JOURNAL OF PHARMA INSIGHTS AND RESEARCH ISSN NO. 3048-5428

REVIEW ARTICLE

Machine Learning for *In Silico* Predictive Modeling of Drug-Excipient Compatibility for Formulation Development

Malini S, Pushpalatha Rathinasabapathy*

Department of Pharmaceutics, Mayor Radhakrishnan College of Pharmacy, Devanampattinam, Tamil Nadu, India

Publication history: Received on 24th March 2026; Revised on 25th April 2026; Accepted on 29th April 2026

Article DOI: 10.69613/ecdvqj46

Abstract: The selection of pharmaceutical excipients is the main determinant of the stability, safety, and therapeutic efficacy of drug products. Unfavorable drug–excipient interactions lead to chemical degradation or physical instability, compromising formulation integrity. Traditional empirical preformulation screening methods, though reliable, are constrained by high costs, low throughput, and prolonged timelines. The use of artificial intelligence and machine learning offers a powerful alternative, enabling rapid, data-driven compatibility predictions. This review discusses about the computational tools specifically DE-Interact, PharmDE, and ExPreSo highlighting their underlying systems, algorithms, and predictive capabilities. DE-Interact utilizes artificial neural networks trained on combined molecular fingerprints to deliver binary compatibility classifications for small molecules. PharmDE functions as a hybrid expert system, combining a curated database of reported incompatibilities with a rule-based inference engine to categorize formulations into high, medium, or low risk. For macromolecular systems, ExPreSo implements supervised ensemble machine learning utilizing protein sequence embeddings, structural features, and product attributes to predict excipient occurrence in stable formulations. While these tools significantly accelerate preformulation screening and minimize resource expenditure, hurdles remain regarding dataset size, representation of rarely used excipients, and regulatory integration. Improving model robustness through expanded, high-fidelity datasets and structural descriptors will be important to establishing fully automated, predictive pipelines in rational pharmaceutical formulation development.

Keywords: AI tools; Drug–excipient compatibility; DE-Interact; PharmDE; ExPreSo.

1. Introduction

In modern pharmaceutical technology, excipients are no longer regarded as mere inactive bulking agents or vehicle substances [1]. Instead, they function as multifunctional enabling components critical to the pharmacokinetic and pharmacodynamic optimization of the active pharmaceutical ingredient (API). Excipients serve diverse, specialized roles, acting as solubilizers, binders, disintegrants, lubricants, glidants, preservatives, buffering agents, and organoleptic modifiers. In solid dosage forms, binders such as polyvinylpyrrolidone or microcrystalline cellulose dictate mechanical strength and compaction, while disintegrants such as croscarmellose sodium facilitate rapid disintegration to ensure timely drug dissolution [2]. Lubricants like magnesium stearate are essential for reducing friction during tableting but must be carefully managed to prevent hydrophobic film formation, which can hinder dissolution rates [3]. In advanced delivery platforms, such as solid dispersions, self-emulsifying drug delivery systems, and nanocarrier-based vehicles, excipients assume a primary role in modulating drug solubility, supersaturation kinetics, and cellular absorption. For instance, polymeric carriers in amorphous solid dispersions such as copovidone or hypromellose acetate succinate prevent recrystallization of poorly soluble APIs by establishing robust intermolecular interactions, such as hydrogen bonding and hydrophobic association [4]. Similarly, in lipidic nanoparticles and polymeric micelles, block copolymers like poloxamers or pegylated lipids modulate surface properties, prolong systemic circulation times, and facilitate targeted delivery. Consequently, any unintentional interaction between the API and these complex excipient matrices directly impacts the thermodynamic state of the drug, its dissolution behavior, and its clinical bioavailability.

The chemical and physical stability of an API is highly sensitive to the functional groups present within both the drug and the surrounding excipient matrix. Chemical incompatibility arises when reactive functional groups undergo thermodynamic reactions, leading to the formation of novel degradation products. The primary chemical degradation pathways include hydrolysis, oxidation, photolysis, isomerization, and polymerization [5]. Hydrolysis represents a major degradative pathway for drugs containing esters, amides, lactams, or imides. The rate of hydrolysis is heavily influenced by the microenvironmental pH, which is dictated by the

* Corresponding author: Pushpalatha Rathinasabapathy

acidic or basic nature of the excipients. For instance, basic excipients can accelerate the hydrolytic cleavage of ester-containing drugs like acetylsalicylic acid [6].

Oxidative degradation typically proceeds via free-radical autoxidation chain mechanisms, catalyzed by trace impurities such as transition metals, peroxides, or oxygen present within excipients. Many common polymeric excipients, including polyethylene glycol and polysorbates, contain residual peroxides from manufacturing, which can induce the oxidation of labile drug groups, such as thioethers, amines, or phenolic hydroxyls [7]. Physical incompatibilities involve non-covalent interactions that do not alter the chemical structure of the drug but modify its physical state, such as crystalline-to-amorphous transitions, polymorphism changes, precipitation, or sorption phenomena [8]. For example, highly hygroscopic excipients like sorbitol or microcrystalline cellulose can absorb atmospheric moisture, raising the local water activity within a solid dosage form and facilitating chemical reactions or phase transformations. Certain active compounds may undergo physical adsorption onto the surfaces of insoluble excipients like colloidal silicon dioxide or dicalcium phosphate, reducing the fraction of drug available for dissolution. Consequently, identifying both chemical and physical interaction profiles during early pre-formulation is essential to ensure long-term product stability.

2. Conventional Empirical Methods for Compatibility Screening

2.1. Solid-State Compatibility

Conventional compatibility screening relies heavily on empirical testing, wherein binary or multi-component mixtures of the API and excipients are prepared in defined ratios [9]. In solid-state compatibility testing, the API and excipient are typically blended in a 1:1 ratio (or a ratio mimicking the final dosage form) to maximize the probability of observing interactions. To simulate worst-case scenarios and accelerate degradation kinetics, these blends are divided into dry mixtures and mixtures containing a defined amount of moisture, typically 4% to 10% w/w water. The samples are then stored under stressed environmental conditions, such as 40ºC and 75% relative humidity, or 50ºC for periods ranging from two to twelve weeks [10]. At designated time points, the samples are retrieved and subjected to comprehensive physical and chemical evaluations. Physical assessment includes visual examination for color changes, liquefaction, gas evolution, or caking. Chemical analysis is performed using highly sensitive analytical methods to quantify drug recovery and identify degradation products, thereby clarifying the kinetics of the degradation process.

