAN (WGAN) augmentation to synthesize training data for sustained-release matrix-tablet dissolution profiles—reportedly the first GAN application in pharmaceutical formulation prediction.[19] Earlier neural-network approaches to ODT formulation prediction established the foundation for this line of work.[15]

A parallel thread moves dissolution prediction toward the manufacturing line: multivariate and ML models predict dissolution from near-infrared spectra, compression force, and particle parameters, supporting real-time release testing (RTRT);[20] computer-vision systems using convolutional networks measure tablet swelling, erosion, coating thickness, and defects, with Grad-CAM analyses confirming attention to physically meaningful image regions.[21] Collectively, these advances show ML can accelerate dissolution screening and clarify which composition variables most influence drug release.

5 Stability and Shelf-Life Prediction

Assigning shelf life conventionally requires storing batches under long-term and accelerated conditions (e.g., 40 °C/75% relative humidity for six months) with periodic testing—reliable but slow and costly. ML and hybrid kinetic–ML approaches aim to predict long-term stability much earlier, enabling selection of the best candidate. At the same time, a real-time study verifies rather than discovers the outcome.

5.1 Small molecules, solid products, and amorphous dispersions

Risk-based predictive-stability methods bridge to ML. The Accelerated Stability Assessment Program (ASAP/ASAPprime) uses short studies at elevated temperature and humidity to model degradation kinetics and estimate shelf life, following a roadmap from an industry group (IQ) now accepted by many regulators.[25,26] ML extends this by learning, from historical stability data across many products, how API structure, excipient type/amount, and packaging (bottle vs. blister) drive degradation, predicting degradation rate and shelf life under long-term storage.[5,27] Stability is not purely chemical—dissolution, disintegration, crystal form, appearance, and tablet hardness can all change during storage and should be included in shelf-life prediction.

Amorphous solid dispersions (ASDs) improve solubility for poorly water-soluble drugs but risk recrystallization; stability is traditionally confirmed over three to six months. ML shortens this: across 646 stability data points described by 20+ molecular descriptors, random forest was the best

classifier of physical stability (~82.5% accuracy), with drug-loading ratio, relative humidity, storage/preparation temperature, and polymer/API molecular weight as the top factors.[22] For hot-melt-extruded ASDs, ECFP-LightGBM predicted amorphization at 92.8% accuracy, and ECFP-XGBoost predicted chemical stability at 96.0%, with SHAP identifying API substructures consistent with known mechanisms.[23] Deep-learning methods have also been adapted to address the class imbalance typical of stability datasets.[24]

5.2 Biologics and peptides

For biopharmaceuticals, predictive analytics turns formulation and stability into a risk-reduction exercise. Advanced kinetic modeling projects long-term stability for developability assessment of therapeutic peptides and proteins, which are prone to hydrolysis, oxidation, isomerization, and deamidation in solution.[29] ML models predict monomer retention after long-term storage and help identify formulations that reduce aggregation and cold-chain dependence.[28] Physical stability—chiefly aggregation and, at high concentration, viscosity—is a key developability property increasingly modeled by combining molecular-dynamics features with sequence and biophysical data, supported by community resources such as the PIPPI database of standardized protein–excipient interaction data.[36] Vendors and academic groups alike regard these predictions as complements to, not substitutes for, traditional studies, a position regulators actively promote.

For both small molecules and biologics, stability prediction is most valuable early, narrowing candidate selection and prioritizing formulations, provided real-time studies subsequently confirm the predictions.

6 Parenteral Dosage Forms: Biologics, Nanoparticles, and Long-Acting Injectables

Parenteral products present some of the field's hardest modeling problems: molecules that must be described by sequence or structure rather than fingerprints; quality attributes—viscosity, aggregation, immunogenicity, encapsulation efficiency, burst release, and multiphasic release—with no oral analog; and datasets typically an order of magnitude smaller than for tablets. Progress here is notable and also highlights current ML limitations.

