

AI and Machine Learning in Pharmaceutical Formulation Development: Applications, Challenges and Future Perspectives

J. Dakshayani¹, B Nagamani², B Renuka³, P Umadevi⁴
^{1,2,3,4} Viswanadha Institute of Pharmaceutical Sciences, Visakhapatnam,

Abstract—Pharmaceutical formulation development is traditionally an iterative and experimentally intensive process involving the selection of active pharmaceutical ingredients, excipients, manufacturing variables, and processing conditions to obtain a product with predefined quality and performance characteristics. The increasing complexity of modern drug-delivery systems has created a need for more efficient and predictive approaches to formulation design. Artificial intelligence (AI), particularly machine learning (ML), deep learning (DL), artificial neural networks (ANNs), Bayesian optimization, and other data-driven approaches, is emerging as an important decision-support technology in pharmaceutical formulation development. These approaches can identify complex relationships between formulation variables and critical quality attributes (CQAs), predict drug-excipient interactions, optimize excipient concentrations, estimate drug release, and reduce the number of experimental trials required for formulation optimization. Recent applications extend to amorphous solid dispersions, nanoparticles, lipid nanoparticles, liposomes, nanocrystals, tablets, controlled-release systems, biologics, and long-acting formulations. Recent literature also highlights active learning, multi-objective optimization, digital twins, generative approaches, and self-driving laboratories as emerging technologies for pharmaceutical development. However, the successful implementation of AI/ML is constrained by limited and heterogeneous datasets, data quality, model interpretability, overfitting, inadequate external validation, biological complexity, reproducibility, regulatory requirements, and the need for experimental confirmation. This review discusses the principles of AI/ML relevant to pharmaceuticals, major applications in formulation development, advantages and limitations, integration with Quality by Design (QbD), regulatory and ethical considerations, and future prospects. AI is best regarded as a complementary decision-support

technology rather than a replacement for pharmaceutical scientists and experimental validation.

Index Terms—Artificial intelligence; machine learning; pharmaceutical formulation; formulation optimization; drug delivery; deep learning; artificial neural networks; Quality by Design; predictive modeling; critical quality attributes.

I. INTRODUCTION

Pharmaceutical formulation development aims to convert an active pharmaceutical ingredient (API) into a safe, stable, effective, manufacturable, and patient-acceptable dosage form. Conventional formulation development frequently depends on systematic experimentation in which formulation composition and process variables are modified and the resulting product characteristics are evaluated. Although Design of Experiments (DoE), Quality by Design (QbD), and statistical optimization have improved this process, modern formulations can contain numerous interacting variables, making formulation development increasingly multidimensional.

Machine learning has emerged as a potential tool for addressing this complexity. Early applications demonstrated that ML could predict formulation properties and guide experimental formulation development, while more recent research has expanded toward advanced optimization, active learning, Bayesian optimization, and automated experimental platforms. A 2025 review described AI applications across pharmaceutical formulation and dosage calculations, including prediction of drug properties, excipient optimization, formulation design, and pharmacokinetic/pharmacodynamic modeling.

MDPI More recent reviews published in 2026 specifically describe AI/ML-guided optimization of drug-delivery systems and its application to solid oral formulations, lipid nanoparticles, biologics, and long-acting injectable systems. The increasing availability of formulation datasets and computational resources therefore creates an opportunity to move from predominantly empirical formulation development toward data-driven and predictive development. However, AI should not be considered a substitute for pharmaceutical experimentation. Current evidence supports its role primarily as a decision-support layer that can help formulate hypotheses, prioritize experiments, identify nonlinear relationships, and explore formulation design spaces. AI provides an invaluable resource that could redefine the entire drug development paradigm. It achieves this by emulating human cognitive functions through intricate algorithms and advanced machine learning techniques, enabling capabilities in data analysis, pattern recognition, and decision-making that were previously unimaginable. To protect patients and ensure equitable treatment outcomes, we must develop robust frameworks for ethical data handling and create diverse datasets that accurately represent the populations we aim to serve. Another major hurdle lies in the regulatory environment. Current regulatory frameworks are often ill-equipped to evaluate and validate AI-driven solutions in drug development, resulting in uncertainty for pharmaceutical companies seeking to innovate. Regulators face the enormous challenge of fostering an environment that encourages technological progress while ensuring patient safety and building public trust. Achieving this delicate balance requires the establishment of clear guidelines for AI model validation, ensuring transparency and explainability in AI-driven decisions, and implementing rigorous post-deployment monitoring to assess the long-term implications of these technologies on patient health. The integration of AI into legacy pharmaceutical workflows also poses significant organizational and infrastructural challenges. Successfully incorporating AI requires a substantial investment in digital infrastructure, ongoing workforce upskilling, and a cultural shift within organizations that are deeply rooted in traditional research methodologies. Furthermore, companies must carefully evaluate the cost-effectiveness and long-term benefits of AI

