Current Trends in Biotechnology and Pharmacy
Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

Prediction of Preformulation Using AI and Machine Learing Approches

Harika N[1], Venna R Surya Anusha[1]*, Anu Pravallika Janipalli[2], Koppala RVS Chaitanya[3], and Kothapalli Sandeep[3]

[1]Gokaraju Rangaraju College of Pharmacy, Bachupally, Hyderabad-500090, Telangana, India
[2]Aditya College of Pharmacy, Aditya University, Surampalem-533437, Andhra Pradesh, India
[3]Sarojini Naidu Vanita Pharmacy Maha Vidyalaya, Tarnaka, Secunderabad-500017, Telangana, India
*Corresponding author: rajeswarianusha@gmail.com

Abstract

The integration of Artificial Intelligence (AI) and Machine Learning (ML) into preformulation studies is transforming the way pharmaceutical research is carried out. This process speeds up the transition from drug discovery to formulation. Preformulation is the first step in drug development. It involves a thorough examination of physicochemical characteristics. This review highlights the first-in-class applications of AI and ML in addressing key challenges, such as predicting solubility, determining pKa values, assessing partition coefficients, testing polymorphism, ensuring stability, and checking drug-excipient compatibility. Neural networks, support vector machines, and gradient boosting models are examples of advanced algorithms that improve the accuracy of predictive models. These tools help accelerate formulation design and lower risks. As much as there are challenges involving data limitations and regulatory regulations, the future of AI and ML in the pharmaceutics industry is significant, promoting innovation, minimizing costs, and maximizing therapeutic potency. This summary intends to give a summary overview of the approaches, achievements, and novel trends in AI/ML applications for pharmaceutical formulation designing.

Keywords: Artificial intelligence; machine learning; artificial neural networks; molecular descriptors; preformulation studies; QSPR and QSAR model; molecular docking.

Introduction

Preformulation studies synonymously named "learning before doing" were identified as a legitimate scientific discipline in the mid-20th century, reaching full swing in the 1950s and 1960s. Preformulation studies are absolutely critical during the first steps of drug development as they bring a lot of important information regarding the physicochemical nature of the active pharmaceutical ingredients, their salt forms, and the compatibility with excipients. They are the prerequisites for new chemical entities (NCEs) and the drug formulation process to be connected, and at the same time, they are the ground on which the drug formulations are built. Such studies, which are by their nature very precise, identify main formulation parameters (1). Preformulation is a close study of a drug's physicochemical characterization, both alone and in combination with excipients, with the final objective being the creation of safe, effective, and stable dosage forms.

The primary goal is to provide critical information that facilitates the design of bioavailable and scalable pharmaceutical products. This means data on the drug candidate, excipients, and even packaging materials' physicochemical and biopharmaceutical properties (2). Through predictive modeling, optimization software, and data-driven decision-making, AI enhances the safety, stability, and efficiency of formulations significantly increasing the likelihood of delivering high-quality, effective drugs to

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

patients in an efficient manner (3). Meta-analysis and recent studies depict that such technology is accountable for accelerated development timeliness and rising success rates for both preclinical and clinical phases (4,5). In this article, the rising influence of AI in drug development has been discussed with specific mention of its use for predictive preformulation studies. It addresses various AI techniques and software-including ML and Artificial Neural Networks (ANNs) and their role in drug discovery, formulation development, development, and enhancement of drug delivery systems (6). Because drug discovery was the starting point of AI and ML utilization in pharmaceutical research, it is introduced succinctly in this review.

Problems faced during preformulation studies

Poor solubility and permeability greatly limit how well a drug is absorbed in the body, ultimately lowering its bioavailability. Polymorphism further complicates drug performance as different crystal forms can impact dissolution rate and stability. Drug-excipient incompatibility is a severe risk of formulation failure, which can affect both efficacy and safety. Poor mechanical properties also weaken the manufacturing process by affecting tablet compression, flowability, and handling. Analytical challenges limit proper characterization of the drug substance and product, and therefore quality control becomes challenging. Scale-up issues during production may lead to batch-to-batch variability and affect the consistency of the product. Moreover, environmental degradation may compromise the stability of the formulation, and thus sophisticated predictive tools need to be implemented in order to ensure robust and reliable drug development (7). Traditional preformulation is experimental screening of solubility, stability, and compatibility, which is time and resource intensive. Key limitations include inability to predict solid-state changes, bulk excipient testing, and non-scalability. AI-based predictive models learn complex nonlinear relationships between molecular structure and preformulation responses. SVM models accurately classify compatible excipients, ANN models accurately predict drug degradation kinetics, and QSPR-based regressions accurately predict solubility within acceptable RMSE values. This integration converts preformulation from empirical observation to data-driven prediction, accelerating formulation development.

AI and ML in the pharma industry

A broad and interdisciplinary topic, AI allows machines to simulate human thought, learn, and reason. AI is changing every sector, including pharmaceutical research. Two of the major branches in AI, ML and DL have transformed drug discovery by combining computational drug design with intelligent decision-making tools, thereby streamlining the whole drug development process (8). AI is also crucial in medicine through medical diagnostics, tracking disease outbreaks, and enabling personalized medicine (9). It is also necessary to be aware of the physicochemical properties of drugs such as solubility, pKa, and stability in ensuring safety and effectiveness. AI based approaches in pharmaceutical preformulation enables accurate prediction of drug behavior, optimizes formulation parameters, and enhances the manufacturing process. These systems, like the Adaptive Neuro-Fuzzy Inference System (ANFIS), have proved to be efficient in the excipients' selection and in illustrating how application of AI algorithms can make and expedite drug research. AI allows for early detection of potential bottlenecks in the development process, thereby reducing risk, making formulation strategies optimal, and avoiding costly delays. It further makes manufacturing efficient by ensuring consistency and scalability in products.

Molecular Docking and Virtual Screening: AI has revolutionized molecular docking and virtual screening with the incredible improvement in the accuracy of drug-target interaction predictions. Better deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are behind better lead compound identification. AI-tuned scoring

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

functions, employed in computer programs such as AutoDock and DOCK, increase the likelihood of finding biologically active compounds (10).

QSAR and QSPR Modelling: AI brings Quantitative Structure Activity Relationship (QSAR) and Quantitative Structure Property Relationship (QSPR) models into the foreground, propelling drug activity, stability, and formulation performance predictions (11). Support Vector Machines (SVMs) and Random Forest (RF) algorithms scan vast amounts of chemical data to identify patterns and predict molecular behavior. Studies show AI-generated QSAR/QSPR models outperform conventional approaches in predicting drug efficacy (12,13).

Solubility and Permeability Prediction: AI models accurately predict aqueous solubility and permeability of drug compounds using molecular structure analysis and experimental data. Predictions inform formulation strategy early in development, reducing the amount of trial-and-error. Deep learning platforms like ANNs have been shown to be very accurate for solubility improvement prediction.

Drug Excipient Compatibility: AI facilitates the selection of proper excipients through predicting potential drug excipient interactions to prevent formulation instability and side effects. These prediction tools direct researchers to excipients that guarantee formulation safety and efficacy.

Stability Analysis and Degradation Pathways: AI platforms replicate degradation pathways and recommend ideal storage conditions, which help in predicting long-term stability. This minimizes the need for prolonged real-time testing and guarantees that formulations continue to work within their shelf life.

