Journal of Molecular Science

Journal of Molecular Science
Volume 35 Issue 4, Year of Publication 2025, Page 1436-1454
DoI-10.004687/1000-9035.2025.188

Journal of Molecular Science
www.jmolecularsci.com ISSN:1000-9035

Predictive Modeling Of Drug Release Behavior In Modified Delivery Systems Using Machine Learning Algorithm

Shivani Gandhi[1,2], Punit Parejiya[3], Shivangi Gandhi[4], Rushikesh Chaudhari[5], Nitin Varshney[6], Heta Shah[7], Urmi Prajapati[8]

[1] PhD Research scholar, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India
[2] Assistant Professor, Faculty of Pharmacy, The Maharaja Sayajirao University of Baroda, Gujarat, India
[3] Department of Pharmaceutics, K.B. Institute of Pharmaceutical Education and Research, Kadi Sarva Vishwavidyalaya, Gandhinagar, Gujarat, India
[4] Department of Computer Science and Engineering, GLS University, Ahmedabad, Gujarat, India
[5] Department of Computer Engineering, Sardar Patel College of engineering and technology, Anand, Gujarat, India
[6] Department of Computer Engineering, Marwadi University, Rajkot, Gujarat, India
[7] Department of Pharmaceutical Quality Assurance, Faculty of Pharmacy, The Maharaja Sayajirao University of Baroda, Gujarat, India
[8] Department of Pharmaceutics, Gandhinagar Institute of Pharmacy, Gandhinagar University, Gujarat, India

Article Information
Received: 24-10-2025
Revised: 17-11-2025
Accepted: 04-12-2025
Published: 23-12-2025

Keywords
Machine learning, Modified release drug delivery systems, Predictive modelling, Formulation optimization.

ABSTRACT

Modified release drug delivery systems (MRDDS) have been developed using impractical, empirical, trial-and-error procedures that are laborious, expensive, and sometimes ineffective. As a novel formulation science tool, machine learning (ML) provides predictive modelling, improved data analysis, and high-throughput formulation variable optimisation. In this review, ML methods like supervised learning, deep neural networks, kernel methods, and ensemble methods are used to MRDDS design and development using experimental, computational, and literature-based datasets. ML methods can predict formulation features such drug release kinetics, polymer–drug interactions, solubility augmentation, and long-term stability. Sustained release matrix tablets, osmotic systems, and nanoparticulate formulations demonstrate that ML methods improve prediction models and minimise experimental workload in MRDDS. Generative models, reinforcement learning, and self-driving laboratories are further ML approaches that can help rethink autonomous formulation design and optimisation. Discussed are data quality, feature selection, model interpretability, and regulatory acceptability issues. This paper shows how machine learning may be used to innovate and revolutionise modified release drug delivery system design for a more timely, cost-effective, and scientifically based pharmaceutical development strategy.

1. INTRODUCTION:

1.1 Overview of Pharmaceutical Formulation and Its Importance

Drug formulation of drugs is a critical step in the pharmaceutical development process that converts active pharmaceutical ingredients (APIs) into dosage forms that are therapeutically useful and have desirable pharmacokinetic and pharmacodynamic properties. The method of formulation will affect how quickly and to what extent the drug is absorbed, its stability and bioavailability, and whether and to what degree the patient adheres to the therapy. Within the scope of drug delivery innovations, we see the advent of a

modified-release drug delivery system (MRDDS), it is a system that releases drugs at predetermined rates, maintains plasma concentrations of drugs in the therapeutic range, and decreases dosing frequency and adverse events[1]. The MRDDS can contain sophisticated matrices, polymers, and excipients to deliver controlled, prolonged or targeted drugs. To create an efficacious formulation, the bioavailability of a drug depends on the drugs physicochemical characteristics of the drugs and the behaviour of the excipients along with additional factors that lie with the process conditions and in vivo biological milieu. Accordingly, an appropriate formulation that considers these factors into consideration, is paramount to ensure that laboratory therapeutics become viable therapeutics in the clinical and socially accepted sense in relation to efficacy, safety, and commercial viability[2].

1.2 Challenge of Traditional Formulation Development

The traditional formulation development process requires an empirical, trial and error process that relies heavily on in vitro and in vivo studies with several iterations. The ultimate trial and error process requires significant resource allotment and typically entails creating hundreds of trial batches to reach an optimized formulation. In the case of MRDDS, the situation is even more complex due to many variables directly affecting drug release kinetics--the type of polymer, viscosity, drug: polymer ratio, and the thickness of the coating are among those affecting the release behaviour. Predicting the consequences of excipient API interactions leads to unpredictable performance, stability, and development costs for the formulation. Barriers to the optimization of formulations include a limited mechanistic understanding of formulation behaviour, fragmented data, and irreproducibility. Conventional statistical approaches, including Design of Experiments (DoE) and Response Surface Methodology (RSM), provide useful information but do not adequately account for the non-linear, multi-dimensional interactions that drive drug release characteristics(3).

1.3 Advancing to Computational and Data-Driven Methods

Computational approaches are often used in pharmaceutical science to mitigate the restrictions posed by these methods. Molecular dynamics simulations, quantitative structure-property relationships (QSPR), and computational fluid dynamics (CFD) have developed novel frameworks to predict processes relevant to drug solubility, diffusion, and polymer behaviour. A limitation of physics-based models is that they often require significant computational resources and, even so, predictive mechanistic assumptions limit their use when novel products or processes include complex interactions. With greater access to high-throughput screening data or ever-advancing digitalization and data storage, the pharmaceutical landscape is also recognizing the transition to data-driven modelling - often better termed data-rich modelling - to better understand hidden relationships among components of the formulation and the metrics of performance. The digital convergence of formulation science has created an ecosystem where artificial intelligence (AI) and machine learning (ML) predication and automated formulation design, reduce the experimental burden and promote reproducibility(4).

1.4 The Contribution of Machine Learning (ML) in Pharmaceutical Sciences

Machine learning, which is a subfield of artificial intelligence provides the computers with the ability to learn from the data and in some cases to make predictions or decisions without explicit programming for it. In pharmaceutical sciences, machine learning has been utilized in several aspects of drug development, such as drug discovery, target identification, toxicity prediction, and formulation development. Insofar as drug delivery is concerned, a machine learning algorithm might be used to predict complicated non-linear relationships between variables in a formulation (e.g., polymer concentration, pH, particle size) with regards to the output (e.g., release rate and bioavailability). Supervised learning approaches such as random forests, support vector machines, and neural networks have proven to be very effective in predicting solubility enhancement, stability prediction, and modeling release kinetics(5). On the other hand, unsupervised and deep learning provide tools for clustering, feature extraction, and representation learning in complex, multidimensional data sets. The use of machine learning in formulation development transforms experimental and empirical methods into a data-driven approach based on enhanced knowledge, and supports experimental planning, facilitates optimal formulation design, and shortens the time to market(6).

1.5 Machine Learning-Aided Modified Release Formulation

The primary goal of this research is to develop ML-aided models for MRDDS that will be predictive, adaptive, and interpretable, and allow route-of-administration dependent modeling with limited experimental data(7). The study will:

1. Develop supervised and mixed machine learning models to predict drug release kinetics and stability profiles(8).
2. Feature engineering and algorithmic approaches are used to develop an

Journal of Molecular Science
Volume 35 Issue 4, Year of Publication 2025, Page 1436-1454
DoI-10.004687/1000-9035.2025.188

understanding of what the formulation components involved in the release mechanisms(8).
3. Different silos of experimental, literature-based, and computational data are combined to produce robust, generalizable predictive models(8).
4. Explainable AI methods can improve the interpretability and regulatory compliance of machine learning models(8).
5. Generative, reinforcement, and transfer learning approaches have been used design novel MRDDS formulations with improved therapeutic efficacy(8).

Approaches enabled by Machine Learning (ML) represent a novel and markedly disruptive professional shift from traditional experimental approaches to novel formulation creation. ML will continue to enhance efficiencies, significant time savings, cost savings, and improved product performance while using predictive modelling that adds reliability to sampling and then uses response modelling to suggest potentially improved formulations based on those modelling results. Converging data science and pharmaceuticals will change the way modified release drug delivery is accomplished through automation, accuracy, and innovation (9).

2. Foundations of Machine Learning in Pharmaceutical Formulation

Machine learning (ML) is the analytical foundation of data-driven formulation science. Machine learning algorithms can extract functional relationships between formulation composition, process parameters and critical quality attributes (CQAs) like drug-release profiles or disintegration time using experimental data. Unlike traditional mechanistic information, machine learning can accelerate the identification of nonlinear, high-dimensional relationships, and provide predictions for modified-release dosage forms (10).

2.1 Principles of Supervised and Unsupervised Learning

In supervised learning, algorithms are trained on labelled datasets that include input features (for example, percentage of polymer, type of excipient, compression pressure) that are related to the outputs of interest (for example, percent medicine released at time points of interest, dissolution rate constants). The model can predict these outcomes for new preparations it has not been trained on.

Commonly used supervised techniques in formulation research include regression (to predict continuous outcomes) and classification (for example, “acceptable” or “non-acceptable” dissolving). When Yang et al. (2019) applied supervised deep-learning algorithms to predict disintegration time and cumulative release of oral sustained-release tablets, they were over 80% accurate(11).

Unsupervised learning captures the patterns in the unlabelled data (i.e., the information is useful for creating predictive models, but the model doesn't realize that this information is present). Algorithms such as Principal Component Analysis (PCA), k-means clustering, and hierarchical clustering will group formulations with similar release behaviours, or similar patterns of variability. For example, when PCA is applied to the excipient composition data, we can determine which properties of the polymers contribute the most to the dissolution, and this can be used to select variables for the supervised models(12).

2.2 Comparative Analysis of Machine Learning and Molecular Modelling Techniques

While rational formulation development is facilitated by machine learning and molecular modelling, the two paradigms have fundamental bases. Molecular modelling (molecular dynamics, quantum chemistry, docking, coarse-grained simulations) is mechanistic because it explicitly defines atomic relationships imposed by physical laws. These models can rigorously explain microscopic behaviours, such as hydrogen bonding between drugs and polymers, or the dynamics of hydration in excipients. The major advantages of mechanistic models are their interpretability and physics based accuracy and limitations are computational, as they can only be run on small systems(13).

In contrast, machine learning was built on data. It picturesquely identifies relationships between formulation quality attributes, or composition, and macroscopic qualities of the observation, one does not have to resolve against a physical equation as the model is trained on data drawn from the underlying behaviours for the observable property and builds relationships through processing thousands of sample formulations. Machine learning efficiently scales to 1000's of trial formulations allowing for rapid predictions of bulk characteristics such as drug release profiles, hardness, or stability.

Some approaches presently track both thinking to couple paradigms into hybrid workflows e.g molecular descriptors (log P, topological polar surface area, rotatable bonds count) calculated from molecular modelling are provided to machine learning regressors to predict dissolution or permeability. Thus, adapting an atomistic understanding of events to a formulation relevant effectiveness (14).

Journal of Molecular Science
Volume 35 Issue 4, Year of Publication 2025, Page 1436-1454
DoI-10.004687/1000-9035.2025.188

2.3 Categories of Machine Learning Algorithms Applicable to Pharmaceutics

The following are the classes of algorithms commonly used in formulation-related research, and comments regarding their applicability, advantages, and disadvantages(15).