implementation in comparison to other foundational improvements, such as enhancing data management capabilities and promoting interoperability across systems and platforms. ^[1, 2, 3]

II. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

2.1 Artificial intelligence

Artificial intelligence refers broadly to computational methods capable of performing tasks that traditionally require aspects of human intelligence, including pattern recognition, prediction, classification, optimization, and decision support. In pharmaceutical formulation development, AI can be used to analyze relationships between:

- **API properties:** Artificial intelligence can analyze the physicochemical and biopharmaceutical properties of an active pharmaceutical ingredient (API) and predict their influence on formulation performance. Important API-related parameters include aqueous solubility, pKa, logP/logD, melting point, particle size, crystal form, polymorphism, hygroscopicity, permeability, density, and chemical stability. Machine-learning models can identify relationships between these properties and formulation outcomes such as dissolution, drug loading, release rate, permeability, and stability. For example, AI can predict whether an API is likely to exhibit poor aqueous solubility and can subsequently assist in selecting an appropriate formulation strategy such as solid dispersion, nanosuspension, lipid-based formulation, cyclodextrin complexation, or nanocarrier systems.
- **Excipient characteristics** Excipients have a major influence on the physical, chemical, and biopharmaceutical performance of a formulation. AI can analyze the characteristics of different excipients and determine their effects on formulation quality.
- **Formulation composition:** Formulation composition refers to the quantitative and qualitative combination of API and excipients used to prepare the dosage form. AI can evaluate the relationship between formulation variables and important product characteristics.

- **Manufacturing parameters:** Manufacturing parameters describe the conditions under which a pharmaceutical product is produced. AI can identify how manufacturing variables affect the final product. Examples include: Mixing time, Mixing speed
- **Process conditions:** Process conditions are closely associated with manufacturing parameters but generally describe the environmental and operational conditions during processing. Important conditions include: Temperature, pH, Relative humidity, Pressure, Processing time
- **Critical quality attributes:** Critical Quality Attributes are measurable physical, chemical, biological, or microbiological characteristics that should remain within appropriate limits to ensure the required quality of a pharmaceutical product.
- **Drug-release characteristics:** Drug release is one of the most important performance characteristics of a pharmaceutical dosage form. AI can analyze experimental dissolution data and predict drug-release behavior.
- **Stability:** Stability studies traditionally require long periods of experimental testing. AI can potentially assist in predicting stability trends from historical and experimental data.
- **Biological performance:** AI can also establish relationships between formulation characteristics and biological performance. Depending on the dosage form, biological performance. ^[4,5]

2.2 Machine learning

Machine learning (ML) is a subset of artificial intelligence in which computational algorithms learn patterns and relationships from experimental or historical data and use these relationships to make predictions, classifications, or optimization decisions. In pharmaceutical formulation development, ML can be particularly useful because formulation performance is often influenced by multiple interacting variables, including API properties, excipient type and concentration, process parameters, and environmental conditions. Depending on the nature and size of the dataset, different ML algorithms can be selected. Commonly used approaches include linear and nonlinear regression,

random forest, support vector machine/support vector regression, artificial neural networks, gradient boosting, decision trees, k-nearest-neighbor methods, Gaussian-process models, and deep neural networks. The choice of algorithm depends on the size and structure of the dataset, the type of prediction required, and the degree of interpretability needed.

2.3 Deep learning

Deep learning uses multilayer neural networks to identify increasingly complex patterns in large datasets. It may be particularly useful where relationships between formulation variables and outputs are highly nonlinear. However, deep learning is not automatically superior to simpler ML models. For small pharmaceutical datasets, simpler models may provide better generalization and greater interpretability.

