RESEARCH PAPER

ARTIFICIAL INTELLIGENCE IN PHARMACEUTICAL FORMULATION DEVELOPMENT: FROM TRIAL-AND-ERROR TO PREDICTIVE AND AUTONOMOUS FORMULATION DESIGN

YUSUF RAZA[1*], SHEFALI ARORA[2], PRIYA ROY[3], MD. RABBAN ALAM[4], DR. MANISH KUMAR[5,] SIDDHI DUBEY[6]

[1]Assistant Professor, Mamta Institute of Education, Chintamanpur, Duraundha, Siwan, Bihar. Affiliated to Bihar University of Health Sciences, Patna, Bihar. E-mail Id: Yusufraza787@gmail.com, Orcid: 0009-0005-3902-3730
[2]Assistant Professor, Ambedkar Institute of Higher Education, Chhitnawa, Danapur, Patna, Bihar. Affiliated to Bihar University of Health Sciences, Patna, Bihar. E-mail Id: shefaliarora19981@gmail.com , Orcid: 0009-0001-8150-4248
[3]Assistant Professor, Ambedkar Institute of Higher Education, Chhitnawa, Danapur, Patna, Bihar. Affiliated to Bihar University of Health Sciences, Patna, Bihar. E-mail Id: priyaroy6937@gmail.com Orcid id- 0009-0008-0925-3772
[4]Assistant Professor, Ambedkar Institute of Higher Education, Chhitnawa, Danapur, Patna, Bihar. Affiliated to Bihar University of Health Sciences, Patna, Bihar. E-mail Id: rabban551@gmail.com Orcid:- 0009-0003-2842-4059
[5]Assistant Professor, Ambedkar Institute of Higher Education, Chhitnawa, Danapur, Patna, Bihar. Affiliated to Bihar University of Health Sciences, Patna, Bihar. E-mail Id: krmanishkr4@gmail.com orcid id :- 0009-0001-1601-8091
[6]Sagar Institute of Pharmacy and Technology (SIPTec), Bhopal, Madhya Pradesh. E-mail: dubeysiddhi812@gmail.com Orcid:- 0009-0003-0216-1772

***Corresponding Author**
Email: Yusufraza787@gmail.com

Received: 5th October, 2026; **Revised:** 10th October, 2026; **Accepted:** 13th October, 2026; **Available Online:** 27th October, 2026

ABSTRACT

Pharmaceutical formulation development has long relied on sequential experimentation in which formulation scientists select active pharmaceutical ingredients (APIs), excipients, compositions, and process parameters, manufacture prototypes, characterize critical quality attributes (CQAs), and iteratively modify unsuccessful formulations. While this approach remains valuable, it can become less efficient when formulation spaces expand and relationships among material attributes, process parameters, and product performance become nonlinear. Artificial intelligence (AI), machine learning (ML), deep learning, Bayesian optimization, natural language processing, and related computational approaches are enabling a transition toward predictive formulation development. These methods can bring together formulation composition, API descriptors, excipient characteristics, process variables, and analytical outcomes to predict CQAs such as dissolution, particle size, viscosity, stability, drug release, encapsulation efficiency, and manufacturability. Recent work has demonstrated ML-assisted excipient selection, prediction of tablet drug-release profiles, AI-assisted liquid formulation development, and Bayesian optimization for formulation experimentation. A promising next step is inverse formulation design, in which the desired product profile is specified first and algorithms identify candidate compositions and process conditions that are most likely to achieve it. Integration with Quality by Design (QbD), Design of Experiments (DoE), Process Analytical Technology (PAT), automated experimentation, digital twins, and robotics could ultimately create closed-loop or autonomous formulation laboratories. However, model reliability is limited by small and biased datasets, inconsistent experimental metadata, lack of external validation, explainability challenges, domain shift, data leakage, and uncertainty concerning regulatory qualification. This review discusses the evolution from empirical formulation development to predictive and autonomous workflows, major AI methods, applications across dosage forms, implementation with QbD and manufacturing, validation and regulatory considerations, research gaps, and future perspectives. The central argument of this review is that AI should support—not replace—formulation science and scientific judgement by reducing unnecessary experiments, prioritizing informative experiments, and accelerating translation of formulation knowledge into robust drug products.

Keywords

Artificial intelligence; machine learning; pharmaceutical formulation; predictive formulation design; formulation optimization; excipient selection; Bayesian optimisation; Quality by Design; drug release; pharmaceutical manufacturing; digital twins; autonomous formulation development

How to cite this article: Raza Y, Arora S, Roy P, Alam MR, Kumar M, Dubey S., Artificial Intelligence in Pharmaceutical Formulation Development: From Trial-and-Error to Predictive and Autonomous Formulation Design. Int J Drug Deliv Technol. 2026;16(88s): 1579-1591; DOI: 10.25258/ijddt.16.88s.156

Source of support: Nil.
Conflict of interest: None.

INTRODUCTION

Pharmaceutical formulation development is the scientific process of converting an active pharmaceutical ingredient into a stable, manufacturable, acceptable, and clinically useful dosage form. The formulation determines how an API is protected, processed, released, absorbed, and ultimately delivered to the intended biological site. Conventional development therefore combines physicochemical characterization, excipient screening, formulation experimentation, process development, analytical testing, stability studies, and scale-up. This work draws on several disciplines, involving pharmaceutics, physical pharmacy, analytical science, materials science, process engineering, biopharmaceutics, and regulatory science.

Even with major advances in formulation science, many development programs still rely on an iterative empirical workflow. Scientists propose a formulation, prepare it, test it, interpret the results, and then modify one or more variables. This approach is not inherently inefficient when the design space is small and the underlying relationships are well understood. However, modern drug products increasingly involve poorly soluble APIs, complex excipient interactions, amorphous systems, lipid-based carriers, biologics, controlled-release platforms, nanomedicines, personalized dosage forms, and advanced manufacturing technologies. These systems create multidimensional design spaces in which simple one-factor-at-a-time experimentation can miss interactions and require large numbers of experiments.