Table 1. Model Solid-State Binary Compatibility Experimental Design Matrix

Condition Identifier	Composition Ratio (API: Excipient)	Moisture Level Added (% w/w)	Thermal and Relative Humidity Stress Profile	Target Exposure Period	Rationale & Analytical Endpoint
Control-A	1:0 (Pure API)	0%	40°C ± 2°C / 75% ± 5% RH	2 to 12 Weeks	Establishes the baseline degradation of the pure drug substance under accelerated thermal and moisture stress via HPLC.
Control-B	0:1 (Pure Excipient)	0%	40°C ± 2°C / 75% ± 5% RH	2 to 12 Weeks	Establishes baseline physical transitions (e.g., melting, deliquescence) and chemical changes of the excipient via DSC and TGA.
Binary-Dry	1:1 (Homogeneous Blend)	0%	40°C ± 2°C / 75% ± 5% RH (Hermetically sealed)	4 to 12 Weeks	Detects solid-state chemical reactivity and physical polymorphism changes in the absence of free moisture via XRPD and FTIR.
Binary-Wet	1:1 (Homogeneous Blend)	5% ± 1%	40°C ± 2°C / 75% ± 5% RH	2 to 4 Weeks	Simulates worst-case high-moisture manufacturing scenarios (e.g., wet granulation) to identify moisture-mediated hydrolysis or solubilization.
Lubricant-Dry	20 : 1 or 50 : 1 (API : Excipient)	0%	50°C (Dry oven)	2 to 4 Weeks	Mimics realistic low-concentration boundary conditions for highly functional excipients like glidants and lubricants (e.g., Magnesium stearate).

2.2. Liquid-State Compatibility

Liquid-state compatibility studies are essential for parenterals, oral liquids, and ophthalmic formulations. These protocols involve dissolving or suspending the API and excipients in an aqueous or co-solvent medium, followed by storage in plain and amber borosilicate glass vials [11]. Liquid-state interactions are highly sensitive to solution pH, ionic strength, buffer capacity, dissolved oxygen, and exposure to light. Degradation pathways in the liquid state, such as photolysis and oxidation, occur at significantly accelerated rates compared to the solid state. Consequently, liquid compatibility protocols incorporate stressed photostability testing under ICH Q1B guidelines, alongside thermal and oxidative challenges [12]. The influence of trace heavy metals, residual initiators in polymeric surfactants, and the buffering capacity of salts must be systematically assessed to prevent drug precipitation or catalytic decomposition. These studies are critical for optimizing the microenvironmental pH of the formulation and selecting appropriate primary packaging materials.

2.3. Analytical Methods for Characterizing Solid-State Interactions

To characterize drug–excipient mixtures comprehensively, a multi-faceted analytical approach is required, as no single instrument can capture all chemical and physical changes. Thermal, spectroscopic, chromatographic, and microscopic techniques are deployed in tandem to gain a holistic view of the interaction landscape [13].

Table 2. Analytical Methods for Characterizing Solid-State Interactions

Analytical Category	Technique	Acronym	Application & Principles
Thermal Analysis	Differential Scanning Calorimetry	DSC	Detects thermodynamic phase transitions, including melting, crystallization, and glass transitions. Shifts or disappearances of endothermic peaks indicate physical dissolution or chemical interaction.
Thermal Analysis	Thermogravimetric Analysis	TGA	Measures mass change as a function of temperature. It quantifies loss of volatile components, dehydration events, and thermal decomposition thresholds.
Spectroscopy	Fourier Transform Infrared Spectroscopy	FTIR	Probes vibrational transitions of functional groups. Changes in absorption bands, such as shifts in carbonyl or hydroxyl stretching, reveal intermolecular hydrogen bonding or covalent alterations.
Spectroscopy	X-ray Powder Diffraction	XRPD	Evaluates long-range crystalline order. It detects phase transformations, polymorphism transitions, and transitions from crystalline to amorphous states.
Spectroscopy	Nuclear Magnetic Resonance	NMR	Provides precise structural and electronic environment data. Solid-state NMR detects subtle changes in chemical shifts, identifying localized intermolecular interactions.
Spectroscopy	Near-Infrared Spectroscopy	NIR	Offers non-destructive, rapid analysis of solid mixtures, monitoring real-time moisture absorption and physical state alterations.
Chromatography	High-Performance Liquid Chromatography	HPLC	Quantifies API recovery and resolves degradation products. It provides precise, quantitative kinetic profile data for chemical interactions.
Chromatography	Thin-Layer Chromatography	TLC	Acts as a rapid, qualitative screening method to identify the emergence of novel chemical impurities or degradation spots.
Other Methods	Hot-Stage Microscopy	HSM	Combines thermal stress with optical observation to visualize morphological changes, melting behaviors, and crystal habit modifications in real time.
Other Methods	Vapor Sorption Analysis	VSA	Quantifies moisture sorption isotherms, characterizing the hygroscopicity of mixtures and its role in accelerating hydrolytic degradation.

3. Artificial Intelligence in Preformulation

3.1. Limitations of Empirical Screening

Empirical screening, though considered the gold standard for regulatory submissions, exhibits several systemic bottlenecks. The primary drawback is the substantial time and financial resource investment required to screen dozens of potential excipients against multiple drug candidates [14]. Stressed storage studies require weeks or months to generate meaningful data, which conflicts with the accelerated timelines of early-phase drug development. These assays require relatively large quantities of high-purity API, which is often in short supply during the early synthesis stages. Empirically stressed conditions can also yield false positives; for instance, thermal stress may induce physical transitions (such as excipient melting) that would never occur under ambient storage, leading to the unnecessary rejection of viable excipients. Conventional methods often fail to elucidate the underlying molecular mechanisms driving the interaction, rendering formulation optimization a trial-and-error process.