2.3.1 Linear and Nonlinear Regression Models

Initial applications included Multiple Linear Regression (MLR) and Partial Least Squares Regression (PLSR) to measure the impact of formulation components on drug-release characteristics. MLR posits a linear relationship between predictors and response; although interpretable, it oversimplifies intricate breakdown behaviour.

PLSR mitigates multicollinearity by transforming data into latent components, while maintaining a linear framework. Polynomial regression or nonlinear least-squares fitting can effectively model curvature in nonlinear responses; however, they frequently require manual feature engineering and pose a risk of overfitting with limited datasets(16).

2.3.2 Tree-Based Models (Random Forest, Gradient Boosting)

Tree-based ensembles, including Random Forest (RF) and Gradient Boosting Machines (GBM, XGBoost, and LightGBM), are highly effective for structured tabular data prevalent in formulation science. Random Forest constructs numerous decision trees utilizing random subsets of data and features; ensemble averaging reduces overfitting and offers internal assessments of feature significance, which is beneficial for pinpointing Critical Material Attributes (CMAs) and Critical Process Parameters (CPPs) within the Quality-by-Design (QbD) paradigm. Gradient boosting incrementally builds trees that rectify prior residuals and frequently attains superior predictive accuracy at the expense of interpretability. These models have exhibited superior efficacy in forecasting dissolution timepoints and discerning predominant excipient factors(17).

2.3.3 Kernel Techniques (Support Vector Machines, Gaussian Process Regression)

Support Vector Machines (SVM) and Gaussian Process Regression (GPR) employ kernel functions to transform inputs into high-dimensional feature spaces. They are effective for small to medium-sized datasets and have been utilized to classify formulation success or predict release processes. GPR provides probabilistic forecasts, delivering mean estimates and confidence intervals beneficial for decision-making under regulatory oversight, where the quantification of uncertainty is crucial. Nonetheless, the computing expense increases cubically with the size of the dataset, hence limiting Gaussian Process Regression to smaller model datasets unless approximations are utilized(18).

2.3.4 Deep Learning Architectures (Artificial Neural Networks, Convolutional Neural Networks, Recurrent Neural Networks)

Deep learning techniques epitomize the cutting edge of machine learning in pharmaceuticals. Artificial Neural Networks (ANNs), consisting of interconnected layers of neurones, approximate nonlinear relationships between the inputs and outputs. Augmenting the number of hidden layers produces Deep Neural Networks (DNNs) proficient in hierarchical feature extraction.

Convolutional Neural Networks (CNNs) are proficient in analyzing spatially organized data, such as microscopy or near-infrared imaging of tablet surfaces, as they discern localized patterns associated with porosity or coating uniformity. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) units, effectively capture temporal dependencies, rendering them suitable for modelling time-series dissolution data and real-time process analytics(19).

2.4 Software Tools and Computational Frameworks (Scikit-learn, TensorFlow, PyTorch)

The deployment of machine learning models in pharmaceuticals has been demonstrated using robust open-source ecosystems. Scikit-learn (Python) offers standardized interfaces for classical methods, including regression, classification, clustering, and model validation, and has been extensively utilized in comparative machine-learning studies for formulation prediction. TensorFlow (Google Brain) and PyTorch (Meta AI) are prominent in deep-learning research, providing GPU acceleration, automatic differentiation, and adaptability for developing intricate neural structures. Keras, constructed using TensorFlow, simplifies the network architecture through intuitive APIs designed for pharmaceutical scientists unfamiliar with programming. Deep Learning, employed by Yang et al. (2019), interacts effortlessly with enterprise settings for extensive deployment. The computational toolkit was augmented by the supporting packages: NumPy, pandas, Matplotlib, and Optuna for hyper-parameter tuning. These frameworks facilitate reproducible modelling, hyper-parameter optimization, and visualization of learning behaviour, thus expediting the implementation of machine learning in both academic and industry formulation laboratories(20).

Journal of Molecular Science
Journal of Molecular Science
Volume 35 Issue 4, Year of Publication 2025, Page 1436-1454
DoI-10.004687/1000-9035.2025.188

2.5 Importance of Interpretability and Explainability of Machine Learning Models

Although prediction accuracy is essential, interpretability and explainability dictate the scientific and regulatory acceptability of machine learning models in pharmaceuticals.

Regulatory agencies, such as the FDA and EMA assert that data-driven models must be comprehensible, auditable, and congruent with mechanistic understanding to adhere to Quality by Design guidelines. Black-box models, particularly deep neural networks, frequently exhibit a deficiency in transparency concerning the impact of certain formulation factors on predictions. To address this, post-hoc interpretability approaches are utilized: feature-importance analysis evaluates the influence of specific formulation factors (e.g., polymer concentration or tablet hardness) on anticipated results. SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) offer localised interpretations by assessing the contribution of each variable to a specific prediction. Partial-dependence graphs illustrate the impact of variations in a single feature on the expected response, with the other features held constant. These interpretability frameworks enable scientists to confirm that model-identified essential elements correspond to established physicochemical principles, thereby enhancing confidence among formulators and regulators. Interpretable machine learning models not only provide predictions but also elucidate the reasons behind the success or failure of specific formulations, thereby connecting data science with domain expertise, an essential requirement for its incorporation into regulated pharmaceutical processes(21).

3.0 Data-Centric workflow for model development

An effective machine learning model developed for drug formulation, particularly modified release systems, relies on the quality, representativeness, and clarity of the data. The predicted accuracy and generalizability of any ML-supported formulation model are contingent upon the quality of the entire workflow from data collection to model interpretation. A data-focused paradigm ensures that each step along the continuum, starting with experimental design and ending with evaluation, is improved to uncover important scientific relationships rather than simply focusing on model sophistication. The following sections provide a thorough presentation of the step that comprise a scientifically robust, data-driven approach to formulation consistent with the Scopus-indexed publications standards(22).

3.1 Data Acquisition Strategies

3.1.1 Experimental Data Generation

The generation of experimental data underpins the domain of ML-assisted drug formulations. Modified release drug delivery systems (MRDDS) should have comprehensive studies completed in the laboratory to assess drug release in the system, solubility, encapsulation efficiency, particle size and modeling, zeta potential and stability, and biopharmaceutical properties under controlled experimental conditions that can be controlled. The experimental data were generated through testing methods such as, but not limited to, USP dissolution studies, HPLC- analysis, or dynamic light scattering.

The experimental data should also include a variety of formulation compositions, processing factors, and experimental conditions to allow the generalization of ML prediction models. Recently, advances in automation, including high-throughput testing and robotics in laboratories, have significantly improved data reproducibility and throughput. Automated workflows reduce human error, ensure controlled conditions throughout the experiment, and can interface with iterative fast cycles of experimentation guided by machine-learning models(23).

3.1.2 Literature and Database Mining

Because of the significant expense and time required for experimental data, the literature and database mining are important additional sources. Data may be found in published journal articles, patents, as well as on the Internet in several curated repositories, such as PubChem, DrugBank, ChEMBL, and the Open Reaction Database. Text-mining and natural language processing (NLP) tools can automate the extraction of quantitative characteristics, including solubility, release rates, and polymer ratios, from the scientific literature. These datasets can be combined in meta-analyses that combine the learnings across studies, providing a larger and more diverse training dataset. Nonetheless, heterogeneity in data owning to different experimental arrangements and reporting formats remains a challenge. Data standardisation must be performed using unit standardization, normalization, and outlier detection, prior to machine learning pipeline inclusion(24).

3.2 Data Preparation and Curation

Data preparation is the process of transforming raw, experimental, and crawled data into a clean and structured machine-readable format. These numerical features may need to be normalized using some methods such as min-max normalization or z-score normalization, so there is a level playing field when training the model.

Journal of Molecular Science

Likewise, for categorical features such as excipient type or formulation method, one-hot or label encoding methods must be used. Statistical thresholds for outlier detection or forests isolation will be useful for removing anomalous data points that affect model learning. If the data are very small, then an enhancement or augmentation of the data, for instance through interpolation of release profiles or simulation, may be useful. Careful curation is important for reproducibility and ensures that ML models can learn actual formulation-performance relationships rather than relationships induced by random noise(25).

3.3 Selection and Engineering of Features for Formulation Variables

Feature selection and engineering are crucial for a good representation of the formulation system in machine-learning modelling. Feature selection and engineering are related to how to express the chemical, physical, and process variables are expressed as numerical or categorical variables that the algorithm can understand(26).

3.3.1 Physicochemical Characteristics of Active Pharmaceutical Ingredients

API-related characteristics are determined by molecular structure and physical properties, which govern solubility, permeability, and release kinetics. Typical descriptors included molecular weight, LogP (lipophilicity), pKa, melting point, polar surface area, counts of hydrogen bond donors/acceptors, and topological indices. These characteristics can be evaluated with cheminformatics software such as RDKit or direct novel empirical measurements. The attributes of excipients within modified release formulations, such as crystallinity, particle size distributions, and kinetics of diffusion in polymer matrices, are particularly important. Feature correlation analysis serves to adjust descriptors and remove redundancy, and will help improve both the interpretability and efficiency of models(27).

3.3.2 Function of Excipients, Polymers, and Surfactants

Excipients are important aspects of consideration in MRDDS as they will modulate the release of the drug, stability of the drug, and mechanical integrity of the system.

Excipients can be defined by their attributes which could include the type of polymer (i.e., HPMC, Eudragit, PVP), molar mass, the viscosity grade, the solubility parameter, the glass transition temperature (Tg), and the hydrophobicity index. The ratios of the drug to excipient or the multiple excipients can be captured as designed variables, such as the drug-to-polymer ratio or as an interaction term in the composite factor. The characteristics of surfactants (like hydrophilic-lipophilic balance (HLB), and critical micelle concentration (CMC)) can be very important in formulations that use emulsions final products or use a lipid dosage form product (lipids being part of a complete formulation). It is also important to remember that categorical factors based on formulations need to be encoded appropriately, and there will be an opportunity to improve chemical descriptors by alternative representation of molecular structures, either as molecular fingerprints or aspects of graph representations methods(28).

3.4 Model Training and Evaluation

3.4.1 Data Partitioning and Cross-Validation

Once a curated dataset is acquired, it must be divided into training, validation, and test sets to evaluate the model's performance fairly. A common division might be 70% training, 15% validation, and 15% testing; however, this depends on the size of the dataset and its application. For smaller datasets, k-fold cross-validation (typically k=5 or 10) is recommended to achieve stability in the evaluation and reduce the variance that can arise from random data partitioning. Stratified sampling ensures that the labels are distributed as in original data. External validation using new data that are out of sample, for example from independent testing or an independent API, establishes credibility in the model. Nested cross-validation allows the optimization of hyperparameters while avoiding data leakage(29).