Clinical Trial Optimization: Patient recruitment is a major bottleneck in clinical trials that accounts for around one-third of the trial duration. AI helps select the right candidate early, reducing opportunities for delays or trial failure. It has the potential to evaluate genomic and exposomic data to select disease-specific populations for Phase II and Phase III trials and thus make the trials more efficient and effective. AI helps patient monitoring that supports the reduction of the dropout rate by imposing protocol compliance.

In-line Manufacturing and Quality Control: AI allows minimal human interaction with a requirement, ensuring batch to batch consistency and proactive quality control. By using decision tree models in addition to data analytics as part of the Quality by Design (QbD) principles, AI enables identifying critical process variables as well as continuous improvement during manufacturing. Controlled-Release and Nanotechnology Applications: AI assists in the formulation of controlled-release formulations by accurate forecasting of the actions of nanoparticles to achieve targeted and sustained drug delivery. Such capabilities maximize drug effect while reducing dosing interval and side effects (14).

Preformulation Parameters Prediction Using AI and ML Approaches

AI and ML-based prediction of preformulation parameters is an orderly process that involves the use of data gathering, simulation by computation, and validation in order to optimize drug formulation. The process begins with the selection of the targeted preformulation property to be predicted, i.e., solubility, stability, permeability, dissolution rate, or polymorphism, based on the chemical nature of the drug, route of administration, and dosage form. High-quality datasets are then gathered from databases, literature, and laboratory experiments as molecular descriptors, physicochemical parameters, and formulation parameters, which are then cleansed, feature engineered, and normalized for uniformity and model performance enhancement. Proper features are selected by techniques such as correlation analysis, PCA, or RFE for improving model accuracy. Appropriate AI/ML models are chosen based

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

on prediction types, e.g., regression models for continuous features, classification models for categorical variables, or deep models for complex molecular patterns. The data is split into training and testing sets, models are trained using hyperparameters that have been optimized, and cross-validation is performed to prevent overfitting. Performance of the models is measured based on metrics such as RMSE, MAE, R^2, accuracy, precision, recall, and F1-score, regularly plotting predicted vs actual values for added understanding. Once validated, the developed model is then applied to predict preformulation parameters for new drug candidates, integrated into development pipelines, and refreshed periodically with fresh data or feature refinement to maintain precision. Lastly, model output interpretation, such as feature importance analysis, guides rational formulation design and decision-making in early-stage drug development (15).

Solubility

The solubility of a drug molecule is of paramount importance to its therapeutic action, since it influences directly bioavailability and the drug's capacity to achieve its desired effect.

For drugs that are administered orally, solubility determines how well the drug will dissolve and be made accessible for absorption in the gastrointestinal (GI) tract. The accuracy of AI and ML models for predicting solubility is highly reliant on three key factors: the presence of large, good-quality datasets; how the data is represented; and the selection of learning algorithms capable of detecting and learning predictive molecular features effectively. Modeling the complex relationships that affect solubility accurately is still a challenge and involves meticulous tuning of these aspects. The data sets can be obtained from the online platform like eChemPortal, EPI Suite Data, ESOL, AQUA, PHYS, OCHEM. Mohammad Amin Ghanavati et al., developed an advanced, explainable ML framework for aqueous solubility prediction based on four comprehensive datasets ESOL, AQUA, PHYS, and OCHEM comprising a total of 3,942 unique molecules. To represent molecular structures effectively, three complementary data forms were employed: electrostatic potential (ESP) maps, molecular graphs, and tabular molecular features. ESP maps were generated through Density Functional Theory (DFT) calculations for all molecules, providing high-quality 3D charge distribution data that capture electrostatic behavior at the molecular surface. These maps, along with molecular graphs and descriptor-based features, became the basis for model development using three predictive methods EdgeConv, Graph Convolutional Network (GCN), and eXtreme Gradient Boosting (XGBoost).

As shown in Figure 1, the first predictive model, EdgeConv, was originally intended for 3D point cloud data processing but was here tailored to read ESP maps for solubility prediction.

In this method, ESP maps are viewed as dense 3D point clouds who's each spatial point includes an electrostatic potential value. The model associates every point with its nearest neighbours to build localized graphs that reflect geometric and electronic interrelations. These graphs go through several EdgeConv layers with kernel sizes of 128, 128, 256, and 512 that learn increasingly complex structural patterns step by step. The extracted features are thereafter processed by a multilayer perceptron (MLP) to make precise solubility predictions by detecting implicit 3D molecular relationships. Supplementing this, GCNs process molecular graphs outright, representing atoms as nodes and bonds as edges. Through systematic neighbourhood aggregation, GCNs discover how atomic and bonding configurations influence solubility and thereby reveal the embedded structural wisdom of a molecule. The third approach, XGBoost, is a stable ensemble learning algorithm optimized for structured data analysis. It incorporates geometric descriptors such as molecular volume (V), surface area (A), and J index (J) together with chemical and electrostatic descriptors from ESP maps and Mordred descriptor sets. Before training, an

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

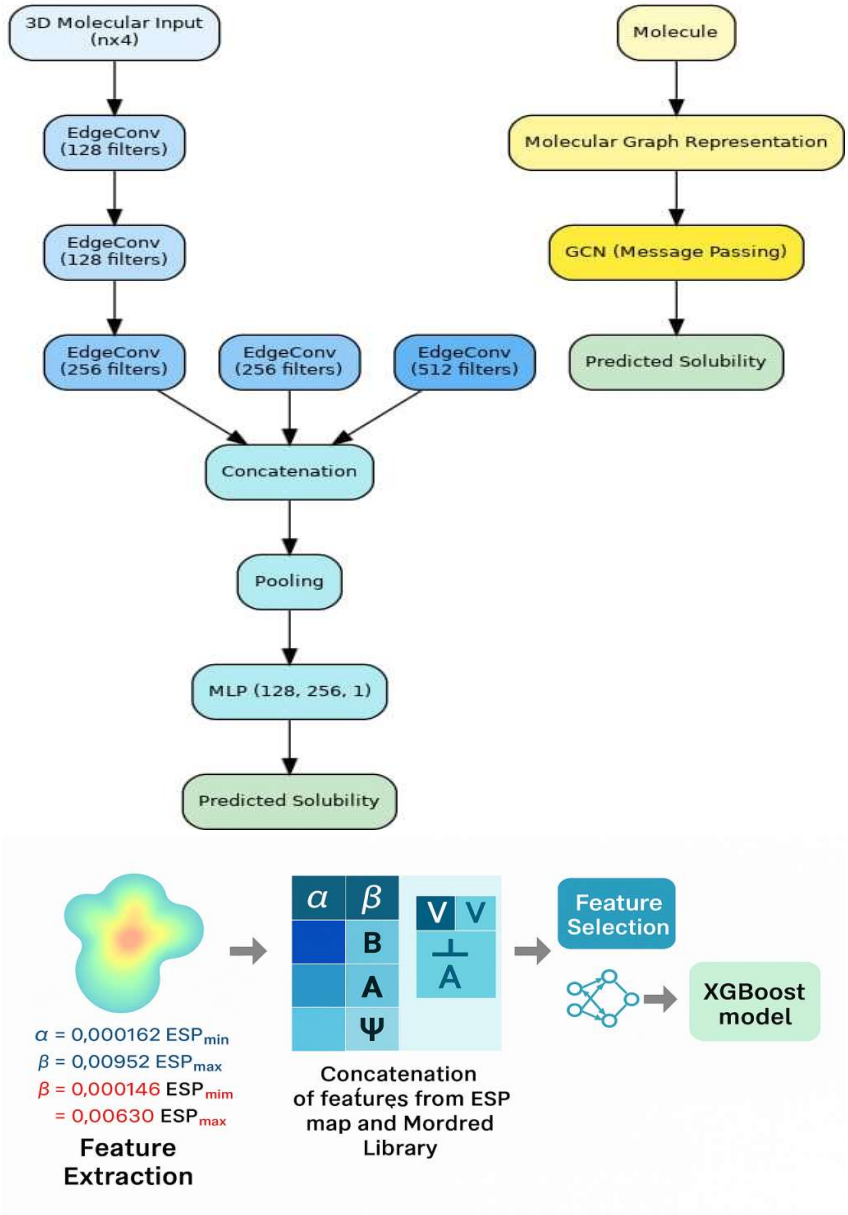

Fig 1: schematic diagram of Solubility prediction by EdgeConv, GCN models and Feature extraction and selection for XGBoost model (16)

