ISSN: 2455-8109

International Journal of Farmacia (IJF)

IJF |Vol. 10 | Issue 3 | Jul – Sep - 2024

www.ijfjournal.com

DOI : https://doi.org/10.61096/ijf.v10.iss3.2024.81-95

Artificial Intelligence and Machine Learning in Pharmaceutical Formulation Development: Predictive Modeling, Optimization, and Future Perspectives

Kalloju Sri Sahethi, Sapavat.Vasantha, G.Madhuri

Scient Institute of Pharmacy
lbrahimpatnam, Ranga Reddy District - 501 506, Telangana
exambranch.ghip@gmail.com

Published on: 05.08.2024
Published by: Futuristic Publications

Abstract:

Background: The pharmaceutical development process is inherently complex, resource-intensive, and time-consuming, with formulation scientists navigating a multidimensional experimental space of drug properties, excipient choices, manufacturing process parameters, and quality attributes whose interrelationships are often non-linear and poorly understood from first principles alone. Artificial intelligence and machine learning offer a transformative shift in how this experimental space is explored, replacing exhaustive trial-and-error experimentation with data-driven predictive models that can identify optimal formulation compositions, predict quality attributes from compositional inputs, and guide experimental design toward the most informative experiments from a sparse dataset.

Objective: This review critically examines the application of artificial intelligence and machine learning methodologies — including artificial neural networks, random forest, support vector machines, Gaussian process regression, deep learning, and generative models — to pharmaceutical formulation development, covering solubility prediction, dissolution modeling, stability forecasting, nanoparticle optimization, and integration with Quality by Design frameworks, based on literature published up to 2023.

Results and Discussion: Machine learning models trained on physicochemical descriptors have achieved prediction accuracy for aqueous solubility, intestinal permeability, and oral bioavailability comparable to or exceeding mechanistic models, with random forest and deep neural network architectures consistently outperforming linear regression and multiple regression approaches on diverse pharmaceutical datasets. Bayesian optimization has emerged as the most efficient strategy for formulation parameter optimization with minimal experimental trials. Natural language processing applied to pharmaceutical literature enables automated extraction of formulation knowledge at a scale inaccessible to manual review.

Conclusion: Artificial intelligence and machine learning are rapidly transitioning from exploratory pharmaceutical research tools to integral components of industrial formulation development pipelines, with demonstrated capabilities in predictive modeling, process optimization, and quality by design implementation that reduce development timelines and experimental burden. The interpretability of models, data quality requirements, and regulatory acceptance of AI-assisted development decisions are the primary barriers requiring systematic resolution.

Keywords: *Artificial intelligence; machine learning; pharmaceutical formulation; neural networks; Bayesian optimization; QSAR; Quality by Design; dissolution prediction; solubility prediction; deep learning*

1. Introduction

Pharmaceutical formulation development involves navigating a complex, high-dimensional experimental space in which the properties of the active pharmaceutical ingredient interact with excipient composition, manufacturing process parameters, and environmental storage conditions to determine the quality, safety, and efficacy of the final drug product. A typical solid dosage form formulation study may involve 10 to 20 excipient variables each at multiple levels, with critical quality attributes including dissolution rate, mechanical strength, content uniformity, and stability to be optimized simultaneously. The traditional approach of one-factor-at-a-time experimentation is woefully inadequate for this dimensionality, while even Design of Experiments approaches based on response surface methodology struggle when the number of variables exceeds five to eight due to the exponential growth in experimental runs required [1,2].

Artificial intelligence encompasses a broad set of computational approaches that enable machines to perform tasks that would ordinarily require human intelligence, including learning from data, identifying patterns, making predictions, and optimizing complex systems. Machine learning, as the dominant practical subset of AI in pharmaceutical applications, focuses specifically on the development of algorithms that improve their performance on a defined task as they are exposed to more training data, without being explicitly programmed with domain-specific rules. This data-driven learning capability makes machine learning particularly well-suited to pharmaceutical formulation problems, where the governing relationships between formulation composition and quality attributes are too complex and context-dependent to be fully described by mechanistic equations [3,4].

The pharmaceutical industry's engagement with machine learning has accelerated substantially over the past decade, driven by the convergence of three enabling factors: the accumulation of large pharmaceutical datasets from historical development programs, regulatory submissions, and published literature; the dramatic increase in computational power available through cloud computing and graphical processing unit acceleration; and the maturation of open-source machine learning software libraries including TensorFlow, PyTorch, and scikit-learn that have made sophisticated model development accessible to pharmaceutical scientists without deep computer science expertise. The publication in 2019 of the FDA's discussion paper on the use of AI and machine learning in drug development signaled regulatory acknowledgment of these methods as legitimate components of pharmaceutical development [5,6].

Beyond formulation optimization, machine learning is being applied across the pharmaceutical development continuum: in preformulation for prediction of physicochemical properties from molecular structure; in manufacturing for real-time process monitoring and control; in stability prediction to identify degradation pathways and accelerate shelf-life studies; in clinical pharmacokinetics for patient-specific dose optimization; and in drug discovery for molecular property prediction and de novo molecular generation. This review focuses specifically on formulation development applications, while acknowledging the interdependencies with upstream drug discovery and downstream manufacturing contexts [7].

This review provides a systematic and critically appraised examination of machine learning methodologies as applied to pharmaceutical formulation development, covering algorithm types and their pharmaceutical applications, data requirements and preprocessing strategies, model validation approaches, integration with Quality by Design frameworks, and specific applications in solubility enhancement, dissolution prediction, nanoparticle formulation optimization, stability prediction, and natural language processing for formulation knowledge extraction. A critical analysis identifies the most significant limitations and misapplications of machine learning in the pharmaceutical literature.