AI offers a complementary way to approach these challenges. Rather than treating every formulation experiment as an isolated observation, AI/ML methods can learn patterns from historical and newly generated data. Recent reviews describe AI applications across formulation optimization, manufacturing, quality control, and pharmaceutical development. A 2024 European Journal of Pharmaceutical Sciences review specifically describes AI use for predicting excipient effects, sustained-release behavior, and physical stability, while a 2025 review emphasizes the emerging data- and knowledge-driven paradigm for formulation development. cite turn0search0 turn0search5

The aim is not to replace formulation scientists with algorithms. A more practical approach is collaboration between scientists and AI: scientists define the target product profile and constraints, AI learns from available evidence, predicts promising regions of the design space, and recommends experiments; experimental results are then used to update the model. This creates a learning cycle in which every well-designed experiment has greater informational value.

2. Scope and Review Approach

This article is structured as a narrative review focused on the application of AI and ML to pharmaceutical formulation development. Recent literature from approximately 2024–2026 was prioritized, while foundational concepts and representative earlier work are included where needed to explain the evolution of the field. Searches were oriented toward combinations of terms including artificial intelligence, machine learning, pharmaceutical formulation, formulation optimization, excipient selection, dissolution prediction, Bayesian optimization, Quality by Design, drug release, stability prediction, automated formulation, and digital twins. Recent peer-reviewed reviews and primary research were emphasized. The article does not claim to be a PRISMA systematic review; a formal systematic review would require a predefined protocol, exhaustive database searching, independent screening, risk-of-bias assessment, and reproducible study selection.

3. From Trial-and-Error to Predictive Formulation Development

The traditional formulation cycle can be summarized as: define a preliminary composition → prepare a prototype → characterize the product → identify failure or performance gaps → modify the formulation → repeat. This approach benefits from direct laboratory observation and expert intuition, but it can become costly when many variables influence the outcome.

AI changes the sequence by introducing a predictive layer before laboratory execution. Historical formulation data can be used to estimate the probability that a candidate formulation will meet predefined CQAs. The scientist can then prioritize candidates with high predicted performance or high information value. The resulting workflow is not simply faster experimentation; it is a different philosophy in which experimental design is informed by learned relationships.

The conceptual progression is: (1) empirical trial-and-error, (2) statistically designed experimentation, (3) ML-assisted prediction, (4) inverse formulation design, (5) active-learning formulation development, and potentially (6) autonomous closed-loop experimentation. Each stage builds on the previous one rather than eliminating it.

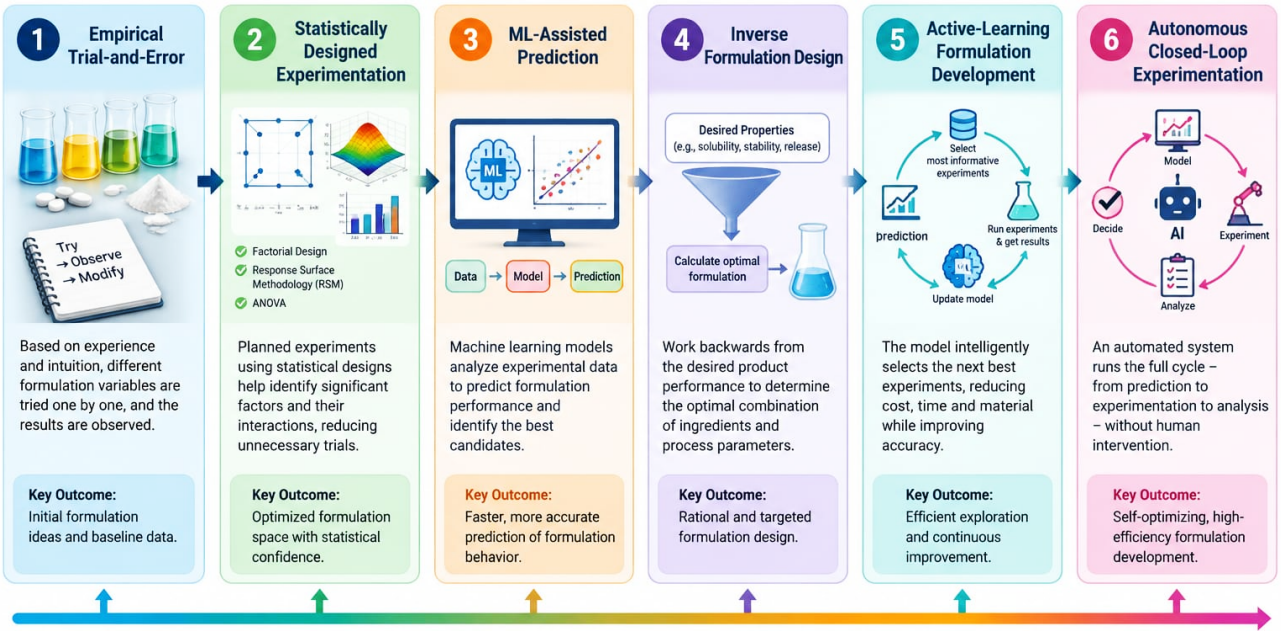


Figure 1. AI-enabled formulation development cycle

4. Why Conventional Formulation Development Is Challenging

4.1 High-dimensional design spaces

A formulation may contain several excipients at different concentrations and may be manufactured using multiple process parameters. The total number of possible combinations increases rapidly as variables are added.

4.2 Nonlinear interactions

API properties, excipient functionality, concentration, mixing, compression, temperature, humidity, and processing history may interact. The effect of one variable may depend on another variable.

4.3 Limited material availability

Early development often has limited API, especially for potent molecules, expensive biologics, or novel modalities. This limits the number of physical experiments that can be performed.

4.4 Long experimental cycles

Some CQAs, especially stability-related outcomes, require long observation periods. Formulation decisions therefore may need to be made before complete long-term data are available.

4.5 Scale-up uncertainty

A formulation that performs well at laboratory scale may behave differently during pilot or commercial manufacturing because of changes in mixing, shear, heat transfer, residence time, compression, or equipment geometry.

4.6 Knowledge fragmentation

Formulation information is distributed across publications, laboratory notebooks, batch records, reports, regulatory documents, and databases. AI can potentially connect these heterogeneous information sources, provided the data are structured and trustworthy.

5. Fundamentals of AI and Machine Learning for Pharmaceutics

AI is a broad term covering computational systems capable of performing tasks that normally require human intelligence. ML is a major AI approach in which algorithms learn relationships from data rather than being programmed with every relationship explicitly. In formulation development, ML can map inputs such as API descriptors, excipient identity, concentrations, process conditions, and environmental variables to outputs such as dissolution, particle size, viscosity, assay, hardness, friability, stability, or encapsulation efficiency.