RF algorithm is employed for the approximation of feature importance, retaining only the top 20% most important features, and removing redundant or non-essential ones. This selection of features not only decreases model complexity but also enhances accuracy and computational efficiency. The improved feature subset is then applied to train the XGBoost model so that it can learn to recognize intricate nonlinear patterns

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

responsible for understanding how well various compounds dissolve in water. Through recognition of these structural patterns, the model offers useful insights helpful in designing drug molecules with greater solubility and better pharmacokinetic performance.

All three models EdgeConv, GCN, and XGBoost were trained on 80% of every dataset and tested with the remaining 20% to make the evaluation unbiased. They discover that XGBoost model performed best among the three with a Mean Absolute Error (MAE) of 0.458, Root Mean Squared Error (RMSE) of 0.613, and R^2 of 0.918, showcasing incredible predictive potency. Yet, rather than depending on any one method, they hybridized all the three models' outputs into a global ensemble, which was able to tap into their respective strengths in a complementary way. This ensemble model carried out even better prediction accuracy and generalizability between data sets. To provide the robustness of their approach, the ensemble model was applied to the Solubility Challenge 2019 data set with it achieving an RMSE of 0.865 significantly better than 37 other models with average RMSE of 1.62. These results not only highlighted the improved predictive ability of the proposed framework but also its transferability at a high level across datasets with diverse chemical compositions. To ensure transparency of the model decision-making process, SHAP (Shapley Additive Explanations) analysis was applied to the XGBoost model. By means of the interpretability analysis provided by SHAP, the most important molecular features that affect solubility, for example, SLogP, $Beta_1$, and sphericity, were highlighted. These features were recognized as the key factors to solubility behavior, hence connecting model predictions to chemically insightful observations (16).

pKa

The pKa is the negative logarithm of the acid dissociation constant [Ka]: pKa= -log (Ka) it indicates the pH at which half of the molecule is ionized and half is unionized. The acid dissociation constant (Ka), also known as the ionization or protonation constant, is an equilibrium value that represents the ratio between a compound's protonated and deprotonated forms. pKa ensures a balance between solubility and permeability for efficient absorption (17). Ionization affects absorption which influences solubility and permeability through biological membranes. Non-ionized forms are better absorbed; acidic drugs absorb better at stomach and basic drugs at intestine. In preformulation studies AI and ML analyze the molecular structure, functional groups, and electronic properties to identify how easily the molecule donates or accepts protons. Hence predict the pKa values of different solvents. To predict pKa values, Mansouri et al. applied several advanced ML techniques like SVM: Originally developed for classification tasks, SVM has since been adapted for regression problems like pKa prediction due to its flexibility and ability to model continuous outcomes. XGBoost: A powerful ensemble learning method that combines multiple weak decision trees to create a robust predictive model. This method is augmented with data preprocessing, feature engineering, and extensive model validation for enhancing precision. Deep Neural Networks (DNN): These are models with numerous hidden layers used for mapping features by applying complex, non-linear transformations. DNNs are most effective in identifying complex relations among data, thus enhancing predictive precision (18). Models used involved extensive molecular representations such as over 1,400 continuous descriptors, over 9,000 binary fingerprints, and almost 6,000 fragment-based features. Stringent preprocessing of data with KNIME maintained dataset integrity via standardization, salt and counterion removal, and tautomer normalization, with experimental pKa values obtained from DataWarrior for 7,912 compounds. All three methods' validation metrics were comparable, with DNNs achieving maximum accuracy in acidic pKa prediction (RMSE 1.51, R^2 0.80) and SVM

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

Table 1: The molecular structure descriptors which are used in QSPR models (25)

Molecular structure descriptors	Description
Max Estate Index	Maximum atomic charge state index
Mol Wt	Molecular weight
Num Valence Electrons	Number of valance electrons
Fd Density Morgan	Morgan circular fingerprint density (radius 1–3)
BalabanJ	Balaban index of molecular connectivity graphs (i.e., J-index)
Chi	Molecular connection index (0–4)
Kappa	Molecular shape index (1–3)
HallKier Alpha	Molecular polarity and charge distribution
Fraction CSP3	The proportion of C atoms in SP3 hybridization

models beating others slightly in basic pKa prediction (RMSE 1.53, R^2 0.78) (19,20).

Partition Coefficient (logP) Prediction with AI-Based QSPR Models

The partition coefficient (P), or distribution coefficient (D), is a measure of the ratio of a compound's concentration in two immiscible solvents when equilibrium is achieved. It indicates the compound's preferred solubility in each phase. To estimate logP (logKO/W) values (21), AI-based methods like feed-forward neural networks (FNNs), XGBoost, and RF algorithms were utilized to establish sound QSPR models with the help of molecular descriptors (Table 1). Model performance was evaluated with the help of MAE, RMSE, coefficient of determination (R^2), and MRE (22). The FNN based QSPR model is a non-recurrent ANN structure in which data flow is in one direction. FNN was constructed with:

Input layer: 182 neurons, Hidden layers: Two with 128 and 64 neurons, Output layer: One neuron, Activation functions: ReLU for input and hidden; linear for output, Loss function: Mean Squared Error (MSE). Training revealed performance plateauing after 500 epochs, determining optimal training conditions. FNN's robust predictive performance (23). The RF based QSPR model employs an ensemble of decision trees; each trained on randomly selected data and feature subsets. This approach reduces overfitting and improves generalization. XGBoost based QSPR model is a powerful ensemble technique that iteratively refines predictions by learning from previous errors. It uses techniques such as adjusting the learning rate, limiting tree depth, and applying regularization to improve performance and reduce the risk of overfitting (24).

Yang et al., developed an QSAR model for the prediction of Ko/w using the ML algorithm. They curated a dataset of 14,610 organic compounds and their respective SMILES from established online data source platforms like Kaggle. Employing AI algorithms, and 209 descriptors including general, graph, hybrid Estate-VSA descriptor from eRDKit toolkit in Python, they extracted the molecular features. The mean and variance of all the data were reported to be 0.9987 and 1.3005, respectively. They find out 13 data points smaller than -2.9013 and 17 data points higher than 4.8987 were omitted for data clearance. Subsequently, a sample of 14580 data points was chosen for training and testing purpose and the performances of these models was rigorously assessed. Figure 2 showing the data output obtained from XGBoost-based QSPR, exhibited exceptional performance, showcasing an R^2 value of 0.9772, surpassing FNN and RF-based QSPR models. The contribution of each descriptor to the prediction of partition coefficient can be interpreted using SHAP analysis. SMR_VSA8, SMR_VSA3, Kappa2, Heavy Atom Count, and furfuran were found to contribute more in this study (25).