2. Scientific Background

2.1 Fundamentals of Machine Learning Relevant to Pharmaceutical Science

Machine learning algorithms are broadly divided into supervised learning, in which a model is trained on labeled input-output pairs to predict outputs for new inputs; unsupervised learning, in which patterns and structure are identified in unlabeled data without predefined output categories; and reinforcement learning, in which an agent learns to make sequential decisions by maximizing cumulative reward from an environment. Pharmaceutical formulation applications predominantly employ supervised learning — predicting drug dissolution from formulation composition, classifying formulations as stable or unstable from stress test data, or predicting API solubility from molecular descriptors — though unsupervised clustering of formulation datasets and reinforcement learning for sequential experimental design are emerging applications [8,9].

The supervised learning workflow consists of dataset curation and cleaning, feature engineering and selection, model selection and hyperparameter tuning, training and cross-validation, and performance

evaluation on a held-out test set. Feature engineering — the transformation of raw pharmaceutical data (molecular structure, formulation composition, process parameters) into numerical representations suitable for model training — is particularly consequential in pharmaceutical applications, where the choice between molecular descriptors (topological, electronic, steric, physicochemical) substantially determines model accuracy. Molecular fingerprints including Morgan (circular) fingerprints and MACCS keys, physicochemical descriptors generated by RDKit or Mordred, and graph-based molecular representations processed by graph neural networks are the primary input representations for structure-property models [10,11].

2.2 Pharmaceutical Data Landscape

The performance of machine learning models in pharmaceutical applications is critically dependent on the size, quality, representativeness, and curation standards of the training dataset. Public pharmaceutical databases including ChEMBL, PubChem, DrugBank, PhysProp, and the ESOL dataset provide tens of thousands of drug-like molecules with measured physicochemical properties that serve as training data for solubility, permeability, and stability prediction models. However, these public databases suffer from measurement inconsistency across laboratories and time periods, variable data quality, and sparse coverage of the chemical space relevant to modern pharmaceutical development, particularly for poorly soluble BCS class II and IV compounds that represent the majority of current drug candidates [12,13].

Industrial pharmaceutical formulation datasets, generated within companies' historical development programs and regulatory submissions, are considerably more internally consistent and pharmacokinetically relevant than public databases, but are proprietary and unavailable for academic model development. This data access asymmetry means that academic machine learning models trained on public data may perform worse than industry models on practically relevant pharmaceutical prediction tasks, while industry models cannot be subjected to independent scientific validation. Federated learning approaches, in which models are trained collaboratively across multiple institutional datasets without sharing raw data, have been proposed as a potential solution to this problem [14].

3. Classification of Machine Learning Methods in Pharmaceutical Formulation

3.1 Traditional Machine Learning Algorithms

Artificial neural networks (ANNs), in their classical multilayer perceptron form with one to three hidden layers, were the first machine learning algorithms to be applied to pharmaceutical formulation problems in the 1990s and remain widely used for their ability to learn nonlinear relationships between formulation variables and quality attributes. ANNs have been applied to prediction of tablet hardness, disintegration time, drug dissolution profiles, and granule flowability from formulation composition and process parameters, consistently outperforming polynomial regression and response surface models on datasets with non-linear variable interactions [15].

Random forest is an ensemble method that constructs a large number of decision trees on bootstrap samples of the training data and aggregates their predictions, reducing overfitting through the averaging effect of many decorrelated trees. Random forest provides both high predictive accuracy and inherent feature importance ranking, enabling identification of the most influential formulation variables without requiring a priori knowledge of their relative importance. It has demonstrated superior performance to ANNs on small-to-medium pharmaceutical datasets (50 to 500 samples) owing to its reduced tendency to overfit on limited data [16,17].

Support vector machines learn a hyperplane in a high-dimensional feature space that maximally separates classes (for classification) or fits training data within a margin (for regression), using kernel functions to handle non-linear relationships without explicitly computing the high-dimensional mapping. SVMs have been applied to prediction of drug-excipient compatibility, classification of solubility classes for BCS biowaivers, and prediction of tablet punch sticking propensity from API and excipient surface energy descriptors [18].

3.2 Deep Learning Architectures

Deep neural networks with many hidden layers (typically five to hundreds) have achieved state-of-the-art performance on pharmaceutical property prediction tasks through their capacity to learn hierarchical feature representations from raw molecular structure inputs without hand-crafted feature engineering. Convolutional neural networks operating on molecular graph representations, recurrent neural networks processing SMILES string representations of molecular structure, and attention-based transformer models pretrained on large chemical databases and fine-tuned on pharmaceutical property datasets have been applied to prediction of aqueous solubility, membrane permeability, metabolic stability, and hERG channel inhibition [19,20].

Graph neural networks represent molecules as graphs in which atoms are nodes and bonds are edges, and learn molecular property predictions by iteratively aggregating and transforming node features from neighboring nodes. This architecture naturally captures the local chemical environment around each atom that determines reactivity, solubility, and other physicochemical properties, and has demonstrated superior accuracy on solubility and bioavailability prediction tasks compared to fingerprint-based models, particularly for structurally diverse molecular datasets [21].

3.3 Bayesian Optimization and Active Learning

Bayesian optimization is a sequential model-based experimental design strategy that uses a probabilistic surrogate model — typically a Gaussian process — to represent uncertainty in the objective function (e.g., dissolution rate, bioavailability, stability) across the formulation design space. At each iteration, an acquisition function balancing exploration of uncertain regions with exploitation of promising regions guides selection of the next experiment to perform. This strategy is particularly valuable in pharmaceutical formulation optimization because it achieves near-optimal formulation identification with dramatically fewer experiments than factorial or response surface designs, typically requiring 20 to 50 experiments to optimize a 5 to 10 variable formulation space [22,23].

3.4 Natural Language Processing

Natural language processing enables extraction of structured pharmaceutical knowledge from unstructured text in published literature, regulatory submissions, and patent databases. Named entity recognition models trained on annotated pharmaceutical text identify drug names, excipient names, process parameters, and quality attributes in sentences, while relation extraction models identify the semantic relationships between these entities (e.g., excipient A at concentration B improves property C). Applied to the PubMed database of pharmaceutical publications, NLP-based knowledge extraction can rapidly survey the formulation literature for a given drug or excipient combination at a scale and speed inaccessible to manual review, potentially identifying overlooked excipient interactions or formulation approaches from historical literature [24].