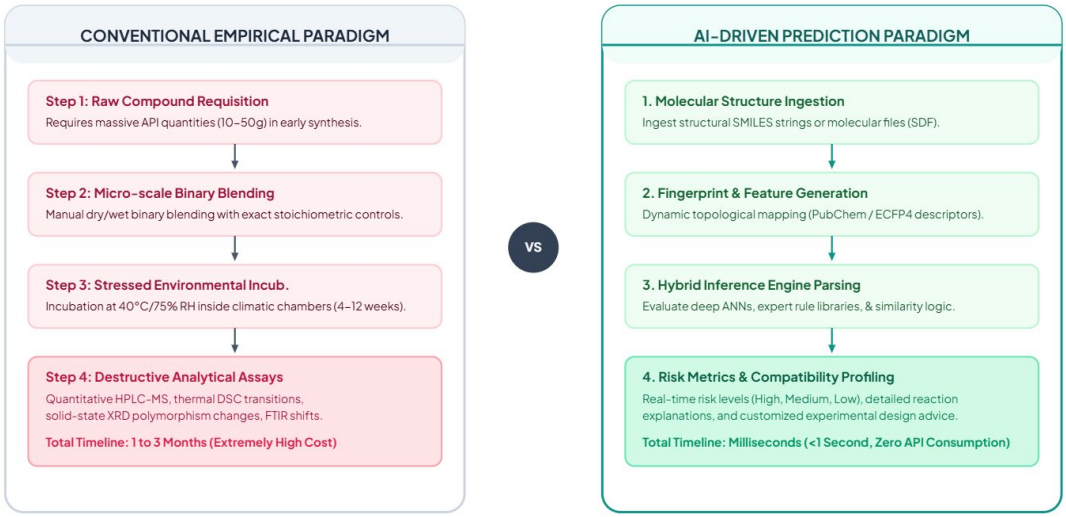

Figure 1. Empirical Screening Vs AI-Driven Prediction Paradigm

3.2. Computational Modeling and Molecular Simulations

To mitigate the limitations of empirical testing, computational chemistry and molecular modeling have emerged as valuable complementary tools. Techniques such as molecular dynamics (MD) simulations and quantum mechanical (QM) calculations allow researchers to simulate drug-excipient interactions at the atomic level. Quantum chemical methods, specifically density functional theory (DFT), can predict the chemical reactivity of functional groups by calculating local properties like molecular electrostatic potentials (MEP) and frontier molecular orbital energies (E_{HOMO} and E_{LUMO}. MD simulations can calculate the thermodynamic binding free energies and solubility parameters (such as Hansen solubility parameters) of drug-polymer mixtures, predicting phase separation, miscibility, and amorphous stability [15]. However, while molecular modeling provides deep structural insights, it is computationally expensive and requires significant expert configuration, making it difficult to use for high-throughput screening of extensive excipient libraries.

3.3. Classification of AI-driven Predictive Methods

Artificial intelligence (AI) and machine learning (ML) offer a highly efficient, high-throughput alternative to both empirical screening and physics-based simulations. Rather than solving complex physical equations or waiting for physical degradation to occur, AI models recognize complex patterns within existing historical datasets to predict compatibility. These computational frameworks can be broadly classified into three categories: data-driven machine learning models (such as deep artificial neural networks), knowledge-based expert systems (incorporating rule-based logic), and hybrid models that combine sequence, structural, and target formulation parameters. These systems can process thousands of candidate pairs in milliseconds by converting molecular structures into numerical descriptors (such as binary fingerprints or graph embeddings) guiding experimental design and focusing empirical resources on high-probability formulations.

4. Machine-Learning Binary Classification

4.1. Deep Artificial Neural Network

The DE-Interact computational platform utilizes a deep artificial neural network (ANN) designed to function as a high-throughput binary classifier for small-molecule drug-excipient pairs [16]. Unlike shallow machine learning architectures that fail to capture complex non-linear chemical structural interactions, DE-Interact implements a multi-layered Feed-Forward Artificial Neural Network (FFANN). This network is engineered to process high-dimensional molecular representations, mapping them to definitive thermodynamic stability outcomes. The underlying computational architecture acts as a digital surrogate for conventional thermal and spectroscopic screenings by identifying functional group features that lead to physical or chemical decomposition.

4.2. PubChem Fingerprint Feature Extraction and Mathematical Concatenation

The initial phase of the DE-Interact workflow requires converting chemical structures into a standardized machine-readable format. This digitization is achieved through the calculation of PubChem molecular fingerprints, which serve as binary structural descriptors [17]. The PubChem fingerprint consists of an 881-bit binary vector:

$$\mathbf{f} \in \{0,1\}^{881}$$

Each position in this vector represents the presence (denoted by 1) or absence (denoted by 0) of a specific structural or chemical feature. These features are hierarchically organized across distinct levels:

- Element counts, which track the abundance of specific atomic species within the molecular framework.
- Ring systems, which define the presence of isolated or fused aromatic, heteroaromatic, and saturated cyclic systems.
- Atom pairs, which characterize the spatial and topological connectivity between pairs of atoms within the molecule.
- Atom neighborhoods, which delineate the localized electronic and steric environments surrounding individual atoms.
- SMARTS patterns, which identify specific, highly reactive substructures and functional arrangements.

To evaluate the compatibility of a specific drug-excipient pair, the model performs a vector concatenation of their individual 881-bit fingerprints. Let f_2 represent the fingerprint of the active pharmaceutical ingredient, and let $f_{excipient}$ represent the fingerprint of the candidate excipient. The combined input vector x_{input} is mathematically expressed as:

$$\mathbf{x}_{\text{input}} = [\mathbf{f}_{\text{drug}} \,\|\, \mathbf{f}_{\text{excipient}}] \in \{0,1\}^{1762}$$

This concatenation yields a 1762-dimensional binary vector representing the structural interface of the two compounds. This vector is then fed directly into the input layer of the neural network.