Table 1. Potential AI applications according to formulation type

Formulation	Potential AI application
Tablets	Hardness, dissolution, disintegration
Capsules	Formulation optimization
Amorphous solid dispersions	Drug-polymer selection
Nanoparticles	Size, PDI, loading, release
Liposomes	Composition and encapsulation
Lipid nanoparticles	Lipid composition and particle characteristics
Nanoemulsions	Droplet size and stability
Nanocrystals	Particle size and dissolution
Controlled-release systems	Release-profile prediction
Long-acting injectables	Release duration and burst prediction
Protein formulations	Stability and aggregation prediction

2.4 Supervised and unsupervised learning

Supervised learning uses labeled input-output datasets. Examples include predicting particle size from formulation variables or predicting dissolution characteristics from tablet composition.

Unsupervised learning identifies patterns or groups without predefined output labels. Clustering may, for example, help identify groups of formulations with similar characteristics. ^[6,7]

2.5 Reinforcement learning and active learning

Reinforcement learning involves algorithm learning through interactions with an environment and optimizing a defined objective. In formulation development, related concepts may eventually support autonomous selection of experiments. Active learning is particularly relevant because it allows a model to identify which new experiment would provide the most useful information. This can reduce unnecessary experimental work and has become an important component of emerging AI-guided formulation workflows.

III. AI/ML WORKFLOW IN PHARMACEUTICAL FORMULATION DEVELOPMENT

AI/ML Workflow in Pharmaceutical Formulation Development translates a pharmaceutical or chemical challenge into a concrete, measurable data science task. A typical AI-assisted formulation-development workflow can be represented as:

Problem definition → Data collection → Data preprocessing → Feature selection → Model development → Model validation → Prediction/optimization → Experimental confirmation → Model refinement

3.1 Problem definition

The first step is to clearly define the formulation problem.

Examples include:

- Predicting drug solubility: Solubility is a primary determinant of a drug's bioavailability (how much of the drug enters systemic circulation to have an effect). Poorly soluble drugs are difficult to formulate. AI models use the molecular structure of the active pharmaceutical ingredient (API)—such as functional groups, molecular weight, and lipophilicity ($\log P$)—to predict aqueous or organic solubility. It gives Quantitative solubility values (mg/mL) or binary classification (e.g., highly soluble vs. poorly soluble)
- Selecting suitable excipients: Excipients are the inactive ingredients (binders, fillers, lubricants, stabilizers) mixed with the API to create the final dosage form. They must be compatible with the API. AI Recommender systems and machine learning algorithms analyze historical

compatibility databases, molecular structures, and functional properties to flag potential API-excipient interactions or degradation paths.

- Optimizing nanoparticle size: Advanced delivery systems (like liposomes or polymeric nanoparticles), particle size dictates how the drug travels through the body, its cellular uptake, and its toxicity. Particle size depends heavily on a complex web of process variables (stirring speed, sonication time, temperature) and formulation variables (polymer concentration, surfactant type). AI maps the non-linear relationships between these manufacturing parameters and the resulting Critical Quality Attributes (CQAs), like mean particle diameter and polydispersity index (PDI)
- Maximizing encapsulation efficiency: Encapsulation efficiency measures the percentage of the total drug that successfully gets trapped inside a delivery vehicle (like a nanoparticle or micelle) versus what is left wasted in the surrounding liquid. Predictive models evaluate how different carrier matrices, drug-to-polymer ratios, and solvent environments affect the drug-retention capacity of the carrier.
- Predicting dissolution: Dissolution is the process by which a solid drug substance dissolves into a liquid medium over time. It mimics how a tablet or capsule behaves in the human stomach or gut. Models simulate the dissolution curve (f_2 similarity factor) by analyzing physical inputs like compression force, particle size distribution, and excipient ratios.
- Optimizing drug release: sustained, controlled, or immediate-release formulations, the drug must release into the body at a specific, predetermined rate. AI acts as an optimization engine. By pairing neural networks with genetic algorithms, it back-calculates the exact formulation composition needed to match a target kinetic release profile.
- Predicting formulation stability: Formulation stability ensures that the product remains safe, effective, and physically intact throughout its shelf life under various temperature and humidity conditions. AI uses chemical degradation kinetics, quantum chemical

descriptors, and environmental data to predict shelf-life, moisture uptake, and potential degradation products.