AI approach	Typical formulation use	Strength	Key limitation
Random forest / gradient boosting	CQA prediction; excipient screening	Handles nonlinear relationships and mixed features	May be difficult to interpret without supporting tools
Support vector machine	Classification or regression of formulation outcomes	Can work with modest-sized datasets	Sensitive to feature scaling and kernel selection
Neural network / deep learning	Complex formulation–property relationships	Flexible representation of nonlinear patterns	May overfit small datasets; needs careful validation
Gaussian process	Uncertainty-aware prediction; Bayesian optimization	Useful for small-data sequential experimentation	Computational burden may rise with dataset size
NLP / large language models	Literature mining; structured knowledge extraction	Can process large text collections	Extracted details require verification
Active learning / inverse design	Next-experiment selection; candidate generation	Can reduce low-value experiments	Depends on trustworthy data and uncertainty estimates

Table 1. Representative AI methods and their potential roles in formulation development.

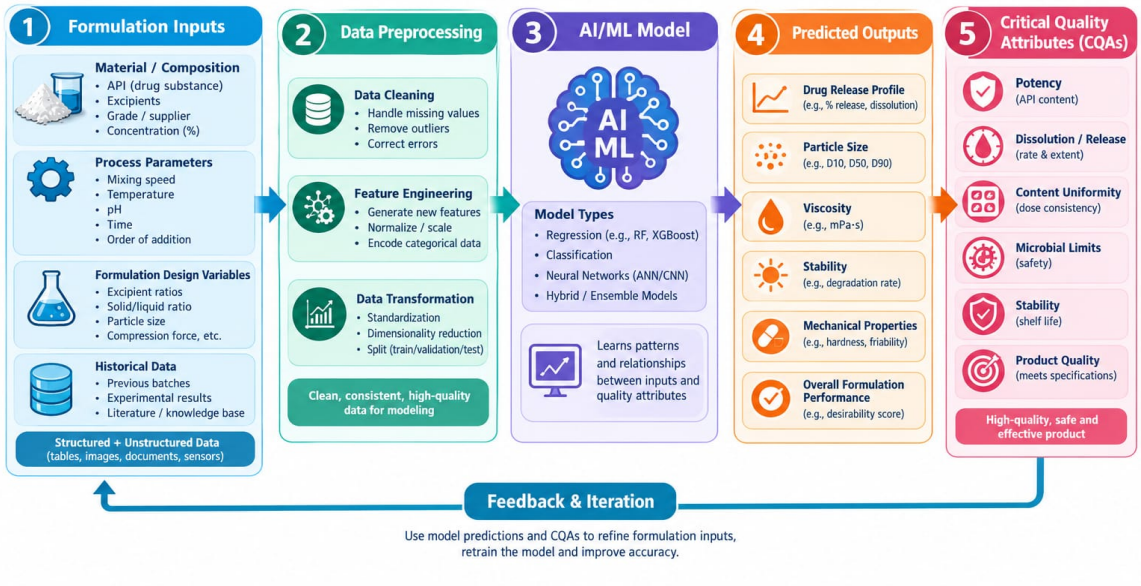


Figure 2. attributes.

5.1 Supervised learning

Supervised learning uses labeled observations in which input variables are paired with known outcomes. Regression models predict continuous variables, whereas

classification models predict categories or probabilities. Random forests, gradient-boosting methods, support vector machines, Gaussian processes, and neural networks are examples relevant to formulation datasets.

5.2 Unsupervised learning

Unsupervised methods identify patterns without a predefined outcome. Clustering and dimensionality-reduction methods can reveal formulation families, detect unusual batches, group excipients according to functional characteristics, or visualize high-dimensional formulation spaces.

5.3 Deep learning

Deep neural networks can represent highly nonlinear relationships and may be useful when large datasets or complex molecular representations are available. More complex deep-learning models are not necessarily better than simpler alternatives. For many pharmaceutical datasets, sample size may be too small to justify highly parameterized architectures.

5.4 Natural language processing

NLP can extract formulation information from scientific publications, patents, reports, and other text. Potential uses include identifying excipients, dosage forms, processing conditions, formulation outcomes, and historical formulation knowledge. Such systems can help build structured datasets, but automated extraction must be validated because pharmaceutical terminology is context-dependent.

5.5 Generative and inverse design

Generative approaches aim to propose new candidates rather than only predict existing observations. In formulation science, an inverse design model can begin with desired CQAs and constraints and search for compositions likely to achieve them. This is particularly attractive when the objective is multi-dimensional, such as maximizing dissolution while minimizing excipient quantity and maintaining manufacturability.

6. AI for Preformulation and API Characterization

Preformulation establishes the physicochemical foundation for dosage-form design. Relevant variables include solubility, pKa, lipophilicity, melting behavior, crystallinity, particle size, hygroscopicity, polymorphism, permeability, and chemical stability. AI can integrate these descriptors to predict formulation-relevant behavior before extensive experimentation.

For poorly soluble APIs, models can estimate solubility or identify formulation strategies such as salt formation, amorphization, particle-size reduction, lipid-based systems, or precipitation inhibitors. For solid-state development, ML may assist in classifying materials according to properties that influence flow, compaction, dissolution, or stability.

At this stage, AI is especially useful for prioritizing options. Instead of testing every possible strategy, the model can rank approaches according to predicted probability of success, allowing formulation scientists to focus experimental resources on the most promising candidates.

7. AI-Based Excipient Selection

Excipient selection is one of the most important opportunities for AI because the number of possible excipients and combinations is large, while experimental screening can be laborious. Excipients influence solubility, wetting, viscosity, flow, compression, release, stability, palatability, preservation, and manufacturability.

A notable 2025 study introduced Excipient Prediction Software (ExPreSo), a supervised ML approach trained using regulatory-approved peptide and protein drug products. The system used protein and product characteristics to predict likely excipient choices and reported good performance for commonly used excipients, illustrating how historical product knowledge can be converted into a formulation-support tool.

For small-molecule products, future systems could combine molecular descriptors of the API with functional descriptors of excipients, prior formulation outcomes, route of administration, target product profile, and regulatory constraints. The model should ideally return not just a ranked list but also uncertainty, rationale, compatibility risks, and experimental recommendations.