Table 1: Classification of machine learning algorithms applied in pharmaceutical formulation with dataset requirements, strengths, and pharmaceutical applications

Algorithm	Type	Dataset Size	Key Strengths	Pharmaceutical Applications
Artificial neural network (ANN)	Supervised (regression/classification)	≥100 samples	Non-linear learning, flexible	Dissolution, tablet hardness, granulation
Random forest	Supervised (ensemble)	50–1000 samples	Feature importance, robust small data	Solubility, stability, bioavailability
Support vector machine	Supervised (kernel-based)	50–500 samples	High-dimensional, margin optimization	Drug-excipient compatibility, BCS class
Gaussian process regression	Probabilistic regression	10–200 samples	Uncertainty quantification, small data	Bayesian optimization of formulations
Deep neural network / GNN	Supervised (deep learning)	>1000 samples	Molecular representation learning	Solubility, permeability, ADMET prediction
Convolutional neural network	Supervised (image/sequence)	>500 samples	Tablet image quality, spectral analysis	Tablet defect detection, NIR/Raman analysis
Natural language processing	Unsupervised/supervised text	Large text corpora	Knowledge extraction from literature	Excipient selection, formulation mining

GNN: graph neural network; BCS: Biopharmaceutics Classification System; ADMET: absorption, distribution, metabolism, excretion, toxicity; NIR: near-infrared.

4. Data Strategy and Feature Engineering for Pharmaceutical ML Models

4.1 Molecular Descriptor Generation

The transformation of drug molecular structure into numerical features suitable for machine learning input is a foundational step in pharmaceutical property prediction. Two-dimensional topological descriptors encode the graph structure of the molecular connectivity, including molecular weight, ring count, hydrogen bond donor and acceptor counts, rotatable bond count, and topological polar surface area. These simple descriptors, readily calculated from SMILES notation, capture pharmacokinetically relevant molecular properties and serve as inputs for Lipinski rule-based classification and BCS solubility class prediction models. Extended connectivity fingerprints (ECFP), also termed Morgan fingerprints, represent the local chemical environment of each atom as a binary bit vector and are among the most widely used molecular representations for structure-property machine learning models [25,26].

Three-dimensional molecular descriptors encoding conformational and electronic properties including molecular volume, surface area, dipole moment, and electron density distribution provide complementary information to 2D descriptors, particularly for properties dependent on molecular shape and charge distribution such as crystal packing energy (relevant to polymorphism and solubility) and membrane permeation. However, 3D descriptor calculation requires conformer generation and energy minimization, which introduces conformational uncertainty and increases computational cost relative to 2D descriptor approaches [27].

4.2 Formulation Composition Encoding

Encoding formulation composition as model inputs requires representation of both qualitative (excipient identity) and quantitative (excipient concentration) variables in a format that captures pharmaceutical relevance. One-hot encoding of excipient identity, combined with quantitative concentration values, is the most straightforward approach but fails to capture the physicochemical relationships between chemically similar excipients. Encoding excipients by their physicochemical properties — solubility parameter, glass transition temperature, molecular weight, viscosity grade — enables generalization across structurally related excipients and potentially allows prediction for excipient-drug combinations not represented in the training data [28].

4.3 Data Augmentation and Transfer Learning

The limited size of most pharmaceutical formulation datasets — typically 50 to 500 observations for a given drug product type — constrains the complexity of machine learning models that can be trained without overfitting. Data augmentation strategies including SMOTE (synthetic minority oversampling technique) for imbalanced classification datasets, Monte Carlo sampling of multivariate normal distributions fitted to existing data, and physics-based simulation of additional datapoints using thermodynamic models have been employed to artificially expand pharmaceutical datasets. Transfer learning — initializing model weights from a model pretrained on a large related dataset, then fine-tuning on the small pharmaceutical dataset — has been applied to pharmaceutical solubility and dissolution prediction using models pretrained on ChEMBL bioactivity data [29,30].

5. Model Development, Training, and Validation

5.1 Cross-Validation and Generalization Assessment

The fundamental goal of machine learning model validation in pharmaceutical science is to estimate how accurately a model will predict quality attributes for new, unseen formulations, not merely to assess how well it reproduces training data. k-fold cross-validation, in which the dataset is partitioned into k subsets with each serving in turn as the validation set while the remaining k-1 subsets are used for training, provides an unbiased estimate of generalization performance for small pharmaceutical datasets where a separate test set would be too small to be statistically meaningful. For pharmaceutical QSAR models, scaffold-based splitting — which assigns molecules with similar core scaffolds to the same fold — provides a more rigorous assessment of model generalization to structurally novel compounds than random splitting, which may produce optimistically biased estimates when structurally similar molecules appear in both training and test sets [31,32].

5.2 Hyperparameter Optimization

Machine learning model hyperparameters — the model configuration settings that are not learned from training data, such as neural network architecture (layer count, neuron count, activation functions), random forest tree

count and maximum depth, and SVM kernel type and regularization parameter — must be optimized using validation data separate from the test set to avoid overfitting. Grid search evaluates all combinations of specified hyperparameter values but becomes computationally intractable for large hyperparameter spaces. Random search evaluates random samples from the hyperparameter space and has been shown to be more efficient than grid search for high-dimensional hyperparameter optimization. Bayesian hyperparameter optimization using Gaussian processes to model the performance surface of the hyperparameter space identifies optimal configurations more efficiently than both grid and random search [33].

5.3 Model Interpretability and Feature Importance

The interpretability of machine learning models — the ability to understand why a model makes a particular prediction in terms of the input features — is a critical requirement for pharmaceutical formulation applications, where the model output must guide actionable formulation decisions and regulatory documentation. Random forest provides inherent feature importance scores based on the mean decrease in prediction accuracy when each feature is permuted. SHAP (SHapley Additive exPlanations) values, derived from cooperative game theory, provide theoretically grounded local interpretability by assigning each input feature a contribution to each individual prediction, enabling identification of which formulation variables most influence a specific quality attribute prediction for a given formulation [34,35].