IV. DATA COLLECTION

Data can originate from:

1. **Published Literature:** Published literature includes scientific information available in research papers, review articles, theses, patents, and conference proceedings. Published literature provides a large amount of previously generated scientific data and can reduce the need to repeat experiments. Drug solubility and permeability, Melting point and pKa, Drug-excipient compatibility

2. **Pharmaceutical Databases:** Pharmaceutical databases are organized collections of information related to drugs, chemicals, excipients, formulations, and biological properties. Databases provide structured and standardized data, which is particularly useful for computational modelling and machine learning.

3. **Laboratory Experiments:** Laboratory experiments generate primary experimental data directly from formulation and evaluation studies. This is often one of the most important sources of data because the researcher generates the data under controlled experimental conditions. Experimental data provide formulation-specific information and can be used to validate AI/ML predictions. **Example:** Drug concentration, Polymer concentration, Surfactant concentration

4. **Manufacturing Records:** Manufacturing records contain information generated during the **actual** production of pharmaceutical products. Manufacturing data help in Process optimization, Process monitoring, Quality prediction, Identification of process variability, Predictive maintenance Continuous manufacturing

5. **Preformulation Studies:** Preformulation studies are performed before developing the final dosage form. They provide information about the physicochemical and biopharmaceutical properties of the drug. Preformulation data form the foundation of formulation development because they help researchers understand the drug before selecting excipients and manufacturing methods. Important

preformulation data include Solubility, pKa, LogP and Melting point

6. **Process Analytical Technology (PAT):** Process Analytical Technology (PAT) refers to the use of analytical tools and measurements to monitor and control pharmaceutical manufacturing processes in real time. PAT can continuously or frequently measure process parameters and product quality attributes. PAT data are useful for Real-time process monitoring, Early detection of deviations, Process control, Continuous manufacturing, Quality-by-Design (QbD), Real-time release testing

7. **Stability Studies:** Stability studies generate data about how the quality of a pharmaceutical product changes with time and environmental conditions. The quality and representativeness of the dataset strongly influence model performance. Recent reviews emphasize data adequacy and reliability as fundamental requirements for useful ML-based formulation development.^[8,9]

4.1 Data preprocessing

Raw pharmaceutical data frequently contain missing values, inconsistent units, duplicated observations, outliers, and categorical variables. Preprocessing may involve:

- **Data Cleaning:** Data cleaning is the initial "housekeeping" phase. It fixes formatting errors, structural inconsistencies, and duplicates that occur when merging datasets from different labs or time periods.
- **Missing-Value Handling:** Pharma datasets are notoriously incomplete. A scientist might have skipped measuring a specific non-critical attribute during an experiment, or an instrument failure might have caused a data gap.
- **Normalization:** Normalization scales numeric features so that their values fall within a specific, tight range—usually 0 to 1 or -1 to 1.
- **Standardization (Z-score Normalization):** Standardization centers the data around a mean of 0 with a standard deviation of 1.
- **Feature Encoding:** Machine learning models speak the language of mathematics, not English. Feature encoding converts alphabetical or textual categories into numeric values.

- **Outlier Assessment:** An outlier is an observation that lies an abnormal distance from other values in a random sample.
- **Dimensionality Reduction:** Pharmaceutical formulation datasets often suffer from the "curse of dimensionality." You might have hundreds of molecular descriptors for a single drug molecule, which can overwhelm a model.

4.2 Feature selection: Features may include API descriptors, physicochemical properties, excipient characteristics, formulation concentrations, process parameters, and environmental conditions. Selecting

relevant features can improve model performance and reduce unnecessary complexity.

4.3 Model training and validation: The dataset is generally divided into training and testing subsets. Cross-validation may be used during model development. Common performance measures include R^2 , Mean absolute error (MAE), Root mean square error (RMSE), Mean absolute percentage error (MAPE), Accuracy, Precision, Recall. However, a high internal validation score alone does not establish that a formulation model will perform reliably on new formulations.