8. Predictive Formulation Design

Predictive formulation design uses computational models to estimate product performance from composition and process inputs. The basic architecture is a mapping from material attributes and process parameters to CQAs. For example, tablet formulation inputs may include API load, filler type, binder concentration, disintegrant concentration, lubricant concentration, compression force, and tablet hardness. Outputs may include disintegration time, tensile strength, friability, and dissolution.

Useful predictive models must perform reliably on formulations beyond those used for training rather than simply memorizing the training dataset. Therefore, model validation should use appropriate train/test separation, cross-validation, external datasets where possible, and domain-specific evaluation. Random splits can be misleading when formulations are highly related; grouped or leave-one-formulation-family-out validation may better represent prospective use.

9. Critical Quality Attributes and Multi-Objective Optimization

Formulation development rarely involves a single objective. A formulation may need rapid dissolution, adequate mechanical strength, acceptable stability, low variability, acceptable viscosity, minimal excipient load, and manufacturability. Optimizing one attribute may worsen another.

Critical quality attribute (CQA)	Example measurement	Potential formulation/process drivers	Possible AI task
Dissolution / drug release	Dissolution profile or kinetic parameters	API solubility, particle size, disintegrant, polymer level, compression	Predict profile; optimize release
Mechanical performance	Hardness, tensile strength, friability	Binder, lubricant, granulation, compression force	Predict strength and failure risk
Particle properties	Size, polydispersity, morphology	Mixing, solvent, stabilizer, lipid/polymer composition	Predict particle attributes
Physical / chemical stability	Assay, degradants, phase change, precipitation	pH, excipients, packaging, temperature, humidity	Rank risk; forecast trends
Rheology / viscosity	Viscosity or flow curve	Polymer concentration, pH, ionic strength, temperature	Predict handling and application properties
Manufacturability	Flow, compressibility, process variability	Material properties, equipment, process settings	Detect risk and support scale-up

Table 2. Examples linking critical quality attributes to formulation variables and AI tasks.

Multi-objective AI can represent these competing requirements using desirability functions, Pareto optimization, constrained optimization, or Bayesian methods. The result may be a group of promising candidate formulations rather than one mathematically optimal solution representing different trade-offs. Scientists can then select a solution according to clinical, manufacturing, cost, and regulatory priorities.

10. Integration of Design of Experiments and Machine Learning

DoE remains highly relevant in the AI era. Statistical DoE provides a structured method for selecting experiments and estimating main effects and interactions. ML can complement DoE when relationships are nonlinear or when the objective is prediction rather than inference alone.

A practical workflow is to use DoE to generate an informative initial dataset, train several candidate models, compare predictive performance, identify influential variables, and then use the model to propose a second experimental round. This hybrid approach avoids the false choice between classical statistics and AI.

Stage	Key activities	Expected output/decision gate
1. Define target	Specify target product profile, route, dose, CQAs, and constraints	Approved target criteria
2. Assemble data	Standardize API, excipient grade, composition, process, methods, and batch metadata	Traceable curated dataset
3. Generate initial data	Use DoE or scientifically justified experiments to cover the design space	Baseline dataset
4. Train and validate	Compare candidate models; separate related batches appropriately; assess uncertainty	Validated model within a stated applicability domain
5. Select next experiments	Use optimization or active learning to balance performance and information gain	Ranked experiment plan
6. Confirm experimentally	Prepare candidates and compare predicted versus measured CQAs	Evidence for or against model generalization
7. Transfer and monitor	Evaluate scale-up, monitor drift, and manage model versions	Controlled lifecycle strategy

Table 3. Practical framework for implementing AI-enabled formulation development.

Effective formulation programs can use DoE to generate structured evidence and ML to model nonlinear patterns and support optimization.

11. Bayesian Optimization and Active Learning

Bayesian optimization is particularly attractive when experiments are expensive. A surrogate model approximates the relationship between formulation variables and the desired outcome, while an acquisition function selects the next experiment by balancing exploitation of promising regions with exploration of uncertain regions.

A 2025 PLOS ONE study on vaccine formulation development illustrates this iterative strategy: a Gaussian-process model suggested experiments, new experimental

results were added to the dataset, and the model was refined repeatedly. Importance analysis and SHAP interpretation were also used to understand model inputs.

Active learning extends this concept by selecting the next experiment because it is expected to provide the greatest information or improve the objective. This can dramatically reduce the number of experiments required when each formulation run is costly.

12. AI in Solid Oral Dosage Forms

Tablets and capsules provide a mature environment for AI applications because formulation composition, process parameters, and performance measurements can often be represented numerically. ML can predict powder flow, tabletability, mechanical strength, disintegration, and dissolution.

A 2025 AAPS Journal study demonstrated ML prediction of drug-release profiles and kinetic parameters from tablet formulations, showing the potential to move beyond single-point dissolution predictions toward full release-profile prediction. Future models may combine material characterization, compaction profiles, tablet geometry, process history, and dissolution data to predict performance under multiple physiological or biorelevant conditions.

13. AI in Liquid and Semi-Solid Formulations

Liquid dosage forms involve variables such as solubility, pH, ionic strength, viscosity, preservatives, suspending agents, surfactants, rheological behavior, and physical stability. AI can help predict viscosity, sedimentation, particle-size behavior, precipitation risk, and stability. A 2025 open-access review in Therapeutic Innovation & Regulatory Science specifically examined AI applications in liquid dosage-form development and highlighted the opportunity to reduce trial-and-error optimization across formulation and manufacturing steps.