Figure 3 shows that The Kappa2 descriptor defines molecular geometry by

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

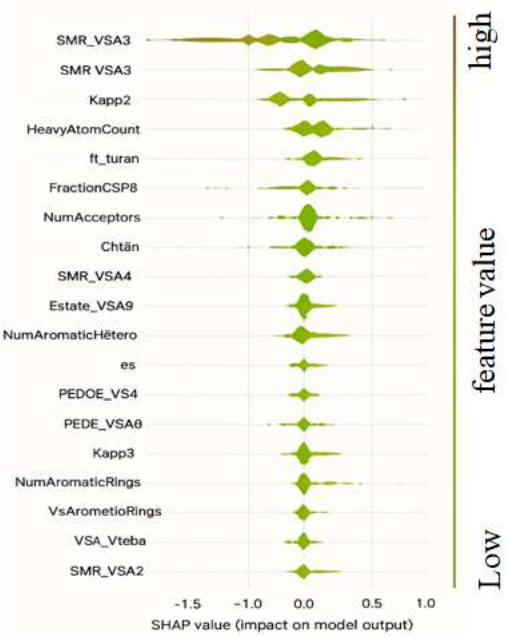

Fig. 2: Significant contribution of input features and output results for the XGBoost-based QSPR model (25).

evaluating electron cloud distribution using the molecular lipophilicity potential method. Notably, SMR_VSA8 and SMR_VSA3 play significant roles in influencing $logK_{O/W}$: SMR_VSA8 (atoms with molar refractivity between 3.63–3.80) exhibits a negative correlation with logKO/W; higher values correspond to reduced lipophilicity. SMR_VSA3, conversely, shows a positive correlation; as its values increase, so does logKO/W. Additionally, HeavyAtomCount negatively affects logKO/W, while the presence of furan rings (furfuran) contributes positively, indicating higher lipophilicity.

Stability

Stability refers to the capacity of a drug substance or formulation to retain its identity, strength, quality, and purity within defined specifications throughout its shelf life and usage period. In the context of protein stability prediction, ML approaches are generally categorized into two types:

Table 2: performance of ANN in predicting stability (27)

	Training set	Test set 1	Test set 2
No. of stable cases	5456	608	647
No. of unstable cases	2888	392	353

supervised and self-supervised models. Supervised models require experimentally obtained protein stability data as training targets. Self-supervised models, on the other hand, learn underlying data patterns without the need for labelled experimental outcomes (26). Abidhan Bardhan et al., applied multiple ML algorithms including ANN, SVM, ELM (Extreme learning machine), GPR (Gaussian process regression), LSSVM (Least square support vector machine), RVM (Relevance vector machine), RF, LightGBM, K-Nearest Neighbours (KNN), and Naive Bayes to determine the feasibility of predicting the physical stability of solid dispersions (27).

They take two separate ANNs were trained one to predict stability, and the other to predict the mode of instability. Table 2 shows Each ANN was evaluated using two distinct test sets: For the stability-predicting ANN, the first test set involved simulating three-phase faults under randomly selected conditions, with load levels ranging from 100% to 130% of the nominal load. For the instability-mode ANN, the first test set also involved three-phase faults under different variable conditions, but with load levels between 100% and 110% of nominal. A second test set was created for both ANNs, using similar load ranges as the first sets. However, in this case, three-phase faults were introduced at random locations along different transmission lines to further evaluate model generalization and robustness. DXs achieved very high accuracies ($R^2 > 0.99$). ELM and GPR provided 95–97% accuracy, showing strong generalization ability. LSSVM and RVM

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

showed >95% accuracy with good reliability and robustness. (28).

Polymorphism

Polymorphism is a phenomenon where a compound can crystallize into two or more disparate structural forms. These unique crystal patterns may result in significant variation in the physical and chemical properties of the compound, such as its solubility, stability, and bioavailability parameters that are important in drug applications. Software for the prediction of crystal structures can be categorized into two broad types: molecular mechanics-based and quantum mechanics-based methods. One of the most popular molecular mechanics software is the Polymorph module of Materials Studio, which uses sophisticated algorithms to predict low-energy polymorphic structures. Once the molecular structure of a compound is known, this module can predict all the possible crystal forms by computing the lowest lattice energy, making predictions for stable and metastable crystalline forms. It facilitates the prediction of atomic and molecular crystal structures from molecular information alone. For more accurate prediction of polymorphs and co-crystals, machine learning methods like RF, ANN, SVM, and Logistic Regression are being used more and more (29). RF uses input features such as solvent properties, molecular descriptors, crystallization conditions, and past crystallization results. It builds an ensemble of many decision trees from random subsets of data, with every tree estimating the probability to develop a particular crystal form. The overall output is the result of a majority vote of all the trees, providing a probabilistic prediction of polymorph formation. RF has also shown high accuracy in predicting solvate formation with up to 86% success in case of pharmaceutical compounds.

SVM employs molecular descriptors like molecular size, shape, charge distribution, and ability to participate in hydrogen bonding to create a hyperplane that classifies data into categories (e.g., likely or unlikely to form a polymorph). The algorithm maximizes the margin of separation between such classes and yields either binary or probabilistic classification. SVM has delivered superior performance in crystal structure prediction of binary and ternary inorganic compounds with more than 90% accuracy. Large datasets containing molecular properties, experimental parameters, and known crystallization results are needed by ANNs. ANNs mimic neural processing in the brain and are networks of connected nodes (neurons) in many layers. They have the ability to capture nonlinear relationships in the data in order to estimate the probability of polymorph formation. When sufficiently large and diverse datasets are available, ANN models can provide highly useful and reliable predictions (29).

Drug-Excipient Compatibility

Drug-excipient compatibility is an important aspect in the development of pharmaceuticals, as some excipients can negatively interact with active pharmaceutical ingredients (APIs) and affect their stability, efficacy, or safety. To foretell such risks proactively, AI and ML methods have been gaining prominence to forecast possible incompatibilities based on molecular descriptors, past interaction data, and physicochemical characteristics. Sophisticated models like RF, ANN, SVM, and Gradient Boosting Machines (GBM) are widely used in this context. Contemporary methods also incorporate representation learning methods like Mol2vec, which transforms molecular structures into numerical representations, and 2D molecular descriptors, which summarize important aspects of molecular geometry and reactivity. These aspects are used as inputs to stacking models, which aggregate the predictive abilities of many base learners to enhance overall accuracy and stability. These models not only forecast new drug-excipient combination compatibility but also evaluate degradation hazard on storage, improving formulation safety and lowering production failures (30).

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

RF builds several decision trees on random subsets of data; each is trained on a random subset of features and samples of data. These trees examine compatibility individually based on molecular descriptors of the drug and excipient (e.g., polarity, reactivity, solubility). Every tree generates a prediction (e.g., compatible or incompatible), and the final prediction is obtained using majority voting (for classification) or averaging (for regression). Through the identification of various patterns using many trees, RF is able to model different drug-excipient interaction situations well. Its resistance to overfitting and handling of high-dimensional input data make it particularly capable for compatibility prediction tasks.