Table 2. AI/ML techniques relevant to pharmaceutical formulation

Technique	Major application	Potential pharmaceutical output
Linear regression	Quantitative prediction	Dissolution, assay
Random forest	Classification/regression	Formulation performance
Support vector machine/regression	Prediction/classification	Solubility, particle characteristics
Artificial neural network	Nonlinear modelling	Drug release, formulation optimization
Deep learning	Complex pattern recognition	Advanced formulation prediction
Bayesian optimization	Experimental optimization	Optimal formulation composition
Active learning	Experiment selection	Reduced experimental burden
Reinforcement learning	Sequential optimization	Adaptive formulation development
Generative AI	Design generation	Candidate formulations/materials
Digital twins	Process simulation	Process/formulation monitoring

V. APPLICATIONS OF AI AND ML IN PHARMACEUTICAL FORMULATION DEVELOPMENT

5.1 Preformulation studies: Preformulation is an important area where AI can assist in predicting API properties and identifying potential formulation challenges. Possible applications include prediction of: Solubility, Melting point, Partition coefficient, Stability, Crystallinity, Hygroscopicity. AI-based prediction can help identify promising formulation strategies before extensive laboratory experimentation.

5.2 Excipient selection and optimization: Excipients strongly influence the performance of pharmaceutical dosage forms. AI models can analyze relationships between APIs, excipient properties, concentrations, and formulation performance. Potential applications include: Selection of suitable binders,

Selection of solubilizers, Selection of surfactants, Selection of polymers. AI can therefore assist in narrowing the formulation design space before laboratory experiments.

5.3 Oral solid dosage forms: Tablets and capsules remain among the most important pharmaceutical dosage forms. AI/ML may be applied to predict: Tablet hardness, Friability, Disintegration time, Dissolution, Compression behavior, Drug content, Content uniformity, Stability, Granule properties. For modified-release tablets, ML can also be used to investigate relationships between polymer type/concentration and drug-release behavior.

5.4 Amorphous solid dispersions: Amorphous solid dispersions are frequently used to improve the apparent solubility and dissolution of poorly water-soluble drugs. Their development involves several interacting variables, including: Polymer selection, Drug-to-polymer ratio, Processing

temperature, Drug loading, Molecular interactions, Storage conditions. AI/ML may assist in predicting formulation performance and identifying suitable drug-polymer combinations.

5.5 Nanoparticulate drug-delivery systems: Nanoparticle formulation is particularly suitable for AI/ML because of its multidimensional nature. Models can potentially predict: Particle size, Polydispersity index, Zeta potential, Encapsulation efficiency, Drug loading, Morphology, Drug release, Physical stability. A 2026 review summarized ML applications across polymeric nanoparticles, lipid nanoparticles, liposomes, solid lipid nanoparticles, nanostructured lipid carriers, nanoemulsions, nanosuspensions, niosomes, nanocrystals, and related systems. ScienceDirect This represents one of the most active areas of AI-assisted pharmaceuticals.^{10,11}

5.6 Lipid nanoparticles

Lipid nanoparticles have gained considerable attention because of their ability to deliver nucleic acids and other therapeutics. AI/ML can potentially optimize: Lipid composition, Ionizable lipid selection, Helper lipid concentration, Cholesterol concentration, PEG-lipid concentration, Particle size, Encapsulation efficiency, Stability, Release characteristic. The complex relationship between lipid composition and nanoparticle performance makes this an attractive area for predictive modeling.

5.7 Controlled- and sustained-release systems: AI can be used to investigate relationships between formulation variables and drug-release profiles. Important variables may include: Polymer type, Polymer concentration, Polymer molecular weight, Drug loading, Tablet geometry, Porosity, Coating thickness, Environmental pH. The objective is to predict the release profile and identify formulations capable of meeting predefined release specifications. Recent AI/ML literature emphasizes release-profile optimization as one of the emerging applications of computational formulation science.

5.8 Long-acting injectable formulations: Long-acting injectable systems present complex formulation and drug-release problems. AI may assist in predicting: Drug loading, Particle characteristics, Polymer degradation, Initial burst release, Sustained release,

Release duration. The application of ML to long-acting injectable systems has been identified as an emerging area within AI-guided drug-delivery optimization.