5.4 Applicability Domain Assessment

Machine learning models are only reliable for predictions within the chemical or formulation space represented in the training data — the model's applicability domain. Predictions for inputs outside the applicability domain may be confidently erroneous, providing a false impression of reliability that can misdirect formulation development. Applicability domain assessment methods include k-nearest neighbor distance in descriptor space (predictions for points far from any training point are flagged as out-of-domain), principal component analysis-based bounding box methods, and leverage-based Williams plot approaches. Reporting applicability domain alongside model predictions should be considered a minimum standard for pharmaceutical machine learning publications, though it remains inconsistently applied in the literature [36].

6. Model Performance Metrics and Benchmarking

6.1 Regression Performance Metrics

For continuous pharmaceutical property predictions including dissolution rate, solubility, and stability parameters, model performance is assessed by the root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2) between predicted and observed values on the test set. RMSE is the most commonly reported metric and penalizes large prediction errors more heavily than MAE through its squared error term, making it sensitive to outliers. The concordance correlation coefficient (CCC) assesses both the precision and accuracy of predictions and is preferred over R^2 for pharmaceutical model comparison as it requires predictions to lie on the line of identity rather than any line of slope one [37].

6.2 Classification Performance Metrics

For pharmaceutical classification problems including BCS class assignment, formulation stability classification, and drug-excipient compatibility prediction, the area under the receiver operating characteristic curve (AUC-ROC), Matthews correlation coefficient, balanced accuracy, and F1 score provide complementary assessments of classifier performance that account for class imbalance — a frequent challenge in pharmaceutical datasets where stable formulations substantially outnumber unstable ones in historical development databases. The AUC-ROC measures the probability that the model ranks a randomly chosen positive example above a randomly chosen negative example, providing a threshold-independent performance measure [38].

6.3 Prospective Validation

The gold standard for pharmaceutical machine learning model validation is prospective experimental confirmation of model predictions, in which the model is used to select candidate formulations for experimental preparation and testing after the model is trained and before the experimental results are known. Prospective validation distinguishes models with genuine predictive capability from those that perform well on retrospective cross-validation but fail to identify practically useful formulations in real experimental campaigns. Several published pharmaceutical formulation studies have demonstrated prospective validation of machine learning-guided formulation selection, reporting significant reduction in the number of experiments required to identify an optimal formulation compared to traditional DoE approaches [39].

7. Integration of Machine Learning with Quality by Design

7.1 Design Space Exploration

The Quality by Design framework established by ICH Q8 requires pharmaceutical developers to define a design space — the multidimensional combination and interaction of process and formulation variables that provide assurance of quality — based on understanding of the relationships between critical material attributes, critical process parameters, and critical quality attributes. Machine learning accelerates design space exploration by predicting CQA values across the full formulation space from a subset of experimentally evaluated points, enabling identification of the design space boundary without exhaustively testing every combination. Gaussian process regression is particularly suited to this application owing to its ability to provide uncertainty estimates alongside predictions, enabling probabilistic design space boundaries that characterize both the mean predicted CQA values and the confidence in those predictions [40,41].

7.2 Real-Time Release Testing and Process Monitoring

Machine learning models trained on spectroscopic data — near-infrared, Raman, or terahertz spectra collected during manufacturing — enable real-time prediction of tablet CQAs including drug content, dissolution rate, hardness, and moisture content without destructive off-line testing. Partial least squares regression, principal component regression, and artificial neural networks have been extensively applied to pharmaceutical NIR calibration model development for content uniformity monitoring in continuous blending. Convolutional neural networks applied to NIR spectra in continuous tablet manufacturing have demonstrated real-time prediction of blend homogeneity with accuracy sufficient for real-time release testing, supporting transition from end-product testing to parametric release [42,43].

7.3 Predictive Stability Modeling

Machine learning models for pharmaceutical stability prediction use accelerated stability data at elevated temperatures and humidity to construct predictive models that extrapolate to long-term stability at intended storage conditions, reducing the time required to establish shelf life. Neural networks and random forest models trained on accelerated degradation kinetics data have been used to predict the identity and relative abundance of degradation products at long-term storage conditions based on stress testing results, enabling early identification of potential stability failures and optimization of formulation composition and packaging to minimize degradation [44].

8. Pharmaceutical Applications of Machine Learning

8.1 Aqueous Solubility Prediction

Aqueous solubility prediction from molecular structure is one of the longest-standing and most extensively benchmarked applications of machine learning in pharmaceutical science, with published models dating to the early 1990s. The ESOL dataset compiled by Delaney, containing experimentally measured log solubility values for 1144 structurally diverse compounds, has become a standard benchmark for comparing model architectures. Deep learning models including directed message passing neural networks and graph attention networks have achieved RMSE values of approximately 0.5 to 0.7 log units on the ESOL benchmark, compared to RMSE values of 0.8 to 1.2 log units for classical linear models. Random forest models trained on RDKit physicochemical descriptors consistently achieve performance competitive with deep learning architectures on this dataset, suggesting that the additional complexity of deep learning may not be warranted for general solubility prediction on small diverse datasets [45,46].

8.2 Dissolution Profile Prediction and IVIVC

Machine learning prediction of drug dissolution profiles from formulation composition and process parameters addresses a central challenge in formulation development: identifying the formulation variables and their levels that produce a dissolution profile meeting the target release specification. ANN models trained on formulation composition and process parameters (tablet hardness, granule size, moisture content) have demonstrated accurate prediction of f2 similarity factors and individual dissolution time points for immediate-release, extended-release, and controlled-release tablet formulations. Integration of machine learning dissolution models with physiologically based pharmacokinetic (PBPK) models enables in vitro-in vivo correlation development without requiring clinical pharmacokinetic studies at every formulation variant, substantially reducing the experimental burden of bioequivalence assessment [47,48].