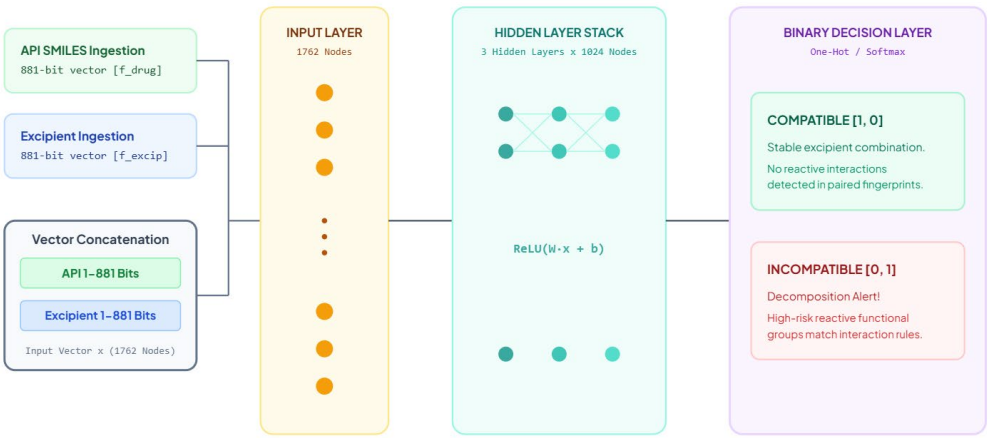

Figure 2. Neural Network Schematic and Feature Processing Stream of the DE-Interact Deep Binary Classifier

4.3. Hidden Layer Parameterization and Optimization

The internal architecture of the DE-Interact network consists of an input layer of 1762 nodes, three hidden layers containing 1024 nodes each, and a binary output layer. The hidden layers perform non-linear feature extraction, identifying complex correlations between different bits in the concatenated vector. Mathematically, the forward propagation through the hidden layers is governed by the following system of equations:

$$z^{(1)} = W^{(1)}x_{input} + b^{(1)}$$

$$a^{(1)} = \mathrm{ReLU}\left(z^{(1)}\right)$$

$$z^{(2)} = W^{(2)}a^{(1)} + b^{(2)}$$

$$a^{(2)} = \mathrm{ReLU}\left(z^{(2)}\right)$$

$$z^{(3)} = W^{(3)}a^{(2)} + b^{(3)}$$

$$a^{(3)} = \mathrm{ReLU}\left(z^{(3)}\right)$$

where $W^{(l)}$ represents the weight matrix for layer l, $b^{(l)}$ represents the bias vector, and ReLU is the Rectified Linear Unit activation function, defined as $\mathrm{ReLU}(v) = \max(0, v)$ to prevent gradient saturation.

The output layer utilizes one-hot encoding to provide a clear binary classification of compatibility. The raw outputs from the final hidden layer are transformed into probabilities using the Softmax activation function:

$$\hat{y} = \mathrm{Softmax}\left(W^{(out)}a^{(3)} + b^{(out)}\right) = \left[p_{\mathrm{compatible}}, p_{\mathrm{incompatible}}\right]$$

An output of [1, 0] indicates thermodynamic compatibility, whereas [0, 1] signals a predicted incompatibility.

To address the challenge of class imbalance since historical literature contains significantly more compatible pairs (3549) than incompatible ones (179) the model is trained using a weighted binary cross-entropy loss function with logits [18]:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[w \cdot y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

where w is a class-weighting factor designed to penalize false negatives. Optimization of the network parameters is achieved via the Root Mean Square Propagation (RMSprop) algorithm over a maximum of 200 epochs. RMSprop dynamically adjusts the learning rate for each parameter by maintaining a running average of the squared gradients, which helps stabilize convergence across complex loss landscapes:

$$v_t = \beta v_{t-1} + (1 - \beta)g_t^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \epsilon}}g_t$$

Here, θ represents the model parameters, g_t is the gradient at step t, β is the decay factor, η is the base learning rate and ε is a smoothing term to prevent division by zero.

4.4. Experimental Validation and Analytical Correlation

The predictive accuracy of the finalized DE-Interact model has been validated using characterized drug-excipient systems, including paracetamol–vanillin, paracetamol–methylparaben, and brinzolamide–polyethylene glycol [16]. During validation, the model predicted an incompatibility for the paracetamol-vanillin binary system based on the combined fingerprints. Subsequent physical characterization using Differential Scanning Calorimetry (DSC) and Fourier Transform Infrared Spectroscopy (FTIR) confirmed

this prediction. DSC thermograms of the paracetamol-vanillin mixture showed a significant shift and broadening of the endothermic melting peak, indicating a physical interaction or mutual dissolution. FTIR spectra confirmed this by showing the disappearance of the characteristic phenolic hydroxyl stretching bands and the appearance of shifted carbonyl bands, indicating robust hydrogen bonding interactions. This correlation between computational predictions and physical data highlights the potential of DE-Interact to replace extensive, labor-intensive lab testing during early formulation screening.

5. Knowledge-Driven Inference and Risk Evaluation Via PharmDE

5.1. Hybrid System

PharmDE is designed as a hybrid expert system that combines a curated relational database with a rule-based inference engine [19]. This structure reflects the principle that chemical structure determines physical and chemical properties. Instead of relying solely on statistical associations, PharmDE uses established chemical rules to predict incompatibility risks. The system includes four core modules: an incompatibility case database, a chemical substance information registry, a specialized rule base, and a RDKit-driven inference engine. This combination allows PharmDE to perform structural searches, evaluate compatibility, and provide detailed explanations of identified risks.

5.2. Compilation and Curatorial Criteria of the Incompatibility Database

The foundation of the PharmDE platform is its incompatibility database, which was constructed by curating hundreds of experimental compatibility studies [19]. The database contains 532 high-fidelity, manually curated data items involving 200 unique active pharmaceutical ingredients and 123 structurally diverse excipients. This information was compiled from 228 peer-reviewed research papers and supplemented with data from the *Handbook of Pharmaceutical Excipients* [20].

Each entry in the database includes detailed experimental metadata, such as the analytical methods used (e.g., DSC, HPLC, FTIR, or XRPD), the specific storage conditions (temperature, humidity, and duration), and the underlying degradation mechanisms (such as hydrolysis, oxidation, or Maillard reactions). To ensure chemical metadata consistency, each API and excipient entry is linked to its Simplified Molecular Input Line Entry System (SMILES) notation, molecular weight, hydrogen bond donor and acceptor counts, and logP values retrieved from PubChem and ChEBL.

5.3. The Rule-Based Logic Engine

The predictive engine of PharmDE relies on a rule base consisting of 60 expert-defined chemical interaction rules and 22 characteristic substructural alerts [19]. These rules represent established chemical reactivities that lead to degradation. The system covers 17 distinct interaction categories, including:

- The Maillard reaction, which occurs between primary or secondary amine drugs and reducing sugar excipients like lactose or dextrose.
- Acylation reactions, where nucleophilic drugs (e.g., those containing primary amines) attack ester or carbonate groups in excipients.
- Transesterification, involving ester-containing APIs and alcohol-bearing polymers like polyethylene glycol.
- Acid-base catalyzed hydrolysis, which affects APIs with esters, lactams, or amides when combined with acidic or basic excipient salts.
- Metal-catalyzed oxidation, triggered by trace transition metal impurities in excipients like talc or dibasic calcium phosphate.