5.9 Biologics and protein formulations: Protein and peptide formulations are highly sensitive to environmental and formulation conditions. Potential applications include prediction of: Aggregation, Protein stability, Excipient effects, Degradation, Denaturation, Storage stability. AI may help identify combinations of stabilizers and formulation conditions that improve protein stability. However, biological systems are highly complex, and computational predictions require rigorous experimental validation.

VI. AI AND QUALITY BY DESIGN

Quality by Design is based on understanding how material attributes and process parameters influence product quality. AI/ML can complement QbD by helping identify relationships between:

Material attributes + process parameters → Critical quality attributes → Product performance

For example: Excipient concentration + mixing speed + compression force → tablet properties → dissolution

In nanomedicine: Lipid composition + mixing parameters → particle characteristics → encapsulation and release

AI may therefore support: Identification of important formulation variables, Design-space exploration, CQA prediction, Optimization of formulation variables, Risk assessment, Process monitoring. Recent literature specifically discusses integrating AI into QbD frameworks to identify critical formulation variables and explore design spaces.

VII. ADVANCED AI APPROACHES IN FORMULATION DEVELOPMENT

7.1 Bayesian optimization

Bayesian optimization can be particularly useful when experiments are expensive. The model predicts the likely outcome of untested formulation conditions and identifies promising experimental conditions for subsequent testing.

This creates an iterative cycle:

Prediction → Experiment → New data → Model update → Next experiment

7.2 Multi-objective optimization

Pharmaceutical formulation development often requires simultaneous optimization of several properties.

For example, a nanoparticle formulation may need: Small particle size, Low polydispersity, High encapsulation efficiency, High stability, Appropriate release profile.

These objectives may conflict with one another. Multi-objective optimization can help identify formulations that provide an acceptable balance between multiple CQAs.

7.3 Active learning

Active learning allows an algorithm to select the next experiment based on the information expected from that experiment. This approach can potentially reduce the number of experiments needed to characterize a formulation space.

7.4 Digital twins

Digital twins are computational representations of physical processes or products that can be updated using experimental or manufacturing data. In pharmaceutical manufacturing, the longer-term objective is to use digital models to simulate formulation and process behavior and support real-time decision-making.

7.5 Self-driving laboratories

The combination of AI, robotics, automated experimentation, and analytical technologies may eventually create semi-autonomous or autonomous formulation-development systems. In such a system: AI proposes experiment → robotic platform prepares formulation → analytical instruments characterize formulation → results return to AI → AI proposes next experiment. Self-driving laboratories have been identified as an important future direction for ML-guided formulation science.^{12,13}

VIII. ADVANTAGES OF AI/ML IN PHARMACEUTICAL FORMULATION

Major potential advantages include:

- Reduction in the number of experimental trials

- Faster formulation screening
- Identification of nonlinear relationships
- Prediction of formulation performance
- Efficient utilization of historical data
- Support for personalized formulations
- Optimization of complex drug-delivery systems
- Improved experimental planning
- Exploration of larger formulation design spaces
- Potential reduction in development resources

AI may therefore complement conventional DoE and QbD rather than replace them.

IX. CHALLENGES AND LIMITATIONS

9.1 Limited datasets: One of the major challenges is the relatively small size of many pharmaceutical formulation datasets. Unlike fields such as image recognition, formulation research may generate only tens or hundreds of experiments for a specific formulation problem. Small datasets can increase the risk of overfitting.

Table 3. Major challenges and possible solutions

Challenge	Possible solution
Small datasets	Multicenter data sharing and standardized databases
Poor data quality	Data-cleaning and standardized reporting
Overfitting	Cross-validation and independent testing
Black-box models	Explainable AI
Heterogeneous datasets	Standardized data structures
Poor generalization	External and prospective validation
Regulatory uncertainty	Documented model lifecycle and human oversight
Biological complexity	Integration of physicochemical and biological data
Limited experimental integration	Automated/robotic experimentation

9.2 Data quality and heterogeneity

Published formulation studies may differ in:

- Experimental methods
- Equipment
- Materials

- Units
 - Reporting standards
 - Experimental conditions
 - Characterization techniques
- Combining heterogeneous datasets without appropriate preprocessing can produce misleading models.