3.4.2 Optimisation of Hyperparameters

Hyperparameters dictate model behaviour and require rigorous optimization for optimal performance. Hyperparameter search algorithms, such as grid or random search, and Bayesian optimization, are extensively used in machine-learning pipelines in the drug development. Adjusting the number of trees in a Random Forest, kernel function in Support Vector Regression or learning rate and depth in Gradient Boosting Machines, can have a meaningful impact on or predictive performance. A hyperparameter search that is grounded in cross-validation ensures that generalisation and overfitting are accounted for. AutoML frameworks, such as Optuna and Hyperopt, can automate such practices to create reproducible optimization(30).

3.5 Model Evaluation Metrics (R^2, RMSE, MAE, Accuracy, Precision, Recall)

The performance must be qantifyied to allow trust in the predictions must be conducted. For regression problems such as the prediction of drug release rate, solubility, and stability, the common metrics of quality are given by the coefficient of determination (R^2), mean absolute error (MAE),

and root mean square error (RMSE). In general, a high R^2 and low error values indicate that the predictive model worked well, while both good predictions and robustness were achieved. In classification concerns -for example, classifying formulations as "ideal," "suboptimal," or "unstable" - accuracy, precision, and recall, as well as the F1-score and area under the ROC curve (AUC), are standard accuracy measures. The ensemble of these metrics can be interpreted together to assess distinctions between categories, as well as the correctness of the predictions. There are also graphical methods of assessments (e.g., parity plots, learning curves, and residual distributions) that may provide deeper insight into predictive behaviour, as well as potential bias(31).

3.6 Interpretability for Understanding Feature Importance

The interpretability of the machine learning model embraces the link between prediction and scientific understanding. The purpose of feature importance capturing is to understand which input parameter(s) have a fluctuating effect on output prediction, while also allowing insights into formulation behaviours. A Random Forest model suggests that the polymer molecular weight and drug-to-polymer ratio are the most influential factors in the release rate. To quantify feature importance, permutation methods, Shapley Additive Explanations (SHAP) and partial dependence plots (PDP) can be utilized. These tools allow the conversion of statistical correlations into understandable scientific trends that can help formulation scientists make informed decisions regarding data. Explainable AI (XAI) techniques are particularly useful in regulated industries (such as pharma) where transparency is a requirement. Furthermore, the interpretations made can guide hypothesis generation and enhance the confidence of researchers in creating new combinations of excipients or new processing conditions. Hence, interpreting a model is not a simple validation stage; it changes machine learning from a prediction tool to a mechanism to explore mechanisms and intelligently design formulations(32).

4.0 Applications of Machine Learning in the Development of Drug Delivery Systems

4.1 Traditional Oral Dosage Form

4.1.1 Early Work on Artificial Neural Networks

The first studies utilizing machine learning (ML) in the manufacturing of oral dosage forms were initiated in the 1990s and focused on the use of artificial neural networks (ANNs) to correlate product performance attributes with formulation composition and processing parameters. Nevertheless, these studies are innovative in demonstrating the value of data-driven models to capture complex and nonlinear relationships between the ratio of excipients, compression force, and product properties to enhance the quality of the traditional designs of experiments (DoE) approach. In most of the first findings, the data datasets were small, and were limited in their validation approaches, but demonstrated the potential of ML to accelerate the screening of formulations and the design of experiments by producing the benchmarking capability of ranked formulations to support laboratory investigation(33).

4.1.2 Forecasting of disintegration, dissolution, and friability

Subsequent research broadened the application of machine learning to forecast essential tablet quality characteristics, including disintegration time, dissolving profiles, tensile strength, and friability. Models, including feedforward neural networks and tree-based ensembles, have been trained using combinations of API descriptors, excipient composition (percent w/w), and process parameters (milling time, compression force). In situations where datasets and validation were robust, machine learning models have provided accurate predictions of dissolution kinetics and disintegration to enable virtual screening of formulations ahead of laboratory tests. These predictive capabilities are useful when rapidly studying candidate formulations, and identifying undesirable situations early in development (e.g., poor compressibility or too rapid disintegration), allowing for laboratory efforts to be directed toward leads of high probability(34).

4.2 Advanced oral formulations and adjustable release systems

4.2.1 Sustained-release matrix tablets

Sustained release matrix tablets require careful optimization of the type of polymer, grade of polymer, amount of drug, and processing conditions to yield the target drug release kinetics. Machine learning underpinnings (tree-based ensembles and neural networks) have been developed to predict release parameters (e.g., percentage released at certain times and release rate constants) based on these factors. Feature importance and sensitivity analyses commonly indicate that the molecular weight of the polymer, the ratio of drug-to-polymer ratios, and the porosity of the matrix are important. When applied wisely through suitable API segmentation and prospective validation, machine learning has been utilized to suggest formulations that achieve target release windows while reducing experimental trials, and to clarify nonintuitive connections among formulation variables(35).

4.2.2 Push-pull osmotic pump mechanisms

Push–pull osmotic pumps exemplify a uniquely mechanistic modified release method, wherein membrane characteristics, core configuration, osmotic agent concentration, and manufacturing tolerances determine the efficacy. Data-driven methodologies have employed machine learning to forecast the combinations of osmogen, membrane composition, and core design will achieve the required zero-order or controlled release profiles. In this area of research, hybrid methods that combine mechanistic descriptors (such as permeability approximations from models) and machine learning descriptors usually outperform models that are purely empirical because they rely on established mass-transfer physics, but at the same time handle manufacturing variability and formulation specifics(36).

4.2.3 Microemulsions and Cyclodextrin Complexes

Microemulsions (lipid-based systems), and host-guest complexes (cyclodextrin inclusion) have elements of composition and physicochemical properties that lend themselves quite well to machine learning. Models have been used to predict solubilization efficacy, phase behaviour domains, particle size, and complexation constants based on excipient properties (HLB, chain length, polarity) and API descriptors. Additionally, efficient screening with machine learning has helped prioritize surfactant/co-surfactant ratios and concentrations for experimental tests, as well as discover cyclodextrin–guest pairs that have an increased probability of enhancing aqueous solubility or stability(37).

4.3 Machine Learning for IVIVC (In-vitro–In-vivo connection) Modelling

IVIVC is a valuable yet complex application of machine learning because it necessitates the connection of *in vitro* dissolution/release profiles with pharmacokinetic consequences in vivo. Machine learning methodologies have integrated comprehensive *in vitro* release profiles, formulation characteristics, dosage data, and subject/study metadata to forecast plasma concentration–time trajectories or summary pharmacokinetic parameters (Cmax, Tmax, AUC). Successful modelling has clearly indicated that the process of curating of the dataset, consideration of relevant features for the study, and validation (e.g., leave-one-API-out or external prospective validation) is important. Once predictive performance is established, ML-based *in vitro*-in vivo prediction can mitigate animal use and direct formulation decision- making in the preclinical phase; however, regulatory approval does 'require some level of transparency and mechanistic basis(38).

4.4 Predictive modelling for solid dispersions and active pharmaceutical ingredient solubility

Solid dispersions and solubility-enhancing methods are key examples of note for machine learning, where important outcomes (physical stability, crystallization propensity, improved dissolution) will be a function of multiple interdependent molecular and formulation variables. Machine learning models which include chemical descriptors (such as changes in glass transition and hydrogen bonding affinity), polymer attributes, and processing have been used to predict the stability and effectiveness of dissolved and solid dispersions. Solubility prediction models can also include excipient and solvent attributes to predict which excipients (hydrotropes, cyclodextrins, surfactants) and formulation conditions will facilitate API solubility. Prediction can be helpful in constraining the experimental space of stable formulations of ASD(39).

4.5 The Role of Machine Learning in Optimising Polymer Ratios and Excipient Interactions

Optimizing polymer ratios and excipient-API interactions is one of the main objectives of MRDDS design; machine learning provides predictive and interpretive approaches to facilitate this process. Based on the assessment of multi-factor datasets, machine learning algorithms can assess synergistic or antagonistic interactions among excipients (e.g., plasticizer-polymer interactions or surfactant-drug micellar solubilization) and determine an appropriate drug-to-polymer ratio to achieve the desired release kinetics. Other feature attribution approaches (e.g. SHAP, partial dependence plots) permit researchers to take model outputs and convert them into chemical and physical type insights about which excipient characteristics (e.g., molecular weight, Tg, hydrophobicity) will affect the modulation of release. In addition, uncertainty estimation can be employed with machine learning (e.g. Gaussian processes, ensembles) to while accounting for uncertainty, prioritizing the formulations that optimize the predicted performance and mitigates risk, and indicating the solutions (potential formulations) through active learning to narrow down the areas of formulation space that need clarification. In summary, machine learning provides added value to excipient selection and ratio optimisation by turning existing and future data into predictive solutions that can be practically verified experimentally in the laboratory (40).

5. Machine Learning Amended Development of Proteins and Biopharmaceuticals

Machine learning (ML) is a revolutionary tool for the development of protein-based and biopharmaceutical products that involve molecular complexity, concern stability, and exposure of proteins to environmental and processing

Journal of Molecular Science
Volume 35 Issue 4, Year of Publication 2025, Page 1436-1454
DoI-10.004687/1000-9035.2025.188

parameters. Traditional formulation approaches often rely on laborious empirical screening of formulations to evaluate stabilizing excipients and/or appropriate processing conditions, leading to lengthy and resource intensive development cycles. By mining historical data to observe stability trends, aggregation trends and release behaviour under solicited and changing conditions, machine learning models can predict variables that derive from protein sequences, protein structures, and excipient properties. Regardless of the ML methods presented in peer reviewed and indexed in Scopus literature, as ML is increasingly featured in decision trees and other methods used in data driven decision making, which enhances formulation robustness and reduces development timeliness and experimental redundancy(41).

5.1 Challenges Associated with Protein Stability and Delivery

In terms of structural complexity, large molecular size, and conformational fragility, proteins and peptides are entirely different from those of small molecule drugs in Table 1. They are subjected to myriad degradation pathways, such as aggregation, deamidation, oxidation, hydrolysis, and denaturation due to environmental conditions (e.g. temperature, pH, ionic strength and mechanical disturbance). Obtaining conformational stability during manufacturing, storage, and administration is a major challenge. Conventional stabilization methods rely on combing different combinations of buffers, cryoprotectants, surfactants and polymers during the experimental phase and are often based on a limited mechanistic understanding of protein degradation pathways. Furthermore, the delivery of therapeutic proteins will require a dedicated carrier system (e.g., polymeric microparticles, liposomes or hydrogels) to ensure a defined release while preserving biological activity. The multifactorial dependencies make protein formulation a difficult optimization problem that is well suited to machine learning modelling that can incorporate additional non-linear and multi-dimensional factors and interactions among formulation parameters and stability outcomes(42).

5.2 Neural Network Models for Predicting Thermal and Aggregation Stability

Artificial neural networks (ANNs) and deep learning systems have shown great effectiveness in predicting the thermal and aggregation stabilities of proteins. Modeling techniques that utilize datasets of physicochemical descriptors based on sequence data (e.g. amino acid composition, hydrophobicity, propensity for secondary structure) combined with formulation variables (e.g., excipient type, excipient concentration, pH, buffering system) can be useful in predicting the melting temperature (Tm), aggregation onset, and turbidity. Multi-layer perceptron (MLP), convolutional neural networks (CNN), and recurrent neural networks (RNN) have all been developed to characterize the relationships between sequence and stability. Deep learning algorithms can learn and extract features from both the primary and tertiary structures of proteins and connect these features to formulation performance metrics. Several papers in the literature indicate that models that incorporate descriptors related to excipient properties (e.g., osmolyte concentration, surfactant class, polymer hydrophobicity) increase predictive power and allow more efficient virtual screening of potential stabilizing components before experimental work is undertaken. Conceptually, interpretable machine learning methods (e.g., SHAP and attention models) allow some degree of identification of residues or excipient descriptors that promote stability and create bridges between data-driven predictions and mechanistic reasoning(42).