These chemical rules are written as SMARTS (SMILES Arbitrary Target Specification) patterns. This format allows the system to identify reactive functional groups within the molecular structures of both the drug and the excipient.

5.4. Determination of Multi-Tiered Risk

The inference engine is written in Python and uses the RDKit chemistry library to perform substructure matching [21]. When a user enters a formulation containing an API and one or more excipients, the engine runs a dual-pathway evaluation.

Table 3. Primary Chemical Incompatibility and Reactive Functional Groups

Degradation Pathway	Primary Reactive API Functional Group	Interfering Excipient Functional Group / Impurity	Representative Degradation Mechanism	Strategic Formulation Mitigation
Maillard Reaction	Primary or secondary aliphatic amines (e.g., Fluoxetine, Atenolol)	Reducing sugars containing active hemiacetal/ketal groups (e.g., Lactose, Dextrose)	Nucleophilic attack of the amine on the carbonyl carbon, followed by dehydration to form a Schiff base and subsequent Amadori rearrangement.	Replacement of reducing sugars with non-reducing polyols (e.g., Mannitol, Isomalt) or microcrystalline cellulose.
Nucleophilic Acylation	Primary aliphatic amines or hydrazine groups	Esters, lactones, or polymeric carbonate impurities (e.g., PVP, Crospovidone residues)	Nucleophilic substitution at the carbonyl center of the excipient, resulting in the formation of stable covalent amide or imide adducts.	Selection of high-purity, low-monomer excipient grades; formulation acidification to protonate the amine.
Transesterification	Ester groups (e.g., Acetylsalicylic acid, Methylphenidate)	Hydroxyl-rich polymers or co-solvents (e.g., Polyethylene glycol, Glycerol, Sorbitol)	Alcoholysis where the excipient hydroxyl group acts as a nucleophile, replacing the alkoxy group of the drug's ester.	Elimination of liquid-state PEG co-solvents; maintaining low microenvironmental relative humidity (<40%).
Hydrolytic Cleavage	Esters, amides, lactams, or imides (e.g., Penicillins, Cephalosporins)	High acidic or basic microenvironmental pH, or high-hygroscopicity excipients	Catalytic proton or hydroxyl ion attack on the carbonyl carbon, accelerating hydrolytic cleavage under high water activity.	Inclusion of microenvironmental pH modifiers (e.g., Citric acid, Sodium carbonate) to keep pH at the point of maximum stability.
Autoxidation	Thioethers, phenols, tertiary amines, unsaturated alkenes	Residual peroxides (ROOH) or trace transition metal ions (Fe^{3+} , Cu^{2+} in excipients (e.g., Polysorbates)	Metal-catalyzed homolytic cleavage of peroxides to form alkoxy and alkyl radicals, which initiate free-radical chain oxidation of the API.	Addition of chelating agents (e.g., EDTA) and chain-breaking antioxidants (e.g., BHT, Propyl gallate); selection of peroxide-free excipient grades.

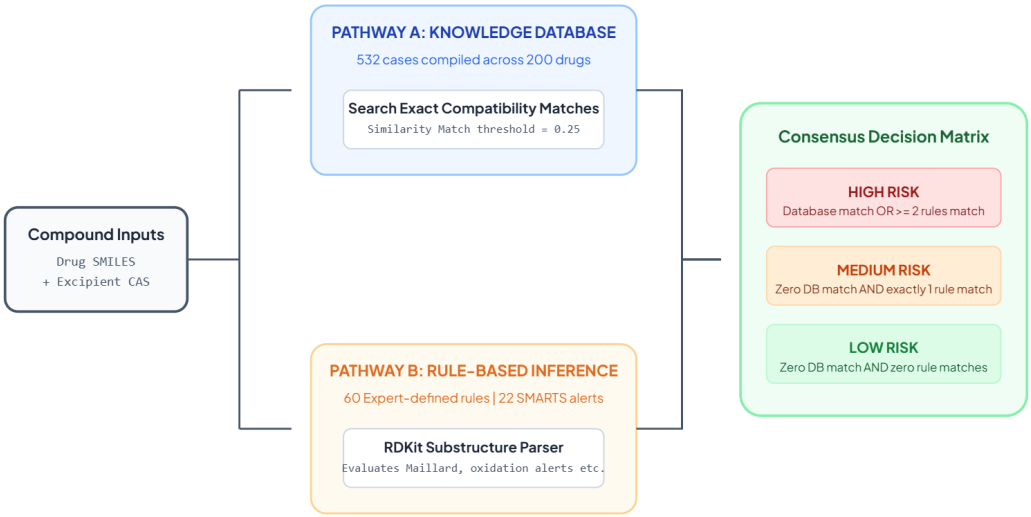

Figure 3. Dual-Pathway Expert System Logic Flow of the PharmDE Hybrid Decision Engine

The final risk recommendation is determined by combining the database search results and the rule matches:

- High Risk: Assigned if the specific drug-excipient combination is already documented as incompatible in the database, or if the substructure search matches two or more chemical interaction rules. This classification indicates a strong probability of degradation, suggesting the excipient should be avoided.
- Medium Risk: Assigned when there is no recorded incompatibility in the database, but the molecular structures match exactly one interaction rule. This suggests a potential risk, and researchers are advised to proceed with targeted empirical validation.
- Low Risk: Assigned when there are no database records of incompatibility and no matching structural interaction rules, indicating a high probability of formulation stability.

This structured risk assessment helps formulation scientists identify potential incompatibilities early, enabling a more targeted and efficient experimental design.

6. Macromolecular Prediction Via ExPreSo

6.1. Challenges in Characterizing Peptide and Protein Co-Formulations

Predicting the stability of macromolecular therapeutics, such as proteins and peptides, presents unique challenges compared to small-molecule drugs [22]. Proteins are larger and structurally complex, making them sensitive to environmental factors like temperature, shear stress, and interfacial tension. Their degradation pathways are primarily physical including unfolding, self-association, and aggregation rather than purely chemical.