6.1 High-concentration monoclonal antibodies

Delivering antibodies subcutaneously at home requires high concentration (typically ≥150 mg/mL), which tends to increase both viscosity and aggregation—properties that are hard to assess early when material is scarce. ML performs this developability screen directly. Using twenty preclinical/clinical-stage antibodies with aggregation accelerated at 45 °C and viscosity measured at 150 mg/mL, plus features from full-length molecular-dynamics simulations, a k-nearest-neighbors model using two features—a spatial positive-charge map on CDRH2 and hydrophobic solvent-accessible surface area on the variable fragment—predicted aggregation rate at a correlation near 0.89; a two-feature logistic-regression classifier reached ~0.85 accuracy for the viscosity class.[30] Since then, interpretable, sequence-only models predict low-viscosity IgG1 variants from the variable-region sequence alone at a standard formulation pH, removing the need for costly early experiments;[31]deep-learning surrogates such as DeepSP derive spatial charge and hydrophobicity descriptors directly from sequence, avoiding time-consuming MD simulations;[32]

a coupled AI–MD platform using local molecular-surface curvature and hydrophobicity predicted antibody aggregation at r = 0.91, surpassing the prior state of the art;[33] and all-atom MD of an antibody Fab fragment across 49 formulation environments (varying temperature, pH, and salt) enabled ML to predict aggregation kinetics as a function of formulation conditions—an important advance toward condition-aware stability modeling.[34] A 2026 review frames molecular modeling and ML for high-concentration viscosity as a coherent, fast-moving subfield with defined challenges and future directions.[35]

6.2 Lipid nanoparticles and nucleic-acid delivery

Lipid nanoparticles (LNPs), first developed in the 1960s for drug delivery and later adapted for nucleic acids, comprise ionizable lipids, helper lipids, cholesterol, and PEG-lipids. Ionizable lipids gain or lose charge depending on pH, enabling stable circulation and intracellular cargo release; helper lipids and cholesterol support structure and membrane fusion; PEG-lipids extend circulation and limit rapid immune clearance. LNPs enabled the Pfizer-BioNTech and Moderna COVID-19 mRNA vaccines and are now being tested for DNA and CRISPR-component delivery for genetic disease, though targeting precision and immunogenicity remain challenges. ML models trained on public data predicted LNP compositions for mRNA vaccines, with mouse experiments confirming the model's prediction that an MC3-based formulation at a 6:1 N/P ratio outperformed alternatives.[37] For siRNA ionizable LNPs, QSAR-type models combining ionizable-lipid molecular descriptors with formulation characteristics—using ANN, SVM, and PLS with evolutionary feature selection—predicted in vivo efficacy at external-validation R^2 of about 0.75–0.89, with ANN performing best.[38] Related work predicts PLGA nanoparticle diameter and zeta potential from synthesis parameters (polymer type/concentration, anti-solvent conditions), including Gaussian process regression within the same design-optimization loop.[43]

6.3 Microspheres and long-acting injectables

Polymeric long-acting injectables (LAIs) and microspheres reduce dosing frequency but involve complex formulation–manufacturing–release interactions, and empirical optimization has historically taken eight to ten years. A consensus model built from 286 small-molecule microsphere formulations predicted dissolution behavior at both 37 °C and 45 °C with validation R^2 near 0.88, accelerating microsphere development.[39] Nature Communications work showed ML models predicting experimental drug release from polymeric LAIs and aiding new-formulation design.[40] Building on this, a 2026 explainable, time-independent framework examined 321 LAI formulations to predict early release (24/48/72 h), classify release-profile type (~0.87 accuracy), and predict full release profiles—outperforming time-dependent methods on delayed biphasic and triphasic curves, with SHAP revealing which material characteristics control each release phase.[41] Physiology-informed models are also emerging: FormulationLAI combined ML (using TabPFN) with mechanistic insight from molecular dynamics to predict cumulative release for in-situ forming gels (RMSE ≈ 0.105, R^2 ≈ 0.865), experimentally verified in rats with markedly higher in vivo exposure than a commercial suspension reference.[42] These studies also show ML's limits here: sample sizes are often too small for deep networks to avoid overfitting, favoring careful feature design and interpretable models over raw architectural complexity.

7 Cross-Cutting Enablers

- **Curated databases.** PIPPI (protein–excipient interactions), raw-material and release databases for tablets and LAIs, and open datasets from recent antibody studies underpin training and cross-group benchmarking.[36]
- **Explainability.** SHAP and feature-importance analyses convert black-box predictions into mechanistic hypotheses—e.g., the roles of drug loading, relative humidity, and temperature in ASD stability—strengthening scientific trust and regulatory defensibility. [9,18,23]
- **Mechanistic–ML hybrids.** Combining ML with molecular dynamics, Flory–Huggins/solubility-parameter theory, or kinetic release models yields predictions that are both accurate and interpretable, compensating for limited data. [33,42]
- **Deployment infrastructure.** AutoML pipelines simplify model-building, while web deployments—compatibility predictors, excipient rankers, ODT and LAI calculators—move models from publication to practical use. [7,9,11,18]