5.3 Machine Learning Prediction of Long-Term Stability Under Variable Conditions

The prediction of long-term stability in Table 1 is a highly beneficial, yet challenging application of machine learning in protein composition. Experimental stability testing under both real-time and accelerated settings may require months or years; however, machine learning models can significantly shorten this duration by predicting the degrading behaviour with minimal early stage data. Regression models, such as Random Forests, Gradient Boosting, and Gaussian Process Regression, have been reliably trained on accelerated stability datasets, demonstrating the effects of temperature, relative humidity, and pH on potency and aggregation(43). Time-series models, such as Long Short-Term Memory (LSTM) networks, have been applied to generate long-term stability profiles from short-term observations, thus providing early detection of potential instability or aggregation risk. The utility of the feature importance from these models continue to show that environmental factors and formulation ingredients (e.g., buffer capacity and surfactant concentrations) are primary predictors of stability. With uncertainty quantification, these models can provide probabilistic confidence intervals around predictions, thereby increasing their applicability to risk-based decision making for the biopharmaceutical development cycle. This means that predictive stability modelling based on machine learning, helped to unambiguously identify the most stable formulations earlier in the development cycle, decreasing formulation optimization time, costs, and R&D resources(44).

5.4 Conjugation to Polymeric Microparticles for Sustained Protein Delivery

Controlled delivery of proteins using polymeric

Journal of Molecular Science

microparticles, principally poly(lactic-co-glycolic acid) (PLGA), is an advanced field, and machine learning will greatly assist in design and optimization. These systems must balance the encapsulation efficiency, release kinetics, and structural viability of the protein antigens in biodegradable polymer systems. Machine learning models can develop predictions for encapsulation efficiency, burst release, and sustained-release profiles based on the polymer molecular weight, lactide ratio, solvent type, emulsifier concentration, and the physicochemical characteristics of the protein. Random Forests, Support Vector Regression, and Deep Neural Networks have demonstrated exceptional capabilities for modelling these datasets, exposing non-linear relationships lost in conventional regression techniques. Feature importance analysis often indicates polymer hydrophobicity, end-capping state, and protein isoelectric point are primary determinants of release behaviour. Hybrid mechanistic–machine learning models, which integrate diffusion or degradation equations as physics-informed restrictions, augment predictive reliability. The knowledge generated by these model(s) allow for accurate formulation design and formulators to select polymer–protein combinations that result in the desired release and still maintain biological activity. As a result, ML-based approaches hasten the translation of sustained-release protein therapies from an experimental idea to a clinical reality, representing the merging of data science and advanced drug delivery technologies(45).

Table 1 Machine Learning Assisted Development of Protein and Biopharmaceutical Formulations

Sr No	Aspect	Description	ML Techniques Used	Applications / Examples	Advantages	Reference
1	Protein Stability Prediction	Predicts degradation, aggregation, or denaturation of proteins under different conditions.	Supervised learning, Regression models, Random Forest, Neural Networks	Predicting thermostability or aggregation hotspots	Reduces experimental screening, accelerates stability optimization	(46)
2	Formulation Component Optimization	Determines optimal excipients, pH, and buffer conditions for stable formulations.	Bayesian Optimization, Decision Trees	Predicts ideal buffer systems and excipient concentrations	Minimizes trial-and-error in formulation design	(47)
3	Protein Excipient Interaction Modeling	Models molecular interactions between protein and stabilizers or adjuvants.	Deep Learning, Molecular Dynamics + ML, Graph Neural Networks	Predicts compatibility of surfactants or sugars	Improves compatibility and prevents denaturation	(48)
4	High-Throughput Screening Data Analysis	Analyzes large datasets from automated formulation experiments.	Clustering, Dimensionality Reduction (PCA, t-SNE)	Classifies stable vs unstable formulations	Extracts meaningful patterns from large datasets	(49)
5	Predicting Aggregation and Viscosity	Estimates formulation viscosity and protein aggregation propensity.	Support Vector Machines (SVM), ANN	Predicts solution behavior during storage	Prevents issues during manufacturing and storage	(50)
6	Accelerated Stability Testing	Uses ML to extrapolate long-term stability from short-term data.	Time-Series Models, Regression	Predicts shelf-life and degradation kinetics	Speeds up stability assessment and reduces cost	(51)
7	Protein Structure Formulation Relationship Analysis	Links 3D protein structure features with formulation behavior.	CNNs, Autoencoders	Identifies structure-dependent formulation sensitivity	Enhances understanding of structure–stability relationship	(52)
8	Quality by Design (QbD) Integration	Integrates ML in QbD workflows for design space identification.	Reinforcement Learning, Predictive Modeling	Predicts critical quality attributes (CQAs) and process parameters	Ensures regulatory compliance and product consistency	(53)
9	Process Monitoring and Control	Real-time monitoring of formulation and filling processes.	Machine Vision, Predictive Maintenance Models	Detects anomalies or contamination	Improves manufacturing efficiency and safety	(54)
10	Data Integration and Knowledge Discovery	Combines multi-omics, formulation, and process data for insight generation.	Multi-modal ML, Data Fusion	Correlates biological data with formulation performance	Enables holistic biopharmaceutical development	(55)

6.0 Machine Learning in Microparticle and Nanoparticle Drug Delivery

In recent years, there has been a significant increase in the utilization of machine learning (ML) for the design of micro - and nano-scaled drug delivery systems. These systems including polymeric microspheres, polymeric nanoparticles, and lipid nanoparticles (LNPs) demonstrate intricate, multivariate formulation-performance correlations that are challenging to optimize experimentally. Machine learning techniques provide a data-driven approach to correlate formulation parameters with essential performance metrics, including encapsulation efficiency (EE), drug loading (DL), and drug-release kinetics(56).

6.1 Analysis of polymeric and lipid-based nano/micro systems

Poly(lactic-co-glycolic acid) (PLGA) and analogous biodegradable polyesters are the most extensively studied carriers for continuous release. Singh et al. (2021) concluded that factors such as polymer molecular weight, lactide: glycolide ratio, and end-group chemistry significantly affect the breakdown rate and release dynamics(57). Simon et al. (2021) and Li et al. (2019) reached analogous conclusions, highlighting the synergistic impact of the particle size and polymer composition on diffusion-controlled release. Lipid nanoparticles (LNPs) and solid lipid nanoparticles (SLNs) have concurrently emerged for the transport of mRNA, siRNA, and poorly soluble small molecules(58).

6.2 Predictive Models for Release Kinetics and Encapsulation Efficiency

The primary machine learning objectives in nanoparticle formulation are (i) to predict encapsulation efficiency and drug loading in Table 2, and (ii) to simulate cumulative drug release(59). Hosni et al. (2025) conducted a thorough analysis of machine learning algorithms in nanoparticle research, revealing that Random Forests (RF) and Gradient Boosting Machines (GBMs) generally achieve good accuracy (R^2 = 0.75 - 0.9) with few data prerequisites. They emphasized that incorporating physically relevant descriptors, such as particle size, polymer molecular weight, and drug log P enhanced model generalization(60). Yang et al. (2021) illustrated that deep neural networks (DNNs) surpassed traditional models, including multiple linear regression (MLR), support-vector machines (SVM), and random forests (RF), to predict multi-time-point dissolution profiles of sustained-release tablets, attaining over 80% accuracy, and introducing the Maximum-Dissimilarity-Function for Intelligent Splitting (MD-FIS) algorithm to mitigate data leakage(61). Their methodological breakthroughs are currently being used in nanoformulations, which are characterized by limited data and significant correlations across characteristics. Seegobin et al. (2024) utilized ensemble learning, namely cubist regression and RF models, to forecast protein release from PLGA microspheres in microsphere systems, attaining an R^2 of approximately 0.69 and validating the significant influences of polymer molecular weight, lactic: glycolic ratio, and drug solubility(62). Subsequent studies by Sivadasan et al. (2021) expanded similar methodologies to hybrid polymeric-lipid particles, demonstrating that the incorporation of process characteristics (e.g., emulsification speed and solvent evaporation rate) enhanced the prediction of both encapsulation efficiency and release profiles(63).

6.3 Case Studies: Microspheres and Lipid Nanoparticles from PLGA

Zawbaa et al. (2016) initiated machine learning modelling of macromolecule release from PLGA microspheres using a well-maintained dataset of 166 formulations. The authors evaluated nine methods and determined that RF had the highest prediction accuracy, and identified significant factors affecting burst release(64). Maksimenko et al. (2019) utilised Gaussian Process Regression (GPR) to model doxorubicin-loaded PLGA particles, employing the model's uncertainty quantification to inform fresh experimental trials, which serves as a prelude to closed-loop experimental design(65). Maharajan et al. (2024) and associates developed an extensive LNP dataset comprising several formulations and utilized XGBoost and GPR to forecast mRNA encapsulation and *in vitro* efficacy. The feature-importance analysis indicated that the microfluidic flow-rate ratio and the percentage of ionizable lipids were the most significant factors(66). In a separate study, Correia et al. (2023) combined the design-of-experiments (DoE) with artificial neural networks (ANNs) to enhance curcumin-loaded solid lipid nanoparticles (SLNs), attaining 93% encapsulation efficiency while reducing experimental runs by 40% compared to traditional DoE methods. Together, these investigations together illustrate that machine learning can proficiently identify nonlinear parameter relationships that govern particle size, encapsulation efficiency, and release kinetics, attributes that would often necessitate substantial empirical optimization(67).

6.4 Model Optimization by Feature Selection Algorithms

Feature selection is crucial for improving the model resilience and interpretability in datasets with numerous associated formulation variables. Filtering techniques were Ge et al. (2021) introduced a comprehensive Fisher - RFE - Logistic (FRL) framework that amalgamates Fisher score ranking, recursive feature elimination (RFE), and logistic regression for biomedical datasets. This method swiftly removes superfluous variables

in formulation science, preserving only those with the highest predictive significance(68). Wrapper and embedded techniques by Figueroa Barraza et al. (2021) integrated internal feature selection by assessing DNN sensitivity coefficients(69), whereas Wang et al. (2022) employed (Recursive Feature Elimination with Cross-Validation) RFECV in conjunction with RF to determine the five principal variables (particle size, polymer ratio, surfactant concentration, stirring speed, and solvent type) that account for 85% of EE variance(70). Evolutionary algorithms, Sarmah et al. (2020) examined the application of genetic algorithms (GA) and particle-swarm optimisation (PSO) for hyperparameter tuning and variable selection in pharmaceutical machine learning pipelines, demonstrating enhancements in prediction accuracy ranging from 5% to 15% compared to manual selection. These methods collectively boost performance and offer mechanistic understanding by identifying the variables that most significantly influence encapsulation or release behavior(71).