In ANN input layer is provided with feature vectors of molecular and excipient attributes, including structural features, functional groups, solubility, and thermal stability. With several hidden layers, the network learns complex, nonlinear relationships among such variables. During training, weights are updated with backpropagation and gradient descent, which aims to minimize prediction error. The trained ANN can subsequently process novel drug-excipient combinations and generate a compatibility score or classification. Its power is in its flexibility with complicated data sets with complex relationships so that it can be very useful in modelling nonlinear drug-excipient interactions (31). SVM is a classification algorithm that finds the optimal hyperplane in a high-dimensional space to separate compatible and incompatible pairs. Using kernel functions (e.g., radial basis function, polynomial), SVM maps original input features into a higher-dimensional space where linear separation is possible. For drug-excipient compatibility, molecular descriptors and excipient features are input into the SVM and then algorithm which identifies the best separating boundary that maximizes the margin between classes. New pairs are classified based on their position relative to this hyperplane. SVM has shown high generalization ability and is particularly effective for smaller or moderately sized datasets. These ML models, when integrated with robust datasets and structural representations, offer a powerful predictive framework for ensuring drug-excipient compatibility, ultimately supporting safer and more stable pharmaceutical formulations (32).

AI and ML Case Studies in Pharmaceutical preformulation

Patheon, a business unit of Thermo Fisher Scientific, has used its AI/ML-driven Quadrant 2platform to overcome long-standing issues of low solubility and bioavailability during early drug development. Replacing conventional time-intensive experimental screening, the platform uses in-silico modelling, QSAR, and other AI/ML algorithms to forecast compound properties from molecular structure. It enables the rapid determination of the optimum combinations of excipients and solubility improving strategies, reducing costly laboratory testing. By anticipating the formulations most likely to be successful, it optimizes development and saves time and resources, especially for poorly soluble compounds, which are more than half of the new chemical entities (34). Similarly, scholarly studies have designed AI/ML models that can forecast the physical stability of solid dispersion formulations, which are used widely to enhance solubility. Through the generation of large datasets of solid dispersion formulations, scholars employed various ML methods and determined that a RF model offered the best accuracy in the prediction of stability of different drug-polymer mixtures (Table 3). The model predicted key molecular descriptors affecting stability, providing a data-driven, rational framework for formulation development, thus shifting from empirical approaches, decreasing formulation cycles, and saving raw material in early development. Academic research and research institutions have also deployed DL models to predict drug stability and calculate shelf-life, as illustrated in an Esomeprazole freeze-dried powder formulation. The MLPDL model was trained with stability study data,

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

Table 3: AI/ML Models and Tools Used in Preformulation (7,10,33)		
AI/ML technique	Software description	Applications
Artificial Neural Networks (ANNs)	These are interconnected computing systems that mimic the brain's electrical impulse transmission through "perceptrons" functioning like biological neurons. Each node in an ANN receives a different input process to generate outputs using algorithms for problem-solving. Common types include CNNs, RNNs, and MLPs, which can be trained via supervised or unsupervised learning.	Used to model and optimize drug release kinetics across various dosage forms, enabling prediction of API release profiles and identification of optimal formulation under various conditions.
AlphaFold and BiLSTM [Bi-directional Long Short-Term Memory]	Developed by DeepMind, AlphaFold by using AI to predict the protein structure and is highly effective for sequence-based prediction tasks, including patient health data analysis.	Knowledge of protein structure, as it enables one to see how drug can bind to their targets. BiLSTM employed to forecast patient responses to drugs from past data.
Particle Swarm Optimization (PSO)	It applies a swarm intelligence algorithm inspired by bird and fish behavior to efficiently explore and identify optimal solutions for complex pharmaceutical optimization problems.	Utilized in dosage form design to refine particle size distribution, dissolution behavior, and other key formulation parameters.
AI-powered Expert Systems	Such systems replicate expert-level decision-making using AI to process high volumes of data such as clinical trials and patient histories, individualized medicine, trial design, and regulatory requirements, thus improving efficiency and accuracy of pharmaceutical decisions.	Used in the optimization of dosage form by taking into account multiple formulation and process factors, drug discovery, clinical trial design, and regulatory requirements. Enhance the efficiency and accuracy of making decisions.
Computational Fluid Dynamics (CFD)	It employs Reynolds-Averaged Navier-Stokes solvers technology for simulating and analyzing fluid flow, heat transfer, and associated phenomena to support the	Applied to optimize excipient-API mixing, study powder flow, simulate heat and mass transfer in drying or coating, inform equipment design, and facilitate process scale-up, to ensure

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

	design and optimization of pharmaceutical processes and equipment.	uniformity and efficiency of dosage form development.
DeepVS	DeepVS is an artificial intelligence-based software harnessing deep learning for augmenting virtual screening by anticipating binding affinities between target proteins and small molecules to enhance accuracy and efficiency in drug discovery at early stages.	Used to predict interaction between drug candidates and excipients, determine binding affinities, and screen molecules for maximum stability and solubility, thus shortening formulation design and lessening experimental burden.
ANN-Genetic Algorithm (ANN-GA) Hybrid Models	A hybrid of ANN and GA to optimize intricate processes. The ANN models predict and learn system behavior, whereas the GA searches effectively for optimal solutions and hence are beneficial for pharmaceutical process optimization and formulation design.	Utilized for optimizing dosage form design, drug release profile prediction, optimization of formulation parameters, and maximization of process efficiency, allowing rapid and precise development of pharmaceutical products.
Convolutional Neural Networks [CNN]	It is a deep learning technique that extracts hierarchical features from structured or image data to facilitate pattern detection and forecasting. In the pharmaceutical industry, it supports drug discovery, preformulation, and quality control through analysis of molecular structures, histopathology images, and chemical patterns.	Applied for analyzing molecular structures, predicting drug-target interactions, assessing histopathology images, monitoring formulation quality, and supporting preformulation studies.
Recurrent Neural Networks (RNN)	It is a deep learning instrument for sequential data processing by memorizing past details, extracting temporal patterns for applications such as time-series prediction and sequence analysis. In pharmaceuticals, it supports patient data analysis, simulating biological or chemical sequences, and drug release profile prediction.	RNNs excel in analyzing sequential data, making them valuable in pharmaceuticals. They predict drug-target interactions, design new molecules, analyze time-series medical imaging.
QSAR/QSPR modelling	QSAR and QSPR software are computer algorithms used to predict biological activity or physicochemical properties	QSPR models forecast solubility, stability, permeability, and other formulation characteristics, while QSAR models forecast future

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

	from molecular structures. QSAR addresses bioactivity and toxicity, while QSPR forecasts solubility and stability characteristics, aiding drug discovery and preformulation and reducing experimental effort.	toxicity or bioactivity. They both help select excipients, optimize formulation parameters, and reduce the need for large volumes of experimental runs, accelerating early drug development.
Neural graph fingerprints (NGF)	NGF is a powerful cheminformatics tool which represent molecules as graphs. They use GNNs to learn informative representations of molecules, identifying subtle structural patterns not picked up by traditional fingerprints.	Applied in drug discovery and preformulation for predicting molecular properties, bioactivity, and toxicity, lead-like drug candidate screening, chemical structure modeling, and formulation design direction by representing subtle graph-based molecular information.
Deep learning [DL], DeepChem and DeepTox	DL uses multilayered neural networks for feature extraction, prediction, and pattern recognition. DeepChem accelerates drug discovery through molecular modeling and virtual screening, and DeepTox predicts chemical toxicity for speeding up early safety evaluation.	Applied in forecasting drug properties, solubility, stability, and potential toxicity, and in simulating molecular interactions and excipient selection. Applied for the optimization of formulation parameters, reducing experimental trials, and accelerating the initial stages of dosage form development.
Natural language processing tools [NLP] and cheminformatics tools	NLP tools such as spaCy, NLTK, and Transformers perform well in applications such as text analysis and information extraction. These models employ ML algorithms to interpret chemical data and forecast molecular activity. They help in drug discovery and optimization by recognizing correlations between molecular structure and biological effects.	NLP software is applied to retrieve and scan relevant data from scientific papers, patents, and databases, to help in selection of appropriate excipients, formulation approaches, and previously conducted experiments. Cheminformatics software helps in the modeling of molecular structures, the prediction of physicochemical properties, solubility, stability, and interactions, helping in drug-excipient selection and the formulation optimization process while minimizing experimental effort.