8.3 Nanoparticle and Lipid Nanoparticle Formulation Optimization

Machine learning optimization of nanoparticle formulations has emerged as a particularly productive application area, driven by the large number of formulation and process variables (lipid composition, drug-to-lipid ratio, flow rate ratio, mixing speed, temperature, PEG density) that influence the critical quality attributes (particle size, PDI, encapsulation efficiency, zeta potential) of PLGA nanoparticles, solid lipid nanoparticles, and lipid nanoparticles. Random forest and ANN models trained on microfluidic LNP production datasets achieved accurate prediction of particle size and encapsulation efficiency from lipid composition and process parameters, with Bayesian optimization using Gaussian process surrogates subsequently identifying optimal formulations for siRNA encapsulation in fewer than 20 experiments [49,50].

8.4 Excipient Selection and Drug-Excipient Compatibility

Machine learning for excipient selection uses databases of excipient physicochemical properties, drug-excipient interaction outcomes, and formulation performance data to recommend excipients for a given drug molecule without requiring extensive compatibility studies. Support vector machine classifiers trained on drug and excipient molecular descriptors together with thermal analysis-derived compatibility outcomes demonstrated accurate prediction of incompatibility between drugs and excipients in a 1000-compound training dataset, achieving AUC-ROC values above 0.85 on held-out test sets. These predictions, combined with the uncertainty quantification of Gaussian process classifiers, could prioritize the most risky drug-excipient combinations for experimental compatibility testing while confidently accepting compatible combinations without testing [51].

8.5 Amorphous Solid Dispersion Development

Machine learning has been applied to several key challenges in amorphous solid dispersion development, including prediction of drug-polymer miscibility, glass transition temperature of drug-polymer blends, physical stability of amorphous dispersions during storage, and supersaturation duration in biorelevant dissolution media. Random forest models trained on drug and polymer physicochemical descriptors (Flory-Huggins interaction parameter, molecular flexibility, glass-forming ability) predicted drug-polymer miscibility with 80 to 90% accuracy on diverse drug-polymer combinations, guiding polymer selection for hot-melt extrusion and spray-drying without exhaustive experimental screening [52,53].

9. Recent Advances

9.1 Generative AI for Formulation Design

Generative adversarial networks and variational autoencoders applied to pharmaceutical formulation data generate novel formulation compositions with predicted properties in regions of the formulation space not represented in the training data, enabling de novo formulation design rather than interpolation within existing formulations. A generative model trained on immediate-release tablet formulation databases generated candidate formulations with predicted dissolution profiles meeting specified target product profiles in a study published in 2022, with a subset of AI-generated formulations confirmed experimentally to meet the dissolution specification in prospective validation. This generative approach to formulation design represents a paradigm shift from optimization within a defined experimental space to generation of fundamentally new formulation concepts guided by desired property targets [54].

9.2 Large Language Models for Pharmaceutical Knowledge

The emergence of large language models with billions of parameters trained on vast text corpora has created new opportunities for pharmaceutical knowledge extraction and synthesis. Models including GPT-3, GPT-4, and domain-specific pharmaceutical variants trained on scientific literature databases have demonstrated capability to answer pharmaceutical formulation questions, summarize complex formulation studies, identify relevant excipients from text descriptions, and generate structured formulation development reports from unstructured experimental notes. The accuracy and reliability of large language model outputs for pharmaceutical applications remain subjects of active investigation, with particular concerns about hallucination of plausible-sounding but factually incorrect information in pharmaceutical contexts where accuracy is clinically consequential [55].

9.3 Multi-Task Learning for ADMET Prediction

Multi-task learning trains a single neural network model to simultaneously predict multiple related pharmaceutical properties — aqueous solubility, membrane permeability, metabolic stability, hERG

inhibition, and plasma protein binding — sharing learned representations across tasks. This approach leverages the mechanistic relationships between these properties (e.g., a drug's lipophilicity influences both membrane permeability and metabolic stability) to improve prediction accuracy for tasks with limited training data by borrowing statistical strength from better-resourced related tasks. Multi-task deep learning models trained on ADMET property datasets from ChEMBL and PubChem Bioassay have demonstrated improved prediction accuracy for underrepresented property endpoints compared to single-task models trained on the same data [56].

9.4 Federated Learning for Cross-Institutional Pharmaceutical Collaboration

Federated learning enables training of machine learning models collaboratively across multiple institutions without centralizing or sharing raw pharmaceutical data, addressing the proprietary data access problem that constrains pharmaceutical ML model development. In a federated pharmaceutical stability prediction study, models trained collaboratively across four pharmaceutical companies using federated averaging of locally computed model updates achieved prediction accuracy comparable to models trained on the combined dataset, demonstrating that data-sharing restrictions need not prevent cross-institutional collaborative model development. The regulatory implications of federated learning for pharmaceutical model validation — particularly the question of how a model trained on distributed data that no single party can fully access is validated — are being actively explored by regulatory agencies [57].

10. Comparative Analysis

The comparison of machine learning approaches with established pharmaceutical formulation optimization methods — one-factor-at-a-time experimentation, factorial design, response surface methodology, and mechanistic modeling — reveals the contexts in which each approach provides superior value and those in which traditional methods remain preferable or necessary. One-factor-at-a-time experimentation is unable to detect or model interactions between formulation variables and is inefficient in terms of experiments per unit of information gained, but remains widely practiced due to its conceptual simplicity. Factorial and response surface designs provide statistically rigorous characterization of variable interactions within a defined experimental space but require a priori specification of the variable space and become practically unmanageable for more than five to eight variables [58].

Machine learning offers advantages over DoE in its ability to learn non-linear variable interactions of arbitrary complexity, its applicability to datasets of irregular structure (historical data with missing values, observational data without designed variation), and its scalability to higher-dimensional variable spaces than DoE methods. DoE retains advantages over machine learning in its well-established statistical theory for confidence interval estimation and hypothesis testing, its compatibility with regulatory expectations for formulation development documentation, and its ability to provide mechanistic insight through analysis of variance and effect estimates. Mechanistic models (PBPK, drug release equations, thermodynamic solubility models) provide physically interpretable predictions that generalize to conditions outside the training data range, a capability that pure empirical machine learning models cannot match [59].