6.5 Comparative Examination of Model Precision and Overfitting Mitigation

Deep architectures, as demonstrated by Emami et al. (2024) can provide enhanced predictive accuracy, especially for multi-output tasks such as whole release curves yet necessitate larger datasets and rigorous regularization. Gaussian-process models provide valuable uncertainty quantification for regulatory filings. Owing to the relatively small size of formulation datasets (fewer than 200 items) and their frequent imbalance, it is imperative to meticulously test model generalisation(72). Hoseini et al. (2025) addressed this issue using the MD-FIS technique to generate realistic train/test splits(73). Vanek et al. (2017) advocated for layered cross-validation and dropout regularisation in deep neural networks(74), whereas Pan et al. (2025) highlighted the need for ensemble averaging and data augmentation via physics-constrained simulations(75). Researchers are increasingly utilising explainability tools, such as SHAP (Shapley additive explanations) and partial-dependence plots, to identify spurious correlations. The assessment criteria, that is RMSE, MAE, and R^2, continue to be conventional metrics; however, Liu et al. (1997) proposed the similarity factor f_2 as a pertinent criterion for evaluating anticipated and experimental dissolution patterns in pharmaceuticals(76). This criteria was subsequently adopted by subsequent research (Stevens et al., 2015) to evaluate release-profile predictions(77).

Table 2 Machine Learning Applications in Microparticle and Nanoparticle Drug Delivery Systems

Sr No	Category	Key Focus Area	Machine Learning Methods	Case Examples	Benefits	Reference
1	Particle Engineering	Prediction of particle size, morphology, and surface charge during synthesis	Regression models, Random Forest, Neural Networks	Predicting nanoparticle diameter and zeta potential from formulation inputs	Enables precise control of delivery profile and targeting efficiency	(78)
2	Formulation Design	Optimization of polymer, surfactant, and solvent ratios	Bayesian Optimization, Decision Trees, Genetic Algorithms	Designing PLGA or chitosan nanoparticles with high encapsulation	Reduces experimental iterations and improves reproducibility	(79)
3	Encapsulation and Drug Loading	Estimation of drug entrapment efficiency and loading capacity	Artificial Neural Networks (ANN), Linear Regression, Support Vector Machines	Predicting encapsulation efficiency for liposomes and polymeric carriers	Enhances stability and maximizes payload	(80)
4	Controlled Drug Release Modeling	Predicting release kinetics and diffusion profiles	Support Vector Regression, Random Forest, Deep Learning	Modeling sustained release of peptide-loaded nanoparticles	Saves time and supports formulation scaling	(81)
5	Stability and Storage Prediction	Evaluation of long-term physicochemical stability	Ensemble Learning, Time-Series Models	Predicting aggregation and crystallization in stored formulations	Accelerates shelf-life determination and reduces waste	(82)
6	Targeting and Delivery Efficiency	Prediction of tissue targeting, bio-distribution, and uptake efficiency	Deep Neural Networks, Graph Neural Networks, Reinforcement Learning	Modeling nanoparticle penetration across the blood–brain barrier	Improves delivery precision and reduces off-target effects	(83)
7	Safety and Toxicity Profiling	Assessment of cytotoxicity, immunogenicity, and hemocompatibility	QSAR Models, Classification Algorithms, Deep Learning	Predicting toxicity of metal or polymeric nanoparticles	Reduces in vivo testing and enhances patient safety	(84)
8	Process Optimization and Scale-Up	Correlation of manufacturing parameters with final product quality	Gaussian Process Regression, DoE + ML	Optimizing spray drying or emulsification conditions	Ensures batch-to-batch consistency and scalability	(85)

9	Image-Based Characterization	Automated particle analysis using imaging data	Convolutional Neural Networks (CNN), Image Recognition	Automated analysis of SEM/TEM images for morphology	Improves accuracy and speed of quality assessment	(86)
10	Intelligent Nanomedicine Design	Integration of ML for smart and personalized delivery systems	Multi-modal Deep Learning, Reinforcement Learning	AI-guided design of patient-specific nanocarriers	Enables precision therapy and predictive formulation design	(87)

7. Further Development of Autonomous Laboratories and Formulating Design

The development of machine-learning (ML)-augmented models for modified release drug delivery system (MRDDS) design represents a significant advancement in pharmaceutical science. This model allows for advanced computational methods to optimize formulation methodologies, resulting in increased delivery efficacy and patient adherence experiences(88).

7.1 Automation and High-Throughput Experimentation

The automation of MRDDS formulation involves the utilization of robotic systems, alongside analytically automated instruments to perform high-throughput studies. These tools provide rapid generation and assessment of multiple formulation variables, such as excipient combinations and processing methodologies. By methodically altering parameters and gathering comprehensive data, researchers can more effectively uncover optimal formulations compared to conventional methods(89).

7.2 The Intersection of Bayesian DL and Reinforcement Learning

Bayesian deep learning offers a probabilistic framework to define uncertainty in prediction, which is useful in the complex context of MRDDS. Its use, in conjunction with reinforcement learning (RL), can allow adaptive decision making to be established in experimental design. RL algorithms learn optimal formulation techniques from interactions with the environment, feedback from RL algorithms, and accordingly update models. The combination of these models can provide new formulations that are optimized for release profiles, stability, and manufacturability(90).

7.3 Generative Models & Discovery of New Formulation Spaces

Generative models, such as generative adversarial networks (GANs) and variational autoencoders (VAEs), are often used to explore new formulation spaces because generative models may be able to identify new combinations of excipients and relevant processing parameters that might not have been previously considered. By simply learning from existing data, generative models can generate innovative formulations with better medication release properties(91).

7.4 Active Learning in Intelligent Experimental Design

Active learning (AL) is an approach in which the model determines which experiments would be the most informative to pursue next. In the context of MRDDS, it can be used to identify formulation variants that yield the greatest insight into the relationship between formulation factors and drug release behaviours. This approach reduces the number of experiments recquired, saving time and resources while accelerating the delivery of optimized formulations.

The use of machine learning approaches in the design of modified release drug delivery systems is a groundbreaking step in pharmaceutical development. It enables researchers to design and optimize formulations that achieve specified therapeutic goals in an efficient manner through automation, probabilistic modelling, generative design and intelligent experimental design(92).

8. Challenges, Limitations, and Future Perspectives

The development of ML-enabled models to develop MRDDS presents great promise, but it involves many challenges and limitations that warrant careful consideration , as shown in Figure 1.

8.1 Lack of Data and Reproducibility Issues

Machine learning approaches rely on a large amount of high-quality data that can be used to predict drug release properties and formulate optimizations. However, data related to MRDDS experiments are moderately limited owing to the financial costs, complexity, and time required to conduct formulation studies. If limited, the lack of data may affect predictive machine learning model overfitting, the ability for model generalizability, and decrease the reproducibility of research findings from laboratory to laboratory. Examples of other variabilities that exacerbate reproducibility include the raw characteristics of the materials, laboratory conditions, and experimental protocols. Not only do these limitations affect transferable

Journal of Molecular Science
Journal of Molecular Science
Volume 35 Issue 4, Year of Publication 2025, Page 1436-1454
DoI-10.004687/1000-9035.2025.188

generalizability but also limit the overall robustness of machine learning predictions(93).

8.2 Transfer Learning and Minimal Data Solutions

To respond to a lack of data, transfer learning and other minimal data solutions are being increasingly used in pharmaceutical machine learning. Transfer learning takes advantage of information from similar formulations or datasets to enhance the predictions of new formulations based on limited experimental data(94).

8.3 Call for Standardisation and Open Access Formulation Databases

One of the primary barriers to the development of ML-based MRDDS is the inconsistent reporting of datasets and the lack of publicly available databases of formulation information being publicly available. Non-standardized reporting of formulation parameters, testing procedures, and metrics similarly hinders the repeatability and comparability of machine learning studies. A centralized , curated and standardized database of formulations would allow the pharmaceutical community to develop richer models and facilitate drug delivery system innovation(95).

8.4 Ethical Issues and Transparency of the Model

Machine-learning models will help inform important resource allocations in drug development, it is important to consider the ethics. Predictive model transparency, interpretability, and explainability are important factors for compliance and confidence in AI-based decisions. Nontransparent models that cannot transparently justify their predictions have implications for patient safety, regulatory approval, and the scientific methods. There is a need for advanced explainable methodologies for AI to ensure accountability and informed decision-making in MRDDS development(96).

8.5 Interdisciplinary Cooperation in Pharmaceutical AI Research

The successful implementation of machine learning in MRDDS will require productive interdisciplinary cooperation among pharmaceutical scientists, formulation chemists, data scientists, and regulatory scientists. Integrative knowledge, which is area knowledge combined with substantial computational methods to develop machine learning models, promotes both mathematical rigor and contemporary relevance to the field. Interdisciplinary cooperation also support the design of pragmatic, effective, and safe formulations through a liaison between predictions made using computational methods and their applications in the pharmaceutical industry(96).

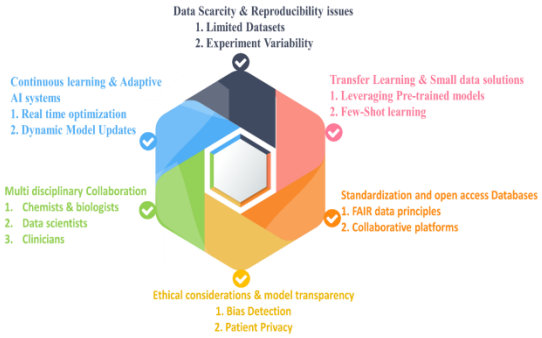

Figure 1 Key Challenges and Strategic Solutions in Machine Learning–Assisted Biopharmaceutical Development.

9.0 DISCUSSION:

A comprehensive data-driven framework was employed to utilize empirical experimentation and predictive modeling to improve the drug delivery design process using information learned from a thorough literature review of the scientific community's use of machine learning for drug design. The literature reviews provided insight into the limitations associated with the traditional trial and error approach to formulating drugs, as well as insight into the ability of applying machine learning techniques to model complex, non-linear relationships that help to control the rate of release of drugs from different modified release delivery systems. The original research articles also provided direction to researchers on how to systematically create data to support their empirical findings, how to determine which variables were relevant to formulating their drug product, and how to create a dataset containing dissolution experiment data for their new drug product. The studies presented here also provided insight into appropriate methods for partitioning datasets and highlighted the importance of validation of models to demonstrate robustness and generalizability, while preventing overfitting, particularly when working with limited pharmaceutical datasets. Studies comparing several different algorithms have demonstrated that the performance of advanced machine-learning algorithms in predicting drug release is generally superior to both classical kinetic and statistical models; thus, it will assist in selecting the appropriate algorithms to use in this current project. Additionally, an important aspect of past research was the necessity of directly comparing actual experimental drug release profiles with those generated from model-prediction outputs using standardised method performance metrics to support the predictive significance of the method. Overall, the information from both the review articles and the research articles included in this work will provide assistance in guiding future formulation developments toward creating an integrated experimental computational framework that

improves prediction quality, supports rational formulation optimisation, reduces the amount of time needed to develop formulations, and provides valuable information to decision makers involved in design and manufacture of modified drug delivery systems as noted in the relevant published literature.