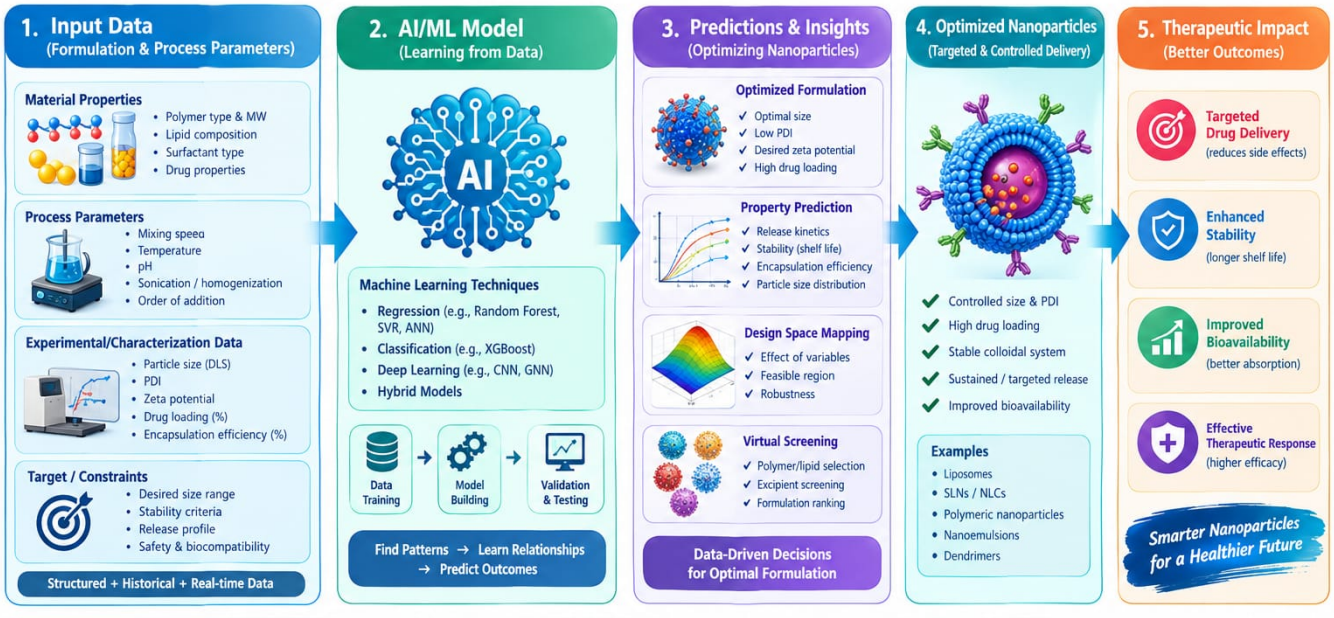
For semisolid products such as creams, gels, and ointments, rheology and texture are particularly important. AI can potentially link composition and process variables to spreadability, viscosity, drug release, and sensory attributes.

14. AI in Nanoparticle and Lipid-Based Formulations

Nanoparticle and lipid-based systems create complex design spaces involving particle size, polydispersity, surface charge, encapsulation efficiency, drug loading, morphology, release rate, and stability. AI can model relationships among lipid composition, polymer properties, solvent conditions, mixing parameters, and nanoparticle characteristics.

The major opportunity is inverse design: specify a target particle size, loading, release profile, or surface characteristic and identify candidate compositions and process conditions. However, the risk of model overfitting is substantial because many nanomedicine datasets are small, heterogeneous, and generated under different laboratory protocols.

Standardized reporting of materials, process conditions, characterization methods, and definitions is therefore essential before large-scale AI models can become broadly transferable.



15. AI for Controlled and Targeted Drug Delivery

Controlled-release systems require precise relationships among matrix properties, drug physicochemical characteristics, polymer composition, geometry, porosity, and environmental conditions. ML can predict release kinetics and help optimize polymer ratios and manufacturing conditions.

For targeted delivery, AI can integrate carrier characteristics with biological data to predict uptake, tissue

distribution, or interactions with biological barriers. Such models should be treated as decision-support tools until prospective experimental and biological validation demonstrates reliable generalization.

16. AI and 3D-Printed Personalized Medicines

Three-dimensional printing creates opportunities for patient-specific dose, geometry, release profile, and combination therapy. AI can assist by mapping printing parameters and formulation composition to tablet geometry, mechanical properties, dose uniformity, and dissolution.

The combination of AI and 3D printing is particularly attractive for personalized medicine because it can support rapid adjustment of formulation and manufacturing

variables. A closed-loop system could potentially read patient or prescription requirements, select an appropriate formulation, predict CQAs, manufacture the dosage form, characterize it, and adjust future production based on accumulated data.

17. AI for Drug Release and Stability Prediction

Drug release and stability are central determinants of dosage-form performance. Conventional stability studies require long periods and multiple storage conditions. AI may assist by learning relationships between formulation composition, packaging, environmental exposure, and degradation or physical changes.

Predictions should support, not replace, the required stability studies. Instead, models can prioritize formulations, identify risk factors, support accelerated development, and help identify unusual behavior. Prospective validation under relevant storage conditions remains necessary.

18. AI within Quality by Design

QbD begins with a predefined target product profile and uses scientific understanding and risk assessment to establish a design space that consistently produces acceptable quality. AI can strengthen QbD by learning nonlinear relationships among material attributes, process parameters, and CQAs.

An AI-enabled QbD workflow can include: target product profile definition → critical quality attribute identification → material and process risk assessment → DoE data generation → ML model development → validation → design-space exploration → confirmation experiments → control-strategy development.

The guiding principle is to keep AI within a scientifically justified design space rather than become an uncontrolled black-box optimiser.

19. AI and Process Analytical Technology

PAT aims to understand and control manufacturing processes through timely measurement of critical variables. Modern manufacturing generates large streams of process data from spectroscopy, imaging, sensors, process controls, and equipment monitoring.

ML can convert these data into predictions of blend uniformity, moisture, particle characteristics, coating performance, or other CQAs. In advanced settings, a model may identify an emerging process deviation before the finished-product test fails.

Integration of PAT with AI can therefore move quality control from predominantly end-product testing toward continuous process understanding and prediction.

20. AI in Scale-Up and Pharmaceutical Manufacturing

Scale-up introduces new sources of variability because equipment geometry and process dynamics change. AI can learn relationships between laboratory and manufacturing-scale parameters, potentially supporting scale translation and process robustness.

Applications include process parameter optimization, predictive maintenance, deviation detection, batch monitoring, and real-time quality prediction. However, scale-up models require datasets representing multiple equipment configurations and operating ranges. A model trained only on one laboratory system may not generalize to a commercial line.

21. Digital Twins and Autonomous Formulation Development

A digital twin is a computational representation of a physical system that is updated using data from the real process. In formulation development, a digital twin could represent a formulation, process, or manufacturing line and simulate the effect of changing inputs.

The longer-term goal is a formulation laboratory capable of running an increasingly autonomous workflow. In such a system, AI selects an experiment; robotic equipment prepares and analyzes the formulation; analytical results return to the model; uncertainty is updated; and the next experiment is automatically selected. This creates a closed-loop cycle of prediction, experimentation, learning, and optimization.