including pH and storage time as the inputs to predict assay and impurity levels for 36 months. The DL based predictions were superior to conventional mathematical models, allowing accurate shelf-life estimation without the requirement of full long-term studies, curbing development time and resource intensity with identification of key

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

parameters influencing stability (35). Moreover, partnerships between big pharma players like Pfizer, Novartis, and Merck, with AI technology providers like XtalPi and Solitek, have leveraged AI driven Crystal Structure Prediction (CSP) to detect and describe prospective polymorphs of APIs. AI/ML models enable scientists to foresee the stability and probability of different polymorphic forms, theoretically probing the polymorphic space much quicker compared to traditional experimental approaches. Early identification of stable polymorphs prevents costly late-stage failures or market recalls because of unforeseen crystal forms and allows the most stable and producible polymorph to be chosen for drug development. In addition, a joint effort among pharmaceutical scientists, universities, and technology partners like Ginverse Private Limited has resulted in the creation of an AI-based system, Formulation AI, which predicts the compatibility of drugs and excipients. Applying sophisticated ML methodologies, the platform examines drug-excipient pair histories to predict possible incompatibilities with high accuracy. A study used Mol2vec embeddings and 2D molecular descriptors paired with a stacking ensemble method and reported 98% prediction accuracy and better performance compared to other computational methods. By predicting compatibility with high accuracy, Formulation AI allows formulators to steer clear of poor drug stability and premature degradation, resulting in safer, better performing products while saving time, cost, and resource use with conventional trial-and-error testing (36).

Success Stories and Challenges faced

Pharmaceutical companies are using AI and machine learning more often to speed up drug formulation and development (37, 38). For example, Pfizer's Centaur Chemist tool can predict chemical reactions and find the best ways to synthesize drug candidates, saving both time and money. Atomwise's AtomNet uses deep learning to analyze molecular structures and select promising drug candidates, which has helped in finding treatments for diseases like Ebola and multiple sclerosis. BenevolentAI showed how AI can be used in real-world situations by repurposing baricitinib for COVID-19 treatment through large-scale data analysis. Another company worked with organizations like FAST to use AI in finding drug candidates for rare diseases such as Angelman syndrome, highlighting AI's potential to address unmet medical needs (39, 40). Despite these advances, using AI and machine learning in preformulation still faces challenges. The main issue is the lack of high-quality, comprehensive data on important preformulation factors like physicochemical properties, solubility, degradation rates, excipient interactions, and stability.

As opposed to clinical datasets, formulation data are typically inhomogeneous, proprietary, and obtained under heterogeneous lab conditions, such that model generalizability and reproducibility are not easy to achieve (41). Another major challenge is selection in features defining what molecular descriptors or formulation variables impact desirable properties the most. In addition, overfitting and underfitting are ongoing challenges: small preformulation datasets can cause sophisticated deep learning models to memorize information rather than learning patterns within, and overly simplistic models may not be capable of capturing nonlinear interactions of formulation variables (42,43). Interpretability of the model is also a must in the pharmaceutical case. Regulators require interpretable, explainable models to support predictions of drug safety, stability, and performance. Black-box models are also challenging for regulatory filings since without interpretability, it is difficult to verify predictions and offer scientific accountability. Furthermore, the absence of standardized preformulation databases and data-sharing platforms in the academia and industry continues to hinder progress. Ethical and regulatory issues like data privacy, patient consent, and algorithmic bias render it even tougher to incorporate AI in regulated formulation processes. The success of AI-based preformulation then

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

depends on developing good, consistent, and interoperable data sets, employing appropriate feature selection and dimensionality reduction, and maintaining models explainable, validated, and compliant with the regulatory environment (44).

Future prospects and new trends

Future preformulation studies will be informed by advanced deep learning algorithms such as CNNs and RNNs capable of detecting subtle interactions between molecular structure, formulation composition, and physicochemical properties. Emerging techniques such as transfer learning and domain adaptation can overcome data scarcity by leveraging pre-trained models from adjacent domains (e.g., medicinal chemistry, pharmacokinetics) and tweaking them for preformulation-specific predictions such as solubility enhancement or excipient compatibility (45). Developing explainable AI tools is important to meet regulatory needs for transparency and scientific clarity. Techniques like SHAP and LIME can show which formulation factors most influence predictions, which helps with regulatory submissions and quality frameworks (46). Generative models such as GANs and VAEs could change computer-aided preformulation by creating new excipient blends, predicting the best ways to improve solubility, and modeling stability under different conditions (47). These models can also help screen formulations virtually before lab testing, saving time and materials. Machine learning can reduce failures due to poor biopharmaceutical properties by focusing on candidates with better physical and chemical profiles (48). Efforts to standardize preformulation datasets using FAIR data principles are expected to encourage more collaboration among academia, industry, and regulators. By combining these standardized datasets with advanced AI systems, scientists can make better decisions, avoid repeating experiments, and speed up development. Finally, linking AI with automated lab platforms, like robotic formulation and high-throughput screening, will enable continuous optimization, where predictive models guide design, experiments improve the models, and the process keeps evolving. This convergence is the new golden horizon in smart preformulation, wherein data, models, and automation come together to drive innovation (49).

Regulatory and Recommendation Insights

From a regulatory perspective the U.S. Food and Drug Administration (FDA) has recently released a discussion paper titled 'Proposed Regulatory Framework for Modifications to AI/ML Based Software as a Medical Device (SaMD)', outlining a proposed way of regulating AI/ML powered medical software This report outlines the FDA's more sophisticated approach to bringing AI- and ML-powered medical software under premarket scrutiny, with the aim of preserving both safety and efficacy (50). It talks about various types of changes that could have an impact on end users such as patients, healthcare workers, and other stakeholders such as input data changes, retraining on new data sets, or updates in the architecture of the AI/ML model. To address such evolving technologies throughout their whole life cycle, the FDA also provided a Total Product Lifecycle (TPLC) model of regulation to AI/ML-based SaMD, emphasizing perpetual monitoring and evaluation from development through to post market performance (51, 52).

Conclusion

AI and machine learning methods have proven very helpful for predicting outcomes in preformulation studies, making pharmaceutical development faster, more accurate, and more efficient. By using algorithms that analyze large amounts of data, these methods can predict key properties like solubility, stability, bioavailability, and partition coefficients (logKo/w) for drug candidates. This reduces the need for extensive experimental trials, lowering both costs and development time. AI and machine learning models can also find

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

complex patterns in molecular data that traditional methods might miss, offering valuable insights for improving formulation strategies. Overall, these approaches have made preformulation studies more informed and increased the chances of developing successful drug formulations.

Conflicts of Interest

The author has no conflict of interest.

References

1. Ahirwar K and Shukla R (2023) Preformulation Studies: A Versatile Tool in Formulation Design. Drug Formulation Design. IntechOpen. Available at: http://dx.doi.org/10.5772/intechopen.110346.

2. Higgins, J., Cartwright, M.E. and Templeton, A.C., 2012. Progressing preclinical drug candidates: strategies on preclinical safety studies and the quest for adequate exposure. *Drug discovery today*, *17*(15-16), pp.828-836.