10. CONCLUSION:

Over the past ten years, significant advancements have occurred in the incorporation of machine learning (ML) and artificial intelligence (AI) into pharmaceutical formulation research. From rudimentary statistical models to advanced deep-learning frameworks, these technologies have proven capable of precisely forecasting dissolution kinetics, optimizing excipient proportions, and engineering controlled-release systems with unparalleled efficacies. Data-driven techniques have transformed formulation development from empirical trial-and-error to predictive, model-informed experimentation consistent with quality-by-design (QbD) and regulatory standards. The evolving paradigm for data-centric pharmaceutical development highlights the establishment of interoperable formulation databases, transparent and interpretable machine learning models, and the seamless integration of computational predictions with automated laboratory equipment. In the coming years, advancements in deep learning, transfer learning, and digital twins are anticipated to facilitate the real-time optimization of modified-release drug systems, whereby the design, simulation, and manufacturing function synergistically within an AI-enabled framework. Ultimately, machine learning-driven formulation science is set to transform pharmaceutical innovation, enabling expedited, intelligent, and more sustainable production of personalized, modified-release therapies.

REFERENCES:

1. Charoo NA, Mohamed EM, Eltigani M, Khan MA, Nutan MTH, Rahman Z. Drug and formulation development processes. In: From Current to Future Trends in Pharmaceutical Technology [Internet]. Elsevier; 2024 [cited 2025 Nov 9]. p. 257–92. Available from: https://linkinghub.elsevier.com/retrieve/pii/B978032391111500007X
2. Moon C, Oh E. Rationale and strategies for formulation development of oral fixed dose combination drug products. Journal of Pharmaceutical Investigation [Internet]. 2016 Dec [cited 2025 Nov 9];46(7):615–31. Available from: http://link.springer.com/10.1007/s40005-016-0286-4
3. Desai D, Wang J, Wen H, Li X, Timmins P. Formulation design, challenges, and development considerations for fixed dose combination (FDC) of oral solid dosage forms. Pharmaceutical Development and Technology [Internet]. 2013 Dec [cited 2025 Nov 9];18(6):1265–76. Available from: http://www.tandfonline.com/doi/full/10.3109/10837450.2012.660699
4. Dong Y, Yang T, Xing Y, Du J, Meng Q. Data-Driven Modeling Methods and Techniques for Pharmaceutical Processes. Processes [Internet]. 2023 July 13 [cited 2025 Nov 9];11(7):2096. Available from: https://www.mdpi.com/2227-9717/11/7/2096
5. Selvaraj C, Chandra I, Singh SK. Artificial intelligence and machine learning approaches for drug design: challenges and opportunities for the pharmaceutical industries. Mol Divers [Internet]. 2022 June [cited 2025 Nov 9];26(3):1893–913. Available from: https://link.springer.com/10.1007/s11030-021-10326-z
6. Karthikeyan A, Priyakumar UD. Artificial intelligence: machine learning for chemical sciences. J Chem Sci [Internet]. 2022 Mar [cited 2025 Nov 9];134(1):2. Available from: https://link.springer.com/10.1007/s12039-021-01995-2
7. Wang H, Tian Z, Wang L, Wang H, Zhang Y, Ban W. Hybrid computational model for analysis of drug release from biomaterial matrix: Machine learning aided material design. Case Studies in Thermal Engineering [Internet]. 2025 Sept [cited 2025 Nov 9];73:106505. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2214157X25007658
8. Aundhia C, Parmar G, Talele C, Shah N, Talele D. Impact of Artificial Intelligence on Drug Development and Delivery. CTMC [Internet]. 2025 Apr [cited 2025 Nov 9];25(10):1165–84. Available from: https://www.eurekaselect.com/232964/article
9. Shah SMA, Soomro A, Hussain K, Rahu MA, Karim S. Innovative Machine Learning Solutions: Investigating Algorithms and Applications. JAET [Internet]. 2024 June 30 [cited 2025 Nov 9];8(1):32–51. Available from: https://jae-tech.com/index.php/jaet/article/view/146
10. Bannigan P, Aldeghi M, Bao Z, Häse F, Aspuru-Guzik A, Allen C. Machine learning directed drug formulation development. Advanced Drug Delivery Reviews [Internet]. 2021 Aug [cited 2025 Nov 9];175:113806. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0169409X21001800
11. Yang X, Song Z, King I, Xu Z. A Survey on Deep Semi-Supervised Learning. IEEE Trans Knowl Data Eng [Internet]. 2023 Sept 1 [cited 2025 Nov 9];35(9):8934–54. Available from: https://ieeexplore.ieee.org/document/9941371/
12. Ros F, Riad R. Feature and Dimensionality Reduction for Clustering with Deep Learning [Internet]. Cham: Springer Nature Switzerland; 2024 [cited 2025 Nov 9]. (Unsupervised and Semi-Supervised Learning). Available from: https://link.springer.com/10.1007/978-3-031-48743-9
13. Kumari M, Tiwari N, Chandra S, Subbarao N. Comparative analysis of machine learning based QSAR models and molecular docking studies to screen potential anti-tubercular inhibitors against InhA of mycobacterium tuberculosis. IJCBDD [Internet]. 2018 [cited 2025 Nov 9];11(3):209. Available from: http://www.inderscience.com/link.php?id=94630
14. Le-Thi-Thu H, Marrero-Ponce Y, Casañola-Martin GM, Cardoso GC, Chávez M del C, Garcia MM, et al. A Comparative Study of Nonlinear Machine Learning for the “In Silico” Depiction of Tyrosinase Inhibitory Activity from Molecular Structure. Molecular Informatics [Internet]. 2011 June [cited 2025 Nov 9];30(6–7):527–37. Available from: https://onlinelibrary.wiley.com/doi/10.1002/minf.201100021
15. Chi HM, Moskowitz H, Ersoy OK, Altinkemer K, Gavin PF, Huff BE, et al. Machine learning and genetic algorithms in pharmaceutical development and manufacturing processes. Decision Support Systems [Internet]. 2009 Dec [cited 2025 Nov 9];48(1):69–80. Available from: https://linkinghub.elsevier.com/retrieve/pii/S016792360900147X
16. Krmar J, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B. Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure - retention relationships modelling in micellar liquid chromatography. Journal of Chromatography A [Internet]. 2020 July [cited 2025 Nov 9];1623:461146. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021967320303861
17. Wei A, Yu K, Dai F, Gu F, Zhang W, Liu Y. Application of Tree-Based Ensemble Models to Landslide Susceptibility Mapping: A Comparative Study. Sustainability [Internet]. 2022 May 23 [cited 2025 Nov 9];14(10):6330.

Available from: https://www.mdpi.com/2071-1050/14/10/6330
18. Colkesen I, Sahin EK, Kavzoglu T. Susceptibility mapping of shallow landslides using kernel-based Gaussian process, support vector machines and logistic regression. Journal of African Earth Sciences [Internet]. 2016 June [cited 2025 Nov 9];118:53–64. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1464343X16300747
19. Patra RC, Saritha G, Raghuwanshi GA, Parthiban P, Ashok V. Recent Advances, Challenges, and Applications of Deep Learning in Healthcare Systems for Medical Diagnosis and Treatment. In: 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS) [Internet]. Tashkent, Uzbekistan: IEEE; 2023 [cited 2025 Nov 9]. p. 716–24. Available from: https://ieeexplore.ieee.org/document/10390124/
20. Yang C, Lan S, Wang L, Shen W, Huang GGQ. Big Data Driven Edge-Cloud Collaboration Architecture for Cloud Manufacturing: A Software Defined Perspective. IEEE Access [Internet]. 2020 [cited 2025 Nov 9];8:45938–50. Available from: https://ieeexplore.ieee.org/document/9020166/
21. Kavitha V, Suresh K, Priyadharshini G, Begum SR, Vidhya R. Interpretable Machine Learning Techniques. In: Kumar A, Ananth Kumar T, Das P, Sharma C, Dubey AK, editors. Explainable Artificial Intelligence in the Healthcare Industry [Internet]. 1st ed. Wiley; 2025 [cited 2025 Nov 9]. p. 185–207. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9781394249312.ch9
22. Ghislat G, Hernandez-Hernandez S, Piyawajanusorn C, Ballester PJ. Data-centric challenges with the application and adoption of artificial intelligence for drug discovery. Expert Opinion on Drug Discovery [Internet]. 2024 Nov [cited 2025 Nov 9];19(11):1297–307. Available from: https://www.tandfonline.com/doi/full/10.1080/17460441.2024.2403639
23. Di Paolo Emilio M. Data Acquisition Systems: From Fundamentals to Applied Design [Internet]. New York, NY: Springer New York; 2013 [cited 2025 Nov 9]. Available from: https://link.springer.com/10.1007/978-1-4614-4214-1
24. Ming-Syan Chen, Jiawei Han, Yu PS. Data mining: an overview from a database perspective. IEEE Trans Knowl Data Eng [Internet]. 1996 Dec [cited 2025 Nov 9];8(6):866–83. Available from: http://ieeexplore.ieee.org/document/553155/
25. Hameed M, Naumann F. Data Preparation: A Survey of Commercial Tools. SIGMOD Rec [Internet]. 2020 Dec 17 [cited 2025 Nov 9];49(3):18–29. Available from: https://dl.acm.org/doi/10.1145/3444831.3444835
26. Uddin MF, Lee J, Rizvi S, Hamada S. Proposing Enhanced Feature Engineering and a Selection Model for Machine Learning Processes. Applied Sciences [Internet]. 2018 Apr 20 [cited 2025 Nov 9];8(4):646. Available from: https://www.mdpi.com/2076-3417/8/4/646
27. Ilyés K, Kovács NK, Balogh A, Borbás E, Farkas B, Casian T, et al. The applicability of pharmaceutical polymeric blends for the fused deposition modelling (FDM) 3D technique: Material considerations–printability–process modulation, with consecutive effects on in vitro release, stability and degradation. European Journal of Pharmaceutical Sciences [Internet]. 2019 Mar [cited 2025 Nov 9];129:110–23. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0928098718305542
28. Kalasz H, Antal I. Drug Excipients. CMC [Internet]. 2006 Sept 1 [cited 2025 Nov 9];13(21):2535–63. Available from: http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8673&volume=13&issue=21&spage=2535
29. Mudalgi S. Optimal Data Splitting Methods [Internet]. VCU Theses and Dissertations. VCU Libraries; 2025 [cited 2025 Nov 9]. Available from: https://scholarscompass.vcu.edu/etd/8134
30. Ali Y, Awwad E, Al-Razgan M, Maarouf A. Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity. Processes [Internet]. 2023 Jan 21 [cited 2025 Nov 9];11(2):349. Available from: https://www.mdpi.com/2227-9717/11/2/349
31. Kaur G, Saini S. Comparative analysis of RMSE and MAP metrics for evaluating CNN and LSTM models. In Mohali, India; 2024 [cited 2025 Nov 9]. p. 040003. Available from: https://pubs.aip.org/aip/acp/article-lookup/doi/10.1063/5.0221565
32. Rudin C, Chen C, Chen Z, Huang H, Semenova L, Zhong C. Interpretable machine learning: Fundamental principles and 10 grand challenges. Statist Surv [Internet]. 2022 Jan 1 [cited 2025 Nov 9];16(none). Available from: https://projecteuclid.org/journals/statistics-surveys/volume-16/issue-none/Interpretable-machine-learning-Fundamental-principles-and-10-grand-challenges/10.1214/21-SS133.full
33. Dey H, Arya N, Mathur H, Chatterjee N, Jadon R. Exploring the Role of Artificial Intelligence and Machine Learning in Pharmaceutical Formulation Design. ijnrph [Internet]. 2024 June 30 [cited 2025 Nov 9];30–41. Available from: https://www.ijnrph.com/index.php/ijnrph/article/view/67
34. Desai J, Dhameliya P, Patel S. Optimizing critical quality attributes of fast disintegrating tablets using artificial neural networks: a scientific benchmark study. Drug Development and Industrial Pharmacy [Internet]. 2024 Dec [cited 2025 Nov 9];50(12):995–1007. Available from: https://www.tandfonline.com/doi/full/10.1080/03639045.2024.2434640
35. Vass P, Szabó E, Domokos A, Hirsch E, Galata D, Farkas B, et al. Scale-up of electrospinning technology: Applications in the pharmaceutical industry. WIREs Nanomed Nanobiotechnol [Internet]. 2020 July [cited 2025 Nov 9];12(4):e1611. Available from: https://wires.onlinelibrary.wiley.com/doi/10.1002/wnan.1611
36. Keraliya RA, Patel C, Patel P, Keraliya V, Soni TG, Patel RC, et al. Osmotic Drug Delivery System as a Part of Modified Release Dosage Form. ISRN Pharmaceutics [Internet]. 2012 July 17 [cited 2025 Nov 9];2012:1–9. Available from: https://www.hindawi.com/journals/isrn/2012/528079/
37. Topuz F, Uyar T. Recent Advances in Cyclodextrin-Based Nanoscale Drug Delivery Systems. WIREs Nanomed Nanobiotechnol [Internet]. 2024 Nov [cited 2025 Nov 9];16(6):e1995. Available from: https://wires.onlinelibrary.wiley.com/doi/10.1002/wnan.1995
38. Suarez-Sharp S, Li M, Duan J, Shah H, Seo P. Regulatory Experience with In Vivo In Vitro Correlations (IVIVC) in New Drug Applications. AAPS J [Internet]. 2016 Nov [cited 2025 Nov 9];18(6):1379–90. Available from: http://link.springer.com/10.1208/s12248-016-9966-2
39. Han R, Xiong H, Ye Z, Yang Y, Huang T, Jing Q, et al. Predicting physical stability of solid dispersions by machine learning techniques. Journal of Controlled Release [Internet]. 2019 Oct [cited 2025 Nov 9];311–312:16–25. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0168365919305140
40. Kumar A, Gupta GD, Raikwar S. Artificial Intelligence Technologies used for the Assessment of PharmaceuticalExcipients. CPD [Internet]. 2024 Feb [cited 2025 Nov 9];30(6):407–9. Available from: https://www.eurekaselect.com/226446/article
41. Khuat TT, Bassett R, Otte E, Grevis-James A, Gabrys B. Applications of Machine Learning in Biopharmaceutical Process Development and Manufacturing: Current Trends, Challenges, and Opportunities [Internet]. arXiv; 2023 [cited 2025 Nov 9]. Available from: https://arxiv.org/abs/2310.09991
42. Akbarian M, Chen SH. Instability Challenges and Stabilization Strategies of Pharmaceutical Proteins. Pharmaceutics [Internet]. 2022 Nov 20 [cited 2025 Nov 9];14(11):2533. Available from: https://www.mdpi.com/1999-4923/14/11/2533
43. Luo Y, Cai J. Deep Learning for the Prediction of Protein Sequence, Structure, Function, and Interaction: Applications, Challenges, and Future Directions. CP [Internet]. 2025 June [cited 2025 Nov 9];21(6):561–79. Available from: https://www.eurekaselect.com/240675/article
44. Kim K, Kim DK, Noh J, Kim M. Stable Forecasting of Environmental Time Series via Long Short Term Memory Recurrent Neural Network. IEEE Access [Internet]. 2018 [cited 2025 Nov 9];6:75216–28. Available from: https://ieeexplore.ieee.org/document/8558483/
45. Bao Z, Kim J, Le Devedec F, Clasky A, Allen C.