Protein-excipient interactions are typically weak and transient, driven by a balance of electrostatic, hydrophobic, and van der Waals forces across the protein's surface. Physics-based modeling, such as all-atom Molecular Dynamics (MD) simulations, is computationally expensive and difficult to scale for high-throughput screening of large excipient libraries. This complexity requires specialized predictive tools designed for the unique structural characteristics of biopharmaceuticals.

6.2. High-Dimensional Feature Space and Protein Language Model

To address these challenges, the ExPreSo (Excipient Prediction Software) tool uses a high-dimensional feature space to characterize the protein, the excipients, and the target product profile [23]. Rather than relying on simple composition metrics, ExPreSo uses advanced protein representation techniques:

The protein's primary structure is encoded using sequence-based descriptors and deep learning representations. Sequence descriptors include amino acid composition, dipeptide frequencies, and basic biophysical properties like the theoretical isoelectric point (pI). To capture complex structural and evolutionary context, the model uses the ProtT5-XL-U50 protein language model [24]. This model, trained on billions of protein sequences, converts each primary sequence into a 1024-dimensional vector:

$$\mathbf{v}_{\mathrm{pLM}} \in \mathbb{R}^{1024}$$

These embeddings encode structural, functional, and biophysical properties without requiring explicit 3D coordinates.

To capture surface properties that drive excipient interactions, ExPreSo incorporates structural features derived from three-dimensional models. These 3D structures are generated using AlphaFold2 for non-antibody proteins, and the Molecular Operating Environment (MOE) Antibody Modeler for monoclonal antibodies (mAbs) [25]. From these structures, the tool calculates localized surface properties, including:

- The total area and spatial distribution of hydrophobic surface patches, which influence self-association and surfactant binding.
- Positive, negative, and overall charge distributions, which dictate pH-dependent electrostatic interactions.
- Dipole moments, which reflect the protein's overall charge asymmetry.

The model combines the attributes from the Target Product Profile (TPP), such as formulation pH, target protein concentration, dosage form (liquid or lyophilized), and the route of administration, ensuring the predictions align with commercial and clinical requirements.

Table 4. High-Dimensional Multi-Scale Features for ExPreSo Protein

Feature Domain	Specific Descriptor Categories	Dimensionality / Mathematical Format	Physical / Biophysical Representation	Influence on Excipient Preferential Interaction
Primary Sequence	Amino Acid Composition & Dipeptide Frequency	420-dimensional vector ()	Abundance and sequential order of polar, hydrophobic, acidic, basic, and aromatic residues.	Determines the protein's overall charge-to-hydrophobicity balance and structural stability.
Protein Language Model (pLM)	ProtT5-XL-U50 Transformer Embeddings	1024-dimensional continuous vector ()	Contextual, evolutionary, and structural representations of the protein sequence.	Captures complex, non-obvious sequence-structure-function relationships that influence stabilization.
3D Structural Geometry	Spatial Coordinates & Solvent Accessible Surface Area (SASA)	Scalar values and spatial density functions	Three-dimensional fold, structural stability, and surface exposure of reactive residues.	Identifies specific surface patches available for physical or chemical interactions.
Surface Biophysics	Hydrophobic Patches & Charge Distribution	3D patch map descriptors	Spatial clustering of hydrophobic residues and localized positive/negative electrostatic potentials.	Directs the binding of surfactants (e.g., Polysorbates) and buffers to prevent aggregation.
Target Product Profile (TPP)	pH, Protein Concentration, and Dosage Form	Categorical and continuous values	Environmental variables: solution pH, concentration (), and physical state (Liquid vs. Lyo).	Dictates the thermodynamic stability landscape, guiding the choice of tonicity modifiers and cryoprotectants.

6.3. The Extra Trees Ensemble Classifier and Dimensionality Reduction

The core predictive engine of ExPreSo is built on the Extra Trees (Extremely Randomized Trees) ensemble learning algorithm [26]. Extra Trees is an ensemble method that fits multiple decision trees on various sub-samples of the dataset. It introduces additional randomness when splitting node-level features, which helps control variance and prevents overfitting on relatively small biopharmaceutical datasets.

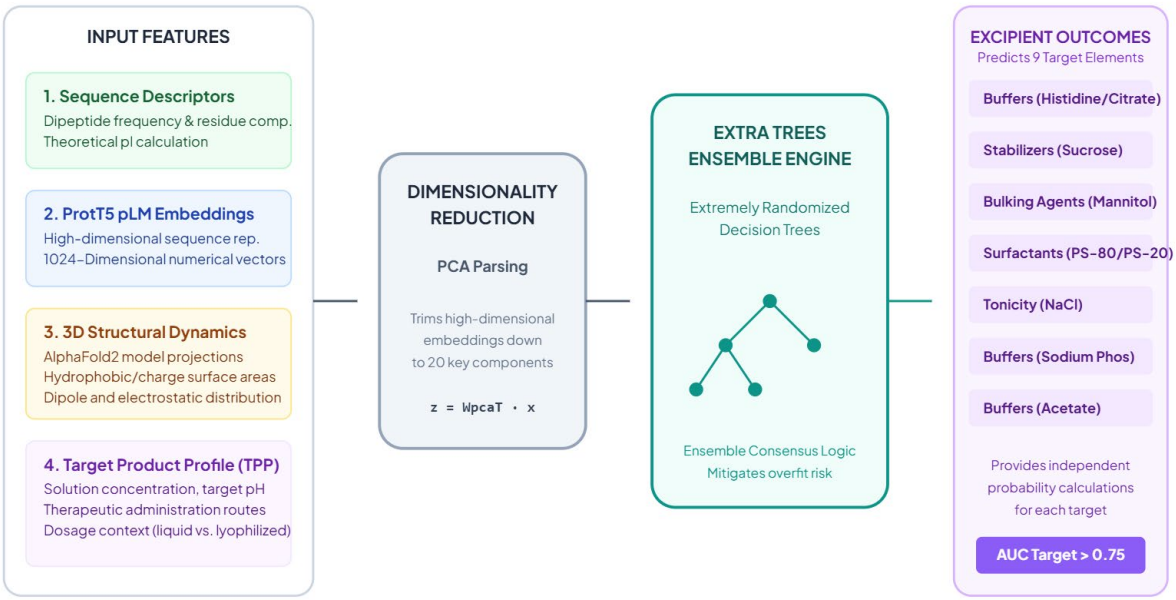

Figure 4. Multi-Scale Feature Integration of the ExPreSo Prediction Engine

To manage the high dimensionality of the combined feature vector (which includes sequence, structural, pLM embeddings, and TPP attributes), ExPreSo uses Principal Component Analysis (PCA) for feature extraction and dimensionality reduction.