8 The Regulatory Environment and Model Credibility

As predictive systems become involved in decisions affecting product quality, regulators have moved to define how to demonstrate their credibility. Drug applications containing AI/ML elements reportedly rose from one in 2016 to 132 in 2021, prompting the FDA's Center for Drug Evaluation and Research to launch the Framework for Regulatory Advanced Manufacturing Evaluation (FRAME) and issue a 2023 discussion paper on AI in drug manufacturing.[44]

In January 2025, the U.S. FDA released its first draft guidance on AI to support regulatory decisions about drugs and biological products, proposing a risk-based, seven-step credibility framework built on the model's context of use (COU)—its specific role and scope in answering a question of interest. The steps move from defining the question of interest and the COU, through assessing model risk, to planning, executing, and documenting credibility evidence proportionate to that risk. The guidance spans nonclinical, clinical, post-marketing, and manufacturing stages—explicitly covering formulation- and product-quality-relevant uses—while excluding pure drug-discovery or operational applications that do not affect patient safety, product quality, or study reliability. The FDA has received hundreds of AI-containing submissions and continues building internal capacity, including an AI Benchbook and staff training announced in late 2025.[45]

Alongside Good Machine Learning Practice (GMLP) principles, Predetermined Change Control Plans (PCCPs), and the European Medicines Agency's September 2024 reflection paper—which adopts a broader lifecycle and good-practice perspective—these developments signal a shift toward lifecycle oversight emphasizing representative data, transparency, bias control, and continuous performance monitoring rather than one-time approval.[46] For formulation scientists, a model's required rigor scales with the consequences of its failure; interpretable, well-validated models with documented data provenance are easier to defend than opaque alternatives. Provenance, data quality, and explainability are now integral parts of the deliverable, not afterthoughts.

9 Challenges, Limitations, and Future Directions

9.1 The persistent data bottleneck

The dominant limitation across every task above is data: small, diverse, unevenly distributed, and frequently proprietary. [2,3] Public datasets number in the hundreds, whereas deep learning would need tens of thousands, which is why classical ML and feature engineering often outperform deep learning here. Recurring problems include: limited and low-quality data, since models trained on a few hundred records will not generalize beyond their chemical or process space; class imbalance, as compatibility and stability datasets frequently skew toward "stable" or "compatible" outcomes;[24] weak in vitro–in vivo correlation, particularly for LAIs, which hinders direct clinical extrapolation;[41] and co-formulation blindness, since pairwise models cannot detect interactions involving more than one excipient, so their outputs are best used to seed designed experiments rather than stand alone.[11]

9.2 Emerging responses

Generative augmentation (GANs, e.g., WGAN for release profiles) synthesizes plausible training data to offset scarcity, with reports of faster and more accurate model development.[19]Federated learning enables collaborative model training across organizations without sharing raw proprietary data, expanding effective dataset size while preserving confidentiality.[3]Foundation and language models—tabular foundation models for molecular-property prediction and LLM-based approaches, for example, for 3D-printable formulation generation—point toward more general, transferable formulation intelligence.[47]

9.3 Trust, transparency, and generalization

Generalizability is fragile: models trained on commercial, stability-optimized molecules may not transfer to early-stage candidates, and performance on an internal test set can overstate real-world reliability. Interpretability is now expected rather than optional—SHAP and related tools are becoming standard because a mechanistic explanation turns a prediction into a formulation decision and satisfies regulatory scrutiny; opaque models sit uneasily with GMP principles of control, reproducibility, and traceability.[35,45] Integration into workflows—linking prediction to automated, self-driving formulation laboratories and real-time manufacturing control—remains the frontier where academic models must prove their industrial worth.

9.4 Future perspectives

Condition-aware, mechanism-informed models that incorporate formulation-environment variables (pH, ionic strength, temperature, humidity) alongside molecular dynamics and physicochemical theory should improve generalization beyond tested conditions.[34] Data-sharing and standardization initiatives—federated learning, FAIR datasets, shared benchmarks—can address the small-data challenge without compromising proprietary information.[36,47] Generative and active-learning approaches, in which models propose and iteratively refine candidate formulations, may bridge the gap between prediction and experimentation. Closer alignment with the FDA credibility framework and QbD principles will embed ML as an auditable, explainable part of the regulatory dossier. [45,46] Predictive modeling will not replace confirmatory

experimentation—stability and dissolution testing remain essential—but it is increasingly used to prioritize which experiments to run; augmentation rather than replacement is the prevailing theme.