Polymer microparticles in an evolving drug delivery landscape: challenges and the role of machine learning. International Journal of Pharmaceutics [Internet]. 2025 Sept [cited 2025 Nov 9];682:125906. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378517325007434
46. Jia L, Yarlagadda R, Reed CC. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. Zhang Y, editor. PLoS ONE [Internet]. 2015 Sept 11 [cited 2025 Nov 9];10(9):e0138022. Available from: https://dx.plos.org/10.1371/journal.pone.0138022
47. Zulbeari N, Wang F, Mustafova SS, Parhizkar M, Holm R. Machine learning strengthened formulation design of pharmaceutical suspensions. International Journal of Pharmaceutics [Internet]. 2025 Jan [cited 2025 Nov 9];668:124967. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378517324012018
48. Wang B, Zhang T, Liu Q, Sutcharitchan C, Zhou Z, Zhang D, et al. Elucidating the role of artificial intelligence in drug development from the perspective of drug-target interactions. Journal of Pharmaceutical Analysis [Internet]. 2025 Mar [cited 2025 Nov 9];15(3):101144. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2095177924002417
49. Jackson PT, Wang Y, Knight S, Chen H, Dorval T, Brown M, et al. Phenotypic Profiling of High Throughput Imaging Screens with Generic Deep Convolutional Features. In: 2019 16th International Conference on Machine Vision Applications (MVA) [Internet]. Tokyo, Japan: IEEE; 2019 [cited 2025 Nov 9]. p. 1–4. Available from: https://ieeexplore.ieee.org/document/8757871/
50. Lai PK, Gallegos A, Mody N, Sathish HA, Trout BL. Machine learning prediction of antibody aggregation and viscosity for high concentration formulation development of protein therapeutics. mAbs [Internet]. 2022 Dec 31 [cited 2025 Nov 9];14(1):2026208. Available from: https://www.tandfonline.com/doi/full/10.1080/19420862.2022.2026208
51. Pinzón JD, Colomé DG. Real-time multi-state classification of short-term voltage stability based on multivariate time series machine learning. International Journal of Electrical Power & Energy Systems [Internet]. 2019 June [cited 2025 Nov 9];108:402–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0142061518320295
52. Tsuchiya Y, Tomii K. Neural networks for protein structure and function prediction and dynamic analysis. Biophys Rev [Internet]. 2020 Apr [cited 2025 Nov 9];12(2):569–73. Available from: http://link.springer.com/10.1007/s12551-020-00685-6
53. Walsh I, Myint M, Nguyen-Khuong T, Ho YS, Ng SK, Lakshmanan M. Harnessing the potential of machine learning for advancing "Quality by Design" in biomanufacturing. mAbs [Internet]. 2022 Dec 31 [cited 2025 Nov 9];14(1):2013593. Available from: https://www.tandfonline.com/doi/full/10.1080/19420862.2021.2013593
54. Ayvaz S, Alpay K. Predictive maintenance system for production lines in manufacturing: A machine learning approach using IoT data in real-time. Expert Systems with Applications [Internet]. 2021 July [cited 2025 Nov 9];173:114598. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0957417421000397
55. Nam Y, Kim J, Jung SH, Woerner J, Suh EH, Lee D gi, et al. Harnessing Artificial Intelligence in Multimodal Omics Data Integration: Paving the Path for the Next Frontier in Precision Medicine. Annual Review of Biomedical Data Science [Internet]. 2024 Aug 23 [cited 2025 Nov 9];7(1):225–50. Available from: https://www.annualreviews.org/content/journals/10.1146/annurev-biodatasci-102523-103801
56. Vardhanapu M, Chaganti PK, Tarigopula P. Characterization and machine learning-based parameter estimation in MQL machining of a superalloy for developed green nano-metalworking fluids. J Braz Soc Mech Sci Eng [Internet]. 2023 Mar [cited 2025 Nov 9];45(3):154. Available from: https://link.springer.com/10.1007/s40430-023-04078-0
57. Singh, Singha P. Effect of Modifications in Poly (Lactide-co-Glycolide) (PLGA) on Drug Release and Degradation Characteristics: A Mini Review. CDD [Internet]. 2021 Dec [cited 2025 Nov 9];18(10):1378–90. Available from: https://www.eurekaselect.com/193278/article
58. Simon AP, Ferreira CH, Santos VAQ, Rodrigues A, Santos JS, Trivinho-Strixino F, et al. Multi-step cefazolin sodium release from bioactive TiO2 nanotubes: Surface and polymer coverage effects. Journal of Materials Research [Internet]. 2021 Apr 14 [cited 2025 Nov 9];36(7):1510–23. Available from: https://link.springer.com/10.1557/s43578-021-00202-9
59. Fahmy OM, Eissa RA, Mohamed HH, Eissa NG, Elsabahy M. Machine learning algorithms for prediction of entrapment efficiency in nanomaterials. Methods [Internet]. 2023 Oct [cited 2025 Nov 9];218:133–40. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202323001378
60. Hosni Z, Achour S, Saadi F, Chen Y, Al Qaraghuli M. Machine learning-driven nanoparticle toxicity. Ecotoxicology and Environmental Safety [Internet]. 2025 July [cited 2025 Nov 9];299:118340. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0147651325006761
61. Ye F, Yang J. A Deep Neural Network Model for Speaker Identification. Applied Sciences [Internet]. 2021 Apr 16 [cited 2025 Nov 9];11(8):3603. Available from: https://www.mdpi.com/2076-3417/11/8/3603
62. Seegobin N, Abdalla Y, Li G, Murdan S, Shorthouse D, Basit AW. Optimising the production of PLGA nanoparticles by combining design of experiment and machine learning. International Journal of Pharmaceutics [Internet]. 2024 Dec [cited 2025 Nov 9];667:124905. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378517324011396
63. Sivadasan D, Sultan MH, Madkhali O, Almoshari Y, Thangavel N. Polymeric Lipid Hybrid Nanoparticles (PLNs) as Emerging Drug Delivery Platform—A Comprehensive Review of Their Properties, Preparation Methods, and Therapeutic Applications. Pharmaceutics [Internet]. 2021 Aug 18 [cited 2025 Nov 9];13(8):1291. Available from: https://www.mdpi.com/1999-4923/13/8/1291
64. Zawbaa HM, Szlęk J, Grosan C, Jachowicz R, Mendyk A. Computational Intelligence Modeling of the Macromolecules Release from PLGA Microspheres—Focus on Feature Selection. Omri A, editor. PLoS ONE [Internet]. 2016 June 17 [cited 2025 Nov 9];11(6):e0157610. Available from: https://dx.plos.org/10.1371/journal.pone.0157610
65. Maksimenko O, Malinovskaya J, Shipulo E, Osipova N, Razzhivina V, Arantseva D, et al. Doxorubicin-loaded PLGA nanoparticles for the chemotherapy of glioblastoma: Towards the pharmaceutical development. International Journal of Pharmaceutics [Internet]. 2019 Dec [cited 2025 Nov 9];572:118733. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0378517319307781
66. Maharjan R, Kim KH, Lee K, Han HK, Jeong SH. Machine learning-driven optimization of mRNA-lipid nanoparticle vaccine quality with XGBoost/Bayesian method and ensemble model approaches. Journal of Pharmaceutical Analysis [Internet]. 2024 Nov [cited 2025 Nov 9];14(11):100996. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2095177924000935
67. Correia AC, Moreira JN, Sousa Lobo JM, Silva AC. Design of experiment (DoE) as a quality by design (QbD) tool to optimise formulations of lipid nanoparticles for nose-to-brain drug delivery. Expert Opinion on Drug Delivery [Internet]. 2023 Dec 2 [cited 2025 Nov 9];20(12):1731–48. Available from: https://www.tandfonline.com/doi/full/10.1080/17425247.2023.2274902
68. Ge C, Luo L, Zhang J, Meng X, Chen Y. FRL: An Integrative Feature Selection Algorithm Based on the Fisher Score, Recursive Feature Elimination, and Logistic Regression to Identify Potential Genomic Biomarkers. Huang T, editor. BioMed Research International [Internet]. 2021 Jan [cited 2025 Nov 9];2021(1):4312850. Available from: https://onlinelibrary.wiley.com/doi/10.1155/2021/4312850
69. Figueroa Barraza J, López Droguett E, Martins MR. Towards Interpretable Deep Learning: A Feature Selection