This process reduces the high-dimensional feature space down to the top 20 principal components:

$$z = W_{PCA}^{T} x_{combined}$$

where W_{PCA} represents the projection matrix containing the principal axes. This feature reduction ensures the model remains robust and generalizable, focusing on the features that explain the most variance in excipient compatibility.

6.4. Operational Variants and Predictive Accuracy

ExPreSo was developed in four distinct variants to accommodate different stages of the preformulation workflow:

- The Fast Model, which relies solely on primary sequence data. This model can generate predictions in milliseconds, making it suitable for rapid, early-stage screening.
- The Protein-Based Model, which uses intrinsic protein features (sequence, structural models, and embeddings) without requiring specific formulation parameters. This allows researchers to evaluate a protein's baseline stability profile.
- The All Features Model, which integrates all sequence, structural, and target product profile inputs to deliver highly personalized predictions for a specific formulation scenario.
- The Interpretable Model, which prioritizes human-understandable biophysical properties (such as charge, hydrophobic patches, and pH) over high-dimensional embeddings, allowing researchers to gain mechanistic insights into why certain excipients are recommended.

The performance of these models has been validated using Monte Carlo and Leave-One-Group-Out (LOGO) cross-validation [23]. Evaluated using the Area Under the Receiver Operating Characteristic (ROC-AUC) and Precision-Recall (PR-AUC) metrics, most excipient models achieved ROC-AUC values greater than 0.75. This high level of accuracy shows the ability of the model to identify suitable stabilizers, buffers, and surfactants, helping to streamline the development of stable protein formulations.

7. Comparison of Advanced Computational Prediction

While DE-Interact, PharmDE, and ExPreSo all aim to accelerate preformulation development, they differ significantly in their underlying logic, training data, input requirements, and clinical target profiles. These distinctions are summarized below:

Table 5. Comparison of Advanced Computational Prediction

Feature Dimension	DE-Interact	PharmDE	ExPreSo
Primary Target Class	Small-molecule drugs and functional excipients.	Small-molecule drugs and diverse excipients.	Macromolecular protein and peptide therapeutics.
Core Architecture	Data-driven feed-forward Artificial Neural Network (ANN).	Hybrid expert system with rule-based inference.	Supervised ensemble machine learning (Extra Trees).
Underlying Data	3,728 curated drug-excipient pairs.	532 literature cases, 60 expert rules, 22 substructures.	335 FDA-approved biopharmaceutical formulations.
Input Representation	Concatenated 1762-bit PubChem fingerprints.	Molecular SMILES, names, or CAS registry numbers.	Sequence, ProtT5 embeddings, 3D structural models.
Primary Methodology	Non-linear feature extraction and deep learning.	SMARTS substructure search and database matching.	PCA dimensionality reduction and ensemble trees.
Model Output	Binary classification: Compatible vs. Incompatible.	Three-tiered risk level: Low, Medium, or High Risk.	Excipient occurrence and compatibility probability.
Interpretability	Moderate (requires post-hoc feature attribution).	High (maps directly to chemical interaction rules).	Moderate to High (via biophysical feature tracking).

These three systems represent complementary approaches to preformulation. DE-Interact is optimized for rapid, high-throughput binary screening of novel small molecules. PharmDE provides detailed, rule-based risk assessments based on established chemical reactivities, which is valuable for identifying specific degradation pathways. ExPreSo is tailored for biopharmaceuticals, where stabilization involves managing complex physical aggregation and self-association pathways rather than simple chemical reactivities. Together, these tools provide a suite of computational methods to guide and streamline formulation design.

8. Regulatory Challenges and Limitations

8.1. Limitations of Current Predictive Systems

Despite their advantages, current computational models for predicting drug-excipient compatibility face several systemic limitations. The most significant challenge is data constraint; machine learning models are inherently limited by the quality, size, and diversity of their training datasets [27]. Historical literature is heavily biased toward successful, stable formulations, leaving a shortage of negative data (documented incompatibilities) to train models effectively. Additionally, these models typically focus on binary mixtures, whereas commercial formulations are multi-component systems. They rarely account for ternary or quaternary interactions, where the presence of a third component can significantly alter stability. Current models also struggle to capture the impact of physical processing parameters such as compaction force, wet granulation moisture levels, or local shear stress which can induce physical transitions or accelerate chemical degradation. Finally, these models generally do not predict the optimal concentrations of excipients or account for batch-to-batch variability in excipient purity, such as trace peroxide levels in polymeric surfactants.

8.2. Regulatory Compliance and Quality by Design (QbD)

Integrating artificial intelligence into pharmaceutical development requires careful alignment with regulatory frameworks, particularly Quality by Design (QbD) principles as outlined in ICH guidelines Q8, Q9, and Q10 [28]. Regulatory bodies, including the FDA and EMA, require that any predictive tool used in drug development be validated, reproducible, and explainable. Under GAMP 5 (Good Automated Manufacturing Practice) guidelines, computer software must undergo rigorous qualification to ensure data integrity and algorithmic reliability [29]. For machine learning models, this requires addressing the "black-box" nature of deep neural networks. Developing interpretable models, such as those utilizing SHAP (SHapley Additive exPlanations) or attention mechanisms, is essential to explain how a model arrived at a specific risk prediction [30]. Establishing these validation standards is critical for computational tools to be accepted as part of formal regulatory filings, such as Investigational New Drug (IND) or New Drug Applications (NDA).

8.3. Forward-Looking Research Directions

To overcome these limitations, future research should focus on expanding high-fidelity, public-access databases that include both stable and unstable formulation outcomes. Developing advanced graph neural networks (GNNs) that represent molecules as 3D stereochemical graphs could improve prediction accuracy compared to 1D sequence or 2D fingerprint representations [31]. Another promising area is the integration of multi-scale modeling, combining deep learning with quantum mechanics and molecular dynamics. In this hybrid approach, machine learning models can perform rapid initial screenings of extensive excipient libraries, while physics-based simulations provide detailed thermodynamic assessments of the top-ranked candidates. As computational power increases and datasets expand, these integrated platforms will become increasingly valuable for guiding rational, automated formulation design.