9.3 Model interpretability

Some advanced ML models behave as "black boxes," making it difficult to understand why a particular formulation is predicted to perform well. Interpretability is particularly important in pharmaceutical science because formulation decisions must be scientifically justified. Recent guidance-oriented literature has highlighted the need for better modelling practices specific to drug-product development. ScienceDirect+1

9.4 Over fitting

A model may perform very well on its training dataset but poorly on previously unseen formulations. Therefore, appropriate: Cross-validation, External validation, Independent test datasets, and Prospective experimental validation are essential.

9.5 Lack of standardized datasets

There is currently no universally standardized formulation-data structure covering all APIs, excipients, processes, and CQAs. Standardized data collection would greatly improve model development and reproducibility.

9.6 Biological complexity

A formulation may demonstrate excellent physicochemical properties but fail to produce the expected biological response. Therefore: **Good** computational prediction $\neq$ guaranteed therapeutic performance. AI models must ultimately be linked with experimental, pharmacokinetic, and biological evidence.

9.7 Regulatory considerations

Regulatory acceptance of AI-supported pharmaceutical development requires confidence in: There is currently no universally standardized formulation-data structure covering all APIs, excipients, processes, and CQAs. Standardized data collection would greatly improve model development and reproducibility. Data integrity, Model

development, Validation, Reproducibility, Change control, Model performance, Traceability, Human oversight. The pharmaceutical industry is therefore moving toward more structured approaches for integrating AI with formulation development rather than treating AI as an independent replacement for established development practices. A 2025 industry-focused review described current adoption as promising but still requiring further refinement and guidance.

X. ETHICAL AND SCIENTIFIC CONSIDERATIONS

AI implementation raises several important considerations: Data bias: Models trained using limited datasets may not generalize to different APIs, excipients, manufacturing processes, or patient populations. Transparency: Researchers should report:

- Dataset size
- Input variables
- Model architecture
- Training methodology
- Validation methodology
- Performance metrics
- External validation
- Experimental confirmation

Reproducibility: AI studies should provide sufficient information to enable independent researchers to reproduce the analysis wherever possible.

Human oversight: AI-generated recommendations should be scientifically reviewed by pharmaceutical scientists before experimental or manufacturing decisions are made.

XI. FUTURE PERSPECTIVES

The future of AI in pharmaceutical formulation is likely to involve closer integration between computational models and experimental science. Important future directions include:

11.1 Generative AI: Generative approaches may eventually assist in designing novel formulation compositions and excipient combinations rather than simply predicting the behavior of existing formulations.

11.2 Large language models: Large language models may support literature mining, extraction of formulation information, experimental protocol generation, and organization of pharmaceutical knowledge. Their role in drug delivery is still emerging. ScienceDirect

11.3 AI-integrated QbD: AI may increasingly become integrated with QbD workflows to support formulation-variable selection and design-space exploration.

11.4 Autonomous formulation development: The combination of AI, robotics, PAT, automated formulation equipment, and high-throughput analytical systems may enable closed-loop formulation optimization.

11.5 Personalized medicines: AI may facilitate formulation selection according to patient-specific characteristics, particularly when combined with 3D printing and other flexible manufacturing technologies.

11.6 Sustainable pharmaceutical development: AI could potentially reduce unnecessary experiments, material consumption, and development waste by improving experimental selection and process optimization.

11.7 Standardized pharmaceutical datasets: The development of standardized, high-quality, interoperable pharmaceutical formulation datasets will probably be one of the most important requirements for reliable AI implementation.