Automation should be introduced in stages. Early systems may automate only experiment selection or data processing, while later systems may integrate formulation preparation, characterization, and model updating. Human oversight remains necessary to protect safety, interpret results, and maintain regulatory accountability.

22. Explainable AI and Model Interpretability

Formulation scientists need to know why an algorithm recommends a formulation. A prediction without an interpretable rationale may be difficult to trust, especially when it contradicts established scientific knowledge.

Feature importance, partial-dependence methods, SHAP values, sensitivity analysis, and surrogate models can help explain predictions. However, explainability methods should not be treated as proof of causality. An important feature in a predictive model may be correlated with the true causal factor rather than directly causing the outcome.

The ideal formulation AI system should therefore provide prediction, uncertainty, influential variables, applicability domain, and a scientifically interpretable explanation.

23. Data Quality, Validation, and Reproducibility

Data quality remains one of the main barriers to reliable AI in formulation development. Even an advanced model cannot overcome measurements that are unreliable or reported inconsistently. Differences in dissolution methods, particle-size measurements, equipment, sampling procedures, excipient grades, environmental conditions, and terminology can introduce hidden variability.

High-quality datasets should include complete formulation composition, API properties, excipient identity and grade, processing parameters, analytical method, experimental conditions, batch information, and outcome measurements. Missing metadata should be explicitly recorded rather than silently imputed.

Validation should include internal cross-validation and, whenever possible, external or prospective validation. Model performance should be reported with appropriate metrics and uncertainty, not only a single coefficient of determination.

24. A Practical Framework for AI-Enabled Formulation Development

Stage	Key Inputs	AI Activity	Experimental Output
1. Target definition	TPP, route, dose, CQAs	Constraint definition and objective setting	Validated target criteria
2. Data assembly	API, excipients, historical batches	Data cleaning and feature engineering	Curated dataset
3. Initial experiments	DoE variables	Baseline model training	Initial predictive model
4. Prediction	Candidate compositions/processes	CQA prediction and uncertainty	Ranked candidates
5. Optimization	Objectives and constraints	Bayesian/multi-objective optimization	Next experiments
6. Confirmation	Selected formulations	Prospective prediction vs observed results	Model validation
7. Scale-up	Process and equipment data	Robustness and transfer modeling	Scale-up strategy
8. Continuous learning	Commercial/process data	Drift detection and model updating	Improved control strategy

25. Regulatory, Ethical, and Governance Considerations

AI-assisted formulation decisions must remain consistent with pharmaceutical quality requirements and applicable regulatory expectations. The regulatory challenge is not simply whether AI is used, but whether the resulting development process is scientifically justified, controlled, validated, documented, and reproducible.

Important governance questions include model version control, training-data provenance, change management, cybersecurity, audit trails, human oversight, software validation, handling of model updates, and responsibility for decisions. A model that changes over time requires a predefined strategy for monitoring and requalification.

Ethically, AI should not create false confidence. Predictive performance must be communicated with uncertainty, and human experts should retain authority over decisions involving safety and product quality.

26. Major Research Gaps

- Large, standardized, interoperable formulation datasets remain scarce.
- Many published models lack prospective external validation.
- Different laboratories use different analytical methods and reporting conventions.
- Small datasets limit the reliability of highly complex deep-learning models.
- Domain shift between research and industrial manufacturing environments is insufficiently studied.
- Uncertainty quantification is often less developed than point prediction.
- Few studies demonstrate end-to-end AI workflows from prediction through laboratory confirmation and scale-up.
- Regulatory expectations for AI-assisted formulation decisions require continued clarification and practical guidance.
- Human–AI interaction and expert acceptance have received less attention than algorithmic accuracy.
- Open benchmark datasets and standardized evaluation protocols are needed to compare models fairly.

27. Future Perspectives

The next decade is likely to move formulation AI from isolated prediction tasks toward integrated development ecosystems. Large language models may assist with knowledge retrieval and experimental planning, while specialized ML models handle numerical prediction. Multimodal systems could integrate text, molecular structures, images, spectra, process data, and formulation tables.

Another important direction is self-driving formulation laboratories. Instead of optimizing only a single formulation parameter, autonomous systems could learn across multiple experiments while respecting constraints on safety, material availability, cost, and

manufacturability.

The most important future advance may therefore be the integration of AI with experimental infrastructure rather than increasingly complicated algorithms alone. A modest but well-validated model connected to high-quality automated experimentation may be more valuable than a sophisticated model trained on inconsistent data.

A mature AI formulation platform should eventually answer five questions: What formulation should be tested? Why was it selected? How certain is the prediction? What experiment will provide the most useful information? And did the experimental result improve the model? These questions define a scientifically useful closed-loop approach.

28. Conclusion

Artificial intelligence is reshaping pharmaceutical formulation development by enabling a transition from predominantly empirical experimentation toward predictive, optimized, and potentially autonomous workflows. AI can support preformulation, excipient selection, formulation composition, drug-release prediction, stability assessment, process development, QbD, PAT, scale-up, and personalized manufacturing. Recent evidence includes ML-based excipient prediction for biopharmaceuticals, ML prediction of tablet drug-release profiles, AI-assisted liquid formulation development, and Bayesian optimization strategies that iteratively select informative experiments.

Nevertheless, AI should be viewed as a scientific augmentation technology rather than a replacement for formulation expertise. The greatest limitations are data quality, model generalizability, interpretability, uncertainty, reproducibility, and regulatory governance. Formulation development is likely to evolve toward a hybrid approach: formulation scientists define clinically and pharmaceutically meaningful objectives; AI identifies patterns and candidate solutions; automated experiments generate new evidence; and validated models continuously improve. The ultimate transition is therefore not simply from trial-and-error to AI, but from trial-and-error to evidence-driven, predictive, experimentally validated, and eventually autonomous formulation design.

29. Suggested Graphical Abstract Concept

Traditional pathway: API characterization → Excipient screening → Prototype formulation → Testing → Failure/adjustment → Re-testing → Optimization.

AI-enabled pathway: Historical + experimental data → Feature engineering → AI/ML prediction → Multi-objective optimization → Candidate formulation → Laboratory confirmation → Model updating.

Future autonomous pathway: Target product profile → AI planner → Robotic formulation → Automated characterization → AI interpretation → Next-best experiment → Confirmed formulation → Scale-up digital twin.