9. Conclusion

Characterizing drug–excipient compatibility is an important part of pharmaceutical preformulation, directly impacting the stability, safety, and clinical efficacy of drug products. While traditional empirical screening methods are resource-intensive, the emergence of artificial intelligence offers a more efficient, predictive alternative. Computational tools like DE-Interact, PharmDE, and ExPreSo show the utility of in silico modeling across diverse formulation scenarios. These platforms can rapidly identify potential incompatibilities by leveraging artificial neural networks, expert rules, and machine learning models trained on sequence and structural features. Although challenges remain regarding dataset limitations, multi-component modeling, and regulatory integration, continuous refinement of these computational frameworks will be important in transitioning from empirical, trial-and-error workflows to automated, rational formulation design. This transition is poised to significantly accelerate drug development timelines and reduce costs, ultimately supporting the delivery of stable and effective therapeutics.

References

[1] Panakanti R, Narang AS. Impact of excipient interactions on drug bioavailability from solid dosage forms. Pharm Res. 2012;29(9):2639-2659.

[2] Patel P, Ahir K, Patel V, Manani L, Patel C. Drug-excipient compatibility studies: first step for dosage form development. Pharma Innov J. 2015;4(5):14-20.

[3] Dave VS, Boyce H, Al-Achi A. Drug-excipient compatibility studies in formulation development: current trends and techniques. AAPS Formul Des Dev Sect Newsl. 2015;9:1-6.

[4] Shah HS, Chaturvedi K, Kuang S, Wang J. Accelerating pre-formulation investigations in early drug product life cycles using predictive methodologies and computational algorithms. Ther Deliv. 2021;12(11):789-797.

[5] Sheskey PJ, Hancock BC, Moss GP, Goldfarb DJ, editors. Handbook of Pharmaceutical Excipients. 9th ed. London: Pharmaceutical Press; 2020.

[6] Connors KA, Amidon GL, Stella VJ. Chemical Stability of Pharmaceuticals: A Handbook for Pharmacists. 2nd ed. New York: John Wiley & Sons; 1986.

[7] Waterman KC, Adami RC, Alsante KM, Antipas AS, Arenson DR, Carrier R, et al. Hydrolysis in pharmaceutical formulations. Pharm Dev Technol. 2002;7(2):113-146.

[8] Crowley PJ. Physical, chemical, and biopharmaceutical characterization of a new drug substance. In: Adeyeye MC, Schacht JH, editors. Preformulation in Solid Dosage Form Development. New York: Informa Healthcare; 2008. p. 115-168.

[9] Jackson K, Young D, Pant S. Experimental protocols for solid-state characterization of active pharmaceutical ingredients. J Pharm Sci. 2013;102(4):1123-1135.

[10] Yoshioka S, Stella VJ. Stability of Drugs and Dosage Forms. New York: Kluwer Academic Publishers; 2002.

[11] Allen LV, Popovich NG, Ansel HC. Ansel's Pharmaceutical Dosage Forms and Drug Delivery Systems. 11th ed. Philadelphia: Wolters Kluwer; 2018.

[12] International Conference on Harmonisation (ICH). ICH Harmonised Tripartite Guideline: Stability Testing of New Drug Substances and Products Q1A(R2). Geneva: ICH; 2003.

[13] Byrn SR, Pfeiffer RR, Stowell JG. Solid-State Chemistry of Drugs. 2nd ed. West Lafayette: SSCI Inc; 1999.

[14] Florence AT, Attwood D. Physicochemical Principles of Pharmacy: In Manufacture, Formulation and Clinical Use. 6th ed. London: Pharmaceutical Press; 2016.

[15] Cloutier TK, Sudrik C, Mody N, Sathish H, Trout BL. Machine learning models of antibody-excipient preferential interactions for use in computational formulation design. Mol Pharm. 2020;17(9):3589-3599.

[16] Patel S, Patel M, Kulkarni M, Patel MS. 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. Int J Pharm. 2023;637:122839.

[17] Dong J, Cao DS, Miao HY, Liu S, Deng BC, Yun YH, et al. ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation. J Cheminform. 2015;7:60.

[18] Vora LK, Gholap AD, Jetha K, Thakur RRS, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics. 2023;15(7):1916.

[19] Wang N, Sun H, Dong J, Ouyang D. PharmDE: a new expert system for drug-excipient compatibility evaluation. Int J Pharm. 2021;607:120962.

[20] Sheskey PJ, Cook WG, Cable CG, editors. Handbook of Pharmaceutical Excipients. 8th ed. London: Pharmaceutical Press; 2017.

[21] Landrum G. RDKit: Open-source cheminformatics software. 2016. Available from: https://www.rdkit.org.

[22] Barata TS, Zhang C, Dalby PA, Brocchini S, Zloh M. Identification of protein-excipient interaction hotspots using computational approaches. Int J Mol Sci. 2016;17(6):853.

[23] Vidal-Henriquez E, Holder T, Lee NF, Pompe C, Teese MG. Machine learning driven acceleration of biopharmaceutical formulation development using Excipient Prediction Software (ExPreSo). Comput Struct Biotechnol J. 2025;27:4517-4525.

[24] Navarro S, Ventura S. Computational methods to predict protein aggregation. Curr Opin Struct Biol. 2022;73:102343.

[25] Chemical Computing Group. Molecular Operating Environment (MOE). 2024. Available from: https://www.chemcomp.com.

[26] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3-42.

[27] Sultana A, Maseera R, Rahamanulla A, Misiriya A. Emergence of artificial intelligence and technology in pharmaceuticals: a review. Future J Pharm Sci. 2023;9(1):65.

[28] International Conference on Harmonisation (ICH). ICH Guidelines Q8 (Pharmaceutical Development), Q9 (Quality Risk Management), Q10 (Pharmaceutical Quality System). Geneva: ICH; 2009.

[29] International Society for Pharmaceutical Engineering (ISPE). GAMP 5: A Risk-Based Approach to Compliant GxP Computerized Systems. Tampa: ISPE; 2008.

[30] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017. p. 4765-4774.

[31] Dong J, Wu Z, Xu H, Ouyang D. FormulationAI: a novel web-based platform for drug formulation development. Brief Bioinform. 2024;25(1):bbad419.