Framework for Prognostics and Health Management Using Deep Neural Networks. Sensors [Internet]. 2021 Sept 1 [cited 2025 Nov 9];21(17):5888. Available from: https://www.mdpi.com/1424-8220/21/17/5888
70. Wang F, Elbadawi M, Tsilova SL, Gaisford S, Basit AW, Parhizkar M. Machine learning predicts electrospray particle size. Materials & Design [Internet]. 2022 July [cited 2025 Nov 9];219:110735. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0264127522003574
71. Sarmah DK. A Survey on the Latest Development of Machine Learning in Genetic Algorithm and Particle Swarm Optimization. In: Kulkarni AJ, Satapathy SC, editors. Optimization in Machine Learning and Applications [Internet]. Singapore: Springer Singapore; 2020 [cited 2025 Nov 9]. p. 91–112. (Algorithms for Intelligent Systems). Available from: http://link.springer.com/10.1007/978-981-15-0994-0_6
72. Emami S, Martínez-Muñoz G. Deep Learning for Multi-Output Regression Using Gradient Boosting. IEEE Access [Internet]. 2024 [cited 2025 Nov 9];12:17760–72. Available from: https://ieeexplore.ieee.org/document/10415168/
73. Hoseini B, Jaafari MR, Golabpour A, Rahmatinejad Z, Karimi M, Eslami S. Machine Learning-Driven Advancements in Liposomal Formulations for Targeted Drug Delivery: A Narrative Literature Review. CDD [Internet]. 2025 Nov [cited 2025 Nov 9];22(9):1287–305. Available from: https://www.eurekaselect.com/231366/article
74. Vaněk J, Zelinka J, Soutner D, Psutka J. A Regularization Post Layer: An Additional Way How to Make Deep Neural Networks Robust. In: Camelin N, Estève Y, Martín-Vide C, editors. Statistical Language and Speech Processing [Internet]. Cham: Springer International Publishing; 2017 [cited 2025 Nov 9]. p. 204–14. (Lecture Notes in Computer Science; vol. 10583). Available from: https://link.springer.com/10.1007/978-3-319-68456-7_17
75. Pan L, Li G, Zhu T, Liu D, Wang Y, Lu Y. Physics-Informed Machine Learning in Design and Manufacturing: Status and Challenges. Journal of Computing and Information Science in Engineering [Internet]. 2025 Oct 9 [cited 2025 Nov 9];1–29. Available from: https://asmedigitalcollection.asme.org/computingengineering/article/doi/10.1115/1.4070100/1225302/Physics-Informed-Machine-Learning-in-Design-and
76. Liu JP, Ma MC, Chow SC. Statistical Evaluation of Similarity Factor f2 as a Criterion for Assessment of Similarity Between Dissolution Profiles. Drug Information J [Internet]. 1997 Oct [cited 2025 Nov 9];31(4):1255–71. Available from: http://link.springer.com/10.1177/009286159703100426
77. Stevens RE, Gray V, Dorantes A, Gold L, Pham L. Scientific and Regulatory Standards for Assessing Product Performance Using the Similarity Factor, f2. AAPS J [Internet]. 2015 Mar [cited 2025 Nov 9];17(2):301–6. Available from: http://link.springer.com/10.1208/s12248-015-9723-y
78. Liu J, Zhang Z, Li X, Zong M, Wang Y, Wang S, et al. Machine learning assisted phase and size-controlled synthesis of iron oxide particles. Chemical Engineering Journal [Internet]. 2023 Oct [cited 2025 Nov 9];473:145216. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1385894723039475
79. Bahrami P, Kazemi P, Mahdavi S, Ghobadi H. A novel approach for modeling and optimization of surfactant/polymer flooding based on Genetic Programming evolutionary algorithm. Fuel [Internet]. 2016 Sept [cited 2025 Nov 9];179:289–98. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0016236116301375
80. Fahmy OM, Eissa RA, Mohamed HH, Eissa NG, Elsabahy M. Machine learning algorithms for prediction of entrapment efficiency in nanomaterials. Methods [Internet]. 2023 Oct [cited 2025 Nov 9];218:133–40. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1046202323001378
81. Alshahrani SM, Alotaibi HF, Alqarni M. Modeling and validation of drug release kinetics using hybrid method for prediction of drug efficiency and novel formulations. Front Chem [Internet]. 2024 June 21 [cited 2025 Nov 9];12:1395359. Available from: https://www.frontiersin.org/articles/10.3389/fchem.2024.1395359/full
82. Bi M, Chen H, Zuo M. Comprehensive quality grading and dynamic prediction of physicochemical indicators of maize during storage based on clustering and time-series prediction models. Journal of Cereal Science [Internet]. 2025 May [cited 2025 Nov 9];123:104187. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0733521025000852
83. Das S, Mazumdar H, Khondakar KR, Kaushik A. Machine learning integrated graphene oxide-based diagnostics, drug delivery, analytical approaches to empower cancer diagnosis. BMEMat [Internet]. 2025 Mar [cited 2025 Nov 9];3(1):e12117. Available from: https://onlinelibrary.wiley.com/doi/10.1002/bmm2.12117
84. Tosca EM, Bartolucci R, Magni P, Poggesi I. Modeling approaches for reducing safety-related attrition in drug discovery and development: a review on myelotoxicity, immunotoxicity, cardiovascular toxicity, and liver toxicity. Expert Opinion on Drug Discovery [Internet]. 2021 Nov 2 [cited 2025 Nov 9];16(11):1365–90. Available from: https://www.tandfonline.com/doi/full/10.1080/17460441.2021.1931114
85. Pathak M, Pokhriyal P, Gandhi I, Khambhampaty S. Implementation of chemometrics, design of experiments, and neural network analysis for prior process knowledge assessment, failure modes and effect analysis, scale-down model development, and process characterization for a chromatographic purification of Teriparatide. Biotechnology Progress [Internet]. 2022 May [cited 2025 Nov 9];38(3):e3252. Available from: https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3252
86. Pathak M, Pokhriyal P, Gandhi I, Khambhampaty S. Implementation of chemometrics, design of experiments, and neural network analysis for prior process knowledge assessment, failure modes and effect analysis, scale-down model development, and process characterization for a chromatographic purification of Teriparatide. Biotechnology Progress [Internet]. 2022 May [cited 2025 Nov 9];38(3):e3252. Available from: https://aiche.onlinelibrary.wiley.com/doi/10.1002/btpr.3252
87. Tivay S, Taghipoor M, Hosseini H, Saeedzadeh M, Valanik S. Next-Generation Drug Delivery: Smart Nanomaterials for Precision Healthcare(focusing oral cancer). JODHN [Internet]. 2025 [cited 2025 Nov 9];2(2):6–18. Available from: https://jodhn.com/index.php/jodhn/article/view/14
88. Ferreira Da Silva R, Moore Ii R, Mintz B, Advincula R, Alnajjar A, Baldwin L, et al. Shaping the Future of Self-Driving Autonomous Laboratories Workshop [Internet]. 2024 Nov [cited 2025 Nov 9] p. ORNL/TM--2024/3714, 2481197. Report No.: ORNL/TM--2024/3714, 2481197. Available from: https://www.osti.gov/servlets/purl/2481197/
89. Vervoort N, Goossens K, Baeten M, Chen Q. Recent advances in analytical techniques for high throughput experimentation. Analytical Science Advances [Internet]. 2021 Apr [cited 2025 Nov 9];2(3–4):109–27. Available from: https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/ansa.202000155
90. Shi Y, Wei P, Feng K, Feng DC, Beer M. A survey on machine learning approaches for uncertainty quantification of engineering systems. Mach Learn Comput Sci Eng [Internet]. 2025 June [cited 2025 Nov 9];1(1):11. Available from: https://link.springer.com/10.1007/s44379-024-00011-x
91. Ramalakshmi S, Asha G. Exploring Generative AI: Models, Applications, and Challenges in Data Synthesis. Asian J Res Com Sci [Internet]. 2024 Dec 13 [cited 2025 Nov 9];17(12):123–36. Available from: https://journalajrcos.com/index.php/AJRCOS/article/view/533
92. Sayin B, Krivosheev E, Yang J, Passerini A, Casati F. A review and experimental analysis of active learning over crowdsourced data. Artif Intell Rev [Internet]. 2021 Oct [cited 2025 Nov 9];54(7):5283–305. Available from: https://link.springer.com/10.1007/s10462-021-10021-3
93. Kumar SA, Ananda Kumar TD, Beeraka NM, Pujar GV, Singh M, Narayana Akshatha HS, et al. Machine Learning and Deep Learning in Data-Driven Decision Making of Drug Discovery and Challenges in High-Quality Data Acquisition in

the Pharmaceutical Industry. Future Med Chem [Internet]. 2022 Feb [cited 2025 Nov 9];14(4):245–70. Available from: https://www.tandfonline.com/doi/full/10.4155/fmc-2021-0243

94. Hutchinson ML, Antono E, Gibbons BM, Paradiso S, Ling J, Meredig B. Overcoming data scarcity with transfer learning [Internet]. arXiv; 2017 [cited 2025 Nov 9]. Available from: https://arxiv.org/abs/1711.05099

95. Mirakhori F, Niazi SK. Harnessing the AI/ML in Drug and Biological Products Discovery and Development: The Regulatory Perspective. Pharmaceuticals [Internet]. 2025 Jan 3 [cited 2025 Nov 9];18(1):47. Available from: https://www.mdpi.com/1424-8247/18/1/47

96. Mathrani A, Susnjak T, Ramaswami G, Barczak A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Computers and Education Open [Internet]. 2021 Dec [cited 2025 Nov 9];2:100060. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2666557321000318