30. Tables for Journal Submission

AI approach	Representative formulation use	Main advantage / limitation
Random forest / gradient boosting	CQA prediction, excipient screening	Strong nonlinear prediction; may be less interpretable
Support vector machine	Classification/regression of formulation outcomes	Useful for smaller datasets; sensitive to feature scaling and kernel choice
Neural networks	Complex formulation-property relationships	Flexible; may require larger datasets
Gaussian process	Bayesian optimization, uncertainty-aware prediction	Excellent for small data; computational burden can increase
NLP / LLM	Literature mining, knowledge extraction, experimental planning	Large-scale knowledge access; extraction errors require verification
Generative/inverse models	Candidate formulation generation	Can explore design space; needs strong constraints and validation
Active learning	Selection of next experiment	Reduces unnecessary experiments; depends on reliable uncertainty estimates

REFERENCES

1. Huanbutta, K., Burapapadh, K., Kraisit, P., et al. (2024). Artificial intelligence-driven pharmaceutical industry: A paradigm shift in drug discovery, formulation development, manufacturing, quality control, and post-market surveillance. European Journal of Pharmaceutical Sciences, 203, 106938. https://doi.org/10.1016/j.ejps.2024.106938
2. Ali, K. A., Mohin, S. K., Mondal, P., et al. (2024). Influence of artificial intelligence in modern pharmaceutical formulation and drug development. Future Journal of Pharmaceutical Sciences, 10, 53. https://doi.org/10.1186/s43094-024-00625-1
3. Wang, X., Liu, Z., Lin, X., et al. (2025). A novel paradigm on data and knowledge-driven drug formulation development:

Opportunities and challenges of machine learning. Journal of Industrial Information Integration, 44, 100796. https://doi.org/10.1016/j.jii.2025.100796

4. Vidal-Henriquez, E., Holder, T., Lee, N. F., et al. (2025). Machine learning driven acceleration of biopharmaceutical formulation development using Excipient Prediction Software (ExPreSo). Computational and Structural Biotechnology Journal, 27, 4517–4525. https://doi.org/10.1016/j.csbj.2025.10.026

5. Nithyanantham, D., Nair, A., & Nayak, U. Y. (2025). Leveraging Artificial Intelligence for Advancements in Liquid Dosage Formulations in the Pharmaceutical Industry. Therapeutic Innovation & Regulatory Science, 59, 1004–1031. https://doi.org/10.1007/s43441-025-00823-w
6. Protopapa, C., Siamidi, A., Eneli, A. A., Elbadawi, M., Vlachou, M., et al. (2025). Machine Learning Predicts Drug Release Profiles and Kinetic Parameters Based on Tablets' Formulations. The AAPS Journal, 27, 124. https://doi.org/10.1208/s12248-025-01101-1
7. Artificial Intelligence (AI) in Pharmaceutical Formulation and Dosage Calculations. (2025). Pharmaceutics, 17(11), 1440. https://doi.org/10.3390/pharmaceutics17111440
8. Bayesian optimization and machine learning for vaccine formulation development. (2025). PLOS ONE, 20(6), e0324205. https://doi.org/10.1371/journal.pone.0324205
9. Artificial Intelligence and Machine Learning in Pharmaceutical Formulation Development: Predictive Modeling, Optimization, and Future Perspectives. (2024). International Journal of Farmacia, 10(3).
10. Artificial intelligence for drug delivery: Yesterday, today and tomorrow. (2026). PMC Review.
11. ICH. (2009). Q8(R2): Pharmaceutical Development. International Council for Harmonisation.
12. ICH. (2023). Q9(R1): Quality Risk Management. International Council for Harmonisation.
13. ICH. (2008). Q10: Pharmaceutical Quality System. International Council for Harmonisation.
14. ICH. (2022). Q13: Continuous Manufacturing of Drug Substances and Drug Products. International Council for Harmonisation.
15. U.S. Food and Drug Administration. (2023). Artificial Intelligence and Machine Learning (AI/ML) in Drug Development: Discussion Paper. FDA.
16. European Medicines Agency. (2023). Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. EMA.
17. 50 Additional Standard References
18. Bannan, C. C., Calabró, G., Kyu, D. Y., & Mobley, D. L. (2016). Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. Journal of Chemical & Engineering Data, 61(12), 4382–4395.
19. Barmpalexis, P., Kanaze, F. I., & Georgarakis, E. (2011). Developing dynamic neural networks for predicting dissolution profiles of solid dosage forms. Journal of Pharmacy and Pharmacology, 63(6), 773–783.
20. Bikiaris, D. N. (2011). Solid dispersions, part I: Recent evolutions and historical review in preparation methods. Expert Opinion on Drug Delivery, 8(11), 1501–1519.
21. Brough, C., & Williams, R. O. (2013). Amorphous solid dispersions and nano-crystal technologies for poorly water-soluble drugs. International Journal of Pharmaceutics, 453(1), 157–166.
22. Bynagari, R., & Dangeti, B. (2023). Application of predictive analytics and machine learning tools in tablet dissolution testing and optimization. Pharmaceutics, 15(4), 1102.
23. Cal, K., & Sollohub, K. (2010). Spray drying technique. II. Review of pharmaceutical applications. Journal of Pharmaceutical Sciences, 99(2), 575–586.
24. Chen, Y., Zhang, Y., & Liu, J. (2022). Machine learning models for predicting drug solubility in co-solvents: A practical evaluation. Molecular Pharmaceutics, 19(5), 1432–1442.
25. Chow, K. T., Zheng, W., & Tan, R. B. (2008). Application of artificial neural networks in tablet compression processes. Drug Development and Industrial Pharmacy, 34(3), 290–298.
26. Elbadawi, M., Muñiz Castro, B., Gavins, F. K., et al. (2021). M3DISEQUAL: A machine learning framework for predicting 3D