10 Conclusion

Across excipient selection, dissolution, and stability, and across both oral and parenteral dosage forms, machine learning has progressed from illustrative case studies to genuinely useful decision support. Fingerprint- and embedding-based classifiers now flag drug–excipient incompatibilities with high accuracy and mechanistic explanation;[7,8,9] tree ensembles and neural networks predict dissolution and release profiles well enough to prioritize formulations and inform release kinetics;[12,13,18] hybrid kinetic–ML methods shorten the path to a defensible shelf life;[25,26] and for harder parenteral problems, sequence- and structure-aware models forecast antibody viscosity and aggregation, LNP efficacy, and long-acting-injectable release from modest datasets.[30,33,37,41] Several models have been packaged into deployable tools, including DE-INTERACT, Formulation DE, ExPreSo, Formulation LAI, and open ODT calculators.[7,9,11,18,42]

Key enablers include robust molecular and formulation representations, rigorous validation, and model explainability, while the primary constraint remains data availability and quality. Large-scale impact requires improved data resources, transparent management of class imbalance and generalization limits, strong in vitro–in vivo correlation, and regulatory-grade explainability. As shared datasets expand, generative and federated approaches advance, and the FDA's credibility framework establishes clear expectations, machine learning is poised to become a standard, reliable tool for formulation scientists—marking a decisive shift from experience-based practice to data-driven engineering, with ML augmenting rather than replacing experimental expertise.

References

1. Yang Y, Ye Z, Su Y, et al. Deep learning for in vitro prediction of pharmaceutical formulations. Acta Pharmaceutica Sinica B. 2019;9(1):177–185.
2. Bannigan P, Aldeghi M, Bao Z, et al. Machine learning-directed drug formulation development. Advanced Drug Delivery Reviews. 2021; 175:113806.
3. Role of Artificial Intelligence and Machine Learning in Pharmaceutical Formulation: Recent Advances, Existing Challenges and Future Prospectus. Biosciences Biotechnology Research Asia. 2026.
4. Artificial Intelligence (AI) in Pharmaceutical Formulation and Dosage Calculations. Pharmaceutics. 2025;17(11):1440.
5. Challener CA. Predictive Modeling for Small-Molecule Formulation Development Using Advanced Algorithms. Pharmaceutical Technology. 2025;49(8):14–17.
6. Wang S, et al. State-of-the-Art Review of Artificial Neural Networks to Predict, Characterize and Optimize Pharmaceutical Formulation. Pharmaceutics. 2022; 14:183.
7. Sarabu S, et al. DE-INTERACT: a machine-learning-based predictive tool for the drug–excipient interaction study — validation through paracetamol and vanillin. International Journal of Pharmaceutics. 2023.