Table 2: Comparative analysis of machine learning versus traditional pharmaceutical formulation optimization approaches

Parameter	OFAT	DoE / RSM	Bayesian Optimization	Machine Learning	Mechanistic Model
Variable interactions	Not detected	Pairwise (RSM)	Fully captured	Fully captured	Explicit
Dimensionality (no. of variables)	Low (1–3)	Moderate (3–8)	High (5–20)	High (>10)	Moderate
Experiments required	Many	Moderate	Few (20–50)	Depends on data	Few
Interpretability	High	High	Moderate	Low–moderate	High
Regulatory acceptance	Established	Well established	Emerging	Emerging	Well established
Extrapolation beyond data	None	Limited	Limited	Poor	Yes (physics-based)

OFAT: one-factor-at-a-time; DoE: Design of Experiments; RSM: response surface methodology.

11. Advantages and Limitations

11.1 Advantages

- Capacity to learn non-linear, high-dimensional relationships between formulation composition, process parameters, and quality attributes that cannot be captured by linear or polynomial regression models
- Bayesian optimization enables identification of optimal formulations with substantially fewer experiments than factorial or response surface designs, reducing material consumption, time, and cost in formulation development
- Predictive models trained on historical formulation data enable rapid screening of candidate formulations in silico before committing to physical experimentation, accelerating the early stages of formulation development
- Feature importance analysis identifies the most influential formulation variables, guiding experimental attention toward the critical few rather than the trivial many and enabling more efficient development programs
- NLP-based extraction of formulation knowledge from published literature and patent databases provides systematic access to historical formulation science at a scale and speed inaccessible to manual literature review
- Real-time process monitoring models trained on inline spectroscopic data enable continuous manufacturing quality control and support the transition from end-product testing to parametric real-time release
- Transfer learning from large chemical databases reduces the data requirements for pharmaceutical property prediction models, enabling accurate predictions from small pharmaceutical-specific datasets

11.2 Limitations

- Machine learning model performance is fundamentally constrained by the size, quality, and representativeness of training data — pharmaceutical formulation datasets are typically small (50 to 500 observations) relative to the dimensionality of the formulation space, risking overfitting and poor generalization
- Most machine learning models are empirical black boxes that provide no mechanistic insight into the physicochemical reasons for observed formulation-property relationships, limiting their utility for rational formulation design beyond interpolation
- Reliable predictions are only generated within the model's applicability domain — predictions for formulations or drug molecules outside the chemical or compositional space of the training data are unreliable but may not be flagged as such by naive model implementations
- The pharmaceutical literature contains numerous examples of machine learning models with overly optimistic performance reported due to data leakage, improper cross-validation (random rather than scaffold-based splitting), or failure to account for temporal bias in historical datasets
- Regulatory agencies have not yet established clear guidance on the validation requirements for machine learning models used in pharmaceutical development decisions, creating uncertainty about the documentation standards expected for submissions incorporating ML-assisted development
- The expertise required to develop, validate, and critically interpret pharmaceutical machine learning models spans molecular informatics, statistics, pharmaceutical science, and software engineering — a breadth of skills rarely concentrated in a single pharmaceutical scientist or team

12. Industrial Implementation and Case Studies

The industrial adoption of machine learning in pharmaceutical formulation development has advanced from exploratory research applications toward integration into standard formulation development workflows in several major pharmaceutical companies by 2023. Pfizer reported the use of machine learning-guided formulation optimization for the development of tablet formulations for multiple drug candidates, with Bayesian optimization reducing the number of experimental formulations required by 30 to 50% compared to traditional DoE approaches in internal benchmarking studies. AstraZeneca developed an in-house

platform for machine learning-guided excipient selection and compatibility prediction, trained on decades of internal formulation development data from across their portfolio [60].

Regulatory agencies have begun to engage actively with the use of AI in pharmaceutical development. The FDA published a discussion paper in 2019 outlining a framework for AI and ML-based software as a medical device, and has received multiple investigational new drug and new drug applications incorporating machine learning-generated data or predictions in support of formulation development decisions. The EMA published a reflection paper on the use of AI in the lifecycle of medicines in 2023, acknowledging both the potential benefits and the validation challenges of ML-assisted pharmaceutical development and recommending a risk-proportionate approach to regulatory expectations [61].

Table 3: Selected pharmaceutical industry applications and case studies of machine learning in formulation development

Company / Group	ML Method	Application	Dataset / Scale	Reported Outcome
Pfizer / AstraZeneca	Bayesian optimization	Tablet formulation optimization	Internal historical data (500+ runs)	30–50% fewer experiments to optimum
Merck KGaA	Random forest + ANN	Dissolution profile prediction	Multi-product tablet database	RMSE < 5% for f2 prediction
MIT / Novartis collaboration	Deep neural network	Continuous manufacturing optimization	Pilot plant production data	Real-time CQA prediction for RTR
Stanford / academic groups	GNN (graph neural network)	Aqueous solubility prediction	ESOL + AqSolDB (10,000+ cpds)	RMSE 0.5–0.7 log S units (ESOL)
Johnson & Johnson	NLP (BERT-based)	Formulation knowledge extraction	PubMed pharmaceutical literature	Automated excipient interaction mining
Multi-company federated (Owkin)	Federated learning	Stability prediction across sites	4-site federated dataset	Accuracy comparable to centralized data

RTR: real-time release; CQA: critical quality attribute; RMSE: root mean squared error; GNN: graph neural network; NLP: natural language processing.

13. Critical Analysis

A critical appraisal of the pharmaceutical machine learning literature reveals several systematic methodological deficiencies that substantially limit the credibility and practical utility of published models. The most prevalent and consequential issue is the use of random train-test splitting in studies evaluating structure-property models, which allows structurally similar molecules to appear in both training and test sets. Given that pharmaceutical activity databases including ChEMBL and ESOL contain multiple closely related analogs with similar property values, random splitting produces test sets that are highly similar to training sets and therefore yield optimistically biased performance estimates. Studies comparing random versus scaffold-based or temporal test set construction have reported that performance metrics for solubility and bioavailability models drop by 0.1 to 0.3 RMSE units when properly dissimilar test sets are used, a difference large enough to change conclusions about model utility for prospective applications [62].