XII. PROPOSED FRAMEWORK FOR AI-ASSISTED FORMULATION DEVELOPMENT

A practical framework can be proposed as follows:

Stage 1: Define QTPP

Stage 2: Identify CQAs

Stage 3: Collect API, excipient and process data

Stage 4: Clean and standardize data

Stage 5: Select meaningful input features

Stage 6: Develop multiple ML models

Stage 7: Cross-validate and externally validate models

Stage 8: Identify optimal formulation region

Stage 9: Experimentally prepare predicted formulations

Stage 10: Compare predicted and experimental CQAs

Stage 11: Update/refine model

Stage 12: Establish formulation design space and control strategy

This approach maintains experimental pharmaceutical science at the center while using AI as a predictive and optimization tool.^{14,15}

XIII. CONCLUSION

Artificial intelligence and machine learning are emerging as important technologies in pharmaceutical formulation development. Their ability to process multidimensional datasets and model nonlinear relationships provides opportunities to improve formulation design, excipient selection, optimization, drug-release prediction, nanocarrier development, and process understanding. Recent developments have expanded the field from conventional predictive modelling toward Bayesian optimization, active learning, multi-objective optimization, digital twins, and self-driving laboratories. Nevertheless, the current evidence does not support viewing AI as an autonomous replacement for formulation scientists. Limited datasets, data heterogeneity, model interpretability, overfitting, lack of standardized reporting, biological complexity, and regulatory considerations remain important barriers. Current reviews consistently emphasize that AI should function as a complementary decision-support layer supported by rigorous experimental validation. Future progress will depend on high-quality standardized datasets, appropriate model validation, explainable AI, integration with QbD, automated experimentation, and multidisciplinary collaboration between formulation scientists, data scientists, engineers, analytical scientists, and regulatory experts. If these challenges are addressed, AI/ML could contribute to a more predictive, efficient, and scientifically rational approach to pharmaceutical formulation development.

REFERENCES

- [1] Machine learning directed drug formulation development. *Advanced Drug Delivery*

- [2] *Reviews*. 2021;175:113806. ScienceDirect
- [3] Bao Z, Bufton J, Hickman RJ, et al. Revolutionizing drug formulation development: The increasing impact of machine learning. *Advanced Drug Delivery Reviews*. 2023;202:115108. ScienceDirect
- [4] Advancing algorithmic drug product development: Recommendations for machine learning approaches in drug formulation. *European Journal of Pharmaceutical Sciences*. 2023;191:106562. ScienceDirect+1
- [5] Bachas, S., Rakocevic, G., Spencer, D., Sastry, A. V., Haile, R., & Sutton, J. M. (2022). Artificial intelligence predictions of binding affinity and naturalness enable antibody optimization. *bioRxiv* 2022, 504181. doi:10.1101/2022.08.16.504181
- [6] Brown, N., Fiscato, M., Segler, M. H. S., & Vaucher, A. C. (2019). GuacaMol: Benchmarking models for de novo molecular design. *J. Chem. Inf. Model.* 59 (3), 1096–1108. doi:10.1021/acs.jcim.8b00839
- [7] A novel paradigm on data and knowledge-driven drug formulation development: Opportunities and challenges of machine learning. *Journal of Industrial Information Integration*. 2025;44:100796. ScienceDirect
- [8] Artificial Intelligence (AI) in Pharmaceutical Formulation and Dosage Calculations. *Pharmaceutics*. 2025;17(11):1440. MDPI
- [9] Warke S, Katari O, Jain S. Current Status on the Convergence of Artificial Intelligence and Formulation Development in Industry: A Review. *AAPS PharmSciTech*. 2025;27(1):44. PubMed
- [10] Machine learning empowered formulation design, optimization and characterization of nanoparticulate drug delivery systems: Current applications, challenges, and future perspectives. *Acta Pharmaceutica Sinica B*. 2026;16(2):665–685. ScienceDirect
- [11] Artificial intelligence and machine learning guided optimization in drug delivery. *Advanced Drug Delivery Reviews*. 2026;232:115781. ScienceDirect
- [12] The growing impact of machine learning on drug formulation science. *Advanced Drug Delivery Reviews*. 2026;229:115761. ScienceDirect
- [13] Integrating artificial intelligence into drug delivery systems: Formulation development and current challenges. *European Journal of Pharmaceutics and Biopharmaceutics*. 2026;226:115133. ScienceDirect+1
- [14] Artificial intelligence for drug delivery: Yesterday, today and tomorrow. *Acta Pharmaceutica Sinica B*. 2026;16(4):2068–2092. ScienceDirect
- [15] Large language models in drug delivery: A review of the current landscape and future perspectives. *Journal of Pharmaceutical Sciences*. 2026;115(2):104147. ScienceDirect