printing of medicines. International Journal of Pharmaceutics, 604, 120715.
27. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. International Conference on Machine Learning, 1050–1059.
28. Galata, D. L., Mesko, A., Dargai, Z., et al. (2021). Applications of machine learning algorithms in pharmaceutical production and quality control. European Journal of Pharmaceutical Sciences, 159, 105718.
29. Han, R., Yang, Y., & Xi, X. (2023). Bayesian optimization for formulation design of lipid nanoparticles for mRNA delivery. ACS Nano, 17(8), 7420–7431.
30. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.
31. Khaw, K. Y., Billa, N., & Ching, S. Y. (2022). Machine learning methods in predicting nanoparticle cell uptake and delivery performance. Nanomedicine: Nanotechnology, Biology and Medicine, 41, 102520.
32. Koras, K., & Binienda, Z. (2024). Active learning paradigms in early-stage formulation screening: Balancing exploration and exploitation. Pharmaceutics, 16(2), 245.
33. Le, T. C., Yuriev, E., & Chalmers, D. K. (2012). Quantitative structure–property relationship (QSPR) modeling of drug release from polymeric matrices. Journal of Controlled Release, 161(2), 516–528.
34. Li, M., & Zhao, Z. (2022). Application of machine learning in solid-state form selection and crystallization process design. Crystallization Growth & Design, 22(10), 6120–6135.
35. Lou, H., Hageman, M. J., & Patel, S. (2021). Application of artificial intelligence and machine learning in preformulation and formulation development. AAPS Open, 7, 12.
36. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774.
37. Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. Journal of the American Statistical Association, 44(247), 335–341.
38. Modaresi, S. M. S., & Taylor, L. S. (2023). Machine learning models predicting the physical stability of amorphous solid dispersions. Molecular Pharmaceutics, 20(3), 1680–1694.
39. Mohapatra, S., & Sahoo, S. K. (2020). Artificial intelligence in nanomedicine and targeted drug delivery systems. Drug Discovery Today, 25(12), 2245–2255.
40. O'Donnell, P. B., & McGinity, J. W. (1997). Preparation of microspheres by the solvent evaporation technique. Advanced Drug Delivery Reviews, 28(1), 25–42.
41. Ong, J. J., Awad, A., & Basit, A. W. (2022). Transformative role of artificial intelligence in 3D printed pharmaceuticals. Advanced Drug Delivery Reviews, 188, 114450.
42. Patel, M. N., & Lazzara, M. J. (2023). Machine learning algorithms for optimizing multi-component polymeric drug delivery systems. Journal of Controlled Release, 355, 340–352.
43. Paul, D. A., & Sanganahalli, U. M. (2022). Natural language processing for literature mining in pharmaceutical formulation databases. Computers in Biology and Medicine, 148, 105890.
44. Peppas, N. A. (1985). Analysis of Fickian and non-Fickian drug release from polymers. Pharmaceutica Acta Helvetiae, 60(4), 110–111.
45. Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian Processes for Machine Learning. MIT Press.
46. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.
47. Rokoch, L., & Maimon, O. (2005). Decision trees. Data Mining and Knowledge Discovery Handbook, 165–192.
48. Rowley, G. (2001). Quantifying standard excipient behavior and interaction through automated testing systems. International Journal of Pharmaceutics, 223(1-2), 11–23.
49. Shah, R. B., Tawde, V. A., & Khan, M. A. (2007). Process Analytical Technology (PAT) and Quality by Design (QbD) in solid oral dosage form manufacturing. International Journal of Pharmaceutics, 338(1-2), 84–92.
50. Siepmann, J., & Peppas, N. A. (2011). Higuchi equation: Derivation, applications,

use and misuse. International Journal of Pharmaceutics, 418(1), 6–12.
51. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. Advances in Neural Information Processing Systems (NeurIPS), 25, 2951–2959.
52. Solanki, N. G., & Serajuddin, A. T. (2020). Modern advancements in amorphous solid dispersion: From formulation screening to industrial execution. Journal of Pharmaceutical Sciences, 109(4), 1389–1407.
53. Sousa, A. S., Serra, J., Estevens, C., Costa, R., & Ribeiro, A. J. (2025). A comparative study of two data-driven modeling approaches to predict drug release from ER matrix tablets. International Journal of Pharmaceutics, 671, 125230.
54. Sun, C. C. (2011). Materials science basis of powder compression. International Journal of Pharmaceutics, 418(2), 290–299.
55. Szepes, A., & Szabó-Révész, P. (2022). Implementation of PAT tools in continuous wet granulation and tableting processes. European Journal of Pharmaceutics and Biopharmaceutics, 170, 110–121.
56. Takayama, K., Fujikawa, M., & Nagai, T. (1999). Artificial neural network as a tool for formulation optimization. Advanced Drug Delivery Reviews, 36(2-3), 245–258.
57. Trenfield, S. J., Awad, A., & Basit, A. W. (2022). Autonomous formulation design using 3D printing and machine learning. Trends in Pharmacological Sciences, 43(6), 461–475.
58. Vora, C., & Joshi, M. (2023). Artificial intelligence in process scale-up and technology transfer of solid oral dosage forms. Drug Development and Industrial Pharmacy, 49(8), 512–524.
59. Wang, Z., & Jiang, J. (2024). Deep generative models for inverse design of multi-functional excipients and pharmaceutical carriers. ACS Central Science, 10(2), 312–324.
60. Yu, L. X., Amidon, G., Khan, M. A., et al. (2014). Understanding Quality by Design (QbD) for pharmaceuticals: Principles and applications. The AAPS Journal, 16(4), 771–783.
61. Zhang, L., & Liu, X. (2024). Explainable AI methods in biopharmaceutical stability assessment: A SHAP-based comparative analysis. Journal of Pharmaceutical and Biomedical Analysis, 239, 115890.
62. Zhao, H., & Chen, Y. (2026). Artificial intelligence in the rational design of lipid nanoparticles for nucleic acid delivery. The Innovation Drug Discovery, 2(1), 100006.
63. Zhou, Q., & Morton, D. A. (2012). Drug–excipient interactions in dry powder inhaler formulations: Mechanism and prediction using computational modeling. European Journal of Pharmaceutical Sciences, 47(3), 545–558.
64. Zidan, A. S., & Khan, M. A. (2021). Continuous manufacturing and PAT integration in hot-melt extrusion processing. International Journal of Pharmaceutics, 592, 120080.
65. Zowada, A., & Sznitowska, M. (2023). Machine learning applications in rheological behavior prediction for topical semi-solid formulations. Pharmaceutics, 15(7), 1890.
66. Zubarev, D. Y., & Santos, R. (2021). Closed-loop autonomous laboratories for material and formulation discovery. Nature Reviews Materials, 6(11), 1001–1015.66. Zhao, H. (2026). Artificial intelligence in the rational design of lipid nanoparticles for targeted drug delivery. The Innovation Drug Discovery, 2(1), 100006.