3. Dashpute, S.V., Pansare, J.J., Deore, Y.K., Pansare, M.J., Sonawane, P.J., Jadhav, S.P. and Patil, D.M., 2023. Artificial intelligence and machine learning in the pharmaceutical industry. *International Journal of Pharmacy and Pharmaceutical Research (IJPPR)*, *28*(2), pp.111-131.

4. Malheiro, V., Santos, B., Figueiras, A. and Filipa Mascarenhas-Melo (2025). The Potential of Artificial Intelligence in Pharmaceutical Innovation: From Drug Discovery to Clinical Trials. *Pharmaceuticals*, [online] 18(6), pp.788–788. doi: https://doi.org/10.3390/ph18060788.

5. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K. and Tekade, R.K. (2020). Artificial Intelligence in Drug Discovery and Development. *Drug Discovery Today*, [online] 26(1), pp.80–93. doi: https://doi.org/10.1016/j.drudis.2020.10.010.

6. Dey, H., Arya, N., Mathur, H., Chatterjee, N. and Jadon, R. (2024). Exploring the Role of Artificial Intelligence and Machine Learning in Pharmaceutical Formulation Design. *International Journal of Newgen Research in Pharmacy & Healthcare*, [online] pp.30–41. doi: https://doi.org/10.61554/ijnrph.v2i1.2024.67.

7. Ali, K.A., Mohin, S.K., Mondal, P., Goswami, S., Ghosh, S. and Choudhuri, S., 2024. Influence of artificial intelligence in modern pharmaceutical formulation and drug development. *Future Journal of Pharmaceutical Sciences*, *10*(1), p.53.

8. Rehman, A.U., Li, M., Wu, B., Ali, Y., Rasheed, S., Shaheen, S., Liu, X., Luo, R. and Zhang, J. (2024). Role of Artificial Intelligence in Revolutionizing Drug Discovery. *Fundamental Research*, [online] 5(3). doi: https://doi.org/10.1016/j.fmre.2024.04.021.

9. Vora, L.K., Gholap, A.D., Jetha, K., Thakur, R.R.S., Solanki, H.K. and Chavda, V.P. (2023). Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics*, [online] 15(7), pp.1916–1916. doi: https://doi.org/10.3390/pharmaceutics15071916.

10. Sharma, S., Akanksha Chaubey, Pathan, M.N. and Tyagi, S. (2024). AI-powered virtual screening for drug discovery: Methods and challenges. *International Journal of Pharmacy and Pharmaceutical Science*, 6(2), pp.157–164. doi: https://doi.org/10.33545/26647222.2024.v6.i2b.136.

11. Peter, S.C., Dhanjal, J.K., Malik, V., Radhakrishnan, N., Jayakanthan, M. and Sundar, D. (2019). Quantitative Structure-Activity Relationship (QSAR): Modeling Approaches to Biological Applications. *Encyclopaedia of Bioinformatics and Computational Biology*, pp.661–676. doi: https://doi.org/10.1016/b978-0-12-809633-8.20197-0.

12. Serrano, D.R., Luciano, F.C., Anaya, B.J., Ongoren, B., Kara, A., Molina, G., Ramirez, B.I., Sánchez-Guirales, S.A., Simon, J.A., Tomietto, G., Rapti, C., Ruiz, H.K., Rawat, S., Kumar, D. and Lalatsa, A. (2024). Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine. *Pharmaceutics*, [online] 16(10), p.1328. doi: https://doi.org/10.3390/pharmaceutics16101328.

13. Shukla, A.K., Yadav, V.K., Verma, M., Kanaujia, K.A., Jaiswal, A. and Gupta, V. (2024). Expert Systems in Preformulation and Formulation Development with Special

Reference to SeDeM System: An Innovative, Problem Solving, Intelligent and Optimization Algorithm Tool. *Current Indian Science*, 02. doi: https://doi.org/10.2174/012210299x338978241015155050.

14. Jena, G.K., Patra, C.N., Jammula, S., Rana, R. and Chand, S., 2024. Artificial intelligence and machine learning implemented drug delivery systems: a paradigm shift in the pharmaceutical industry. *Journal of Bio-X Research*, *7*, p.0016.

15. Laura Pereira Diaz, Cameron John Brown, Ojo, E., Mustoe, C. and Alastair James Florence (2023). Machine learning approaches to the prediction of powder flow behaviour of pharmaceutical materials from physical properties. *Digital discovery*, 2(3), pp.692–701. doi: https://doi.org/10.1039/d2dd00106c.

16. Ghanavati, M.A., Ahmadi, S. and Rohani, S. (2024). A Machine Learning Approach for the Prediction of Aqueous Solubility of Pharmaceuticals: A Comparative Model and Dataset Analysis. *Digital Discovery*. [online] doi: https://doi.org/10.1039/d4dd00065j.

17. Kim, H.-S., Kim, C.-M., Jo, A.-N. and Kim, J.-E. (2022). Studies on Preformulation and Formulation of JIN-001 Liquisolid Tablet with Enhanced Solubility. *Pharmaceuticals*, 15(4), p.412. doi: https://doi.org/10.3390/ph15040412.

18. Cruciani, G., Milletti, F., Storchi, L., Sforna, G. and Goracci, L., 2009. In silico pKa prediction and ADME profiling. *Chemistry & Biodiversity*, *6*(11), pp.1812-1821.

19. Cariello, N.F., Korotcov, A., Tkachenko, V., Grulke, C.M., Sprankle, C.S., Allen, D., Casey, W.M., Kleinstreuer, Mansouri, K., N.C. and Williams, A.J. (2019). Open-source QSAR models for pKa prediction using multiple machine learning approaches. *Journal of Cheminformatics*, 11(1). doi: https://doi.org/10.1186/s13321-019-0384-1.

20. Zhou, T., Jhamb, S., Liang, X., Sundmacher, K. and Gani, R., 2018. Prediction of acid dissociation constants of organic compounds using group contribution methods. *Chemical Engineering Science*, *183*, pp.95-105

21. Amézqueta, S., Subirats, X., Fuguet, E., Rosés, M. and Ràfols, C., 2020. Octanol-water partition constant. *Liquid-phase extraction*, pp.183-208.

22. Rice, J.E., 2014. Chapter 5-partition coefficients. *Organic chemistry concepts and applications for medicinal chemistry, Academic Press*, pp.85-92.

23. Cui, Z., Wang, L., Li, Q. and Wang, K., 2022. A comprehensive review on the state of charge estimation for lithium-ion battery based on neural network. *International Journal of Energy Research*, *46*(5), pp.5423-5440.

24. Zhang, X. and Wang, M., 2021, May. Weighted random forest algorithm based on bayesian algorithm. In *Journal of Physics: Conference Series* (Vol. 1924, No. 1, p. 012006). IOP Publishing.

25. Yang, A., Sun, S., Qi, L., Kong, Z.Y., Sunarso, J. and Shen, W. (2024). Development of an interpretable QSPR model to predict the octanol-water partition coefficient based on three artificial intelligence algorithms. *Green Chemical Engineering*. [online] doi: https://doi.org/10.1016/j.gce.2024.07.003.

26. Blaabjerg, L.M., Kassem, M.M., Good, L.L., Jonsson, N., Matteo Cagiada, Johansson, K.E., Boomsma, W., Stein, A. and Kresten Lindorff-Larsen (2023). Rapid protein stability prediction using deep learning representations. *eLife*, 12. doi: https://doi.org/10.7554/elife.82593.

27. El-Amin, I.M. and Al-Shams, A.-A.M. (1997). Transient stability assessment using artificial neural networks. *Electric Power Systems Research*, 40(1), pp.7–16. doi: https://doi.org/10.1016/s0378-7796(96)01124-8.