The publication bias in pharmaceutical machine learning is particularly pronounced: studies reporting machine learning models with high prediction accuracy are substantially more likely to be published than studies reporting negative results or models with marginal performance, creating a distorted impression of the field's overall predictive capability. Multiple independent benchmarking studies that have systematically compared many models on the same dataset have found that the best-performing model on any given dataset is often highly dataset-specific, with different model architectures winning on different pharmaceutical property prediction tasks. The implication is that claims of universal superiority for any particular machine learning architecture should be treated with considerable skepticism in the absence of prospective multi-dataset benchmarking [63].

The conflation of retrospective cross-validation performance with prospective utility is perhaps the most practically important methodological gap in the pharmaceutical machine learning literature. A model with high cross-validation R^2 may still fail to identify the best formulation in a prospective experimental

campaign if its prediction errors are systematically correlated with the experimental design variables used to train it, or if the optimal formulation lies outside the training data manifold. The published literature contains very few examples of genuinely prospective validation studies in which the ML model is trained, the model predictions are used to select experiments, and the experimental results are then compared against what traditional DoE would have found — the type of study that would most directly demonstrate the practical value of ML-assisted formulation development [64].

Finally, the pharmaceutical machine learning community has not yet reached consensus on the minimum dataset requirements for models of different types to generalize reliably. Rules of thumb such as requiring at least 10 training observations per model parameter, or at least 50 to 100 observations per input feature, are widely cited but have not been systematically validated for pharmaceutical formulation datasets whose variable structure differs substantially from the image and natural language datasets on which these rules were originally developed. Systematic empirical investigation of the sample size requirements for reliable pharmaceutical ML models, stratified by dataset type, model architecture, and prediction task, would be a valuable contribution to the field [65].

14. Conclusion

Artificial intelligence and machine learning have established themselves as genuinely useful tools in pharmaceutical formulation development, with demonstrated capabilities in aqueous solubility prediction, dissolution profile modeling, nanoparticle formulation optimization, excipient selection, stability prediction, and real-time process monitoring that offer practical value beyond what traditional statistical and mechanistic approaches alone can provide. The integration of Bayesian optimization with pharmaceutical formulation screening has shown the most convincing prospective validation, with multiple studies demonstrating identification of optimal or near-optimal formulations with significantly fewer experimental trials than conventional screening approaches.

The adoption of machine learning in industrial pharmaceutical formulation development has accelerated considerably in the period leading up to 2023, transitioning from academic research demonstrations to integrated workflows within major pharmaceutical companies' formulation development platforms. Regulatory agencies have begun to develop frameworks for evaluating AI-assisted pharmaceutical development decisions, signaling institutional readiness to accept ML-supported submissions when appropriate validation evidence is provided.

A critical and dispassionate reading of the pharmaceutical machine learning literature, however, reveals that the field is characterized by systematic methodological deficiencies — improper test set construction, publication bias toward positive results, conflation of retrospective cross-validation with prospective utility, and insufficient attention to applicability domain — that collectively inflate the apparent predictive power of published models relative to what would be achieved in practice. Addressing these deficiencies through community-wide adoption of rigorous benchmarking standards, mandatory prospective validation for clinically consequential models, transparent reporting of applicability domain, and open sharing of pharmaceutical formulation datasets is essential to ensure that the genuine promise of AI-assisted pharmaceutical formulation development is realized rather than overstated.

References

1. Prausnitz MR, Langer R. Transdermal drug delivery. Nat Biotechnol. 2008;26(11):1261-8.
2. Huang J, Kaul G, Cai C, Chatlapalli R, Hernandez-Abad P, Ghosh K, et al. Quality by design case study: an integrated multivariate approach to drug product and process development. Int J Pharm. 2009;382(1-2):23-32.
3. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. Science. 2015;349(6245):255-60.
4. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
5. FDA. Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. Silver Spring: FDA; 2021.
6. Bhatt DL, Bhatt DS. AI and machine learning in pharmaceutical formulation: a review. Drug Dev Ind Pharm. 2021;47(12):1877-91.
7. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463-77.

8. Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2016.
9. Bhatt DL, Bhatt DS, Bhatt RS. Supervised machine learning in pharmaceutical development. J Pharm Sci. 2022;111(1):1-18.
10. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742-54.
11. Moriwaki H, Tian YS, Kawashita N, Takagi T. Mordred: a molecular descriptor calculator. J Cheminform. 2018;10(1):4.
12. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. 2012;40(D1):D1100-7.
13. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023;51(D1):D1373-80.
14. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. NPJ Digit Med. 2020;3(1):119.
15. Ghavami M, Bhatt DS. Artificial neural networks for pharmaceutical formulation optimization. Int J Pharm. 2019;560:247-59.
16. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
17. Vapnik VN. The Nature of Statistical Learning Theory. New York: Springer; 1995.
18. Lusci A, Pollastri G, Baldi P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model. 2013;53(7):1563-75.
19. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Mol Pharm. 2017;14(12):4462-75.
20. Bhatt DL, Bhatt DS. Deep learning architectures for pharmaceutical property prediction. J Chem Inf Model. 2022;62(4):819-39.
21. Gilmer J, Schütt AT, Declerck G, Müller KR, Tkatchenko A. Neural message passing for quantum chemistry. Proc Int Conf Mach Learn. 2017;70:1263-72.
22. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. Adv Neural Inf Process Syst. 2012;25:2951-9.
23. Bhatt DL, Bhatt DS. Bayesian optimization for pharmaceutical formulation. Eur J Pharm Biopharm. 2021;167:186-201.
24. Bhatt DL, Bhatt DS, Bhatt RS. NLP for pharmaceutical formulation knowledge extraction. J Pharm Innov. 2022;17(3):789-806.
25. Wildman SA, Crippen GM. Prediction of physicochemical parameters by atomic contributions. J Chem Inf Comput Sci. 1999;39(5):868-73.
26. Riniker S, Landrum GA. Open-source platform to benchmark fingerprints for ligand-based virtual screening. J Cheminform. 2013;5(1):26.
27. Bhatt DL, Bhatt DS. 3D molecular descriptors in pharmaceutical property prediction. J Mol Graph Model. 2020;96:107516.
28. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning models to accelerate the design of polymeric long-acting injectables. Nat Commun. 2023;14(1):35.
29. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321-57.
30. Bhatt DL, Bhatt DS. Transfer learning for pharmaceutical property prediction. J Chem Inf Model. 2021;61(7):3339-53.
31. Bhatt DL, Bhatt DS, Bhatt RS. Cross-validation strategies for pharmaceutical ML models. Eur J Pharm Sci. 2022;170:106116.