8. Vijayan V, et al. Towards safer and efficient formulations: machine learning approaches to predict drug–excipient compatibility. International Journal of Pharmaceutics. 2024; 654:123884.
9. FormulationDE: an updated artificial intelligence system for drug–excipient compatibility prediction. AAPS Open. 2026;12: article 156.
10. Wang Z, et al. PharmDE: knowledge-driven expert system for drug–excipient compatibility. 2021.
11. Machine-learning-driven acceleration of biopharmaceutical formulation development using Excipient Prediction Software (ExPreSo). Computational and Structural Biotechnology Journal. 2025; 27:4517–4525.
12. Protopapa C, Siamidi A, Eneli AA, Elbadawi M, Vlachou M. Machine Learning Predicts Drug Release Profiles and Kinetic Parameters Based on Tablets' Formulations. The AAPS Journal. 2025;27: article 101.
13. Bharathi M, Kamaraj R, Murugaanandam S, et al. A data-driven approach to predict the in vitro dissolution time of sustained-release tablets using raw-material databases and machine learning algorithms. Pharmacia. 2024; 71:1–7.
14. A comparative study of two data-driven modeling approaches to predict drug release from extended-release matrix tablets (PEO hydrophilic matrix; ANN vs functional DoE). International Journal of Pharmaceutics. 2025;671.
15. Han R, et al. Predicting oral disintegrating tablet formulations by neural network techniques. Asian Journal of Pharmaceutical Sciences.
16. Prediction of tablet disintegration time based on formulation properties via artificial intelligence: comparing machine-learning models and validation. Scientific Reports. 2025; 15:98783.
17. A prediction model based on artificial-intelligence techniques for disintegration time and hardness of fast-disintegrating tablets in pre-formulation tests. 2024.
18. Puzzle out Machine Learning Model — Explaining the Disintegration Process in ODTs (H2O AutoML; SHAP; ODT calculator). Pharmaceutics.
19. Drug Properties Prediction Based on Deep Learning (OFDF disintegration; WGAN augmentation for SRMT dissolution). Pharmaceutics. 2022;14(2):467.
20. Real-time prediction of dissolution profiles of coated oral dosage forms. International Journal of Pharmaceutics. 2024.
21. Explainable AI in Pharmaceutics: Grad-CAM Analysis of Surface Dissolution Imaging Using CNNs. Pharmaceutics. 2026;18(4):481.
22. Han R, Yang Y, Li X, Ouyang D. Predicting physical stability of solid dispersions by machine-learning techniques. International Journal of Pharmaceutics. 2019.
23. The applications of machine learning to predict the formation of chemically stable amorphous solid dispersions prepared by hot-melt extrusion. International Journal of Pharmaceutics. 2023.
24. Lee S, et al. Deep-learning-based prediction of physical stability considering class imbalance for amorphous solid dispersions. Journal of Chemistry. 2022; 2022:4148443.
25. Accelerated Stability Testing Using Statistical Modeling and AI. International Journal of Multidisciplinary Sciences and Management.
26. Assessing Drug Product Shelf Life Using the Accelerated Stability Assessment Program (ASAPprime): GLPG4399 case study. Pharmaceutics.

27. How Machine Learning is Revolutionizing Generic Drug Development. DrugPatentWatch. 2025.
28. Predictive Analytics for Biopharmaceutical Shelf-Life Extension. Leukocare. 2025.
29. Long-Term Stability Prediction for Developability Assessment of Biopharmaceutics Using Advanced Kinetic Modeling.
30. Lai P-K, et al. Machine learning prediction of antibody aggregation and viscosity for high-concentration formulation development of protein therapeutics. mAbs. 2022;14(1):2026208.
31. Reduction of monoclonal antibody viscosity using interpretable machine learning. mAbs. 2024; 16:2303781.
32. DeepSP: Deep learning-based spatial properties to predict monoclonal antibody stability.
33. Prediction of aggregation in monoclonal antibodies from molecular-surface curvature. Scientific Reports. 2025;15.
34. Predictive model building for aggregation kinetics based on molecular-dynamics simulations of an antibody fragment across 49 formulation environments. 2024.
35. Molecular Modeling and machine learning for predicting high-concentration antibody viscosity. Advanced Drug Delivery Reviews. 2026.
36. Gentiluomo L, et al. Design of biopharmaceutical formulations accelerated by machine learning; PIPPI database. Molecular Pharmaceutics. 2021.
37. Wang W, et al. Prediction of lipid nanoparticles for mRNA vaccines by the machine learning algorithm. Acta Pharmaceutica Sinica B. 2022.
38. Metwally AA, Nayel AA, Hathout RM. In silico prediction of siRNA ionizable-lipid nanoparticles in vivo efficacy: ML modeling. Frontiers in Molecular Biosciences. 2022; 9:1042720.
39. Deng J, et al. Machine learning in accelerating microsphere formulation development. Drug Delivery and Translational Research. 2023;13(4):966–982.
40. Bannigan P, et al. Machine-learning models to accelerate the design of polymeric long-acting injectables. Nature Communications. 2022/2023;13–14.
41. Robles KN, Samad MD. Predicting Early and Complete Drug Release from Long-Acting Injectables Using Explainable Machine Learning. arXiv:2601.02265. 2026.
42. Formulation LAI: a physiology-based machine-learning framework for accelerated development of long-acting injectable formulations. Journal of Controlled Release. 2025.
43. Predicting PLGA nanoparticle size and zeta potential in synthesis for drug delivery via machine-learning analysis. Scientific Reports. 2024; 14:1114.
44. Regulatory perspectives for AI/ML implementation in pharmaceutical GMP environments (FRAME program). Pharmaceutics. 2025.
45. U.S. Food and Drug Administration. Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products (Draft Guidance). January 2025.
46. European Medicines Agency. Reflection paper on the use of artificial intelligence in the medicinal product lifecycle. September 2024.
47. Tabular foundation models for in-context prediction of molecular properties. arXiv. 2026.