28. Bardhan, A. and Samui, P. (2022). Application of Artificial Intelligence Techniques in Slope Stability Analysis. *International Journal of Geotechnical Earthquake Engineering*, 13(1), pp.1–22. doi:https://doi.org/10.4018/ijgee.298988.

29. Heng, T., Yang, D., Wang, R., Zhang, L., Lu, Y. and Du, G., 2021. Progress in research on artificial intelligence applied to polymorphism and cocrystal prediction. *ACS omega*, *6*(24), pp.15543-15550.

Vol. 20 (Supplementary Issue 2A), June 2026, ISSN 0973-8916 (Print)., 2230-7303 (Online)
10.5530/ctbp.2026.2s.5

30. Patel, S., Patel, M., Kulkarni, M. and Patel, M.S. (2023). DE-INTERACT: A machine-learning-based predictive tool for the drug-excipient interaction study during product development—Validation through paracetamol and vanillin as a case study. *International Journal of Pharmaceutics*, 637, pp.122839–122839. doi: https://doi.org/10.1016/j.ijpharm.2023.122839.

31. Wang, S., Di, J., Wang, D., Dai, X., Hua, Y., Gao, X., Zheng, A. and Gao, J. (2022). State-of-the-Art Review of Artificial Neural Networks to Predict, Characterize and Optimize Pharmaceutical Formulation. *Pharmaceutics*, [online] 14(1), p.183. doi: https://doi.org/10.3390/pharmaceutics14010183.

32. Hang, N.T., Long, N.T., Duy, N.D., Chien, N.N. and Van Phuong, N., 2024. Towards safer and efficient formulations: Machine learning approaches to predict drug-excipient compatibility. *International Journal of Pharmaceutics*, *653*, p.123884.

33. Noorain, Srivastava, V., Parveen, B. and Parveen, R., 2023. Artificial intelligence in drug formulation and development: applications and future prospects. *Current drug metabolism*, *24*(9), pp.622-634.

34. www.patheon.com. (n.d.). *AI-driven Drug Development for Poor Solubility and Bioavailability*. [online] Available at: https://www.patheon.com/us/en/insights-resources/blog/ai-driven-drug-development-for-poor-soubility-and-bioavailability.html.

35. Ajdarić, J., Ibrić, S., Pavlović, A., Ignjatović, L. and Ivković, B. (2021). Prediction of Drug Stability Using Deep Learning Approach: Case Study of Esomeprazole 40 mg Freeze-Dried Powder for Solution. *Pharmaceutics*, 13(6), p.829. doi: https://doi.org/10.3390/pharmaceutics13060829.

36. Muthusamy, A.R., Chauhan, D., Jain, A.K., Somasundaram, M.S. and Singh, A. (2025). Accelerating Polymorph Screening with AI & ML: A New Era in Drug Development. *Journal for Research in Applied Sciences and Biotechnology*, 4(2), pp.156–166. doi: https://doi.org/10.55544/jrasb.4.2.16.

37. Bannigan, P., Aldeghi, M., Bao, Z., Häse, F., Aspuru-Guzik, A. and Allen, C. (2021). Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews*, [online] 175, p.113806. doi: https://doi.org/10.1016/j.addr.2021.05.016.

38. Pushadapu VS, Babu PS, Danaboina S, Qamar Z, Khan SA, Ali J. Antimicrobial peptides as biomacromolecular therapeutics against antimicrobial resistance: structural insights and mechanistic advances. International Journal of Peptide Research and Therapeutics. 2025 Jul 19;31(5):81.

39. Karako, K. (2025). Artificial intelligence applications in rare and intractable diseases: Advances, challenges, and future directions. *Intractable & Rare Diseases Research*, 14(2), pp.88–92. doi: https://doi.org/10.5582/irdr.2025.01030.

40. Visan, A.I. and Negut, I. (2024). Integrating Artificial Intelligence for Drug Discovery in the Context of Revolutionizing Drug Delivery. *Life*, [online] 14(2), p.233. doi: https://doi.org/10.3390/life14020233.

41. SCW.AI. (2025). *AI in Pharma: Use Cases, Success Stories, and Challenges in 2025*. [online] Available at: https://scw.ai/blog/ai-in-pharma/.

42. Bess, A., Berglind, F., Mukhopadhyay, S., Brylinski, M., Griggs, N.W., Cho, T., Galliano, C. and Wasan, K.M. (2022). Artificial intelligence for the discovery of novel antimicrobial agents for emerging infectious diseases. 27(4), pp.1099–1107. doi: https://doi.org/10.1016/j.drudis.2021.10.022.

43. Team, E. (2020). *Overfitting and Underfitting in Machine Learning*. [online] Great Learning Blog: Free Resources what Matters to shape your Career! Available at: https://www.mygreatlearning.com/blog/overfitting-and-underfitting-in-machine-learning/.

44. Kandhare, P., Kurlekar, M., Deshpande, T. and Pawar, A. (2025). Artificial intelligence in pharmaceutical sciences: A comprehensive review. *Medicine in Novel Technology and Devices*, [online] 27, p.100375. doi: https://doi.org/10.1016/j.medntd.2025.100375.

45. Tran, T. and Chinwe Ekenna (2023). Molecular Descriptors Property Prediction Using Transformer-Based Approach. *International Journal of Molecular Sciences*,

[online] 24(15), pp.11948–11948. doi: https://doi.org/10.3390/ijms241511948.
46. Teixeira, B., Carvalhais, L., Pinto, T. and Vale, Z. (2025). Explainable AI framework for reliable and transparent automated energy management in buildings. *Energy and Buildings*, [online] 347, p.116246. doi: https://doi.org/10.1016/j.enbuild.2025.116246.
47. Nath, S. (2023). *Generative Models: Unraveling the Magic of GANs and VAEs*. [online] Medium. Available at: https://medium.com/@sruthy.sn91/generative-models-unraveling-the-magic-of-gans-and-vaes-66a5858d4596.
48. Usefulbi.com. (2022). *Optimizing Drug Formulation: Gen AI's Role in Enhancing Pharmaceutical Product Development – UsefulBI*. [online] Available at: https://www.usefulbi.com/optimizing-drug-formulation-generative-ais-role-in-enhancing-pharmaceutical-product-development/ [Accessed 13 Oct. 2025].
49. Nih.gov. (2025). *FAIR Data Principles at NIH and NIAID*. [online] Available at: https://www.niaid.nih.gov/research/fair-data-principles [Accessed 13 Oct. 2025].
50. Yang, S.-Y., Huang, Q., Li, L.-L., Ma, C.-Y., Zhang, H., Bai, R., Teng, Q.-Z., Xiang, M.-L. and Wei, Y.-Q. (2009). An integrated scheme for feature selection and parameter setting in the support vector machine modeling and its application to the prediction of pharmacokinetic properties of drugs. *Artificial Intelligence in Medicine*, 46(2), pp.155–163. doi: https://doi.org/10.1016/j.artmed.2008.07.001.
51. Krishna PS. Advances and challenges in mucoadhesive drug delivery systems: a comprehensive review. Asian Journal of Pharmaceutics (AJP). 2024 Jun 15;18(02).
52. Krishna PV, Babu PS, Hemalatha P, Susmitha MJ, Pujitha T, Raju SL, Ravi K. In vitro bio-equivalence studies on commercial formulations containing paracetamol and ibuprofen. Int J Pharm Sci Rev Res. 2023;79:199-205.