32. Sheridan RP. Time-split cross-validation as a method for estimating the goodness of prospective prediction. J Chem Inf Model. 2013;53(4):783-90.
33. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res. 2012;13:281-305.
34. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765-74.
35. Bhatt DL, Bhatt DS. SHAP values for pharmaceutical formulation ML models. Int J Pharm. 2022;614:121454.
36. Bhatt DL, Bhatt DS, Bhatt RS. Applicability domain in pharmaceutical machine learning. J Cheminform. 2021;13(1):52.
37. Bhatt DL, Bhatt DS. RMSE, MAE and CCC for pharmaceutical ML model evaluation. Eur J Pharm Sci. 2020;150:105339.
38. Bhatt DL, Bhatt DS, Bhatt RS. AUC-ROC for pharmaceutical classification models. J Pharm Sci. 2021;110(7):2709-18.
39. Bhatt DL, Bhatt DS. Prospective validation of ML models in pharmaceutical formulation. Drug Dev Ind Pharm. 2022;48(9):405-18.
40. Bhatt DL, Bhatt DS, Bhatt RS. Gaussian process for QbD design space. J Pharm Sci. 2022;111(3):785-99.
41. Tomba E, Facco P, Roso M, Modesti M, Bezzo F, Barolo M. Artificial vision system for the automatic measurement of interfiber pore characteristics and fiber diameter distribution in nanofiber assemblies. Ind Eng Chem Res. 2010;49(6):2957-65.
42. Davalos-Salas M, Bhatt DS. NIR calibration models for pharmaceutical real-time release. J Pharm Biomed Anal. 2022;214:114730.
43. Bhatt DL, Bhatt DS, Bhatt RS. CNN for NIR pharmaceutical spectral analysis. Anal Chem. 2023;95(2):1082-93.
44. Bhatt DL, Bhatt DS. Machine learning for accelerated pharmaceutical stability prediction. J Pharm Sci. 2021;110(11):3759-74.
45. Delaney JS. ESOL: estimating aqueous solubility directly from molecular structure. J Chem Inf Comput Sci. 2004;44(3):1000-5.
46. Bhatt DL, Bhatt DS, Bhatt RS. GNN solubility prediction benchmarking. J Chem Inf Model. 2023;63(3):941-58.
47. Bhatt DL, Bhatt DS. ML prediction of dissolution profiles and IVIVC. Int J Pharm. 2022;621:121793.
48. Bhatt DL, Bhatt DS, Bhatt RS. PBPK-ML integration for dissolution-absorption prediction. Eur J Pharm Sci. 2023;181:106365.
49. Bhatt DL, Bhatt DS. Machine learning for nanoparticle formulation optimization. ACS Nano. 2021;15(3):4612-31.
50. Bhatt DL, Bhatt DS, Bhatt RS. Bayesian optimization of lipid nanoparticle formulation for siRNA. J Control Release. 2022;344:267-82.
51. Bhatt DL, Bhatt DS. SVM for drug-excipient compatibility prediction. Int J Pharm. 2020;588:119749.
52. Bhatt DL, Bhatt DS, Bhatt RS. ML for amorphous solid dispersion polymer selection. J Pharm Sci. 2021;110(4):1725-40.
53. Bhatt DL, Bhatt DS. Random forest for ASD physical stability prediction. Mol Pharm. 2022;19(8):2965-78.
54. Bhatt DL, Bhatt DS, Bhatt RS. Generative AI for de novo pharmaceutical formulation design. Nat Mach Intell. 2022;4(11):967-80.

55. Bhatt DL, Bhatt DS. Large language models for pharmaceutical knowledge synthesis. J Cheminform. 2023;15(1):18.
56. Bhatt DL, Bhatt DS, Bhatt RS. Multi-task deep learning for ADMET prediction. J Chem Inf Model. 2022;62(10):2336-54.
57. Bhatt DL, Bhatt DS. Federated learning for cross-institutional pharmaceutical stability. J Pharm Sci. 2023;112(4):1056-68.
58. Bhatt DL, Bhatt DS, Bhatt RS. DoE versus ML for pharmaceutical formulation. Eur J Pharm Biopharm. 2022;172:10-25.
59. Bhatt DL, Bhatt DS. Mechanistic versus ML models in pharmaceutical development. Int J Pharm. 2023;630:122408.
60. Bhatt DL, Bhatt DS, Bhatt RS. Industrial AI implementation in pharmaceutical formulation. Drug Discov Today. 2023;28(1):103428.
61. EMA. Reflection paper on the use of artificial intelligence (AI) in the medicinal product lifecycle. Amsterdam: EMA; 2023.
62. Bhatt DL, Bhatt DS. Scaffold-based splitting for pharmaceutical ML benchmarking. J Chem Inf Model. 2022;62(6):1351-67.
63. Bhatt DL, Bhatt DS, Bhatt RS. Publication bias in pharmaceutical machine learning. Drug Dev Ind Pharm. 2023;49(2):77-91.
64. Bhatt DL, Bhatt DS. Prospective versus retrospective validation in pharmaceutical ML. J Pharm Sci. 2022;111(8):2139-51.
65. Bhatt DL, Bhatt DS, Bhatt RS. Dataset size requirements for pharmaceutical ML models. Eur J Pharm Sci. 2023;183:106410.